DISCOVERING STATISTICS USING IBM SPSS STATISTICS

6TH EDITION



ANDY FIELD

DISCOVERING STATISTICS USING IBM SPSS STATISTICS

CATISFIED CUSTOMERS

WWW.INSTAGRAM.COM/DISCOVERING_CATISTICS/

























DISCOVERING STATISTICS USING IBM SPSS STATISTICS

6TH EDITION

ANDY FIELD

Sage



1 Oliver's Yard
55 City Road
London EC1Y 1SP

2455 Teller Road
Thousand Oaks
California 91320

Unit No 323-333, Third Floor, F-Block
International Trade Tower
Nehru Place, New Delhi – 110 019

8 Marina View Suite 43-053
Asia Square Tower 1
Singapore 018960

© Andy Field 2024

Throughout the book, screen shots and images from IBM® SPSS® Statistics software ('SPSS') are reprinted courtesy of International Business Machines Corporation, © International Business Machines Corporation. SPSS Inc. was acquired by IBM in October 2009.

Apart from any fair dealing for the purposes of research, private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, this publication may not be reproduced, stored or transmitted in any form, or by any means, without the prior permission in writing of the publisher, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publisher.

Editor: Jai Seaman
Production editor: Ian Antcliff
Copyeditor: Richard Leigh
Proofreader: Richard Walshe
Marketing manager: Ben Griffin-Sherwood
Cover design: Wendy Scott
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in Italy

Library of Congress Control Number: 2023939237

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-5296-3001-5
ISBN 978-1-5296-3000-8 (pbk)

At Sage we take sustainability seriously. Most of our products are printed in the UK using responsibly sourced papers and boards. When we print overseas we ensure sustainable papers are used as measured by the Paper Chain Project grading system. We undertake an annual audit to monitor our sustainability.

CONTENTS

PREFACE xv
HOW TO USE THIS BOOK xxi
THANK YOU xxv
SYMBOLS USED IN THIS BOOK xxviii
A BRIEF MATHS OVERVIEW xxx

1 WHY IS MY EVIL LECTURER FORCING ME TO LEARN STATISTICS? 1
2 THE SPINE OF STATISTICS 49
3 THE PHOENIX OF STATISTICS 105
4 THE IBM SPSS STATISTICS ENVIRONMENT 149
5 VISUALIZING DATA 197
6 THE BEAST OF BIAS 243
7 NON-PARAMETRIC MODELS 321
8 CORRELATION 373
9 THE LINEAR MODEL (REGRESSION) 411
10 CATEGORICAL PREDICTORS: COMPARING TWO MEANS 477
11 MODERATION AND MEDIATION 521
12 GLM 1: COMPARING SEVERAL INDEPENDENT MEANS 557
13 GLM 2: COMPARING MEANS ADJUSTED FOR OTHER PREDICTORS (ANALYSIS OF COVARIANCE) 615
14 GLM 3: FACTORIAL DESIGNS 651
15 GLM 4: REPEATED-MEASURES DESIGNS 695
16 GLM 5: MIXED DESIGNS 751
17 MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA) 783
18 EXPLORATORY FACTOR ANALYSIS 823
19 CATEGORICAL OUTCOMES: CHI-SQUARE AND LOGLINEAR ANALYSIS 883
20 CATEGORICAL OUTCOMES: LOGISTIC REGRESSION 925
21 MULTILEVEL LINEAR MODELS 979

EPILOGUE 1029
APPENDIX 1033
GLOSSARY 1045
REFERENCES 1079
INDEX 1095

EXTENDED CONTENTS

PREFACE xv
HOW TO USE THIS BOOK xxi
THANK YOU xxv
SYMBOLS USED IN THIS BOOK xxviii
A BRIEF MATHS OVERVIEW xxx

1 WHY IS MY EVIL LECTURER FORCING ME TO LEARN STATISTICS? 1
1.1 What the hell am I doing here? I don't belong here 3
1.2 The research process 3
1.3 Initial observation: finding something that needs explaining 4
1.4 Generating and testing theories and hypotheses 5
1.5 Collecting data: measurement 9
1.6 Collecting data: research design 16
1.7 Analysing data 22
1.8 Reporting data 40
1.9 Jane and Brian's story 43
1.10 What next? 45
1.11 key terms that I've discovered 45
Smart Alex's tasks 46

2 THE SPINE OF STATISTICS 49
2.1 What will this chapter tell me? 50
2.2 What is the SPINE of statistics? 51
2.3 Statistical models 51
2.4 Populations and samples 55
2.5 The linear model 56
2.6 P is for parameters 58
2.7 E is for estimating parameters 66
2.8 S is for standard error 71
2.9 I is for (confidence) interval 74
2.10 N is for null hypothesis significance testing 81
2.11 Reporting significance tests 100
2.12 Jane and Brian's story 101
2.13 What next? 103
2.14 Key terms that I've discovered 103
Smart Alex's tasks 103

3 THE PHOENIX OF STATISTICS 105
3.1 What will this chapter tell me? 106
3.2 Problems with NHST 107
3.3 NHST as part of wider problems with science 117
3.4 A phoenix from the EMBERS 123
3.5 Sense, and how to use it 124
3.6 Preregistering research and open science 125
3.7 Effect sizes 126
3.8 Bayesian approaches 135
3.9 Reporting effect sizes and Bayes factors 145
3.10 Jane and Brian's story 145
3.11 What next? 145
3.12 Key terms that I've discovered 146
Smart Alex's tasks 147

4 THE IBM SPSS STATISTICS ENVIRONMENT 149
4.1 What will this chapter tell me? 150
4.2 Versions of IBM SPSS Statistics 151
4.3 Windows, Mac OS and Linux 151
4.4 Getting started 152
4.5 The data editor 153
4.6 Entering data into IBM SPSS Statistics 157
4.7 SPSS syntax 174
4.8 The SPSS viewer 174
4.9 Exporting SPSS output 182
4.10 Saving files and restore points 182
4.11 Opening files and restore points 184
4.12 A few useful options 185
4.13 Extending IBM SPSS Statistics 187
4.14 Jane and Brian's story 190
4.15 What next? 191
4.16 Key terms that I've discovered 192
Smart Alex's tasks 192

5 VISUALIZING DATA 197
5.1 What will this chapter tell me? 198
5.2 The art of visualizing data 199
5.3 The SPSS Chart Builder 202
5.4 Histograms 205
5.5 Boxplots (box–whisker diagrams) 211
5.6 Visualizing means: bar charts and error bars 214
5.7 Line charts 226
5.8 Visualizing relationships: the scatterplot 227
5.9 Editing plots 236
5.10 Brian and Jane's story 239
5.11 What next? 240
5.12 Key terms that I've discovered 240
Smart Alex's tasks 240

6 THE BEAST OF BIAS 243
6.1 What will this chapter tell me? 244
6.2 Descent into statistics hell 245
6.3 What is bias? 265
6.4 Outliers 266
6.5 Overview of assumptions 269
6.6 Linearity and additivity 271
6.7 Spherical errors 271
6.8 Normally distributed something or other 275
6.9 Checking for bias and describing data 279
6.10 Reducing bias with robust methods 303
6.11 A final note 316
6.12 Jane and Brian's story 317
6.13 What next? 317
6.14 Key terms that I've discovered 319
Smart Alex's tasks 319

7 NON-PARAMETRIC MODELS 321
7.1 What will this chapter tell me? 322
7.2 When to use non-parametric tests 323
7.3 General procedure of non-parametric tests using SPSS 324
7.4 Comparing two independent conditions: the Wilcoxon rank-sum test and Mann–Whitney test 325
7.5 Comparing two related conditions: the Wilcoxon signed-rank test 337
7.6 Differences between several independent groups: the Kruskal–Wallis test 346
7.7 Differences between several related groups: Friedman's ANOVA 359
7.8 Jane and Brian's story 368
7.9 What next? 370
7.10 Key terms that I've discovered 370
Smart Alex's tasks 370

8 CORRELATION 373
8.1 What will this chapter tell me? 374
8.2 Modelling relationships 375
8.3 Data entry for correlation analysis 384
8.4 Bivariate correlation 384
8.5 Partial and semi-partial correlation 397
8.6 Comparing correlations 404
8.7 Calculating the effect size 405
8.8 How to report correlation coefficients 405
8.9 Jane and Brian's story 406
8.10 What next? 409
8.11 Key terms that I've discovered 409
Smart Alex's tasks 409

9 THE LINEAR MODEL (REGRESSION) 411
9.1 What will this chapter tell me? 412
9.2 The linear model (regression) … again! 413
9.3 Bias in linear models 422
9.4 Generalizing the model 425
9.5 Sample size and the linear model 429
9.6 Fitting linear models: the general procedure 431
9.7 Using SPSS to fit a linear model with one predictor 432
9.8 Interpreting a linear model with one predictor 433
9.9 The linear model with two or more predictors (multiple regression) 436
9.10 Using SPSS to fit a linear model with several predictors 441
9.11 Interpreting a linear model with several predictors 447
9.12 Robust regression 466
9.13 Bayesian regression 469
9.14 Reporting linear models 471
9.15 Jane and Brian's story 471
9.16 What next? 474
9.17 Key terms that I've discovered 474
Smart Alex's tasks 475

10 CATEGORICAL PREDICTORS: COMPARING TWO MEANS 477
10.1 What will this chapter tell me? 478
10.2 Looking at differences 479
10.3 A mischievous example 479
10.4 Categorical predictors in the linear model 482
10.5 The *t*-test 484
10.6 Assumptions of the *t*-test 492
10.7 Comparing two means: general procedure 492
10.8 Comparing two independent means using SPSS 493
10.9 Comparing two related means using SPSS 502
10.10 Reporting comparisons between two means 515
10.11 Between groups or repeated measures? 516
10.12 Jane and Brian's story 516
10.13 What next? 517
10.14 Key terms that I've discovered 518
Smart Alex's tasks 518

11 MODERATION AND MEDIATION 521
11.1 What will this chapter tell me? 522
11.2 The PROCESS tool 523
11.3 Moderation: interactions in the linear model 523
11.4 Mediation 539
11.5 Jane and Brian's story 554
11.6 What next? 554
11.7 Key terms that I've discovered 555
Smart Alex's tasks 555

12 GLM 1: COMPARING SEVERAL INDEPENDENT MEANS 557
12.1 What will this chapter tell me? 558
12.2 A puppy-tastic example 559
12.3 Compare several means with the linear model 560
12.4 Assumptions when comparing means 574
12.5 Planned contrasts (contrast coding) 577
12.6 *Post hoc* procedures 590
12.7 Effect sizes when comparing means 592
12.8 Comparing several means using SPSS 592
12.9 Output from one-way independent ANOVA 599
12.10 Robust comparisons of several means 607
12.11 Bayesian comparison of several means 609
12.12 Reporting results from one-way independent ANOVA 610
12.13 Jane and Brian's story 610
12.14 What next? 611
12.15 Key terms that I've discovered 612
Smart Alex's tasks 612

13 GLM 2: COMPARING MEANS ADJUSTED FOR OTHER PREDICTORS (ANALYSIS OF COVARIANCE) 615
13.1 What will this chapter tell me? 616
13.2 What is ANCOVA? 617
13.3 The general linear model with covariates 618
13.4 Effect size for ANCOVA 622
13.5 Assumptions and issues in ANCOVA designs 622
13.6 Conducting ANCOVA using SPSS 627
13.7 Interpreting ANCOVA 634
13.8 The non-parallel slopes model and the assumption of homogeneity of regression slopes 641
13.9 Robust ANCOVA 644
13.10 Bayesian analysis with covariates 645
13.11 Reporting results 647
13.12 Jane and Brian's story 647
13.13 What next? 647
13.14 Key terms that I've discovered 648
Smart Alex's tasks 649

14 GLM 3: FACTORIAL DESIGNS 651
14.1 What will this chapter tell me? 652
14.2 Factorial designs 653
14.3 A goggly example 653
14.4 Independent factorial designs and the linear model 655
14.5 Interpreting interaction plots 659
14.6 Simple effects analysis 662
14.7 *F*-statistics in factorial designs 663
14.8 Model assumptions in factorial designs 668

14.9 Factorial designs using SPSS 669
14.10 Output from factorial designs 675
14.11 Robust models of factorial designs 683
14.12 Bayesian models of factorial designs 687
14.13 More effect sizes 689
14.14 Reporting the results of factorial designs 691
14.15 Jane and Brian's story 692
14.16 What next? 693
14.17 Key terms that I've discovered 693
Smart Alex's tasks 693

15 GLM 4: REPEATED-MEASURES DESIGNS 695
15.1 What will this chapter tell me? 696
15.2 Introduction to repeated-measures designs 697
15.3 Emergency! The aliens are coming! 698
15.4 Repeated measures and the linear model 698
15.5 The ANOVA approach to repeated-measures designs 700
15.6 The *F*-statistic for repeated-measures designs 704
15.7 Assumptions in repeated-measures designs 709
15.8 One-way repeated-measures designs using SPSS 709
15.9 Output for one-way repeated-measures designs 714
15.10 Robust tests of one-way repeated-measures designs 721
15.11 Effect sizes for one-way repeated-measures designs 724
15.12 Reporting one-way repeated-measures designs 725
15.13 A scented factorial repeated-measures design 726
15.14 Factorial repeated-measures designs using SPSS 727
15.15 Interpreting factorial repeated-measures designs 733
15.16 Reporting the results from factorial repeated-measures designs 747
15.17 Jane and Brian's story 747
15.18 What next? 748
15.19 Key terms that I've discovered 748
Smart Alex's tasks 749

16 GLM 5: MIXED DESIGNS 751
16.1 What will this chapter tell me? 752
16.2 Mixed designs 753
16.3 Assumptions in mixed designs 753
16.4 A speed-dating example 754
16.5 Mixed designs using SPSS 756
16.6 Output for mixed factorial designs 762
16.7 Reporting the results of mixed designs 776
16.8 Jane and Brian's story 779
16.9 What next? 780
16.10 Key terms that I've discovered 781
Smart Alex's tasks 781

17 MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA) 783
17.1 What will this chapter tell me? 784
17.2 Introducing MANOVA 785
17.3 The theory behind MANOVA 787
17.4 Practical issues when conducting MANOVA 800
17.5 MANOVA using SPSS 802
17.6 Interpreting MANOVA 804
17.7 Reporting results from MANOVA 810
17.8 Following up MANOVA with discriminant analysis 811
17.9 Interpreting discriminant analysis 814
17.10 Reporting results from discriminant analysis 817
17.11 The final interpretation 818
17.12 Jane and Brian's story 819
17.13 What next? 819
17.14 Key terms that I've discovered 821
Smart Alex's tasks 821

18 EXPLORATORY FACTOR ANALYSIS 823
18.1 What will this chapter tell me? 824
18.2 When to use factor analysis 825
18.3 Factors and components 826
18.4 Discovering factors 832
18.5 An anxious example 841
18.6 Factor analysis using SPSS 847
18.7 Interpreting factor analysis 853
18.8 How to report factor analysis 866
18.9 Reliability analysis 868
18.10 Reliability analysis using SPSS 873
18.11 Interpreting reliability analysis 875
18.12 How to report reliability analysis 878
18.13 Jane and Brian's story 879
18.14 What next? 880
18.15 Key terms that I've discovered 880
Smart Alex's tasks 881

19 CATEGORICAL OUTCOMES: CHI-SQUARE AND LOGLINEAR ANALYSIS 883
19.1 What will this chapter tell me? 884
19.2 Analysing categorical data 885
19.3 Associations between two categorical variables 885
19.4 Associations between several categorical variables: loglinear analysis 895
19.5 Assumptions when analysing categorical data 897
19.6 General procedure for analysing categorical outcomes 898
19.7 Doing chi-square using SPSS 899
19.8 Interpreting the chi-square test 902

19.9 Loglinear analysis using SPSS 910
19.10 Interpreting loglinear analysis 913
19.11 Reporting the results of loglinear analysis 919
19.12 Jane and Brian's story 920
19.13 What next? 921
19.14 Key terms that I've discovered 922
Smart Alex's tasks 922

20 CATEGORICAL OUTCOMES: LOGISTIC REGRESSION 925
20.1 What will this chapter tell me? 926
20.2 What is logistic regression? 927
20.3 Theory of logistic regression 927
20.4 Sources of bias and common problems 939
20.5 Binary logistic regression 943
20.6 Interpreting logistic regression 952
20.7 Interactions in logistic regression: a sporty example 962
20.8 Reporting logistic regression 971
20.9 Jane and Brian's story 973
20.10 What next? 975
20.11 Key terms that I've discovered 975
Smart Alex's tasks 976

21 MULTILEVEL LINEAR MODELS 979
21.1 What will this chapter tell me? 980
21.2 Hierarchical data 981
21.3 Multilevel linear models 984
21.4 Practical issues 998
21.5 Multilevel modelling using SPSS 1008
21.6 How to report a multilevel model 1025
21.7 A message from the octopus of inescapable despair 1026
21.8 Jane and Brian's story 1026
21.9 What next? 1026
21.10 Key terms that I've discovered 1027
Smart Alex's tasks 1028

EPILOGUE 1029
APPENDIX 1033
GLOSSARY 1045
REFERENCES 1079
INDEX 1095

PREFACE

Introduction

Many behavioural and social science students (and researchers for that matter) despise statistics. Most of us have a non-mathematical background, which makes understanding complex statistical equations very difficult. Nevertheless, the Evil Horde (a.k.a. professors) force our non-mathematical brains to apply themselves to what is the very complex task of becoming a statistics expert. The end result, as you might expect, can be quite messy. The one weapon that we have is the computer, which allows us to neatly circumvent the considerable hurdle of not understanding mathematics. Computer programs such as IBM SPSS Statistics, SAS, R, JASP and the like provide an opportunity to teach statistics at a conceptual level without getting too bogged down in equations. The computer to a goat-warrior of Satan is like catnip to a cat: it makes them rub their heads along the ground and purr and dribble ceaselessly. The only downside of the computer is that it makes it really easy to make a complete fool of yourself if you don't understand what you're doing. Using a computer without any statistical knowledge at all can be a dangerous thing. Hence this book.

My first aim is to strike a balance between theory and practice: I want to use the computer as a tool for teaching statistical concepts in the hope that you will gain a better understanding of both theory and practice. If you want theory and you like equations then there are certainly more technical books. However, if you want a stats book that also discusses digital rectal stimulation then you have just spent your money wisely.

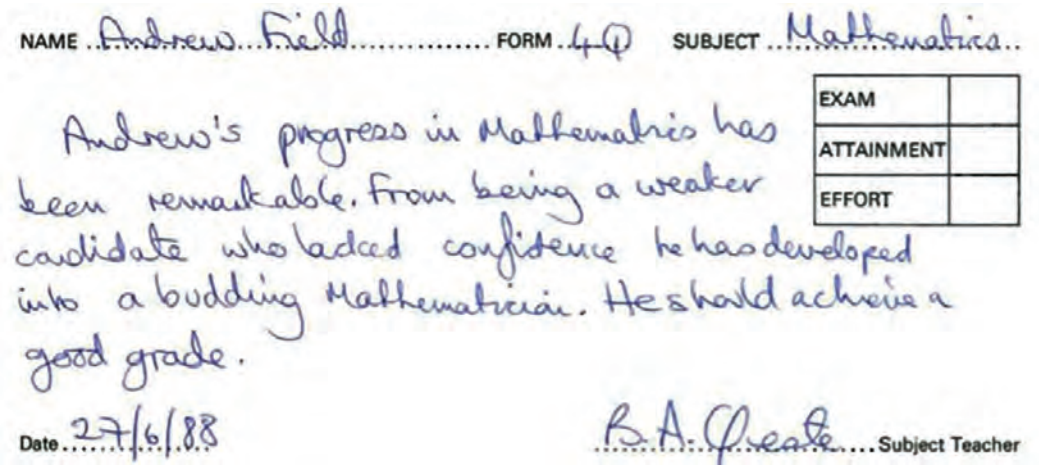
Too many books create the impression that there is a 'right' and 'wrong' way to do statistics. Data analysis is more subjective than you might think. Therefore, although I make recommendations, within the limits imposed by the senseless destruction of rainforests, I hope to give you enough background in theory to enable you to make your own decisions about how best to conduct your analysis.

A second (ridiculously ambitious) aim is to make this the only statistics book that you'll ever need to buy (sort of). It's a book that I hope will become your friend from your first year at university right through to your professorship. The start of the book is aimed at first-year undergraduates (Chapters 1–10), and then we move onto second-year undergraduate level material (Chapters 6, 9 and 11–16) before a dramatic climax that should keep postgraduates tickled (Chapters 17–21). There should be something for everyone in each chapter, and to help you gauge the difficulty of material, I flag the level of each section within each chapter (more on that later).

My final and most important aim is to make the learning process fun. I have a sticky history with maths. This extract is from my school report at the age of 11:

MATHEMATICS ADDL. MATHS.	43	59	27=	D	C	His work shows lack of discipline in thought and presentation. I hope it will mature next year.
CHEMISTRY						

The '27=' in the report is to say that I came equal 27th with another student out of a class of 29. That's pretty much bottom of the class. The 43 is my exam mark as a percentage. Oh dear. Four years later (at 15) this was my school report:



NAME Andrew Field FORM 4(1) SUBJECT Mathematics

EXAM	
ATTAINMENT	
EFFORT	

Andrew's progress in Mathematics has been remarkable. From being a weaker candidate who lacked confidence he has developed into a budding Mathematician. He should achieve a good grade.

Date 27/6/88 R.A. Greatorex Subject Teacher

The catalyst of this remarkable change was a good teacher: my brother, Paul. I owe my life as an academic to Paul's ability to teach me stuff in an engaging way – something my maths teachers failed to do. Paul's a great teacher because he cares about bringing out the best in people, and he was able to make things interesting and relevant to me. Everyone should have a brother Paul to teach them stuff when they're throwing their maths book at their bedroom wall, and I will attempt to be yours.

I strongly believe that people appreciate the human touch, and so I inject a lot of my own personality and sense of humour (or lack of) into *Discovering Statistics Using …* books. Many of the examples in this book, although inspired by some of the craziness that you find in the real world, are designed to reflect topics that play on the minds of some students (i.e., dating, music, celebrity, people doing weird stuff). There are also some examples that are there simply because they made me laugh. So, the examples are light-hearted (some have said 'smutty', but I prefer 'light-hearted') and by the end, for better or worse, I think you will have some idea of what goes on in my head on a daily basis. I apologize to those who don't like it, but, come on, what's not funny about a man putting an eel up his anus?

I never believe that I meet my aims, but previous editions have been popular. I enjoy the rare luxury of having complete strangers emailing me to tell me how wonderful I am. With every new edition, I fear that the changes I make will poison the magic. Let's see what you're going to get and what's different this time around.

What do you get for your money?

This book takes you on a journey (I try my best to make it a pleasant one) not just of statistics but also of the weird and wonderful contents of the world and my brain. It's full of daft examples, bad jokes, and some smut. Aside from the smut, I have been forced reluctantly to include some academic content. It contains most of what I know about statistics.

- Everything you'll ever need to know: I want this book to be good value for money so it guides you from complete ignorance (Chapter 1 tells you the basics of doing research) to multilevel

linear modelling (Chapter 21). No book can contain everything, but I think this one has a fair crack. It's pretty good for developing your biceps also.
- Pointless faces: You'll notice that the book is riddled with 'characters', some of them my own. You can find out more about the pedagogic function of these 'characters' in the next section.
- Data sets: There about 132 data files associated with this book on the companion website. Not unusual in itself for a statistics book, but my data sets contain more sperm (not literally) than other books. I'll let you judge for yourself whether this is a good thing.
- My life story: Each chapter is book-ended by a chronological story from my life. Does this help you to learn about statistics? Probably not, but it might provide light relief between chapters.
- SPSS tips: SPSS does confusing things sometimes. In each chapter, there are boxes containing tips, hints and pitfalls related to SPSS.
- Self-test questions: Given how much students hate tests, I thought the best way to damage sales was to scatter tests liberally throughout each chapter. These range from questions to test what you have just learned to going back to a technique that you read about several chapters before and applying it in a new context.
- Companion website: There's the official companion site, but also the one that I maintain. They contain a colossal amount of additional material, which no one reads. The section about the companion website lets you know what you're ignoring.
- Digital stimulation: not the aforementioned type of digital stimulation, but brain stimulation. Many of the features on the companion website will be accessible from tablets and smartphones, so that when you're bored in the cinema you can read about the fascinating world of heteroscedasticity instead.
- Reporting your analysis: Every chapter has a guide to writing up your analysis. How one writes up an analysis varies a bit from one discipline to another, but my guides should get you heading in the right direction.
- Glossary: Writing the glossary was so horribly painful that it made me stick a vacuum cleaner into my ear to suck out my own brain. You can find my brain in the bottom of the vacuum cleaner in my house.
- Real-world data: Students like to have 'real data' to play with. I trawled the world for examples of research that I consider to be vaguely interesting. Every chapter has a real research example.

What do you get that you didn't get last time?

I suppose if you have spent your hard-earned money on the previous edition it's reasonable that you want a good reason to spend more of your hard-earned money on this edition. In some respects it's hard to quantify all of the changes in a list: I'm a better writer than I was 5years ago, so there is a lot of me rewriting things because I think I can do it better than before. I spent 6 months solidly on the updates, so suffice it to say that a lot has changed; but anything you might have liked about the previous edition probably hasn't changed. Every chapter has had a thorough edit/rewrite. First of all, a few general things across chapters:

- IBM SPSS compliance: This edition was written using version 29 of IBM SPSS Statistics. IBM release new editions of SPSS Statistics more often than I bring out new editions of this book, so, depending on when you buy the book, it may not reflect the latest version. This shouldn't worry you because the procedures covered in this book are unlikely to be affected (see Section 4.2).
- Theory: Chapter 6 was completely rewritten to be the main source of theory for the general linear model (although I pre-empt this chapter with gentler material in Chapter 2). In general, whereas I have shied away from being too strict about distinguishing parameters from their estimates, my recent teaching experiences have convinced me that I can afford to be a bit more precise without losing readers.

- Effect sizes: IBM SPSS Statistics will now, in some situations, produce Cohen's *d* so there is less 'hand calculation' of effect sizes in the book, and I have tended to place more emphasis on partial eta squared in the general linear model chapters.

Here is a chapter-by-chapter run down of the more substantial changes:

- Chapter 1 (Doing research): tweaks, not substantial changes.
- Chapter 2 (Statistical theory): I expanded the section on null hypothesis significance testing to include a concrete example of calculating a *p*-value. I expanded the section on estimation to include a description of maximum likelihood. I expanded the material on probability density functions.
- Chapter 3 (Current thinking in statistics): The main change is that I rewrote the section on the intentions of the researcher and *p*-values to refer back to the new example in Chapter 2. I hope this makes the discussion more concrete.
- Chapter 4 (IBM SPSS Statistics): Obviously reflects changes to SPSS since the previous edition. I expanded the sections on currency variables and data formats, and changed the main example in the chapter. The introduction of the workbook format now means the chapter has sections on using SPSS in both classic and workbook mode. The structure of the chapter has changed a bit as a result.
- Chapter 5 (visualizing data): No substantial changes, I tweaked a few examples.
- Chapter 6 (Bias and model assumptions): This chapter was entirely rewritten. It now does the heavy lifting of introducing the linear model. The first half of the chapter includes some more technical material (which can be skipped) on assumptions of ordinary least squares estimation. I removed most of the material on the *split file* command and frequencies to focus more on the *explore* command.
- Chapter 7 (Nonparametric models): No substantial changes to content other than updates to the SPSS material (which has changed quite a bit).
- Chapter 8 (Correlation): Lots changed in SPSS (e.g., you can obtain confidence intervals for correlations). I overhauled the theory section to link to the updated Chapter 6.
- Chapter 9 (The linear model): Some theory moved out into Chapter 6, so this chapter now has more focus on the 'doing' than the theory.
- Chapter 10 (*t*-tests): I revised some theory to fit with the changes to Chapter 6.
- Chapter 11 (Mediation and moderation): I removed the section on dummy variables and instead expanded the section on dummy coding in Chapter 12. Removing this material made space for a new example using two mediators. I updated all the PROCESS tool material.
- Chapters 12 (GLM 1): I expanded the section on dummy coding (see previous chapter). The effect size material is more focused on partial eta squared.
- Chapter 13 (GLM 2): I framed the material on homogeneity of regression slopes more in terms of fitting parallel slopes and non-parallel slopes models in an attempt to clarify what assumptions we are and are not making with ANCOVA models. I expanded the section on Bayesian models.
- Chapter 14 (GLM 3): I restructured the theoretical material on interactions and simple effects to bring it to the front of the chapter (and to link back to Chapter 11). In SPSS you can now run simple effects through dialog boxes and perform *post hoc* tests on interactions, so I replaced the sections on using syntax and expanded my advice on using these methods. I removed the Labcoat Leni section based on work by Nicolas Guéguen because of concerns that have been raised about his research practices and the retraction of some of his other studies.
- Chapter 15 (GLM 4): I changed both examples in this chapter (so it's effectively a complete rewrite) to be about preventing an alien invasion using sniffer dogs.
- Chapters 16–17 (GLM 5 and MANOVA): These chapters have not changed substantially.

- Chapter 18 (Factor analysis): This chapter has had some theory added. In particular, I have added sections on parallel analysis (including how to conduct it). I have also expanded the reliability theory section. Although the examples are the same, the data file itself has changed (for reasons related to adding the sections on parallel analysis).
- Chapters 19 (Categorical data): No major changes here.
- Chapter 20 (Logistic regression): I have removed the section on multinomial logistic regression to make room for an expanded theory section on binary logistic regression. I felt like the chapter covered a lot of ground without actually giving students a good grounding in what logistic regression does. I had lots of ideas about how to rewrite the theory section, and I'm very pleased with it, but something had to make way. I also changed the second example (penalty kicks) slightly to allow me to talk about interactions in binary logistic regression and to reinforce how to interpret logistic models (which I felt was lacking in previous editions).
- Chapter 21 (Multilevel models): Wow, this was a gateway to a very unpleasant dimension for me. This chapter is basically a complete rewrite. I expanded the theory section enormously and also included more practical advice. To make space the section on growth models was removed, but it's fair to say that I think this version will give readers a much better grounding in multilevel models. The main example changed slightly (new data, but still on the theme of cosmetic surgery).

Goodbye

The first edition of this book was the result of two years (give or take a few weeks to write up my Ph.D.) of trying to write a statistics book that I would enjoy reading. With each new edition I try not just to make superficial changes but also to rewrite and improve everything (one of the problems with getting older is you look back at your past work and think you can do things better). This book has consumed the last 25 years or so of my life, and each time I get a nice email from someone who found it useful I am reminded that it is the most useful thing I'll ever do in my academic life. It began and continues to be a labour of love. It still isn't perfect, and I still love to have feedback (good or bad) from the people who matter most: you.

Andy



www.facebook.com/profandyfield

@ProfAndyField

www.youtube.com/user/ProfAndyField

https://www.instagram.com/discovering_catistics/

www.discoverspss.com

HOW TO USE THIS BOOK

When the publishers asked me to write a section on 'How to use this book' it was tempting to write 'Buy a large bottle of Olay anti-wrinkle cream (which you'll need to fend off the effects of ageing while you read), find a comfy chair, sit down, fold back the front cover, begin reading and stop when you reach the back cover.' I think they wanted something more useful.☺

What background knowledge do I need?

In essence, I assume that you know nothing about statistics, but that you have a very basic grasp of computers (I won't be telling you how to switch them on, for example) and maths (although I have included a quick overview of some concepts).

Do the chapters get more difficult as I go through the book?

Yes, more or less: Chapters 1–10 are first-year degree level, Chapters 9–16 move into second-year degree level, and Chapters 17–21 discuss more technical topics. However, my aim is to tell a statistical story rather than worry about what level a topic is at. Many books teach different tests in isolation and never really give you a grasp of the similarities between them; this, I think, creates an unnecessary mystery. Most of the statistical models in this book are the same thing expressed in slightly different ways. I want the book to tell this story, and I see it as consisting of seven parts:

- Part 1 (Doing research): Chapters 1–4.
- Part 2 (Exploring data and fitting models): Chapters 5–7.
- Part 3 (Linear models with continuous predictors): Chapters 8–9.
- Part 4 (Linear models with continuous or categorical predictors): Chapters 10–16.
- Part 5 (Linear models with multiple outcomes): Chapter 17–18.
- Part 6 (Linear models with categorical outcomes): Chapters 19–20.
- Part 7 (Linear models with hierarchical data structures): Chapter 21.

This structure might help you to see the method in my mind. To help you on your journey I've also coded sections within chapters with a letter icon that gives you an idea of the difficulty of the material. It doesn't mean you can skip the sections, but it might help you to contextualize sections that you find challenging. It's based on a wonderful categorization system using the letters A to D to indicate increasing difficulty. Each letter also conveniently stands for a word that rhymes with and describes the likely state of your brain as you read these sections:

- Ⓐ is for brain **Attain**: I'd hope that everyone can get something out of these sections. These sections are aimed at readers just starting a non-STEM undergraduate degree.
- Ⓑ is for brain **Bane**. These sections are aimed at readers with a bit of statistics knowledge, but they might at times be a source of unhappiness for your brain. They are aimed at readers in the second year of a non-STEM degree.

- Ⓒ is for brain **Complain**. These topics are quite difficult. They are aimed at readers completing the final year of a non-STEM undergraduate degree or starting a Masters degree.
- Ⓓ is for brain **Drain**. These are difficult topics that will drain many people's brains including my own. The 'D' might instead stand for **Dogbane** because if your brain were a dog, these sections would kill it. Anything that hurts dogs is evil, draw your own conclusions about these sections.

Why do I keep seeing silly faces everywhere?



Brian Haemorrhage: Brian is a really nice guy, and he has a massive crush on Jane Superbrain. He's seen her around the university campus and whenever he sees her, he gets a knot in his stomach. He soon realizes that the only way she'll date him is if he becomes a statistics genius (and changes his surname). So, he's on a mission to learn statistics. At the moment he knows nothing, but he's about to go on a journey that will take him from statistically challenged to a genius, in 1,144 pages. Along his journey he pops up and asks questions, and at the end of each chapter he flaunts his newly found knowledge to Jane in the hope that she'll notice him.



Confusius: The great philosopher Confucius had a lesser-known brother called Confusius. Jealous of his brother's great wisdom and modesty, Confusius vowed to bring confusion to the world. To this end, he built the confusion machine. He puts statistical terms into it, and out of it come different names for the same concept. When you see Confusius he will be alerting you to statistical terms that mean (essentially) the same thing.



Correcting Cat: This cat lives in the ether and appears to taunt the Misconception Mutt by correcting his misconceptions. He also appears when I want to make a bad cat-related pun. He exists in loving memory of my own ginger cat who after 20 years as the star of this book sadly passed away, which he promised me he'd never do. You can't trust a cat.



Cramming Sam: Samantha thinks statistics is a boring waste of time. She wants to pass her exam and forget that she ever had to know anything about normal distributions. She appears and gives you a summary of the key points that you need to know. If, like Samantha, you're cramming for an exam, she will tell you the essential information to save you having to trawl through hundreds of pages of my drivel.



Jane Superbrain: Jane is the cleverest person in the universe. She has acquired a vast statistical knowledge, but no one knows how. She is an enigma, an outcast, and a mystery. Brian has a massive crush on her. Jane appears to tell you advanced things that are a bit tangential to the main text. Can Brian win his way into her heart? You'll have to read and find out.



Labcoat Leni: Leni is a budding young scientist and he's fascinated by real research. He says, 'Andy, I like an example about using an eel as a cure for constipation as much as the next guy, but all of your data are made up. We need some real examples, buddy!' Leni walked the globe, a lone data warrior in a thankless quest for real data. When you see Leni you know that you will get some real data, from a real research study, to analyse. Keep it real.



Misconception Mutt: The Misconception Mutt follows his owner to statistics lectures and finds himself learning about stats. Sometimes he gets things wrong, though, and when he does something very strange happens. A ginger cat materializes out of nowhere and corrects him.



Oditi's Lantern: Oditi believes that the secret to life is hidden in numbers and that only by large-scale analysis of those numbers shall the secrets be found. He didn't have time to enter, analyse, and interpret all the data in the world, so he established the cult of undiscovered numerical truths. Working on the principle that if you gave a million monkeys typewriters, one of them would re-create Shakespeare, members of the cult sit at their computers crunching numbers in the hope that one of them will unearth the hidden meaning of life. To help his cult Oditi has set up a visual vortex called 'Oditi's Lantern'. When Oditi appears it is to implore you to stare into the lantern, which basically means there is a video tutorial to guide you.



Oliver Twisted: With apologies to Charles Dickens, Oliver, like the more famous fictional London urchin, asks 'Please, Sir, can I have some more?' Unlike Master Twist though, Master Twisted always wants more statistics information. Who wouldn't? Let us not be the ones to disappoint a young, dirty, slightly smelly boy who dines on gruel. When Oliver appears he's telling you that there is additional information on the companion website. (It took a long time to write, so someone please actually read it.)



Satan's Personal Statistics Slave: Satan is a busy boy – he has all of the lost souls to torture in hell, then there are the fires to keep fuelled, not to mention organizing enough carnage on the planet's surface to keep Norwegian black metal bands inspired. Like many of us, this leaves little time for him to analyse data, and this makes him very sad. So, he has his own personal slave, who, also like some of us, spends all day dressed in a gimp mask and tight leather pants in front of SPSS analysing Satan's data. Consequently, he knows a thing or two about SPSS, and when Satan's busy spanking a goat, he pops up in a box with SPSS tips.



Smart Alex: Alex was aptly named because she's, like, super smart. She likes teaching people, and her hobby is posing people questions so that she can explain the answers to them. Alex appears at the end of each chapter to pose you some questions. Her answers are on the companion website.

What is on the companion websites?

My exciting website

The data files for the book and the solutions to the various tasks and pedagogic features in this book are at www.discoverspss.com, a website that I wrote and maintain. This website contains:

- Self-test solutions: where the self-test is not answered within the chapter, the solution is on this website.
- Smart Alex answers: Each chapter ends with a set of tasks for you to test your newly acquired expertise. The website contains detailed answers. Will I ever stop writing?

- Datasets: we use a lot of bizarre but hopefully interesting datasets throughout this book, all of which can be downloaded so that you can practice what you learn.
- Screencasts: whenever you see Oditi it means that there is a screencast to accompany the chapter. Although these videos are hosted on my YouTube channel (www.youtube.com/user/ProfAndyField), which I have amusingly called µ-Tube (see what I did there?), they are also embedded in interactive tutorials hosted at my site.
- Labcoat Leni solutions: there are full answer to the Labcoat Leni tasks.
- Oliver Twisted's pot of gruel: Oliver Twisted will draw your attention to the hundreds of pages of more information that we have put online so that (1) the planet suffers a little less, and (2) you won't die when the book falls from your bookshelf onto your head.
- Cyberworms of knowledge: I have used nanotechnology to create cyberworms that crawl down your broadband, wifi or 5G, pop out of a port on your computer, tablet, or phone, and fly through space into your brain. They rearrange your neurons so that you understand statistics. You don't believe me? You'll never know for sure unless you visit my website …

Sage's less exciting website for lecturers and professors

In addition, the publishers have an official companion website that contains a bunch of useful stuff for lecturers and professors, most of which I haven't had any part in. Their site will also link to my resources above, so you can use the official site as a one-stop shop. To enter Sage's world of delights, go to https://study.sagepub.com/field6e. Once you're there, Sage will flash up subliminal messages that make you use more of their books, but you'll also find resources for use with students in the classroom and for assessment.

- Testbank: there is a comprehensive testbank of multiple-choice questions for instructors to use for substantive assessment. It comes in a lovely file that you can upload into your online teaching system (with answers separated out). This enables you to assign questions for exams and assignments. Students can see feedback on correct/incorrect answers, including pointers to areas in the book where the right answer can be found.
- Chapter based multiple-choice questions: organized to mirror your journey through the chapters, which allow you to test students' formative understanding week-by-week and to see whether you are wasting your life, and theirs, trying to explain statistics. This should give them the confidence to waltz into the examination. If everyone fails said exam, please don't sue me.
- Resources for other subject areas: I am a psychologist, and although I tend to base my examples around the weird and wonderful, I do have a nasty habit of resorting to psychology when I don't have any better ideas. My publishers have recruited some non-psychologists to provide data files and an instructor's testbank of multiple-choice questions for those teaching in business and management, education, sport sciences and health sciences. You have no idea how happy I am that I didn't have to write those.
- PowerPoint decks: I can't come and teach your classes for you (although you can watch my lectures on YouTube). Instead, I like to imagine a *Rocky*-style montage of lecturers training to become a crack team of highly skilled and super-intelligent pan-dimensional beings poised to meet any teaching-based challenge. To assist in your mission to spread the joy of statistics I have provided PowerPoint decks for each chapter. If you see something weird on the slides that upsets you, then remember that's probably my fault.

Happy reading, and don't get distracted by social media.

THANK YOU

Colleagues: This book (in the SPSS, SAS, and R versions) wouldn't have happened if not for Dan Wright's unwarranted faith in a then postgraduate to write the first SPSS edition. Numerous other people have contributed to previous editions of this book. I don't have room to list them all, but particular thanks to Dan (again), David Hitchin, Laura Murray, Gareth Williams, Lynne Slocombe, Kate Lester, Maria de Ridder, Thom Baguley, Michael Spezio, J. W. Jacobs, Ann-Will Kruijt, Johannes Petzold, E.-J. Wagenmakers, and my wife Zoë who have given me invaluable feedback during the life of this book. Special thanks to Jeremy Miles. Part of his 'help' involves ranting on at me about things I've written being, and I quote, 'bollocks'. Nevertheless, working on the SAS and R versions of this book with him influenced me enormously. He's also been a very nice person to know over the past few years (apart from when he's ranting on at me about …). For this edition I'm grateful to Nick Brown, Bill Browne, and Martina Sladekova for feedback on some sections and Hanna Eldarwish for her help with the web materials. Thanks to the following for permission to include data from their fascinating research: Rebecca Ang, Philippe Bernard, Michael V. Bronstein, Hakan Çetinkaya, Tomas Chamorro-Premuzic, Graham Davey, Mike Domjan, Gordon Gallup, Sarah Johns, Eric Lacourse, Nate Lambert, Sarah Marzillier, Karlijn Massar, Geoffrey Miller, Peter Muris, Laura Nichols, Nick Perham, Achim Schützwohl, Mirjam Tuk, and Lara Zibarras.

Not all contributions are as tangible as those above. Very early in my career Graham Hole made me realize that teaching research methods didn't have to be dull. My approach to teaching has been to steal his good ideas and he has had the good grace not to ask for them back! He is a rarity in being brilliant, funny, and nice. Over the past few years (in no special order) Danielle Evans, Lincoln Colling, Jennifer Mankin, Martina Sladekova, Jenny Terry, and Vlad Costin from the methods team at the University of Sussex have been the most inspirational and supportive colleagues you could hope to have. It is a privilege to be one of their team.

Readers: I appreciate everyone who has taken time to write nice things about this book in emails, reviews, and on websites and social media; the success of this book has been in no small part due to these people being so positive and constructive in their feedback. I always hit motivational dark times when I'm writing, but feeling the positive vibes from readers always gets me back on track (especially the photos of cats, dogs, parrots, and lizards with my books☺). I continue to be amazed and bowled over by the nice things that people say about the book (and disproportionately upset by the less positive things).

Software: This book wouldn't exist without the generous support of International Business Machines Corporation (IBM) who kindly granted permission for me to include screenshots and images from SPSS. I wrote this edition on MacOS but used Windows for the screenshots. Mac and Mac OS are trademarks of Apple Inc., registered in the United States and other countries; Windows is a registered trademark of Microsoft Corporation in the United States and other countries. I don't get any incentives for saying this (perhaps I should, hint, hint …) but the following software packages are invaluable to me when writing: TechSmith's (www.techsmith.com) Camtasia (which I use to produce videos) and Snagit (which I use for screenshots) for Mac; the Omnigroup's (www.omnigroup.com) OmniGraffle, which I use to create most of the diagrams and flowcharts (it is awesome); and R and R Studio, which I use for data visualizations, creating data, and building my companion website.

Publishers: My publishers, Sage, are rare in being a large, successful company that manages to maintain a family feel. For this edition I was particularly grateful for them trusting me enough to leave me alone to get on with things because my deadline was basically impossible. Now I have emerged from my attic, I'm fairly sure I'm going to be grateful to Jai Seaman and Hannah Cooper for what they have been doing and will do to support the book. I extremely grateful to have again had the rigorous and meticulous eyes and brain of Richard Leigh to copyedit this edition. My long-suffering production editor, Ian Antcliff, deserves special mention not only for the fantastic job he does but also for being the embodiment of calm when the pressure is on. I'm also grateful to Karen and Ziyad who don't work directly on my books but are such an important part of my fantastic relationship with Sage.

We've retained the characters and designs that James Iles produced in the previous edition. It's an honour to have his artwork in another of my books.

Music: I always write listening to music. For this edition I predominantly enjoyed (my neighbours less so): AC/DC, A Forest of Stars, Ashenspire, Courtney Marie Andrews, Cradle of Filth, The Damned, The Darkness, Deafheaven, Deathspell Omega, Deep Purple, Europe, Fishbone, Ghost, Helloween, Hexvessel, Ihsahn, Iron Maiden, Janes Addiction, Jonathan Hultén, Judas Priest, Katatonia, Lingua Ignota, Liturgy, Marillion, Mastodon, Meshuggah, Metallica, Motörhead, Nick Drake, Opeth, Peter Gabriel, Queen, Queensryche, Satan, Slayer, Smith/Kotzen, Tesseract, The Beyond, Tribulation, Wayfarer, and ZZ Top.

Friends and family: All this book-writing nonsense requires many lonely hours of typing. I'm grateful to some wonderful friends who drag me away from the dark places that I can tend to inhabit. My eternal gratitude goes to Graham Davey, Ben Dyson, Mark Franklin, and their lovely families for reminding me that there is more to life than work. Particular thanks this time to Darren Hayman, my partner in the strange and beautiful thing that we call 'listening club'. Thanks to my parents and Paul and Julie for being my family. Special cute thanks to my niece and nephew, Oscar and Melody: I hope to teach you many things that will annoy your parents.

Chicken-flavoured thanks to Milton the spaniel for giving the best hugs. For someone who spends his life writing, I'm constantly surprised at how incapable I am of finding words to express how wonderful my wife Zoë is. She has a never-ending supply of patience, love, support, and optimism (even when her husband is a grumpy, sleep-deprived, withered, self-doubting husk). I never forget, not even for a nanosecond, how lucky I am. Finally, thanks to Zach and Arlo for the realization of how utterly pointless work is and for the permanent feeling that my heart has expanded to bursting point from trying to contain my love for them.

DEDICATION

Like the previous editions, this book is dedicated to my brother Paul and my cat Fuzzy (now in the spirit cat world), because one of them was an intellectual inspiration and the other woke me up in the morning by sitting on me and purring in my face until I gave him cat food: mornings were considerably more pleasant when my brother got over his love of cat food for breakfast.

SYMBOLS USED IN THIS BOOK

Mathematical operators

Σ	This symbol (called sigma) means 'add everything up'. So, if you see something like $\sum x_i$ it means 'add up all of the scores you've collected'.
Π	This symbol means 'multiply everything'. So, if you see something like $\prod x_i$ it means 'multiply all of the scores you've collected'.
$\sqrt{x}$	This means 'take the square root of x'.

Greek symbols

α	Alpha, the probability of making a Type I error
β	Beta, the probability of making a Type II error
β_i	Beta, the standardized regression coefficient
ε	Epsilon, usually stands for 'error', but is also used to denote sphericity
η^2	Eta squared, an effect size measure
μ	Mu, the mean of a population of scores
ρ	Rho, the correlation in the population; also used to denote Spearman's correlation coefficient
σ	Sigma, the standard deviation in a population of data
σ^2	Sigma squared, the variance in a population of data
$\sigma_{\overline{X}}$	Another variation on sigma, which represents the standard error of the mean
τ	Kendall's tau (non-parametric correlation coefficient)
ϕ	Phi, a measure of association between two categorical variables, but also used to denote the dispersion parameter in logistic regression
χ^2	Chi-square, a test statistic that quantifies the association between two categorical variables
χ^2_F	Another use of the letter chi, but this time as the test statistic in Friedman's ANOVA, a non-parametric test of differences between related means
ω^2	Omega squared (an effect size measure). This symbol also means 'expel the contents of your intestine immediately into your trousers'; you will understand why in due course.

English symbols

b_i	The regression coefficient (unstandardized); I tend to use it for any coefficient in a linear model
df	Degrees of freedom
e_i	The residual associated with the ith person
F	F-statistic (test statistic used in ANOVA)
H	Kruskal–Wallis test statistic
k	The number of levels of a variable (i.e., the number of treatment conditions), or the number of predictors in a regression model
ln	Natural logarithm
MS	The mean squared error (mean square): the average variability in the data
N, n, n_i	The sample size. N usually denotes the total sample size, whereas n usually denotes the size of a particular group
P	Probability (the probability value, p-value, or significance of a test are usually denoted by p)
r	Pearson's correlation coefficient
r_s	Spearman's rank correlation coefficient
r_b, r_{pb}	Biserial correlation coefficient and point-biserial correlation coefficient, respectively
R	The multiple correlation coefficient
R^2	The coefficient of determination (i.e., the proportion of data explained by the model)
s	The standard deviation of a sample of data
s^2	The variance of a sample of data
SS	The sum of squares, or sum of squared errors to give it its full title
SS_A	The sum of squares for variable A
SS_M	The model sum of squares (i.e., the variability explained by the model fitted to the data)
SS_R	The residual sum of squares (i.e., the variability that the model can't explain – the error in the model)
SS_T	The total sum of squares (i.e., the total variability within the data)
t	Testicle statistic for a t-test. Yes, I did that deliberately to check whether you're paying attention.
T	Test statistic for Wilcoxon's matched-pairs signed-rank test
U	Test statistic for the Mann–Whitney test
W_s	Test statistic for Rilcoxon's wank-sum test. See what I did there? It doesn't matter because no one reads this page.
$\bar{X}$	The mean of a sample of scores
z	A data point expressed in standard deviation units

A BRIEF MATHS OVERVIEW

There are good websites that can help you if any of the maths in this book confuses you. Use a search engine to find something that suits you.

Working with expressions and equations

Here are three important things to remember about working with equations:

1 Two negatives make a positive: Although in life two wrongs don't make a right, in mathematics they do. When we multiply a negative number by another negative number, the result is a positive number. For example, $(-2) \times (-4) = 8$.
2 A negative number multiplied by a positive one makes a negative number: If you multiply a positive number by a negative number then the result is another negative number. For example, $2 \times (-4) = -8$, or $(-2) \times 6 = -12$.
3 BODMAS and PEMDAS: These two acronyms are different ways of remembering the order in which mathematical operations are performed. BODMAS stands for Brackets, Order, Division, Multiplication, Addition, and Subtraction; whereas PEMDAS stems from Parentheses, Exponents, Multiplication, Division, Addition, and Subtraction. Having two widely used acronyms is confusing (especially because multiplication and division are the opposite way around), but they mean the same thing:

    - Brackets/Parentheses: When solving any expression or equation, you deal with anything in brackets/parentheses first.
    - Order/Exponents: Having dealt with anything in brackets, you next deal with any order terms/exponents. These refer to power terms such as squares. Four squared, or 4^2, used to be called 4 raised to the order of 2, hence the word 'order' in BODMAS. These days the term 'exponents' is more common (so by all means use BEDMAS as your acronym if you find that easier).
    - Division and Multiplication: The next things to evaluate are any division or multiplication terms. The order in which you handle them is from the left to the right of the expression/equation. That's why BODMAS and PEMDAS can list them the opposite way around, because they are considered at the same time (so, BOMDAS or PEDMAS would work as acronyms too).
    - Addition and Subtraction: Finally deal with any addition or subtraction. Again, go from left to right doing any addition or subtraction in the order that you meet the terms. (So, BODMSA would work as an acronym too but it's hard to say.)

Let's look at an example of BODMAS/PEMDAS in action: what would be the result of $1 + 3 \times 5^2$? The answer is 76 (not 100 as some of you might have thought). There are no brackets, so the first thing is to deal with the order/exponent: 5^2 is 25, so the equation becomes $1 + 3 \times 25$. Moving from left to right, there is no division, so we do the multiplication: 3×25, which gives us 75. Again, going from

left to right we look for addition and subtraction terms. There are no subtractions so the first thing we come across is the addition: 1 + 75, which gives us 76 and the expression is solved. If I'd written the expression as $(1 + 3) \times 5^2$, then the answer would have been 100 because we deal with the brackets first: $(1 + 3) = 4$, so the expression becomes 4×5^2. We then deal with the order/exponent (5^2 is 25), which results in $4 \times 25 = 100$.

Logarithms

The logarithm is the inverse function to exponentiation, which means that it reverses exponentiation. In particular, the natural logarithm is the logarithm of a number using something called Euler's number ($e \approx 2.718281$) as its base. So, what is the natural logarithm in simple terms? It is the number of times that you need to multiply e by itself to get that number. For example, the natural logarithm of 54.6 is approximately 4, because you need to multiply e by itself 4 times to get 54.6.

$$e^4 = 2.718281^4 \approx 54.6$$

Therefore,

$$\ln(54.6) \approx 4$$

In general, if x and y are two values,

$$y = \ln(x) \text{ is equivalent to } x \approx e^y.$$

For example,

$$4 \approx \ln(54.6) \text{ is equivalent to } 54.6 \approx e^4.$$

There are two rules relevant to this book that apply to logarithms. First, when you add the natural logarithms of two values, the result is the natural logarithm of their products (i.e. the values multiplied). In general,

$$\ln(x) + \ln(y) = \ln(xy).$$

Second, when you subtract the natural logarithms of two values, the result is the natural logarithm of their ratio (i.e. the first value divided by the second). In general,

$$\ln(x) - \ln(y) = \ln\left(\frac{x}{y}\right).$$

Matrices

We don't use matrices much in this book, but they do crop up. A matrix is a grid of numbers arranged in columns and rows. A matrix can have many columns and rows, and we specify its dimensions using numbers. In general people talk of an $i \times j$ matrix in which i is the number of rows and j is the number of columns. For example, a 2×3 matrix is a matrix with two rows and three columns, and a 5×4 matrix is one with five rows and four columns.

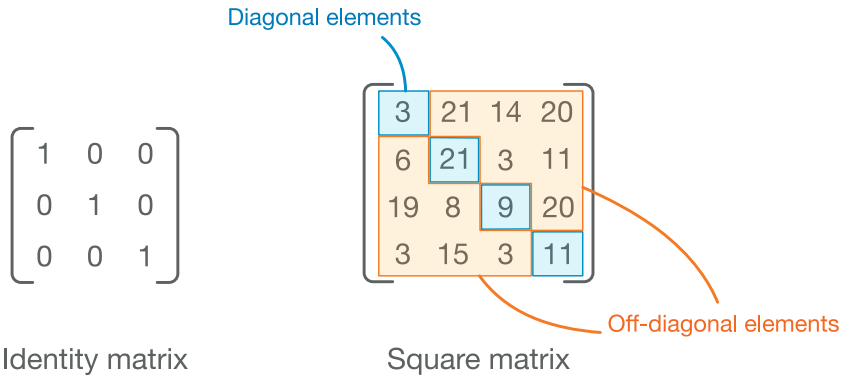
$$\begin{bmatrix} 2 & 5 & 6 \\ 3 & 5 & 8 \end{bmatrix}$$

2×3 matrix

$$\begin{bmatrix} 3 & 21 & 14 & 20 \\ 6 & 21 & 3 & 11 \\ 19 & 8 & 9 & 20 \\ 3 & 15 & 3 & 11 \\ 23 & 1 & 14 & 11 \end{bmatrix}$$

5×4 matrix

The values within a matrix are *components* or *elements* and the rows and columns are *vectors*. A *square matrix* has an equal number of columns and rows. When using square matrices we sometimes refer to the diagonal components (i.e., the values that lie on the diagonal line from the top left component to the bottom right component) and the off-diagonal ones (the values that do not lie on the diagonal). In the square matrix below, the diagonal components are 3, 21, 9 and 11 and the off-diagonal components are the other values. An *identity matrix* is a square matrix in which the diagonal elements are 1 and the off-diagonal elements are 0. Hopefully, the concept of a matrix is less scary than you thought it might be: it is not some magical mathematical entity, merely a way of representing data – like a spreadsheet.





WHY IS MY EVIL LECTURER FORCING ME TO LEARN STATISTICS?

1.1 What the hell am I doing here? I don't belong here 3
1.2 The research process 3
1.3 Initial observation: finding something that needs explaining 4
1.4 Generating and testing theories and hypotheses 5
1.5 Collecting data: measurement 9
1.6 Collecting data: research design 16
1.7 Analysing data 22
1.8 Reporting data 40
1.9 Jane and Brian's story 43
1.10 What next? 45
1.11 Key terms that I've discovered 45
Smart Alex's tasks 46

Letters A to D indicate increasing difficulty

I was born on 21 June 1973. Like most people, I don't remember anything about the first few years of life, and, like most children, I went through a phase of annoying my dad by asking 'Why?' every five seconds. With every question, the word 'dad' got longer and whinier: 'Dad, why is the sky blue?', 'Daaad, why don't worms have legs?', 'Daaaaaaaaad, where do babies come from?' Eventually, my dad could take no more and whacked me around the face with a golf club.[1]

My torrent of questions reflected the natural curiosity that children have: we all begin our voyage through life as inquisitive little scientists. At the age of 3, I was at my friend Obe's party (just before he left England to return to Nigeria, much to my distress). It was a hot day, and there was an electric fan blowing cold air around the room. My 'curious little scientist' brain was working through what seemed like a particularly pressing question: 'What happens when you stick your finger in a fan?' The answer, as it turned out, was that it hurts – a lot.[2] At the age of 3, we intuitively know that to answer questions you need to collect data, even if it causes us pain.

My curiosity to explain the world never went away, which is why I'm a scientist. The fact that you're reading this book means that the inquisitive 3-year-old in you is alive and well and wants to answer new and exciting questions too. To answer these questions you need 'science', and science has a pilot fish called 'statistics' that hides under its belly eating ectoparasites. That's why your evil lecturer is forcing you to learn statistics. Statistics is a bit like sticking your finger into a revolving fan blade: sometimes it's very painful, but it does give you answers to interesting questions. I'm going to try to convince you in this chapter that statistics is an important part of doing research. We will overview the whole research process, from why we conduct research in the first place, through how theories are generated, to why we need data to test these theories. If that doesn't convince you to read on then maybe the fact that we discover whether Coca-Cola kills sperm will. Or perhaps not.



Figure 1.1 When I grow up, please don't let me be a statistics lecturer

1 Accidentally – he was practicing in the garden when I unexpectedly wandered behind him at the exact moment he took a back swing. It's rare that a parent enjoys the sound of their child crying, but on this day it filled my dad with joy because my wailing was tangible evidence that he hadn't killed me, which he thought he might have done.

2 In the 1970s fans didn't have helpful protective cages around them to prevent foolhardy 3-year-olds sticking their fingers into the blades.

1.1 What the hell am I doing here? I don't belong here Ⓐ

You're probably wondering why you have bought this book. Maybe you liked the pictures, maybe you fancied doing some weight training (it *is* heavy), or perhaps you needed to reach something in a high place (it *is* thick). The chances are, though, that given the choice of spending your hard-earned cash on a statistics book or something more entertaining (a nice novel, a trip to the cinema, etc.) you'd choose the latter. So, why have you bought the book (or downloaded an illegal PDF of it from someone who has way too much time on their hands if they're scanning 900 pages for fun)? It's likely that you obtained it because you're doing a course on statistics, or you're doing some research, and you need to know how to analyse data. It's possible that you didn't realize when you started your course or research that you'd have to know about statistics but now find yourself inexplicably wading, neck high, through the Victorian sewer that is data analysis. The reason why you're in the mess that you find yourself in is that you have a curious mind. You might have asked yourself questions like why people behave the way they do (psychology) or why behaviours differ across cultures (anthropology), how businesses maximize their profit (business), how the dinosaurs died (palaeontology), whether eating tomatoes protects you against cancer (medicine, biology), whether it is possible to build a quantum computer (physics, chemistry), or whether the planet is hotter than it used to be and in what regions (geography, environmental studies). Whatever it is you're studying or researching, the reason why you're studying it is probably that you're interested in answering questions. Scientists are curious people, and you probably are too. However, it might not have occurred to you that to answer interesting questions, you need data and explanations for those data.

The answer to 'what the hell are you doing here?' is simple: to answer interesting questions you need data. One of the reasons why your evil statistics lecturer is forcing you to learn about numbers is that they are a form of data and are vital to the research process. Of course there are forms of data other than numbers that can be used to test and generate theories. When numbers are involved the research involves **quantitative methods**, but you can also generate and test theories by analysing language (such as conversations, magazine articles and media broadcasts). This research involves **qualitative methods** and it is a topic for another book not written by me. People can get quite passionate about which of these methods is *best*, which is a bit silly because they are complementary, not competing, approaches and there are much more important issues in the world to get upset about. Having said that, all qualitative research is rubbish.[3]

1.2 The research process Ⓐ

How do you go about answering an interesting question? The research process is broadly summarized in Figure 1.2. You begin with an observation that you want to understand, and this observation could be anecdotal (you've noticed that your cat watches birds when they're on TV but not when jellyfish are on)[4] or could be based on some data (you've got several cat owners to keep diaries of their cat's TV habits and noticed that lots of them watch birds). From your initial observation you consult relevant theories and generate



3 This is a joke. Like many of my jokes, there are people who won't find it remotely funny, but this is what you have signed up for I'm afraid.

4 In his younger days my cat (RIP) actually did climb up and stare at the TV when birds were being shown.

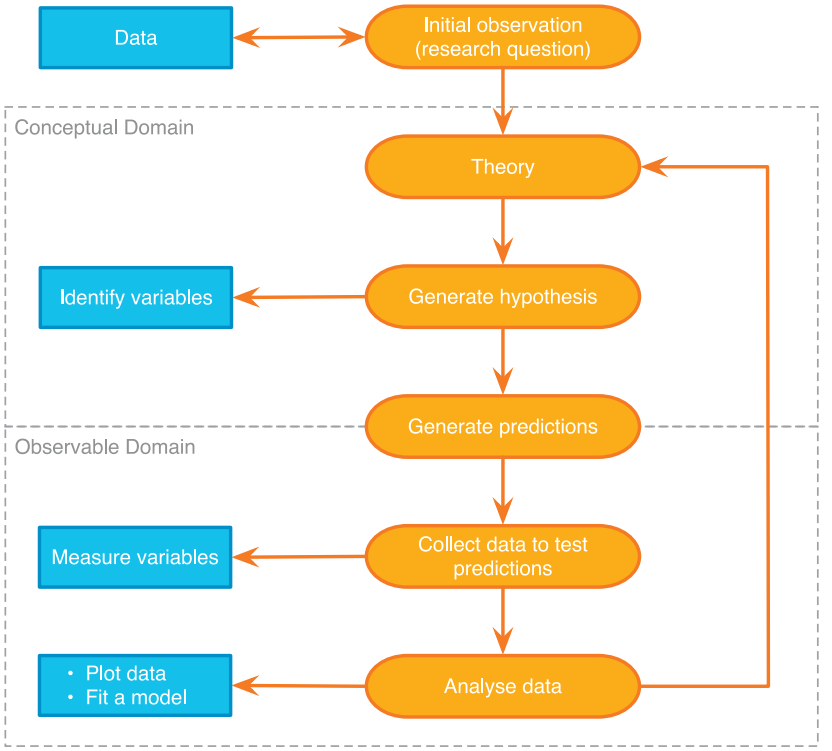


Figure 1.2 The research process

explanations (hypotheses) for those observations, from which you can make predictions. To test your predictions you need data. First you collect some relevant data (and to do that you need to identify things that can be measured) and then you analyse those data. The analysis of the data may support your hypothesis, or generate a new one, which in turn might lead you to revise the theory. As such, the processes of data collection and analysis and generating theories are intrinsically linked: theories lead to data collection/analysis and data collection/analysis informs theories. This chapter explains this research process in more detail.

1.3 Initial observation: finding something that needs explaining Ⓐ

The first step in Figure 1.2 was to come up with a question that needs an answer. I spend rather more time than I should watching reality TV. Over many years I used to swear that I wouldn't get hooked on reality TV, and yet year upon year I would find myself glued to the TV screen waiting for the next contestant's meltdown (I am a psychologist, so really this is just research). I used to wonder why there is so much arguing in these shows, and why so many contestants have really unpleasant personalities (my money is on narcissistic personality disorder).[5] A lot of scientific endeavour starts this way: not by watching reality TV, but by observing something in the world and wondering why it happens.

5 This disorder is characterized by (among other things) a grandiose sense of self-importance, arrogance, lack of empathy for others, envy of others and belief that others envy them, excessive fantasies of brilliance or beauty, the need for excessive admiration by and exploitation of others.

Having made a casual observation about the world (a lot of reality TV contestants have extreme personalities and argue a lot), I need to collect some data to see whether this observation is true (and not a biased observation). To do this, I need to identify one or more **variables** that quantify the thing I'm trying to measure. There's one variable in this example: 'narcissistic personality disorder'. I could measure this variable by giving contestants one of the many well-established questionnaires that measure personality characteristics. Let's say that I did this and I found that 75% of contestants had narcissistic personality disorder. These data support my observation: a lot of reality TV contestants have narcissistic personality disorder.

1.4 Generating and testing theories and hypotheses Ⓐ

The next logical thing to do is to explain these data (Figure 1.2). The first step is to look for relevant theories. A **theory** is an explanation or set of principles that is well substantiated by repeated testing and explains a broad phenomenon. We might begin by looking at theories of narcissistic personality disorder, of which there are currently very few. One theory of personality disorders in general links them to early attachment (put simplistically, the bond formed between a child and their main caregiver). Broadly speaking, a child can form a secure (a good thing) or an insecure (not so good) attachment to their caregiver, and the theory goes that insecure attachment explains later personality disorders (Levy et al., 2015). This is a theory because it is a set of principles (early problems in forming interpersonal bonds) that explains a general broad phenomenon (disorders characterized by dysfunctional interpersonal relations). There is also a critical mass of evidence to support the idea. The theory also tells us that those with narcissistic personality disorder tend to engage in conflict with others despite craving their attention, which perhaps explains their difficulty in forming close bonds.

Given this theory, we might generate a **hypothesis** about our earlier observation (see Jane Superbrain Box 1.1). A hypothesis is a proposed explanation for a fairly narrow phenomenon or set of observations. It is not a guess, but an informed, theory-driven attempt to explain what has been observed. Both theories and hypotheses seek to explain the world, but a theory explains a wide set of phenomena with a small set of well-established principles, whereas a hypothesis typically seeks to explain a narrower phenomenon and is, as yet, untested. Both theories and hypotheses exist in the conceptual domain, and you cannot observe them directly.

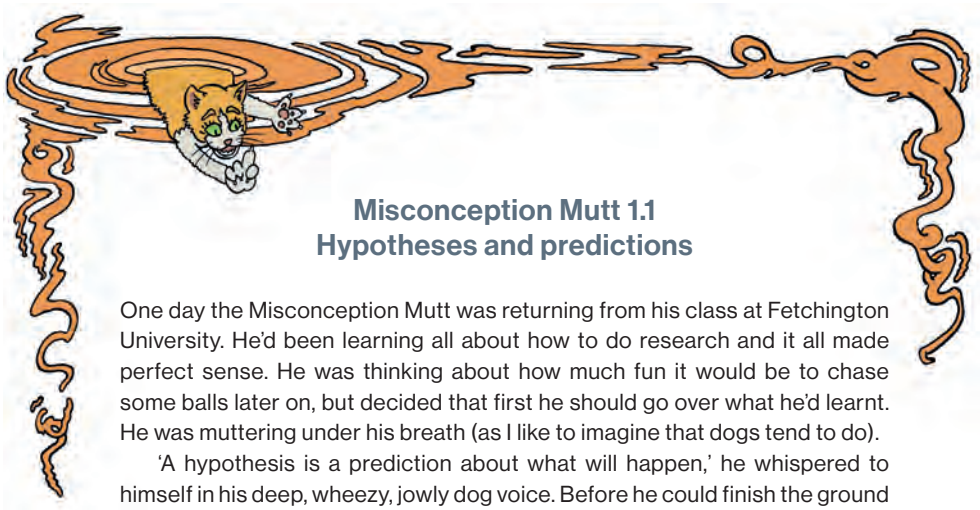
To continue the example, having studied the attachment theory of personality disorders, we might decide that this theory implies that people with personality disorders seek out the attention that a TV appearance provides because they lack close interpersonal relationships. From this we can generate a hypothesis: people with narcissistic personality disorder use reality TV to satisfy their craving for attention from others. This is a conceptual statement that explains our original observation (that rates of narcissistic personality disorder are high on reality TV shows).

To test this hypothesis, we need to move from the conceptual domain into the observable domain. That is, we need to operationalize our hypothesis in a way that enables us to collect and analyse data that have a bearing on the hypothesis (Figure 1.2). We do this using predictions. Predictions emerge from a hypothesis (Misconception Mutt 1.1), and transform it from something unobservable into something that is observable. If our hypothesis is that people with narcissistic personality disorder use reality TV to satisfy their craving for attention from others, then a prediction we could make based on this hypothesis is that people with narcissistic personality disorder are more likely to audition for reality TV than those without. In making this prediction we can move from the conceptual domain into the observable domain, where we can collect evidence.

In this example, our prediction is that people with narcissistic personality disorder are more likely to audition for reality TV than those without. We can measure this prediction by getting a team of clinical psychologists to interview each person at a reality TV audition and diagnose them as having

narcissistic personality disorder or not. The population rates of narcissistic personality disorder are about 1%, so we'd be able to see whether the rate of narcissistic personality disorder is higher at the audition than in the general population. If it is higher then our prediction is correct: a disproportionate number of people with narcissistic personality disorder turned up at the audition. Our prediction, in turn, tells us something about the hypothesis from which it derived.

This is tricky stuff, so let's look at another example. Imagine that, based on a different theory, we generated a different hypothesis. I mentioned earlier that people with narcissistic personality disorder tend to engage in conflict, so a different hypothesis is that producers of reality TV shows select people who have narcissistic personality disorder to be contestants because they believe that conflict makes good TV. As before, to test this hypothesis we need to bring it into the observable domain by generating a prediction from it. The prediction would be that (assuming no bias in the number of people with narcissistic personality disorder applying for the show) a disproportionate number of people with narcissistic personality disorder will be selected by producers to go on the show.



Misconception Mutt 1.1
Hypotheses and predictions

One day the Misconception Mutt was returning from his class at Fetchington University. He'd been learning all about how to do research and it all made perfect sense. He was thinking about how much fun it would be to chase some balls later on, but decided that first he should go over what he'd learnt. He was muttering under his breath (as I like to imagine that dogs tend to do).

'A hypothesis is a prediction about what will happen,' he whispered to himself in his deep, wheezy, jowly dog voice. Before he could finish the ground before him became viscous, as though the earth had transformed into liquid. A slightly irritated-looking ginger cat rose slowly from the puddle.

'Don't even think about chasing me,' he said in his whiny cat voice.

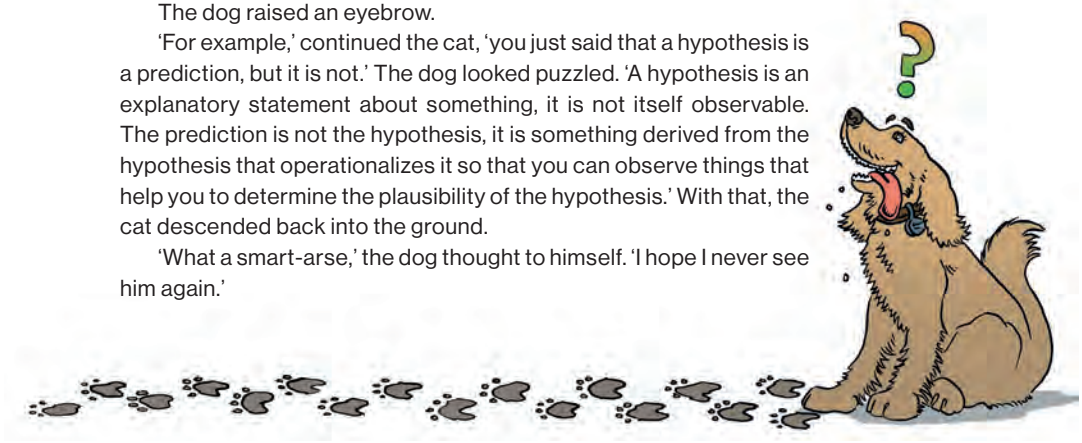
The mutt twitched as he inhibited the urge to chase the cat. 'Who are you?' he asked.

'I am the Correcting Cat,' said the cat wearily. 'I travel the ether trying to correct people's statistical misconceptions. It's very hard work – there are a lot of misconceptions about.'

The dog raised an eyebrow.

'For example,' continued the cat, 'you just said that a hypothesis is a prediction, but it is not.' The dog looked puzzled. 'A hypothesis is an explanatory statement about something, it is not itself observable. The prediction is not the hypothesis, it is something derived from the hypothesis that operationalizes it so that you can observe things that help you to determine the plausibility of the hypothesis.' With that, the cat descended back into the ground.

'What a smart-arse,' the dog thought to himself. 'I hope I never see him again.'





Jane Superbrain 1.1
When is a prediction not a prediction? Ⓐ

A good theory should allow us to make statements about the state of the world. Statements about the world are good things: they allow us to make sense of our world, and to make decisions that affect our future. One current example is climate change. Being able to make a definitive statement that global warming is happening, and that it is caused by certain practices in society, allows us to change these practices and, hopefully, avert catastrophe. However, not all statements can be tested using science. Scientific statements are ones that can be verified with reference to empirical evidence, whereas non-scientific statements are ones that cannot be empirically tested. So, statements such as 'The Led Zeppelin reunion concert in London in 2007 was the best gig ever',[6] 'Lindt chocolate is the best food' and 'This is the worst statistics book in the world' are all non-scientific; they cannot be proved or disproved. Scientific statements can be confirmed or disconfirmed empirically. 'Watching *Curb Your Enthusiasm* makes you happy', 'Having sex increases levels of the neurotransmitter dopamine' and 'Velociraptors ate meat' are all things that can be tested empirically (provided you can quantify and measure the variables concerned). Non-scientific statements can sometimes be altered to become scientific statements, so 'The Beatles were the most influential band ever' is non-scientific (because it is probably impossible to quantify 'influence' in any meaningful way) but by changing the statement to 'The Beatles were the best-selling band ever' it becomes testable (we can collect data about worldwide album sales and establish whether the Beatles have, in fact, sold more records than any other music artist). Karl Popper, the famous philosopher of science, believed that non-scientific statements had no place in science. Good theories and hypotheses should, therefore, produce predictions that are scientific statements.



Imagine we collected the data in Table 1.1, which shows how many people auditioning to be on a reality TV show had narcissistic personality disorder or not. In total, 7662 people turned up for the audition. Our first prediction (derived from our first hypothesis) was that the percentage of people with narcissistic personality disorder will be higher at the audition than the general level in the population. We can see in the table that of the 7662 people at the audition, 854 were diagnosed with the disorder: this is about 11% (854/7662 × 100), which is much higher than the 1% we'd expect in the general population. Therefore, prediction 1 is correct, which in turn supports hypothesis 1. The second prediction was that the producers of reality TV have a bias towards choosing people with narcissistic personality disorder. If we look at the 12 contestants that they selected, 9 of them had the disorder (a massive 75%). If the producers did not have a bias we would have expected only 11% of

6 It was pretty awesome.

the contestants to have the disorder (the same rate as was found when we considered everyone who turned up for the audition). The data are in line with the second prediction, which supports our second hypothesis. Therefore, my initial observation that contestants have personality disorders was verified by data, and then using theory I generated specific hypotheses that were operationalized by generating predictions that could be tested using data. Data are *very* important.

Table 1.1 The number of people at the TV audition split by whether they had narcissistic personality disorder and whether they were selected as contestants by the producers

	No Disorder	Disorder	Total
Selected	3	9	12
Rejected	6805	845	7650
Total	6808	854	7662

I would now be smugly sitting in my office with a contented grin on my face because my hypotheses were well supported by the data. Perhaps I would quit while I was ahead and retire. It's more likely, though, that having solved one great mystery, my excited mind would turn to another. I would lock myself in a room to watch more reality TV. I might wonder at why contestants with narcissistic personality disorder, despite their obvious character flaws, enter a situation that will put them under intense public scrutiny.[7] Days later, the door would open, and a stale odour would waft out like steam rising from the New York subway. Through this green cloud, my bearded face would emerge, my eyes squinting at the shards of light that cut into my pupils. Stumbling forwards, I would open my mouth to lay waste to my scientific rivals with my latest profound hypothesis: 'Contestants with narcissistic personality disorder believe that they will win'. I would croak before collapsing on the floor. The prediction from this hypothesis is that if I ask the contestants if they think that they will win, the people with a personality disorder will say 'yes'.

Let's imagine I tested my hypothesis by measuring contestants' expectations of success in the show, by asking them, 'Do you think you will win?' Let's say that 7 of 9 contestants with narcissistic personality disorder said that they thought that they would win, which confirms my hypothesis. At this point I might start to try to bring my hypotheses together into a theory of reality TV contestants than revolves around the idea that people with narcissistic personalities are drawn towards this kind of show because it fulfils their need for approval and they have unrealistic expectations about their likely success because they don't realize how unpleasant their personalities are to other people. In parallel, producers tend to select contestants with narcissistic tendencies because they tend to generate interpersonal conflict.

One part of my theory is untested, which is the bit about contestants with narcissistic personalities not realizing how others perceive their personality. I could operationalize this hypothesis through a prediction that if I ask these contestants whether their personalities were different from those of other people they would say 'no'. As before, I would collect more data and ask the contestants with narcissistic personality disorder whether they believed that their personalities were different from the norm. Imagine that all 9 of them said that they thought their personalities *were* different from the norm. These data contradict my hypothesis. This is known as **falsification**, which is the act of disproving a hypothesis or theory.

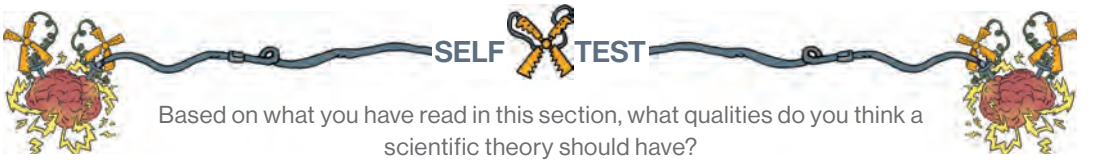
It's unlikely that we would be the only people interested in why individuals who go on reality TV have extreme personalities. Imagine that these other researchers discovered that: (1) people with narcissistic personality disorder think that they are more interesting than others; (2) they also think that

7 One of the things I like about many reality TV shows in the UK is that the most odious people tend to get voted out quickly, which gives me faith that humanity favours the nice.

they deserve success more than others; and (3) they also think that others like them because they have 'special' personalities.

This additional research is even worse news for my theory: if contestants didn't realize that they had a personality different from the norm then you wouldn't expect them to think that they were more interesting than others, and you certainly wouldn't expect them to think that others will *like* their unusual personalities. In general, this means that this part of my theory sucks: it cannot explain the data, predictions from the theory are not supported by subsequent data, and it cannot explain other research findings. At this point I would start to feel intellectually inadequate and people would find me curled up on my desk in floods of tears, wailing and moaning about my failing career (no change there then).

At this point, a rival scientist, Fester Ingpant-Stain, appears on the scene adapting my theory to suggest that the problem is not that personality-disordered contestants don't realize that they have a personality disorder (or at least a personality that is unusual), but that they falsely believe that this special personality is perceived positively by other people. One prediction from this model is that if personality-disordered contestants are asked to evaluate what other people think of them, then they will overestimate other people's positive perceptions. You guessed it – Fester Ingpant-Stain collected yet more data. He asked each contestant to fill out a questionnaire evaluating all of the other contestants' personalities, and also complete the questionnaire about themselves but answering from the perspective of each of their housemates. (So, for every contestant there is a measure of both what they thought of every other contestant, and what they believed every other contestant thought of them.) He found out that the contestants with personality disorders did overestimate their housemates' opinions of them; conversely, the contestants without personality disorders had relatively accurate impressions of what others thought of them. These data, irritating as it would be for me, support Fester Ingpant-Stain's theory more than mine: contestants with personality disorders do realize that they have unusual personalities but believe that these characteristics are ones that others would feel positive about. Fester Ingpant-Stain's theory is quite good: it explains the initial observations and brings together a range of research findings. The end result of this whole process (and my career) is that we should be able to make a general statement about the state of the world. In this case we could state that 'reality TV contestants who have personality disorders overestimate how much other people like their personality characteristics'.



Based on what you have read in this section, what qualities do you think a scientific theory should have?

1.5 Collecting data: measurement Ⓐ

In looking at the process of generating theories and hypotheses, we have seen the importance of data in testing those hypotheses or deciding between competing theories. This section looks at data collection in more detail. First we'll look at measurement.

1.5.1 Independent and dependent variables Ⓐ

To test hypotheses we need to measure variables. Variables are things that can change (or vary); they might vary between people (e.g., IQ, behaviour) or locations (e.g., unemployment) or even time (e.g., mood, profit, number of cancerous cells). Most hypotheses can be expressed in terms of two variables:

a proposed cause and a proposed outcome. For example, if we take the scientific statement 'Coca-Cola is an effective spermicide'[8] then the proposed cause is Coca-Cola and the proposed effect is dead sperm. Both the cause and the outcome are variables: for the cause we could vary the type of drink, and for the outcome, these drinks will kill different amounts of sperm. The key to testing scientific statements is to measure these two variables.

A variable that we think is a cause is known as an **independent variable** (because its value does not depend on any other variables). A variable that we think is an effect is called a **dependent variable** because the value of this variable depends on the cause (independent variable). These terms are very closely tied to experimental methods in which the cause is manipulated by the experimenter (as we will see in Section 1.6.2). However, researchers can't always manipulate variables (for example, if you wanted see whether smoking causes lung cancer you wouldn't lock a bunch of people in a room for 30 years and force them to smoke). Instead, they sometimes use correlational methods (Section 1.6), for which it doesn't make sense to talk of dependent and independent variables because all variables



Cramming Sam's Tips 1.1
Variables

When doing and reading research you're likely to encounter these terms:

- *Independent variable*: A variable thought to be the cause of some effect. This term is usually used in experimental research to describe a variable that the experimenter has manipulated.
- *Dependent variable*: A variable thought to be affected by changes in an independent variable. You can think of this variable as an outcome.
- *Predictor variable*: A variable thought to predict an outcome variable. This term is basically another way of saying independent variable. (Although some people won't like me saying that; I think life would be easier if we talked only about predictors and outcomes.)
- *Outcome variable*: A variable thought to change as a function of changes in a predictor variable. For the sake of an easy life this term could be synonymous with 'dependent variable'.



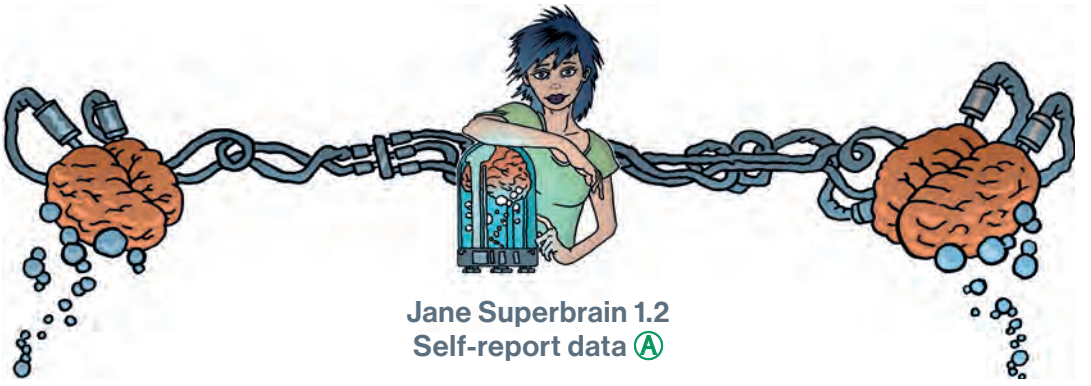
8 There is a long-standing urban myth that a post-coital douche with the contents of a bottle of Coke is an effective contraceptive. Unbelievably, this hypothesis has been tested and Coke does affect sperm motility, and some types of Coke are more effective than others – Diet Coke is best apparently (Umpierre et al., 1985). In case you decide to try this method out, it's worth mentioning that despite the effects on sperm motility a Coke douche is ineffective at preventing pregnancy.

are essentially dependent variables. I prefer to use the terms **predictor variable** and **outcome variable** in place of dependent and independent variable. This is not a personal whimsy: in experimental work the cause (independent variable) is a predictor and the effect (dependent variable) is an outcome, and in correlational work we can talk of one or more (predictor) variables predicting (statistically at least) one or more outcome variables.

1.5.2 Levels of measurement Ⓐ

Variables can take on many different forms and levels of sophistication. The relationship between what is being measured and the numbers that represent what is being measured is known as the **level of measurement**. Broadly speaking, variables can be categorical or continuous, and can have different levels of measurement.

A **categorical variable** is made up of categories. A categorical variable that you should be familiar with already is your species (e.g., human, domestic cat, fruit bat). You are a human or a cat or a fruit bat: you cannot be a bit of a cat and a bit of a bat, and neither a batman nor a catwoman exist other than in graphic novels and movies. A categorical variable is one that names distinct entities. In its simplest form it names just two distinct types of things, for example dog or cat. This is known as a **binary variable**. Other examples of binary variables are being alive or dead, pregnant or not, and responding 'yes' or 'no' to a question. In all cases there are just two categories and an entity can be placed into only one of the two categories. When two things that are equivalent in some sense are given the same name (or number), but there are more than two possibilities, the variable is said to be a **nominal variable**.



Jane Superbrain 1.2
Self-report data Ⓐ

A lot of self-report data are ordinal. Imagine two judges on a TV talent show were asked to rate Billie's singing on a 10-point scale. We might be confident that a judge who gives a rating of 10 found Billie more talented than one who gave a rating of 2, but can we be certain that the first judge found her five times more talented than the second? What if both judges gave a rating of 8; could we be sure they found her equally talented? Probably not: their ratings will depend on their subjective feelings about what constitutes talent (the quality of singing? stage craft? dancing?). For these reasons, in any situation in which we ask people to rate something subjective (e.g., their preference for a product, their confidence about an answer, how much they have understood some medical instructions) we should probably regard these data as ordinal, although many scientists do not.



It should be obvious that if the variable is made up of names it is pointless to do arithmetic on them (if you multiply a human by a cat, you do not get a hat). However, sometimes numbers are used to denote categories. For example, the numbers worn by players in a sports team. In rugby, the numbers on shirts denote specific field positions, so the number 10 is always worn by the fly-half[9] and the number 2 is always the hooker (the battered-looking player at the front of the scrum). These numbers do not tell us anything other than what position the player plays. We could equally have shirts with FH and H instead of 10 and 2. A number 10 player is not necessarily better than a number 2 (most managers would not want their fly-half stuck in the front of the scrum!). It is equally daft to try to do arithmetic with nominal scales where the categories are denoted by numbers: the number 10 takes penalty kicks, and if the coach found that his number 10 was injured he would not get his number 4 to give number 6 a piggy-back and then take the kick. The only way that nominal data can be used is to consider frequencies. For example, we could look at how frequently number 10s score compared to number 4s.

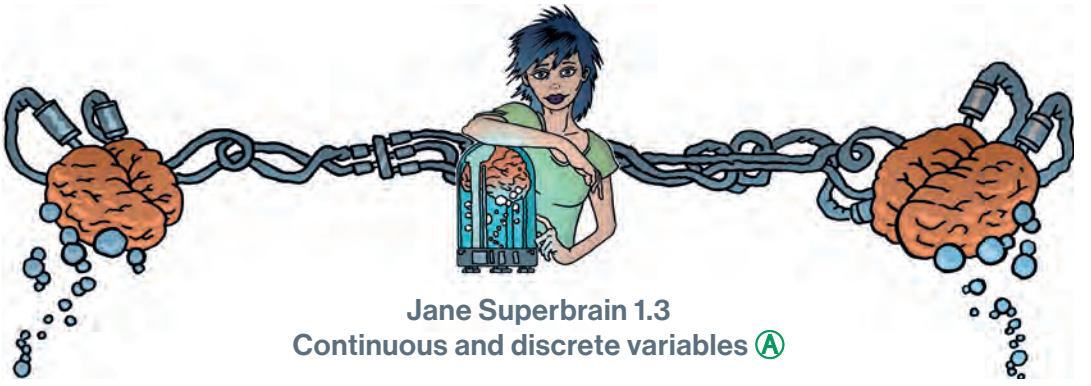
The categorical variables we have considered up to now have been unordered (e.g., different brands of Coke with which you're trying to kill sperm), but they can be ordered too (e.g., increasing concentrations of Coke with which you're trying to skill sperm). When categories are ordered, the variable is known as an **ordinal variable**. Ordinal data tell us not only that things have occurred, but also the order in which they occurred. However, these data tell us nothing about the differences between values. In TV shows like *American Idol* and *The Voice*, hopeful singers compete to win a recording contract. They are hugely popular shows, which could (if you take a depressing view) reflect the fact that Western society values 'opportunity' more than hard work.[10] Imagine that the three winners of a particular series were Billie, Freema and Gemma. Although the names of the winners don't provide any information about where they came in the contest, labelling them according to their performance does: first (Billie), second (Freema) and third (Gemma). These categories are ordered. In using ordered categories we now know that Billie (who won) was better than both Freema (who came second) and Gemma (third). We still know nothing about the differences between categories. We don't, for example, know how much better the winner was than the runners-up: Billie might have been an easy victor, getting many more votes than Freema and Gemma, or it might have been a very close contest that she won by only a single vote. Ordinal data, therefore, tell us more than nominal data (they tell us the order in which things happened) but they still do not tell us about the differences between points on a scale.

The next level of measurement moves us away from categorical variables and into continuous variables. A **continuous variable** is one that gives us a score for each person and can take on any value on the measurement scale that we are using. The first type of continuous variable that you might encounter is an **interval variable**. Interval data are considerably more useful than ordinal data and most of the statistical tests in this book require data measured at this level or above. To say that data are interval, we must be certain that equal intervals on the scale represent equal differences in the property being measured. For example, on *www.ratemyprofessors.com* students are encouraged to rate their lecturers on several dimensions (some of the lecturers' rebuttals of their negative evaluations are worth a look). Each dimension (helpfulness, clarity, etc.) is evaluated using a 5-point scale. For this scale to be interval it must be the case that the difference between helpfulness ratings of 1 and 2 is the same as the difference between (say) 3 and 4, or 4 and 5. Similarly, the difference in helpfulness between ratings of 1 and 3 should be identical to the difference between ratings of 3 and 5. Variables like this that look interval (and are treated as interval) are often ordinal – see Jane Superbrain Box 1.2.

9 Unlike, for example, NFL American football, where a quarterback could wear any number from 1 to 19.

10 I am not bitter at all about ending up as an academic despite years spent learning musical instruments and trying to create original music.

Ratio variables go a step further than interval data by requiring that in addition to the measurement scale meeting the requirements of an interval variable, the ratios of values along the scale should be meaningful. For this to be true, the scale must have a true and meaningful zero point. In our lecturer ratings this would mean that a lecturer rated as 4 would be twice as helpful as a lecturer rated as 2 (who would in turn be twice as helpful as a lecturer rated as 1). The time to respond to something is a good example of a ratio variable. When we measure a reaction time, not only is it true that, say, the difference between 300 and 350 ms (a difference of 50 ms) is the same as the difference between 210 and 260 ms or between 422 and 472 ms, but it is also true that distances along the scale are divisible: a reaction time of 200 ms is twice as long as a reaction time of 100 ms and half as long as a reaction time of 400 ms. Time also has a meaningful zero point: 0 ms means a complete absence of time.



Jane Superbrain 1.3
Continuous and discrete variables Ⓐ

The distinction between discrete and continuous variables can be blurred. For one thing, continuous variables can be measured in discrete terms; for example, when we measure age we rarely use nanoseconds but use years (or possibly years and months). In doing so we turn a continuous variable into a discrete one (the only acceptable values are whole numbers of years). Also, we often treat discrete variables as if they were continuous. For example, the number of partners that you have had is a discrete variable (it will be, in all but the very weirdest cases, a whole number). However, you might read a magazine that says, 'the average number of partners that people in their twenties have has increased from 4.6 to 8.9'. This assumes that the variable is continuous, and of course these averages are meaningless: no one in their sample actually had 8.9 partners.



Continuous variables can be, well, continuous (obviously) but also discrete. This is quite a tricky distinction (Jane Superbrain Box 1.3). A truly continuous variable can be measured to any level of precision, whereas a **discrete variable** can take on only certain values (usually whole numbers, also known as integers) on the scale. Let me explain. The ratings of lecturers on a 5-point scale are an example of a discrete variable. The range of the scale is 1–5, but you can enter only values of 1, 2, 3, 4 or 5; you cannot enter a value of 4.32 or 2.18. Although a continuum plausibly exists underneath the scale (i.e., a rating of 3.24 makes sense), the actual values that the variable takes on are constrained to whole numbers. A true continuous variable would be something like age, which can be measured at an infinite level of precision (you could be 34 years, 7 months, 21 days, 10 hours, 55 minutes, 10 seconds, 100 milliseconds, 63 microseconds, 1 nanosecond old).



Cramming Sam's Tips 1.2
Levels of measurement

Variables can be split into categorical and continuous, and within these types there are different levels of measurement:

- Categorical (entities are divided into distinct categories):
  - Binary variable: There are only two categories (e.g., dead or alive).
  - Nominal variable: There are more than two categories (e.g., whether someone is an omnivore, vegetarian, vegan, or fruitarian).
  - Ordinal variable: The same as a nominal variable but the categories have a logical order (e.g., whether people got a fail, a pass, a merit or a distinction in their exam).
- Continuous (entities get a distinct score):
  - Interval variable: Equal intervals on the variable represent equal differences in the property being measured (e.g., the difference between 6 and 8 is equivalent to the difference between 13 and 15).
  - Ratio variable: The same as an interval variable, but the ratios of scores on the scale must also make sense (e.g., a score of 16 on an anxiety scale means that the person is, in reality, twice as anxious as someone scoring 8). For this to be true, the scale must have a meaningful zero point.



1.5.3 Measurement error Ⓐ

It's one thing to measure variables, but it's another thing to measure them accurately. Ideally we want our measure to be calibrated such that values have the same meaning over time and across situations. Weight is one example: we would expect to weigh the same amount regardless of who weighs us, or where we take the measurement (assuming it's on Earth and not in an anti-gravity chamber). Sometimes variables can be measured directly (profit, weight, height) but in other cases we are forced to use indirect measures such as self-report, questionnaires and computerized tasks (to name a few).

It's been a while since I mentioned sperm, so let's go back to our Coke as a spermicide example. Imagine we took some Coke and some water and added them to two test tubes of sperm. After several minutes, we measured the motility (movement) of the sperm in the two samples and discovered no difference. As you might expect given that Coke and sperm rarely top scientists' research lists, a few years pass before another scientist, Dr Jack Q. Late, replicates the study. Dr Late finds that sperm motility is worse in the Coke sample. There are two measurement-related issues that could explain

his success and our failure: (1) Dr Late might have used more Coke in the test tubes (sperm might need a critical mass of Coke before they are affected); (2) Dr Late measured the outcome (motility) differently than us.

The former point explains why chemists and physicists have devoted many hours to developing standard units of measurement. If we had reported that we'd used 100 ml of Coke and 5 ml of sperm, then if Dr Late used the same quantities we'd know that our studies are comparable because millilitres are a standard unit of measurement. Direct measurements such as the millilitre provide an objective standard: 100 ml of a liquid is known to be twice as much as only 50 ml.

The second reason for the difference in results between the studies could have been to do with how sperm motility was measured. Perhaps in our study we measured motility using absorption spectrophotometry, whereas Dr Late used laser light-scattering techniques.[11] Perhaps his measure is more sensitive than ours.

There will often be a discrepancy between the numbers we use to represent the thing we're measuring and the actual value of the thing we're measuring (i.e., the value we would get if we could measure it directly). This discrepancy is known as **measurement error**. For example, imagine that you know as an absolute truth that you weigh 83 kg. One day you step on the bathroom scales and they read 80 kg. There is a difference of 3 kg between your actual weight and the weight given by your measurement tool (the scales): this is a measurement error of 3 kg. Although properly calibrated bathroom scales should produce only very small measurement errors (despite what we might want to believe when it says we have gained 3 kg), self-report measures will produce larger measurement error because factors other than the one you're trying to measure will influence how people respond to our measures. For example, if you were completing a questionnaire that asked you whether you had stolen from a shop, would you admit it, or might you be tempted to conceal this fact?

1.5.4 Validity and reliability Ⓐ

One way to try to ensure that measurement error is kept to a minimum is to determine properties of the measure that give us confidence that it is doing its job properly. The first property is **validity**, which is whether an instrument measures what it sets out to measure. The second is **reliability**, which is whether an instrument can be interpreted consistently across different situations.

Validity refers to whether an instrument measures what it was designed to measure (e.g., does your lecturer helpfulness rating scale actually measure lecturers' helpfulness?); a device for measuring sperm *motility* that actually measures sperm *count* is not valid. Things like reaction times and physiological measures are valid in the sense that a reaction time does in fact measure the time taken to react and skin conductance does measure the conductivity of your skin. However, if we're using these things to infer other things (e.g., using skin conductance to measure anxiety) then they will be valid only if there are no factors other than the one we're interested in that can influence them.

Criterion validity is whether an instrument measures what it claims to measure through comparison to objective criteria. In an ideal world, you assess this criterion by relating scores on your measure to real-world observations. For example, we could take an objective measure of how helpful lecturers were and compare these observations to students' ratings of helpfulness on *ratemyprofessors.com*. When data are recorded simultaneously using the new instrument and existing criteria, then this is said to assess **concurrent validity**; when data from the new instrument are used to predict observations at a later point in time, this is said to assess **predictive validity**.

Assessing criterion validity (whether concurrently or predictively) is often impractical because objective criteria that can be measured easily may not exist. Also, with measuring something like attitudes you might be interested in the person's perception of reality and not reality itself (you might

11 In the course of writing this chapter I have discovered more than is healthy about the measurement of sperm motility.

not care whether a person *was* treated unjustly but whether they *think* they were treated unjustly). With self-report measures/questionnaires we can talk about **content validity**, which is the degree to which individual items represent the construct being measured and tap all aspects of it.

Validity is a necessary but not sufficient condition of a measure. A second consideration is reliability, which is the ability of the measure to produce the same results under the same conditions. To be valid the instrument must first be reliable. The easiest way to assess reliability is to test the same group of people twice: a reliable instrument will produce similar scores at both points in time (**test–retest reliability**). Sometimes, however, you will want to measure something that does vary over time (e.g., moods, blood-sugar levels, productivity). Statistical methods can also be used to determine reliability (we will discover these in Chapter 18).



1.6 Collecting data: research design Ⓐ

We've looked at the question of *what* to measure and discovered that to answer scientific questions we measure variables (which can be collections of numbers or words). We also saw that to get accurate answers we need accurate measures. We move on now to look at research design: *how* data are collected. If we simplify things quite a lot then there are two ways to test a hypothesis: either by observing what naturally happens, or by manipulating some aspect of the environment and observing the effect it has on the variable that interests us. In **correlational** or **cross-sectional research** we observe what naturally goes on in the world without directly interfering with it, whereas in **experimental research** we manipulate one variable to see its effect on another.

1.6.1 Correlational research methods Ⓐ



In correlational research we observe natural events; we can do this either by taking a snapshot of many variables at a single point in time, or by measuring variables repeatedly at different time points (known as **longitudinal research**). For example, we might measure pollution levels in a stream and the numbers of certain types of fish living there; lifestyle variables (alcohol intake, exercise, food intake) and disease (cancer, diabetes); workers' job satisfaction under different managers; or children's school performance across regions with different demographics. Correlational research provides a very natural view of the question we're researching because we're not influencing what happens and the measures of the variables should not be biased by the researcher being there (this is an important aspect of **ecological validity**).

At the risk of sounding absolutely obsessed with using Coke as a contraceptive (I'm not, but the discovery that people in the 1950s and 1960s actually thought this was a good idea has intrigued me), let's return to that example. If we wanted to answer the question 'Is Coke an effective contraceptive?' we could administer questionnaires about sexual practices (quantity of sexual activity, use of contraceptives, use of fizzy drinks as contraceptives, pregnancy, that sort of thing). By looking at these variables we could see which variables correlate with pregnancy and, in particular, whether those reliant on Coca-Cola as a form of contraceptive were more likely to end up pregnant than those using other contraceptives, and less likely than those using no contraceptives at all. This is the only way to answer a question like this because we cannot manipulate any of these variables particularly easily.

For example, it would be unethical to require anyone to do something (like using Coke as a contraceptive) that might lead to an unintended pregnancy. There is a price to pay with observational research, which relates to causality: correlational research tells us nothing about the causal influence of variables.

1.6.2 Experimental research methods Ⓐ

Most scientific questions imply a causal link between variables; we have seen already that dependent and independent variables are named such that a causal connection is implied (the dependent variable *depends* on the independent variable). Sometimes the causal link is very obvious, as in the research question 'Does low self-esteem cause dating anxiety?' Sometimes the implication might be subtler; for example, in 'Is dating anxiety all in the mind?' the implication is that a person's mental outlook causes them to be anxious when dating. Even when the cause – effect relationship is not explicitly stated, most research questions can be broken down into a proposed cause (in this case mental outlook) and a proposed outcome (dating anxiety). Both the cause and the outcome are variables: for the cause the degree to which people perceive themselves in a negative way varies; and for the outcome, the degree to which people feel anxious on dates varies. The key to answering the research question is to uncover how the proposed cause and the proposed outcome relate to each other; are the people with lower self-esteem the same people who are more anxious on dates?

David Hume, an influential philosopher, defined a cause as 'An object precedent and contiguous to another, and where all the objects resembling the former are placed in like relations of precedency and contiguity to those objects that resemble the latter' (Hume, 1739).[12] This definition implies that (1) the cause needs to precede the effect, and (2) causality is equated to high degrees of correlation between contiguous events. In our dating example, to infer that low self-esteem caused dating anxiety, it would be sufficient to find that lower levels of self-esteem co-occur with higher levels of anxiety when on a date, and that the low self-esteem emerged before the dating anxiety.

In correlational research, variables are often measured simultaneously. The first problem that this creates is that there is no information about the contiguity between different variables: we might find from a questionnaire study that people with low self-esteem have high dating anxiety but we wouldn't know whether the low self-esteem or high dating anxiety came first. Longitudinal research addresses this issue to some extent but there is a still a problem with Hume's idea that causality can be inferred from corroborating evidence, which is that it doesn't distinguish between what you might call an 'accidental' conjunction and a causal one. For example, it could be that both low self-esteem and dating anxiety are caused by a third variable. For example, difficulty with social skills might make you feel generally worthless and also ramp up the pressure you feel in dating situations. Perhaps, then, low self-esteem and dating anxiety co-occur (meeting Hume's definition of cause) because difficulty with social skills causes them both.

This example illustrates an important limitation of correlational research: the ***tertium quid*** ('a third person or thing of indeterminate character'). For example, a correlation has been found between sexual infidelity and poorer mental health (e.g., Wenger & Frisco, 2021). However, there are external factors likely to contribute to both; for example, the quality of a romantic relationship plausibly affects both the likelihood of infidelity and the mental health of those within the partnership. These extraneous factors are called **confounding variables** or 'confounds' for short.

The shortcomings of Hume's definition led John Stuart Mill (1865) to suggest that in addition to a correlation between events, all other explanations of the cause–effect relationship must be ruled out. To rule out confounding variables, Mill proposed that an effect should be present when the cause is present and that when the cause is absent the effect should be absent also. In other words, the only way to infer causality is through comparing two controlled situations: one in which the cause is present and one in which the cause is absent. This is what *experimental methods* strive to do: to

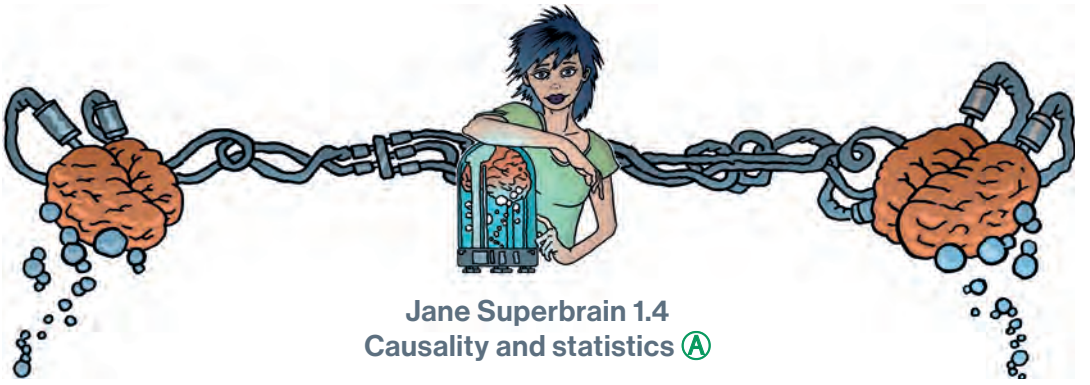
12 As you might imagine, his view was a lot more complicated than this definition, but let's not get sucked down that particular wormhole.

provide a comparison of situations (usually called *treatments* or *conditions*) in which the proposed cause is present or absent.

As a simple case, we might want to look at the effect of feedback style on learning about statistics. I might, therefore, randomly split[13] some students into three different groups in which I change my style of feedback in the seminars on my course:

- **Group 1 (supportive feedback)**: During seminars I congratulate all students in this group on their hard work and success. Even when they get things wrong, I am supportive and say things like, 'That was very nearly the right answer, you're coming along really well' and then give them a piece of chocolate.
- **Group 2 (harsh feedback)**: This group receives seminars in which I give relentless verbal abuse to all of the students even when they give the correct answer. I demean their contributions and am patronizing and dismissive of everything they say. I tell students that they are stupid, worthless and shouldn't be doing the course at all. In other words, this group receives normal university-style seminars☺.
- **Group 3 (no feedback)**: Students get no feedback at all.

The thing that I have manipulated is the feedback style (supportive, harsh or none). As we have seen, this variable is known as the independent variable and in this situation it is said to have three levels, because it has been manipulated in three ways (i.e., the feedback style has been split into three types:



Jane Superbrain 1.4
Causality and statistics Ⓐ

People sometimes get confused and think that certain statistical procedures allow causal inferences and others don't. This isn't true, it's the fact that in experiments we manipulate the causal variable systematically to see its effect on an outcome (the effect). In correlational research we observe the co-occurrence of variables; we do not manipulate the causal variable first and then measure the effect, therefore we cannot compare the effect when the causal variable is present against when it is absent. In short, we cannot say which variable causes a change in the other; we can merely say that the variables co-occur in a certain way. The reason why some people think that certain statistical tests allow causal inferences is that historically certain tests (e.g., ANOVA, *t*-tests) have been used to analyse experimental research, whereas others (e.g., regression, correlation) have been used to analyse correlational research (Cronbach, 1957). As you'll discover, these statistical procedures are, in fact, mathematically identical.



13 This random assignment of students is important, but we'll get to that later.

supportive, harsh and none). The outcome in which I am interested is statistical ability, and I could measure this variable using a statistics exam after the last seminar. As we have seen, this outcome variable is the dependent variable because we assume that these scores will depend upon the type of teaching method used (the independent variable). The critical thing here is the inclusion of the 'no feedback' group because this is a group in which our proposed cause (feedback) is absent, and we can compare the outcome in this group against the two situations in which the proposed cause is present. If the statistics scores are different in each of the feedback groups (cause is present) compared to the group for which no feedback was given (cause is absent) then this difference can be attributed to the type of feedback used. In other words, the style of feedback used caused a difference in statistics scores (Jane Superbrain Box 1.4).

1.6.3 Two methods of data collection Ⓐ

When we use an experiment to collect data, there are two ways to manipulate the independent variable. The first is to test different entities. This method is the one described above, in which different groups of entities take part in each experimental condition (a **between-groups**, **between-subjects**, or **independent design**). An alternative is to manipulate the independent variable using the same entities. In our motivation example, this means that we give a group of students supportive feedback for a few weeks and test their statistical abilities and then give this same group harsh feedback for a few weeks before testing them again, and then finally give them no feedback and test them for a third time (a **within-subject** or **repeated-measures design**). As you will discover, the way in which the data are collected determines the type of test that is used to analyse the data.

1.6.4 Two types of variation Ⓐ

Imagine we were trying to see whether you could train chimpanzees to run the economy. In one training phase they are sat in front of a chimp-friendly computer and press buttons that change various parameters of the economy; once these parameters have been changed a figure appears on the screen indicating the economic growth resulting from those parameters. Now, chimps can't read (I don't think), so this feedback is meaningless. A second training phase is the same except that if the economic growth is good, they get a banana (if growth is bad they do not) – this feedback is valuable to the average chimp. This is a repeated-measures design with two conditions: the same chimps participate in condition 1 *and* in condition 2.

Let's take a step back and think what would happen if we did *not* introduce an experimental manipulation (i.e., there were no bananas in the second training phase so condition 1 and condition 2 were identical). If there is no experimental manipulation then we expect a chimp's behaviour to be similar in both conditions. We expect this because external factors such as age, sex, IQ, motivation and arousal will be the same for both conditions (a chimp's biological sex etc. will not change from when they are tested in condition 1 to when they are tested in condition 2). If the performance measure (i.e., our test of how well they run the economy) is reliable, and the variable or characteristic that we are measuring (in this case ability to run an economy) remains stable over time, then a participant's performance in condition 1 should be very highly related to their performance in condition 2. So, chimps who score highly in condition 1 will also score highly in condition 2, and those who have low scores for condition 1 will have low scores in condition 2. However, performance won't be *identical*; there will be small differences in performance created by unknown factors. This variation in performance is known as **unsystematic variation**.

If we introduce an experimental manipulation (i.e., provide bananas as feedback in one of the training sessions), then we do something different to participants in condition 1 than in condition 2. So, the *only* difference between conditions 1 and 2 is the manipulation that the experimenter has made (in this case that the chimps get bananas as a positive reward in one condition but not in the

other).[14] Therefore, any differences between the means of the two conditions are probably due to the experimental manipulation. So, if the chimps perform better in one training phase than in the other then this *has* to be due to the fact that bananas were used to provide feedback in one training phase but not the other. Differences in performance created by a specific experimental manipulation are known as **systematic variation**.

Now let's think about what happens when we use different participants – an independent design. In this design we still have two conditions, but this time different participants participate in each condition. Going back to our example, one group of chimps receives training without feedback, whereas a second group of different chimps does receive feedback on their performance via bananas.[15] Imagine again that we didn't have an experimental manipulation. If we did nothing to the groups, then we would still find some variation in behaviour between the groups because they contain different chimps who will vary in their ability, motivation, propensity to get distracted from running the economy by throwing their own faeces, and other factors. In short, the factors that were held constant in the repeated-measures design are free to vary in the independent design. So, the unsystematic variation will be bigger than for a repeated-measures design. As before, if we introduce a manipulation (i.e., bananas) then we will see additional variation created by this manipulation. As such, in both the repeated-measures design and the independent design there are always two sources of variation:

- **Systematic variation**: This variation is due to the experimenter doing something in one condition but not in the other condition.
- **Unsystematic variation**: This variation results from random factors that exist between the experimental conditions (natural differences in ability, the time of day, etc.).

Statistical tests are often based on the idea of estimating how much variation there is in performance, and comparing how much of this is systematic to how much is unsystematic.

In a repeated-measures design, differences between two conditions can be caused by only two things: (1) the manipulation that was carried out on the participants, or (2) any other factor that might affect the way in which an entity performs from one time to the next. The latter factor is likely to be fairly minor compared to the influence of the experimental manipulation. In an independent design, differences between the two conditions can also be caused by one of two things: (1) the manipulation that was carried out on the participants, or (2) differences between the characteristics of the entities allocated to each of the groups. The latter factor in this instance is likely to create considerable random variation both within each condition and between them. When we look at the effect of our experimental manipulation, it is always against a background of 'noise' created by random, uncontrollable differences between our conditions. In a repeated-measures design this 'noise' is kept to a minimum and so the effect of the experiment is more likely to show up. This means that, other things being equal, repeated-measures designs are more sensitive to detect effects than independent designs.

1.6.5 Randomization Ⓐ

In both repeated-measures and independent designs it is important to try to keep the unsystematic variation to a minimum. By keeping the unsystematic variation as small as possible we get a more sensitive measure of the experimental manipulation. Generally, scientists use the **randomization** of

14 Actually, this isn't the only difference because by condition 2 they have had some practice (in condition 1) at running the economy; however, we will see shortly that these practice effects are easily minimized.

15 Obviously I mean they received a banana as a reward for their correct response and not that the bananas develop little banana mouths that sing them a little congratulatory song.

entities to treatment conditions to achieve this goal. Many statistical tests work by identifying the systematic and unsystematic sources of variation and then comparing them. This comparison allows us to see whether the experiment has generated considerably more variation than we would have got had we just tested participants without the experimental manipulation. Randomization is important because it eliminates most other sources of systematic variation, which allows us to be sure that any systematic variation between experimental conditions is due to the manipulation of the independent variable. We can use randomization in two different ways depending on whether we have an independent or repeated-measures design.

Let's look at a repeated-measures design first. I mentioned earlier (in a footnote) that when the same entities participate in more than one experimental condition they are naive during the first experimental condition but they come to the second experimental condition with prior experience of what is expected of them. At the very least they will be familiar with the dependent measure (e.g., the task they're performing). The two most important sources of systematic variation in this type of design are:

- **Practice effects**: Participants may perform differently in the second condition because of familiarity with the experimental situation and/or the measures being used.
- **Boredom effects**: Participants may perform differently in the second condition because they are tired or bored from having completed the first condition.

Although these effects are impossible to eliminate completely, we can ensure that they produce no systematic variation between our conditions by **counterbalancing** the order in which a person participates in a condition.

We can use randomization to determine in which order the conditions are completed. That is, we determine randomly whether a participant completes condition 1 before condition 2, or condition 2 before condition 1. Let's look at the teaching method example and imagine that there were just two conditions: no feedback and harsh feedback. If the same participants were used in all conditions, then we might find that statistical ability was higher after the harsh feedback. However, if every student experienced the harsh feedback after the no feedback seminars then they would enter the harsh condition already having a better knowledge of statistics than when they began the no feedback condition. So, the apparent improvement after harsh feedback would not be due to the experimental manipulation (i.e., it's not because harsh feedback works), but because participants had attended more statistics seminars by the end of the harsh feedback condition compared to the no feedback one. We can ensure that the number of statistics seminars does not introduce a systematic bias by assigning students randomly to have the harsh feedback seminars first or the no feedback seminars first.

If we turn our attention to independent designs, a similar argument can be applied. We know that participants in different experimental conditions will differ in many respects (their IQ, attention span, etc.). We know that these confounding variables contribute to the variation between conditions, but we need to make sure that these variables contribute to the unsystematic variation and *not* to the systematic variation. For example, imagine you wanted to test the effectiveness of teaching analytic thinking skills as a way to reduce beliefs in fake news. You have two groups: the first attends workshop that aims to teach critical and analytic thinking skills, the second does not (this is the control group). Delusion-proneness refers to a person's tendency to endorse unusual ideas (conspiracy theories, unusual explanations for events, etc.) and people scoring high on this trait tend to find it more difficult to engage in critical and analytic thinking (Bronstein et al., 2019). Let's imagine that you go to a local meetup of the Conspiracy Theory Society and recruit them into your analytic thinking workshop, and then go to a local sports club to recruit the control group. You find that your analytic thinking workshop does not seem to reduce beliefs in fake news. However, this finding could be because (1) analytic thinking does not reduce the tendency to believe fake news, or (2) the conspiracy theorists who you recruited to the workshop were unable or unwilling to learn analytic thinking. You have no way to dissociate these explanations because the groups varied not only on whether they attended the workshop, but also on their delusion-proneness (and, therefore, ability to engage in analytic thought).

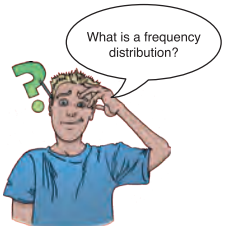
Put another way, the systematic variation created by the participants' delusion-proneness cannot be separated from the effect of the experimental manipulation. The best way to reduce this eventuality is to allocate participants to conditions randomly: by doing so you minimize the risk that groups differ on variables other than the one you want to manipulate.



1.7 Analysing data Ⓐ

The final stage of the research process is to analyse the data you have collected. When the data are quantitative this involves both looking at your data graphically (Chapter 5) to see what the general trends in the data are, and fitting statistical models to the data (all other chapters). Given that the rest of book is dedicated to this process, we'll begin here by looking at a few common ways to look at and summarize the data you have collected.

1.7.1 Frequency distributions Ⓐ



Once you've collected some data a very useful thing to do is to plot how many times each score occurs. A **frequency distribution**, or **histogram**, plots values of observations on the horizontal axis, with a bar the height of which shows how many times each value occurred in the data set. Frequency distributions are very useful for assessing properties of the distribution of scores. We will find out how to create these plots in Chapter 1.

Frequency distributions come in many different shapes and sizes. It is quite important, therefore, to have some general descriptions for common types of distributions. First, we could look at whether the data are distributed symmetrically around the centre of all scores. In other words, if we drew a vertical line through the centre of the distribution does it look the same on both sides? A well-known example of a symmetrical distribution is the **normal distribution**, which is characterized by a bell-shaped curve (Figure 1.3). This shape implies that the majority of scores lie around the centre of the distribution (so the largest bars on the histogram are around the central value). As we get further away from the centre the bars get smaller, implying that as scores start to deviate from the centre their frequency decreases. As we move still further away from the centre scores become very infrequent (the bars are very short). Many naturally occurring things have this shape of distribution. For example, most men in the UK are around 175 cm tall;[16] some are a bit taller or shorter, but most cluster around this value. There will be very few men who are really tall (i.e., above 205 cm) or really short (i.e., under 145 cm).

There are two main ways in which a distribution can deviate from normal: (1) lack of symmetry (called **skew**) and (2) tail density (called **kurtosis**). Skewed distributions are not symmetrical and instead the most frequent scores (the tall bars on the graph) are clustered at one end of the scale. So, the typical pattern is a cluster of frequent scores at one end of the scale and the frequency of scores tailing off towards the other end of the scale. A skewed distribution can be either **positively skewed**

16 I am exactly 180 cm tall. In my home country this makes me smugly above average. However, in the Netherlands the average male height is 185 cm (a massive 10 cm higher than the UK), so I feel like a child when I visit.

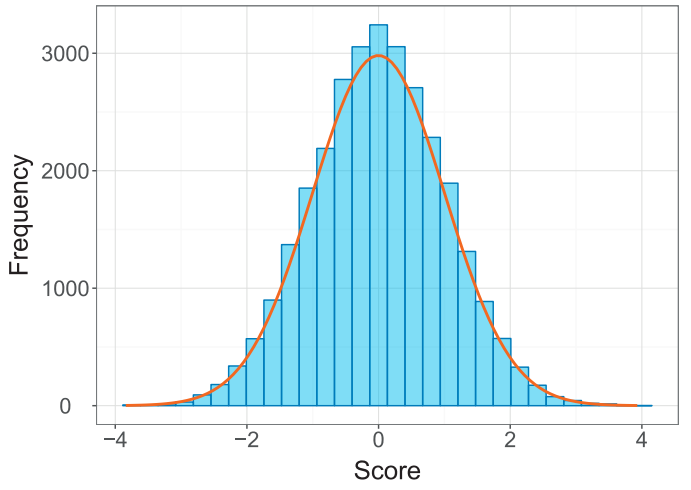


Figure 1.3 A 'normal' distribution (the curve shows the idealized shape)

(the frequent scores are clustered at the lower end and the tail points towards the higher or more positive scores) or **negatively skewed** (the frequent scores are clustered at the higher end and the tail points towards the lower or more negative scores). Figure 1.4 shows examples of these distributions.

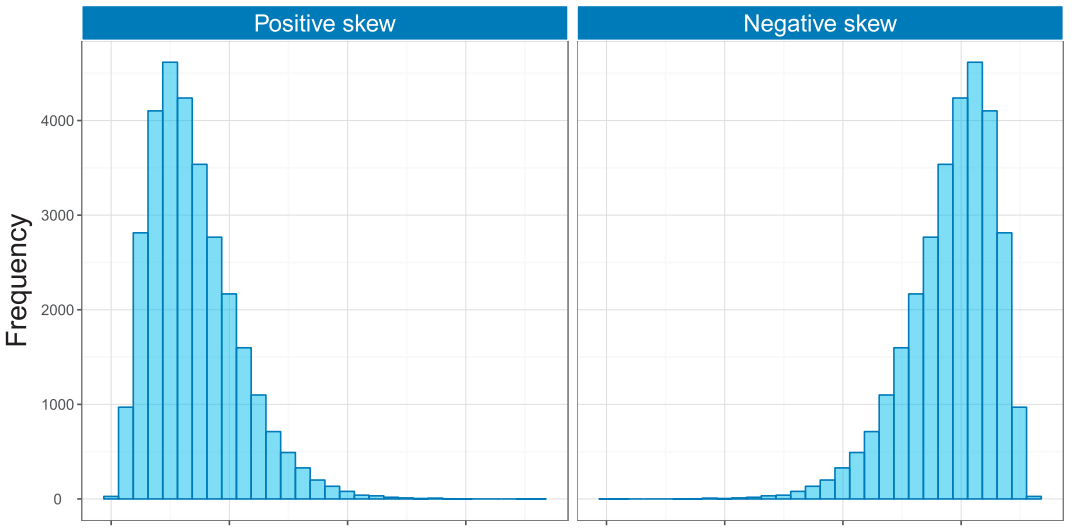


Figure 1.4 A positively (left) and negatively (right) skewed distribution

Distributions also vary in their **kurtosis**. Kurtosis, despite sounding like some kind of exotic disease, refers to the degree to which scores cluster at the ends of the distribution (known as the *tails*). A distribution with *positive kurtosis* has too many scores in the tails (a so-called heavy-tailed distribution). This is known as a **leptokurtic** distribution. In contrast, a distribution with *negative kurtosis* has too few scores in the tails (a so-called light-tailed distribution). This distribution is called **platykurtic**. Heavy-tailed distributions can often look more pointy than normal whereas light-tailed distributions sometimes look flatter, which has led to the widespread conflation (including earlier editions of this book) of kurtosis with how pointed or flat a distribution is; however, kurtosis can only reasonably be seen as a measure of tail density (Westfall, 2014).

When there is no deviation from normal (i.e., the distribution is symmetrical and the tails are as they should be), the value of skew is 0 and the value of kurtosis is 3. However, IBM SPSS Statistics expresses kurtosis in terms of 'excess kurtosis'; that is, how different kurtosis is from the expected value of 3. Therefore, when using SPSS Statistics software you can interpret values of skew or excess kurtosis that differ from 0 as indicative of a deviation from normal: Figure 1.5 shows distributions with kurtosis values of +2.6 (left panel) and –0.09 (right panel).

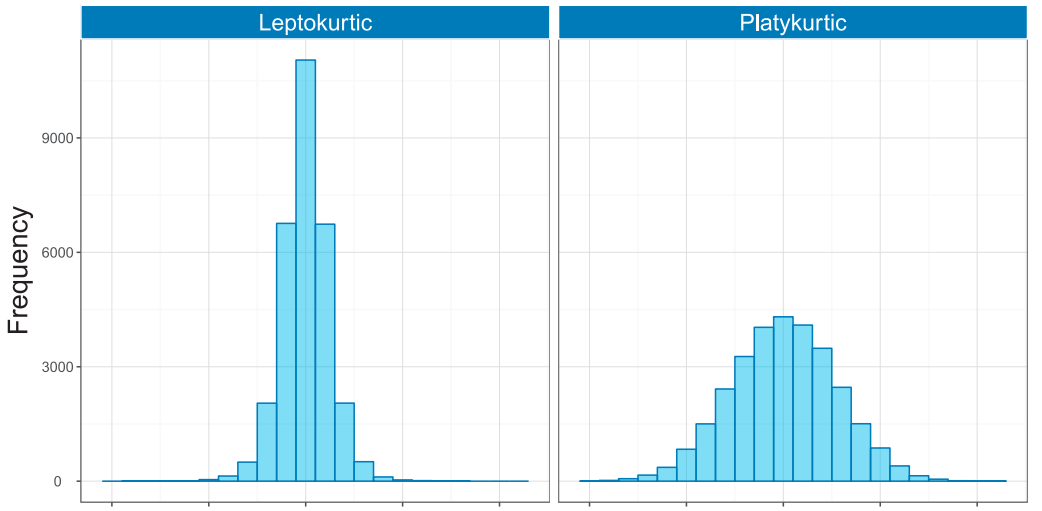


Figure 1.5 Distributions with positive kurtosis (leptokurtic, left) and negative kurtosis (platykurtic, right)

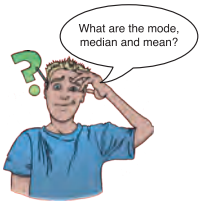
1.7.2 The mode Ⓐ

We can calculate where the centre of a frequency distribution lies (known as the **central tendency**) using three measures commonly used: the mean, the mode and the median. Other methods exist, but these three are the ones you're most likely to come across.

The **mode** is the score that occurs most frequently in the data set. This is easy to spot in a frequency distribution because it will coincide with the tallest bar. To calculate the mode, place the data in ascending order (to make life easier), count how many times each score occurs, and the score that occurs the most is the mode. One problem with the mode is that it can take on several values. For example, Figure 1.6 shows an example of a distribution with two modes (there are two bars that are the highest), which is said to be **bimodal**, and three modes (data sets with more than two modes are **multimodal**). Also, if the frequencies of certain scores are very similar, then the mode can be influenced by only a small number of cases.

1.7.3 The median Ⓐ

Another way to quantify the centre of a distribution is to look for the middle score when scores are ranked in order of magnitude. This is called the **median**. Imagine we looked at the number of followers that 11 users of the social networking website Instagram had. Figure 1.7 shows the number of followers for each of the 11 users: 57, 40, 103, 234, 93, 53, 116, 98, 108, 121, 22.



To calculate the median, we first arrange these scores into ascending order: 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 234.

Next, we find the position of the middle score by counting the number of scores we have collected (n), adding 1 to this value, and then dividing by 2. With 11 scores, this gives us

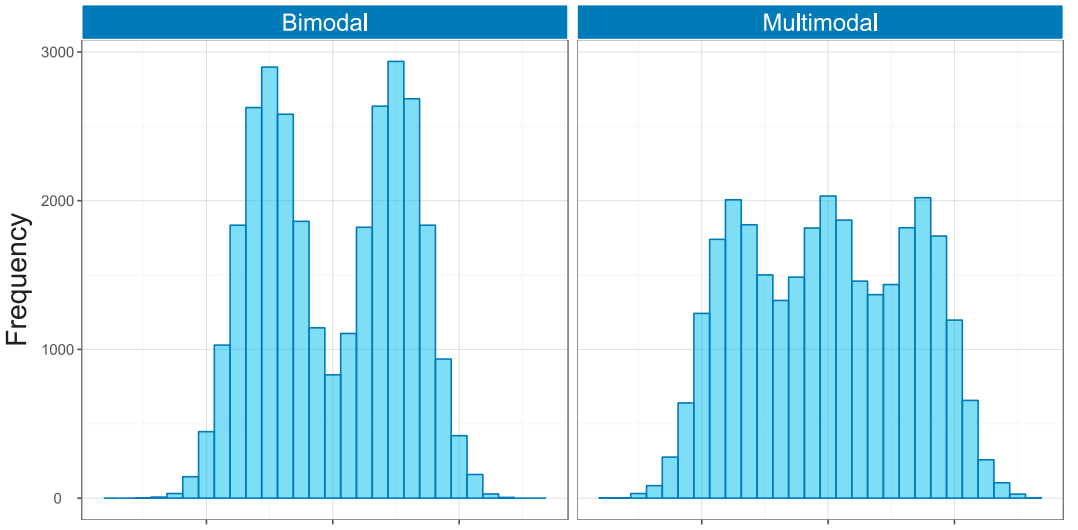


Figure 1.6 Examples of bimodal (left) and multimodal (right) distributions

$(n + 1)/2 = (11 + 1)/2 = 12/2 = 6$. Then, we find the score that is positioned at the location we have just calculated. So, in this example we find the sixth score (see Figure 1.7).

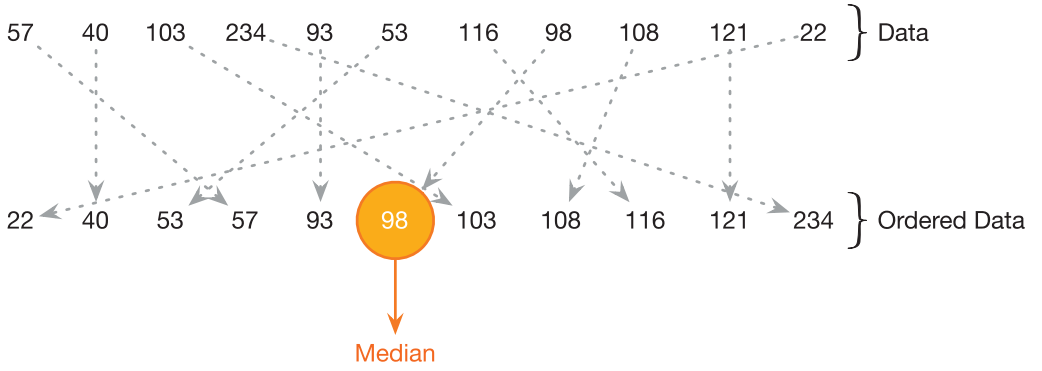


Figure 1.7 The median is simply the middle score when you order the data

This process works very nicely when we have an odd number of scores (as in this example), but when we have an even number of scores there won't be a middle value. Let's imagine that we decided that because the highest score was so big (almost twice as large as the next biggest number), we would ignore it. (For one thing, this person is far too popular and we hate them.) We have only 10 scores now. Figure 1.8 shows this situation. As before, we rank-order these scores: 22, 40, 53, 57, 93, 98, 103, 108, 116, 121. We then calculate the position of the middle score, but this time it is $(n + 1)/2 = 11/2 = 5.5$, which means that the median is halfway between the fifth and sixth scores. To get the median we add these two scores and divide by 2. In this example, the fifth score in the ordered list was 93 and the sixth score was 98. We add these together (93 + 98 = 191) and then divide this value by 2 (191/2 = 95.5). The median number of friends was, therefore, 95.5.

The median is relatively unaffected by extreme scores at either end of the distribution: the median changed only from 98 to 95.5 when we removed the extreme score of 234. The median is also relatively unaffected by skewed distributions and can be used with ordinal, interval and ratio data (it cannot, however, be used with nominal data because these data have no numerical order).

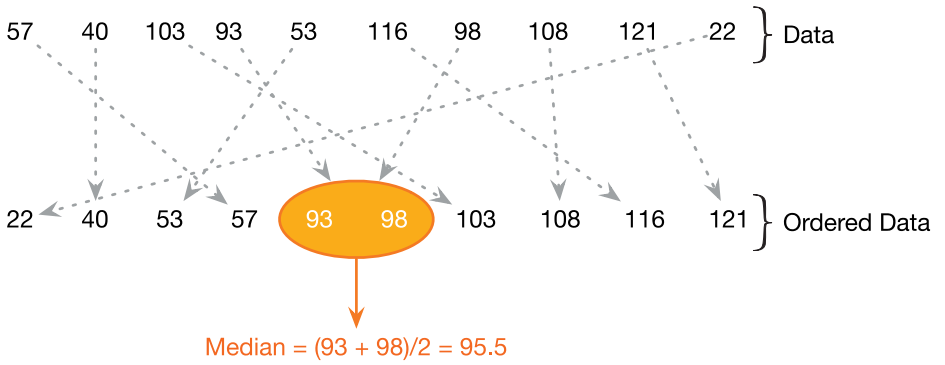


Figure 1.8 When the data contain an even number of scores, the median is the average of the middle two values

1.7.4 The mean Ⓐ

The **mean** is the measure of central tendency that you are most likely to have heard of because it is the average score, and the media love an average score.[17] To calculate the mean we add up the scores and then divide by the total number of scores we have. We can write this in equation form as:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

This equation may look complicated, but the top half means 'add up all of the scores' (the x_i means 'the score of a particular person'; we could replace the letter i with each person's name), and the bottom bit means divide this total by the number of scores you have got (n). Let's calculate the mean for the Instagram data. First, we first add up all the scores:

$$\sum_{i=1}^{n} x_i = 22 + 40 + 53 + 57 + 93 + 98 + 103 + 108 + 116 + 121 + 234 = 1045 \tag{1.2}$$

We then divide by the number of scores (in this case 11):

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1045}{11} = 95 \tag{1.3}$$



17 I originally wrote this on 15 February, and to prove my point the BBC website ran a headline on that day about how PayPal estimates that Britons will spend an average of £71.25 each on Valentine's Day gifts. However, uSwitch.com said that the average spend would be only £22.69. Always remember that the media are full of lies and contradictions.

The mean is 95 friends, which is not a value we observed in our actual data. In this sense the mean is a statistical model – more on this in the next chapter.

If you calculate the mean without our most popular person (i.e., excluding the value 234), the mean drops to 81.1 friends. This reduction illustrates one disadvantage of the mean: it can be influenced by extreme scores. In this case, the person with 234 followers on Instagram increased the mean by about 14 followers; compare this difference with that of the median. Remember that the median changed very little – from 98 to 95.5 – when we excluded the score of 234, which illustrates how the median is typically less affected by extreme scores than the mean. While we're being negative about the mean, it is also affected by skewed distributions and can be used only with interval or ratio data.

If the mean is so lousy then why is it so popular? One very important reason is that it uses every score (the mode and median ignore most of the scores in a data set). Also, the mean tends to be stable in different samples (more on that later too).



Cramming Sam's Tips 1.3
Central tendency

- The mean is the sum of all scores divided by the number of scores. The value of the mean can be influenced quite heavily by extreme scores.
- The median is the middle score when the scores are placed in ascending order. It is not as influenced by extreme scores as the mean.
- The mode is the score that occurs most frequently.



1.7.5 The dispersion in a distribution Ⓐ

It's also essential to quantify the spread, or dispersion, of scores. The easiest way to look at dispersion is to take the largest score and subtract from it the smallest score. This is known as the **range** of scores. For our Instagram data we saw that if we order the scores we get 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 234. The highest score is 234 and the lowest is 22; therefore, the range is 234 – 22 = 212. The range is affected dramatically by extreme scores because it uses only the highest and lowest score.



Compute the range but excluding the score of 234.

From the self-test you'll see that without the extreme score the range is less than half the size: it drops from 212 (with the extreme score included) to 99.

One way around this problem is to calculate the range excluding values at the extremes of the distribution. One convention is to cut off the top and bottom 25% of scores and calculate the range of the middle 50% of scores – known as the **interquartile range**. Let's do this with the Instagram data. First we need to calculate values called **quartiles**. Quartiles are the three values that split the sorted data into four equal parts. First we calculate the median, which is also called the *second quartile*, which splits our data into two equal parts. We already know that the median for these data is 98. The **lower quartile** is the median of the lower half of the data and the **upper quartile** is the median of the upper half of the data.[18] As a rule of thumb the median is not included in the two halves when they are split (this is convenient if you have an odd number of values), but you can include it (although which half you put it in is another question). Figure 1.9 shows how we would calculate these values for the Instagram data. Like the median, if each half of the data had an even number of values in it then the upper and lower quartiles would be the average of two values in the data set (therefore, the upper and lower quartile need not be values that actually appear in the data). Once we have worked out the values of the quartiles, we can calculate the interquartile range, which is the difference between the upper and lower quartile. For the Instagram data this value would be 116 – 53 = 63. The advantage of the interquartile range is that it isn't affected by extreme scores at either end of the distribution. However, the problem with it is that you lose half of the scores.

It's worth noting that quartiles are special cases of things called **quantiles**. Quantiles are values that split a data set into equal portions. Quartiles are quantiles that split the data into four equal parts, but there are other quantiles such as **percentiles** (points that split the data into 100 equal parts), **noniles** (points that split the data into nine equal parts) and so on.

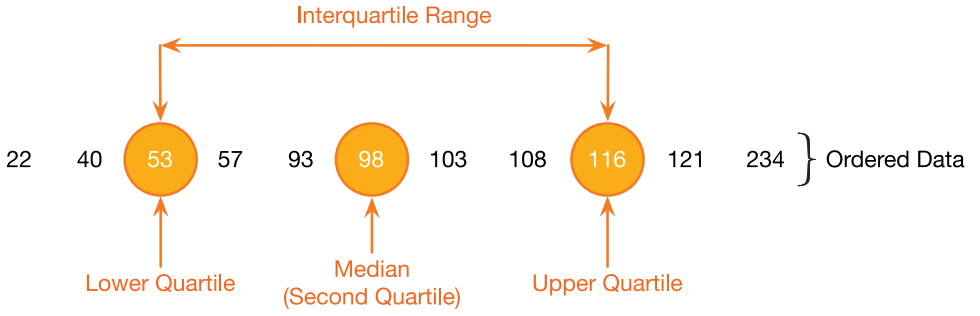


Figure 1.9 Calculating quartiles and the interquartile range



Twenty-one heavy smokers were put on a treadmill at the fastest setting. The time in seconds was measured until they fell off from exhaustion:

18, 16, 18, 24, 23, 22, 22, 23, 26, 29, 32, 34, 34, 36, 36, 43, 42, 49, 46, 46, 57

Compute the mode, median, mean, upper and lower quartiles, range and interquartile range

18 This description is huge oversimplification of how quartiles are calculated, but gives you the flavour of what they represent. If you want to know more, have a look at Hyndman and Fan (1996).

If we want to use all the data rather than half of it, we can calculate the spread of scores by looking at how different each score is from the centre of the distribution. If we use the mean as a measure of the centre of a distribution then we can calculate the difference between each score and the mean, which is known as the **deviance**:

$$\text{deviance} = x_i - \bar{x} \tag{1.4}$$

If we want to know the total deviance then we could add up the deviances for each data point. In equation form, this would be:

$$\text{total deviance} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right) \tag{1.5}$$

The sigma symbol (Σ) means 'add up all of what comes after', and the 'what comes after' in this case are the deviances. So, this equation means 'add up all of the deviances'.

Let's try this with the Instagram data. Table 1.2 shows the number of followers for each person in the Instagram data, the mean, and the difference between the two. Note that because the mean is at the centre of the distribution, some of the deviations are positive (scores greater than the mean) and some are negative (scores smaller than the mean). Consequently, when we add the scores up, the total is zero. Therefore, the 'total spread' is nothing. This conclusion is as silly as a tapeworm thinking they can have a coffee with the King of England if they don a bowler hat and pretend to be human. Everyone knows that the King drinks tea.

To overcome this problem, we could ignore the minus signs when we add the deviations up. There's nothing wrong with doing this, but people tend to square the deviations, which has a similar effect (because a negative number multiplied by another negative number becomes positive). The final column of Table 1.2 shows these squared deviances. We can add these squared deviances up to get the **sum of squared errors, SS** (often just called the *sum of squares*); unless your scores are all identical, the resulting value will be bigger than zero, indicating that there is some deviance from the mean. As an equation, we would write:

$$\text{sum of squared errors (SS)} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{1.6}$$

in which the sigma symbol means 'add up all of the things that follow' and what follows is the squared deviances (or squared errors as they're more commonly known).

We can use the sum of squares as an indicator of the total dispersion, or total deviance of scores from the mean. The problem with using the total is that its size will depend on how many scores we have in the data. The sum of squares for the Instagram data is 32,246, but if we added another 11 scores that value would increase (other things being equal, it will more or less double in size). The total dispersion is a bit of a nuisance then because we can't compare it across samples that differ in size. Therefore, it can be useful to work not with the *total* dispersion, but the *average* dispersion, which is known as the **variance**. We have seen that an average is the total of scores divided by the number of scores, therefore, the variance is the sum of squares divided by the number of observations (N). Actually, we normally divide the SS by the number of observations minus 1 (the reason why is explained in the next chapter and Jane Superbrain Box 2.5):

$$\text{variance}\left(s^2\right) = \frac{\text{SS}}{N-1} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{N-1} = \frac{32246}{10} = 3224.6 \tag{1.7}$$

Table 1.2 Table showing the deviations of each score from the mean

Number of Followers (x_i)	Mean ($\bar{x}$)	Deviance ($x_i - \bar{x}$)	Deviance squared $(x_i - \bar{x})^2$
22	95	−73	5329
40	95	−55	3025
53	95	−42	1764
57	95	−38	1444
93	95	−2	4
98	95	3	9
103	95	8	64
108	95	13	169
116	95	21	441
121	95	26	676
234	95	139	19321
		$\sum_{i=1}^{n} x_i - \bar{x} = 0$	$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 32246$

As we have seen, the variance is the average error between the mean and the observations made. There is one problem with the variance as a measure: it gives us a measure in units squared (because we squared each error in the calculation). In our example we would have to say that the average error in our data was 3224.6 followers squared. It makes very little sense to talk about followers squared, so we often take the square root of the variance to return the 'average error' back the original units of measurement. This measure is known as the **standard deviation** and is the square root of the variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N - 1}} = \sqrt{3224.6} = 56.79 \tag{1.8}$$

The sum of squares, variance and standard deviation are all measures of the dispersion or spread of data around the mean. A small standard deviation (relative to the value of the mean itself) indicates that the data points are close to the mean. A large standard deviation (relative to the mean) indicates that the data points are distant from the mean. A standard deviation of 0 would mean that all the scores were the same. Figure 1.10 shows the overall ratings (on a 5-point scale) of two lecturers after each of five different lectures. Both lecturers had an average rating of 2.6 out of 5 across the lectures. However, the standard deviation of ratings for the first lecturer is 0.55. This value is relatively small compared to the mean, suggesting that ratings for this lecturer were consistently close to the mean rating. There was a small fluctuation, but generally her lectures did not vary in popularity. This can be seen in the left plot where the ratings are not spread too widely around the mean. The second lecturer, however, had a standard deviation of 1.82. This value is relatively large compared to the mean. From the right plot we can see that this large value equates to ratings for this second lecturer that are more spread from the mean than the first: for some lectures she received very high ratings, and for others her ratings were appalling.

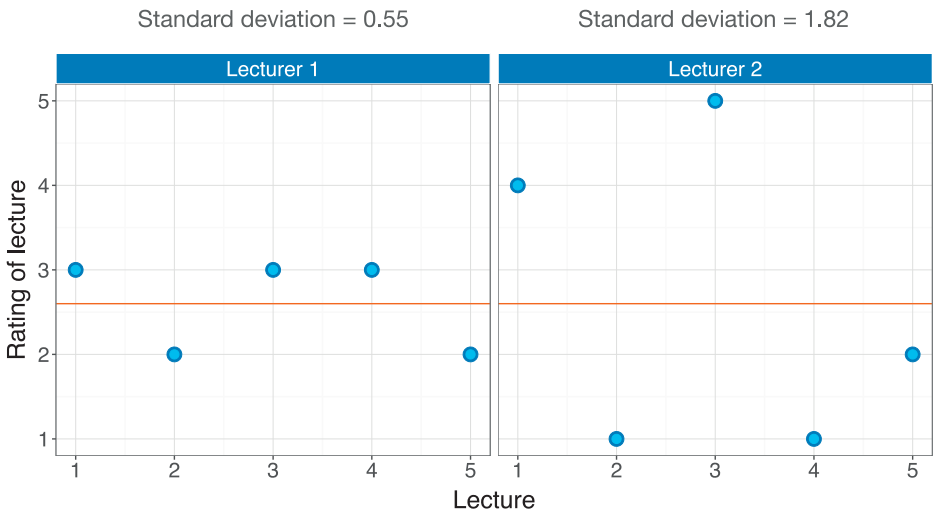


Figure 1.10 Graphs illustrating data that have the same mean but different standard deviations

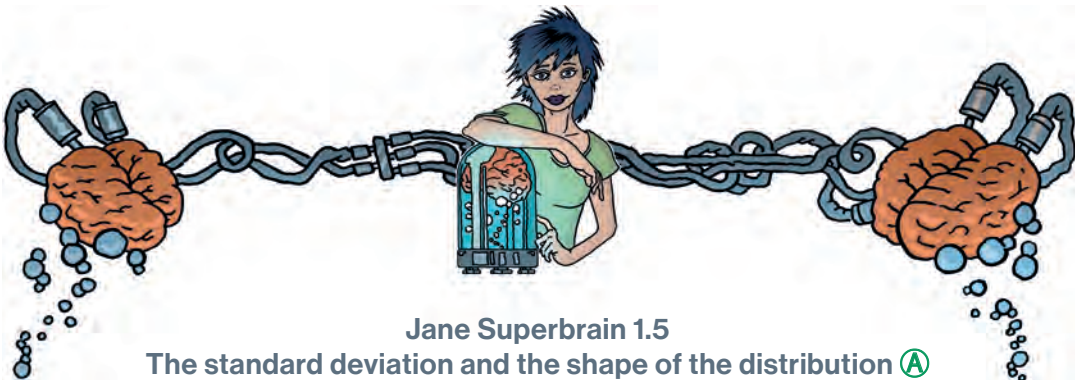
1.7.6 Using a frequency distribution to go beyond the data Ⓐ

Another way to think about frequency distributions is not in terms of how often scores actually occurred, but how likely it is that a score would occur (i.e., probability). The word 'probability' causes most people's brains to overheat (myself included) so it seems fitting that we use an example about throwing buckets of ice over our heads. Internet memes sometimes follow the shape of a normal distribution, which we discussed a while back. A good example of this is the ice bucket challenge from 2014. You can check Wikipedia for the full story, but it all started (arguably) with golfer Chris Kennedy tipping a bucket of iced water on his head to raise awareness of the disease amyotrophic lateral sclerosis (ALS, also known as Lou Gehrig's disease).[19] The idea is that you are challenged and have 24 hours to post a video of you having a bucket of iced water poured over your head and also challenge at least three other people. If you fail to complete the challenge your forfeit is to donate to charity (in this case ALS). In reality many people completed the challenge *and* made donations.

The ice bucket challenge is a good example of a meme: it ended up generating something like 2.4 million videos on Instagram and 2.3 million on YouTube. I mentioned that memes sometimes follow a normal distribution and Figure 1.12 shows this: the insert shows the 'interest' score from Google Trends for the phrase 'ice bucket challenge' from August to September 2014.[20] The 'interest' score that Google calculates is a bit hard to unpick but essentially reflects the relative number of times that the term 'ice bucket challenge' was searched for on Google. It's not the total number of searches, but the relative number. In a sense it shows the trend of the popularity of searching for 'ice bucket challenge'. Compare the line with the perfect normal distribution in Figure 1.3 – they look similar, don't they? Once it got going (about 2–3 weeks after the first video) it went viral and its popularity increased rapidly, reaching a peak at around 21 August (about 36 days after Chris Kennedy got the ball rolling). After this peak, its popularity rapidly declined as people tired of the meme.

19 Chris Kennedy did not invent the challenge, but he's believed to be the first to link it to ALS. There are earlier reports of people doing things with ice-cold water in the name of charity but I'm focusing on the ALS challenge because it is the one that spread as a meme.

20 You can generate the inset graph for yourself by going to Google Trends, entering the search term 'ice bucket challenge' and restricting the dates shown to August 2014 to September 2014.



Jane Superbrain 1.5
The standard deviation and the shape of the distribution Ⓐ

The variance and standard deviation tell us about the shape of the distribution of scores. If the mean is an accurate summary of the data then most scores will cluster close to the mean and the resulting standard deviation is small relative to the mean. When the mean is a less accurate summary of the data, the scores cluster more widely around the mean and the standard deviation is larger. Figure 1.11 shows two distributions that have the same mean (50) but different standard deviations. One has a large standard deviation relative to the mean (*SD* = 25) and this results in a flatter distribution that is more spread out, whereas the other has a small standard deviation relative to the mean (*SD* = 15), resulting in a pointier distribution in which scores close to the mean are very frequent but scores further from the mean become increasingly infrequent. The message is that as the standard deviation gets larger, the distribution gets fatter. This can make distributions look platykurtic or leptokurtic when, in fact, they are not.

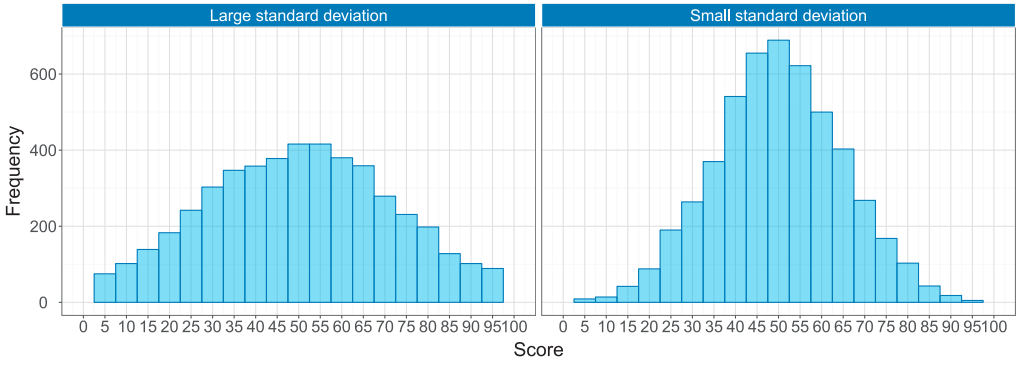


Figure 1.11 Two distributions with the same mean, but large and small standard deviations



The main histogram in Figure 1.12 shows the same pattern but reflects something a bit more tangible than 'interest scores'. It shows the number of videos posted on YouTube relating to the ice bucket challenge on each day after Chris Kennedy's initial challenge. There were 2323 thousand in total (2.32 million) during the period shown. In a sense it shows approximately how many people took up the



Labcoat Leni's Real Research 1.1
Is Friday the 13th unlucky? Ⓐ

Scanlon, T. J., et al. (1993). *British Medical Journal, 307*, 1584–1586.

Many of us are superstitious, and a common superstition is that Friday the 13th is unlucky. Most of us don't literally think that someone in a hockey mask is going to kill us, but some people are wary. Scanlon and colleagues, in a tongue-in-cheek study (Scanlon et al., 1993), looked at accident statistics at hospitals in the South West Thames region of the UK. They took statistics both for Friday the 13th and Friday the 6th (the week before) in different months in 1989, 1990, 1991 and 1992. They looked at both emergency admissions of accidents and poisoning, and also transport accidents.

Date	Accidents and Poisoning		Traffic Accidents	
	Friday 6th	Friday 13th	Friday 6th	Friday 13th
October 1989	4	7	9	13
July 1990	6	6	6	12
September 1991	1	5	11	14
December 1991	9	5	11	10
March 1992	9	7	3	4
November 1992	1	6	5	12

Calculate the mean, median, standard deviation and interquartile range for each type of accident and on each date. You can view the solutions at www.discoverspss.com/solutions/

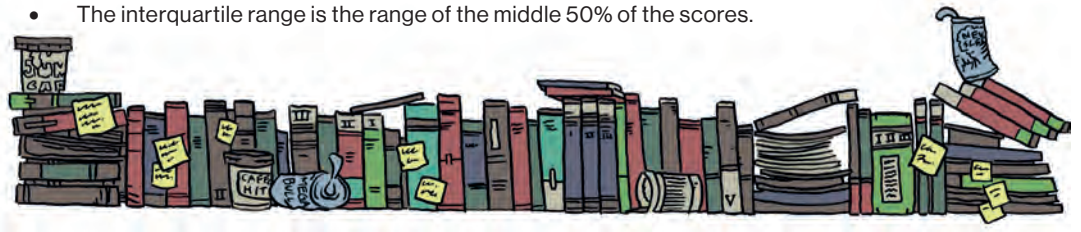
challenge each day.[21] You can see that nothing much happened for 20 days, and early on relatively few people took up the challenge. By about 30 days after the initial challenge things are hotting up (well, cooling down really) as the number of videos rapidly accelerated from 29,000 on day 30 to

21 Very approximately indeed. I have converted the Google interest data into videos posted on YouTube by using the fact that I know 2.33 million videos were posted during this period and making the (not unreasonable) assumption that behaviour on YouTube will have followed the same pattern over time as the Google interest score for the challenge.



Cramming Sam's Tips 1.4
Dispersion

- The deviance or error is the distance of each score from the mean.
- The sum of squared errors is the total amount of error in the mean. The errors/deviances are squared before adding them up.
- The variance is the average distance of scores from the mean. It is the sum of squares divided by the number of scores. It tells us about how widely dispersed scores are around the mean.
- The standard deviation is the *square root of the variance*. It is the variance converted back to the original units of measurement of the scores used to compute it. Large standard deviations relative to the mean suggest data are widely spread around the mean; whereas small standard deviations suggest data are closely packed around the mean.
- The range is the distance between the highest and lowest score.
- The interquartile range is the range of the middle 50% of the scores.



196,000 on day 35. At day 36 the challenge hits its peak (204,000 videos posted), after which the decline sets in as it becomes 'yesterday's news'. By day 50 it's only people like me, and statistics lecturers more generally, who don't check Instagram for 50 days, who suddenly become aware of the meme and want to get in on the action to prove how down with the kids we are. It's too late though: people at that end of the curve are uncool, and the trendsetters who posted videos on day 25 call us lame and look at us dismissively. It's OK though, because we can plot sick histograms like the one in Figure 1.12; take that, hipsters!

I digress. We can think of frequency distributions in terms of probability. To explain this, imagine that someone asked you, 'How likely is it that a person posted an ice bucket video after 60 days?' What would your answer be? Remember that the height of the bars on the histogram reflects how many videos were posted. Therefore, if you looked at the frequency distribution before answering the question you might respond 'not very likely' because the bars are very short after 60 days (i.e., relatively few videos were posted). What if someone asked you, 'How likely is it that a video was posted 35 days after the challenge started?' Using the histogram, you might say, 'It's relatively likely' because the bar is very high on day 35 (so quite a few videos were posted). Your inquisitive friend is on a roll and asks, 'How likely is it that someone posted a video 35 to 40 days after the challenge started?' The bars representing these days are shaded orange in Figure 1.12. The question about the likelihood of a video being posted 35–40 days into the challenge is really asking, 'How big is the orange area of Figure 1.12 compared to the total size of all bars?' We can find out the size of the dark blue region by adding the values of the bars (196 + 204 + 196 + 174 + 164 + 141 = 1075); therefore, the orange area represents 1075 thousand videos. The total size of all bars is the total number of videos

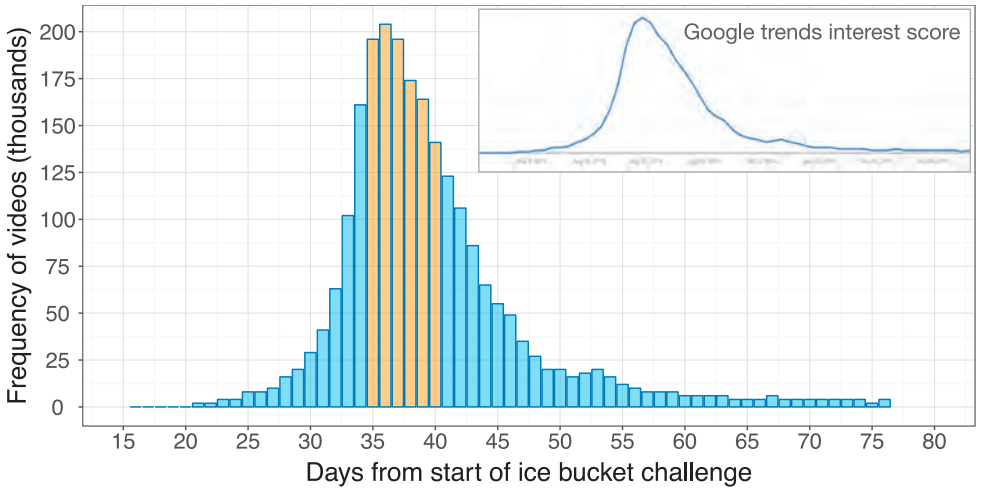


Figure 1.12 Frequency distribution showing the number of ice bucket challenge videos on YouTube by day since the first video (the inset shows the actual Google Trends data on which this example is based)

posted (i.e., 2323 thousand). If the orange area represents 1075 thousand videos, and the total area represents 2323 thousand videos, then if we compare the orange area to the total area we get 1075/2323 = 0.46. This proportion can be converted to a percentage by multiplying by 100, which gives us 46%. Therefore, our answer might be, 'It's quite likely that someone posted a video 35–40 days into the challenge because 46% of all videos were posted during those 6 days.' A very important point here is that the area of the bars relates directly to the probability of an event occurring.

Hopefully these illustrations show that we can use the frequencies of different scores, and the area of a frequency distribution, to estimate the probability that a particular score will occur. A probability value can range from 0 (there's no chance whatsoever of the event happening) to 1 (the event will definitely happen). So, for example, when I talk to my publishers I tell them there's a probability of 1 that I will have completed the revisions to this book by July. However, when I talk to anyone else, I might, more realistically, tell them that there's a 0.10 probability of me finishing the revisions on time (or put another way, a 10% chance, or 1 in 10 chance that I'll complete the book in time). In reality, the probability of my meeting the deadline is 0 (not a chance in hell). If probabilities don't make sense to you then you're not alone; just ignore the decimal point and think of them as percentages instead (i.e., a 0.10 probability that something will happen is a 10% chance that something will happen) or read the chapter on probability in my other excellent textbook (Field, 2016).

I've talked in vague terms about how frequency distributions can be used to get a rough idea of the probability of a score occurring. However, we can be precise. For any distribution of scores we could, in theory, calculate the probability of obtaining a score of a certain size – it would be incredibly tedious and complex to do it, but we could. To spare our sanity, statisticians have identified several common distributions. For each one they have worked out mathematical formulae (known as **probability density functions (PDFs)**) that specify idealized versions of these distributions. We could draw such a function by plotting the value of the variable (x) against the probability of it occurring (y).[22] The resulting curve is known as a **probability distribution**; for a normal distribution (Section 1.7.1) it would look like Figure 1.13, which has the characteristic bell shape that we saw already in Figure 1.3.

22 Actually we usually plot something called the density, which is closely related to the probability.

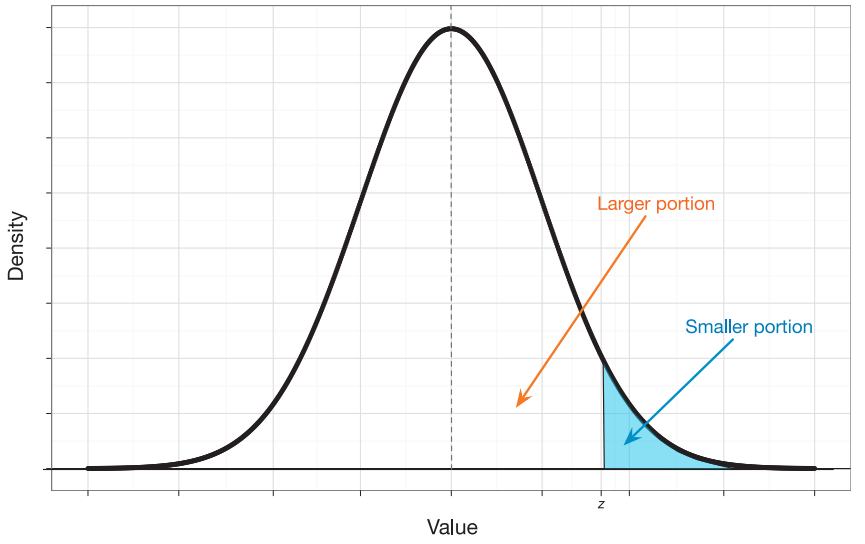


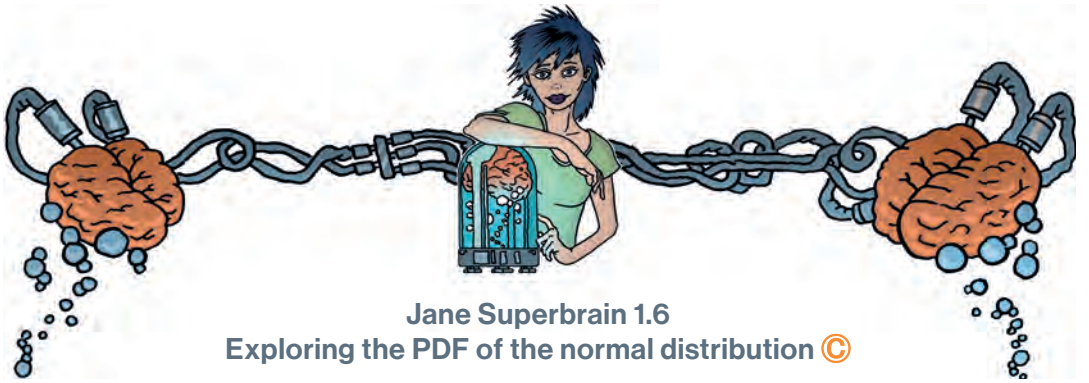
Figure 1.13 The normal probability distribution

A probability distribution is like a histogram except that the lumps and bumps have been smoothed out so that we see a nice smooth curve. However, like a frequency distribution, the area under this curve tells us something about the probability of a value occurring. Just like we did in our ice bucket example, we could use the area under the curve between two values to tell us how likely it is that a score fell within a particular range. For example, the blue shaded region in Figure 1.13 corresponds to the probability of a score being z or greater. The normal distribution is not the only distribution that has been precisely specified by people with enormous brains. There are many distributions that have characteristic shapes and have been specified with a PDF. We'll encounter some of these other distributions throughout the book, for example, the *t*-distribution, chi-square (χ^2) distribution and *F*-distribution. For now, the important thing to remember is that these distributions are all defined by equations that enable us to calculate the probability of obtaining a given score precisely (Jane Superbrain Box 1.6). For example, under the assumption that a normal distribution is a good approximation of the distribution of the scores we have, we can calculate the exact probability of a particular score occurring.

As we have seen, distributions can have different means and standard deviations. This isn't a problem for the PDF because the mean and standard deviation are featured in the equation (Jane Superbrain Box 1.6). However, it is a problem for us because PDFs are difficult enough to spell, let alone use to compute probabilities. Therefore, to avoid a brain meltdown we often use a standard normal distribution, which has a mean of 0 and a standard deviation of 1 (Jane Superbrain Box 1.6). Using the standard normal distribution has the advantage that we can pretend that the PDF doesn't exist and use tabulated probabilities (as in the Appendix) instead. The obvious problem is that not all of the data we collect will have a mean of 0 and standard deviation of 1. For example, for the ice bucket data the mean is 39.68 and the standard deviation is 7.74. However, any distribution of scores can be converted into a set of scores that have a mean of 0 and a standard deviation of 1. First, to centre the data around zero, we take each score (X) and subtract from it the mean of all scores ($\bar{X}$). To ensure the data have a standard deviation of 1, we divide the resulting score by the standard deviation (s), which we recently encountered. The resulting scores are denoted by the letter z and are known as **z-scores**. In equation form, the conversion that I've just described is:



$$z = \frac{X - \bar{X}}{s} \tag{1.11}$$



Jane Superbrain 1.6
Exploring the PDF of the normal distribution ©

The probability density function for the normal distribution is

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty. \tag{1.9}$$

It looks terrifying. The $-\infty < x < \infty$ part tells us that the function works on any scores between plus and minus infinity and isn't part of the equation. The left-hand side of the equation tells us the information that the function needs: it needs a score (x), the mean of the distribution (μ) and the standard deviation of the distribution (σ). The right-hand side uses this information to calculate the corresponding probability for the score entered into the function. Some of the right-hand side creates the general bell shape of the curve (the exponential function, e, does most of the heavy lifting), and some of it determines the position and width of the curve. Specifically, this part of the equation is the part most relevant to us:

$$\frac{x-\mu}{\sigma}.$$

The top of this fraction is the difference between each score and the mean of the distribution, and it centres the curve at the mean. It doesn't affect the shape of the curve, but it moves it horizontally: as the mean increases the curve shifts to the right, and as the mean decreases the curve shifts to the left. The bottom of the fraction scales the curve by the standard deviation and affects its width. As the standard deviation increases, so does the width of the bell shape, and conversely the bell shape gets narrower as the standard deviation decreases.

To see the equation in action, imagine we had scores distributed with a mean of 0 and standard deviation of 1, and we want to know the probability of $x = 2$. We can plug these numbers into the equation:

$$f(x=2;\mu=0,\sigma=1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{2-0}{1}\right)^2}=\frac{1}{\sqrt{2\pi}}e^{-2}=0.054$$

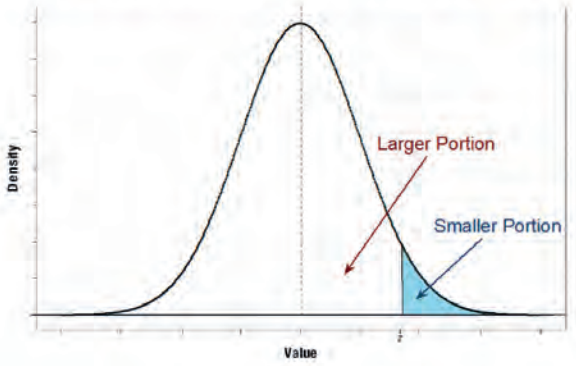
The probability of a score of 2 is $p = 0.054$. Note that we could have plugged in any values for the mean and standard deviation, but using a mean of 0 and standard deviation of 1 simplifies the equation to

$$f(z;\mu=0,\sigma=1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}, \qquad -\infty < z < \infty. \tag{1.10}$$

In this special case, it is common to refer to the score with the letter z rather than x and refer to it as a z-score.



A.1 Table of the standard normal distribution



z	Larger Portion	Smaller Portion	y
.00	.50000	.50000	.3989
.01	.50399	.49601	.3989
.02	.50798	.49202	.3989
.03	.51197	.48803	.3988
.04	.51595	.48405	.3986
.25	.59871	.40129	.3867
.26	.60257	.39743	.3857
.27	.60642	.39358	.3847
.28	.61026	.38974	.3836
.29	.61409	.38591	.3825
1.50	.93319	.06681	.1295
1.51	.93448	.06552	.1276
1.52	.93574	.06426	.1257
1.53	.93699	.06301	.1238
1.54	.93822	.06178	.1219
1.55	.93943	.06057	.1200
1.56	.94062	.05938	.1182
1.57	.94179	.05821	.1163
1.58	.94295	.05705	.1145
1.59	.94408	.05592	.1127
1.60	.94520	.05480	.1109
1.61	.94630	.05370	.1092
1.62	.94738	.05262	.1074
1.63	.94845	.05155	.1057
1.64	.94950	.05050	.1040
1.65	.95053	.04947	.1023
1.95	.97441	.02559	.0596
1.96	.97500	.02500	.0584
1.97	.97558	.02442	.0573
1.98	.97615	.02385	.0562
1.99	.97670	.02330	.0551
2.00	.97725	.02275	.0540
2.01	.97778	.02222	.0529
2.02	.97831	.02169	.0519
2.03	.97882	.02118	.0508
2.04	.97932	.02068	.0498
2.05	.97982	.02018	.0488
2.06	.98030	.01970	.0478
2.07	.98077	.01923	.0468
2.08	.98124	.01876	.0459
2.09	.98169	.01831	.0449
2.10	.98214	.01786	.0440
2.57	.99492	.00508	.0147
2.58	.99506	.00494	.0143
2.59	.99520	.00480	.0139
2.60	.99534	.00466	.0136
2.61	.99547	.00453	.0132
2.62	.99560	.00440	.0129
2.63	.99573	.00427	.0126
2.96	.99846	.00154	.0050
2.97	.99851	.00149	.0048
2.98	.99856	.00144	.0047
2.99	.99861	.00139	.0046
3.00	.99865	.00135	.0044
...	...	...	...
3.25	.99942	.00058	.0020

Figure 1.14 Using tabulated values of the standard normal distribution

This equation might look familiar because, apart from having used different symbols for the mean and standard deviation, it forms part of the PDF of the normal distribution (Jane Superbrain Box 1.6).

Having converted a set of scores to z-scores, we can refer to the table of probability values that have been calculated for the standard normal distribution, which you'll find in the Appendix. Why is this table important? Well, if we look at our ice bucket data, we can answer the question 'What's the probability that someone posted a video on day 60 or later?' First we convert 60 into a *z*-score. We saw that the mean was 39.68 and the standard deviation was 7.74, so our score of 60 expressed as a *z*-score is:

$$z = \frac{60 - 39.68}{7.74} = 2.63 \tag{1.12}$$

We can now use this value, rather than the original value of 60, to compute an answer to our question.

Figure 1.14 shows (an edited version of) the tabulated values of the standard normal distribution from the Appendix of this book. This table gives us a list of values of z, and the density (y) for each value of z, but, most important, it splits the distribution at the value of z and tells us the size of the two areas under the curve that this division creates. For example, when z is 0, we are at the mean or centre of the distribution so it splits the area under the curve exactly in half. Consequently, both areas have a size of 0.5 (or 50%). However, any value of z that is not zero will create different sized areas, and the table tells us the size of the larger and smaller portion. For example, if we look up our *z*-score of 2.63, we find that the smaller portion (i.e., the area above this value, or the blue area in Figure 1.14) is 0.0043, or only 0.43%. I explained before that these areas relate to probabilities, so in this case we could say that there is only a 0.43% chance that a video was posted 60 days or more after the challenge started. By looking at the larger portion (the area below 2.63) we get 0.9957, or put another way, there's a 99.57% chance that an ice bucket video was posted on YouTube within 60 days of the challenge starting. Note that these two proportions add up to 1 (or 100%), so the total area under the curve is 1.

Another useful thing we can do (you'll find out just how useful in due course) is to work out limits within which a certain percentage of scores fall. With our ice bucket example, we looked at how likely it was that a video was posted between 35 and 40 days after the challenge started; we could ask a similar question such as, 'What is the range of days between which the middle 95% of videos were posted?'

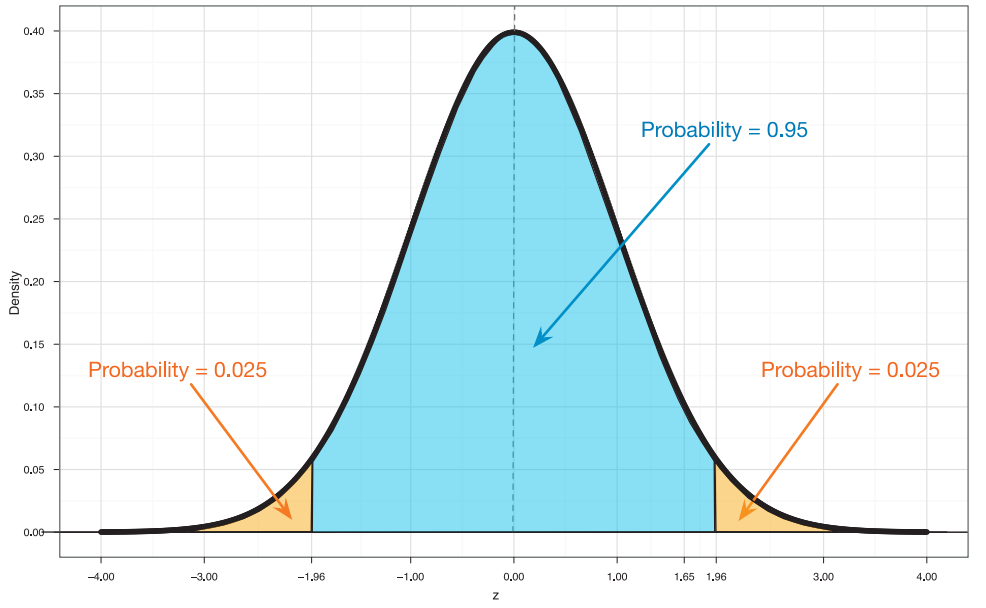
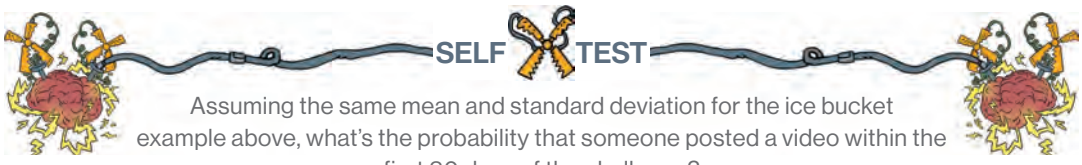


Figure 1.15 The probability density function of a normal distribution

To answer this question we need to use the table the opposite way around. We know that the total area under the curve is 1 (or 100%), so to discover the limits within which 95% of scores fall we're asking 'what is the value of *z* that cuts off 5% of the scores?' It's not quite as simple as that because if we want the *middle* 95%, then we want to cut off scores from both ends. Given the distribution is symmetrical, if we want to cut off 5% of scores overall but we want to take some from both extremes of scores, then the percentage of scores we want to cut from each end will be 5%/2 = 2.5% (or 0.025 as a proportion). If we cut off 2.5% of scores from each end then in total we'll have cut off 5% scores, leaving us with the middle 95% (or 0.95 as a proportion) – see Figure 1.15. To find out what value of *z* cuts off the top area of 0.025, we look down the column 'smaller portion' until we reach 0.025, we then read off the corresponding value of *z*. This value is 1.96 (see Figure 1.14) and because the distribution is symmetrical around zero, the value that cuts off the bottom 0.025 will be the same but a minus value (–1.96). Therefore, the middle 95% of *z*-scores fall between –1.96 and 1.96. If we wanted to know the limits between which the middle 99% of scores would fall, we could do the same: now we would want to cut off 1% of scores, or 0.5% from each end. This equates to a proportion of 0.005. We look up 0.005 in the *smaller portion* part of the table and the nearest value we find is 0.00494, which equates to a *z*-score of 2.58 (see Figure 1.14). This tells us that 99% of *z*-scores lie between –2.58 and 2.58. Similarly (have a go), you can show that 99.9% of them lie between –3.29 and 3.29. Remember these values (1.96, 2.58 and 3.29) because they'll crop up time and time again.



Assuming the same mean and standard deviation for the ice bucket example above, what's the probability that someone posted a video within the first 30 days of the challenge?

1.7.7 Fitting statistical models to the data Ⓐ

Having looked at your data (and there is a lot more information on different ways to do this in Chapter 1), the next step of the research process is to fit a statistical model to the data. That is to go where eagles dare, and no one should fly where eagles dare; but to become scientists we have to, so the rest of this book attempts to guide you through the various models that you can fit to the data.

1.8 Reporting data Ⓐ

1.8.1 Dissemination of research Ⓐ

Having established a theory and collected and started to summarize data you might want to tell other people what you have found. This sharing of information is a fundamental part of being a scientist. As discoverers of knowledge, we have a duty of care to the world to present what we find in a clear and unambiguous way, and with enough information that others can challenge our conclusions. It is good practice, for example, to make your data available to others and to be open with the resources you used. Initiatives such as the Open Science Framework (https://osf.io) make this easy to do. Tempting as it may be to cover up the more unsavoury aspects of our results, science is about truth, openness and willingness to debate your work.

Scientists tell the world about our findings by presenting them at conferences and in articles published in a scientific **journals**. A scientific journal is a collection of articles written by scientists on a vaguely similar topic. A bit like a magazine but more tedious. These articles can describe



Cramming Sam's Tips 1.5
Distributions and *z*-scores

- A frequency distribution can be either a table or a chart that shows each possible score on a scale of measurement along with the number of times that score occurred in the data.
- Scores are sometimes expressed in a standard form known as *z*-scores.
- To transform a score into a *z*-score you subtract from it the mean of all scores and divide the result by the standard deviation of all scores.
- The sign of the *z*-score tells us whether the original score was above or below the mean; the value of the *z*-score tells us how far the score was from the mean in standard deviation units.



new research, review existing research, or might put forward a new theory. Just like you have magazines such as *Modern Drummer*, which is about drumming, or *Vogue*, which is about fashion (or Madonna, I can never remember which), you get journals such as *Journal of Anxiety Disorders*, which publishes articles about anxiety disorders, and *British Medical Journal*, which publishes articles about medicine (not specifically British medicine, I hasten to add). As a scientist, you submit your work to one of these journals and they will consider publishing it. Not everything a scientist writes will be published. Typically, your manuscript will be given to an 'editor' who will be a fairly eminent scientist working in that research area who has agreed, in return for their soul, to make decisions about whether or not to publish articles. This editor will send your manuscript out to review, which means they send it to other experts in your research area and ask those experts to assess the quality of the work. Often (but not always) the reviewer is blind to who wrote the manuscript. The reviewers' role is to provide a constructive and even-handed overview of the strengths and weaknesses of your article and the research contained within it. Once these reviews are complete the editor reads them, and assimilates the comments with his or her own views on the manuscript and decides whether to publish it (in reality, you'll be asked to make revisions at least once before a final acceptance).

The review process is an excellent way to get useful feedback on what you have done, and very often throws up things that you hadn't considered. The flip side is that when people scrutinize your work they don't always say nice things. Early on in my career I found this process quite difficult: often you have put months of work into the article and it's only natural that you want your peers to receive it well. When you do get negative feedback, and even the most respected scientists do, it can be easy to feel like you're not good enough. At those times, it's worth remembering that if you're not affected by criticism then you're probably not human; every scientist I know has moments of self-doubt.

1.8.2 Knowing how to report data Ⓐ

An important part of publishing your research is how you present and report your data. You will typically do this through a combination of plots (see Chapter 1) and written descriptions of the data. Throughout this book I will give you guidance about how to present data and write up results. The difficulty is that different disciplines have different conventions. In my area of science (psychology) we typically follow the publication guidelines of the American Psychological Association or APA (American Psychological Association, 2019), but even within psychology different journals have their own idiosyncratic rules about how to report data. Therefore, my advice will be broadly based on the APA guidelines with a bit of my own personal opinion thrown in when there isn't a specific APA 'rule'. However, when reporting data for assignments or for publication it is always advisable to check the specific guidelines of your tutor or the journal.

Despite the fact that some people would have you believe that if you deviate from any of the 'rules' in even the most subtle of ways then you will unleash the four horsemen of the apocalypse onto the world to obliterate humankind, the 'rules' are no substitute for common sense. Although some people treat the APA style guide like a holy sacrament, its job is not to lay down intractable laws, but to offer a guide so that everyone is consistent in what they do. It does not tell you what to do in every situation but does offer (mostly) sensible guiding principles that you can extrapolate to the situations you encounter.

1.8.3 Some initial guiding principles Ⓐ

When reporting data your first decision is whether to use text, a plot/graph or a table. You want to be succinct so you shouldn't present the same values in multiple ways; if you have a plot showing some results then don't also produce a table of the same results: it's a waste of space. The APA gives the following guidelines:

- ✓ Choose a mode of presentation that optimizes the understanding of the data.
- ✓ If you present three or fewer numbers then try using a sentence.
- ✓ If you need to present between 4 and 20 numbers then consider a table.
- ✓ If you need to present more than 20 numbers then a graph is often more useful than a table.

Of these, I think the first is most important: I can think of countless situations where I would want to use a plot rather than a table to present 4–20 values because it will show up the pattern of data most clearly. Similarly, I can imagine some plots presenting more than 20 numbers being an absolute mess. This takes me back to my point about rules being no substitute for common sense, and the most important thing is to present the data in a way that makes it easy for the reader to digest. We'll look at how to plot data in Chapter 5 and we'll look at tabulating data in various chapters when we discuss how best to report the results of particular analyses.

A second general issue is how many decimal places to use when reporting numbers. The guiding principle from the APA (which I think is sensible) is that the fewer decimal places the better, which means that you should round as much as possible but bear in mind the precision of the measure you're reporting. This principle again reflects making it easy for the reader to understand the data. Let's look at an example. Sometimes when a person doesn't respond to someone, they will ask 'what's wrong, has the cat got your tongue?' Actually, my cat had a large collection of carefully preserved human tongues that he kept in a box under the stairs. Periodically he'd get one out, pop it in his mouth and wander around the neighbourhood scaring people with his big tongue. If I measured the difference in length between his actual tongue and his fake human tongue, I might report this difference as 0.0425 metres, 4.25 centimetres, or 42.5 millimetres. This example illustrates three points: (1) I needed a different number of decimal places (4, 2 and 1,

respectively) to convey the same information in each case; (2) 4.25 cm is probably easier for someone to digest than 0.0425 m because it uses fewer decimal places; and (3) my cat was odd. The first point demonstrates that it's not the case that you should always use, say, two decimal places; you should use however many you need in a particular situation. The second point implies that if you have a very small measure it's worth considering whether you can use a different scale to make the numbers more palatable.

Finally, every set of guidelines will include advice on how to report specific analyses and statistics. For example, when describing data with a measure of central tendency, the APA suggests you use *M* (capital M in italics) to represent the mean but is fine with you using mathematical notation (e.g., $\bar{X}$, $\hat{\mu}$). However, you should be consistent: if you use *M* to represent the mean you should do so throughout your report. There is also a sensible principle that if you report a summary of the data such as the mean, you should also report the appropriate measure of the spread of scores. Then people know not just the central location of the data, but also how spread out they were. Therefore, whenever we report the mean, we typically report the standard deviation also. The standard deviation is usually denoted by *SD*, but it is also common to simply place it in parentheses as long as you indicate that you're doing so in the text. Here are some examples from this chapter:

- ✓ Andy has 2 followers on Instagram. On average, a sample of other users (N = 11), had considerably more, M = 95, SD = 56.79.
- ✓ The average number of days it took someone to post a video of the ice bucket challenge was $\bar{X}$ = 39.68, SD = 7.74.
- ✓ By reading this chapter we discovered that (*SD* in parentheses), on average, people have 95 (56.79) followers on Instagram and on average it took people 39.68 (7.74) days to post a video of them throwing a bucket of iced water over themselves.

Note that in the first example, I used *N* to denote the size of the sample. This is a common abbreviation: a capital *N* represents the entire sample and a lower case *n* represents a subsample (e.g., the number of cases within a particular group).

Similarly, when we report medians, there is a specific notation (the APA suggest *Mdn*) and we should report the range or interquartile range as well (the APA doesn't have an abbreviation for either of these terms, but IQR is commonly used for the interquartile range). Therefore, we could report:

- ✓ Andy has 2 followers on Instagram. A sample of other users (N = 11) typically had more, Mdn = 98, IQR = 63.
- ✓ Andy has 2 followers on Instagram. A sample of other users (N = 11) typically had more, Mdn = 98, *range* = 212.

1.9 Jane and Brian's story Ⓐ

Brian had a crush on Jane. He'd seen her around campus a lot, always rushing with a big bag and looking sheepish. She was an enigma, no one had ever spoken to her or knew why she scuttled around the campus with such purpose. Brian found her quirkiness sexy. As she passed him on the library stairs, Brian caught her shoulder. She looked horrified.

'Sup,' he said with a smile.

Jane looked sheepishly at the bag she was carrying.

'Fancy a brew?' Brian asked.

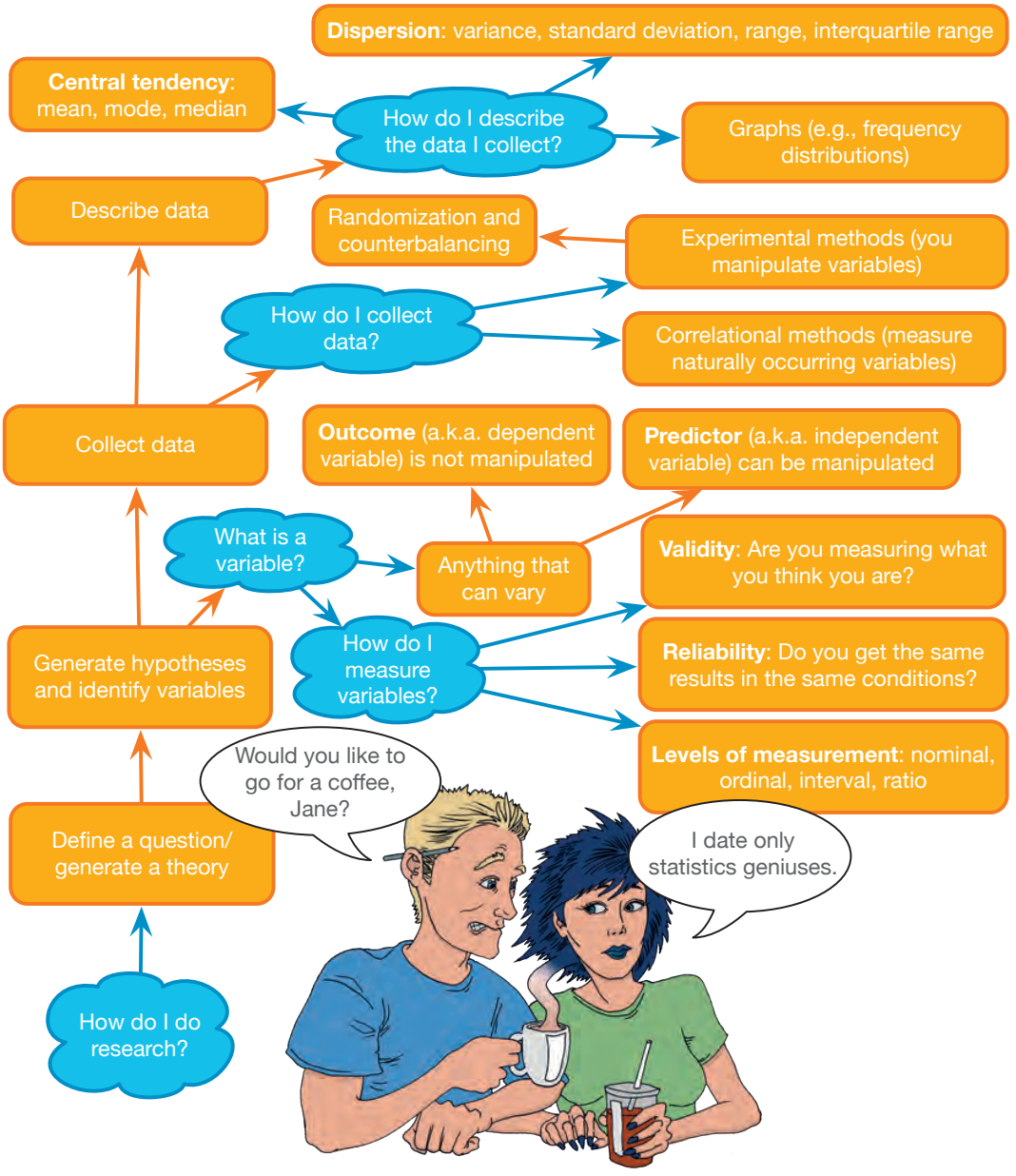


Figure 1.16 What Brian learnt from this chapter

Jane looked Brian up and down. He looked like he might be clueless ... and Jane didn't easily trust people. To her surprise, Brian tried to impress her with what he'd learnt in his statistics lecture that morning. Maybe she was wrong about him, maybe he was a statistics guy ... that would make him more appealing – after all, stats guys always told the best jokes.

Jane took his hand and led him to the Statistics section of the library. She pulled out a book called *An Adventure in Statistics* and handed it to him. Brian liked the cover. Jane turned and strolled away enigmatically.

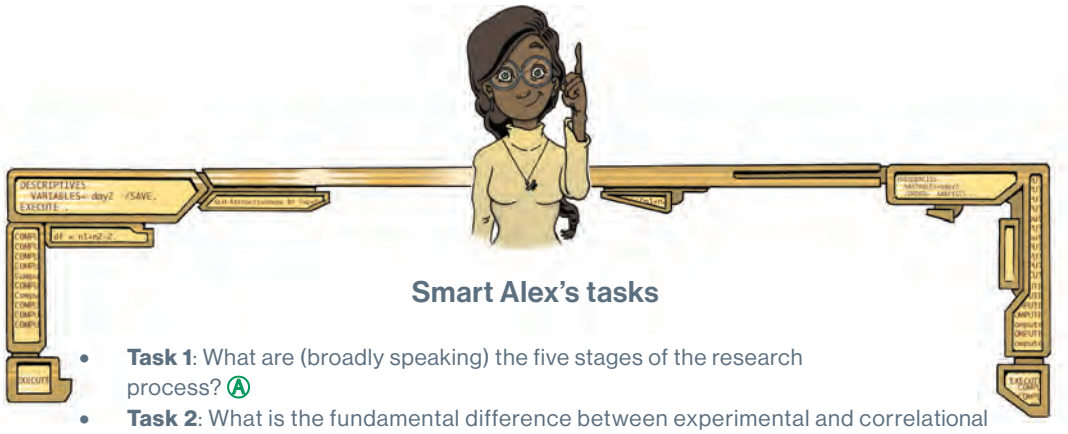
1.10 What next? Ⓐ

It is all very well discovering that if you stick your finger into a fan or get hit around the face with a golf club it hurts, but what if these are isolated incidents? It's better if we can somehow extrapolate from our data and draw more general conclusions. Even better, perhaps we can start to make predictions about the world: if we can predict when a golf club is going to appear out of nowhere then we can better move our faces. The next chapter looks at fitting models to the data and using these models to draw conclusions that go beyond the data we collected.

My early childhood wasn't all full of pain, on the contrary it was filled with a lot of fun: the nightly 'from how far away can I jump into bed' competition (which sometimes involved a bit of pain) and being carried by my brother and dad to bed as they hummed Chopin's *Marche Funèbre* before lowering me between two beds as though being buried in a grave. It was more fun than it sounds.

1.11 Key terms that I've discovered

Between-groups design
Between-subjects design
Bimodal
Binary variable
Boredom effect
Categorical variable
Central tendency
Concurrent validity
Confounding variable
Content validity
Continuous variable
Correlational research
Counterbalancing
Criterion validity
Cross-sectional research
Dependent variable
Deviance
Discrete variable
Ecological validity
Experimental research
Falsification
Frequency distribution
Histogram
Hypothesis
Independent design
Independent variable
Interquartile range
Interval variable
Journal
Kurtosis
Leptokurtic
Level of measurement
Longitudinal research
Lower quartile
Mean
Measurement error
Median
Mode
Multimodal
Negative skew
Nominal variable
Nonile
Normal distribution
Ordinal variable
Outcome variable
Percentile
Platykurtic
Positive skew
Practice effect
Predictive validity
Predictor variable
Probability density function (PDF)
Probability distribution
Qualitative methods
Quantile
Quantitative methods
Quartile
Randomization
Range
Ratio variable
Reliability
Repeated-measures design
Second quartile
Skew
Standard deviation
Systematic variation
Sum of squared errors
Tertium quid
Test–retest reliability
Theory
Unsystematic variance
Upper quartile
Validity
Variables
Variance
Within-subject design
z-scores



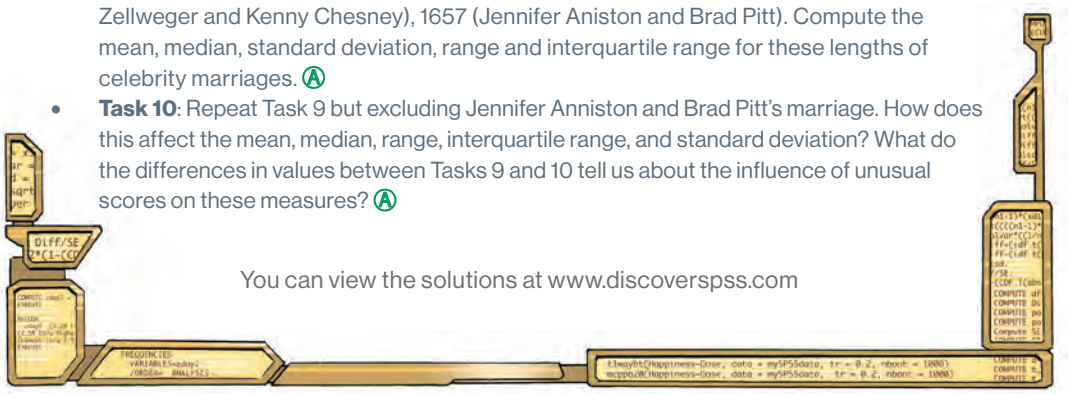
Smart Alex's tasks

- **Task 1**: What are (broadly speaking) the five stages of the research process? Ⓐ
- **Task 2**: What is the fundamental difference between experimental and correlational research? Ⓐ
- **Task 3:** What is the level of measurement of the following variables? Ⓐ
  - The number of downloads of different bands' songs on iTunes
  - The names of the bands that were downloaded
  - Their positions in the download chart
  - The money earned by the bands from the downloads
  - The weight of drugs bought by the bands with their royalties
  - The type of drugs bought by the bands with their royalties
  - The phone numbers that the bands obtained because of their fame
  - The gender of the people giving the bands their phone numbers
  - The instruments played by the band members
  - The time they had spent learning to play their instruments
- **Task 4**: Say I own 857 CDs. My friend has written a computer program that uses a webcam to scan the shelves in my house where I keep my CDs and measure how many I have. His program says that I have 863 CDs. Define measurement error. What is the measurement error in my friend's CD-counting device? Ⓐ
- **Task 5**: Sketch the shape of a normal distribution, a positively skewed distribution and a negatively skewed distribution. Ⓐ
- **Task 6**: In 2011 I got married and we went to Disney World in Florida for our honeymoon. We bought some bride and groom Mickey Mouse hats and wore them around the parks. The staff at Disney are really nice and upon seeing our hats would say 'congratulations' to us. We counted how many times people said congratulations over 7 days of the honeymoon: 5, 13, 7, 14, 11, 9, 17. Calculate the mean, median, sum of squares, variance and standard deviation of these data. Ⓐ
- **Task 7**: In this chapter we used an example of the time taken for 21 heavy smokers to fall off a treadmill at the fastest setting (18, 16, 18, 24, 23, 22, 22, 23, 26, 29, 32, 34, 34, 36, 36, 43, 42, 49, 46, 46, 57). Calculate the sum of squares, variance and standard deviation of these data. Ⓐ
- **Task 8**: Sports scientists sometimes talk of a 'red zone', which is a period during which players in a team are more likely to pick up injuries because they are fatigued. When a player hits the red zone it is a good idea to rest them for a game or two. At a prominent London football club that I support, they measured how many consecutive games the 11 first team players could manage before hitting the red zone: 10, 16, 8, 9, 6, 8, 9, 11, 12, 19, 5. Calculate the mean, standard deviation, median, range and interquartile range. Ⓐ
- **Task 9**: Celebrities always seem to be getting divorced. The (approximate) lengths of some celebrity marriages in days are: 240 (J-Lo and Cris Judd), 144 (Charlie Sheen and Donna Peele), 143 (Pamela Anderson and Kid Rock), 72 (Kim Kardashian, if you can call her a celebrity), 30 (Drew Barrymore and Jeremy Thomas), 26 (W. Axl Rose and Erin Everly), 2 (Britney Spears and Jason Alexander), 150 (Drew Barrymore again, but this time with Tom Green), 14 (Eddie Murphy and Tracy Edmonds), 150 (Renée

Zellweger and Kenny Chesney), 1657 (Jennifer Aniston and Brad Pitt). Compute the mean, median, standard deviation, range and interquartile range for these lengths of celebrity marriages. Ⓐ

- **Task 10**: Repeat Task 9 but excluding Jennifer Anniston and Brad Pitt's marriage. How does this affect the mean, median, range, interquartile range, and standard deviation? What do the differences in values between Tasks 9 and 10 tell us about the influence of unusual scores on these measures? Ⓐ

You can view the solutions at www.discoverspss.com





THE SPINE OF STATISTICS

2.1 What will this chapter tell me? 50
2.2 What is the SPINE of statistics? 51
2.3 Statistical models 51
2.4 Populations and samples 55
2.5 The linear model 56
2.6 P is for parameters 58
2.7 E is for estimating parameters 66
2.8 S is for standard error 71
2.9 I is for (confidence) interval 74
2.10 N is for null hypothesis significance testing 81
2.11 Reporting significance tests 100
2.12 Jane and Brian's story 101
2.13 What next? 103
2.14 Key terms that I've discovered 103
Smart Alex's Tasks 103

Letters A to D indicate increasing difficulty

2.1 What will this chapter tell me? Ⓐ

Although I had learnt a lot about golf clubs randomly appearing out of nowhere and hitting me around the face, I still felt that there was much about the world that I didn't understand. For one thing, could I learn to predict the presence of these golf clubs that seemed inexplicably drawn towards my head? A child's survival depends upon being able to predict reliably what will happen in certain situations; consequently they develop a model of the world based on the data they have (previous experience) and they then test this model by collecting new data/experiences.

At about 5 years old I moved from nursery to primary school. My nursery school friends were going to different schools and I was terrified about meeting new children. I had a model of the world that children who I didn't know wouldn't like me. A hypothesis from this model was 'if I talk to a new child they will reject me'. I arrived in my new classroom, and as I'd feared, it was full of scary children waiting to reject me. In a fairly transparent ploy to make me think that I'd be spending the next 6 years building sandcastles, the teacher told me to play in the sandpit. While I was nervously trying to discover whether I could build a pile of sand high enough to bury my head in it, a boy came to join me. His name was Jonathan Land. Much as I didn't want to, like all scientists I needed to collect new data and see how well my model fit them. So, with some apprehension, I spoke to him. Within an hour he was my new best friend (5-year-olds are fickle …) and I loved school. We remained close friends all through primary school.

It turned out that my model of 'new children won't like me' was a poor fit of these new data, and inconsistent with the hypothesis that if I talk to a child they will reject me. If I were a good scientist I'd have updated my model and moved on, but instead I clung doggedly to my theory. Some (bad) scientists do that too. So that you don't end up like them I intend to be a 5-year-old called Jonathan guiding you through the sandpit of statistical models. Thinking like a 5-year-old comes quite naturally to me so it should be fine. We will edge sneakily away from the frying pan of research methods and trip accidentally into the fires of statistics hell. We will start to see how we can use the properties of data to go beyond our observations and to draw inferences about the world at large. This chapter and the next lay the foundation for the rest of the book.



Figure 2.1 The face of innocence

2.2 What is the SPINE of statistics? Ⓐ

To many students, statistics is bewildering mass of different statistical tests, each with their own set of equations. The focus is often on 'difference'. It feels like you need to learn a lot of different stuff. What I hope to do in this chapter is to focus your mind on some core concepts that many statistical models have in common. In doing so, I want to set the tone for you to focus on the *similarities* between statistical models rather than the differences. If your goal is to use statistics as a tool, rather than to bury yourself in the theory, then I think this approach makes your job a lot easier. In this chapter, I will first argue that most statistical models are variations on the very simple idea of predicting an outcome variable from one or more predictor variables. The mathematical form of the model changes, but it usually boils down to a representation of the relations between an outcome and one or more predictors. If you understand that then there are five key concepts to get your head around. If you understand these, you've gone a long way towards understanding any statistical model that you might want to fit. They are the SPINE of statistics, which is a clever acronym for:

- Standard error
- Parameters
- Interval estimates (confidence intervals)
- Null hypothesis significance testing (NHST)
- Estimation

I cover each of these topics, but not in this order because PESIN doesn't work nearly so well as an acronym.[1]

2.3 Statistical models Ⓐ

We saw in the previous chapter that scientists are interested in discovering something about a phenomenon that we assume exists (a 'real-world' phenomenon). These real-world phenomena can be anything from the behaviour of interest rates in the economy to the behaviour of undergraduates at the end-of-exam party. Whatever the phenomenon, we collect data from the real world to test predictions from our hypotheses about that phenomenon. Testing these hypotheses involves building statistical models of the phenomenon of interest.



Let's begin with an analogy. Imagine an engineer wishes to build a bridge across a river. That engineer would be pretty daft if she built any old bridge, because it might fall down. Instead, she uses theory: what materials and structural features are known to make bridges strong? She might look at data to support these theories: have existing bridges that use these materials and structural features survived? She uses this information to construct an idea of what her new bridge will be (this is a 'model'). It's expensive and impractical for her to build a full-sized version of her bridge, so she builds a scaled-down version. The model may differ from reality in several ways – it will be smaller, for a start – but the engineer will try to build a model that best fits the situation of interest based on the data available. Once the model has been built, it can be used to predict things about the real world: for example, the engineer might test whether the bridge can withstand strong winds by placing her model in a wind tunnel. It is important that the model accurately represents the real world, otherwise any conclusions she extrapolates to the real-world bridge will be meaningless.

1 There is another, more entertaining, acronym, but I decided not to use it because in a séance with Freud he advised me that it could lead to pesin envy.

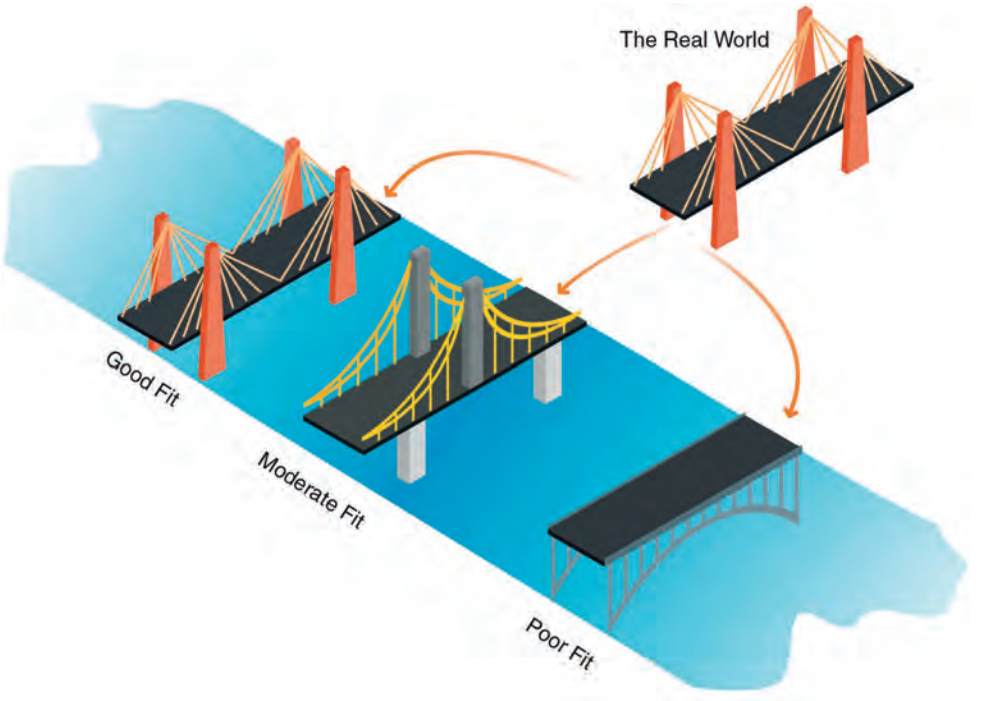
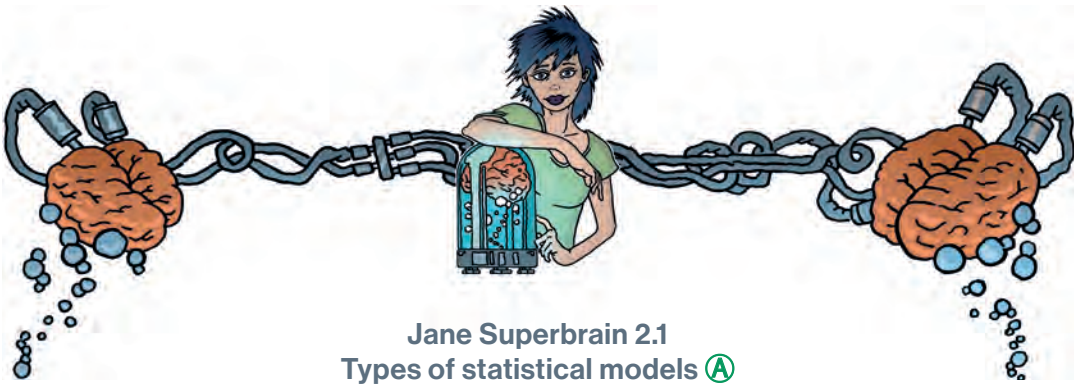


Figure 2.2 Fitting models to real-world data (see text for details)

Scientists do much the same: they build (statistical) models of real-world processes to predict how these processes operate under certain conditions (see Jane Superbrain Box 2.1). Unlike engineers, we don't have access to the real-world situation and so we can only *infer* things about psychological, societal, biological or economic processes based upon the models we build. However, like the engineer, our models need to be as accurate as possible so that the predictions we make about the real world are accurate too; the statistical model should represent the data collected (the *observed data*) as closely as possible. The degree to which a statistical model represents the data collected is known as the **fit** of the model.

Figure 2.2 shows three models that our engineer has built to represent her real-world bridge. The first is an excellent representation of the real-world situation and is said to be a *good fit*. If the engineer uses this model to make predictions about the real world then, because it so closely resembles reality, she can be confident that her predictions will be accurate. If the model collapses in a strong wind, then there is a good chance that the real bridge would collapse also. The second model has some similarities to the real world: the model includes some of the basic structural features, but there are some big differences too (e.g., the absence of one of the supporting towers). We might consider this model to have a *moderate fit* (i.e., there are some similarities to reality but also some important differences). If our engineer uses this model to make predictions about the real world then her predictions could be inaccurate or even catastrophic. For example, perhaps the model predicts that the bridge will collapse in a strong wind, so after the real bridge is built it gets closed every time a strong wind occurs, creating 100-mile tailbacks with everyone stranded in the snow, feasting on the crumbs of old sandwiches that they find under the seats of their cars. All of which turns out to be unnecessary because the real bridge was safe – the prediction from the model was wrong because it was a bad representation of reality. We can have a little confidence in predictions from this model, but not complete confidence. The final model is completely different than the real-world situation; it bears no structural similarities to the real bridge and is a *poor fit*. Any predictions based on this model are likely to be completely inaccurate. Extending this analogy to science, if our model is a poor fit to the observed data then the predictions we make from it will be poor too.



Jane Superbrain 2.1
Types of statistical models Ⓐ

Scientists (especially behavioural and social ones) tend to use **linear models**, which in their simplest form are models that summarize a process using a straight line. As you read scientific research papers you'll see that they are riddled with 'analysis of variance (ANOVA)' and 'regression', which are identical statistical systems based on the linear model (Cohen, 1968). In fact, most of the chapters in this book explain this 'general linear model'.

Imagine we were interested in how people evaluated dishonest acts.[2] Participants evaluate the dishonesty of acts based on watching videos of people confessing to those acts. Imagine we took 100 people and showed them a random dishonest act described by the perpetrator. They then evaluated the honesty of the act (from 0 = appalling behaviour to 10 = it's OK really) and how much they liked the person (0 = not at all, 10 = a lot).

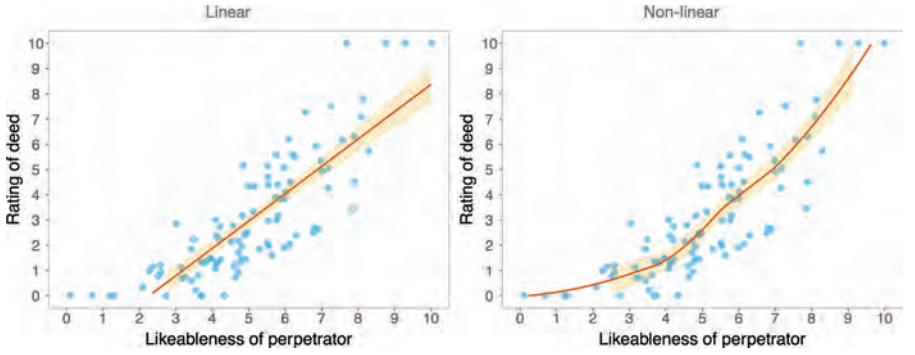


Figure 2.3 A scatterplot of the same data with a linear model fitted (left), and with a non-linear model fitted (right)

We can represent these hypothetical data on a scatterplot in which each dot represents an individual's rating on both variables (see Section 5.8). Figure 2.3 shows two versions of the same data. We can fit different models to the same data: on the left we have a linear (straight) model and on the right a non-linear (curved) one. Both models show that the more likeable a perpetrator is, the more positively people view their dishonest act. However, the curved line shows a subtler pattern: the trend to be more forgiving of likeable people kicks in when the likeableness rating rises above 4. Below 4 (where the perpetrator is not very likeable) all deeds are rated fairly low (the red line is quite flat), but above 4 (as the perpetrator becomes likeable) the slope of the line becomes steeper, suggesting that as likeableness rises above this value, people become increasingly forgiving of dishonest acts. Neither of the two models is necessarily correct, but one model will fit the data better than the other; this is why it is important for us to assess how well a statistical model fits the data.

2 This example came from a media story about The Honesty Lab set up by Stefan Fafinski and Emily Finch. However, this project no longer seems to exist and I can't find the results published anywhere. I like the example though, so I kept it.

Linear models tend to get fitted to data over non-linear models because they are less complex, because non-linear models are often not taught (despite 900 pages of statistics hell, I don't get into non-linear models in this book) and because scientists sometimes stick to what they know. This could have had interesting consequences for science: (1) many published statistical models might not be the ones that fit best (because the authors didn't try out non-linear models); and (2) findings might have been missed because a linear model was a poor fit and the scientists gave up rather than fitting non-linear models (which perhaps would have been a 'good enough' fit). It is useful to plot your data first: if your plot seems to suggest a non-linear model then don't apply a linear model because that's all you know; fit a nonlinear model (after complaining to me about how I don't cover them in this book).



Although it's easy to visualize a model of a bridge, you might be struggling to understand what I mean by a 'statistical model'. Even a brief glance at some scientific articles will transport you into a terrifying jungle of different types of 'statistical model': you'll encounter tedious-looking names like *t*-test, ANOVA, regression, multilevel models and structural equation modelling. It might make you yearn for social media where the distinction between opinion and evidence need not trouble you. Fear not, though; I have a story that may help.

Many centuries ago there existed a cult of elite mathematicians. They spent 200 years trying to solve an equation that they believed would make them immortal. However, one of them forgot that when you multiply two minus numbers you get a positive number and instead of achieving eternal life they unleashed Cthulhu from his underwater city. It's amazing how small computational mistakes in maths can have these sorts of consequences. Anyway, the only way they could agree to get Cthulhu to return to his entrapment was by promising to infect the minds of humanity with confusion. They set about this task with gusto. They took the simple and elegant idea of a statistical model and reinvented it in hundreds of seemingly different ways (Figures 2.4 and 2.6). They described each model as though it were completely different from the rest. 'Ha!' they thought, 'that'll confuse students!' And confusion did indeed infect the minds of students. The statisticians kept their secret that all statistical models could be described in one simple, easy-to-understand equation locked away in a wooden box with Cthulhu's head burned into the lid. 'No one will open a box with a big squid head burnt into it', they reasoned. They were correct, until a Greek fisherman stumbled upon the box and, thinking it contained some vintage calamari, opened it. Disappointed with the contents, he sold the script inside the box on eBay. I bought it for €3 plus postage. This was money well spent, because it means that I can now give you the key that will unlock the mystery of statistics for ever. Everything in this book (and statistics generally) boils down to the following equation:

$$\text{outcome}_i = (\text{model}) + \text{error}_i \tag{2.1}$$

This equation means that the data we observe can be predicted from the model we choose to fit plus some amount of error.[3] The 'model' in the equation will vary depending on the design of your study,

3 The little *i* (e.g., outcome_i) refers to the *i*th score. Imagine, we had three scores collected from Andy, Zach and Zoë. We could replace the *i* with their name, so if we wanted to predict Zoë's score we could change the equation to: $\text{outcome}_{\text{Zoë}} = \text{model} + \text{error}_{\text{Zoë}}$. The *i* reflects the fact that the value of the outcome and the error can be different for each (in this case) person.



Figure 2.4 Thanks to the Confusion machine a simple equation is made to seem like lots of unrelated tests

the type of data you have and what it is you're trying to achieve with your model. Consequently, the model can also vary in its complexity. No matter how long the equation that describes your model might be, you can just close your eyes, reimagine it as the word 'model' (much less scary) and think of the equation above: we predict an outcome variable from some model (that may or may not be hideously complex) but we won't do so perfectly so there will be some error in there too. Next time you encounter some sleep-inducing phrase like 'hierarchical growth model' just remember that in most cases it's just a fancy way of saying 'predicting an outcome variable from some other variables'.

2.4 Populations and samples

Before we get stuck into a specific form of statistical model, it's worth remembering that scientists are usually interested in finding results that apply to an entire **population** of entities. For example, psychologists want to discover processes that occur in all humans, biologists might be interested in processes that occur in all cells, and economists want to build models that apply to all salaries. A population can be very general (all human beings) or very narrow (all male ginger cats called Bob). Usually, scientists strive to infer things about general populations rather than narrow ones. For example, it's not very interesting to conclude that psychology students with brown hair who own a pet hamster named George recover more quickly from sports injuries if the injury is massaged (unless you happen to be a psychology student with brown hair who has a pet hamster named George, like René Koning).[4] It will have a much wider impact if we can conclude that *everyone's* (or most people's) sports injuries are aided by massage.

Remember that our bridge-building engineer could not make a full-sized model of the bridge she wanted to build and instead built a small-scale model and tested it under various conditions. From the results obtained from the small-scale model she inferred things about how the full-sized bridge would respond. The small-scale model may respond differently than a full-sized version of the bridge, but the larger the model, the more likely it is to behave in the same way as the full-sized bridge. This metaphor can be extended to scientists: we rarely, if ever, have access to every member of a population (the real-sized bridge). Psychologists cannot collect data from every human being and ecologists cannot observe every male ginger cat called Bob. Therefore, we collect data from a smaller subset of the population

4 A brown-haired psychology student with a hamster called Sjors (Dutch for George, apparently) who emailed me to weaken my foolish belief that I'd generated an obscure combination of possibilities.



Jane Superbrain 2.2
What is a population? Ⓐ

It's convenient to talk about the 'population' as a living breathing entity. In all sorts of contexts (psychology, economics, biology, ecology, immunology, etc.) it can be useful to think of using a sample of data to represent 'the population' in a literal sense. That is, to think that our sample allows us to tap into 'what we would find if only we had collected data from the entire population'. That is how we tend to interpret models estimated in samples, and it is also how I write about populations when trying to explain things in Chapter 6. However, when we use a sample to estimate something in the population what we're really doing is testing (in the sample) whether it's reasonable to describe a process in a particular way. For example, in Chapter 1 we discussed how memes sometimes follow a normal distribution and I presented sample data in Figure 1.12 regarding the number of ice bucket challenge videos on YouTube since the first video. If our question is 'do internet memes follow a normal distribution?' then the 'population' isn't data from everyone on the planet, it is the normal curve. We're basically saying that internet memes can be described by the equation that defines a normal curve (Jane Superbrain Box 1.6):

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty$$

The 'population' is that equation. We then use the sample data to ask 'is that equation a reasonable model of the process (i.e. the life cycle of memes)?' If the sample data suggest that the population model is reasonable (that is, the equation for the normal curve is a decent enough approximation of the life cycle of memes) then we might draw a general conclusion like 'internet memes follow a normal curve in general and not just in our sample'. We will return to this point in Chapter 6.



known as a **sample** (the scaled-down bridge) and use these data to infer things about the population as a whole (Jane Superbrain Box 2.2). The bigger the sample, the more likely it is to reflect the whole population. If we take several random samples from the population, each of these samples will give us slightly different results but, on average, the results from large samples should be similar.

2.5 The linear model

I mentioned in Jane Superbrain Box 2.1 that a very common statistical model is the linear model, so let's have a look at it. A simple form of the linear model is given by

$$\text{outcome}_i = b_0 + b_1 X_{1i} + \text{error}_i \quad (2.2)$$

More formally, we usually denote the outcome variable with the letter Y, and error with the Greek letter epsilon (ε):

$$Y_i = b_0 + b_1 X_{1i} + \varepsilon_i \quad (2.3)$$

This model can be summed up as 'we want to predict the values of an outcome variable (Y_i) from a predictor variable (X_i)'. As a brief diversion, let's imagine that instead of b_0 we use the letter c, and instead of b_1 we use the letter m. Let's also ignore the error term. We could predict our outcome as follows:

$$Y_i = mx + c$$

Or if you're American, Canadian or Australian let's use the letter b instead of c:

$$Y_i = mx + b$$

Perhaps you're French, Dutch or Brazilian, in which case let's use a instead of m:

$$Y_i = ax + b$$

Do any of *these* equations look familiar? If not, there are two explanations: (1) you didn't pay enough attention at school; or (2) you're Latvian, Greek, Italian, Swedish, Romanian, Finnish, Russian or from some other country that has a different variant of the equation of a straight line. The fact that there are multiple ways to write the same equation using different letters illustrates that the symbols or letters in an equation are arbitrary choices (Jane Superbrain Box 2.3). Whether we write $mx + c$ or $b_1X + b_0$ doesn't really matter; what matters is what the symbols represent. So, what do the symbols represent?

I have talked throughout this book about fitting 'linear models', and linear simply means 'straight line'. All the equations above are forms of the equation of a straight line. Any straight line can be defined by two things: (1) the slope (or gradient) of the line (usually denoted by b_1); and (2) the point at which the line crosses the vertical axis of the graph (known as the *intercept* of the line, b_0). b_1 and b_0 will crop up throughout this book, where you see them referred to generally as b (without any subscript) or b_i (meaning the b associated with variable i). Figure 2.5 (left) shows a set of lines that have the same intercept but different gradients. For these three models, b_0 is the same in each but b_1 is different for each line. The b attached to a predictor variable represents the rate of change of the outcome variable as the predictor changes. In Section 3.7.4 we will see how relationships can be positive or negative (and I don't mean whether you and your partner argue all the time). A model with a positive b_1 describes a positive relationship, whereas a line with a negative b_1 describes a negative relationship. Looking at Figure 2.5 (left), the purple line describes a positive relationship, whereas the bile-coloured line describes a negative relationship. A steep slope will have a large value of b, which suggests the outcome changes a lot for a small change in the predictor (a large rate of change), whereas a relatively flat slope will have a small value of b, suggesting the outcome changes very little as the predictor changes (a small rate of change).

Figure 2.5 (right) also shows models that have the same gradients (b_1 is the same in each model) but different intercepts (b_0 is different in each model). The intercept determines the vertical height of the line and it represents *the overall level of the outcome variable when predictor variables are zero*.

As we will see in Chapter 9, the linear model quite naturally expands to predict an outcome from more than one predictor variable. Each new predictor variable gets added to the right-hand side of the equation and is assigned a b:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni} + \varepsilon_i \quad (2.4)$$

In this model, we're predicting the value of the outcome for a particular entity (i) from the value of the outcome when the predictors have a value of zero (b_0) and the entity's score on each of the predictor variables (X_{1i}, X_{2i}, up to and including the last predictor, X_{ni}). As with the one-predictor model, b_0 represents the overall level of the outcome variable when predictor variables are zero. Each predictor variable has a parameter (b_1, b_2) attached to it, which tells us something about the relationship between that predictor and the outcome when the other predictors are constant.

To sum up, we can use a linear model (i.e., a straight line) to summarize the relationship between two variables: the b attached to the predictor variable tells us what the model looks like (its shape) and the intercept (b_0) locates the model in geometric space.

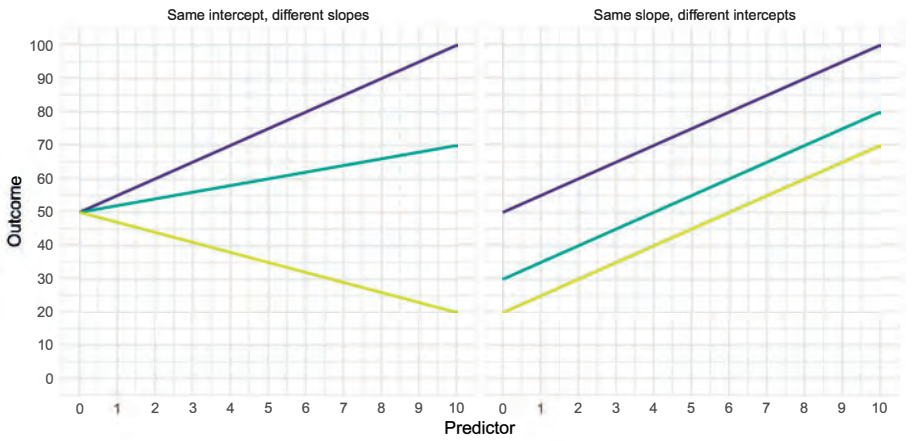


Figure 2.5 Lines that share the same intercept but have different gradients, and lines with the same gradients but different intercepts

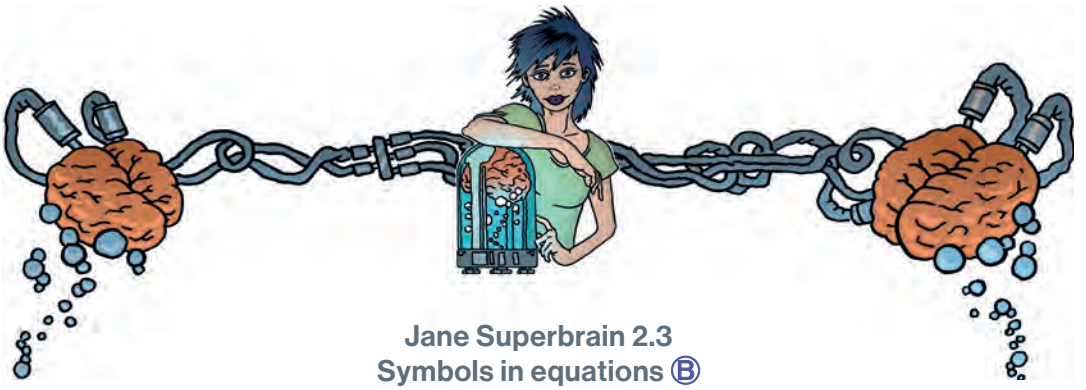
2.6 P is for parameters Ⓐ

We've just discovered the general form of the linear model, and we'll get into more detail from Chapter 9 onwards. We saw that the linear model (like all statistical models) is made up of variables and things that we have labelled as b. The bs in the model are known as **parameters**, the 'P' in the SPINE of statistics. Variables are measured constructs that vary across entities. In contrast, parameters are not measured and are (usually) constants believed to represent some fundamental truth about the variables in the model.

In Jane Superbrain Box 2.1 we looked at an example of people rating the honesty of acts (from 0 = appalling behaviour to 10 = it's OK really) and how much they liked the perpetrator (0 = not at all, 10 = a lot). Now, let's imagine that we are interested in knowing some 'fundamental truths' about ratings of the honesty of acts. Let's start as simply as we can and imagine that we are interested only in the overall level of 'honesty'. Put another way, what is a typical rating of honesty? Our model has no predictor variables and only an outcome and a single parameter (the intercept, b_0):

$$\begin{aligned} Y_i &= b_0 + \varepsilon_i \\ \text{honesty}_i &= b_0 + \varepsilon_i \end{aligned} \tag{2.5}$$

With no predictor variables in the model, the value of b_0 is the mean (average) honesty levels. Statisticians try to confuse you by giving different estimates of parameters different symbols and letters; for example, you might see the mean written as $\bar{X}$. However, symbols are arbitrary (Jane Superbrain Box 2.3) so try not to let that derail you.



Jane Superbrain 2.3
Symbols in equations Ⓑ

Throughout this book, model parameters are typically represented with the letter *b*; for example, Eq. 2.3 is written as

$$Y_i = b_0 + b_1X_i + \varepsilon_i$$

Although the choice of *b* to represent parameters is not unusual, you'll sometimes see this equation written as

$$Y_i = \beta_0 + \beta_1X_i + \varepsilon_i$$

The only difference is that this equation uses the Greek letter β instead of *b*. We could use the Greek letter gamma (γ):

$$Y_i = \gamma_0 + \gamma_1X_i + \varepsilon_i$$

These three equations are mathematically equivalent, but they use different symbols to represent the model parameters. These symbols are arbitrary, there is nothing special about them. We could use skulls to represent the parameters and a ghost to represent the outcome if we like, which might be more representative of our emotional states when thinking about statistical models:[5]

$$\text{👻}_i = \text{💀}_0 + \text{💀}_1X_i + e_i$$

Tempting though it is to kick-start the trend to represent parameters with spooky stuff, reluctantly we'll stick with convention and use letters. The main point to bear in mind as we start to learn about how to interpret the various parameters is that these interpretations are unchanged by whether we represent them with *b*, β, γ, a skull or an ice-skating kitten.



5 This equation was created using the *halloweenmath* package for LaTeX, which I wish I had invented.

Often we are not just interested in the typical value of the outcome variable, we want to predict it from a different variable. In this example, we might want to predict honesty ratings from ratings of how likeable the perpetrator is. The model expands to include this predictor variable:

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{honesty}_i &= b_0 + b_1 \text{likeableness}_i + \varepsilon_i \end{aligned} \tag{2.6}$$

Now we're predicting the value of the outcome for a particular entity (i) not just from the value of the outcome when the predictor variables are equal to zero (b_0) but from the entity's score on the predictor variable (X_i). The predictor variable has a parameter (b_1) attached to it, which tells us something about the relationship between the predictor (X_i), in this case likeableness ratings, and outcome, in this case honesty ratings. If we want to predict an outcome from two predictors then we can add another predictor to the model Eq. 2.7. For example, imagine we also had independent ratings of the severity of the dishonest act. We could add these ratings to the model as a predictor:

$$\begin{aligned} Y_i &= b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i \\ \text{honesty}_i &= b_0 + b_1 \text{likeableness}_i + b_2 \text{severity}_i + \varepsilon_i \end{aligned} \tag{2.7}$$

In this model, we're predicting the value of the outcome for a particular entity (i) from the value of the outcome when likeableness and severity are both zero (b_0) and the entity's score on likeableness and severity. The parameter b_1 quantifies the change in honesty as likeableness changes (when severity is constant) and b_2 quantifies the change in honesty as severity changes (when likeableness is constant). We'll expand on these ideas in Chapter 9.

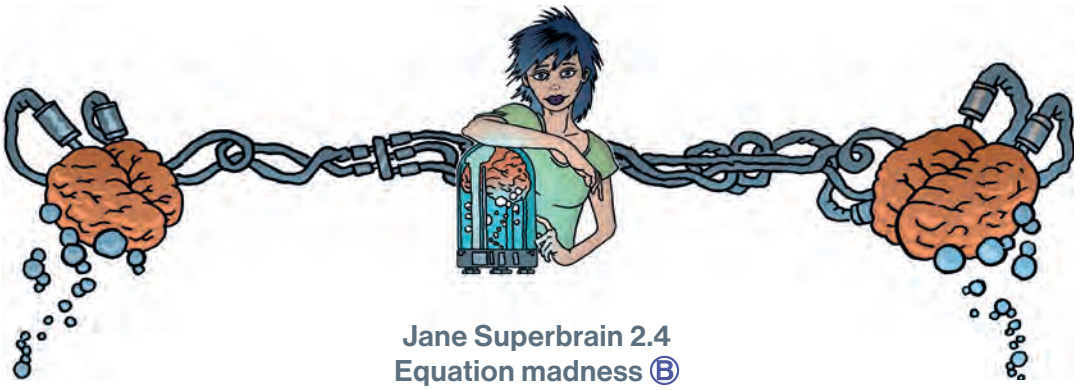
Hopefully what you can take from this section is that this book boils down to the idea that we can predict values of an outcome variable based on a model. The form of the model changes, there will always be some error in prediction, and there will always be parameters that tell us about the shape or form of the model.

To work out what the model looks like we estimate the parameters (i.e., the value(s) of b). You'll hear the phrase 'estimate the parameter' or 'parameter estimates' a lot in statistics, and you might wonder why we use the word 'estimate'. Surely statistics has evolved enough that we can compute exact values of things and not merely estimate them? As I mentioned before, we're interested in drawing conclusions about a population (to which we don't have access). In other words, we want to know what our model might look like in the whole population. Given that our model is defined by parameters, this amounts to saying that we're not interested in the parameter values in our sample, but we care about the parameter values in the population. The problem is that we don't know what the parameter values are in the population because we didn't measure the population, we measured only a sample. However, we can use the sample data to *estimate* what the population parameter values are likely to be. That's why we use the word 'estimate', because when we calculate parameters based on sample data they are only estimates of what the true parameter value is in the population.

Let's reflect on what we mean by a population (Jane Superbrain Box 2.2) by returning to the model that predicts honesty ratings from likeableness ratings:

$$\text{honesty}_i = b_0 + b_1 \text{likeableness}_i + \varepsilon_i \tag{2.8}$$

This equation is our population; we are saying 'the relationship between ratings of likeableness and honesty can be reasonably modelled by this equation'. We then use the sample data to (1) see whether the equation is a reasonable model of the relationship and, if so, (2) what are the likely values of the b_0 and b_1. If the model (equation) is a reasonable approximation of the process (the relationship between ratings of likeableness and honesty) then we'd use the estimated values of the b_0 and b_1 within the model to make predictions beyond the sample. In other words, the sample estimates of the bs are assumed to reflect the 'true' values. Let's make these ideas a bit more concrete by going back to the simplest linear model: the mean.



Jane Superbrain 2.4
Equation madness Ⓑ

The linear model can be presented in different ways depending on whether we're referring to the population model, the sample model or discussing the values predicted by the model. To take the example of the linear model with one predictor variable, the population model is referred to as

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

(although as mentioned before, β is often used instead of b). The i refers a generic entity in the population, not an entity from/about which we collected data. When referring to the sample model, we make a few changes:

$$y_i = \hat{b}_0 + \hat{b}_1 x_i + e_i$$

First, we add hats to the bs to make clear that these are estimates of the population values from the data, we typically use lower-case y and x to refer to the variables, and we use e to refer to errors because we're referring to observed errors in the sample. In this equation the i refers to a specific entity whose data was collected. Finally, we sometimes refer to values of the outcome variable predicted by the model, in which case we write

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$

Notice that the outcome variable now has a hat to indicate that it is a value estimated (or predicted) from the model, and the error term disappears because predicted values have no error (because they are not observed).



2.6.1 The mean as a statistical model Ⓐ

We encountered the mean in Section 1.7.4, where I briefly mentioned that it was a statistical model because it is a hypothetical value and not necessarily one that is observed in the data. For example, if we took five statistics lecturers and measured the number of friends that they had, we might find the following data: 1, 2, 3, 3 and 4. If we want to know the mean number of friends, this can be calculated by adding the values we obtained, and dividing by the number of values measured: $(1 + 2 + 3 + 3 + 4)/5 = 2.6$.

It is impossible to have 2.6 friends, so the mean value is a *hypothetical* value: it is a model created to summarize the data and there will be error in prediction. As in Eq. 2.5 the model is:

$$\begin{aligned} Y_i &= b_0 + \varepsilon_i \\ \text{friends}_i &= b_0 + \varepsilon_i \end{aligned} \tag{2.9}$$

in which the parameter, b_0, is the mean of the outcome. The important thing is that we can use the value of the mean (or any parameter) computed in our sample to estimate the value in the population (which is the value in which we're interested). We give estimates little hats to compensate them for the lack of self-esteem they feel at not being true values. Who doesn't love a hat?

$$\text{friends}_i = \hat{b}_0 + e_i \tag{2.10}$$

When you see equations where these little hats are used, try not to be confused: all the hats are doing is making explicit that the values underneath them are estimates. Imagine the parameter as wearing a little baseball cap with the word 'estimate' printed along the front. We'll talk more about notation in Chapter 6. For now, run with it. In the case of the mean, we estimate the population value by assuming that it is the same as the value in the sample (in this case 2.6).



Figure 2.6 Thanks to the Confusion machine there are lots of terms that basically refer to error

2.6.2 Assessing the fit of a model: sums of squares and variance revisited Ⓐ

It's important to assess the fit of any statistical model (to return to our bridge analogy, we need to know how representative the model bridge is of the bridge that we want to build). With most statistical models we can determine whether the model represents the data well by looking at how different the scores we observed in the data are from the values that the model predicts. For example, let's look what happens when we use the model of 'the mean' to predict how many friends the first lecturer has. The first lecture was called Andy; it's a small world. We observed that lecturer 1 had one friend and the model (i.e., the mean of all lecturers) predicted 2.6. By rearranging Eq. 2.1 we see that this is an error of −1.6:[6]

6 Remember that I'm using the symbol $\hat{b}_0$ to represent the mean. If this upsets you then replace it (in your mind) with the more traditionally used symbol, $\bar{X}$.

$$\begin{aligned} y_{\text{lecturer 1}} &= \hat{b}_0 + e_{\text{lecturer 1}} \\ e_{\text{lecturer 1}} &= y_{\text{lecturer 1}} - \hat{b}_0 \\ &= 1 - 2.6 \\ &= -1.6 \end{aligned} \tag{2.11}$$

You might notice that all we have done here is calculate the deviance, which we encountered in Section 1.7.5. The *deviance* is another word for *error* (Figure 2.6). A more general way to think of the deviance or error is by rearranging Eq. 2.1:

$$\text{deviance}_i = \text{outcome}_i - \text{model}_i \tag{2.12}$$

In other words, the error or deviance for a particular entity is the score predicted by the model for that entity (which we denote by $\hat{y}_i$ because we need a hat to remind us that it is an estimate) subtracted from the corresponding observed score (y_i):

$$\text{deviance}_i = y_i - \hat{y}_i \tag{2.13}$$

Figure 2.7 shows the number of friends that each statistics lecturer had, and the mean number that we calculated earlier on. The line representing the mean represents the predicted value from the model ($\hat{y}_i$). The predicted value is the same for all entities and has been highlighted for the first two lecturers. The dots are the observed data (y_i), which have again been highlighted for the first two lecturers. The diagram also has a series of vertical lines that connect each observed value to the mean value. These lines represent the error or deviance of the model for each lecturer (e_i). The first lecturer, Andy, had only 1 friend (a glove puppet of a pink hippo called Professor Hippo) and we have already seen that the error for this lecturer is –1.6; the fact that it is a negative number shows that our model *overestimates* Andy's popularity: it predicts that he will have 2.6 friends but, in reality, he has only 1 (bless!)

We know the accuracy or 'fit' of the model for a particular lecturer, Andy, but we want to know the fit of the model *overall*. We saw in Section 1.7.5 that we can't add deviances because some errors are positive and others negative and so we'd get a total of zero:

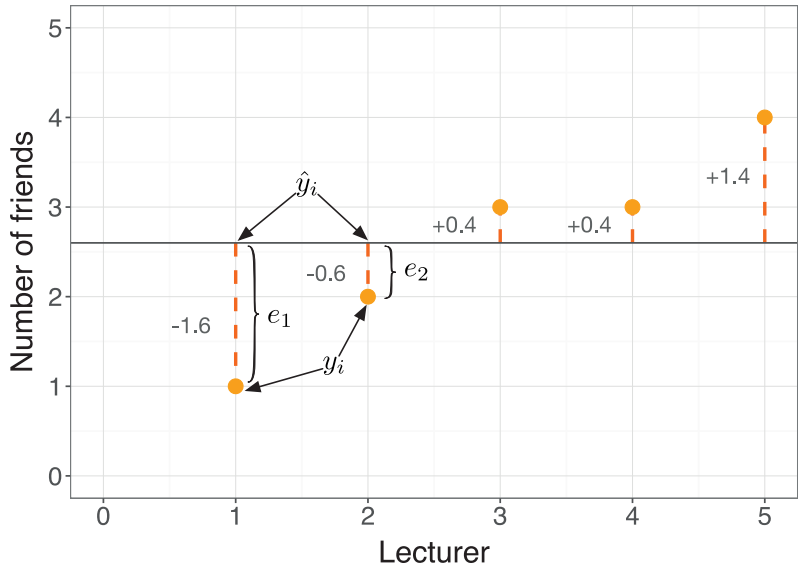


Figure 2.7 Plot showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends

$$\begin{aligned}\text{total error} &= \text{sum of error} \\ &= \sum_{i=1}^{n}(\text{outcome}_i - \text{model}_i) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= (-1.6)+(-0.6)+(0.4)+(0.4)+(1.4) = 0\end{aligned} \tag{2.14}$$

We also saw in Section 1.7.5 that one way around this problem is to square the errors. This would give us a value of 5.2:

$$\begin{aligned}\text{sum of squared errors (SS)} &= \sum_{i=1}^{n}(\text{outcome}_i - \text{model}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= (-1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2 \\ &= 2.56 + 0.36 + 0.16 + 0.16 + 1.96 \\ &= 5.20\end{aligned} \tag{2.15}$$

Does this equation look familiar? It ought to, because it's the same as Eq. 1.6 for the sum of squares in Section 1.7.5; the only difference is that Eq. 1.6 was specific to when our model is the mean, so the 'model' was replaced with the symbol for the mean ($\bar{x}$), and the outcome was replaced by the letter x, which is a commonly used to represent a score on a variable:

$$\sum_{i=1}^{n}(\text{outcome}_i - \text{model}_i)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.16}$$

Another example of differences in notation creating the impression that similar things are different. However, when we're thinking about models more generally, this illustrates that we can think of the total error in terms of this general equation:

$$\text{total error} = \sum_{i=1}^{n}(\text{observed}_i - \text{model}_i)^2 \tag{2.17}$$

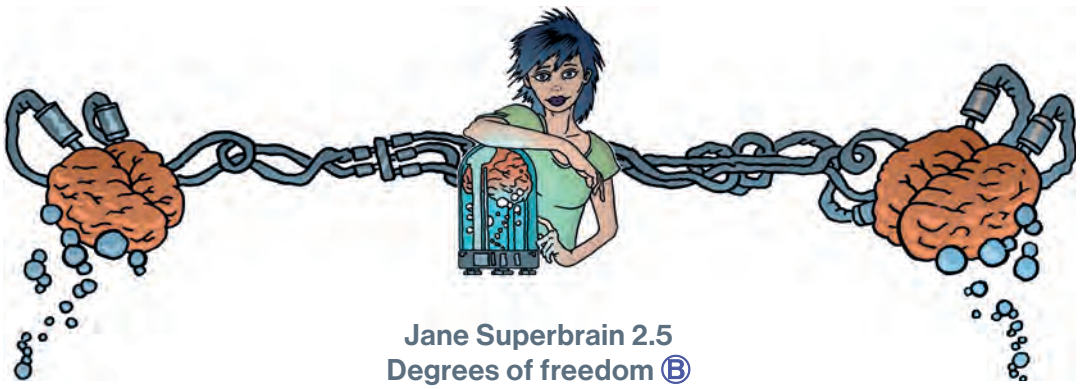
This equation shows how something we have used before (the sum of squares) can be used to assess the total error in any model (not just the mean).

We saw in Section 1.7.5 that although the sum of squared errors (SS) is a good measure of the accuracy of our model, it depends upon the quantity of data that has been collected – the more data points, the higher the SS. We also saw that we can overcome this problem by using the average error, rather than the total. To compute the average error we divide the sum of squares (i.e., the total error) by the number of values (N) that we used to compute that total. We again come back to the problem that we're usually interested in the error in the model in the population (not the sample). To estimate the mean error in the population we need to divide not by the number of scores contributing to the total, but by the **degrees of freedom** (df), which is the number of scores used to compute the total adjusted for the fact that we're trying to estimate the population value (Jane Superbrain Box 2.5):

$$\begin{aligned}\text{mean squared error} &= \frac{SS}{df} \\ &= \frac{\sum_{i=1}^{n}(\text{outcome}_i - \text{model}_i)^2}{N-1} \\ &= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N-1}\end{aligned} \tag{2.18}$$

Does this equation look familiar? Again, it ought to, because it's a more general form of the equation for the variance (Eq. 1.7). We can see this if we replace the predicted model values ($\hat{y}_i$) with the symbol for the mean ($\bar{x}$), and use the letter x (instead of y) to represent a score on the outcome. Lo and behold, the equation transforms into that of the variance:

$$\text{mean squared error} = \frac{SS}{df} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1} = \frac{5.20}{4} = 1.30 \tag{2.19}$$



Jane Superbrain 2.5
Degrees of freedom Ⓑ

The concept of degrees of freedom (*df*) is very difficult to explain. I'll begin with an analogy. Imagine you're the manager of a sports team (I'll try to keep it general so you can think of whatever sport you follow, but in my mind I'm thinking about soccer). On the morning of the game you have a team sheet with (in the case of soccer) 11 empty slots relating to the positions on the playing field. Different players have different positions on the field that determine their role (defence, attack, etc.) and to some extent their physical location (left, right, forward, back). When the first player arrives, you have the choice of 11 positions in which to place this player. You place their name in one of the slots and allocate them to a position (e.g., striker) and, therefore, one position on the pitch is now occupied. When the next player arrives, you have 10 positions that are still free: you have 'a degree of freedom' to choose a position for this player (they could be put in defence, midfield, etc.). As more players arrive, your choices become increasingly limited: perhaps you have enough defenders so you need to start allocating some people to attack, where you have positions unfilled. At some point you will have filled 10 positions and the final player arrives. With this player you have no 'degree of freedom' to choose where he or she plays – there is only one position left. In this scenario, there are 10 degrees of freedom: for 10 players you have a degree of choice over where they play, but for one player you have no choice. The degrees of freedom are one less than the number of players.

In statistical terms the degrees of freedom relate to the number of observations that are free to vary. If we take a sample of four observations from a population, then these four scores are free to vary in any way (they can be any value). We use these four sampled values to estimate the mean of the population. Let's say that the mean of the sample was 10 and, therefore, we estimate that the population mean is also 10. The value of this parameter is now fixed: we have held one parameter constant. Imagine we now want to use this sample of four observations to estimate the mean squared error in the population. To do this,

we need to use the value of the population mean, which we estimated to be the fixed value of 10. With the mean fixed, are all four scores in our sample free to be sampled? The answer is no, because to ensure that the population mean is 10 only three values are free to vary. For example, if the values in the sample we collected were 8, 9, 11, 12 (mean = 10) then the first value sampled could be any value from the population, say 9. The second value can also be any value from the population, say 12. Like our football team, the third value sampled can also be any value from the population, say 8. We now have values of 8, 9 and 12 in the sample. The final value we sample, the final player to turn up to the soccer game, cannot be any value in the population, it *has to be* 11 because this is the value that makes the mean of the sample equal to 10 (the population parameter that we have held constant). Therefore, if we hold one parameter constant then the degrees of freedom must be one fewer than the number of scores used to calculate that parameter. This fact explains why when we use a sample to estimate the mean squared error (or indeed the standard deviation) of a population, we divide the sums of squares by $N - 1$ rather than N alone. There is a lengthier explanation in Field (2016).



To sum up, we can use the sum of squared errors and the mean squared error to assess the fit of a model. The variance, which measures the 'fit' of the mean, is an example of this general principle of squaring and adding errors to assess fit. This general principle can be applied to more complex models than the mean. Sums of squared errors (or average squared errors) give us an idea of how well a model fits the data: large values relative to the model indicate a lack of fit. Think back to Figure 1.10, which showed students' ratings of five lectures given by two lecturers. These lecturers differed in their mean squared error:[7] lecturer 1 had a smaller mean squared error than lecturer 2. Compare their plots: the ratings for lecturer 1 were consistently close to the mean rating, indicating that the mean is a good representation of the observed data – it is a good fit. The ratings for lecturer 2, however, were more spread out from the mean: for some lectures, she received very high ratings, and for others her ratings were terrible. Therefore, the mean is not such a good representation of the observed scores – it is a poor fit.

2.7 E is for estimating parameters Ⓐ

We have seen that models are defined by parameters, and these parameters need to be estimated from the data that we collect. Estimation is the E in the SPINE of statistics. We used an example of the mean because it was familiar, and we will continue to use it to illustrate a general principle about how parameters are estimated. Let's imagine that one day we walked down the road and fell into a hole. Not just any old hole, but a hole created by a rupture in the space-time continuum. We slid down the hole, which turned out to be a sort of U-shaped tunnel under the road, and we emerged out of the other end to find that not only were we on the other side of the road, but we'd gone back in time a few hundred years. Consequently, statistics had not been invented and neither had the equation to compute the mean. Happier times, you might think. A tiny newt with a booming baritone voice accosts us, demanding to know the average number of friends that a lecturer has. This kind of stuff happens to me *all the time*. If we didn't know the equation for computing the mean, how might we

7 I reported the standard deviation but this value is the square root of the variance, and therefore, the square root of the mean square error.

answer his question? We could guess and see how well our guess fits the data. Remember, we want the value of the parameter $\hat{b}_0$ in Eq. 2.10:

$$\text{friends}_i = \hat{b}_0 + e_i$$

We know already that we can rearrange this equation to give us the error for each person:

$$e_i = \text{friends}_i - \hat{b}_0$$

If we add the squared error for each person then we'll get the sum of squared errors, which we can use as a measure of 'fit'. Imagine we begin by guessing that the mean number of friends that a lecturer has is 2. We'll call this guess $\hat{b}_1$. We can compute the error for each lecturer by subtracting this value from the number of friends they actually had. We then square this value to get rid of any minus signs, and we add up these squared errors. Table 2.1 shows this process, and we find that by guessing a value of 2, we end up with a total squared error of 7. Now let's take another guess, which we'll call $\hat{b}_2$, and this time we'll guess the value is 4. Again, we compute the sum of squared errors as a measure of 'fit'. This model (i.e., guess) is worse than the last because the total squared error is larger than before: it is 15. We could carry on guessing different values of the parameter and calculating the error for each guess. If we were nerds with nothing better to do, we could play this guessing game for all possible values of the mean ($\hat{b}_0$), but you're probably too busy being rad hipsters and doing whatever it is that rad hipsters do. I, however, am a badge-wearing nerd, so I have plotted the results of this process in Figure 2.8, which shows the sum of squared errors that you would get for various potential values of the parameter $\hat{b}_0$. Note that, as we just calculated, when $\hat{b}_0$ is 2 we get an error of 7, and when it is 4 we get an error of 15. The shape of the line is interesting because it curves to a minimum value – a value that produces the lowest sum of squared errors given the data. The value of b at the lowest point of the curve is 2.6, and it produces an error of 5.2. Do these values seem familiar? They should because they are the values of the mean and sum of squared errors that we calculated earlier. This example illustrates that the equation for the mean is designed to estimate that parameter to minimize the error. In other words, it is the value that has the least squared error. The estimation process we have just explored, which uses the principle of minimizing the sum of squared errors, is known generally as the **method of least squares** or **ordinary least squares** (OLS). The mean is, therefore, an example of a least squares estimator. Of course, when using least squares estimation we don't iterate through every possible value of a parameter, we use some clever maths (differentiation) to get us directly to the minimum value of the squared error curve for the data.

Table 2.1 Guessing the mean

Number of Friends (x_i)	$\hat{b}_1$	Squared error $(x_i - \hat{b}_1)^2$	$\hat{b}_2$	Squared error $(x_i - \hat{b}_2)^2$
1	2	1	4	9
2	2	0	4	4
3	2	1	4	1
3	2	1	4	1
4	2	4	4	0
		$\sum_{i=1}^{n}(x_i - b_1)^2 = 7$		$\sum_{i=1}^{n}(x_i - b_2)^2 = 15$

The other very widely used estimation method that we meet in this book is **maximum likelihood estimation**. This method is a bit trickier to explain but, in a general sense, there are similar principles at play. With least squares estimation the goal is to find (mathematically) values for parameters

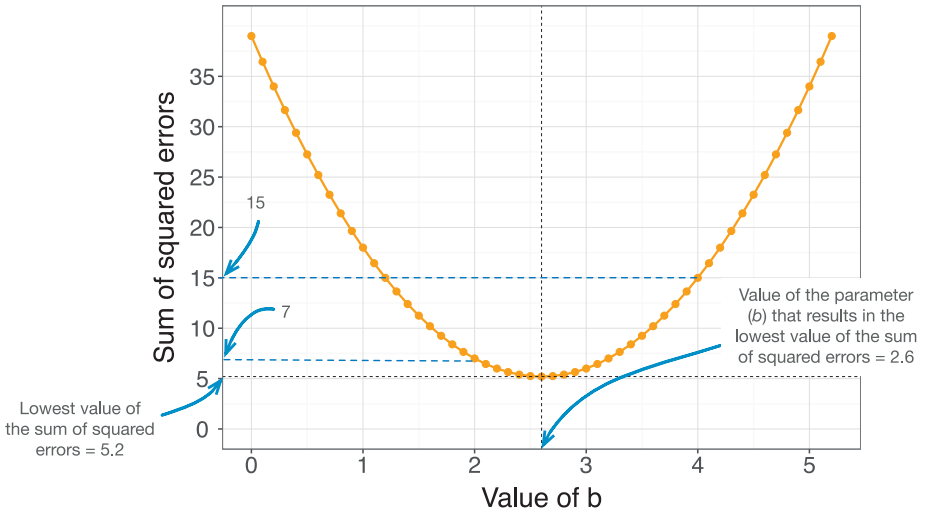


Figure 2.8 Plot showing the sum of squared errors for different 'guesses' of the mean

that *minimize* the sum of squared error, whereas in maximum likelihood estimation the goal is to find values that *maximize* something called the **likelihood**. The likelihood of a score is the output of the probability density function that describes the population of scores. For example, if the population is normally distributed the likelihood of a score is defined by the probability density function for the normal distribution, which is the rather hideous equation from Jane Superbrain Box 1.6, which I'll repeat below to ruin your day, mwha ha ha ha:

$$l_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \tag{2.20}$$

in which, x_i is a score, μ is the population mean and σ is the population standard deviation. Having seen that equation, let's engage in some meditation to restore inner peace and calm. Fortunately, we can ignore most of the equation because the bit that matters is this:

$$\frac{x_i - \mu}{\sigma} \tag{2.21}$$

which describes the standardized distance of a score from the population mean.[8] Notice the format is the same as that of a *z*-score (Eq. 1.11): the top half of the equation represents the difference between the score and the mean and the bottom half standardizes this difference by dividing by the population standard deviation.

Imagine that you have some scores and you happen to know the population mean (let's say it's $\mu = 3$) and variance (for argument's sake, $\sigma^2 = 1$). You can plug these values into Eq. 2.20 along with each score (x_i) to find the likelihood of the score. If you were to do this for every possible value of x_i and plot the results you would see a normal curve (because we used the likelihood function for the normal distribution). This idea is illustrated in Figure 2.9. The curve is generated using Eq. 2.20; it represents the likelihoods for each score. For any score we can use the curve to read off the associated likelihood. For example, a score of 1 (or indeed 5) has a likelihood of 0.054 whereas a score of 4 (or indeed 2) has a likelihood of 0.242. You can think of these likelihoods as *relative* probabilities.

8 Most of the rest of the equation is there to make sure that the area under the distribution sums to 1.

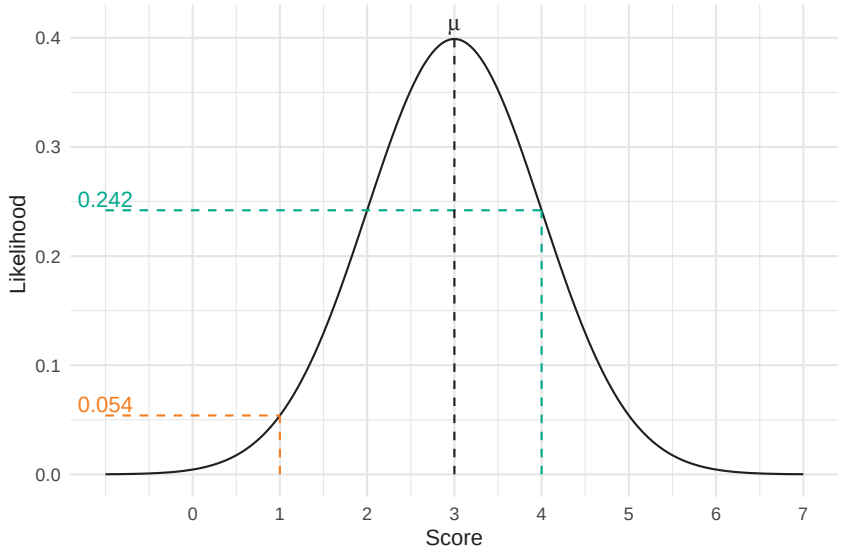


Figure 2.9 Illustration of the likelihood

They don't tell us the probability of a score occurring, but they do tell us the probability *relative* to other scores. For example, we know that scores of 4 occur more often than scores of 1 because the likelihood for a score of 4 (0.242) is higher than for a score of 1 (0.054). Notice that as scores get closer to the population mean the likelihood increases such that it reaches its maximum value at the population mean. At this point Eq. 2.21 is zero because $x_i - \mu$ is zero and Eq. 2.20 reduces to the part to the left of e (because $e^0 = 1$). Conversely, as scores get further from the population mean, $x_i - \mu$ gets larger, the values from Eq. 2.21 also get larger and consequently the values from Eq. 2.20 become smaller (because e is being raised to an increasingly negative power).

This is all well and good but, of course, we don't know the population mean and variance; these are precisely the things we're trying to estimate from the data. So, although this discussion is useful for understanding the likelihood, it is hypothetical and we need to move on to think about how we use likelihoods to estimate these values. In fact, maximum likelihood estimation is conceptually quite similar to least squares estimation. Least squares estimation tries to find parameter values that minimize squared error. So, it uses squared error as a metric for the fit of a given parameter value to the data and tries to minimize it because smaller squared error equates to better fit. Maximum likelihood estimation uses the likelihood as an indicator of fit rather than squared error, and because large likelihood values indicate better fit we maximize it (not minimize it). Least squares estimation uses the sum of squared error to quantify error across all scores, but you can't sum likelihoods.[9] Instead, you can multiply them to quantify the combined likelihood across all scores.

Let's repeat the thought exercise that we did for least squares where we 'guess' the parameter value as being 2 (μ_1) or 4 (μ_2), calculate the likelihood for each score and then work out the combined likelihood across all scores. Table 2.2 shows this process (compare it to Table 2.1). We have the scores for the number of friends as before, and again we're 'guessing' parameter values of 2 and 4. The values in the column L_1 come from Eq. 2.20. For each score we plug in values of $\mu = 2$, $\sigma^2 = 1.3$ (which normally we'd be estimating too but to keep things simple I've used the variance of the five observed scores) and the score itself (x_i). The bottom row shows the result of multiplying the values in the column together. In other words it's the combined likelihood across all five scores.

9 Actually you can if you convert them to log values, which is very often done. The result is known as the log-likelihood.

Table 2.2 Guessing the mean (again)

Number of Friends (x_i)	μ_1	L_1	μ_2	L_2
1	2	0.238	4	0.011
2	2	0.350	4	0.075
3	2	0.238	4	0.238
3	2	0.238	4	0.238
4	2	0.075	4	0.350
		$\prod_{i=1}^{n} L_1 = 3.54 \times 10^{-4}$		$\prod_{i=1}^{n} L_2 = 1.64 \times 10^{-5}$

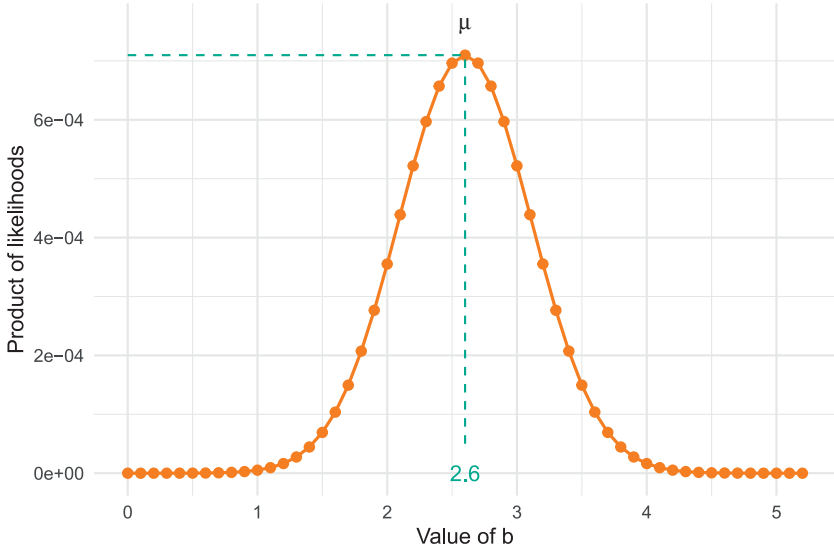


Figure 2.10 Plot showing the likelihood for different 'guesses' of the mean

The values in the column L_2 are generated in the same way except using $\mu = 4$. We can see that our guess of 2 is better than our guess of 4 because it generated a larger combined likelihood (3.54×10^{-4} is greater than 1.64×10^{-5}).

As we did before, we could repeat this process for every possible 'guess' and see which one has the largest likelihood (i.e. which one maximizes the likelihood). This idea is shown in Figure 2.10. Notice that the 'guess' that maximizes the likelihood is 2.6, the same value as when using least squares estimation. In fact there are situations where the least squares estimate and the maximum likelihood estimate are the same thing. Of course, when using maximum likelihood estimation we don't iterate through every possible value of a parameter, we use some clever maths (differentiation again) to get us directly to the value that maximizes the combined likelihood.

Throughout this book we will fit lots of different models, with parameters other than the mean that need to be estimated, and where we're estimating more than a single parameter. Although the maths that generates the estimates gets more complex, the processes follow the principles of minimizing squared error or maximizing the likelihood: the estimation methods will give you the parameter that has the least squared error or maximizes the likelihood given the data you have. It's worth reiterating that this is not the same thing as the parameter being accurate, unbiased or representative of the population: it could be the best of a bad bunch.

2.8 S is for standard error Ⓐ

We have looked at how we can fit a statistical model to a set of observations to summarize those data. It's one thing to summarize the data that you have actually collected, but in Chapter 1 we saw that good theories should say something about the wider world. It is one thing to be able to say that a sample of high-street stores in Brighton improved profits by placing cats in their store windows, but it's more useful to be able to say, based on our sample, that all high-street stores can increase profits by placing cats in their window displays. To do this we need to go beyond the data, and to go beyond the data we need to begin to look at how representative our samples are of the population of interest. This idea brings us to the S in the SPINE of statistics: the *standard error*.

In Chapter 1 we saw that the standard deviation tells us about how well the mean represents the sample data. However, if we're using the sample mean to estimate this parameter in the population, then we need to know how well it represents the value in the population, especially because samples from a population differ. Imagine that we were interested in the student ratings of all lecturers (so lecturers in general are the population). We could take a sample from this population, and this sample would be one of many possible ones. If we were to take several samples from the same population, then each sample would have its own mean, and some of these sample means will be different. Figure 2.11 illustrates the process of taking samples from a population. Imagine for a fleeting second that we eat some magic beans that transport us to an astral plane where we can see for a few short, but beautiful, seconds the ratings of all lecturers in the world. We're in this astral plane just long enough to compute the mean of these ratings (which, given the size of the population, implies we're there for quite some time). Thanks to our astral adventure we know, as an absolute fact, that the mean of all ratings is 3 (this is the *population mean*, μ, the parameter that we're trying to estimate).

Back in the real world, where we don't have magic beans, we also don't have access to the population, so we use a sample. In this sample we calculate the average rating, known as the *sample mean*, and discover it is 3; that is, lecturers were rated, on average, as 3. 'That was fun,' we think to ourselves, 'let's do it again.' We take a second sample and find that lecturers were rated, on average, as only 2. In other words, the sample mean is different in the second sample than in the first. This difference illustrates **sampling variation**: that is, samples vary because they contain different members of the population; a sample that, by chance, includes some very good lecturers will have a higher average than a sample that, by chance, includes some awful lecturers.

Imagine that we're so excited by this sampling malarkey that we take another seven samples so that we have nine in total (as in Figure 2.11). If we plotted the resulting sample means as a frequency distribution, or histogram,[10] we would see that three samples had a mean of 3, means of 2 and 4 occurred in two samples each, and means of 1 and 5 occurred in only one sample each. The end result is a nice symmetrical distribution known as a **sampling distribution**. A sampling distribution is the frequency distribution of sample means (or whatever parameter you're trying to estimate) from the same population. You need to imagine that we're taking hundreds or thousands of samples to construct a sampling distribution – I'm using nine to keep the diagram simple. The sampling distribution is a bit like a unicorn: we can imagine what one looks like, we can appreciate its beauty, and we can wonder at its magical feats, but the sad truth is that you'll never see a real one. They both exist as ideas rather than physical things. You would never go out and actually collect thousands of samples and draw a frequency distribution of their means. Instead, very clever statisticians have used maths to describe what these distributions look like and how they behave. Likewise, you'd be ill-advised to search for unicorns.

The sampling distribution of the mean tells us about the behaviour of samples from the population, and you'll notice that it is centred at the same value as the mean of the population (i.e., 3). Therefore, if

10 This is a graph of possible values of the sample mean plotted against the number of samples that have a mean of that value – see Section 2.9.1 for more details.

we took the average of all sample means we'd get the value of the population mean. We can use the sampling distribution to tell us how representative a sample is of the population. Think back to the standard deviation. We used the standard deviation as a measure of how representative the mean was of the observed data. A small standard deviation represented a scenario in which most data points were close to the mean, whereas a large standard deviation represented a situation in which data points were widely spread from the mean. If our 'observed data' are *sample* means then the standard deviation of these sample means would similarly tell us how widely spread (i.e., how representative) sample means are around their average. Bearing in mind that the average of the sample means is the same as the population mean, the standard deviation of the sample means would therefore tell us how widely sample means are spread around the population mean: put another way, it tells us whether sample means are typically representative of the population mean.

The standard deviation of sample means is known as the **standard error of the mean (SE)** or **standard error** for short. In the land where unicorns exist, the standard error could be calculated

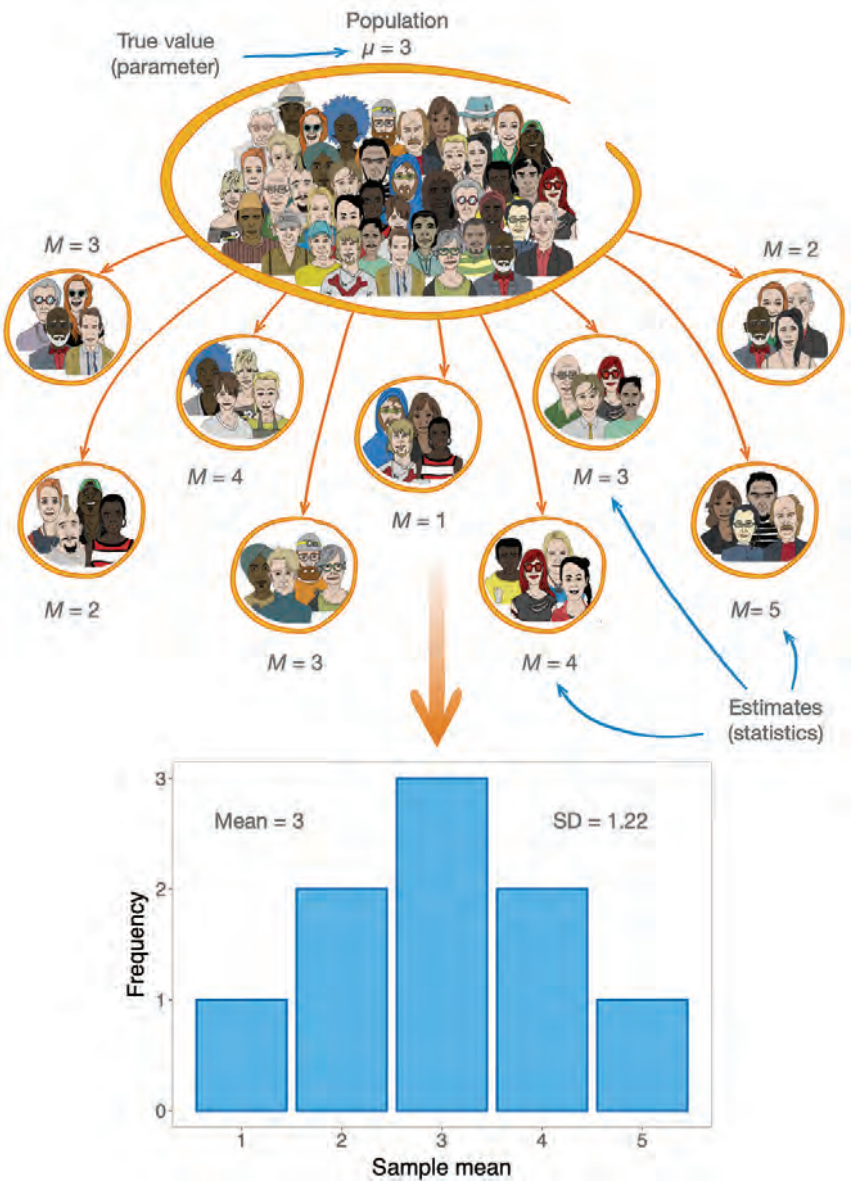
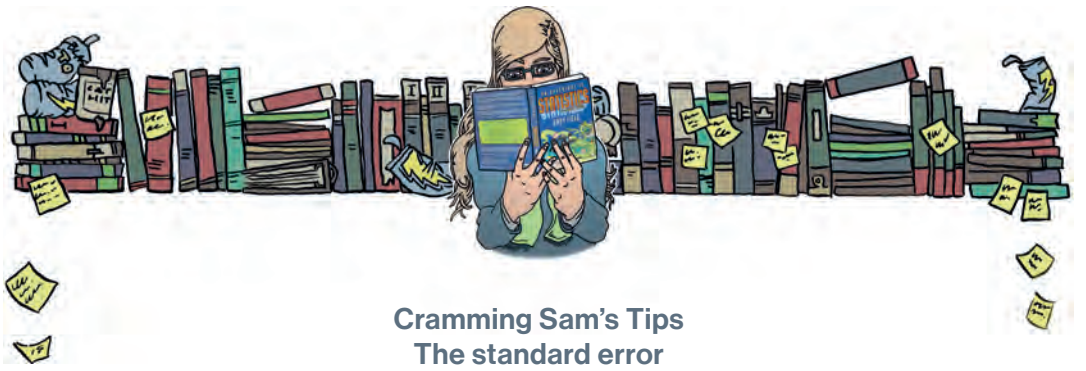


Figure 2.11 Illustration of the standard error (see text for details)

by taking the difference between each sample mean and the overall mean, squaring these differences, adding them up and then dividing by the number of samples. Finally, the square root of this value would need to be taken to get the standard deviation of sample means: the standard error. In the real world, we compute the standard error from a mathematical approximation. Some exceptionally clever statisticians have demonstrated something called the **central limit theorem** which tells us that as samples get large (usually defined as greater than 30), the sampling distribution has a normal distribution with a mean equal to the population mean, and a standard deviation given by

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}} \tag{2.22}$$

We will return to the central limit theorem in more detail in Chapter 6, but I've mentioned it here because it tells us that if our sample is large we can use Eq. 2.22 to approximate the standard error (because it is the standard deviation of the sampling distribution).[11] When the sample is relatively small (fewer than 30) the sampling distribution is not normal: it has a different shape, known as a *t*-distribution, which we'll come back to later. A final point is that our discussion here has been about the mean, but everything we have learnt about sampling distributions applies to other parameters too: any parameter that can be estimated in a sample has a hypothetical sampling distribution and standard error.



Cramming Sam's Tips
The standard error

- The standard error of the mean is the standard deviation of sample means. As such, it is a measure of how representative of the population a sample mean is likely to be. A large standard error (relative to the sample mean) means that there is a lot of variability between the means of different samples and so the sample mean we have might not be representative of the population mean. A small standard error indicates that most sample means are similar to the population mean (i.e., our sample mean is likely to accurately reflect the population mean).

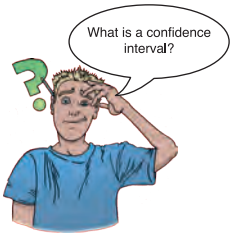


11 In fact, it should be the *population* standard deviation (σ) that is divided by the square root of the sample size; however, it's rare that we know the standard deviation of the population and for large samples this equation is a reasonable approximation.

2.9 I is for (confidence) interval Ⓑ

The 'I' in the SPINE of statistics is for 'interval': confidence interval, to be precise. As a brief recap, we usually use a sample value as an estimate of a parameter (e.g., the mean) in the population. We've just seen that the estimate of a parameter (e.g., the mean) will differ across samples, and we can use the standard error to get some idea of the extent to which these estimates differ across samples. We can also use this information to calculate boundaries within which we believe the population value will fall. Such boundaries are called **confidence intervals**. Although what I'm about to describe applies to any parameter, we'll stick with the mean to keep things consistent with what you have already learnt.

2.9.1 Calculating confidence intervals Ⓑ



Domjan et al. (1998) examined the learnt release of sperm in Japanese quail. The basic idea is that if a quail is allowed to copulate with a female quail in a certain context (an experimental chamber) then this context will serve as a cue to a mating opportunity and this in turn will affect semen release (although during the test phase the poor quail were tricked into copulating with a terry cloth with an embalmed female quail head stuck on top).[12] Anyway, if we look at the mean amount of sperm released in the experimental chamber, there is a true mean (the mean in the population); let's imagine it's 15 million sperm. Now, in our sample, we might find the mean amount of sperm released was 17 million. Because we don't know what the true value of the mean is (the population value), we don't know how good (or bad) our sample value of 17 million is as an estimate of it. So rather than fixating on a single value from the sample (the **point estimate**), we could use an **interval estimate** instead: we use our sample value as the midpoint, but set a lower and upper limit around it. So, we might say, we think the true value of the mean sperm release is somewhere between 12 million and 22 million sperm (note that 17 million falls exactly between these values). Of course, in this case the true value (15 million) does fall within these limits. However, what if we'd set smaller limits – what if we'd said we think the true value falls between 16 and 18 million (again, note that 17 million is in the middle)? In this case the interval does not contain the population value of the mean.

Let's imagine that you were particularly fixated with Japanese quail sperm, and you repeated the experiment 100 times using different samples. Each time you did the experiment you constructed an interval around the sample mean as I've just described. Figure 2.12 shows this scenario: the dots represent the mean for each sample with the lines sticking out of them representing the intervals for these means. The true value of the mean (the mean in the population) is 15 million and is shown by a vertical line. The first thing to note is that the sample means are different from the true mean (this is because of sampling variation as described earlier). Second, although most of the intervals do contain the true mean (they cross the vertical line, meaning that the value of 15 million sperm falls somewhere between the lower and upper boundaries), a few do not.

The crucial thing is to construct the intervals in such a way that they tell us something useful. For example, perhaps we might want to know how often, in the long run, an interval contains the true value of the parameter we're trying to estimate (in this case, the mean). This is what a confidence interval does. Typically, we look at 95% confidence intervals, and sometimes 99% confidence intervals, but they all have a similar interpretation: they are limits constructed such that for a certain percentage of samples (be that 95% or 99%) the true value of the population parameter falls within these limits. So, when you see a 95% confidence interval for a mean, think of it like this: if we'd

12 The male quails didn't seem to mind too much. Read into that what you will.

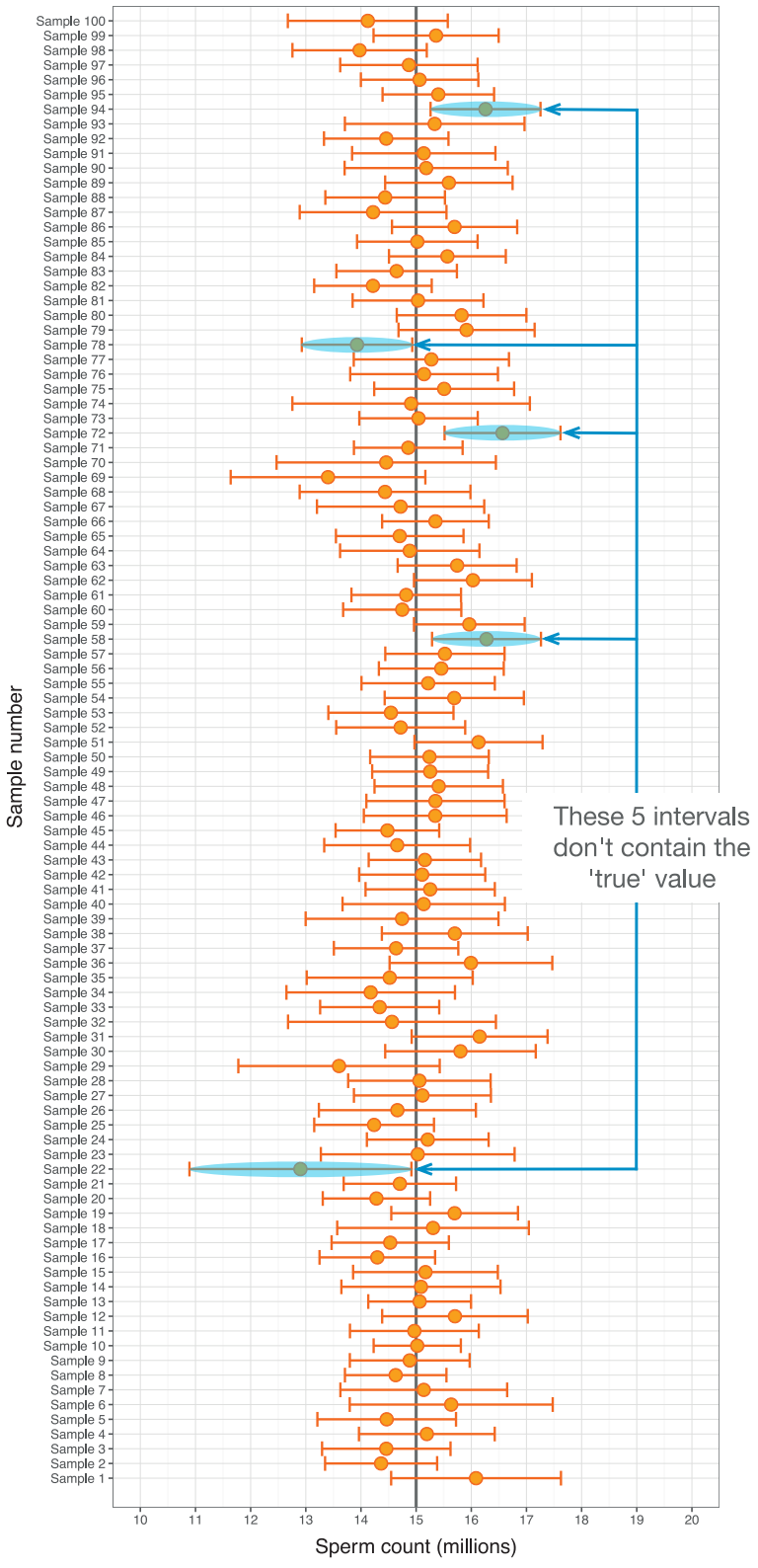


Figure 2.12 The confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 100 different samples (vertical axis)

collected 100 samples, and for each sample calculated the mean and a confidence interval for it (a bit like in Figure 2.12), then for 95 of these samples, the confidence interval contains the value of the mean in the population, but in 5 of the samples the confidence interval does not contain the population mean. The trouble is, you do not know whether the confidence interval from a particular sample is one of the 95% that contain the true value or one of the 5% that do not (Misconception Mutt 2.1).

To calculate the confidence interval, we need to know the limits within which 95% of sample means will fall. We know (in large samples) that the sampling distribution of means will be normal, and the standard normal distribution has been precisely defined such that it has a mean of 0 and a standard deviation of 1. We can use this information to compute the probability of a score occurring, or the limits between which a certain percentage of scores fall (see Section 1.7.6). It was no coincidence that when I explained this in Section 1.7.6 I used the example of how we would work out the limits between which 95% of scores fall; that is precisely what we need to know if we want to construct a 95% confidence interval. We discovered in Section 1.7.6 that 95% of *z*-scores fall between –1.96 and 1.96. This means that if our sample means were normally distributed with a mean of 0 and a standard error of 1, then the limits of our confidence interval would be –1.96 and +1.96. Luckily, we know from the central limit theorem that in large samples (above about 30) the sampling distribution *will* be normally distributed (see Section 6.8.2). It's a pity then that our mean and standard deviation are unlikely to be 0 and 1; except we have seen that we can convert a distribution of scores to have a mean of 0 and standard deviation of 1 using Eq. 1.11:

$$\frac{X - \bar{X}}{s} = z$$

If we know that our limits are –1.96 and 1.96 as *z*-scores, then to find out the corresponding scores in our raw data we can replace *z* in the equation (because there are two values, we get two equations):

$$\frac{X - \bar{X}}{s} = 1.96 \qquad \frac{X - \bar{X}}{s} = -1.96$$

We rearrange these equations to discover the value of X:

$$X - \bar{X} = 1.96 \times s \qquad X - \bar{X} = -1.96 \times s$$

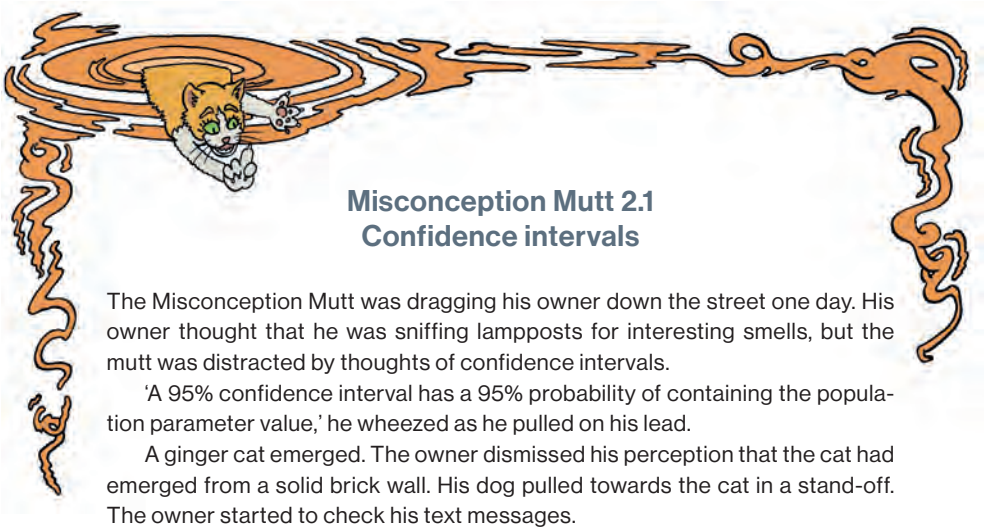
$$X = \bar{X} + (1.96 \times s) \qquad X = \bar{X} - (1.96 \times s)$$

Therefore, the confidence interval can be calculated once the standard deviation (s in the equation) and mean ($\bar{X}$ in the equation) are known. However, we use the standard error and not the standard deviation because we're interested in the variability of *sample* means, not the variability in observations within the sample. The lower boundary of the confidence interval is, therefore, the mean minus 1.96 times the standard error, and the upper boundary is the mean plus 1.96 standard errors:

$$\begin{aligned} &\text{lower boundary of confidence interval} = \bar{X} - (1.96 \times SE) \\ &\text{upper boundary of confidence interval} = \bar{X} + (1.96 \times SE) \end{aligned} \tag{2.23}$$

As such, the point estimate of the parameter (in this case the mean) is always in the centre of the confidence interval. We know that 95% of confidence intervals contain the population mean, so we can assume this confidence interval contains the true mean. Under this assumption (and we should be clear that the assumption might be wrong), if the interval is small, the sample mean must be very close to the true mean. Conversely, if the confidence interval is very wide then the sample mean could be very

different from the true mean, indicating that it is a bad representation of the population. Although we have used the mean to explore confidence intervals, you can construct confidence intervals around any parameter. Confidence intervals will come up time and time again throughout this book.



Misconception Mutt 2.1
Confidence intervals

The Misconception Mutt was dragging his owner down the street one day. His owner thought that he was sniffing lampposts for interesting smells, but the mutt was distracted by thoughts of confidence intervals.

'A 95% confidence interval has a 95% probability of containing the population parameter value,' he wheezed as he pulled on his lead.

A ginger cat emerged. The owner dismissed his perception that the cat had emerged from a solid brick wall. His dog pulled towards the cat in a stand-off. The owner started to check his text messages.

'You again?' the mutt growled.

The cat considered the dog's reins and paced around smugly displaying his freedom. 'I'm afraid you will see very much more of me if you continue to voice your statistical misconceptions,' he said. 'They call me the Correcting Cat for a reason.'

The dog raised his eyebrows, inviting the feline to elaborate.

'You can't make probability statements about confidence intervals,' the cat announced.

'Huh?' said the mutt.

'You said that a 95% confidence interval has a 95% probability of containing the population parameter. It is a common mistake, but this is not true. The 95% reflects *a long-run* probability.'

'Huh?'

The cat raised his eyes to the sky. 'It means that if you take repeated samples and construct confidence intervals, then 95% of them will contain the population value. That is not the same as a particular confidence interval for a specific sample having a 95% probability of containing the value. In fact, for a specific confidence interval the probability that it contains the population value is either 0 (it does not contain it) or 1 (it does contain it). You have no way of knowing which it is.' The cat looked pleased with himself.

'What's the point of that?' the dog asked.

The cat pondered the question. 'It is important if you want to control error,' he eventually answered. 'If you assume that the confidence interval contains the population value then you will be wrong only 5% of the time if you use a 95% confidence interval.'

The dog sensed an opportunity to annoy the cat. 'I'd rather know how likely it is that the interval contains the population value,' he said.

'In which case, you need to become a Bayesian,' the cat said, disappearing indignantly into the brick wall.

The mutt availed himself of the wall, hoping it might seal the cat in for ever.

2.9.2 Calculating other confidence intervals Ⓑ

The example above shows how to compute a 95% confidence interval (the most common type). However, we sometimes want to calculate other types of confidence interval such as a 99% or 90% interval. The 1.96 and –1.96 in Eq. 2.23 are the limits within which 95% of *z*-scores occur. If we wanted to compute confidence intervals for a value other than 95% then we need to look up the value of *z* for the percentage that we want. For example, we saw in Section 1.7.6 that *z*-scores of –2.58 and 2.58 are the boundaries that cut off 99% of scores, so we could use these values to compute 99% confidence intervals. In general, we could say that confidence intervals are calculated as:

$$\begin{aligned}\text{lower boundary of confidence interval} &= \bar{X} - \left(z_{\frac{1-p}{2}} \times SE\right)\\ \text{upper boundary of confidence interval} &= \bar{X} + \left(z_{\frac{1-p}{2}} \times SE\right)\end{aligned} \tag{2.24}$$

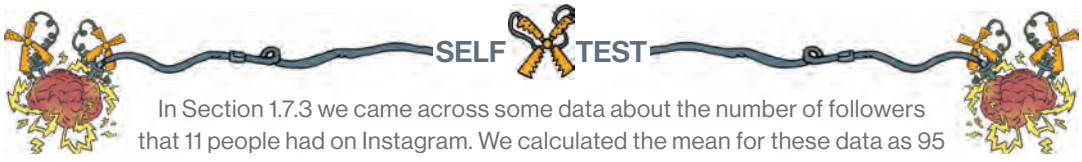
in which p is the probability value for the confidence interval. So, if you want a 95% confidence interval, then you want the value of *z* for (1–0.95)/2 = 0.025. Look this up in the 'smaller portion' column of the table of the standard normal distribution (look back at Figure 1.14) and you'll find that *z* is 1.96. For a 99% confidence interval we want *z* for (1–0.99)/2 = 0.005, which from the table is 2.58 (Figure 1.14). For a 90% confidence interval we want *z* for (1–0.90)/2 = 0.05, which from the table is 1.64 (Figure 1.14). These values of *z* are multiplied by the standard error (as above) to calculate the confidence interval. Using these general principles, we could work out a confidence interval for any level of probability that takes our fancy.

2.9.3 Calculating confidence intervals in small samples Ⓑ

The procedure that I have just described is fine when samples are large, because the central limit theorem tells us that the sampling distribution will be normal. However, for small samples the sampling distribution is not normal – it has a *t*-distribution. The *t*-distribution is a family of probability distributions that change shape as the sample size gets bigger (when the sample is very big, it has the shape of a normal distribution). To construct a confidence interval in a small sample we use the same principle as before but instead of using the value for *z* we use the value for *t*:

$$\begin{aligned}\text{lower boundary of confidence interval} &= \bar{X} - (t_{n-1} \times SE)\\ \text{upper boundary of confidence interval} &= \bar{X} + (t_{n-1} \times SE)\end{aligned} \tag{2.25}$$

The $n - 1$ in the equations is the degrees of freedom (see Jane Superbrain Box 2.5) and tells us which of the *t*-distributions to use. For a 95% confidence interval we find the value of *t* for a two-tailed test with probability of 0.05, for the appropriate degrees of freedom.

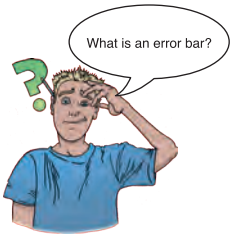


In Section 1.7.3 we came across some data about the number of followers that 11 people had on Instagram. We calculated the mean for these data as 95 and standard deviation as 56.79.

- Calculate a 95% confidence interval for this mean.
- Recalculate the confidence interval assuming that the sample size was 56.

2.9.4 Showing confidence intervals visually Ⓑ

Confidence intervals provide us with information about a parameter, and, therefore, you often see them displayed on plots. (We will discover more about how to create these plots in Chapter 5.) When the parameter is the mean, the confidence interval is usually displayed using something called an error bar, which looks like the letter 'I'. An error bar can represent the standard deviation, or the standard error, but it often shows the 95% confidence interval.



We have seen that any two samples can have slightly different means (or whatever parameter you've estimated) and the standard error tells us a little about how different we can expect sample means to be. We have seen that the 95% confidence interval is an interval constructed such that in 95% of samples the true value of the population mean will fall within its limits. Therefore, under the assumption that our particular confidence interval is one of the 95% capturing the population mean, by comparing confidence intervals of different means (or other parameters) we can get some idea about whether they are likely to have come from the same or different populations. (We can't be entirely sure because we don't know whether our particular confidence intervals are ones that contain the population value or not.)

Taking our previous example of quail sperm, imagine we had a sample of quail and the mean sperm release had been 9 million sperm with a confidence interval of 2 to 16. Therefore, if this is one of the 95% of intervals that contains the population value then the population mean is between 2 and 16 million sperm. What if we now took a second sample of quail and found the confidence interval ranged from 4 to 15? This interval overlaps a lot with our first sample (Figure 2.13). The fact that the confidence intervals overlap in this way tells us that these means could plausibly come from the same population: in both cases, if the intervals contain the true value of the mean (and they are constructed such that in 95% of studies they will), and both intervals overlap considerably, then they contain many similar values. It's very plausible that the population values reflected by these intervals are similar or the same.

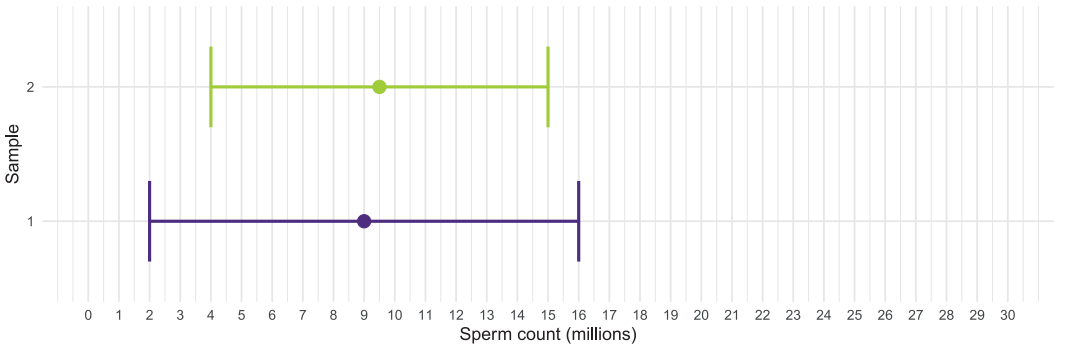


Figure 2.13 Two overlapping 95% confidence intervals

What if the confidence interval for our second sample ranged from 18 to 28? If we compared this to our first sample we'd get Figure 2.14. These confidence intervals don't overlap at all, so one confidence interval, which is likely to contain the population mean, tells us that the population mean is somewhere between 2 and 16 million, whereas the other confidence interval, which is also likely to contain the population mean, tells us that the population mean is somewhere between 18 and 28 million. This contradiction suggests two possibilities: (1) our confidence intervals both contain the population mean, but they come from different populations (and, therefore, so do our samples); or (2) both samples come from the same population but one (or both) of the confidence intervals doesn't

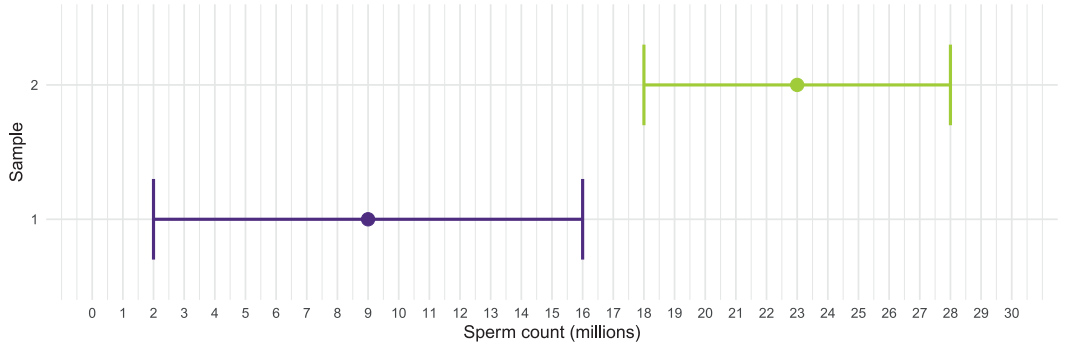


Figure 2.14 Two 95% confidence intervals that don't overlap

contain the population mean (because, as you know, in 5% of cases they don't). If we've used 95% confidence intervals then we know that the second possibility is unlikely (this happens only 5 times in 100 or 5% of the time), so the first explanation is more plausible.

I can hear you all thinking 'so what if the samples come from a different population?' Well, it has a very important implication in experimental research. When we do an experiment, we introduce some form of manipulation between two or more conditions (see Section 1.6.2). If we have taken two random samples of people, and we have tested them on some measure, then we expect these people to belong to the same population. We'd also, therefore, expect their confidence intervals to reflect the same population value for the mean. If their sample means and confidence intervals are so different



Cramming Sam's Tips
Confidence intervals

- A confidence interval for the mean (or any parameter we've estimated) is a range of scores constructed such that the population value falls within this range in 95% of samples.
- The confidence interval is *not* an interval within which we are 95% confident that the population value will fall.
- There is *not* a 95% probability of the population value falling within the limits of a 95% confidence interval. The probability is either 1 (the population value falls in the interval) or 0 (the population value does not fall in the interval), but we don't know which of these probabilities is true for a particular confidence interval.



as to suggest that they come from different populations, then this is likely to be because our experimental manipulation has induced a difference between the samples. Therefore, error bars showing 95% confidence intervals are useful, because if the bars of any two means do not overlap (or overlap by only a small amount) then we can infer that these means are from different populations – they are significantly different. We will return to this point in Section 2.10.10.

2.10 N is for null hypothesis significance testing Ⓐ

In Chapter 1 we saw that research was a six-stage process (Figure 1.2). This chapter has looked at the final stage:

- Analyse the data: fit a statistical model to the data – this model will test your original predictions. Assess this model to see whether it supports your initial predictions

I have shown that we can use a sample of data to estimate what's happening in a larger population to which we don't have access. We have also seen (using the mean as an example) that we can fit a statistical model to a sample of data and assess how well it fits. However, we have yet to see how fitting models like these can help us to test our research predictions. How do statistical models help us to test complex hypotheses such as 'is there a relationship between the amount of gibberish that people speak and the amount of vodka jelly they've eaten?', or 'does reading this chapter improve your knowledge of research methods?' This brings us to the 'N' in the SPINE of statistics: null hypothesis significance testing.

Null hypothesis significance testing (NHST) is a cumbersome name for an equally cumbersome process. NHST is the most commonly taught approach to testing research questions with statistical models. It arose out of two different approaches to the problem of how to use data to test theories: (1) Ronald Fisher's idea of computing probabilities to evaluate evidence and (2) Jerzy Neyman and Egon Pearson's idea of competing hypotheses.

2.10.1 Fisher's *p*-value Ⓐ



Fisher (1925/1991) (Figure 2.15) described an experiment designed to test a claim by a woman that she could determine, by tasting a cup of tea, whether the milk or the tea was added first to the cup. Fisher thought that he should give the woman some cups of tea, some of which had the milk added first and some of which had the milk added last, and see whether she could distinguish them correctly. The woman would know that there are an equal number of cups in which milk was added first or last but wouldn't know in which order the cups were placed. If we take the simplest situation in which there are only two cups, then the woman has a 50% chance of guessing correctly. If she did guess correctly we wouldn't be particularly convinced that she can tell the difference between cups in which the milk was added first and those in which it was added last, because even by guessing she would be correct half of the time. But what if we complicated things by having six cups? There are 20 orders in which these cups can be arranged and the woman would guess the correct order only 1 time in 20 (or 5% of the time). If she got the order correct we would be much more convinced that she could genuinely tell the difference (and bow down in awe of her finely tuned palate). If you'd like to know more about Fisher and his tea-tasting antics see David Salsburg's excellent book *The Lady Tasting Tea* (Salsburg, 2002). For our purposes the take-home point is that only when there was a very small probability that the woman could complete the tea task by guessing alone would we conclude that she had genuine skill in detecting whether milk was poured into a cup before or after the tea.

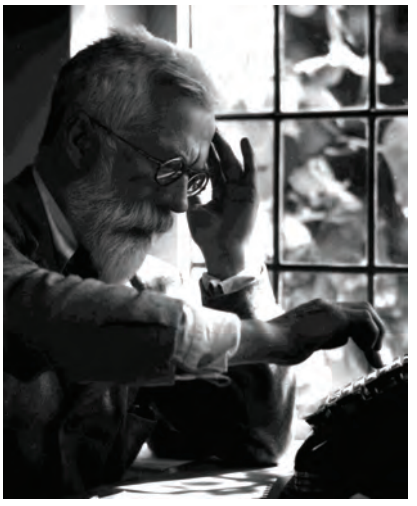


Figure 2.15 Sir Ronald A. Fisher

It's no coincidence that I chose the example of six cups above (where the tea-taster had a 5% chance of getting the task right by guessing), because scientists tend to use 5% as a threshold for their beliefs: only when there is a 5% chance (or 0.05 probability) of getting the result we have (or one more extreme) if no effect exists are we convinced that the effect is genuine.[13] Fisher's basic point was that you should calculate the probability of an event and evaluate this probability within the research context. Although Fisher felt that $p = 0.01$ would be strong evidence to back up a hypothesis, and perhaps $p = 0.20$ would be weak evidence, he never said $p = 0.05$ was in any way a magic number. Fast-forward 100 years or so, and everyone treats 0.05 as though it *is* a magic number.

2.10.2 Types of hypothesis Ⓐ

In contrast to Fisher, Neyman and Pearson believed that scientific statements should be split into testable hypotheses. The hypothesis or prediction from your theory would normally be that an effect will be present. This hypothesis is called the **alternative hypothesis** and is denoted by H_1. (It is sometimes also called the **experimental hypothesis**, but because this term relates to a specific type of methodology it's probably best to use 'alternative hypothesis'.) There is another type of hypothesis called the **null hypothesis**, which is denoted by H_0. This hypothesis is the opposite of the alternative hypothesis and so usually states that an effect is absent.

When I write my thoughts are often drawn towards chocolate. I believe that I would eat less of it if I could stop thinking about it. However, according to Morewedge et al. (2010), that's not true. In fact, they found that people ate less of a food if they had previously imagined eating it. Imagine we did a similar study. We might generate the following hypotheses:

- Alternative hypothesis: if you imagine eating chocolate you will eat less of it.
- Null hypothesis: if you imagine eating chocolate you will eat the same amount as normal.

The null hypothesis is useful because it gives us a baseline against which to evaluate how plausible our alternative hypothesis is. We can evaluate whether we think the data we have collected are more likely, given the null or alternative hypothesis. A lot of books talk about accepting or rejecting these hypotheses, implying that you look at the data and either accept the null hypothesis (and therefore reject the alternative) or accept the alternative hypothesis (and reject the null). In fact, this isn't quite right because the way that scientists typically evaluate these hypotheses using *p*-values (which we'll come onto shortly) doesn't provide evidence for such black-and-white decisions. So, rather than talking about accepting or rejecting a hypothesis we should talk about 'the chances of obtaining the result we have (or one more extreme) assuming that the null hypothesis is true'.

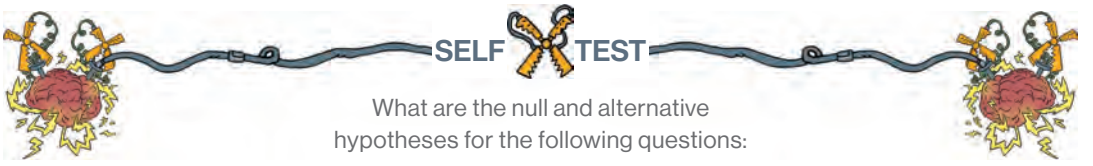
Imagine in our study that we took 100 people and measured how many pieces of chocolate they usually eat (day 1). On day 2, we got them to imagine eating chocolate and again measured how much chocolate they ate that day. Imagine that we found that 75% of people ate less chocolate on the second day than the first. When we analyse our data, we are really asking, 'Assuming that imagining eating chocolate has no effect whatsoever, is it likely that 75% of people would eat less chocolate on the second day?' Intuitively the answer is that the chances are very low: if the null hypothesis is true, then everyone should eat the same amount of chocolate on both days. Therefore, we are very unlikely to have got the data that we did if the null hypothesis were true.

13 Of course it might not be true – we're just prepared to believe that it is.

What if we found that only 1 person (1%) ate less chocolate on the second day? If the null hypothesis is true and imaging eating chocolate has no effect whatsoever on consumption, then no people should eat less on the second day. The chances of getting these data if the null hypothesis is true are quite high. The null hypothesis is quite plausible given what we have observed.

When we collect data to test theories we work in these terms: we cannot talk about the null hypothesis being true or the experimental hypothesis being true, we can talk only in terms of the probability of obtaining a particular result or statistic if, hypothetically speaking, the null hypothesis were true. It's also worth remembering that our alternative hypothesis is likely to be one of many possible models that we could fit to the data, so even if we believe it to be more likely than the null hypothesis, there may be other models of the data that we haven't considered that are a better fit, which again means that we cannot talk about the hypothesis as being definitively true or false, but we can talk about its plausibility relative to other hypotheses or models that we have considered.

Hypotheses can be directional or non-directional. A directional hypothesis states that an effect will occur, but it also states the direction of the effect. For example, 'if you imagine eating chocolate you will eat less of it' is a one-tailed hypothesis because it states the direction of the effect (people will eat less). A non-directional hypothesis states that an effect will occur, but it doesn't state the direction of the effect. For example, 'imagining eating chocolate affects the amount of chocolate you eat' does not tell us whether people will eat more, or less.



What are the null and alternative hypotheses for the following questions:

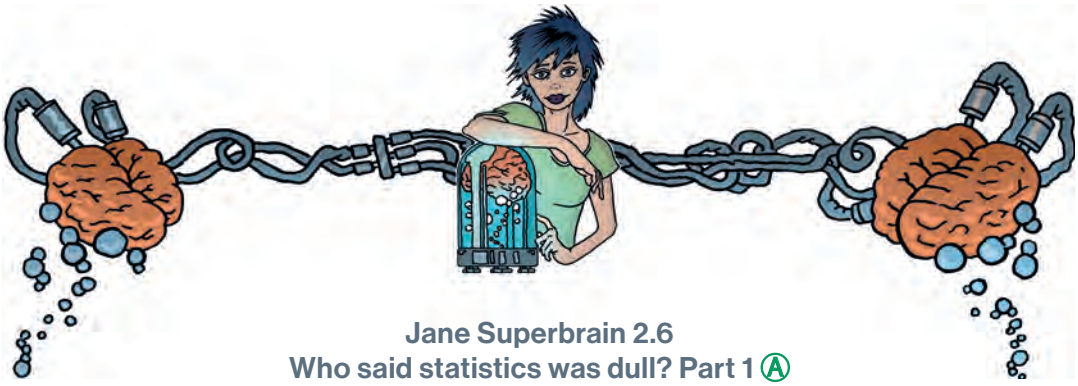
- 'Is there a relationship between the amount of gibberish that people speak and the amount of vodka jelly they've eaten?'
- 'Does reading this chapter improve your knowledge of research methods?'

2.10.3 The process of NHST Ⓐ

NHST is a blend of Fisher's idea of using the probability value p as an index of the weight of evidence against a null hypothesis, and Jerzy Neyman and Egon Pearson's idea of testing a null hypothesis against an alternative hypothesis (Neyman & Pearson, 1933). There was no love lost between these competing statisticians (Jane Superbrain Box 2.6). NHST is a system designed to tell us whether the alternative hypothesis is likely to be true – it helps us to decide whether to confirm or reject our predictions.

Figure 2.16 outlines the steps in NHST. As we have seen before, the process starts with a research hypothesis that generates a testable prediction. These predictions are decomposed into a null (there is no effect) and alternative hypothesis (there is an effect). At this point you decide upon the long-run error rate that you are prepared to accept, alpha (α). In other words, how often are you prepared to be wrong? This is the significance level, the probability of accepting an effect in our population as true, when no such effect exists (it is known as the Type I error rate, which we'll discuss in due course). It is important that we fix this error rate before we collect data, otherwise we are cheating (see Jane Superbrain Box 2.7). You should determine your error rate based on the nuances of your research area, and what it is you're trying to test (Maier & Lakens, 2022). Put another way, it should be a meaningful decision. In reality, it is not: everyone uses 0.05 (a 5% error rate) with barely a thought for what it means or why they're using it. Go figure.

Having not given the thought you should have to your error rate, you choose a sampling distribution. This involves working out what statistical model to fit to the data that will test your hypothesis,



Jane Superbrain 2.6
Who said statistics was dull? Part 1 Ⓐ

Students often think that statistics is dull, but back in the early 1900s it was anything but dull, with prominent figures entering into feuds on a regular basis. Ronald Fisher and Jerzy Neyman had a particularly impressive feud. On 28 March 1935 Neyman delivered a talk to the Royal Statistical Society, at which Fisher was present, in which he criticized some of Fisher's most important work. Fisher directly attacked Neyman in his discussion of the paper at the same meeting: he more or less said that Neyman didn't know what he was talking about and didn't understand the background material on which his work was based. He said, 'I put it to you, sir, that you are a fool, a man so incapacitated by stupidity that in a battle of wits with a single-cell amoeba, the amoeba would fancy its chances.' He didn't really say that, but he opened the discussion without proposing a vote of thanks which would have been equally as rude in those days.

Relations soured so much that while they both worked at University College London, Neyman openly attacked many of Fisher's ideas in lectures to his students. The two feuding groups even took afternoon tea (a common practice in the British academic community of the time and one, which, frankly, we should reinstate) in the same room but at different times. The truth behind who fuelled these feuds is, perhaps, lost in the mists of time, but Zabell (1992) makes a sterling effort to unearth it. Basically, the founders of modern statistical methods, despite being super-humanly intelligent,[14] acted like a bunch of squabbling children.



looking at what parameters that model has, and then deciding on the shape of the sampling distribution attached to those parameters. Let's take my example of whether thinking about chocolate is related to consumption. You could measure how much people think about chocolate during the day and how much of it they eat in the same day. If the null hypothesis is true (there is no effect) then there should be no relationship at all between these variables. If thinking about chocolate reduces consumption then we'd expect a negative relationship between the two. One model we could fit that tests this hypothesis is the linear model that I described earlier in which we predict consumption (the outcome) from thought about chocolate (the predictor). Our model is basically Eq. 2.6, but I'll replace the outcome and letter X with our variable names:

$$\text{consumption}_i = b_0 + b\text{Thought}_i + \varepsilon_i \tag{2.26}$$

14 Fisher, in particular, was a world leader in genetics, biology and medicine as well as possibly the most original mathematical thinker ever (Barnard, 1963; Field, 2005; Savage, 1976). Unfortunately, he was also a big fan of eugenics, which goes to show that intelligence is not mutually exclusive from stupidity.

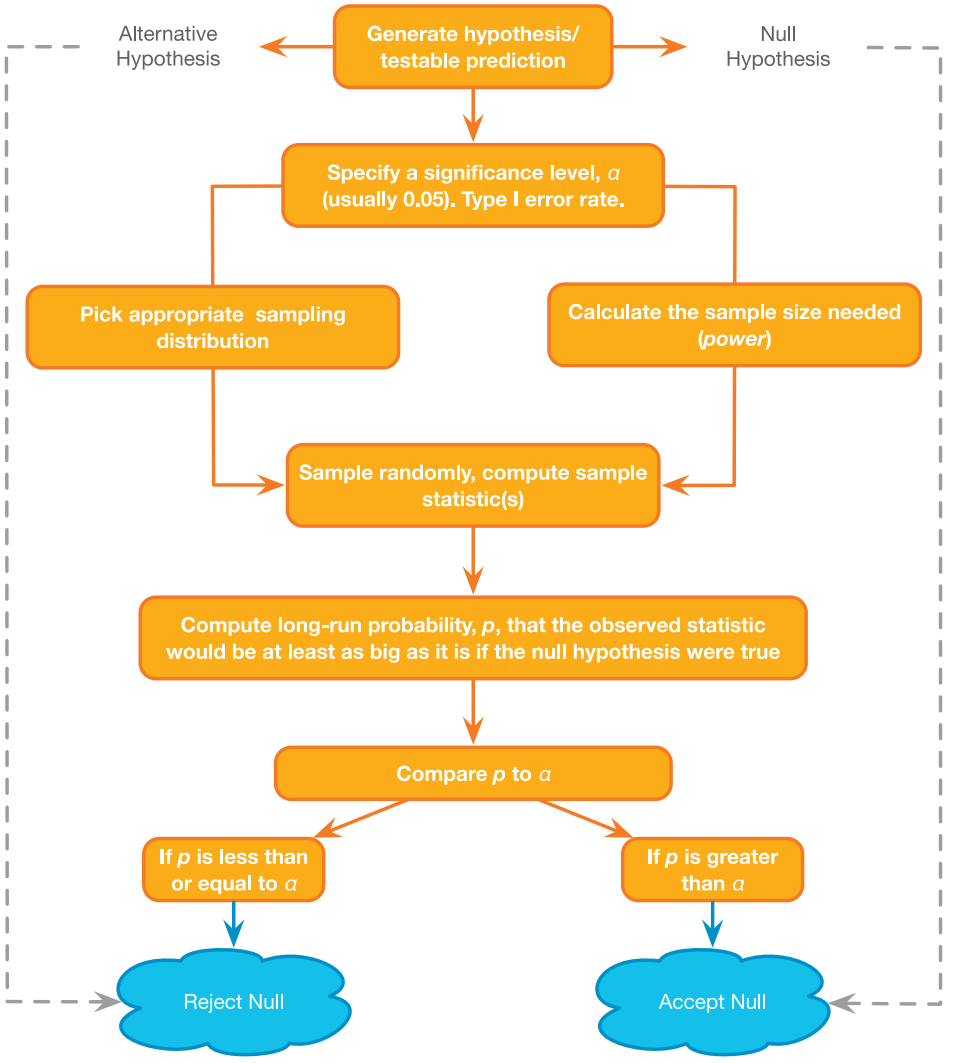
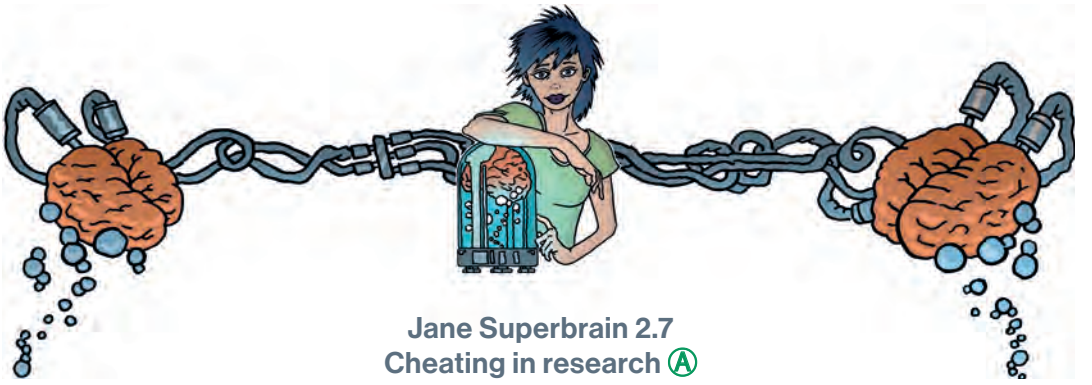


Figure 2.16 Flow chart of null hypothesis significance testing

The parameter, b, attached to the variable *thought* tests our hypothesis: it quantifies the size and strength of relationship between thinking and consuming. If the null is true, b will be zero, otherwise it will be a value different from 0, the size and direction of which depends on what the relationship between the thought and consumption variables is. It turns out (see Chapter 6) that this parameter has a sampling distribution that has a t-distribution. So, that's what we'd use to test our hypothesis. We also need to establish how much data to collect to stand a reasonable chance of finding the effect we're looking for. This is called the power of the test, and I'll elaborate on this concept shortly.

Now the fun begins and you collect data. You fit the statistical model that tests your hypothesis to the data. In the chocolate example, we'd estimate the parameter that represents the relationship between thought and consumption and its confidence interval. It's usually also possible to compute a test statistic that maps the parameter to a long-run probability value (the p-value). In our chocolate example, we can compute a statistic known as t, which has a specific sampling distribution from which we can get a probability value (p). This probability value tells us how likely we would be to get a value of t at least as big as the one we have if the null hypothesis is true. As I keep mentioning, this p is a long-run probability: it is computed by working out how often you'd get specific values of the test statistic (in this case t) if you repeated your exact sampling process an infinite number of times.

It is important that you collect the amount of data that you set out to collect, otherwise the *p*-value you obtain will not be correct. It is possible to compute a *p*-value representing the long-run probability of getting a *t*-value at least as big as the one you have in repeated samples of, say, 80, but there is no way of knowing the probability of getting a *t*-value at least as big as the one you have in repeated samples of 74, where the intention was to collect 80 but term ended and you couldn't find any more participants. If you cut data collection short (or extend it) for this sort of arbitrary reason, then whatever *p*-value you end up with is certainly not the one you want. Again, it's cheating: you're changing your team after you have placed your bet, and you will likely end up with research egg on your face when no one can replicate your findings.



Jane Superbrain 2.7
Cheating in research Ⓐ

NHST works only if you generate your hypotheses and decide on your criteria for whether an effect is significant before collecting the data. Imagine I wanted to place a bet on who would win the soccer World Cup. Being English, I might bet on England to win the tournament. To do this I'd: (1) place my bet, choosing my team (England) and odds available at the betting shop (e.g., 6/4); (2) see which team wins the tournament; (3) collect my winnings (or, having bet on England, more likely not).

To keep everyone happy, this process needs to be equitable: the betting shops set their odds such that they're not paying out too much money (which keeps them happy), but so that they do pay out sometimes (to keep the customers happy). The betting shop can offer any odds before the tournament has ended, but it can't change them once the tournament is over (or the last game has started). Similarly, I can choose any team before the tournament, but I can't then change my mind halfway through, or after the final game.

The research process is similar: we can choose any hypothesis (soccer team) before the data are collected, but we can't change our minds halfway through data collection (or after data collection). Likewise we have to decide on our probability level (or betting odds) before we collect data. *If* we do this, the process works. However, researchers sometimes cheat. They don't formulate hypotheses before they conduct their experiments, they change them when the data are collected (like me changing my team after the World Cup is over), or worse still they decide on them after the data are collected (see Chapter 3). With the exception of procedures such as *post hoc* tests, this is cheating. Similarly, researchers can be guilty of choosing which significance level to use after the data are collected and analysed, like a betting shop changing the odds after the tournament.

If you change your hypothesis or the details of your analysis you increase the chance of finding a significant result, but you also make it more likely that you will publish results that other researchers can't reproduce (which is embarrassing). If, however, you follow the rules carefully and do your significance testing at the 5% level you at least know that in the long run at most only 1 result out of every 20 will risk this public humiliation. (Thanks to David Hitchin for this box, and apologies to him for introducing soccer into it.)



Having hopefully stuck to your original sampling frame and obtained the appropriate p-value, you compare it to your original alpha value (usually 0.05). If the p you obtain is less than or equal to the original α, scientists typically use this as grounds to reject the null hypothesis outright; if the p is greater than α then they accept that the null hypothesis is plausibly true (they reject the alternative hypothesis). We can never be completely sure that either hypothesis is correct; all we can do is to calculate the probability that our model would fit at least as well as it does if there were no effect in the population (i.e., the null hypothesis is true). As this probability decreases, we gain greater confidence that the alternative hypothesis is more plausible than the null hypothesis. This overview of NHST is a lot to take in, so we will revisit a lot of the key concepts in detail the next few sections.

2.10.4 A dating example ©

To make NHST a little more concrete, let's look at an example that illustrates some parts of the process:

1. Breaking a hypothesis into null and alternative hypotheses
2. Selecting a sampling distribution that represents what happens when the null is true
3. Sampling some data and establishing a 'test statistic'
4. Using the sampling distribution of the null to establish the probability of obtaining a test statistic at least as large as the one obtained

In reality, things are a little more complicated than this example. The model here is simpler than most of the models you will use, and I'm not going to get into power and alpha levels. There's some maths in this section, but you don't need to understand it. Provided you can follow the logic of the example then you can safely accept the maths.

Imagine you have just put your profile onto a dating app and you are interested in how popular it has been relative to the norm. We can split this interest into null and alternative hypotheses:

- Null hypothesis: your profile has been as popular as the norm
- Alternative hypothesis: because you're a confident type of person you hypothesize that your profile will be more popular than the norm.

On this dating app, when someone sees your profile they either swipe it away (i.e., reject you) or click a heart icon (i.e., show that they are interested in you as a potential date). In other words, there are two possible outcomes: swipe (0) or like (1). We can assume that the actions of one person who sees your profile don't influence the actions of anyone else, so these events are independent. To operationalize your hypotheses we decide to use the number of people who like your profile as an outcome variable.

Before we collect data, we fix the sample size N. In this example, we will sample the first 10 people to view your profile and observe how many of them swipe it or like it (in real research we'd use a sample much larger than this). The sample size is, therefore, fixed at $N = 10$. If we treat the profile being liked as a success, then we're operationalizing the hypotheses by asking 'how many successes occur in every 10 trials?' It turns out we can model this situation using something called a binomial distribution, which is described by two parameters: the number of trials n, and the probability p of a success. For argument's sake, let's assume that we know from historic data across all users of the app that this probability is 0.4, that is, in general for every 10 people that look at a profile, 4 people will click on the heart. The binomial distribution with $n = 10$ and $p = 0.4$ is shown in Figure 2.17. This is the distribution under the null. Remember that our null was that our profile was just as popular as the norm. The distribution in Figure 2.17 shows the 'norm'; that is, it shows the probability with which a profile is liked by each of 0, 1, 2, 3, …, 10 people for every 10 people who view it. This is the sampling distribution mentioned in Figure 2.16. It represents the plausibility of different outcomes under the null.

Having generated hypotheses, fixed our sample size and selected an appropriate sampling distribution to represent the null, we collect some data. We post our profile to the app and we count how many of the first 10 people to view it click the heart (i.e. like it). Eight people click the heart. Our 'test statistic' is, therefore, $x = 8$. The question is, therefore: how probable is this observation under the null hypothesis?

We can use Figure 2.17 to answer this question. Assuming a sample size of 10 (that is, 10 people viewing a profile) the distribution shows the probability that each of 0, 1, 2, 3, …, 10 people who view a profile like it. The red bars represent 8 or more people (out of 10) liking a profile under the null, and the blue bars represent 7 or fewer people (out of 10) liking a profile. Intuitively, we can see that the probability that 8 or more people will like a profile is low (the area occupied by the red bars is much smaller than that of the blue bars). The probabilities on the plot can be calculated as follows:[15]

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0,1,\ldots,n$$
$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{2.27}$$

Note that we will substitute actual numbers for values of n and p and that ! is the symbol for the factorial function (so $x!$ means that you multiply all the whole numbers up to and including x).

The probability that 8 out of 10 people will like a profile is calculated as 0.0106:

$$\begin{aligned} P(8) &= \binom{10}{8} 0.4^8 (1-0.4)^{10-8}, \quad x = 0,1,\ldots,n \\ &= \frac{10!}{8!(10-8)!} 0.4^8 0.6^2 \\ &= \frac{10\times 9\times \cancel{8\times 7\times 6\times 5\times 4\times 3\times 2\times 1}}{\cancel{8\times 7\times 6\times 5\times 4\times 3\times 2\times 1}\times(2\times 1)} 0.4^8 0.6^2 \\ &= \frac{10\times 9}{2} 0.00065536 \times 0.36 \\ &= 0.0106 \end{aligned} \tag{2.28}$$

This is the corresponding value on the y-axis on Figure 2.17. (Notice with the expression for the factorial terms in Eq. 2.28 that you can automatically cancel out a lot of terms from the top and bottom, so in future equations I'll skip the step of writing out these terms in full.)

However, we don't want to know the probability of exactly 8 people liking a profile; we want to know the probability of 8 *or more* people liking it. Therefore, we also need to work out the probabilities that a profile is liked by 9 or 10 people. These values are calculated as follows:

$$\begin{aligned} P(9) &= \binom{10}{9} 0.4^9 (1-0.4)^{10-9}, \quad x = 0,1,\ldots,n \\ &= \frac{10!}{9!(10-9)!} 0.4^9 0.6^1 \\ &= \frac{10}{1} 0.000262144 \times 0.6 \\ &= 0.0016 \end{aligned} \tag{2.29}$$

15 An explanation of this equation is beyond the scope of this book, but you'll find an explanation in numerous more technical books.

$$
\begin{aligned}
P(10) &= \binom{10}{10} 0.4^{10}(1-0.4)^{10-10}, \quad x = 0,1,\ldots,n \\
&= \frac{10!}{10!0!} 0.4^{10} 0.6^{0} \\
&= 1 \times 0.0001048576 \times 1 \\
&= 0.0001
\end{aligned}
\tag{2.30}
$$

Notice that as the number of people liking a profile (x) increases, the probabilities decrease (which is also seen in Figure 2.17 by the bars getting shorter). To get the probability of at least 8 of the 10 people liking a profile we add these probabilities. It turns out that the probability of at least 8 of 10 people liking a profile is $p = 0.0123$:

$$
\begin{aligned}
p(x \geq 8) &= p(8) + p(9) + p(10) \\
&= 0.0106 + 0.0016 + 0.0001 \\
&= 0.0123
\end{aligned}
\tag{2.31}
$$

This probability is small. In other words, under the null, it is highly unlikely that 8 or more out of 10 people would like a profile on this dating app. Given that you have observed a value of 8 for your own profile, and that this is an unlikely event under the null, you might decide that your observed value is sufficiently unlikely and conclude that your profile is rated differently to the norm. This example hopefully makes some of the ideas of NHST a little more concrete.

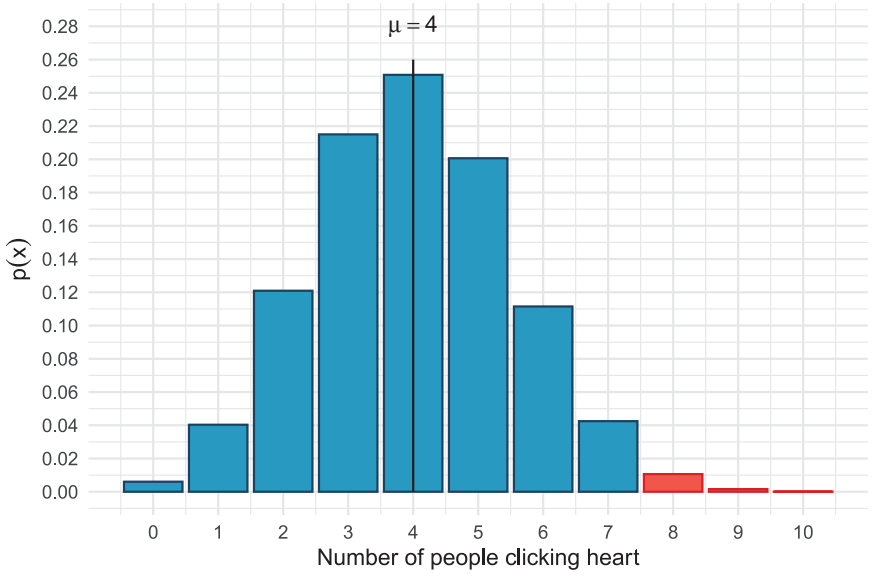


Figure 2.17 Binomial distribution with $n = 10$ and $p = 0.4$

2.10.5 Test statistics Ⓐ

I mentioned that NHST relies on fitting a model to the data and then evaluating the probability of this model given the assumption that no effect exists. I mentioned in passing that the fit of a model, or the parameters within it, are typically mapped to a probability value through a test statistic. I was deliberately vague about what the 'test statistic' is, so let's lift the veil of secrecy. To do this we need to return to the concepts of systematic and unsystematic variation that we encountered in Section 1.6.4.

Systematic variation is variation that can be explained by the model that we've fitted to the data (and, therefore, due to the hypothesis that we're testing). Unsystematic variation is variation that cannot be explained by the model that we've fitted. In other words, it is error, or variation not attributable to the effect we're investigating. The simplest way, therefore, to test whether the model fits the data, or whether our hypothesis is a good explanation of the data we have observed, is to compare the systematic variation against the unsystematic variation (see Chapters 6 and 10 in particular for a more detailed explanation). In effect, we look at a signal-to-noise ratio: we compare how good the model/hypothesis is against how bad it is (the error):

$$\text{Test Statistic} = \frac{\text{signal}}{\text{noise}} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}} \tag{2.32}$$

Likewise, the best way to test a parameter is to look at the size of the parameter relative to the background noise (the sampling variation) that produced it. Again, it's a signal-to-noise ratio: the ratio of how big a parameter is to how much it can vary across samples:

$$\text{Test Statistic} = \frac{\text{signal}}{\text{noise}} = \frac{\text{size of parameter}}{\text{sampling variation in the parameter}} = \frac{\text{effect}}{\text{error}} \tag{2.33}$$

The ratio of effect relative to error is a **test statistic**, and you'll discover later in the book there are lots of them: t, F and χ^2, to name three. Each test statistic has its own equation, but they are all, crudely speaking, the ratio of signal to noise. The reason why this ratio is so useful is intuitive really: if our model is good then we'd expect it to be able to explain more variance than it can't explain. In this case, the test statistic will be greater than 1 (but not necessarily significant). Similarly, larger test statistics will be produced by larger parameters (bigger effects) that have smaller sampling variation (i.e. are more likely to represent the population).

A test statistic is a statistic for which we know how frequently different values occur. I mentioned the t-distribution, chi-square (χ^2) distribution and F-distribution in Section 1.7.6 and said that they are all defined by an equation (the PDF) that enables us to calculate the probability of obtaining a given score. Therefore, if a test statistic comes from one of these distributions we can calculate the probability of obtaining a certain value (just as we could estimate the probability of getting a score of a certain size from a frequency distribution in Section 1.7.6). This probability is the p-value that Fisher described and in NHST it is used to estimate how likely (in the long run) it is that we would get a test statistic at least as big as the one we have *if there were no effect* (i.e., the null hypothesis were true).

Test statistics can be a bit scary, so let's imagine that they're cute kittens. Kittens are typically very small (about 100g at birth on average), but every so often a cat will give birth to a big one (say, 150g). A 150g kitten is rare, so the probability of finding one is very small. Conversely, 100g kittens are very common, so the probability of finding one is quite high. Test statistics are the same as kittens in this respect: small ones are quite common and large ones are rare. So, if we do some research (i.e., give birth to a kitten) and calculate a test statistic (weigh the kitten) we can calculate the probability of obtaining a value/weight at least that large. The more variation our model explains compared to the variance it can't explain, the bigger the test statistic will be (i.e., the more the kitten weighs), and the more unlikely it is to occur if there were no effect (like our 150 g kitten). Like kittens, as test statistics get bigger the probability of them occurring becomes smaller. If this probability falls below a certain value ($p < 0.05$ if we blindly apply the conventional 5% error rate), we presume that the test statistic is as large as it is because our model explains a sufficient amount of variation to reflect a genuine effect in the real world (the population). The test statistic is said to be *statistically significant*. Given that the statistical model that we fit to the data reflects the hypothesis that we set out to test, then a significant test statistic tells us that the model would be unlikely to fit this well if the there was no effect in the population (i.e., the null hypothesis was true). Typically, this is taken as a reason to reject the null hypothesis and gain confidence that the alternative hypothesis is true. If, however, the probability of

obtaining a test statistic at least as big as the one we have (if the null hypothesis were true) is too large (typically $p > 0.05$) then the test statistic is said to be non-significant and is used as grounds to reject the alternative hypothesis (see Section 3.2.1 for a discussion of what statistically significant does and does not mean).

2.10.6 One- and two-tailed tests Ⓑ

We saw in Section 2.10.2 that hypotheses can be directional (e.g., 'the more someone reads this book, the more they hate its author') or non-directional (i.e., 'reading more of this book could increase or decrease the reader's hate of its author'). A statistical model that tests a directional hypothesis is called a **one-tailed test**, whereas one testing a non-directional hypothesis is known as a **two-tailed test**.



Imagine we wanted to discover whether reading this book increased or decreased hatred of me. If we have no directional hypothesis then there are three possibilities. (1) People who read this book hate me more than those who don't, so the difference (the mean for those reading the book minus the mean for non-readers) is positive. Put another way, as the amount of time spent reading this book increases, so does the hatred of me – a positive relationship. (2) People who read this book hate me less than those who don't, so the difference (the mean for those reading the book minus the mean for non-readers) is negative. Alternatively, as the amount of time spent reading this book increases, the hatred of me decreases – a negative relationship. (3) There is no difference between readers and non-readers in their hatred of me – the mean for readers minus the mean for non-readers is exactly zero. There is no relationship between reading this book and hating me. This final option is the null hypothesis. The direction of the test statistic (i.e., whether it is positive or negative) depends on whether the difference, or direction of relationship, is positive or negative. Assuming there is a positive difference or relationship (the more you read, the more you hate me), then to detect this difference we take account of the fact that the mean for readers is bigger than for non-readers (and so derive a positive test statistic). However, if we've predicted incorrectly and reading this book makes readers hate me *less*, then the test statistic will be negative instead.

What are the consequences of this? Well, if at the 0.05 level we needed to get a test statistic bigger than, say, 10 and the one we got was actually –12, then we would reject the hypothesis even though a difference does exist. To avoid this, we can look at both ends (or tails) of the distribution of possible test statistics. This means we will catch both positive and negative test statistics. However, doing this has a price because to keep our criterion probability of 0.05 we split this probability across the two tails: we have 0.025 at the positive end of the distribution and 0.025 at the negative end. Figure 2.18 shows this situation – the orange tinted areas are the areas above the test statistic needed at a 0.025 level of significance. Combine the probabilities (i.e., add the two tinted areas together) at both ends and we get 0.05, our criterion value.

If we have made a prediction, then we put all our eggs in one basket and look only at one end of the distribution (either the positive or the negative end, depending on the direction of the prediction we make). In Figure 2.18, rather than having two small orange tinted areas at either end of the distribution that show the significant values, we have a bigger area (the blue tinted area) at only one end of the distribution that shows significant values. Note that this blue area contains within it one of the orange areas as well as an extra bit of blue area. Consequently, we can just look for the value of the test statistic that would occur if the null hypothesis were true with a probability of 0.05. In Figure 2.18, the blue tinted area is the area above the positive test statistic needed at a 0.05 level of significance (1.64); this value is smaller than the value that begins the area for the 0.025 level of significance (1.96). This means that if we make a specific prediction then we need a smaller test statistic to find a significant result (because we are looking in only one tail of the distribution), but if our prediction happens to be in the wrong direction then we won't detect the effect that does exist. This final point is very important, so let me rephrase it: if you do a one-tailed test and the results turn out to be in the opposite

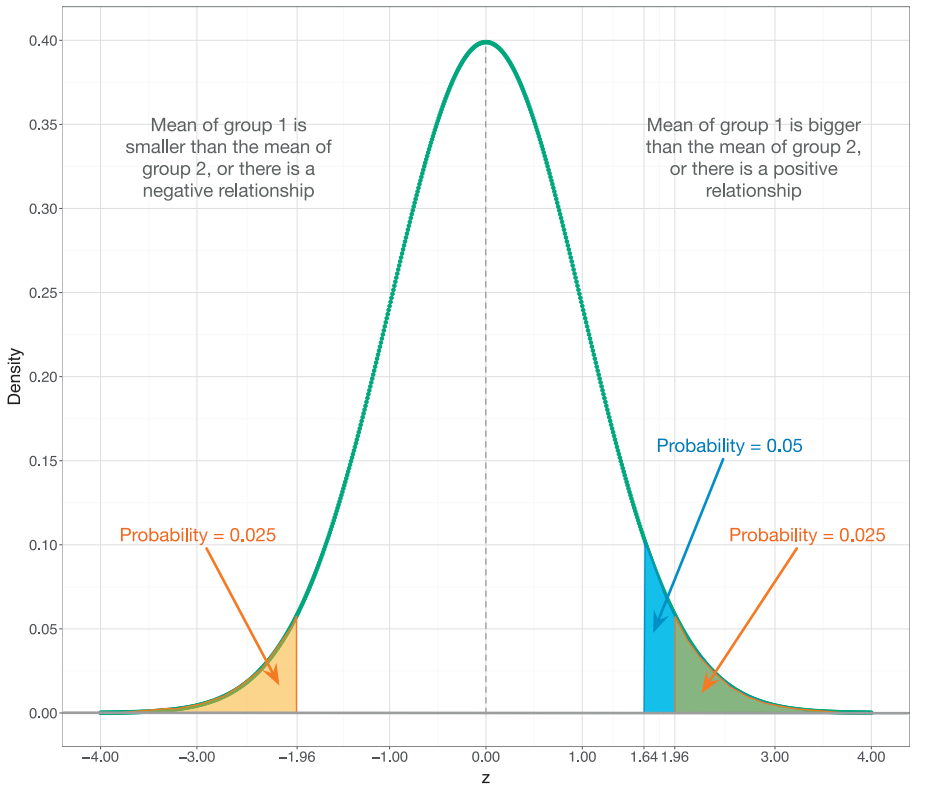


Figure 2.18 Diagram to show the difference between one- and two-tailed tests

direction to what you predicted you must ignore them, resist all temptation to interpret them, and accept (no matter how much it pains you) the null hypothesis. If you *don't* do this, then you have done a two-tailed test using a different level of significance from the one you set out to use (and Jane Superbrain Box 2.7 explains why that is a bad idea).

I have explained one- and two-tailed tests because people expect to find them explained in statistics textbooks. However, there are a few reasons why you should think long and hard about whether one-tailed tests are a good idea. Wainer (1972) quotes John Tukey (one of the great modern statisticians) as responding to the question 'Do you mean to say that one should *never* do a one-tailed test?' by saying 'Not at all. It depends upon to whom you are speaking. *Some people will believe anything*' (emphasis added). Why might Tukey have been so sceptical?

As I have said already, if the result of a one-tailed test is in the opposite direction to what you expected, *you cannot and must not reject the null hypothesis*. In other words, you must completely ignore that result even though it is poking you in the arm and saying 'look at me, I'm intriguing and unexpected'. The reality is that when scientists see interesting and unexpected findings their instinct is to want to explain them. Therefore, one-tailed tests are like a mermaid luring a lonely sailor to his death by being beguiling and interesting: they lure lonely scientists to their academic death by throwing up irresistible and unpredicted results.

One context in which a one-tailed test *could* be used, then, is if a result in the opposite direction to that expected would result in the same action as a non-significant result (Lombardi & Hurlbert, 2009; Ruxton & Neuhaeuser, 2010). There are some limited circumstances in which this might be the case. First, if a result in the opposite direction would be theoretically meaningless or impossible to explain even if you wanted to (Kimmel, 1957). Second, imagine you're testing a new drug to treat depression. You predict it will be better than existing drugs. If it is not better than existing drugs (non-significant p) you would not approve the drug; however, if it was significantly worse than existing

drugs (significant p but in the opposite direction) you would also not approve the drug. In both situations, the drug is not approved.

Finally, one-tailed tests encourage cheating. If you do a two-tailed test and find that your p is 0.06, then you would conclude that your results were not significant (because 0.06 is bigger than the critical value of 0.05). Had you done this test one-tailed, however, the p you would get would be half of the two-tailed value (0.03). This one-tailed value would be significant at the conventional level (because 0.03 is less than 0.05). Therefore, if we find a two-tailed p that is just non-significant, we might be tempted to pretend that we'd always intended to do a one-tailed test because our 'one-tailed' p-value is significant. But we can't change our rules after we have collected data (Jane Superbrain Box 2.7) so we must conclude that the effect is not significant. Although scientists hopefully don't do this sort of thing deliberately, people do get confused about the consequences of their statistical decisions. Two surveys of practice in ecology journals concluded that 'all uses of one-tailed tests in the journals surveyed seemed invalid' (Lombardi & Hurlbert, 2009) and that only 1 in 17 papers using one-tailed tests were justified in doing so (Ruxton & Neuhaeuser, 2010).

One way around the temptation to cheat is to pre-register your study (which we'll discuss in detail in the following chapter). At a simple level, pre-registration means that you commit publicly to your analysis strategy before collecting data. This could be a simple statement on your own website, or, as we shall see, a formally submitted article outlining your research intentions. One benefit of pre-registering your research is that it then becomes transparent if you change your analysis plan (e.g., by switching from a two-tailed to a one-tailed test). It is much less tempting to halve your p-value to take it below 0.05 if the world will know you have done it!

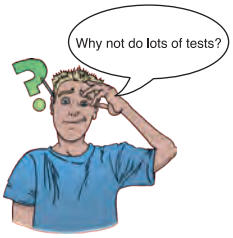
2.10.7 Type I and Type II errors Ⓐ

Neyman and Pearson identified two types of errors that we can make when we test hypotheses. When we use test statistics to tell us about the true state of the world, we're trying to see whether there is an effect in our population. There are two possibilities: there is, in reality, an effect in the population, or there is, in reality, no effect in the population. We have no way of knowing which of these possibilities is true; however, we can look at test statistics and their associated probability to help us to decide which of the two is more likely. There are two mistakes we can make: a Type I and a Type II error. A **Type I error** occurs when we believe that there is an effect in our population, when in fact there isn't. If we use the conventional criterion for alpha then the probability of this error is 0.05 (or 5%) when there is no effect in the population – this value is the **α-level** that we encountered in Figure 2.16. Assuming there is no effect in our population, if we replicated our data collection process 100 times we could expect that on five occasions we would obtain a test statistic large enough to make us think that there was an effect in the population even though there isn't. The opposite is a **Type II error**, which occurs when we believe that there is no effect in the population when, in reality, there is. This would occur when we obtain a small test statistic (perhaps because there is a lot of natural variation between our samples). In an ideal world, we want the probability of this error to be very small (if there is an effect in the population then it's important that we can detect it). Cohen (1992) suggests that the maximum acceptable probability of a Type II error would be 0.2 (or 20%) – this is called the **β-level**. That would mean that if we took 100 samples of data from a population in which an effect exists, we would fail to detect that effect in 20 of those samples (so we'd miss 1 in 5 genuine effects).

There is a trade-off between these two errors: if we lower the probability of accepting an effect as genuine (i.e., make α smaller) then we increase the probability that we'll reject an effect that does genuinely exist (because we've been so strict about the level at which we'll accept that an effect is genuine). The exact relationship between the Type I and Type II error is not straightforward because they are based on different assumptions: to make a Type I error there must be no effect in the population, whereas to make a Type II error the opposite is true (there must be an effect that

we've missed). So, although we know that as the probability of making a Type I error decreases, the probability of making a Type II error increases, the exact nature of the relationship is usually left for the researcher to make an educated guess (Howell, 2012, gives a great explanation of the trade-off between errors).

2.10.8 Inflated error rates Ⓐ



As we have seen, if a test uses a 0.05 level of significance then the chances of making a Type I error are only 5%. Logically, then, the probability of no Type I errors is 0.95 (95%) for each test. However, in science it's rarely the case that we can get a definitive answer to our research question using a single test on our data: we often need to conduct several tests. For example, imagine we wanted to look at factors that affect how viral a video becomes on YouTube. You might predict that the amount of humour and innovation in the video will be important factors. To test this, you might look at the relationship between the number of hits and measures of both the humour content and the innovation. However, you probably ought to also look at whether innovation and humour content are related too. Therefore, you would need to do three tests. If we assume that each test is independent (which in this case they won't be, but it enables us to multiply the probabilities) then the overall probability of no Type I errors will be $(0.95)^3 = 0.95 \times 0.95 \times 0.95 = 0.857$, because the probability of no Type I errors is 0.95 for each test and there are three tests. Given that the probability of no Type I errors is 0.857, then the probability of making at least one Type I error is this number subtracted from 1 (remember that the maximum probability of any event occurring is 1). So, the probability of at least one Type I error is $1 - 0.857 = 0.143$, or 14.3%. Therefore, across this group of tests, the probability of making a Type I error has increased from 5% to 14.3%, a value greater than the criterion that is typically used. This error rate across statistical tests conducted on the same data is known as the **familywise** or **experimentwise error rate**. Our scenario with three tests is relatively simple, and the effect of carrying out several tests is not too severe, but imagine that we increased the number of tests from three to 10. The familywise error rate can be calculated using the following equation (assuming you use a 0.05 level of significance):

$$\text{familywise error} = 1 - (0.95)^n \tag{2.34}$$

In this equation n is the number of tests carried out on the data. With 10 tests carried out, the familywise error rate is $1 - 0.95^{10} = 0.40$, which means that there is a 40% chance of having made at least one Type I error.

To combat this build-up of errors, we can adjust the level of significance for individual tests such that the overall Type I error rate (α) across all comparisons remains at 0.05. There are several ways in which the familywise error rate can be controlled. The most popular (and easiest) way is to divide α by the number of comparisons, k:

$$P_{Crit} = \frac{\alpha}{k} \tag{2.35}$$

Therefore, if we conduct 10 tests, we use 0.005 as our criterion for significance. In doing so, we ensure that the cumulative Type I error remains below 0.05. This method is known as the **Bonferroni correction**, because it uses an inequality described by Carlo Bonferroni, but despite the name its modern application to confidence intervals can be attributed to Olive Dunn (Figure 2.19). There is a trade-off for controlling the familywise error rate and that is a loss of statistical power, which is the next topic on our agenda.



Figure 2.19 The king and queen of correction

2.10.9 Statistical power Ⓑ

We have seen that it is important to control the Type I error rate so that we don't too often mistakenly think that an effect is genuine when it is not. The opposite problem relates to the Type II error, which is how often we miss effects in the population that genuinely exist. If we set the Type II error rate high then we potentially miss a lot of genuine effects, but if we set it low we will be less likely to miss them. The ability of a test to detect an effect is known as its statistical **power** (not to be confused with statistical powder, which is an illegal substance that makes you better understand statistics). The power of a test is the probability that a given test will find an effect, assuming that one exists in the population. This is the opposite of the probability that a given test will *not* find an effect assuming that one exists in the population, which as we have seen is the β-level (i.e., Type II error rate). Therefore, the power of a test can be expressed as $1 - \beta$. Given that Cohen (1988, 1992) recommends a 0.2 probability of failing to detect a genuine effect (see above), the corresponding level of power would be 1 – 0.2, or 0.8. Therefore, we typically aim to achieve a power of 0.8, or put another way, an 80% chance of detecting an effect if one genuinely exists. The power of a statistical test depends on:[16]

1 The size of the effect (see Section 3.5). Self-evidently large effects will be easier to spot than small ones.
2 The α-level. The stricter we are about the threshold to accept a test as 'significant', the harder it will be to 'find' an effect. This strictness is reflected in the α-level. This brings us back to our point in the previous section about correcting for multiple tests. If we use a more conservative Type I error rate for each test (such as a Bonferroni correction) then the probability of rejecting an effect that does exist is increased (we're more likely to make a Type II error). In other words, when we apply a Bonferroni correction we reduce the power of each test to detect an effect.
3 The sample size. We saw earlier in this chapter that larger samples are better approximations of the population; therefore, they have less sampling error. Remember that test statistics are basically a signal-to-noise ratio, so given that large samples have less 'noise' they make it easier to find the 'signal'.

16 It will also depend on whether the test is a one- or two-tailed test (see Section 3.10.5), but, as we have seen, you'd normally do a two-tailed test.

Given that power ($1 - \beta$), the α-level, sample size and the size of the effect are all linked; if we know three of these things, then we can find out the remaining one. There are two things that scientists do with this knowledge:

1. **Calculate the power of a test**: Given that we've conducted our experiment, we will already have selected a value of α, we can estimate the effect size based on our sample data, and we will know how many participants we used. Therefore, we can use these values to calculate $1 - \beta$, the power of our test. If this value turns out to be greater than or equal to 0.8 we can bask in the warm glow of having achieved sufficient power to detect any effects that might have existed. If the resulting value is less than 0.8, then we might want to replicate the experiment using more participants to increase the power.
2. **Calculate the sample size necessary to achieve a given level of power**: We can set the value of α and $1 - \beta$ to be whatever we want (normally, 0.05 and 0.8, respectively). We can also estimate the likely effect size in the population by using data from past research. Even if no one had previously done the exact research that we intend to do, we can still estimate the likely effect size based on similar data. Given this information, we can calculate how many participants we would need to detect the effect of interest (based on the values of α and $1 - \beta$ that we've chosen).

The point of calculating the power of a test after the experiment has always been lost on me a bit: if you find a non-significant effect then you didn't have enough power, if you found a significant effect then you did. Using power to calculate the necessary sample size is the more common and, in my opinion, more useful thing to do. The actual computations are very cumbersome, but there are computer programs available that will do them for you. *G*Power* is a free tool (https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower), there is a package *pwr* that can be used in the open source statistics package *R*, and various websites including https://powerandsamplesize.com. There are also commercial software packages such as *nQuery* (https://www.statsols.com/nquery), *Power and Precision* (www.power-analysis.com) and *PASS* (www.ncss.com/software/pass). Also, Cohen (1988) provides extensive tables for calculating the number of participants for a given level of power (and vice versa).

2.10.10 Confidence intervals and statistical significance Ⓑ

I mentioned earlier (Section 2.9.4) that if 95% confidence intervals for the mean don't overlap then we might conclude that the means come from different populations, and, therefore, they are significantly different. This comment alluded to the fact that there is a relationship between statistical significance and confidence intervals. Cumming and Finch (2001) have three guidelines that are shown in Figure 2.20:

1. 95% confidence intervals that just about touch end-to-end (as in the top left panel of Figure 2.20) represent a p-value for testing the null hypothesis of no differences of approximately 0.01.
2. If there is a gap between the upper end of one 95% confidence interval and the lower end of another (as in the top right panel of Figure 2.20) then $p < 0.01$.
3. A p-value of 0.05 is represented by *moderate* overlap between the bars (the bottom panels of Figure 2.20).

These guidelines are poorly understood by many researchers. In one study (Belia et al., 2005), 473 researchers from medicine, psychology and behavioural neuroscience were shown a graph of means and confidence intervals for two independent groups and asked to move one of the error bars up or down on the graph until they showed a 'just significant difference' (at $p < 0.05$). The sample ranged from new researchers to very experienced ones but, surprisingly, this experience did not predict

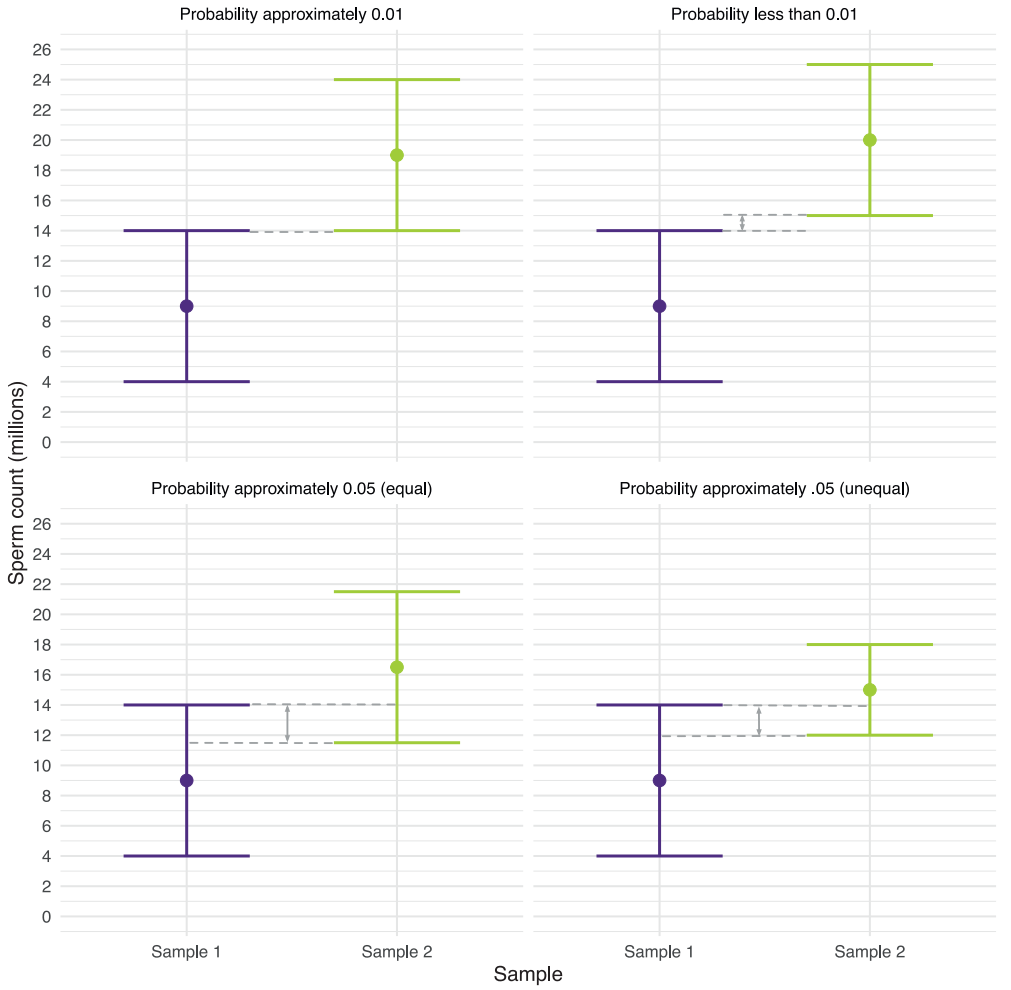


Figure 2.20 The relationship between confidence intervals and statistical significance

their responses. Only a small percentage of researchers could position the confidence intervals correctly to show a just significant difference (15% of psychologists, 20% of behavioural neuroscientists and 16% of medics). The most frequent response was to position the confidence intervals more or less at the point where they stop overlapping (i.e., a *p*-value of approximately 0.01). Very few researchers (even experienced ones) realized that moderate overlap between confidence intervals equates to the standard *p*-value of 0.05 for accepting significance.

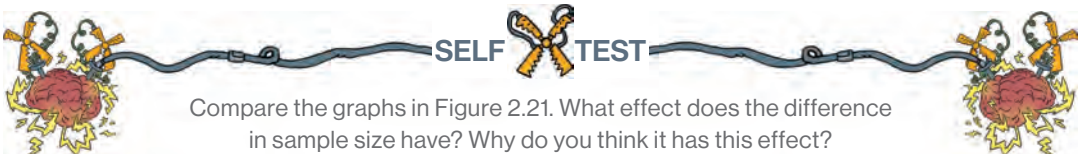
What do we mean by moderate overlap? Cumming defines it as half the length of the average margin of error, MOE (Cumming, 2012; Cumming & Calin-Jageman, 2017). The MOE is half the length of the confidence interval (assuming it is symmetric), so it's the length of the bar sticking out in one direction from the mean. In the bottom left of Figure 2.20 the confidence interval for sample 1 ranges from 4 to 14 so has a length of 10 and an MOE of half this value (i.e., 5). For sample 2, it ranges from 11.5 to 21.5 so again a distance of 10 and an MOE of 5. The average MOE is, therefore $(5 + 5)/2 = 5$. Moderate overlap would be half of this value (i.e., 2.5). This is the amount of overlap between the two confidence intervals in the bottom left of Figure 2.20. Basically, if the confidence intervals are the same length, then $p = 0.05$ is represented by an overlap of about a quarter of the confidence interval. In the more likely scenario of confidence intervals with different lengths, the interpretation of overlap is more difficult. In the bottom right of Figure 2.20 the confidence interval for sample 1 again ranges from 4 to 14 so has a length of 10 and an MOE of 5. For sample 2, it ranges from 12 to 18 so a distance of 6 and

an MOE of half this value, 3. The average MOE is, therefore (5 + 3)/2 = 4. Moderate overlap would be half of this value (i.e., 2) and this is what we get in bottom right of Figure 2.20: the confidence intervals overlap by 2 points on the scale, which equates to a p of around 0.05.

2.10.11 Sample size and statistical significance Ⓑ

When we discussed power, we saw that it is intrinsically linked with the sample size. Given that power is the ability of a test to find an effect that genuinely exists, and we 'find' an effect by having a statistically significant result (i.e., a $p < 0.05$), there is also a connection between the sample size and the p-value associated with a test statistic. We can demonstrate this connection with two examples. Apparently male mice 'sing' to female mice to try to attract them as mates (Hoffmann et al., 2012), I'm not sure what they sing, but I like to think it might be 'This mouse is on fire' by AC/DC, or perhaps 'Mouses of the holy' by Led Zeppelin, or even 'The mouse Jack built' by Metallica. It's probably not 'Terror and Hubris in the mouse of Frank Pollard' by Lamb of God. That would be weird. Anyway, many a young man has spent time wondering how best to attract female mates, so to help them out, imagine we did a study in which we got two groups of 10 heterosexual young men and got them to go up to a woman that they found attractive and either engage them in conversation (group 1) or sing them a song (group 2). We measured how long it was before the woman ran away. Imagine we repeated this experiment but using 100 men in each group.

Figure 2.21 shows the results of these two experiments. The summary statistics from the data are identical: in both cases the singing group had a mean of 10 and a standard deviation of 3, and the conversation group had a mean of 12 and a standard deviation of 3. Remember that the only difference between the two experiments is that one collected 10 scores per sample, and the other 100 scores per sample.



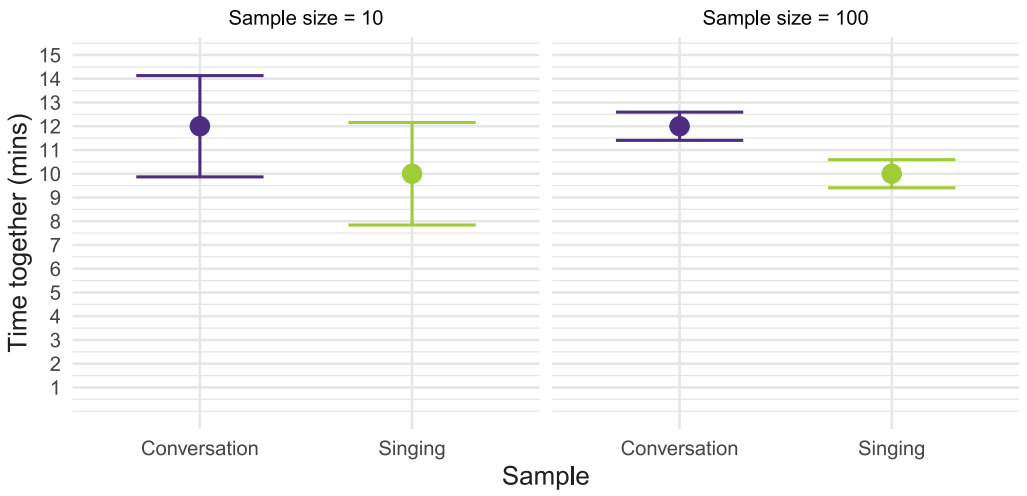


Figure 2.21 Graph showing two data sets with the same means and standard deviations but based on different-sized samples

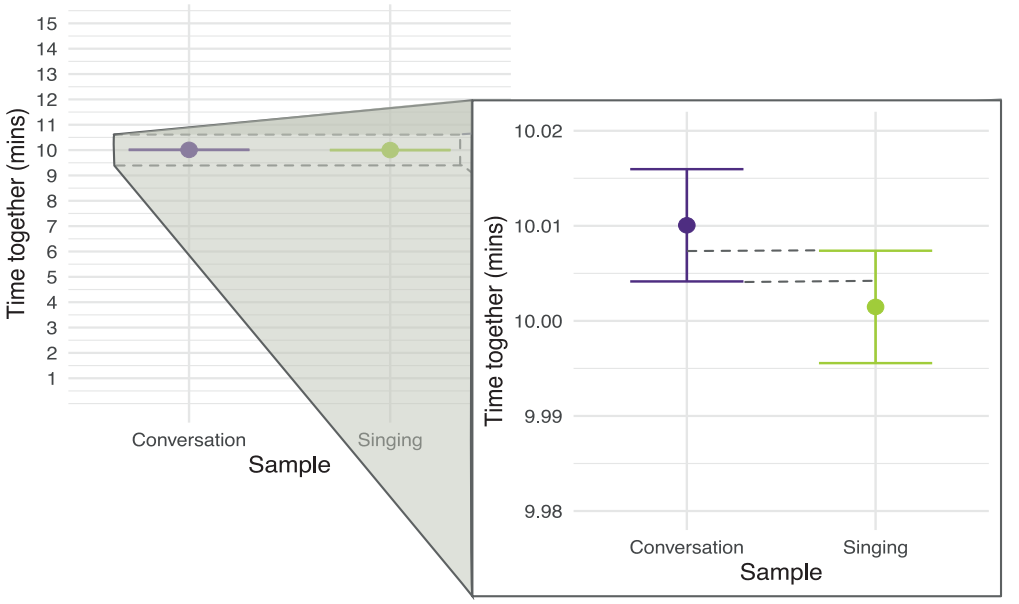


Figure 2.22 A very small difference between means based on an enormous sample size (n = 1,000,000 per group)

Notice in Figure 2.21 that the means for each sample are the same in both plots, but the confidence intervals are much narrower when the samples contain 100 scores than when they contain only 10 scores. You might think that this is odd given that I said that the standard deviations are all the same (i.e., 3). If you think back to how the confidence interval is computed it is the mean plus or minus 1.96 times the standard error. The standard error is the standard deviation divided by the square root of the sample size (see Eq. 2.22), therefore as the sample size gets larger, the standard error gets smaller, and consequently the confidence interval becomes narrower.

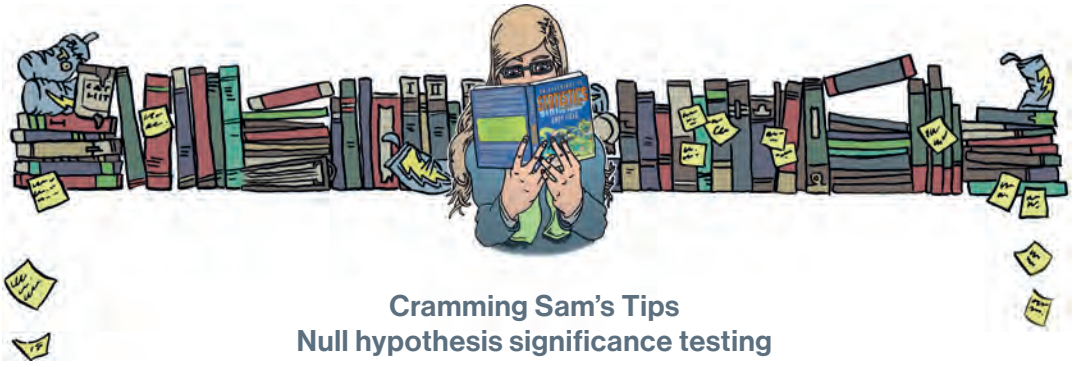
We saw in the previous section that if the confidence intervals of two samples are the same length then a p of around 0.05 is represented by an overlap of about a quarter of the confidence interval. Therefore, we can see that even though the means and standard deviations are identical in both plots, the study that has only 10 scores per sample is not significant (the bars overlap quite a lot; in fact p = 0.15) but the study that has 100 scores per sample shows a highly significant difference (the bars don't overlap at all; the p for these data is less than 0.001). Remember, the means and standard deviations are *identical* in the two graphs, but the sample size affects the standard error and hence the significance.

Taking this relationship to the extreme, we can illustrate that with a big enough sample even a completely meaningless difference between two means can be deemed significant with a $p < 0.05$. Figure 2.22 shows such a situation. This time, the singing group has a mean of 10.00 (SD = 3) and the conversation group has a mean of 10.01 (SD = 3): a difference of 0.01 – a very small difference indeed. The main plot looks very odd: the means look identical and there are no confidence intervals. In fact, the confidence intervals are so narrow that they merge into a single line. The figure also shows a zoomed image of the confidence intervals (note that by zooming in the values on the vertical axis range from 9.98 to 10.02, so the entire range of values in the zoomed image is only 0.04). As you can see, the sample means are 10 and 10.01 as mentioned before[17] but by zooming in we can see the confidence intervals. Note that the confidence intervals show an overlap of about a quarter which equates to a significance value of about p = 0.05 (for these data the actual value of p is 0.044). How is it possible that we have two sample means that are almost identical (10 and 10.01), and have the same

17 The mean of the singing group looks bigger than 10, but this is only because we have zoomed in so much that its actual value of 10.00147 is noticeable.

standard deviations, but are deemed significantly different? The answer is again the sample size: there are 1 million cases in each sample, so the standard errors are minuscule.

This section has made two important points. First, the sample size affects whether a difference between samples is deemed significant or not. *In large samples, small differences can be significant, and in small samples large differences can be non-significant.* This point relates to power: large samples have more power to detect effects. Second, even a difference of practically zero can be deemed 'significant' if the sample size is big enough. Remember that test statistics are effectively the ratio of signal to noise, and the standard error is our measure of 'sampling noise'. The standard error is estimated from the sample size, and the bigger the sample size, the smaller the standard error. Therefore, bigger samples have less 'noise', so even a tiny signal can be detected.



Cramming Sam's Tips
Null hypothesis significance testing

- NHST is a widespread method for assessing scientific theories. The basic idea is that we have two competing hypotheses: one says that an effect exists (the *alternative hypothesis*) and the other says that an effect doesn't exist (the *null hypothesis*). We compute a test statistic that represents the alternative hypothesis and calculate the probability that we would get a value as big as the one we have if the null hypothesis were true. If this probability is less than 0.05 we reject the idea that there is no effect, say that we have a *statistically significant* finding and throw a little party. If the probability is greater than 0.05 we do not reject the idea that there is no effect, we say that we have a *non-significant* finding and we look sad.
- We can make two types of error: we can believe that there is an effect when, in reality, there isn't (a *Type I error*) and we can believe that there is not an effect when, in reality, there is (a *Type II error*).
- The power of a statistical test is the probability that it will find an effect when one exists.
- The significance of a test statistic is directly linked to the sample size: the same effect will have different *p*-values in different-sized samples, small differences can be deemed 'significant' in large samples, and large effects might be deemed 'non-significant' in small samples.



2.11 Reporting significance tests Ⓑ

In Section 1.8 we looked at some general principles for reporting data. Now that we have learnt a bit about fitting statistical models, we can add to these guiding principles. We learnt in this chapter that we can construct confidence intervals (usually 95% ones) around a parameter such as the mean. A 95%

confidence interval contains the population value in 95% of samples, so if your sample is one of those 95% the confidence interval contains useful information about the population value. It is important to tell readers the type of confidence interval used (e.g., 95%) and in general we use the format [*lower boundary*, *upper boundary*] to present the values. So, if we had a mean of 30 and the confidence interval ranged from 20 to 40, we might write $M = 30$, 95% CI [20, 40]. If we were reporting lots of 95% confidence intervals it might be easier to state the level at the start of our results and just use the square brackets:

✓ 95% confidence intervals are reported in square brackets. Fear reactions were higher, $M = 9.86$ [7.41, 12.31], when Andy's cat Fuzzy wore a fake human tongue compared to when he did not, $M = 6.58$ [3.47, 9.69].

We also saw that when we fit a statistical model we calculate a test statistic and a *p*-value associated with it. Scientists typically conclude that an effect (our model) is significant if this *p*-value is less than 0.05. APA style is to remove the zero before the decimal place (so you'd report $p = .05$ rather than $p = 0.05$) but because many other journals don't have this idiosyncratic rule this is an APA rule that I don't follow in this book. Historically, people would report *p*-values as being either less than or greater than 0.05. They would write things like:

× Fear reactions were significantly higher when Andy's cat Fuzzy wore a fake human tongue compared to when he did not, $p < 0.05$.

If an effect was very significant (e.g., if the *p*-value was less than 0.01 or even 0.001), they would also use these two criteria to indicate a 'very significant' finding:

× The number of cats intruding into the garden was significantly less when Fuzzy wore a fake human tongue compared to when he didn't, $p < 0.01$.

Non-significant effects would be reported in much the same way (note this time the *p* is reported as greater than 0.05):

× Fear reactions were not significantly different when Fuzzy wore a David Beckham mask compared to when he did not, $p > 0.05$.

In the days before computers it made sense to use these standard benchmarks for reporting significance because it was a bit of a pain to compute exact significance values (Jane Superbrain Box 3.1). However, computers make computing *p*-values a piece of *p*s, so we should report exact *p*-values because it gives the reader more information than knowing that the *p*-value was less or more than an arbitrary threshold like 0.05. The possible exception is the threshold of 0.001. If we find a *p*-value of 0.0000234 then for the sake of everyone's sanity it would be reasonable to report $p < 0.001$:

✓ Fear reactions were significantly higher when Andy's cat Fuzzy wore a fake human tongue compared to when he did not, $p = 0.023$.
✓ The number of cats intruding into the garden was significantly less when Fuzzy wore a fake human tongue compared to when he did not, $p = 0.007$.

2.12 Jane and Brian's story Ⓐ

Brian was feeling a little deflated after his encounter with Jane. She had barely said a word to him. Was he that awful to be around? He was enjoying the book she'd handed to him though. The whole thing had spurred him on to concentrate more during his statistics lectures. Perhaps that's what she wanted.

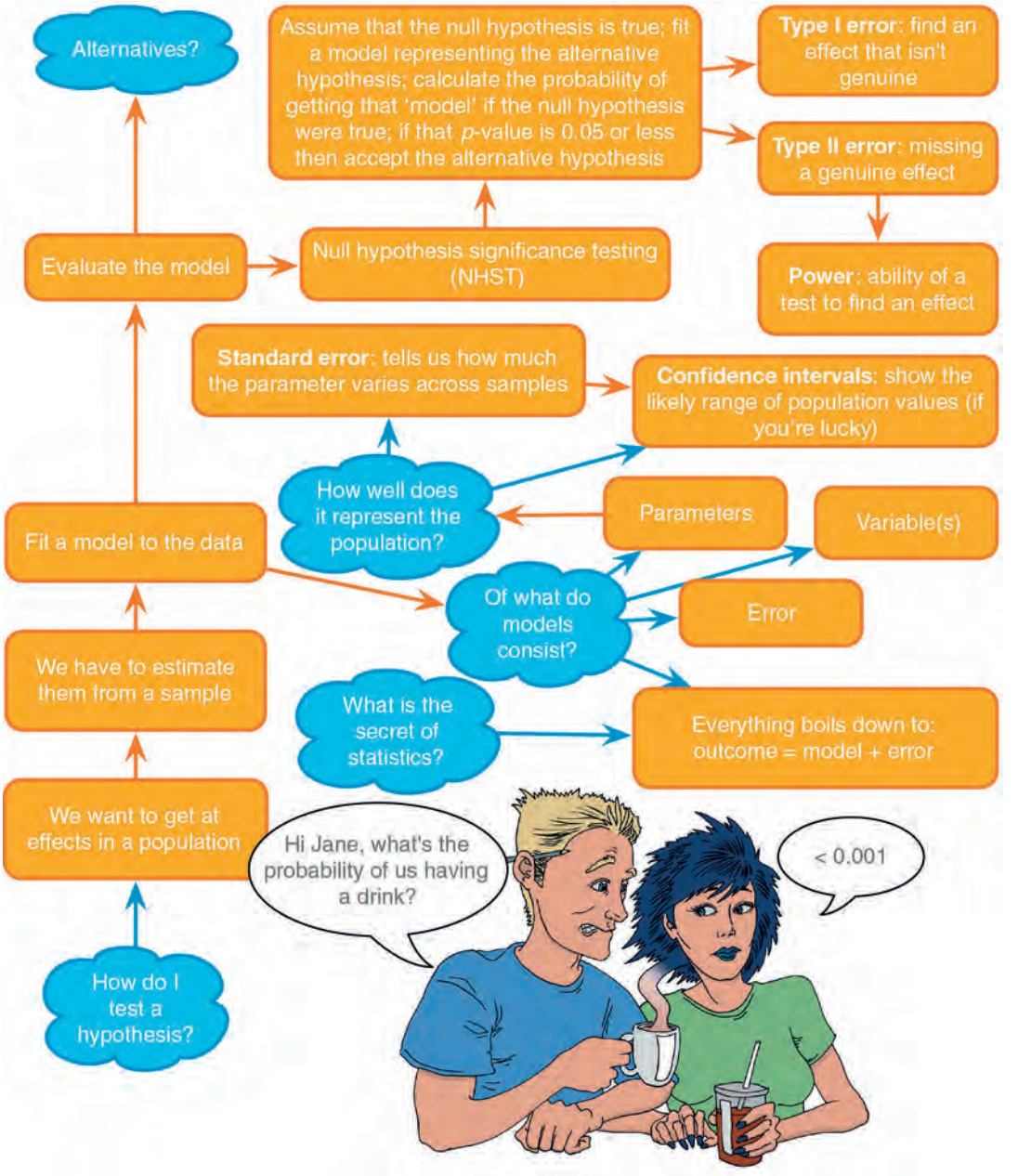


Figure 2.23 What Brian learnt from this chapter

He'd seen Jane flitting about campus. She always seemed to have a massive bag with her. It seemed to contain something large and heavy, judging from her posture. He wondered what it was as he day-dreamed across the campus square. His thoughts were broken by being knocked to the floor.

'Watch where you're going,' he said angrily.

'I'm so sorry …' the dark-haired woman replied. She looked flustered. She picked up her bag as Brian looked up to see it was Jane. Now *he* felt flustered.

Jane looked as though she wanted to expand on her apology but the words had escaped. Her eyes darted as though searching for them.

Brian didn't want her to go, but in the absence of anything to say he recited his statistics lecture to her. It was weird. Weird enough that as he finished she shrugged and ran off.

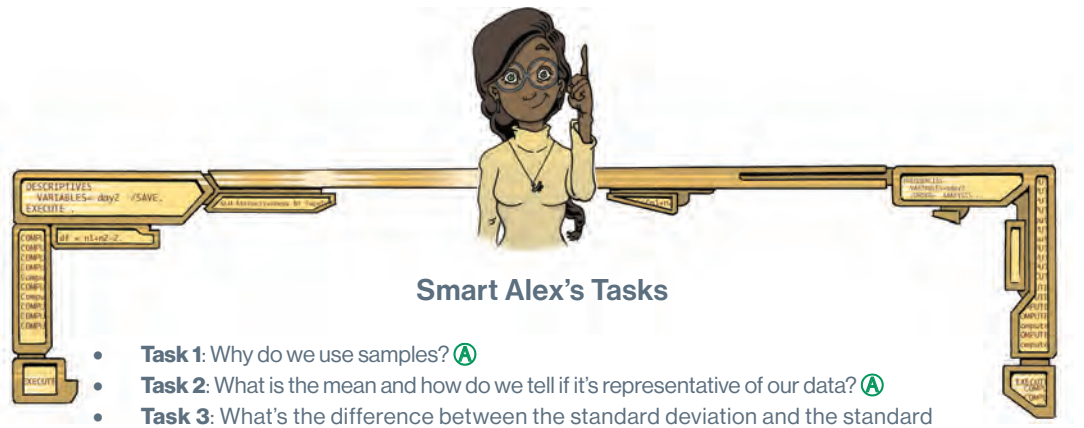
2.13 What next? Ⓐ

Nursery school was the beginning of an educational journey that I am a still on several decades later. As a child, your belief systems are very adaptable. One minute you believe that sharks can miniaturize themselves, swim up pipes from the sea to the swimming pool you're currently in before restoring their natural size and eating you; the next minute you don't, simply because your parents say it's not possible. At the age of 3 any hypothesis is plausible, every way of life is acceptable, and multiple incompatible perspectives can be accommodated. Then a bunch of adults come along and force you to think more rigidly. Suddenly, a cardboard box is not a high-tech excavator, it's a cardboard box, and there are 'right' and 'wrong' ways to live your life. As you get older, the danger is that – left unchecked – you plunge yourself into a swimming-pool-sized echo chamber of your own beliefs leaving your cognitive flexibility in a puddle at the edge of the pool. If only sharks *could* compress themselves ...

Before you know it, you're doggedly doing things and following rules that you can't remember the reason for. One of my early beliefs was that my older brother Paul (more on him later…) was 'the clever one'. Far be it from me to lay the blame at anyone's feet for this belief, but it probably didn't help that members of my immediate family used to say things like 'Paul is the clever one, but at least you work hard'. Like I said, over time, if nothing challenges your view you can get very fixed in a way of doing things or a mode of thinking. If you spend your life thinking you're not 'the clever one', how can you ever change that? You need something unexpected and profound to create a paradigm shift. The next chapter is all about breaking ways of thinking that scientists have been invested in for a long time.

2.14 Key terms that I've discovered

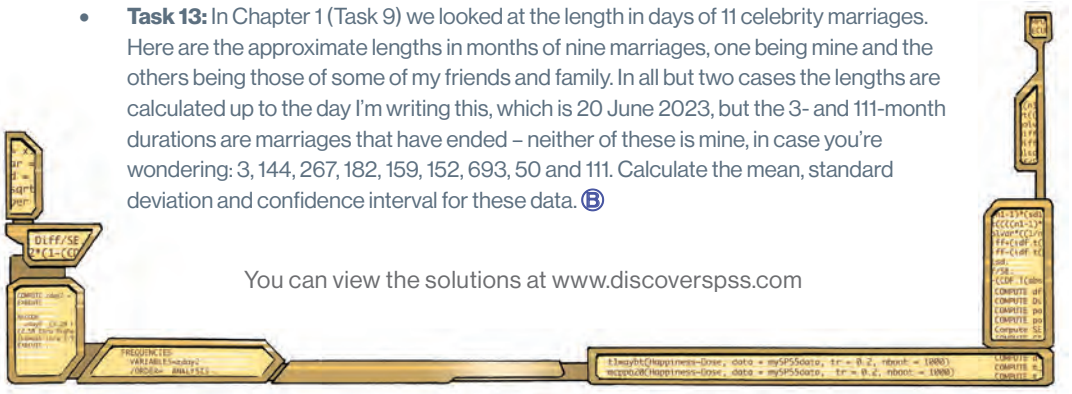
α-level
Alternative hypothesis
β-level
Bonferroni correction
Central limit theorem
Confidence interval
Degrees of freedom
Experimental hypothesis
Experimentwise error rate
Familywise error rate
Fit
Interval estimate
Likelihood
Linear model
Maximum likelihood
Method of least squares
Null hypothesis
One-tailed test
Ordinary least squares
Parameter
Point estimate
Population
Power
Sample
Sampling distribution
Sampling variation
Standard error
Standard error of the mean (SE)
Test statistic
Two-tailed test
Type I error
Type II error



Smart Alex's Tasks

- **Task 1**: Why do we use samples? Ⓐ
- **Task 2**: What is the mean and how do we tell if it's representative of our data? Ⓐ
- **Task 3**: What's the difference between the standard deviation and the standard error? Ⓐ

- **Task 4**: In Chapter 1 we used an example of the time taken for 21 heavy smokers to fall off a treadmill at the fastest setting (18, 16, 18, 24, 23, 22, 22, 23, 26, 29, 32, 34, 34, 36, 36, 43, 42, 49, 46, 46, 57). Calculate the standard error and 95% confidence interval of these data. Ⓑ
- **Task 5**: What do the sum of squares, variance and standard deviation represent? How do they differ? Ⓐ
- **Task 6**: What is a test statistic and what does it tell us? Ⓐ
- **Task 7**: What are Type I and Type II errors? Ⓐ
- **Task 8**: What is statistical power? Ⓑ
- **Task 9**: Figure 2.21 shows two experiments that looked at the effect of singing versus conversation on how much time a woman would spend with a man. In both experiments the means were 10 (singing) and 12 (conversation), the standard deviations in all groups were 3, but the group sizes were 10 per group in the first experiment and 100 per group in the second. Compute the values of the confidence intervals displayed in the figure. Ⓑ
- **Task 10**: Figure 2.22 shows a similar study to that in Task 9, but the means were 10 (singing) and 10.01 (conversation), the standard deviations in both groups were 3 and each group contained 1 million people. Compute the values of the confidence intervals displayed in the figure. Ⓑ
- **Task 11:** In Chapter 1 (Task 8) we looked at an example of how many games it took a sportsperson before they hit the 'red zone' Calculate the standard error and confidence interval for those data. Ⓑ
- **Task 12**: At a rival club to the one I support, they similarly measured the number of consecutive games it took their players before they reached the red zone. The data are: 6, 17, 7, 3, 8, 9, 4, 13, 11, 14, 7. Calculate the mean, standard deviation and confidence interval for these data. Ⓑ
- **Task 13:** In Chapter 1 (Task 9) we looked at the length in days of 11 celebrity marriages. Here are the approximate lengths in months of nine marriages, one being mine and the others being those of some of my friends and family. In all but two cases the lengths are calculated up to the day I'm writing this, which is 20 June 2023, but the 3- and 111-month durations are marriages that have ended – neither of these is mine, in case you're wondering: 3, 144, 267, 182, 159, 152, 693, 50 and 111. Calculate the mean, standard deviation and confidence interval for these data. Ⓑ



You can view the solutions at www.discoverspss.com



THE PHOENIX OF STATISTICS

3.1 What will this chapter tell me? 106
3.2 Problems with NHST 107
3.3 NHST as part of wider problems with science 117
3.4 A phoenix from the EMBERS 123
3.5 Sense, and how to use it 124
3.6 Preregistering research and open science 125
3.7 Effect sizes 126
3.8 Bayesian approaches 135
3.9 Reporting effect sizes and Bayes factors 145
3.10 Jane and Brian's story 145
3.11 What next? 146
3.12 Key terms that I've discovered 147
Smart Alex's Tasks 147

Letters A to D indicate increasing difficulty

3.1 What will this chapter tell me? Ⓐ

I concluded the previous chapter with a self-pitying tale about my family thinking that my brother, Paul, was 'the clever one'. Perhaps you're anticipating an anecdote about how, on my fourth birthday, I presented my parents with a time machine that I had invented (tweaking Einstein's relativity theory in the process) and invited them to take a trip back a few years so that they could unsay this. That didn't happen; apart from anything, my parents meeting themselves would probably have caused a rift in space-time or something and we'd all be leaking into each other's parallel universes. Which, of course, might be fun if we ended up in a parallel universe where statistics hasn't been invented.

There was only one culturally appropriate way to deal with my family's low expectations of me: emotional repression. I silently internalized their low opinion of my intellect into a deep-seated determination to prove them wrong. The plan started slowly: I distracted myself with playing football and guitar. Then, at some point during my teens, when the resentment had festered like the contents of a can of surströmming, I rose like a phoenix of statistics and worked myself into the ground for 20 years to get a degree, PhD, publish tonnes of research papers, get grants and write textbooks. If you do enough of that stuff eventually you get to call yourself 'Professor', which sounds posh, like you might be clever. 'Who is the clever one now?' I thought to myself as I watched my brother enjoying a life in consultancy free from stress-related psychological problems and earning a lot more money than me.

The point is that I'd got locked into a pattern of thinking in which my internal value as a human was connected to my academic achievements. As long as I 'achieved', I could justify taking up valuable space on this planet. I'd been invested in this 'habit' for close to 40 years. Like null hypothesis significance testing from the previous chapter, I'd found a system, a recipe to follow, that rewarded me. I thought I needed to achieve because that's how I'd been taught to think; I defined myself by academic success because I'd spent years defining myself by academic success. I never questioned the system, I never looked for its weaknesses, I never investigated an alternative.



Figure 3.1 'Hi! Can we interest you in a paradigm shift?'

Then Zach and Arlo came along[1] and changed everything. There's something about the

1 This turn of phrase makes it sound a bit like I opened my wardrobe one morning and two cheeky grins emanated out of the darkness. My wife certainly would have preferred that mode of delivery. Although their arrival was not in any way unexpected, their endless capacity to melt my heart was.

overwhelming, unconditional love that these two little humans evoke that shows up a career in academia for the hollow, pointless waste of time that it has been. Has twenty years of my research made even a minuscule difference to anyone anywhere? No, of course it hasn't. Would the world be worse off if I never existed to write statistics books? No: there are a lot of people better at statistics than I am. Do my family now think I'm cleverer than my brother? Obviously not. I thought my 'system' was a way to the truth, but its logic was flawed. Gazing into Zach and Arlo's eyes, I see two hearts that would be irrevocably broken if I ceased to exist. If I work a 60-hour week 'achieving', the world will not care, but two little boys *will* care that they don't see their dad. If I write 100 more articles the world won't change, but if I read Zach and Arlo 100 more stories their world *will* be better. Some people do science that changes the world, but I'm not one of them. I'll never be an amazing scientist, but I *can* be an amazing dad.

That was my personal paradigm shift and this chapter is about a statistical one. We shift away from the dogmatic use of null hypothesis significance testing (NHST) by exploring its limitations and how they feed into wider issues in science. We then discover some alternatives. It's possible that this chapter might come across as me trying to deter you from science by banging on about how awful it is. Science *is* awesome, and I believe that mostly it's done well by people who care about finding out the truth, but it's not perfect. Those of you who don't plan to be scientists need to know how to scrutinize scientific work, so that when politicians and the media try to convince you of things using science, you have the skills to read the science with a critical mind, aware of what may have influenced the results and conclusions. Many of you *will* want to become scientists, and for you this chapter is here because science needs you to make it better. Unhelpful scientific culture won't be overturned by the 'set-in-their-ways' professors, but it will be by the brilliant minds of the future. That's you, because you are better than some of the dodgy stuff that currently happens.

3.2 Problems with NHST Ⓑ

We saw in the previous chapter that NHST is the dominant method for testing theories using statistics. It is compelling because it offers a rule-based framework for deciding whether to believe a hypothesis. It is also appealing to teach because even if your students don't understand the logic behind NHST, most of them can follow the rule that a $p \leq 0.05$ is 'significant' and a $p > 0.05$ is not. No one likes to get things wrong and NHST seems to provide an easy way to disentangle the 'correct' conclusion from the 'incorrect' one. Like timid bakers we want a delicious fruitcake rather than egg on our face, so we diligently follow the steps in the NHST recipe. Unfortunately, when you bite into an NHST cake it tastes like one of those fermented sharks that some Icelandic people bury in small holes for 3 months to maximize their rancidness.[2] Here are two quotes about NHST that sum up just what a putrefying shark it is:

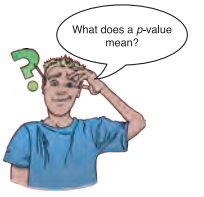
2 I probably shouldn't knock *hákarl* until I've tried it, but if I'm ever in Iceland I'm going to play my vegetarian trump card to the max to avoid doing so. Apologies to any Icelandic readers for being such a massive, non-Viking, wuss. I'm rubbish at growing beards and invading European countries too.

> The almost universal reliance on merely refuting the null hypothesis is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology. (Meehl, 1978, p. 817)
>
> NHST; I resisted the temptation to call it Statistical Hypothesis Inference Testing. (Cohen, 1994, p. 997)

This section rakes the stones, gravel and sand to reveal the decaying shark carcass that is NHST. We momentarily gag on its aroma of urine-infused cheese before … OK, I've gone too far with the shark thing, haven't I?

3.2.1 Misconceptions about statistical significance Ⓑ

People have devoted entire books to the problems with NHST (for example, Ziliak & McCloskey, 2008), but we have space only to scratch the surface. We'll start with three common misconceptions (in no particular order) about what a statistically significant result (typically defined as a *p*-value less than or equal to 0.05) allows you to conclude. For a good paper on NHST misconceptions, see Greenland et al. (2016).



Misconception 1: A significant result means that the effect is important

Statistical significance is not the same thing as importance because the *p*-value from which we determine significance is affected by sample size (Section 2.10.11). Therefore, don't be fooled by that phrase 'statistically significant', because very small and unimportant effects will be statistically significant if sufficiently large amounts of data are collected (Figure 2.22), and very large and important effects will be missed if the sample size is too small.

Misconception 2: A non-significant result means that the null hypothesis is true

Actually, no. If the *p*-value is greater than 0.05 then you could decide to reject the alternative hypothesis,[3] but this is not the same as the null hypothesis being true. A non-significant result tells us only that the effect is not big enough to be found (given our sample size), it doesn't tell us that the effect size is zero. The null hypothesis is *a hypothetical* construct. We can assume an effect of zero in a population to calculate a probability distribution under the null hypothesis, but it isn't reasonable to think that a real-world effect is zero to an infinite number of decimal places. Even if a population effect had a size of 0.000000001, that is not the same as zero, and given a big enough sample that effect could be detected and deemed 'statistically significant' (Cohen, 1990, 1994). Therefore, a non-significant result should never be interpreted as 'no difference between means' or 'no relationship between variables'.

Misconception 3: A significant result means that the null hypothesis is false

Wrong again. A significant test statistic is based on probabilistic reasoning, which limits what we can conclude. Cohen (1994) points out that formal reasoning relies on an initial statement of fact followed by a statement about the current state of affairs, and an inferred conclusion. This syllogism illustrates what he meant:

- If a *person plays the flute*, then the person is not *a member of the band Iron Maiden*.

3 You shouldn't because your study might be underpowered, but people do.

    - This person is a member of Iron Maiden.
        - Therefore, this person does not play the flute.

The syllogism starts with a statement of fact that allows the end conclusion to be reached because you can deny the person plays the flute (the antecedent) by denying that 'not a member of Iron Maiden' (the consequent) is true. A comparable version of the null hypothesis is:

- If the *null hypothesis* is correct, then this *test statistic value* cannot occur.
    - This test statistic value has occurred.
        - Therefore, the null hypothesis is not correct.

This is all very nice except that the null hypothesis is not characterized by a statement of fact such as 'if the null hypothesis is correct, then this test statistic cannot occur'; instead it reflects a statement of probability such as 'if the null hypothesis is correct, then this test statistic is highly unlikely'. Not starting out with a statement of fact messes up the consequent logic. The syllogism becomes:

- If the *null hypothesis* is correct, then it is highly unlikely *to get this test statistic value*.
    - This test statistic value has occurred.
        - Therefore, the null hypothesis is highly unlikely.

If, like me, logic makes your brain pulsate in an unpleasant way then it might not be obvious why the syllogism no longer makes sense, so let's convert it to something more tangible by replacing the phrase 'null hypothesis' with 'person plays guitar', and the phrase 'to get this test statistic value' with 'the person is a member of the band Iron Maiden'. Let's see what we get:

- If *person plays guitar* is correct, then it is highly unlikely that *the person is a member of the band Iron Maiden*.
    - This person is a member of Iron Maiden.
        - Therefore, *person plays guitar* is highly unlikely.

Let's break this syllogism down. The first statement is true. On Earth, there are (very approximately) 50 million people who play guitar and only three of them are in Iron Maiden. Therefore, the probability of someone being in Iron Maiden given that they play guitar is about 3/50 million or 6×10^{-8}. In other words, it *is* very unlikely. The consequent logical statements assert that given someone *is* a member of Iron Maiden it is unlikely that they play guitar, which is demonstrably false. Iron Maiden have six band members and three of them play guitar (I'm not including bass guitar in any of this), so, given that a person is in Iron Maiden, the probability that they play guitar is 3/6 or 0.5 (50%), which is quite likely (not unlikely). This example illustrates a common fallacy in hypothesis testing.

To sum up, although NHST is the result of trying to find a system that can test which of two competing hypotheses (the null or the alternative) is likely to be correct, it fails because the significance of the test provides no evidence about either hypothesis.

3.2.2 All-or-nothing thinking Ⓑ

Perhaps the biggest practical problem created by NHST is that it encourages all or nothing thinking: if $p \leq 0.05$ then an effect is significant, but if $p > 0.05$ it is not. One scenario that illustrates the ridiculousness of this thinking is where you have two effects, based on the same sample sizes,

and one has a $p = 0.0499$, and the other a $p = 0.0501$. If you apply the NHST recipe book, then the first effect is significant and the second is not. You'd reach completely opposite conclusions based on p-values that differ by only 0.0002. Other things being equal, these p-values would reflect basically the same-sized effect, but the 'rules' of NHST encourage people to treat them as completely opposite.

There is nothing magic about the criterion of $p < 0.05$, it is merely a convenient rule of thumb that has become popular for arbitrary reasons (see Jane Superbrain Box 3.1). When I outlined the NHST process (Section 2.10.3) I said that you should set an idiosyncratic alpha that is meaningful for your research question (Maier & Lakens, 2022), but that most people choose 0.05 without thinking about why they are doing so.

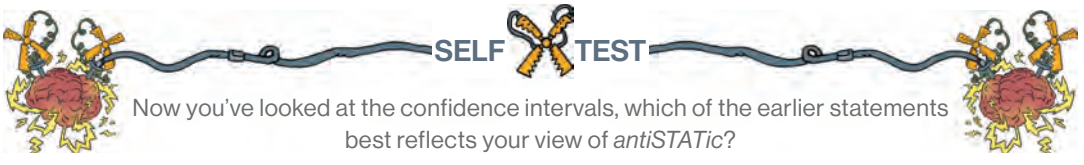
Let's look at how the recipe-book nature of NHST encourages us to think in these black-and-white terms, and how misleading that can be. Students are often very scared of statistics. Imagine that a scientist claimed to have found a cure for statistics anxiety: a potion containing badger sweat, a tear from a newborn child, a teaspoon of Guinness, some cat saliva and sherbet. She called it *antiSTATic*. Treat yourself by entertaining the fiction that 10 researchers all did a study in which they compared anxiety levels in students who had taken *antiSTATic* to those who had taken a placebo potion (water). If *antiSTATic* doesn't work, then there should be a difference of zero between these group means (the null hypothesis), but if it does work then those who took *antiSTATic* should be less anxious than those who took the placebo (which will show up in a positive difference between the groups). The results of the 10 studies are shown in Figure 3.2 along with the p-value within each study.



Based on what you have learnt so far, which of the following statements best reflects your view of *antiSTATic*?

a The evidence is equivocal; we need more research.
b All the mean differences show a positive effect of *antiSTATic*; therefore, we have consistent evidence that *antiSTATic* works.
c Four of the studies show a significant result ($p < 0.05$), but the other six do not. Therefore, the studies are inconclusive: some suggest that *antiSTATic* is better than placebo, but others suggest there's no difference. The fact that more than half of the studies showed no significant effect means that *antiSTATic* is not (on balance) more successful in reducing anxiety than the control.
d I want to go for C, but I have a feeling it's a trick question.

Based on what I have told you about NHST you should have answered C: only four of the 10 studies have a 'significant' result, which isn't very compelling evidence for *antiSTATic*. Now pretend you know nothing about NHST, look at the confidence intervals and think about what we know about overlapping confidence intervals.



Now you've looked at the confidence intervals, which of the earlier statements best reflects your view of *antiSTATic*?

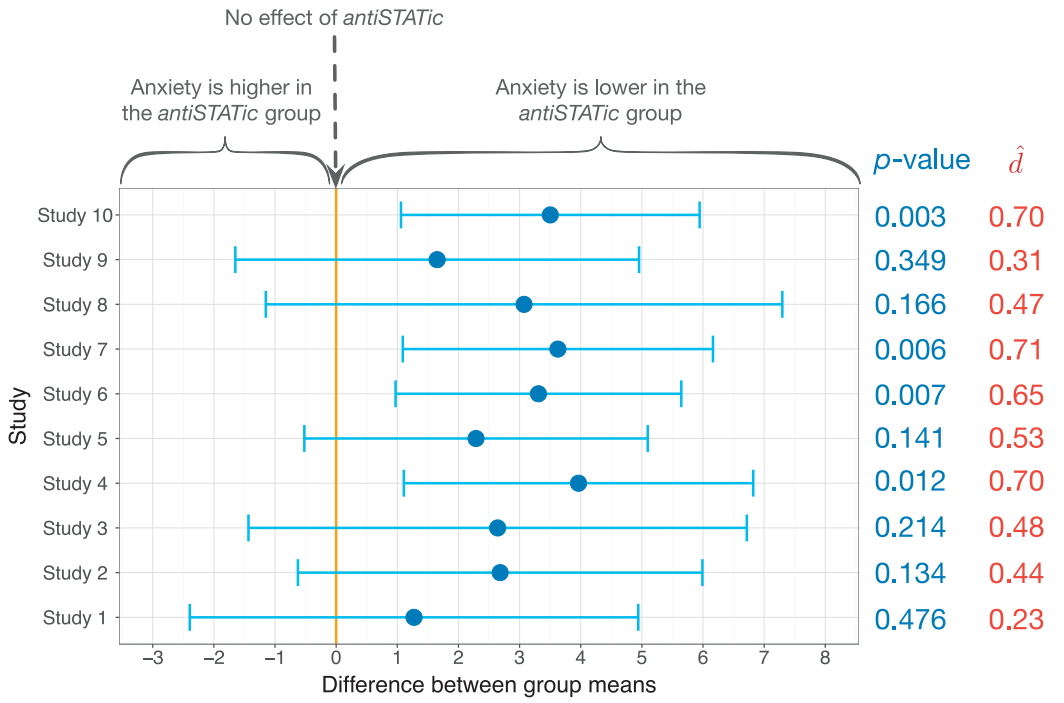
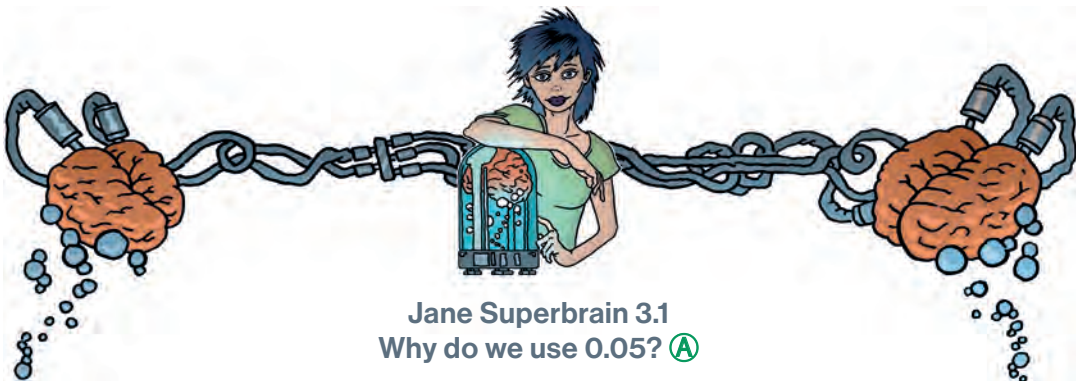


Figure 3.2 Results of 10 different studies looking at the difference between two interventions. The circles show the mean difference between groups

I would hope that some of you have changed your mind to option B. If you're still sticking with option C, then let me try to convince you otherwise. First, 10 out of 10 studies show a positive effect of *antiSTATic* (none of the mean differences are below zero), and even though sometimes this positive effect is not always 'significant', it is consistently positive. The confidence intervals overlap with each other substantially in all studies, suggesting consistency across the studies: they all throw up (potential) population effects of a similar size. Remember that the confidence interval will contain the actual population value in 95% of samples. Look at how much of each confidence interval falls above zero across the 10 studies: even in studies for which the confidence interval includes zero (implying that the population effect might be zero) most of the interval is greater than zero. Again, this suggests very consistent evidence that the population value could be greater than zero (i.e., *antiSTATic* works). Therefore, looking at the confidence intervals rather than focusing on significance allows us to see the consistency in the data and not a bunch of apparently conflicting results (based on NHST): in all studies the effect of *antiSTATic* was positive, and taking all 10 studies into account, there's good reason to think that the population effect is plausibly greater than zero.

3.2.3 *p*-values are influenced by the intentions of the scientist ©

Another problem is that the *p*-value and the decision rules applied to it depend on what the researcher intended to do before collecting data (Wagenmakers, 2007). Let's look at an example. In Section 2.10.4 we looked at a dating example to illustrate the process of obtaining a *p*-value. Have a look back at that section. To summarize the example, you had recently put your profile onto a dating app and were interested in how popular it had been relative to the norm. We derived the following hypotheses:



Jane Superbrain 3.1
Why do we use 0.05? Ⓐ

Given that the criterion of 95% confidence, or a 0.05 probability, is so ubiquitous in NHST, you'd expect a very solid justification for it, wouldn't you? Think again. The mystery of how the 0.05 criterion came to be is complicated. Fisher believed that you calculate the probability of an event and evaluate this probability within the research context. Although Fisher felt that $p = 0.01$ would be strong evidence to back up a hypothesis, and perhaps a $p = 0.20$ would be weak evidence, he objected to Neyman's use of an alternative hypothesis (among other things). Conversely, Neyman objected to Fisher's exact probability approach (Berger, 2003; Lehmann, 1993). The confusion created by the schism between Fisher and Neyman was like a lightning bolt bringing life to NHST: a bastard child of both approaches. I use the word 'bastard' advisedly.

During the decades in which Fisher and Neyman's ideas were stitched together into a horrific Frankenstein, the probability level of 0.05 rose to prominence. The likely reason is that in the days before computers, scientists had to compare their test statistics against published tables of 'critical values' (they did not have software to calculate exact probabilities for them). These critical values had to be calculated by people like Fisher, who produced tables of these values in his influential textbook *Statistical Methods for Research Workers* (Fisher, 1925).[4] To save space, Fisher tabulated critical values for specific probability values (0.05, 0.02 and 0.01). This book had a monumental impact (to get some idea of its influence 25 years after publication see Mather, 1951; Yates, 1951) and the tables of critical values were widely used – even Fisher's rivals, Neyman and Pearson, admitted that the tables influenced them (Lehmann, 1993). This disastrous combination of researchers confused about the Fisher and Neyman–Pearson approaches and the availability of critical values for only certain levels of probability created a trend to report test statistics as being significant at the now infamous $p \leq 0.05$ and $p \leq 0.01$ (because critical values were readily available at these probabilities). Despite this, Fisher believed that the dogmatic use of a fixed level of significance was silly:

> no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (Fisher, 1956)



4 You can read this online at http://psychclassics.yorku.ca/Fisher/Methods/

- Null hypothesis: your profile has been as popular as the norm.
- Alternative hypothesis: because you're a confident type of person you hypothesize that your profile will be more popular than the norm.

On this dating app when someone sees your profile they either swipe it away (i.e., reject you) or click a heart icon (i.e., are interested in you as a potential date). Before we collected any data, we fixed the sample size at $N = 10$. We then took a sample of the first 10 people to view your profile and counted the number who liked it. In this example the intention of the researcher is to fix N; that is, to collect data until a certain number of observations have been collected (in this case 10).

We discovered that in general profiles are liked about 40% of the time, so for every 10 views of a profile, on average, 4 people will like it. We also discovered that we could use something called a binomial distribution with $n = 10$ and $p = 0.4$ to model this situation (Figure 2.17 is reproduced in the left panel of Figure 3.3). Remember that our null hypothesis was that our profile was as popular as the norm, so this is what we might label the 'test statistic under the null hypothesis', or t_{null} for short (Kruschke, 2013). In our example, the distribution represents the relative probabilities of every possible value of t_{null} and the likelihood that it will occur (compared to other values of t_{null}) given that 10 observations are made. In other words, given an infinite number of replications of our experiment, using an identical sampling process involving a fixed N (in this case 10), these are the relative probabilities of every possible value of t_{null}. Our observed t_{null} was 8, that is, we observed that 8 of the 10 people who viewed our profile liked it. We then compared this observed value to the distribution of t_{null}. We found that the probability of observing a value of t_{null} at least as large as 8 was $p(x \geq 8) = 0.0123$. This p-value is a long-run probability: it is the relative frequency of the most recent observed t compared to all possible values of t_{null}.

Now imagine a parallel universe where you're in a bit of a rush, so even though you intended to collect data from the first 10 people that view your profile, you cut the experiment short after an hour, at which point only 9 people had viewed your profile. Your intended sample size of 10 has dropped to 9. No big deal, you think, it's just one person, and if p-values represent some kind of objective truth about the world then this won't affect them. Let's see. We can again use a binomial distribution and follow the same calculations as in Section 2.10.4, the difference being that the parameters of that distribution have changed: whereas if we'd stuck to our original intention of sampling 10 profile viewers we'd use a binomial distribution with $n = 10$ and $p = 0.4$ to model the situation, we now use a binomial distribution with $n = 9$ and $p = 0.4$. The centre panel of Figure 3.3 shows this new distribution alongside the distribution we would have used had we stuck to our original intention in the left panel. Note that the red area (the area representing the p-value) is a smaller area in the centre panel than in the left panel. Also note that the mean of the new distribution is different than our original distribution.[5] We can work out the size of the red area using Eq. 2.27, just like we did in Chapter 1; the only difference is that 10 becomes 9 in the equations. We work out the probability of observing 8 (Eq. 3.1) and 9 (Eq. 3.2) likes of a profile. For the binomial distribution with $n = 9$, it is impossible to observe 10 likes and so the probability associated with this value is 0 (basically we don't include it in the calculation). To get the total probability we add these probabilities up (Eq. 3.3). The resulting p-value is $p(x \geq 8) = 0.0038$, which is quite a bit smaller than the probability we would have got had we stuck with our original intention to collect 10 observations:

5 For a binomial distribution the expected value (the mean) is $\mu = np$. Our intentions changing (that is, collecting 9 observations rather than 10) does not affect p (we still believe the overall rate of liking a profile is 0.4) but it does affect n. So, with 10 observations $\mu = 10 \times 0.4 = 4$, but with 9 observations $\mu = 9 \times 0.4 = 3.6$.

$$
\begin{aligned}
P(8) &= \binom{9}{8} 0.4^8 (1-0.4)^{9-8}, \quad x = 0,1,\ldots,n \\
&= \frac{9!}{8!(9-8)!} 0.4^8 0.6^1 \\
&= \frac{9 \times \cancel{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}}{\cancel{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \times (1)} 0.4^8 0.6 \\
&= 9 \times 0.00065536 \times 0.6 \\
&= 0.0035
\end{aligned} \tag{3.1}
$$

$$
\begin{aligned}
P(9) &= \binom{9}{9} 0.4^9 (1-0.4)^{9-9}, \quad x = 0,1,\ldots,n \\
&= \frac{9!}{9!(9-9)!} 0.4^9 0.6^0 \\
&= 1 \times 0.000262144 \times 1 \\
&= 0.0003
\end{aligned} \tag{3.2}
$$

$$
\begin{aligned}
p(x \geq 8) &= p(8) + p(9) + p(10) \\
&= 0.0035 + 0.0003 + 0 \\
&= 0.0038
\end{aligned} \tag{3.3}
$$

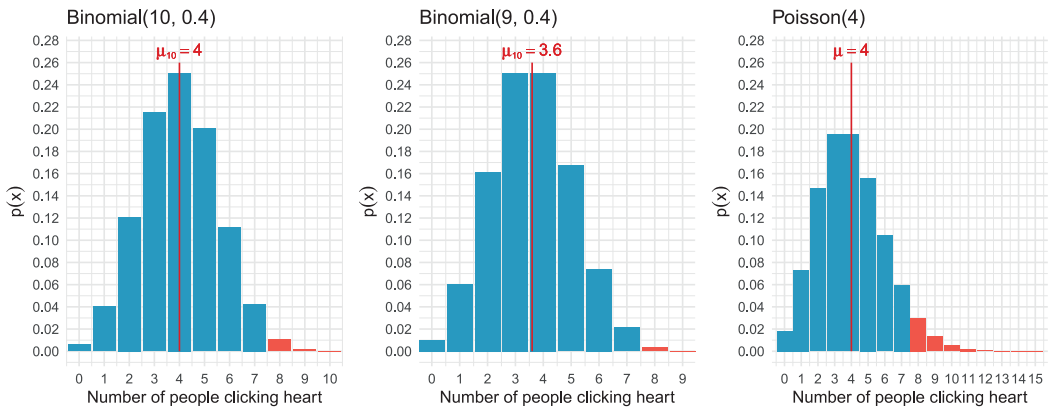


Figure 3.3 Binomial distributions with $n = 10$ (left) and $n = 9$ (middle) and $p = 0.4$, and a Poisson(4) distribution (right)

By changing our intentions and collecting 9 observations instead of 10, we have ended up with a different *p*-value. However, we have stuck with the decision of using a sampling distribution that fixes *N*, albeit we've abused the word 'fixes' by changing our minds about the value of *N*. In fact, we changed *N* because we decided to cut the data collection short after an hour. The binomial distribution is an appropriate model of how many events occur in a fixed number of trials, but it is *not* a good model of how many events occur in a fixed period of *time*. For that, the Poisson distribution is more appropriate. This distribution is defined by a single parameter, known as the rate (λ), which specifies the expected number of events in a given time and is the mean of the distribution. We might consider a Poisson (λ = 4) distribution to be a reasonable model of our new data collection process because its expected value matches that of our original sampling distribution.[6] This distribution

6 The expected value of a Poisson distribution is the same as its parameter $\mu = \lambda$.

is shown in the right panel of Figure 3.3. We can compare it to the other two distributions in that diagram, which we have already considered. Note that the red area (the area representing the *p*-value) is larger than for the other two distributions. The distribution is also flatter than the other two; that is, small and large numbers of likes are more probable in this distribution than in the binomial distributions. Also, for the binomial distributions we were modelling how many likes were received out of *N* views of the profile, so it is only possible to observe values up to and including *N*. We can see this because the *x*-axes are truncated at 10 and 9, respectively. However, the Poisson distribution is modelling the number of likes within an hour and so it is not bounded; it is possible to have more than 10 profile likes in an hour, albeit with decreasing probability. In fact, we could work out the probability of observing as many as 100 or more likes in an hour using this distribution; the probability of such a high number of likes would be zero to many decimal places but not actually identical to zero. We can calculate the probability of a score (x) from a Poisson distribution using the following equation:

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}, \quad x = 0,1,\ldots \tag{3.4}$$

in which μ is the rate parameter (and mean), which in this case is 4, and e is the base of natural logarithms (a constant value approximately equal to 2.718). However, to calculate the red area in Figure 3.3 would be very tedious because, unlike the binomial distribution, the value of x is not constrained to be less than n, so we'd need to calculate p(x) for values of x from 8 to infinity (or at least a large enough value of x to give us a decent approximation). There is an easier way. It turns out that because these distributions represent relative probabilities, the area of the plot sums to 1. This fact makes sense because the probability of x occurring across all potential values of x is 1 (x will certainly be within the range of possible values of x). Given that the area of the distribution is equal to 1, the red and blue areas must sum to 1, so we can calculate the red area by subtracting the blue area from 1. Calculating the blue area still involves calculating the probabilities for $x = 0$ to $x = 8$, but that's a vast improvement on an infinite number of calculations.[7] Strap in. Eq. 3.5 shows the calculation of the probability of observing a value less than or equal to 8, which is the blue area in Figure 3.3 (right panel), and Eq. 3.6 calculates the probability of observing a value at least as large as 8 by subtracting the probability calculated in Eq. 3.5 from 1. The resulting p-value is $p(x \geq 8) = 0.0214$, which is larger than both probabilities computed using the binomial distribution.

$$\begin{aligned}
P(x \leq 8) &= p(0) + p(1) + p(2) + \ldots + p(7) + p(8) \\
&= \frac{e^{-4}4^{0}}{0!} + \frac{e^{-4}4^{1}}{1!} + \frac{e^{-4}4^{2}}{2!} + \ldots + \frac{e^{-4}4^{7}}{7!} + \frac{e^{-4}4^{8}}{8!} \\
&= e^{-4}\left(1 + 4 + \frac{4^{2}}{2!} + \frac{4^{3}}{3!} + \frac{4^{4}}{4!} + \frac{4^{5}}{5!} + \frac{4^{6}}{6!} + \frac{4^{7}}{7!} + \frac{4^{8}}{8!}\right) \\
&= e^{-4}(5 + 8 + 10.667 + 10.667 + 8.533 + 5.689 + 3.251 + 1.625) \\
&= 53.432e^{-4} \\
&= 0.9786
\end{aligned} \tag{3.5}$$

7 For the binomial distribution we included $p(8)$ in the calculation for the *red* area of the plot, so you might wonder why it's included here for the *blue* area of the Poisson distribution. The reason is that the binomial distribution models x as discrete integer values. So $x = 7$ and $x = 8$ are plausible values but $x = 7.9$ is not. Therefore, if we are calculating $P(x \geq 8)$ we want to include 8 in that calculation, but not in the calculation for $P(x < 8)$. The Poisson distribution, however, models x as a continuous variable so $x = 7.999999$ is a plausible value in the distribution (but not in real life as the number of views has to be an integer). Therefore to calculate $P(x < 8)$ we need to include all values of x up to 8 (otherwise, we're missing the part of the distribution between $x = 7$ and $x = 8$).

$$\begin{aligned} P(x \geq 8) &= 1 - p(x < 8) \\ &= 1 - 0.9786 \\ &= 0.0214 \end{aligned} \tag{3.6}$$

All of this is a fairly long-winded way of demonstrating that the *p*-value is affected by the intentions of the experimenter (see Kruschke, 2014, for a more detailed explanation). We started with the same research question and hypotheses about our dating profile, but when we stuck to our original sampling intentions of sampling 10 observations we ended up with $p = 0.0123$, when we collected only 9 observations we got $p = 0.0037$ and when we changed the sampling distribution to model the reason why we collected fewer observations we got $p = 0.0214$. The *p*-value we get and report has been affected by factors unrelated to whether our hypothesis is true.

You might reasonably wonder why we should care about fairly small differences in *p*, and some would argue we shouldn't, but in an ideal world we don't want our metric for evaluating a hypothesis being influenced by something unrelated to that hypothesis. Remember that we set out to collect 10 observations but collected only 9 because we decided to collect data for only 1 hour. In situations where data collection is affected by things like participant availability, researchers tend to fit the same models that they planned to fit but using a different sample size to the one they planned to collect. In our specific example, we want the *p*-value to be the relative frequency of our test statistic compared to all possible t_{null}s from experiments of size 10 but we'd end up comparing it to all possible t_{null}s from experiments of size 9. The space of possible t_{null}s has been influenced by an arbitrary variable such as the length of time during which we collected data (Kruschke, 2013). The *p*-value that we end up with carries information about the hypothesis, but it now also contains information about the time frame for data collection. Also, by modelling the data with the same sampling distribution that we planned to use (binomial) we have modelled the data in a way that does not reflect our data collection plan (because the Poisson distribution provides a better model of a sampling process where data collection stops after a fixed period of time). Of course, there's nothing to stop researchers from switching to a more appropriate model, but they typically don't, and if they do they have introduced a researcher degree of freedom (see Section 3.3.2).

When performing NHST we apply a decision rule to the observed *p*-value. We compare it to a predefined alpha (typically 0.05), which is also the probability of a Type I error. Alpha is an **empirical probability**, which is defined as the proportion of events that have the outcome in which you're interested in an indefinitely large collection of events (Dienes, 2011). For example, if you define the collection as everyone who has ever eaten fermented Icelandic shark, then the empirical probability of retching will be the proportion of people (who have ever eaten fermented Icelandic shark) who retched. The crucial point is that this probability is a *long-run probability*: it applies to the collection and not to the individual events. You can talk about there being a 0.1 probability of retching when someone is about to eat putrefying shark, but the individuals who ate the shark either retched or didn't, so their *individual* probability of retching was either 0 (they didn't retch) or 1 (they retched). Similarly, the Type I error rate is a long-run probability meaning that *across repeated identical experiments* the probability of making a Type I error is 0.05. The probability does not apply to an individual study, in which you either have ($p = 1$) or have not ($p = 0$) made a Type I error. The probability of a Type II error (β) is also a long-run probability. If you set it at 0.2 then over an infinite number of identical experiments you can expect to miss an effect that genuinely exists in 20% of experiments. However, in an individual study, once the data have been collected, the probability is not 0.2, it is either 1 (you have missed the effect) or 0 (you have not missed the effect).

The fact that decisions by the researcher affect the value of *p* potentially affects the probabilities of Type I and Type II errors for the particular study, but not necessarily the long-term probabilities of these errors. For example, if you intended to collect 50 observations and that yielded $p = 0.06$ but you ended up collecting 30 observations and observing $p = 0.04$, then the decision rule would lead you to reject the null when you would have accepted it had to stuck to your sampling plan. Your conclusion flips and so would the Type I and Type II error probabilities *for the individual study*. However, the

probabilities of making a Type I and Type II error over all independent experiments in your research career remain 0.05 and 0.2 respectively if you were to always compute p based on a fixed sample size (even if that sample size differs from your original intention). That might sound reassuring, but it doesn't change the fact that your sampling distribution might not reflect your sampling procedure. For example, when we changed from sampling a fixed number of observations (10 participants) to sampling for a fixed time (1 hour) there was a known alternative distribution that we could use to model this situation (the Poisson distribution). However, what if data collection stops early because of a technical fault with your brain scanner, or what if a school visit to your university gives you an unexpected pool of willing participants for your developmental psychology study? There are no distributions that show us the probability of events when brain scanners fail, or when schoolchildren unexpectedly visit. We simply can't compute p to reflect an idiosyncratic sampling process. These situations might seem trite, but they illustrate the point that if we deviate from our sampling plan we add noise to the p-value.

3.3 NHST as part of wider problems with science Ⓑ



There are direct consequences attached to the problems we have just looked at. For example, the consequences of the misconceptions of NHST are that scientists overestimate the importance of their effects (misconception 1); ignore effects that they falsely believe don't exist because of 'accepting the null' (misconception 2); and pursue effects that they falsely believe exist because of 'rejecting the null' (misconception 3). Given that a lot of science is directed at informing policy and practice, the practical implications could be things like developing treatments that, in reality, have trivial efficacy, or not developing ones that have potential. NHST also plays a role in some wider issues in science.

To understand why, we need a trip to the seedy underbelly of science. Science should be objective, and it should be driven, above all else, by a desire to find out truths about the world. It should not be self-serving, especially not if that gets in the way of the truth. Unfortunately, scientists compete for scarce resources to do their work: research funding, jobs, lab space, participant time and so on. It is easier to get these scarce resources if you are 'successful', and being 'successful' can be tied up with NHST (as we shall see). Also, scientists are not emotionless robots (seriously!) but are people who have, for the most part, spent their whole lives succeeding at academic pursuits (school grades, college grades, etc.). They also work in places where they are surrounded by incredibly intelligent and 'successful' people and they probably don't want to be known as 'Professor Thicko whose experiments never work'. Before you get out the world's tiniest violin to play me a tune, I'm not looking for sympathy, but it does get wearing feeling inferior to people with more grants, better papers and more ingenious research programmes, trust me. What does this have to do with NHST? Let's find out.

3.3.1 Incentive structures and publication bias Ⓑ

Imagine that two scientists, Beth and Danielle, are both interested in the psychology of people with extreme views. If you've spent any time on social media, you're likely to have encountered this sort of scenario: someone expresses an opinion, someone else responds with some unpleasant insult followed by the directly opposing view, and there ensues a pointless and vile exchange where neither person manages to convince the other of their position. Beth and Danielle want to know whether people with extreme views literally can't see the grey areas. They set up a test in which participants see words displayed in various shades of grey and for each word they try to match its colour by clicking along a gradient of greys from near white to near black. Danielle found that politically moderate participants are significantly more accurate (the shades of grey they chose more closely matched the

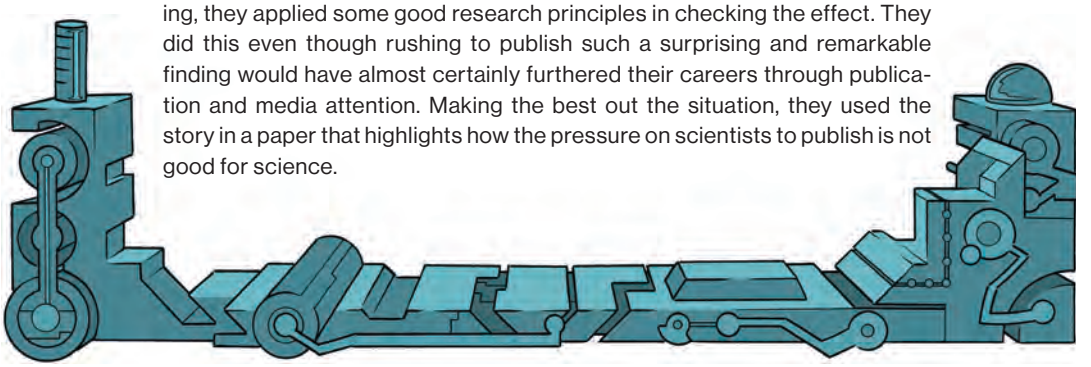
colour of the word) than participants with extreme political views to the left or right. Beth found no significant differences in accuracy between groups with different political views. This is a real study (Labcoat Leni's Real Research 3.1).

What are the consequences for our two scientists? Danielle has an interesting, surprising and media-friendly result, she writes it up and submits it to a journal for publication. Beth does not have a sexy result but being a positive kind of person, and trusting the rigour of her study, she decides to send it to a journal anyway. The chances are that Danielle's article will be published but Beth's will not because significant findings are about seven times more likely to be published than non-significant ones (Coursol & Wagner, 1986). This phenomenon is called **publication bias**. In my own discipline, psychology, over 90% of journal articles report significant results (Fanelli, 2010b, 2012). This bias is driven partly by reviewers and editors rejecting articles with non-significant results (Hedges, 1984), and partly by scientists not submitting articles with non-significant results because they are aware of this editorial bias (Dickersin et al., 1992; Greenwald, 1975).



Labcoat Leni's Real Research 3.1
Researcher degrees of freedom: a sting in the tale

In the main text, I described Beth and Danielle who both conducted a study in which people of differing political views were asked to judge the shade of grey of words. Danielle's study showed a significant effect, Beth's did not. This story is true, but the researchers were called Brian, Jeffrey and Matt (Nosek et al., 2012). The first time they ran the study (N = 1979), like Danielle, they found a significant effect with a p-value of 0.01. Even though the study tested a theory-driven hypothesis, the result surprised them. They had recently read about researcher degrees of freedom and they were aware that their sample size was large enough to detect even a very small effect. So instead of trying to publish the study, they replicated it. In the replication (N = 1300), like Beth's study, the effect was far from significant (p = 0.59). Although the replication doesn't completely rule out a genuine effect, it suggests that no one should get too excited before there is more evidence either way. Nosek and colleagues did the correct thing: rather than rushing to publish their initial finding, they applied some good research principles in checking the effect. They did this even though rushing to publish such a surprising and remarkable finding would have almost certainly furthered their careers through publication and media attention. Making the best out the situation, they used the story in a paper that highlights how the pressure on scientists to publish is not good for science.



To return to Danielle and Beth, assuming they are equal in all other respects, Danielle now has a better-looking track record than Beth: she will be a stronger candidate for jobs, research funding, and internal promotion. Danielle's study was no different than Beth's (except for the results) but doors are now a little more open for her. The effect of a single research paper may not be dramatic, but over time, and over research papers, it can be the difference between a 'successful' career and an 'unsuccessful' one.

The current incentive structures in science are individualistic rather than collective. Individuals are rewarded for 'successful' studies that can be published and can therefore form the basis for funding applications or tenure; 'success' is, therefore, defined largely by results being significant. If a person's career as a scientist depends on significant results, then they might feel pressure to get significant results. In the USA, scientists from institutions that publish a large amount are more likely to publish results supporting their hypothesis (Fanelli, 2010a). Given that a good proportion of hypotheses ought to be wrong, and these wrong hypotheses ought to be distributed across researchers and institutions, Fanelli's work implies that those working in high-stress 'publish or perish' environments have fewer wrong hypotheses. One explanation is that high-stress environments attract better scientists who derive better-quality hypotheses, but an alternative is that the incentive structures in academia encourage people in high-pressure environments to cheat more. Scientists wouldn't do that, though.

3.3.2 Researcher degrees of freedom Ⓑ

As well as reviewers and editors tending to reject articles reporting non-significant results, and scientists not submitting them, there are other ways that scientists contribute to publication bias. The first is by selectively reporting their results to focus on significant findings and exclude non-significant ones. At the extreme, this could entail not including details of other experiments that had results that contradict the significant finding. The second is that researchers might capitalize on **researcher degrees of freedom** (Simmons et al., 2011) to show their results in the most favourable light possible. 'Researcher degrees of freedom' refers to the fact that a scientist has many decisions to make when designing and analysing a study. We have already seen some decisions related to NHST that apply here: the alpha level, the level of power and how many participants should be collected. There are many others, though: which statistical model to fit, how to deal with extreme scores, which control variables to consider, which measures to use, when to stop collecting data and so on. Focusing on analysis alone, when 29 different research teams were asked to answer the same research question (are soccer referees more likely to give red cards to players with dark skin than to those with light skin?) using the same data set, 20 teams found a significant effect, whereas nine did not. There was also wide variety in the analytical models that the teams used to address the question (Silberzahn et al., 2018; Silberzahn & Uhlmann, 2015). These researcher degrees of freedom could be misused to, for example, exclude cases to make the result significant.

Fanelli (2009) assimilated data[8] from 18 published studies containing surveys about questionable research practices and found that for scientists reporting on their own behaviour, across all studies, on average 1.97% admitted fabricating or falsifying data, or altering results to improve the outcome. That's a small amount, but other questionable practices were admitted across studies by 9.54% and there were higher rates for things like dropping observations based on a gut feeling (15.3%), using inappropriate research designs (13.5%), not publishing key findings (12.1%), allowing industry funders to write the first draft of the report (29.6%) and influence when the study is terminated (33.7%). The last two practices are important because companies that fund research often have a conflict of interest in the results and so might spin them, or terminate data collection earlier or later than planned, to suit their agenda. Plans for data collection should be determined before the study, not adjusted during it. The percentages reported by Fanelli will almost certainly be underestimates because

8 Using something called meta-analysis, which you'll discover later in this chapter.

scientists were reporting on their own behaviour, and, therefore, admitting to activities that harm their credibility. Fanelli also assimilated studies in which scientists reported on *other scientists'* behaviour. He found that, on average, 14.12% had observed fabricating or falsifying data, or altering results to improve the outcome, and on average a disturbingly high 28.53% reported other questionable practices. Fanelli details response rates to specific research practices mentioned in the studies he assimilated.[9] It's noteworthy that there were high rates of scientists responding that they were aware of others failing to report contrary data in an article (69.3%), choosing a statistical technique that provided a more favourable outcome (45.8%), reporting only significant results (58.8%) and excluding data based on a gut feeling (20.3%).

Questionable research practices are not necessarily the fault of NHST, but NHST nurtures these temptations by fostering black-and-white thinking in which significant results garner much greater personal rewards than non-significant ones. For example, having put in months of hard work planning a project and collecting data, it's easy to imagine that if your analysis spews out a difficult-to-publish *p*-value of 0.08, it might be tempting to change your analytic decisions to see whether the *p*-value drops below the easier-to-publish 0.05 threshold. Doing so creates scientific noise.

3.3.3 *p*-hacking and HARKing Ⓑ

p-hacking (Simonsohn et al., 2014)[10] and hypothesizing after the results are known or **HARKing** (Kerr, 1998) are researcher degrees of freedom that relate closely to NHST. *p*-hacking refers to researcher degrees of freedoms that lead to the selective reporting of significant *p*-values. It is a broad term that encompasses some of the practices we have already discussed such as trying multiple analyses and reporting only the one that yields significant results, deciding to stop collecting data at a point other than when the predetermined (prior to data collection) sample size is reached, and including (or not) data based on the effect they have on the *p*-value. The term also encompasses practices such as including (or excluding) variables in an analysis based on how those variables affect the *p*-value, measuring multiple *outcomes* or *predictor* variables but reporting only those for which the effects are significant, merging groups of variables or scores to yield significant results, and transforming or otherwise manipulating scores to yield significant *p*-values. HARKing refers to the practice in research articles of presenting a hypothesis that was made *after* data collection as though it were made *before* data collection.

Let's return to Danielle, our researcher interested in whether perceptions of the shade of grey differed across people of different political extremes. Imagine that she had recorded not only the political views of her participants, but also a bunch of other variables about their lifestyle and personality such as what music they like, their openness to new experiences, how tidy they are, various mental health questionnaires measuring different things, their sex, gender identity, sexual orientation, age and so on. You get the picture. At the end of data collection, Danielle has her measure of how accurately participants perceive the colours of the grey words and, let's say, 20 other variables. She does 20 different analyses to see whether each of these 20 variables predicts perception of the colour grey. The only one that significantly predicts it is 'political group', so she reports this effect and doesn't mention the other 19 analyses. This is *p*-hacking. She then tries to explain this finding to herself by looking at the literature on embodiment (the link between the mind and body) and decides that the results might be because the perception of grey is an 'embodiment' of political extremism. When she writes

9 See Table S3 of the article (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005738#s5).

10 Simonsohn et al. coined the phrase '*p*-hacking', but the problem of selective reporting and trying multiple analyses to obtain significance is an old problem that has variously been termed *fishing*, *cherry-picking*, *data snooping*, *data mining*, *data diving*, *selective reporting*, and *significance chasing* (De Groot, 1956; Peirce, 1878). Nevertheless, I enjoy the imagery of a maniacal scientist driven so mad in the pursuit of a *p*-value less than 0.05 that they hack at their computer with a machete.

up the report she pretends that the motivation for the study was to test a hypothesis about embodiment of political views. This is HARKing.

The example above illustrates that HARKing and *p-hacking* are not mutually exclusive: a person might *p*-hack to find a significant result that they then HARK. If you were so bad at comedy that you had to write statistics textbooks for a living, you might call this *p*-harking.

Both *p*-hacking and HARKing are cheating the system of NHST (explained in Section 2.10.3). In both cases, it means that you're not controlling the Type I error rate (because you're deviating from the process that ensures that it is controlled) and therefore you have no idea how many Type I errors you will make in the long run (although it will certainly be more than 5%). You are making it less likely that your findings will replicate. More important, it is morally dubious to put spurious results, or scientific noise, into the public domain. It wastes a lot of people's time and money.

Scientists have looked at whether there is general evidence for *p*-hacking in science. There have been different approaches to the problem but, broadly speaking, they have focused on comparing the distribution of *p*-values you would expect to get if *p*-hacking doesn't happen with the distributions you might expect if it does. To keep things simple, one example is to extract the reported *p*-values across a range of studies on a topic, or within a discipline, and plot their frequency. Figure 3.4 reproduces some of the data from Masicampo and Lalande (2012), who extracted the reported *p*-values from 12 issues of three prominent psychology journals. The line shows a ***p*-curve** which is the number of *p*-values you would expect to get for each value of *p*. In this paper the *p*-curve was derived from the data; it was a statistical summary of *p*-values that they extracted from the articles, and shows that smaller *p*-values are more frequently reported (left of the figure) than larger, non-significant ones (right of the figure). You'd expect this because of the bias towards publishing significant results. You can also compute *p*-curves based on statistical theory, whereby the curve is affected by the effect size and the sample size but has a characteristic shape (Simonsohn et al., 2014). When there is no effect (effect size = 0) the curve is flat – all *p*-values are equally likely. For effect sizes greater than 1, the curve has an exponential shape, which looks similar to curve in Figure 3.4, because smaller *p*s (reflecting more significant results) occur more often than larger *p*s (reflecting less significant results).

The dots in Figure 3.4 are the number of *p*-values reported for each value of *p* that Masicampo and Lalande (2012) extracted from a journal called *Psychological Science*. Note that the dots are generally

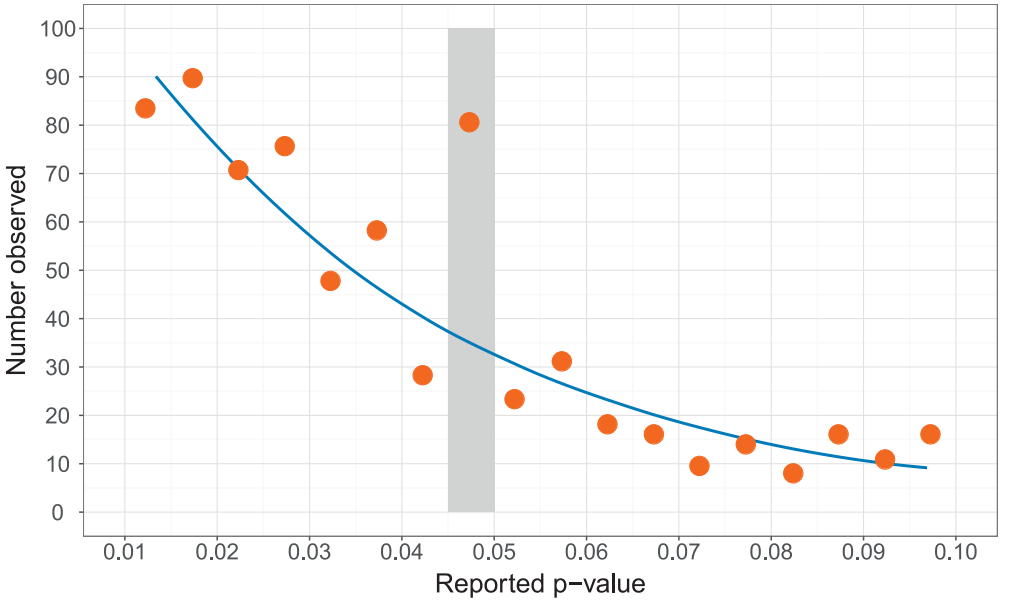


Figure 3.4 Reproduction of the data in Figure 2 from Masicampo and Lalande (2012)

close to the line: that is, the frequency of reported *p*-values matches what the model (the line) predicts. The interesting part of this graph (yes, there is one) is the shaded box, which shows *p*-values just under the 0.05 threshold. The dot in this part of the graph is much higher than the line. This shows that there are a lot more *p*-values just under the 0.05 threshold reported in *Psychological Science* than you would expect based on the curve. Masicampo and Lalande argue that this shows *p*-hacking: researchers are engaging in practices to nudge their *p*-values below the threshold of significance and that's why these values are over-represented in the published literature. Others have used similar analyses to replicate this finding in psychology (Leggett et al., 2013) and science more generally (Head et al., 2015). The analysis of *p*-curves looks for *p*-hacking by examining any *p*-value reported within and across the articles studied. Often it is an automated process using a computer to search journal articles and extract anything that looks like a *p*-value.

Another approach is to focus on *p*-values in individual studies that report multiple experiments. The logic here is that if a scientist reports, say, four studies that examine the same effect, then you can work out, based on the size of effect being measured and sample size of the studies, what the probability is that you would get a significant result in all four studies. A low probability means that it is highly unlikely that the researcher would get these results: they are 'too good to be true' (Francis, 2013), the implication being that *p*-hacking or other dubious behaviour may have occurred. Research using these **tests of excess success** (TES) also finds evidence for *p*-hacking in psychology, epigenetics and science generally (Francis, 2014b, 2014a; Francis et al., 2014).

All of this paints a pretty bleak picture of science, but it also illustrates the importance of statistical literacy and a critical mind. Science is only as good as the people who do it, and what you will learn in this book (and others) more than anything else are the skills to evaluate scientific results and data for yourself. Having thoroughly depressed you, it's worth ending this section on a cheery note. I was tempted to tell you a tale of a small army of puppies found bouncing on their owner's bed to the sound of 'The chase is better than the cat' by Motörhead. While you process that image, I'll mention that some of the work that seems to suggest rampant *p*-hacking is not without its problems. For example, Lakens (2015) has shown that the *p*-curves used by Masicampo and Lalande probably don't resemble what you would actually see if *p*-hacking were taking place; when you use *p*-curves that better resemble *p*-hacking behaviour, the evidence vanishes. Hartgerink et al. (2016) noted that *p*-values are often misreported in articles, and that studies of *p*-hacking typically look for a 'bump' just below the 0.05 threshold (as shown Figure 3.4) rather than more subtle forms of excess at the 0.05 threshold. Looking at 258,050 tests from 30,701 psychology research articles, Hartgerink et al. found evidence for a 'bump' in only three of the eight journals that they looked at. They found widespread rounding errors in the *p*-values reported and when they corrected these errors the 'bump' remained in only one of the eight journals examined. Perhaps their most important conclusion was that it was very difficult to extract enough information (even in this huge study) to model accurately what *p*-hacking might look like, and therefore ascertain whether it had happened. This sentiment is echoed by other researchers, who have pointed out that the shape of the *p*-curve may indicate very little about whether *p*-hacking has occurred (Bishop & Thompson, 2016; Bruns & Ioannidis, 2016). There is also an argument that TES aren't useful because they don't test what they set out to test (e.g., Morey, 2013; Vandekerckhove et al., 2013).

You might reasonably wonder why I've wasted your time convincing you that scientists are all duplicitous, self-interested, nefarious hacks only to then say 'actually, they might not be'. If you spend enough time in science you'll realize that this is what it's like: as soon as you believe one thing, someone will come along and change your mind. That's OK though, and remaining open-minded is good because it helps you not to get sucked into dubious research behaviours. Given that you will be the future of science, what matters is not what has happened until now, but what you go out into the world and do in the name of science. To that end, I hope to have convinced you not to *p*-hack, or HARK, or fudge your data and so on. You are better than that, and science will be better for it.



Cramming Sam's Tips 3.1
Problems with NHST

- A lot of scientists misunderstand NHST. A few examples of poorly understood things related to significance testing are:
  - A significant effect is not necessarily an important one.
  - A non-significant result does not mean that the null hypothesis is true.
  - A significant result does not mean that the null hypothesis is false.
- NHST encourages all-or-nothing thinking whereby an effect with a *p*-value just below 0.05 is perceived as important whereas one with a *p*-value just above 0.05 is perceived as unimportant.
- NHST is biased by researchers deviating from their initial sampling frame (e.g., by stopping data collection earlier than planned).
- There are lots of ways that scientists can influence the *p*-value. These are known as *researcher degrees of freedom* and include selective exclusion of data, fitting different statistical models but reporting only the one with the most favourable results, stopping data collection at a point other than that decided at the study's conception, and including only control variables that influence the *p*-value in a favourable way.
- Incentive structures in science that reward publication of significant results also reward the use of researcher degrees of freedom.
- *p*-hacking refers to practices that lead to the selective reporting of significant *p*-values, most commonly trying multiple analyses and reporting only the one that yields significant results.
- Hypothesizing after the results are known (HARKing) occurs when scientists present a hypothesis that was made after data analysis as though it were made at the study's conception.



3.4 A phoenix from the EMBERS Ⓑ

The pitfalls of NHST and the way that scientists use it have led to a shift in the pervasive view of how to evaluate hypotheses. It's not quite the paradigm shift to which I alluded at the start of this chapter, but certainly the tide is turning. The ubiquity of NHST is strange given that its problems have been acknowledged for decades (e.g., Rozeboom, 1960). In a post on a discussion board of the American Statistical Association (ASA), George Cobb noted a circularity: the reason why so many colleges and graduate schools still teach $p = 0.05$ is that it's what the scientific community uses, and the reason why the scientific community still (predominantly) uses $p = 0.05$ is that it's what they were taught in college or graduate school (Wasserstein & Lazar, 2016). This cycle of habit probably extends to teaching

too, in that statistics teachers will tend to teach what they know, which will be what they were taught, which is typically NHST. Therefore, their students learn NHST too, and those that go on to teach statistics will pass NHST down to their students. It doesn't help that because most scientists are taught NHST and that's all they know, they use it in their research papers, which means that lecturers probably *must* teach it so that their students can understand research papers. NHST is a hard habit to break because it requires a lot of people making the effort to broaden their statistical horizons.

Things are changing, though. In my discipline (psychology), an American Psychological Association (APA) task force produced guidelines for the reporting of data in its journals. This report acknowledged the limitations of NHST while appreciating that a change in practice would not happen quickly; consequently, it didn't recommend against NHST but suggested instead that scientists report things like confidence intervals and effect sizes to help them (and readers) evaluate the research findings without dogmatic reliance on *p*-values (Wilkinson & Task Force on Statistical Inference, 1999). At the extreme, in 2015 the journal *Basic and Applied Social Psychology* banned *p*-values from its articles, but that was silly because it throws the baby out with the bathwater. The ASA has published some guidelines about *p*-values (Wasserstein & American Statistical Association, 2016), which we'll get to shortly. In the following sections, we'll explore some of the ways you might address the problems with NHST. It will be like a statistical phoenix rising from the EMBERS of significant testing (but not in that order):

- **E**ffect sizes
- **M**eta-analysis
- **B**ayesian **E**stimation
- **R**egistration
- **S**ense

3.5 Sense, and how to use it Ⓑ

It is easy to blame NHST for the evils in the scientific world, but some of the problem is not the process so much as how people misuse and misunderstand it. As we have seen (Jane Superbrain Box 3.1), Fisher never encouraged people to brainlessly use a 0.05 level of significance and throw a party for effects with *p*-values below that threshold while hiding away those with *p*s above the threshold in a dungeon of shame. Nor did anyone suggest that we evolve a set of incentive structures in science that reward dishonesty and encourage the use of researcher degrees of freedom. You *can* blame NHST for being so difficult to understand, but you can also blame scientists for not working harder to understand it. The first way to combat the problems of NHST is, therefore, to apply sense when you use it. A statement by the ASA on *p*-values (Wasserstein & American Statistical Association, 2016) offers a set of six principles for scientists using NHST.

1. The ASA points out that *p*-values *can* indicate how incompatible the data are with a specified statistical model. In the case of what we've been discussing this means we can use the value of *p* (its exact value, not whether it is above or below an arbitrary threshold) to indicate how incompatible the data are with the null hypothesis. Smaller *p*s indicate more incompatibility with the null. You are at liberty to use the degree of incompatibility to inform your own beliefs about the relative plausibility of the null and alternative hypotheses, as long as …
2. …you don't interpret *p*-values as a measure of the probability that the hypothesis in question is true. They are also not the probability that the data were produced by random chance alone.
3. Scientific conclusions and policy decisions *should not* be based only on whether a *p*-value passes a specific threshold. Basically, resist the black-and-white thinking that *p*-values encourage.
4. Don't *p*-hack. The ASA says that '*p*-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values

(typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable.' Be fully transparent about the number of hypotheses explored during the study, and all data collection decisions and statistical analyses.

5 Don't confuse statistical significance with practical importance. A *p*-value does not measure the size of an effect and is influenced by the sample size, so you should never interpret a *p*-value in any way that implies that it quantifies the size or importance of an effect.

6 Finally, the ASA notes that 'By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.' In other words, even if your small *p*-value suggests that your data are compatible with the alternative hypothesis or your large *p*-value suggests that your data are compatible with the null, there may be many other hypotheses (untested and perhaps not conceived of) that are more compatible with the data than the hypotheses you have tested.

If you have these principles in mind when you use NHST then you will avoid some of the common pitfalls when interpreting and communicating your research.

3.6 Preregistering research and open science Ⓐ

The ASA's fourth principle is a call for transparency in what is reported. This notion of transparency is gaining considerable momentum because it is a simple and effective way to protect against some of the pitfalls of NHST. There is a need for transparency in scientific intentions both before the study is conducted and after it is complete (whether that is through sharing your data or stating clearly any deviations from your intended protocol). These goals are encapsulated by the term **open science**, which refers to a movement to make the process, data and outcomes of research freely available to everyone. Part of this movement is about providing free access to scientific journals, which have traditionally been available only to individuals and institutions who pay to subscribe. The part most relevant to our discussion, though, is **preregistration** of research, which refers to the practice of making all aspects of your research process (rationale, hypotheses, design, data processing strategy, data analysis strategy) publicly available before data collection begins. This can be done in a **registered report** in an academic journal (e.g., Chambers et al., 2015; Nosek & Lakens, 2014), or more informally (e.g., on a public website such as the Open Science Framework). A formal registered report is a submission to an academic journal that outlines an intended research protocol (rationale, hypotheses, design, data processing strategy, data analysis strategy) before data are collected. The submission is reviewed by relevant experts just as a submission of a completed study would be (Section 1.8). If the protocol is deemed to be rigorous enough and the research question novel enough, the protocol is accepted by the journal typically with a guarantee to publish the findings no matter what they are.

Other initiatives that facilitate open science principles include the **Peer Reviewers' Openness Initiative** (Morey et al., 2016), which asks scientists to commit to the principles of open science when they act as expert reviewers for journals. Signing up is a pledge to review submissions only if the data, stimuli, materials, analysis scripts and so on are made publicly available (unless there is a good reason, such as a legal requirement, not to). Another, the Transparency and Openness Promotion (TOP) guidelines, is a set of standards for open science principles that can be applied to journals. The eight standards cover citations (1), preregistration of study and analysis protocols (2 and 3), replication transparency with data, analysis scripts, design and analysis plans, and research materials (4–7) and replication (8). For each standard, there are levels defined from 0 to 3. For example, in data transparency level 0 is defined as the journal merely encouraging data-sharing or saying nothing, and level 3 is data posted on a trusted repository and results reproduced independently before publication (Nosek et al., 2015). Using these standards, a journal's level of commitment to open science can be 'badged'. As junior scientists you can contribute by aiming to prioritize 'open' materials (questionnaires, stimuli, etc.) over equivalent proprietary ones, and you can make your own materials available when you publish your own work.

Open science practices help to combat many of the wider problems into which NHST feeds. For example, preregistering research encourages adherence to an agreed protocol, thus discouraging the misuse of researcher degrees of freedom. Better still, if the protocol has been reviewed by experts then the scientist gets useful feedback *before* collecting data (rather than after, by which time it is too late to change things). This feedback can be on the methods and design, but also on the analytic plan, which ought to improve the study and in turn science. Also, by guaranteeing publication of the results – no matter what they are – registered reports should reduce publication bias and discourage questionable research practices aimed at nudging *p*-values below the 0.05 threshold. Of course, by having a public record of the planned analysis strategy, deviations from it will be transparent. In other words, *p*-hacking and HARKing will be discouraged. Data being in the public domain also makes it possible to check for researcher degrees of freedom that may have influenced the results.

None of these developments change the flaws inherent in NHST but they tighten up the way that it is used, and the degree to which scientists can conceal misuses of it. Finally, by promoting methodological rigour and theoretical importance above the results themselves (Nosek & Lakens, 2014), open science initiatives push incentive structures towards quality not quantity, documenting research rather than 'story telling' and collaboration rather than competition. Individualists need not fear either because there is evidence that practising open science principles is associated with higher citation rates of your papers, more media exposure and better-quality feedback on your work from experts (McKiernan et al., 2016). It's a win–win.

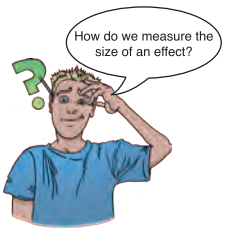
3.7 Effect sizes Ⓑ

One of the problems we identified with NHST was that significance does not tell us about the importance of an effect. The solution to this criticism is to measure the size of the effect that we're testing in a standardized way. When we measure the size of an effect (be that an experimental manipulation or the strength of a relationship between variables) it is known as an **effect size**. Much has been written about how to define the term 'effect size' (Kelley & Preacher, 2012), but put simply it is a measure of the magnitude and direction of an observed effect. Effect sizes add information that you don't get from a *p*-value, so reporting them is a habit well worth getting into. Broadly speaking there are two types of effect size: raw and standardized. We'll start by discussing that distinction before looking at the theory behind three of the most commonly used effect sizes: Cohen's $\hat{d}$; Pearson's correlation coefficient, *r*; and the odds ratio. We'll look at how to compute these effect sizes in relevant sections later on in the book.

3.7.1 Raw effect sizes Ⓐ

A raw effect size is the observed effect expressed in the original (or raw) units of measurement. As we progress through this book you will see that every statistical model we come across is a variation of a linear model (see Chapter 9). We will discover that within this model we have a variable that we're trying to predict (an outcome variable) and variables that we are using to predict it (predictor variables). Within the linear model every predictor variable has a value *b* attached to it. This value is a raw effect size. It expresses the relationship between a predictor variable and an outcome variable in the original (raw) units of measurement. As we will see, this *b* can represent both relationships between variables (Chapter 9) and differences between mean outcome scores (Chapter 10). There's a lot to be said for raw effect sizes: they are directly interpretable and easy to understand (because they are in the original units of measurement).

Think back to our example of singing as a mating strategy (Section 2.10.11). Remember we had 10 men singing to women, and a different 10 starting conversations with them. The outcome was how



long it took before the woman ran away.[11] Remember that the singing group had a mean of 10 and a standard deviation of 3, and the conversation group had a mean of 12 and a standard deviation of 3. If we wanted to quantify the effect between the singing and conversation groups, the raw effect size would be the difference between means. The conversation group had a mean of 12 minutes (before the recipient ran away), and the singing group a mean of 10 minutes. So, the effect of singing compared to conversation is 10 – 12 = –2 minutes. This is the raw effect size. Singing had a detrimental effect on how long the recipient stayed, by –2 minutes. The advantage of this measure is that it is expressed in the units of measurement for the outcome variable. In this example, this is convenient because 'minutes' mean something to us: we can all imagine what an extra 2 minutes of time spent with someone would be like. We can also have an idea of what 2 minutes with someone is relative to the amount of time we usually spend talking to random people. The question is whether this value is meaningful within the context of the research. Within the context that, on average, people who have conversations talk for 12 minutes, 2 minutes fewer (when the person was singing) is not inconsequential. The simplicity of raw effect sizes and the fact they are interpretable because they relate directly to the variables that they quantify makes them (arguably) the most useful way to quantify effects (Baguley, 2009). However, scientists love a standardized effect size.

3.7.2 Standardized effect sizes Ⓐ

Standardized effect sizes are raw effect sizes that have been converted into a standard unit of measurement. Instead of expressing the effect in the original units of measurement, it is expressed in these standardized units (usually standard deviations). Scientists like this idea because 'standardized' effect sizes can be compared across different studiesW that have measured different variables, or have used different scales of measurement. For example, an effect size based on reaction time in milliseconds could be compared to an effect size based on heart rates. The downside is that standardized effect sizes can be difficult to interpret because most of us don't have an intuitive grasp of what it means to have an effect size of, for example, half a standard deviation. What does that mean in the real world? In converting to standardized units we (often) lose an intuitive grasp of what the effect means. In the absence of a direct understanding of what the effect means, people (understandably) try to find meaning by adopting pointless benchmarks for whether an effect has meaning (more on this later). Nevertheless, standardized effect sizes are common, and can be useful.

3.7.3 Cohen's *d* Ⓑ

Let's go back to our singing as a dating strategy example. I mentioned that one of the strengths of the raw effect size was that it is directly interpretable. I noted that '2 minutes' less time had intrinsic meaning – we can all relate to what 2 minutes less of conversation would be like. However, if we'd measured what the recipient thought of the singer rather than how much time they spent with them, then interpretation is trickier: 2 units of 'thought' or 'positivity' or whatever is less tangible to us than 2 minutes of time. So, the strength of raw effect sizes can also be a weakness. It depends a bit on the units of measurement. A second inconvenience is that although the difference between means gives us an indication of the 'signal', it does not tell us about the 'noise' in the measure. Is 2 minutes of time a lot relative to the 'normal' amount of time spent talking to strangers?

11 This fictional study describes males trying to attract females by singing because the idea came from replicating, in humans, research showing that male mice 'sing' to female mice to attract them (Hoffmann et al., 2012). However, adapt the examples to whatever sex-pairings best reflect your life.

We can adjust for both limitations by standardizing the effect. We saw in Chapter 1 that the standard deviation is a measure of 'error' or 'noise' in the data, and we saw in Section 1.7.6 that if we divide by the standard deviation then the result is a score expressed in standard deviation units (i.e., a *z*-score). Therefore, if we divide the difference between means by the standard deviation we get a signal-to-noise ratio but we also get a value that is expressed in standard deviation units (and can, therefore, be compared in different studies that used different outcome measures). What I have just described is **Cohen's $\hat{d}$**, and we can express it formally as:

$$\hat{d} = \frac{\bar{X}_1 - \bar{X}_2}{s} \tag{3.7}$$

I have put a hat on the d to remind us that we're interested in the effect size in the population, but because we can't measure that directly, we estimate it from the sample.[12] We've seen these hats before; they mean 'estimate of'. Therefore, $\hat{d}$ is the difference between means divided by the standard deviation. However, we had two means and, therefore, two standard deviations, so which one should we use? Sometimes we assume that group variances (and therefore standard deviations) are equal (see Chapter 6) and under this assumption we can pick a standard deviation from either group because it won't matter much. In our example of singing for a date, the standard deviations were identical in the two groups (SD = 3) so whichever one we pick, we get:

$$\frac{\bar{X}_{\text{Singing}} - \bar{X}_{\text{Conversation}}}{s} = \frac{10 - 12}{3} = -0.667 \tag{3.8}$$

This effect size means that if a person sang rather than having a normal conversation, the time the recipient stayed was reduced by 0.667 standard deviations. That's quite a bit.

Cohen (1988, 1992) made some widely used suggestions about what constitutes a large or small effect: $\hat{d}$ = 0.2 (small), 0.5 (medium) and 0.8 (large). For our singing data this would mean we have a medium to large effect size. However, as Cohen acknowledged, these benchmarks encourage the kind of lazy thinking that we were trying to avoid and ignore the context of the effect such as the measurement instruments and norms within the research area. Lenth put it nicely when he said that when we interpret effect sizes we're not trying to sell T-shirts: 'I'll have the Metallica tour effect size in a medium, please' (Baguley, 2004; Lenth, 2001).

When groups do not have equal standard deviations, there are two common options. First, use the standard deviation of the control group or baseline. This option makes sense because any intervention or experimental manipulation might be expected to change not just the mean but also the spread of scores. Therefore, the control group/baseline standard deviation will be a more accurate estimate of the 'natural' standard deviation for the measure you're using. In our singing study, we would use the conversation group standard deviation because you wouldn't normally go up to someone and start singing. Therefore, d would represent the amount of time less that a person spent with someone who sang at them compared to someone who talked to them relative to the normal variation in time that people spend with strangers who talk to them.

The second option is to pool the standard deviations of the two groups using (if your groups are independent) this equation:

12 The value for the population is expressed as:

$$d = \frac{\bar{\mu}_1 - \bar{\mu}_2}{\sigma}$$

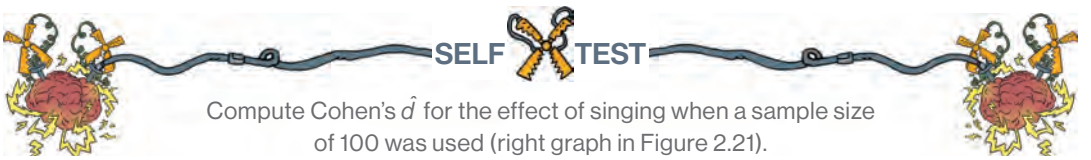
It's the same equation, but because we're dealing with population values rather than sample values, the hat over the d is gone, the means are expressed with μ and the standard deviation with σ. You will more frequently see people refer to d without its hat because they are using a different convention of expressing population values with Greek letters (in this case δ) and estimates from samples as Latin letters (d). Personally, I think this other convention requires more mental effort, so I tend to use hats: anything with a hat on is an estimate.

$$s_p = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}} \tag{3.9}$$

in which N is the sample size of each group and s is the standard deviation. For the singing data, because the standard deviations and sample sizes are the same in the two groups this pooled estimate will be 3, the same as the standard deviation:

$$s_p = \sqrt{\frac{(10 - 1)3^2 + (10 - 1)3^2}{10 + 10 - 2}} = \sqrt{\frac{81 + 81}{18}} = \sqrt{9} = 3 \tag{3.10}$$

When the group standard deviations are different, this pooled estimate can be useful; however, it changes the meaning of $\hat{d}$ because we're now comparing the difference between means against all the background 'noise' in the measure, not just the noise that you would expect to find in normal circumstances.



Compute Cohen's $\hat{d}$ for the effect of singing when a sample size of 100 was used (right graph in Figure 2.21).

If you did the self-test you should have got the same result as before: –0.667. That's because the difference in sample size did not affect the means or standard deviations, and therefore will not affect the effect size. Other things being equal, effect sizes are not affected by sample size, unlike *p*-values. Therefore, by using effect sizes we overcome one of the major problems with NHST. The situation is more complex because, like any parameter, you will get better estimates of the population value in large samples than small ones. So, although the sample size doesn't affect the computation of your effect size in the sample, it does affect how closely the sample effect size matches that of the population (the *precision*).



Compute Cohen's $\hat{d}$ for the effect in Figure 2.22. The exact mean of the singing group was 10, and for the conversation group was 10.01. In both groups the standard deviation was 3.

If you did the self-test then you will have found that the effect size for our larger study was $\hat{d}$ = –0.003. In other words, very small indeed. Remember that when we looked at *p*-values, this very small effect was deemed statistically significant.



Look at Figure 2.21 and Figure 2.22. Compare what we concluded about these three data sets based on *p*-values with what we conclude using effect sizes.

When we looked at the data sets in Figure 2.21 and Figure 2.22 and their corresponding *p*-values, we concluded:

- Figure 2.21: Two experiments with identical means and standard deviations yield completely opposite conclusions when using a *p*-value to interpret them (the study based on 10 scores per group was not significant but the study based on 100 scores per group was).
- Figure 2.22: Two virtually identical means are deemed to be significantly different based on a *p*-value.

However, using effect sizes to guide our interpretations, we conclude:

- Figure 2.21: Two experiments with identical means and standard deviations yield identical conclusions when using an effect size to interpret them (both studies had $\hat{d} = -0.667$).
- Figure 2.22: Two virtually identical means are deemed to be not very different at all based on an effect size ($\hat{d} = -0.003$, which is tiny).

To sum up, effect sizes offer us something potentially less misleading than NHST.

3.7.4 Pearson's *r* Ⓑ

Let's move onto Pearson's correlation coefficient, *r*.[13] We'll cover this statistic in Chapter 8. For now, all you need to know is that it is a measure of the strength of a relationship between two continuous variables, or between one continuous variable and a categorical variable containing two categories. Like Cohen's $\hat{d}$ it is a standardized effect size; unlike it, *r* is constrained to vary from –1 (a perfect negative relationship) through 0 (no relationship) to +1 (a perfect positive relationship).

Imagine we continued our interest in whether singing was an effective dating strategy. This time, however, we focused not on whether the person sang or not, but instead on whether the duration of the performance mattered. Rather than having two groups (singing versus conversation), all the participants in our study sang a song, but for different durations ranging from a 1-minute edit to a 10-minute extended remix. Once the performance had stopped, we timed how long the lucky recipient hung around for a chat. Figure 3.5 shows five different outcomes of this study. The top left and middle panels show that the longer the song was, the less time the recipient stayed to chat. This is called a negative relationship: as one variable increases, the other decreases. A perfect negative relationship (top left) has $r = -1$ and means that as you increase the length of the song, you decrease the subsequent conversation by a proportionate amount. A slightly smaller negative relationship (top middle, $r = -0.5$) means that an increase in the duration of the song also decreases the subsequent conversation, but by a smaller amount. A positive correlation coefficient (bottom row) shows the opposite trend: as the song increases in length so does the subsequent conversation. In general terms, as one variable increases, so does the other. If the positive relationship is perfect ($r = 1$, bottom left), then the increases are by a proportionate amount, whereas values less than 1 reflect increases in one variable that are not equivalent to the increase in the other. A correlation coefficient of 0 (top right) shows a situation where there is no relationship at all: as the length of the song increases there is no fluctuation whatsoever in the duration of the subsequent conversation.

Another point to note from Figure 3.5 is that the strength of the correlation reflects how tightly packed the observations are around the model (straight line) that summarizes the relationship between variables. With a perfect relationship ($r = -1$ or 1) the observed data basically fall exactly on

13 I'm going to break with the convention of using hats for estimates when referring to *r* because although it is an estimate of the population effect size (denoted ρ), I don't think I've ever seen anyone use $\hat{r}$. However, don't be fooled by the lack of hat, *r* is an estimate of the population effect size.

the line (the model is a perfect fit to the data), but for a weaker relationship (r = –0.5 or 0.5) the observed data are scattered more widely around the line.

Although the correlation coefficient is typically known as a measure of the relationship between two continuous variables (as we have just described), it can also be used to quantify the difference in means between two groups. Remember that r quantifies the relationship between two variables; it turns out that if one of those variables is a categorical variable that represents two groups where one group is coded with a 0 and the other is coded with a 1, then what you get is a standardized measure of the difference between two means (a bit like Cohen's $\hat{d}$). I'll explain this in Chapter 8, but for now trust me.

Like with $\hat{d}$, Cohen (1988, 1992) suggested some 'T-shirt sizes' for r:

- ***r* = 0.10 (small effect)**: In this case the effect explains 1% of the total variance. (You can convert r to the proportion of variance by squaring it – see Section 8.4.2.)
- ***r* = 0.30 (medium effect)**: The effect accounts for 9% of the total variance.
- ***r* = 0.50 (large effect)**: The effect accounts for 25% of the variance.

It's worth bearing in mind that r is not measured on a linear scale, so an effect with r = 0.6 isn't twice as big as one with r = 0.3. Also, as with $\hat{d}$, although it's tempting to wheel out these 'canned' effect sizes when you can't be bothered to think properly about your data, you ought to evaluate an effect size within the context of your specific research question.

There are many reasons to like r as an effect size measure, one of them being that it is constrained to lie between 0 (no effect) and 1 (a perfect effect).[14] However, there are situations in which $\hat{d}$ may be favoured; for example, when group sizes are very discrepant r can be quite biased compared to $\hat{d}$ (McGrath & Meyer, 2006).

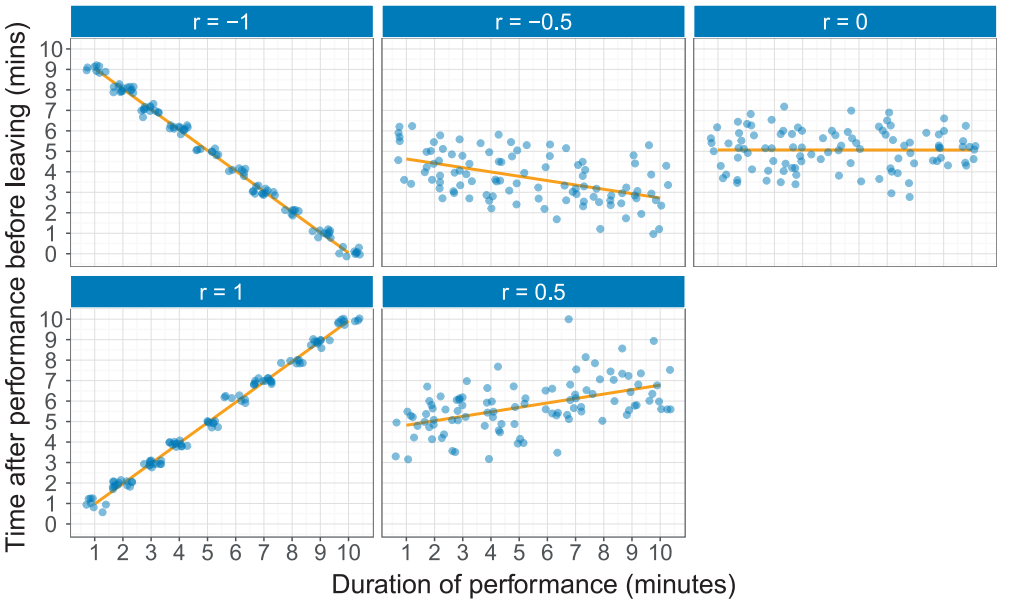


Figure 3.5 Different relationships shown by different correlation coefficients

14 The correlation coefficient can also be negative (but not below –1), which is useful because the sign of r tells us about the direction of the relationship, but when quantifying group differences the sign of r merely reflects the way in which the groups were coded (see Chapter 9).

3.7.5 The odds ratio Ⓑ

The final effect size we'll look at is the **odds ratio**, which is a popular effect size that uses raw frequencies or counts. Imagine a final scenario for our dating research in which we had groups of people who either sang a song or started up a conversation (like our Cohen's $\hat{d}$ example). However, this time, the outcome was not the length of time before the recipient ran away, instead at the end of the song or conversation the recipient was asked 'would you go on a date with me?' and their response ('yes' or 'no') recorded.

Here we have two categorical variables (singing versus conversation and date versus no date) and the outcome is a count (the number of recipients in each combination of those categories). Table 3.1 summarizes the data in a 2×2 **contingency table**, which is a table representing the cross-classification of two or more *categorical variables*. The levels of each variable are arranged in a grid, and the number of observations falling into each category is contained in the cells of the table. In this example, we see that the categorical variable of whether someone sang or started a conversation is represented by rows of the table, and the categorical variable for the response to the question about a date is represented by columns. This creates four cells representing the combinations of those two variables (singing–yes, singing–no, conversation–yes, conversation–no) and the numbers in these cells are the frequencies of responses in each combination of categories. From the table we can see that 12 recipients said 'yes' to a date after they were sung to compared to 26 after a conversation. Eighty-eight recipients said 'no' after singing compared to 74 after a conversation. Looking at the column and row totals, we can see that there were 200 date-pairs in the study; in 100 of them the strategy was to sing, whereas in the remainder, it was a conversation. In total, 38 recipients agreed to a date and 162 did not.

Table 3.1 A contingency table of some fabricated data about singing and dates

		Response to date question		
		Yes	No	Total
Dating behaviour	Singing	12	88	100
	Conversation	26	74	100
	Total	38	162	200

To understand these counts, we might ask how much more likely a person was to say 'yes' to a singer than to someone who starts a conversation. To quantify this effect, we need to calculate some odds. The **odds** of an event occurring are defined as the probability of an event occurring divided by the probability of that event not occurring:

$$\text{odds} = \frac{P(\text{event})}{P(\text{no event})} \tag{3.11}$$

To begin with, we want to know the odds of a 'yes' response to singing, which will be the probability of 'yes' response to a singer divided by the probability of a 'no' response to a singer. The probability of a 'yes' response for singers is the number of yes responses divided by the total number of singers, and the probability of a 'no' response for singers is the number of no responses divided by the total number of singers:

$$\begin{aligned} \text{Odds}_{\text{yes to a singer}} &= \frac{P(\text{yes to a singer})}{P(\text{no to a singer})} \\ &= \frac{\text{Yes responses to singers} / \text{number of singers}}{\text{No responses to singers} / \text{number of singers}} \\ &= \frac{12/100}{88/100} \end{aligned} \tag{3.12}$$

To get both probabilities you divide by the total number of singers (in this case 100), so the result of Eq. 3.12 is the same as dividing the number of 'yes' responses to a singer by the corresponding number of 'no' responses, which gives us 0.14:

$$\begin{aligned}\text{Odds}_{\text{yes to a singer}} &= \frac{\text{Number of yes responses to a singer}}{\text{Number of no responses to a singers}} \\ &= \frac{12}{88} \\ &= 0.14\end{aligned} \tag{3.13}$$

Next, we want to know the odds of a 'yes' response if the person started a conversation ('talkers'). We compute this in the same way: it is, the *probability* of 'yes' response divided by the *probability* of a 'no' response for talkers, which is the same as dividing the *number* of 'yes' responses to talkers by the corresponding *number* of 'no' responses:

$$\begin{aligned}\text{Odds}_{\text{yes to a talker}} &= \frac{\text{Number of yes responses to a talker}}{\text{Number of no responses to a talker}} \\ &= \frac{26}{74} \\ &= 0.35\end{aligned} \tag{3.14}$$

Finally, the odds ratio is the odds of a yes to a singer divided by the odds of a yes to a talker:

$$\begin{aligned}\text{Odds Ratio} &= \frac{\text{Odds}_{\text{yes to a singer}}}{\text{Odds}_{\text{yes to a talker}}} \\ &= \frac{0.14}{0.35} \\ &= 0.4\end{aligned} \tag{3.15}$$

This ratio tells us that the odds of a 'yes' response were 0.4 times more likely to a singer than to someone who started a conversation. If the odds ratio is 1 then it means that the odds of one outcome are equally as likely as the odds of the other. Because 0.4 is less than 1 we know that the odds of a 'yes' response after singing are *worse* than the odds after a conversation. We can flip this odds ratio on its head to see the relative odds of a 'yes' after a conversation compared to after singing by dividing 1 by the odds ratio. Doing so gives us 1/0.4 = 2.5. We can state the same relationship then by saying that the odds of a 'yes' response to a person who started a conversation were 2.5 times more likely than to one who sang.

Once you understand what odds are, the odds ratio becomes a very intuitive way to quantify the effect. In this case, if you were looking for love, and someone told you that the odds of a date after a conversation were 2.5 times those after singing, then you probably wouldn't need to finish this book to know it would be wise to keep your Justin Bieber impersonation, no matter how good it is, to yourself. It's always good to keep your Justin Bieber impersonation to yourself.

3.7.6 Effect sizes compared to NHST Ⓑ

Effect sizes overcome many of the problems associated with NHST:

- They encourage interpreting effects on a continuum and not applying a categorical decision rule such as 'significant' or 'not significant'. This is especially true if you ignore the 'canned effect size' benchmarks.
- Effect sizes are affected by sample size (larger samples yield better estimates of the population effect size), but, unlike *p*-values, there is no decision rule attached to effect sizes (see above) so the interpretation of effect sizes is not confounded by sample size (although it is important in

contextualizing the degree to which an effect size might represent the population). Because of this effect sizes are less affected than *p*-values by things like early or late termination of data collection, or sampling over a time period rather than until a set sample size is reached.

- Of course, there are still some researcher degrees of freedom (not related to sample size) that researchers could use to maximize (or minimize) effect sizes, but there is less incentive to do so because effect sizes are not tied to a decision rule in which effects either side of a certain threshold have qualitatively opposite interpretations.

3.7.7 Meta-analysis Ⓑ

I have alluded many times to how scientists often test similar theories and hypotheses. An important part of science is replicating results, and it is rare that a single study gives a definitive answer to a scientific question. In Section 3.2.2 we looked at an example of 10 experiments that had all explored whether a potion called *antiSTATic* reduces statistics anxiety compared to a placebo (water). The summary of these studies was shown in Figure 3.2. Earlier, we saw that, based on *p*-values, we would conclude that there were inconsistent results: four studies show a significant effect of the potion and six do not. However, based on the confidence intervals we would conclude the opposite: that the findings across the studies were quite consistent and that it was likely that the effect in the population was positive. Also shown in this figure, although you wouldn't have known what they were at the time, are the values of Cohen's $\hat{d}$ for each study.



Look back at Figure 3.2. Based on the effect sizes, is your view of the efficacy of the potion more in keeping with what we concluded based on *p*-values or based on confidence intervals?

The 10 studies summarized in Figure 3.2 have $\hat{d}$s ranging from 0.23 (other things being equal, smallish) to 0.71 (other things being equal, fairly large). The effect sizes are all positive: no studies showed worse anxiety after taking *antiSTATic*. Therefore, the effect sizes are very consistent: all studies show positive effects and *antiSTATic*, at worst, had an effect of about a quarter of a standard deviation, and, at best, an effect of almost three-quarters of a standard deviation. Our conclusions are remarkably like what we concluded when we looked at the confidence intervals, that is, there is consistent evidence of a positive effect in the population. Wouldn't it be nice if we could use these studies to get a definitive estimate of the effect in the population? Well, we can, and this process is known as **meta-analysis**. It sounds hard, doesn't it?

What wouldn't be hard would be to summarize these 10 studies by taking an average of the effect sizes:

$$\begin{aligned}\bar{\hat{d}} &= \frac{\sum_{i=1}^{k} \hat{d}_i}{n} \\ &= \frac{0.23 + 0.44 + 0.48 + 0.70 + 0.53 + 0.65 + 0.71 + 0.47 + 0.31 + 0.70}{10} \\ &= 0.52\end{aligned} \tag{3.16}$$

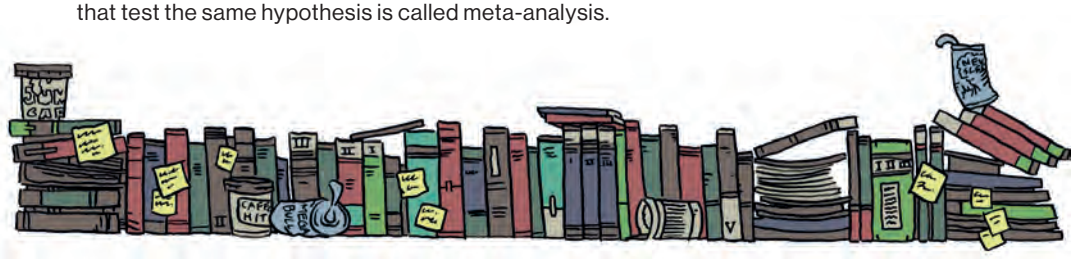
Congratulations, you have done your first meta-analysis – well, sort of. It wasn't that hard, was it? There's more to it than that but at a very basic level a meta-analysis involves computing effect sizes for a series of studies that investigated the same research question, and taking an average of those effect

sizes. At a less simple level we don't use a conventional average, we use what's known as a weighted average: in a meta-analysis each effect size is weighted by its precision (i.e., how good an estimate of the population it is) before the average is computed. By doing this, large studies, which will yield effect sizes that are more likely to closely approximate the population, are given more 'weight' than smaller studies, which should have yielded imprecise effect size estimates. Because the aim of meta-analysis is not to look at *p*-values and assess 'significance', it overcomes the same problems of NHST that we discussed for effect sizes. Of course, it has its own problems, but let's leave them for another book.



Cramming Sam's Tips 3.2
Effect sizes and meta-analysis

- An effect size is a way of measuring the size of an observed effect, usually relative to the background error.
- Cohen's $\hat{d}$ is the difference between two means divided by the standard deviation of the mean of the control group, or by a pooled estimated based on the standard deviations of both groups.
- Pearson's correlation coefficient, *r*, is a versatile effect size measure that can quantify the strength (and direction) of relationship between two continuous variables, and can also quantify the difference between groups along a continuous variable. It ranges from −1 (a perfect negative relationship) through 0 (no relationship at all) to +1 (a perfect positive relationship).
- The odds ratio is the ratio of the *odds* of an event occurring in one category compared to another. An odds ratio of 1 indicates that the *odds* of a particular outcome are equal in both categories.
- Estimating the size of an effect in the population by combining effect sizes from different studies that test the same hypothesis is called meta-analysis.



3.8 Bayesian approaches ©

The final alternative to NHST is based on a different philosophy on analysing data called **Bayesian statistics**. Bayesian statistics is a topic for an entire textbook (I particularly recommend these ones: Kruschke, 2014; McElreath, 2020) so we'll explore only the key concepts. Bayesian statistics is about using the data you collect to update your beliefs about a model parameter or a hypothesis. In some senses, NHST is also about updating your beliefs, but Bayesian approaches model the process explicitly.



To illustrate this idea, imagine you've got a crush on someone in your class. This is in no way autobiographical, but imagine you love stats and heavy metal music, and this other person seems way cooler than that. They hang around with the trendsetters and you haven't noticed any tendencies for them to wear Iron Maiden T-shirts, not even in a 'fashion statement' kind of way. This person never notices you. What is your belief that this person has a romantic interest in you? It's probably low, maybe a 10% chance.

A couple of days later you're in a class and out of the corner of your eye you notice your crush looking at you. Naturally you avoid eye contact for fear that it will make the universe explode, but in your peripheral vision you spot them looking at you every so often. Towards the end of class, your curiosity gets the better of you and, certain that they're looking somewhere else, you turn to them. To your horror, they're looking directly back at you and your eyes meet. *The* universe doesn't explode, but *yours* does. They smile at you. It's a sweet, friendly smile that melts your heart a little bit.

You have some new data about your situation – how will it affect your original belief? The data contained lots of positive signs so maybe you now think there's a 30% chance that they like you. Your belief after inspecting the data is different from your belief before you looked at it: you have updated your beliefs based on new information. That's a sensible way to live your life and is the essence of Bayesian statistics.

3.8.1 Bayesian statistics and NHST Ⓑ

There are important differences between Bayesian statistics and the classical methods that this book focuses on. NHST evaluates the probability of getting a test statistic at least as large as the one you have, given the null hypothesis. In doing so, you quantify the probability of the data obtained given a hypothesis is true:[15] p(data|hypothesis). Specifically, you ask what is the probability of my test statistic (data) or one bigger, given that the *null* hypothesis is true, p(test statistic|null hypothesis). This isn't a test of the null hypothesis, it's a test of the data given the null hypothesis is true. To test the null hypothesis, you need to answer the question 'what is the probability of the hypothesis given the data we have collected, p(hypothesis|data)?' In the case of the null we want the probability of the null hypothesis given the observed test statistic, p(null hypothesis|test statistic).

A simple example will illustrate that p(data|hypothesis) is not the same as p(null hypothesis|data). The probability that you are a professional actor or actress given that you have appeared in a blockbuster movie, p(actor or actress|been in a blockbuster), is very high because I suspect nearly everyone appearing in a blockbuster movie is a professional actor or actress. However, the inverse probability that you have been in a blockbuster movie given that you are a professional actor or actress, p(been in a blockbuster|actor or actress), is very small because the vast majority of actors and actresses do not end up in a blockbuster movie. Let's look at this distinction in more detail.

As you're probably aware, a significant proportion of people are scaly green lizard aliens disguised as humans. They live happily and peacefully among us causing no harm to anyone, and contributing lots to society by educating humans through things like statistics textbooks. I've said too much. The way the world is going right now, it's only a matter of time before people start becoming intolerant of the helpful alien lizards and try to eject them from their countries. Imagine you have been accused of being a green lizard alien. The government has a hypothesis that you are an alien. They sample your DNA and compare it to an alien DNA sample that they have. It turns out that your DNA matches the alien DNA.

Table 3.2 illustrates your predicament. Imagine that you live on a small island of 2000 inhabitants. One of those people (but not you) *is* an alien, and his or her DNA will match that of an alien. It is not possible to be an alien and for your DNA *not* to match, so that cell of the table contains a zero. Now, the remaining 1999 people are not aliens and the majority (1900) have DNA that does not match that of the alien sample that the government has; however, a small number do (99), including you.

15 I'm simplifying the situation because, in reality, NHST asks about the data *and more extreme data*.

Table 3.2 Contingency table of whether people are humans or green lizard aliens based on whether their DNA matches a sample of alien DNA that the government has

		DNA sample matches alien DNA		
		Match	No Match	Total
Accused	Green lizard alien	1	0	1
	Human	99	1900	1999
	Total	100	1900	2000

We'll start by look at the probability of the data given the hypothesis. This is the probability that the DNA matches given that the person is, in fact, an alien, $p(\text{match}|\text{alien})$. The conditional probability of a DNA match given the person is an alien would be the probability of being an alien *and* having a DNA match, $p(\text{alien} \cap \text{match})$, divided by the probability of being an alien, $p(\text{alien})$. There was one person from the population of 2000 who was an alien and had a DNA match, and there was also one person out of 2000 who was an alien, so the conditional probability is 1:

$$p(A \mid B) = \frac{p(\text{B} \cap \text{A})}{p(\text{B})}$$

$$p(\text{match}|\text{alien}) = \frac{p(\text{alien} \cap \text{match})}{p(\text{alien})} = \frac{1/2000}{1/2000} = \frac{0.0005}{0.0005} = 1 \tag{3.17}$$

Given the person is an alien, their DNA *must* match the alien DNA sample.

Asking what the probability of the data (DNA match) is given the hypothesis is true (they are an alien) is a daft question: if you already know the person is an alien, then it is a certainty that their DNA will match the alien sample. Also, if you already knew they were an alien then you would not need to bother collecting data (DNA). Calculating the probability of the data given that the hypothesis is true does not tell you anything useful about that hypothesis because it is conditional on the assumption that the hypothesis is true, which it might not be. Draw your own conclusions about NHST, which is based on this logic.

If you were a government employee tasked with detecting green lizards pretending to be humans, the question that matters is: given the data (the fact that your DNA matches) what is the probability of the theory that you *are* an alien? The conditional probability of a person being an alien given that their DNA matches would be the probability of being an alien and having a DNA match, $p(\text{match} \cap \text{alien})$, divided by the probability of a DNA sample matching the alien sample in general, $p(\text{match})$. There is one person in the population of 2000 who is an alien and has a DNA match, and there were 100 out of 2000 DNA samples that match that of an alien:

$$p(\text{alien}|\text{match}) = \frac{p(\text{match} \cap \text{alien})}{p(\text{match})} = \frac{1/2000}{100/2000} = \frac{0.0005}{0.05} = 0.01 \tag{3.18}$$

This illustrates the importance of asking the right question. If the government agent followed the logic of NHST, he or she would believe you were certainly a green lizard alien because the probability of you having a DNA match given you are an alien is 1. However, by asking what the probability is that you are an alien given that your DNA matches, they would realize that there is only a 0.01 probability (i.e. 1% chance) that you are an alien given that your DNA matches. Bayesian statistics addresses a more useful question than NHST: what is the probability of your hypothesis being true given the data collected?

3.8.2 Bayes' theorem ©

The conditional probabilities that we just discussed can be obtained using Bayes' theorem, which states that the conditional probability of two events can be obtained from their individual probabilities and the inverse conditional probability (Eq. 3.19). We can replace the letter A with our model or hypothesis, and replace letter B with the data collected to get a sense of how Bayes' theorem might be useful in testing hypotheses:

$$\begin{aligned}\text{Bayes' theorem} &= p(A|B) = \frac{p(B|A)\times p(A)}{p(B)} \\ &= p(\text{Model}|\text{Data}) = \frac{p(\text{Data}|\text{Model})\times p(\text{Model})}{p(\text{Data})} \\ &= \text{Posterior probability} = \frac{\text{likelihood}\times\text{prior probability}}{\text{marginal likelihood}}\end{aligned} \tag{3.19}$$

The terms in this equation have special names that we'll explore in detail now.

The **posterior probability** is our belief in a hypothesis (or parameter, but more on that later) after we have considered the data (hence it is *posterior* to considering the data). In the alien example, it is our belief that a person is an alien given that their DNA matches alien DNA, p(alien|match). This is the value that we are interested in finding out: the probability of our hypothesis given the data.

The **prior probability** is our belief in a hypothesis (or parameter) before considering the data. In our alien example, it is the government's belief in your guilt before they consider whether your DNA matches or not. This would be the base rate for aliens, p(aliens), which in our example is 1 in 2000, or 0.0005.

The **marginal likelihood**, or *evidence*, is the probability of the observed data, which in this example is the probability of matching DNA, p(match). The data show that there were 100 matches in 2000 cases, so this value is 100/2000, or 0.05. The **likelihood** is the probability that the observed data could be produced given the hypothesis or model being considered. In the alien example, it is, therefore, the probability that you would find that the DNA matched given that someone was in fact an alien, p(match|alien), which we saw before is 1.

If we put this together, our belief in you being a lizard alien given that your DNA matches that of a lizard alien is a function of how likely a match is if you *were* an alien (p = 1), our prior belief in you being an alien (p = 0.0005) and the probability of getting a DNA match (0.05). Our belief, having considered the data is 0.01: there is a 1% chance that you are an alien:

$$\begin{aligned}p(\text{alien}|\text{match}) &= \frac{p(\text{match}|\text{alien})\times p(\text{alien})}{p(\text{match})} \\ &= \frac{1\times 0.0005}{0.05} = 0.01\end{aligned} \tag{3.20}$$

This is the same value that we calculated in Eq. 3.18, all of which shows that Bayes' theorem is another route to obtaining the posterior probability.

Note that our prior belief in you being an alien (before we knew the DNA results) was 1 in 2000 or a probability of 0.0005. Having examined the data, that your DNA matched an alien, our belief increases to 1 in 100 or 0.01. After knowing that your DNA matched an alien we are more convinced that you are an alien than before we knew, which is as it should be; however, this is a far cry from being *certain* that you are an alien because your DNA matched.

3.8.3 Priors on parameters ©

In the alien example our belief was a hypothesis (you are an alien). We can also use Bayesian logic to update beliefs about parameter values (e.g., we can produce a Bayesian estimate of a *b*-value in a linear model by updating our prior beliefs in the value of that *b* using the data we collect).

Let's return to the example of you having a crush on someone in your class. One day, before you notice them smiling at you, you're lamenting your unrequited love to a friend and, being a stats enthusiast too, they ask you to estimate how much your crush likes you from 0 (hates you) to 10 (wants to marry you). '1', you reply. Self-esteem was never your strong point. They then ask, 'is it possible that it's a 5?' You don't think it is. What about a 4? No chance, you think. You turn your nose up at the possibility of a 3 but concede that there's an outside chance that it could be a 2, and you're sure it won't be a 0 because your crush doesn't seem like the hateful type. Through this process you've established that you think that your crush will like you somewhere between 0 and 2 on the scale, but you're most confident of a score of 1. As in the alien example, you can represent this belief *prior* to examining more data, but as a distribution, not a single probability value.

Immediately after the class in which you notice your crush smiling at you, you rush to your friend's room to report on this exciting development. Your friend asks you again to estimate where your crush's feelings lie on the 0–10 scale. '2', you reply. You're less confident than before that it is a 1, and you're prepared to entertain the possibility that it could be 3. Your beliefs after you have examine the data are represented by a posterior probability *distribution*.

Unlike a belief in a hypothesis where the prior probability is a single value, when estimating a parameter, the prior probability is a *distribution* of possibilities. Figure 3.6 shows two examples of a **prior distribution**. The top one mirrors our example and is said to be an **informative prior distribution**. We know from what we have learnt about probability distributions that they represent the plausibility of values: the peak is the value with the highest probability (the most plausible value) and the tails represent values with low probability. The top distribution is centred on 1, the value that you were most confident about when your friend first asked you to estimate your crush's feelings. This is the value you think is the most probable. However, the fact that there is a curve around this point shows that you are prepared to entertain other values with varying degrees of certainty. For example, you believe that values above 1 (the peak) are decreasingly likely until by the time the value is 2 your belief is close to 0: you think that it is impossible that your crush would like you *more than* 2 on the scale. The same is true if we look at values below 1: you think that values below the peak are decreasingly likely until we reach a rating of 0 (they hate you) at which point the probability of this belief is 0. Basically, you think that it is impossible that they hate you. Your self-esteem is low, but not *that* low. To sum up, you feel strongest that your crush likes you 1 on the scale, but you're prepared to accept that it could be any value between 0 and 2; as you approach those extremes your beliefs become weaker and weaker. This distribution is informative because it narrows down your beliefs: we know, for example, that you are not prepared to believe values above 2. We also know from the shape of the distribution that you are most confident of a value between about 0.5 and 1.5.

Figure 3.6 shows another type of prior distribution known as an **uninformative prior distribution**. It is a flat line, which means that you're prepared to believe all possible outcomes with equal probability. In our example, this means you're equally prepared to believe that your crush hates you (0) as you are that they love you (10) or sort of like you (5) or any value in between. Basically, you don't have a clue and you're prepared to believe anything. It's kind of a nice place to be when it comes to unrequited love. This prior distribution is called uninformative because the distribution doesn't tell us anything useful about your beliefs: it doesn't narrow them down.

Next you examine some data. In our example, you observe the person on whom you have a crush. These data are then mashed together (using Bayes' theorem) with your prior distribution to create a

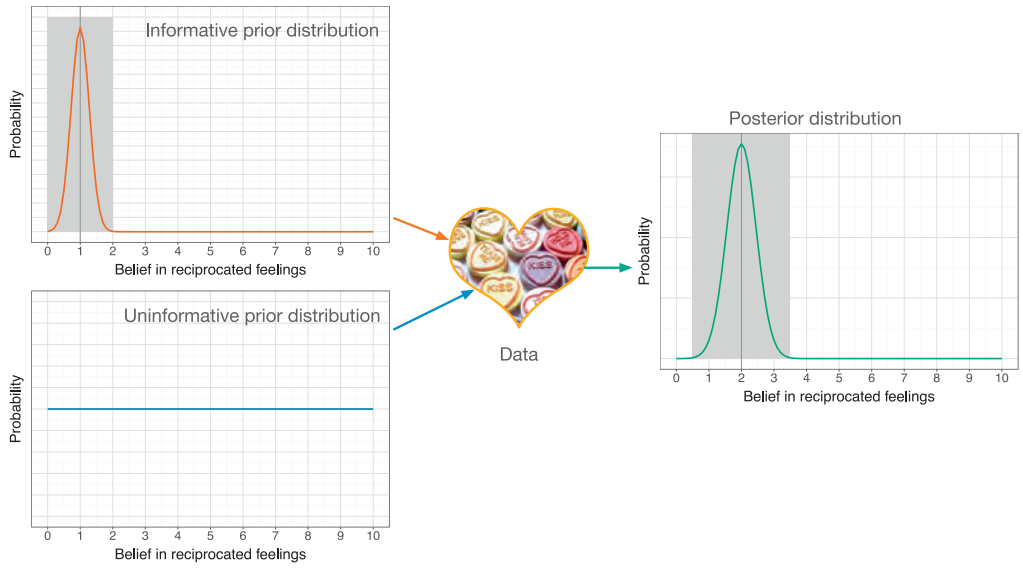


Figure 3.6 The process of updating your beliefs with Bayesian statistics

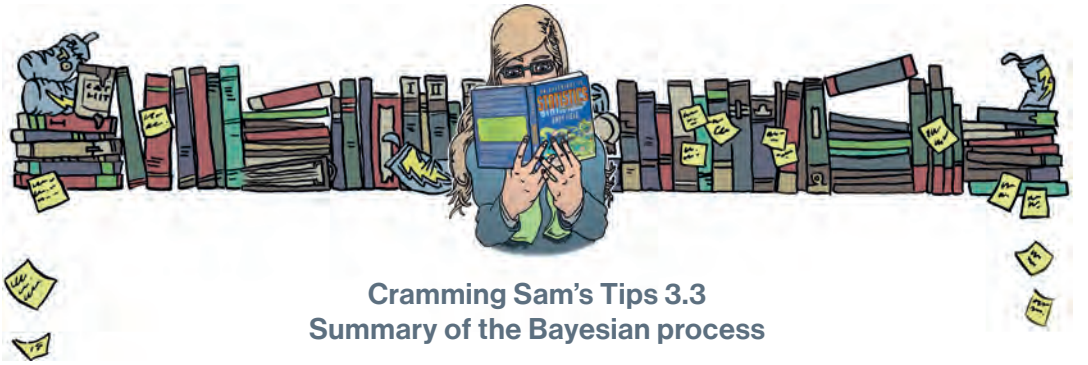
new distribution: the **posterior distribution**. Basically, a toad called the Reverend Toadmas Lillipad-Bayes sticks out his big sticky tongue and grabs your prior distribution and pops it in his mouth, he then lashes the tongue out again and grabs the observed data. He swirls the numbers around in his mouth, does that funny expanding throat thing that toads sometimes do, before belching out the posterior distribution. I think that's how it works.

Figure 3.6 shows the resulting posterior. It differs from the prior distribution. How it differs depends on both the original prior and the data. So, in Figure 3.6, if the data were the same you'd get different posteriors from the two prior distributions (because we're putting different initial beliefs into the toad's mouth). In general, if the prior is uninformative then the posterior will be heavily influenced by the data. This is sensible because if your prior belief is very open-ended then the data have a lot of scope to shape your posterior beliefs. If you start off not knowing what to believe, then your posterior beliefs should reflect the data. If the prior is strongly informative (i.e., you have very constrained beliefs) then the data will influence the posterior less than for an uninformative prior. The data are working against the prior. If the data are consistent with your already narrow beliefs, then the effect would be to narrow your belief further.

By comparing the prior and posterior distributions in Figure 3.6 we can speculate about what the data might have looked like. For the informative prior the data were probably quite a bit more positive than your initial beliefs (just like our story in which you observed positive signs from your crush) because your posterior belief is more positive than the prior: after inspecting the data, you're most willing to believe that your crush would rate you as a 2 (rather than your prior belief of 1). Because your prior belief was narrow (your beliefs were very constrained) the data haven't managed to pull your belief too far away from its initial state. The data have, however, made your range of beliefs wider than it was: you're now prepared to believe values in a range of about 0.5 to 3.5 (the distribution has got wider). In the case of the uninformative prior, the data that would yield the posterior in Figure 3.6 would be data distributed like the posterior: in other words, the posterior looks like the data. The posterior is more influenced by the data than the uninformative prior.

We can use the posterior distribution to quantify the plausible values of a parameter (whether that be a value such as your belief in what your crush thinks about you, or a *b*-value in a linear model).

If we want a point estimate (a single value) of the parameter, we can use the value where the probability distribution peaks. In other words, use the value that we think is most probable. For an interval estimate we can use the values that enclose a percentage of the posterior distribution. The interval is known as a **credible interval**, which is the limits between which 95% (or whatever percentage interests you) of values fall in the posterior distribution fall. Unlike a confidence interval, which for a given sample may or may not contain the true value, a credible interval can be turned into a probability statement such as 'there is a 95% probability that the parameter of interest falls within the limits of the interval'.



Cramming Sam's Tips 3.3
Summary of the Bayesian process

1. Define a prior that represents your subjective beliefs about a hypothesis (the prior is a single value) or a parameter (the prior is a distribution of possibilities). The prior can range from completely uninformative, which means you are prepared to believe pretty much anything, to strongly informative, which means your initial beliefs are quite narrow and specific.
2. Inspect the relevant data. In our frivolous example, this was observing the behaviour of your crush. In science, the process would be a bit more formal than that.
3. Bayes' theorem is used to update the prior distribution with the data. The result is a posterior probability, which can be a single value representing your new belief in a hypothesis, or a distribution that represents your beliefs in plausible values of a parameter, after seeing the data.
4. A posterior distribution can be used to obtain a point estimate (perhaps the peak of the distribution) or an interval estimate (a boundary containing a certain percentage, for example 95%, of the posterior distribution) of the parameter in which you were originally interested.



3.8.4 Bayes factors ©

We can use Bayes' theorem to compare two competing hypotheses. We'll return to the alien example (Section 3.8.1). In NHST terms, the alternative hypothesis equates to something happening. In this case, it would be that you are an alien. We know that the probability of you being an alien is 0.01. In NHST we also have a null hypothesis that reflects no effect or that nothing happened. The equivalent here is the hypothesis that you are human given the DNA match, $p(\text{human}|\text{match})$. What would this probability be?'



First, we need to know the probability of a DNA match given the person was human, $p(\text{match}|\text{human})$. There were 1999 humans and 99 had a DNA match, so this probability would be 99/1999, or 0.0495. Next, we need to know the probability of being human, $p(\text{human})$. There were 1999 out of 2000 people who are human (Table 3.2), so the probability would be 1999/2000, or 0.9995. Finally, we need the probability of a DNA match, $p(\text{match})$, which we worked out before as 0.05. Putting these values into Bayes' theorem, we get a probability of 0.99:

$$\begin{aligned} p(\text{human}|\text{match}) &= \frac{p(\text{match}|\text{human}) \times p(\text{human})}{p(\text{match})} \\ &= \frac{0.0495 \times 0.9995}{0.05} = 0.99 \end{aligned} \tag{3.21}$$

We know, then, that the probability of you being an alien given the data is 0.01 or 1%, but the probability of you being human given the data is 0.99 or 99%. We can evaluate the evidence of whether you are alien or human (given the data we have) using these values. In this case, you are far more likely to be human than you are to be an alien.

We know the posterior probability of being human given a DNA match is 0.99, and we previously worked out that the posterior probability of being alien given a DNA match is 0.01. We can compare these values by computing their ratio, which is known as the **posterior odds**. In this example, the value is 99, which means that you are 99 times more likely to be human than alien despite your DNA matching the alien sample:

$$\text{posterior odds} = \frac{p(\text{hypothesis 1}|\text{data})}{p(\text{hypothesis 2}|\text{data})} = \frac{p(\text{human}|\text{match})}{p(\text{alien}|\text{match})} = \frac{0.99}{0.01} = 99 \tag{3.22}$$

This example of the posterior odds shows that we can use Bayes' theorem to compare two hypotheses. NHST compares a hypothesis about there being an effect (the alternative hypothesis) with one that the effect is zero (the null hypothesis). We can use Bayes' theorem to compute the probability of both hypotheses given the data, which means we can compute a posterior odds of the alternative hypothesis relative to the null. This value quantifies how much more likely the alternative hypothesis is given the data relative to null (given the data). Eq. 3.23 shows how this can be done:

$$\begin{aligned} \frac{p(\text{alternative}|\text{data})}{p(\text{null}|\text{data})} &= \frac{p(\text{data}|\text{alternative}) \times p(\text{alternative}) / p(\text{data})}{p(\text{data}|\text{null}) \times p(\text{null}) / p(\text{data})} \\ \frac{p(\text{alternative}|\text{data})}{p(\text{null}|\text{data})} &= \frac{p(\text{data}|\text{alternative})}{p(\text{data}|\text{null})} \times \frac{p(\text{alternative})}{p(\text{null})} \\ \text{posterior odds} &= \text{Bayes factor} \times \text{prior odds} \end{aligned} \tag{3.23}$$

The first line shows Bayes' theorem for the alternative hypothesis divided by Bayes' theorem for the null. Because both incorporate the probability of the data, $p(\text{data})$, which is the marginal likelihood for both the alternative and null hypotheses, these terms drop out in the second line. The final line shows some special names that are given to the three components of the equation. On the left, we

have the posterior odds, which we have just explained. On the right, we have the **prior odds**, which reflect the probability of the alternative hypothesis to the null *before* you look at the data.

We have seen that priors reflect subjective beliefs before looking at the data, and the priors odds are no different. A value of 1 would reflect a belief that the null and alternative hypotheses are equally likely. This value might be appropriate if you were testing a completely novel hypothesis. Usually, though, our hypothesis would be based on past research and we would have a stronger belief in the alternative hypothesis than the null, in which case your prior odds might be greater than 1.

The final term in Eq. 3.23 is the **Bayes factor**. It represents the degree to which our beliefs change because of looking at the data (i.e., the change in our prior odds). A Bayes factor less than 1 supports the null hypothesis because it suggests that our beliefs in the alternative (relative to the null) have weakened. Looking at what the Bayes factor is (Eq. 3.23), a value less than 1 also means that the probability of the data given the null is higher than the probability of the data given the alternative hypothesis. Therefore, it makes sense that your beliefs in the alternative (relative to the null) should weaken. For the opposite reason, a Bayes factor greater than 1 suggests that the observed data are more likely given the alternative hypothesis than the null and your belief in the alternative (relative to the null) strengthens. A value of exactly 1 means that the data are equally likely under the alternative and null hypotheses and your prior beliefs are unchanged by the data.

For example, a Bayes factor of 10 means that the observed data are ten times more likely under the alternative hypothesis than the null hypothesis. Like effect sizes, some benchmarks have been suggested, but need to be used with extreme caution: values between 1 and 3 are considered evidence for the alternative hypothesis that is 'barely worth mentioning', values between 3 and 10 are considered 'evidence that has substance' for the alternative hypothesis, and values greater than 10 are strong evidence for the alternative hypothesis (Jeffreys, 1961).

3.8.5 Benefits of Bayesian approaches Ⓑ

We've seen already that Bayesian hypothesis testing asks a more sensible question than NHST, but it's worth reflecting on how it overcomes the problems associated with NHST.



First of all, we looked at various misconceptions around what conclusions you can draw from NHST about the null and alternative hypothesis. Bayesian approaches specifically evaluate the evidence for the null hypothesis so, unlike NHST, you *can* draw conclusions about the likelihood that the null hypothesis is true.

Second, *p*-values are confounded by the sample size and the stopping rules applied to data collection. Sample size is not an issue for Bayesian analysis. Theoretically you can update your beliefs based on one new data point. That single data point might not have much influence over a strong prior belief, but it could do. Imagine that you think my lizard alien example is ridiculous. You are utterly convinced that no green lizard alien people live among us. You have extensive data from every day of your life on which your belief is based. One day during a statistics lecture, you notice your professor licking her eyeball with a strangely serpentine tongue. After the lecture, you follow her, and when she thinks no one can see she peels her face off to reveal scaly green cheeks. This one new observation might substantially change your beliefs. However, perhaps you can explain this away as a prank. You turn to run across campus to tell your friends, and as you do you face a crowd of hundreds of students all peeling away their faces to reveal their inner space lizard. After soiling your pants, my guess is that

your prior belief would change substantially. My point is that the more data you observe, the stronger the effect it can have on updating your beliefs, but that's the only role that sample size plays in Bayesian statistics: bigger samples provide more information.

This point relates to the intentions of the researcher too. Your prior belief can be updated based on any amount of new information, therefore you do not need to determine how much data to collect before the analysis. It does not matter when you stop collecting data or what your rule is for collecting data, because any new data relevant to your hypothesis can be used to update your prior beliefs. Finally, Bayesian analysis is focused on estimating parameter values (which quantify effects) or evaluating relative evidence for the alternative hypothesis (Bayes factors) and so there is no black-and-white thinking involved, only estimation and interpretation. Thus, behaviour such as *p*-hacking is circumvented. Pretty cool, eh? I think Chicago put it best in that cheesy song: 'I'm addicted to you Bayes, you're a hard habit to break'.

There are down sides too. The major objection to Bayesian statistics is the reliance on a prior, which is a subjective decision and open to researcher degrees of freedom. For example, you can imagine a scenario where someone tweaks the prior to get a bigger Bayes factor or a parameter value that they like the look of. This criticism is valid, but I also see it as a strength in that the decision about the prior *must* be explicit: I don't see how you can report Bayesian statistics without explaining your decisions about the prior. In doing so readers are fully informed of what researcher degrees of freedom you might have employed.



Cramming Sam's Tips 3.4
Bayes factors

- Bayes' theorem can be used to update your prior belief in a hypothesis based on the observed data.
- The probability of the alternative hypothesis given the data relative to the probability of the null hypothesis given the data is quantified by the *posterior odds*.
- A *Bayes factor* is the ratio of the probability of the data given the alternative hypothesis to that for the null hypothesis. A Bayes factor greater than 1 suggests that the observed data are more likely given the alternative hypothesis than given the null. Values less than 1 suggest the opposite. Values between 1 and 3 reflect evidence for the alternative hypothesis that is 'barely worth mentioning', values between 1 and 3 are evidence that 'has substance', and values between 3 and 10 are 'strong' evidence (Jeffreys, 1961).



3.9 Reporting effect sizes and Bayes factors Ⓑ

If you're going to use NHST then because *p*-values depend on things like the sample size it is highly advisable to report effect sizes, which quantify the size of the observed effect, as well as *p*-values. Using the examples from the previous chapter (Section 2.11), we could report significance tests like this (note the presence of exact *p*-values and effect sizes):[16]

- ✓ Fear reactions were significantly higher when Andy's cat Fuzzy wore a fake human tongue compared to when he did not, $p = 0.023$, $\hat{d} = 0.54$.
- ✓ The number of cats intruding into the garden was significantly less when Fuzzy wore a fake human tongue compared to when he did not, $p = 0.007$, $\hat{d} = 0.76$.
- ✓ Fear reactions were not significantly different when Fuzzy wore a Billie Eilish mask compared to when he did not, $p = 0.18$, $\hat{d} = 0.22$.

You could also report Bayes factors, which are usually denoted by BF_{01}. Currently, because most journals expect to find NHST, it is not that common to see papers that report purely Bayesian statistics (although the number is increasing), and instead scientists augment their *p*-values with a Bayes factor. As you will see, Bayes factors are computed with an estimate of error (a percentage) and it's useful to report this value as well. You might report something like:

- ✓ The number of cats intruding into the garden was significantly less when Fuzzy wore a fake human tongue compared to when he did not, $p = 0.007$, $\hat{d} = 0.76$, $BF_{01} = 5.67 \pm 0.02\%$.

3.10 Jane and Brian's story Ⓐ

Brian had been ruminating about his last encounter with Jane. It hadn't gone well. He had a plan though. His lecturer had casually mentioned some problems with significance testing but glossed over them the way that lecturers do when they don't understand something. Brian decided to go to the library, where he read up on this thing called Bayesian statistics. 'If my lecturer doesn't understand the limitations of NHST, let alone Bayesian estimation,' Brian thought, 'Jane has got to be impressed with what I've been reading.' Brian noted everything down. A lot of it was too complex, but if he could get across some buzzwords, then maybe, just maybe, Jane would say more than a few words to him.

Brian waited on the library steps to see whether he could spot Jane in the campus square. Sure enough, he saw her figure in the distance. He took one final look at his notes, and as she walked past him he said, 'Hey Jane. 'Sup?'

Jane stopped. She looked down at the guy she'd bumped into last week. Why wouldn't he leave her alone? All she ever wanted was for people to leave her alone. She smiled a fake smile.

'Everyone say's you're really smart,' Brian continued. 'Maybe you can help me with some stuff I've been reading?'

Notes shaking in his hand, Brian told Jane everything he could about significance testing, effect sizes and Bayesian statistics.

16 If you use APA format, you need to drop the zero before the decimal point, so report $p = .023$, not $p = 0.023$.

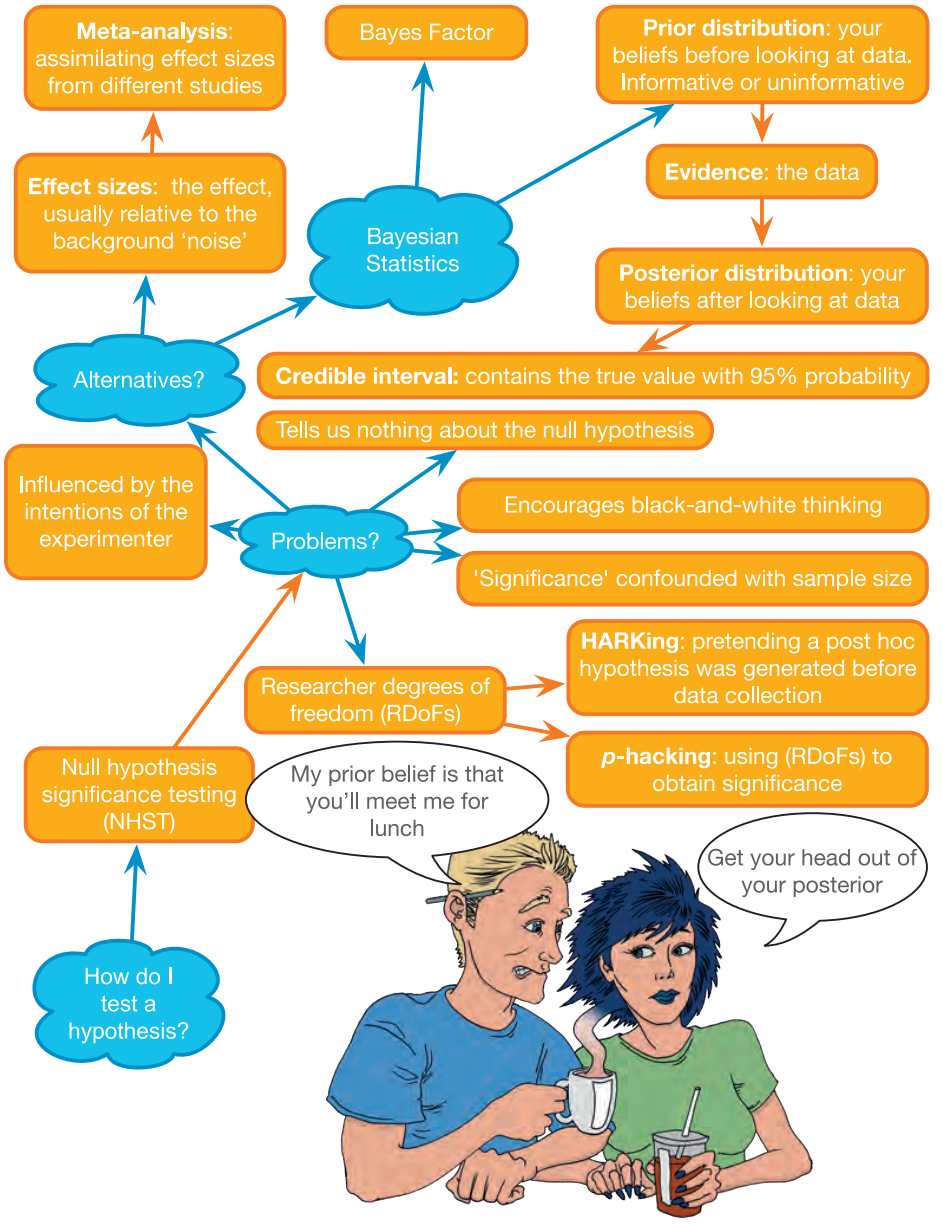


Figure 3.7 What Brian learnt from this chapter

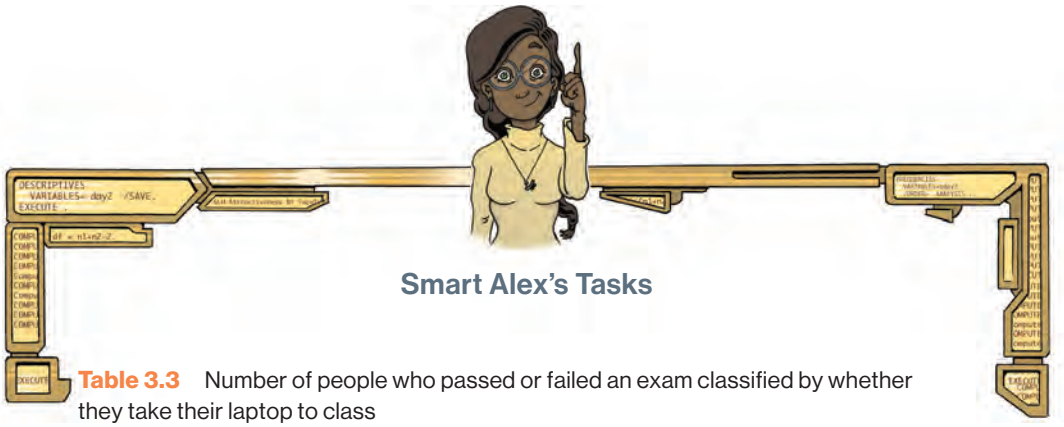
3.11 What next? Ⓐ

We began this chapter by seeing how an early belief that I couldn't live up to my brother's vast intellect panned out. We skipped forward in my life story, so that I could crowbar a photo of my children into the book. It was a pleasant diversion but now we must return to my own childhood. At age 4 I wasn't that concerned about the crippling weight of low expectation from my family. I had more pressing issues, namely talk of me leaving nursery (note: not 'being thrown out') and going to primary school (or 'elementary school' as I believe it's called in the USA). I can't remember how I felt about leaving nursery, but given how massively neurotic I am, it's hard to believe that I was anything other than anxious. Nursery was a safe and nurturing place to be. The notion of 'school' was hard for me to imagine. It

seemed like I was about to be dropped into a bewildering new environment where I'd be forced to interact with other children, learn to read, write and count, sing songs, draw, play sports, I'd be forced to ingest warm milk, I'd have a teacher, and I'd get to experience the very latest technology such as writing on a board with a new-fangled thing known as chalk. New environments can seem scary, and IBM SPSS Statistics might be a new environment for you. Fear not, we'll hold hands and face it together.

3.12 Key terms that I've discovered

Bayes factor
Bayesian statistics
Cohen's $\hat{d}$
Contingency table
Credible interval
Effect size
Empirical probability
HARKing
Informative prior distribution
Likelihood
Marginal likelihood
Meta-analysis
Odds
Odds ratio
Open science
p-curve
p-hacking
Peer Reviewers' Openness Initiative
Posterior distribution
Posterior odds
Posterior probability
Preregistration
Prior distribution
Prior odds
Prior probability
Publication bias
Registered reports
Researcher degrees of freedom
Tests of excess success
Uninformative prior distribution



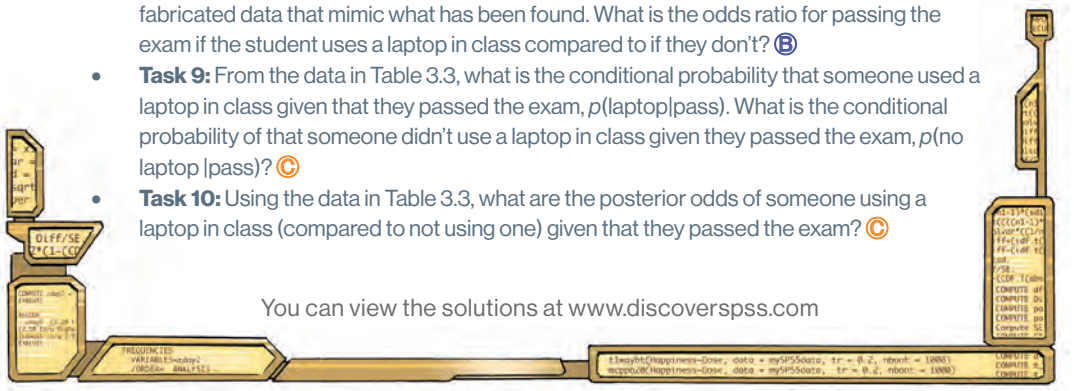
Smart Alex's Tasks

Table 3.3 Number of people who passed or failed an exam classified by whether they take their laptop to class

		Exam result		
		Pass	Fail	Total
Laptop in class?	Yes	24	16	40
	No	49	11	60
	Total	73	27	100

- **Task 1**: What is an effect size and how is it measured? Ⓑ
- **Task 2**: In Chapter 1 (Task 8) we looked at an example of how many games it took a sportsperson before they hit the 'red zone', then in Chapter 2 we looked at data from a rival club. Compute and interpret Cohen's $\hat{d}$ for the difference in the mean number of games it took players to become fatigued in the two teams mentioned in those tasks. Ⓑ
- **Task 3**: Calculate and interpret Cohen's $\hat{d}$ for the difference in the mean duration of the celebrity marriages in Chapter 1 (Task 9) and mine and my friends' marriages in Chapter 2 (Task 13). Ⓑ

- **Task 4:** What are the problems with null hypothesis significance testing? Ⓑ
- **Task 5:** What is the difference between a confidence interval and a credible interval? Ⓑ
- **Task 6:** What is meta-analysis? Ⓑ
- **Task 7:** Describe what you understand by the term 'Bayes factor'. Ⓑ
- **Task 8:** Various studies have shown that students who use laptops in class often do worse on their modules (Payne-Carter et al., 2016; Sana et al., 2013). Table 3.3 shows some fabricated data that mimic what has been found. What is the odds ratio for passing the exam if the student uses a laptop in class compared to if they don't? Ⓑ
- **Task 9:** From the data in Table 3.3, what is the conditional probability that someone used a laptop in class given that they passed the exam, *p*(laptop|pass). What is the conditional probability of that someone didn't use a laptop in class given they passed the exam, *p*(no laptop |pass)? Ⓒ
- **Task 10:** Using the data in Table 3.3, what are the posterior odds of someone using a laptop in class (compared to not using one) given that they passed the exam? Ⓒ



You can view the solutions at www.discoverspss.com



THE IBM SPSS STATISTICS ENVIRONMENT

4.1 What will this chapter tell me? 150
4.2 Versions of IBM SPSS Statistics 151
4.3 Windows, Mac OS and Linux 151
4.4 Getting started 152
4.5 The data editor 153
4.6 Entering data into IBM SPSS Statistics 157
4.7 SPSS syntax 174
4.8 The SPSS viewer 174
4.9 Exporting SPSS output 182
4.10 Saving files and restore points 182
4.11 Opening files and restore points 184
4.12 A few useful options 185
4.13 Extending IBM SPSS Statistics 187
4.14 Jane and Brian's story 190
4.15 What next? 191
4.16 Key terms that I've discovered 192
Smart Alex's tasks 192

Letters A to D indicate increasing difficulty

4.1 What will this chapter tell me? Ⓐ

Aged 7, we started a school project about 'the future' in which I wrote about robots (Figure 4.1). Astute child that I was, I noticed several important differences between robots and humans: they have lots of wires in them, you can control them and they will do whatever you want and – perhaps most important – if you get angry with them you can press their self-destruct button. I'm ashamed to say that I have contemplated a world where humans have self-destruct buttons rather more than the 7-year-old me pondered on the morality of blowing up a robot every time you got out of bed on the wrong side. I've also considered how nice it would be if I could control people, especially if I could control them into being a bit more tolerant and kind, and by 'a bit' I mean 'quite a lot'. Both current and 7-year-old me could see the use of robots to make the world a better place, it's just that for 7-year-old me the world being a better place consisted of being delivered an endless supply of tea and sweets.

Sadly the humanoid robotic tea and sweet delivery system has yet to materialize in the many decades since school. (Unless you are one of my two children, who live with robots called 'mum' and 'dad' who are under their control and deliver endless supplies of things that they want.) What has materialized is the statistical equivalent: IBM SPSS Statistics software. Like a robot, it has wires in it (well, the computer that runs it), it will do anything you want it to (not the tea and sweets thing, but …), and when you get angry with it you can press its self-destruct button. Sometimes, it presses its own self-destruct button just for kicks and giggles. This chapter introduces you to your new robotic statistics slave: IBM SPSS Statistics. Over subsequent chapters, SPSS Statistics will help us to build a utopian understanding of statistics, or to press our self-destruct buttons. Time will tell.

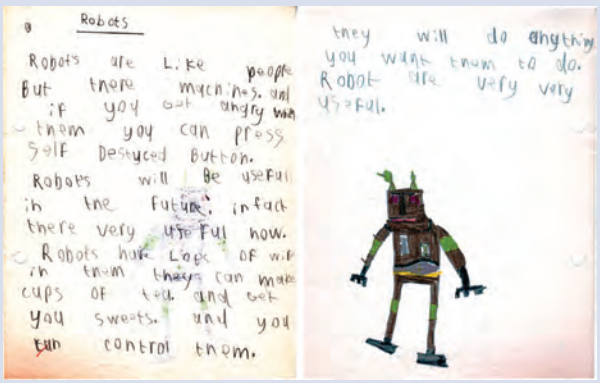


Figure 4.1 'Is he alive or dead? Has he thoughts within his head?'

4.2 Versions of IBM SPSS Statistics Ⓐ



This book is based primarily on version 29 of *IBM SPSS Statistics* (I generally call it SPSS for short). New versions of SPSS typically come out annually, and with new versions there will be improvements. However, this book covers only a small proportion of the functionality of SPSS, and focuses on tools that are stable and have been in the software a long time. Consequently, improvements made in new versions of SPSS are unlikely to impact the contents of this book. With a bit of common sense, you can get by with a book that doesn't explicitly cover the latest version (or the version you're using). So, although this edition was written using version 29, it will happily cater for earlier versions (certainly back to version 25), and most likely for versions 30 onwards (unless IBM does a major overhaul to keep me on my toes).

IBM SPSS Statistics comes in four flavours:[1]

- **Base**: Most of the functionality covered in this book is in the base package. The exceptions are logistic regression, loglinear analysis, Bayesian statistics and exact tests.
- **Standard**: This has everything in the base package but also covers logistic regression, loglinear analysis and Bayesian statistics (which are in this book) and generalized linear models (which are not in this book).
- **Professional**: This has everything in the standard edition, but with missing-value imputation and decision trees and forecasting (not covered in this text)
- **Premium**: This has everything in the professional package but also exact tests (which we cover in this book), and structural equation modelling, neural networks and complex sampling (which we don't cover).

There is also a subscription model where you can buy monthly access to a base package and, for an extra fee, add-ons for:

- **Custom tables and advanced statistics users**: This adds the functionality that comes with the standard package above.
- **Complex sampling & testing users**: This adds some of the functionality that comes with the professional package above such as missing-value analysis.
- **Forecasting & decision trees users**: This adds some more of the functionality from the professional package (forecasting and decision trees) and neural network predictive models (in the premium package above).

If you are subscribing then most of the contents of this book appear in the base subscription package, with a few things (e.g., Bayesian statistics and logistic regression) requiring the advanced statistics add-on.

4.3 Windows, Mac OS and Linux Ⓐ

SPSS Statistics works on Windows, Mac OS and Linux (and Unix-based operating systems such as IBM AIX, HP-UX and Solaris). SPSS Statistics is built on a program called Java, which means that the Windows, Mac OS and Linux versions differ very little (if at all). They look a bit different but only in

1 You can look at a detailed feature list at https://www.ibm.com/downloads/cas/ERYRD6G5 (which is a URL that seems likely to change during the lifespan of this book, but is linked from https://www.ibm.com/products/spss-statistics/details)

the way that, say, Mac OS looks different from Windows anyway.[2] I have taken the screenshots from Windows because that's the operating system that most readers will use, but you can use this book if you have a Mac (or Linux). In fact, I wrote this book on Mac OS.

4.4 Getting started Ⓐ

SPSS mainly uses two windows: the **data editor** (this is where you input your data and carry out statistical functions) and the **viewer** (this is where the results of any analysis appear). When SPSS loads, the start-up window in Figure 4.2 appears. At the top left is a box labelled *New Files*, where you can select to open an empty data editor window, or begin a database query (something not covered in this book). Underneath, in the box labelled *Recent Files* is a pane displaying any restore points you have. A restore point is essentially a snapshot of your working environment at a specific time (Section 4.10). If you want to reload a previous workstate then select it from this list and click Open. If you click on Recent Files the pane changes to list SPSS data files (on the current computer) on which you've recently worked. If you want to open an existing file, select it from the list and click Open. If you want to open a file that isn't in the list select Open another file... and click Open to browse to the file you want (see Section 4.11). The dialog box also has an overview of what's new in this release and contains links to tutorials and support, and a link to the online developer community. If you don't want this dialog to appear when SPSS starts up then select ☑ Don't show this dialog in the future.

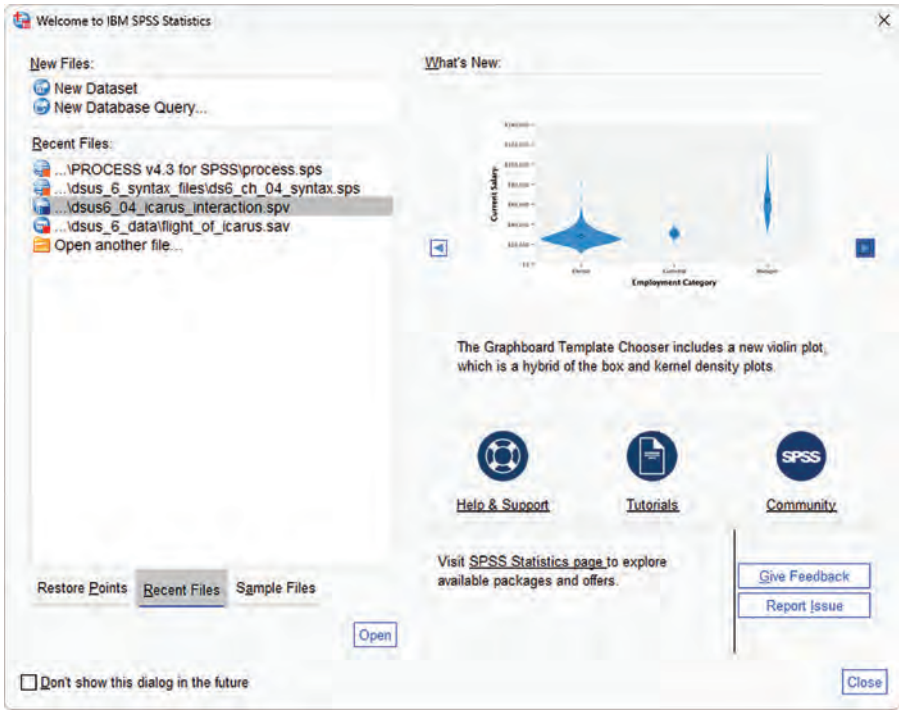


Figure 4.2 The start-up window of IBM SPSS

2 You can get the Mac OS version to display itself like the Windows version, but I have no idea why you'd want to do that.

4.5 The data editor Ⓐ

Unsurprisingly, the data editor window is where you enter and view data (Figure 4.3). At the top of this window (or the top of the screen on a Mac) is a menu bar like ones you've probably seen in other apps. As I am sure you're aware, you can navigate menus by using your mouse/trackpad to move the on-screen arrow to the menu you want and pressing (*clicking*) the left mouse button once. The click will reveal a list of menu items in a list, which again you can click using the mouse. In SPSS if a menu item is followed by a ❯ then clicking on it will reveal another list of options (a *submenu*) to the right of that menu item; if it doesn't then clicking on it will activate a window known as a *dialog box*. Any window in which you have to provide information or a response (i.e., 'have a dialog' with the computer) is a dialog box. When referring to selecting items in a menu I will use the menu item names connected by arrows to indicate moving down items or through submenus. For example, if I were to say that you should select the *Save As …* option in the *File* menu, you will see *File* ❯ *Save As …*.

The data editor has a **data view** and a **variable view**. The data view is for entering data, and the variable view is for defining characteristics of the variables within the data editor. To switch between the views, select one of the tabs at the bottom of the data editor (Data View Variable View); the highlighted tab indicates which view you're in (although it's obvious). Let's look at some features of the data editor that are consistent in both views. Both the data view and variable view are made up of lots of *cells*, which are boxes in which data values can be placed. When a cell is active it becomes highlighted (as in Figure 4.3). You can move around the data editor, from cell to cell, using the arrow keys ← ↑ ↓ → (on the right of the keyboard) or by clicking on the cell that you wish to activate. To enter a number into the data editor move to the cell in which you want to place the data value, type the value, then press the appropriate arrow button for the direction in which you wish to move. So, to enter a row of data, move to the far left of the row, type the first value and then press → (this process inputs the value and moves you into the next cell on the right).

Next, let's look at the menus.

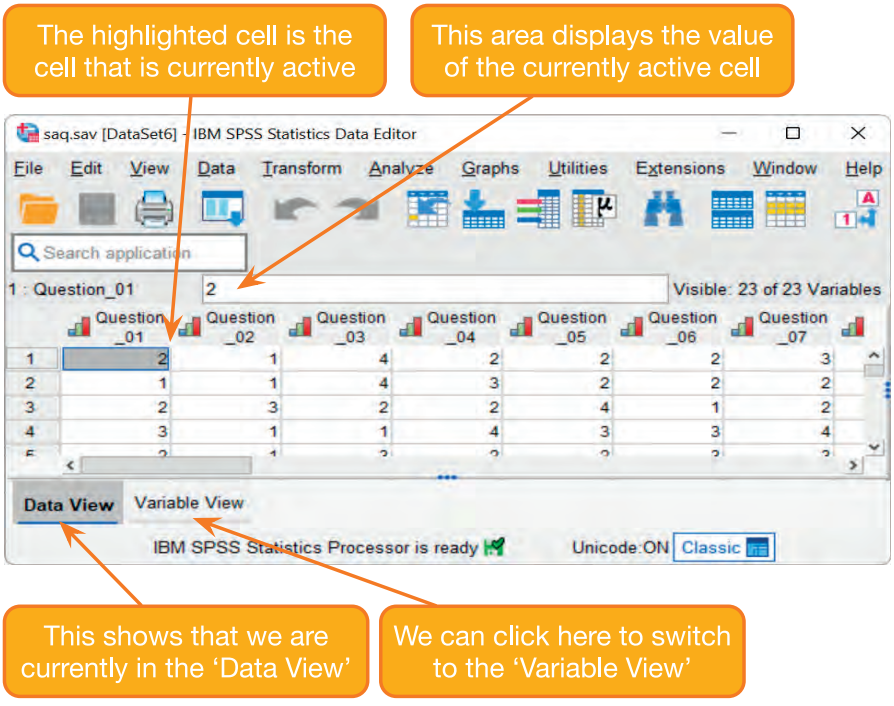


Figure 4.3 The SPSS data editor

Some letters are underlined within menu items in Windows, which tells you the *keyboard shortcut* for accessing that item. With practice these shortcuts are faster than using the mouse. In Windows, menu items can be activated by simultaneously pressing *Alt* on the keyboard and the underlined letter. So, to access the *File* > *Save As …* menu item you would simultaneously press *Alt* and F on the keyboard to activate the *File* menu, then, keeping your finger on the *Alt* key, press A. In Mac OS, keyboard shortcuts are listed in the menus; for example, you can save a file by simultaneously pressing ⌘ and S (I denote these shortcuts by ⌘ + S). Below is a brief reference guide to each of the menus:

- *File* This menu contains all the options that you expect to find in *File* menus: you can save data, graphs or output, open previously saved files and print graphs, data or output.
- *Edit* This menu contains edit functions for the data editor. For example, it is possible to *cut* and *paste* blocks of numbers from one part of the data editor to another (which is handy when you realize that you've entered lots of numbers in the wrong place). You can insert a new variable into the data editor (i.e., add a column) using Insert Variable, and add a new row of data between two existing rows using Insert Cases. Other useful options for large data sets are the ability to skip to a particular row (Go to Case…) or column (Go to Variable…) in the data editor. Finally, although for most people the default preferences are fine, you can change them by selecting Options….
- *View* This menu deals with system specifications such as whether you have grid lines on the data editor, or whether you display value labels (exactly what value labels are will become clear later).
- *Data* This menu is all about manipulating the data in the data editor. Some of the functions we'll use are the ability to split the file (Split File…) by a grouping variable (see Section 7.5.2), to run analyses on only a selected sample of cases (Select Cases…), to weight cases by a variable (Weight Cases…) which is useful for frequency data (Chapter 19), and to convert the data from wide format to long or vice versa (Restructure…).
- *Transform* This menu contains items relating to manipulating variables in the data editor. For example, if you have a variable that uses numbers to code groups of cases then you might want to switch these codes around by changing the variable itself (Recode into Same Variables…) or creating a new variable (Recode into Different Variables…). You can also create new variables from existing ones (e.g., you might want a variable that is the sum of 10 existing variables) using the *compute* function (Compute Variable…); see Section 6.10.4.
- *Analyze* The fun begins here, because statistical procedures lurk in this menu. Below is a rundown of the bits of the statistics menu that we'll use in this book:
    - *Descriptive Statistics* We'll use this for conducting descriptive statistics (mean, mode, median, etc.), frequencies and general data exploration. We'll use *Crosstabs…* for exploring frequency data and performing tests such as chi-square, Fisher's exact test and Cohen's kappa (Chapter 19).
    - *Compare Means* We'll use this menu for *t*-tests (related and unrelated – Chapter 10) and one-way independent ANOVA (Chapter 12).
    - *General Linear Model* This menu is for linear models involving categorical predictors, typically experimental designs in which you have manipulated a predictor variable using different cases (independent design), the same cases (repeated measures deign) or a combination of these (mixed designs). It also caters for multiple outcome variables, such as in multivariate analysis of variance (MANOVA; see Chapters 13–17).
    - *Mixed Models* We'll use this menu in Chapter 21 to fit a multilevel linear model and growth curve.
    - *Correlate* It doesn't take a genius to work out that this is where measures of correlation hang out, including bivariate correlations such as Pearson's r, Spearman's rho (ρ) and Kendall's tau (τ) and partial correlations (see Chapter 8).
    - *Regression* There are a variety of regression techniques available in SPSS, including simple linear regression, multiple linear regression (Chapter 9) and logistic regression (Chapter 20).

  - *Loglinear* Loglinear analysis is hiding in this menu, waiting for you, and ready to pounce like a tarantula from its burrow (Chapter 19).
  - *Dimension Reduction* You'll find factor analysis here (Chapter 18).
  - *Scale* We'll use this menu for reliability analysis in Chapter 18.
  - *Nonparametric Tests* Although, in general, I'm not a fan of these tests, in Chapter 7 I bury my principles in a coffin of hypocrisy to cover the Mann–Whitney test, the Kruskal–Wallis test, Wilcoxon's test and Friedman's ANOVA.

- *Graphs* This menu is used to access the Chart Builder (discussed in Chapter 5), which is your gateway to, among others, bar charts, histograms, scatterplots, box–whisker plots, pie charts and error bar graphs.
- *Utilities* There's plenty of useful stuff here but we don't get into it. I will mention that Data File Comments... is useful for writing notes about the data file to remind yourself of important details that you might forget (where the data come from, the date they were collected and so on). Also, you can define sets of variables using Define Variable Sets..., which can be useful in large data sets where you might want to work on a subset of the data. Having defined a set of variables (and named it), you can ask SPSS to only display this variable set in the data editor and in dialog boxes.
- *Extensions* Use this menu to access other IBM software that augments SPSS Statistics. Because most people won't have these add-ons (including me) I'm not going to discuss them in the book.
- *Window* This menu allows you to switch from window to window. So, if you're looking at the output and you wish to switch back to your data sheet, you can do so using this menu. There are icons to shortcut most of the options in this menu so it isn't particularly useful.
- *Help* Use this menu to access extensive searchable help files.

At the top of the data editor window are a set of *icons* (see Figure 4.3) that are shortcuts to frequently used facilities in the menus. Using the icons saves you time. Below is a brief list of these icons and their functions.



Use this icon to open a previously saved file (if you are in the data editor SPSS assumes you want to open a data file; if you are in the output viewer, it will offer to open a viewer file).



Use this icon to save files. It will save the file you are currently working on (be it data, output or syntax). If the file hasn't already been saved it will produce the *Save Data* As dialog box.



Use this icon for printing whatever you are currently working on (either the data editor or the output). The exact print options will depend on your printer. By default SPSS prints everything in the output window so a useful way to save trees is to print only a selection of the output (see SPSS Tip 4.5).



The big arrow on this icon implies to me that clicking it activates a miniaturizing ray that shrinks you before sucking you into a cell in the data editor, where you will spend the rest of your days cage-fighting decimal points. It turns out my intuition is wrong though and this icon opens the *Case* tab of the *Go To* dialog box, which enables you to go to a specific case (row) in the data editor. This shortcut is useful for large data files. For example, if we were analysing a survey with 3000 respondents, and wanted to look at the responses of participant 2407, rather than tediously scrolling down the data editor to find row 2407 we could click this icon, enter 2407 in the response box and click Go (Figure 4.4, left).

Go to a case



Go to a variable

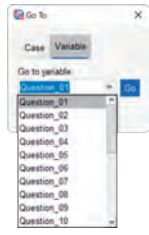


Figure 4.4 The *Go To* dialog box for a case (left) and a variable (right)



Clicking on this icon activates a list of the last 12 dialog boxes that were used; select any box from the list to reactivate the dialog box. This icon is a useful shortcut if you need to repeat parts of an analysis.



As well as data files with huge numbers of cases, you sometimes have ones with huge numbers of variables. Like the previous icon, clicking this one opens the *Go To* dialog box but in the *Variable* tab, which enables you to go to a specific variable (column) in the data editor. For example, the data file we use in Chapter 18 (**saq.sav**) contains 23 variables and each variable represents a question on a questionnaire and is named accordingly. If we wanted to go to Question 15, rather than getting wrist cramp by scrolling across the data editor to find the column containing the data for Question 15, we could click this icon, scroll down the variable list to Question 15 and click Go (Figure 4.4, right).



Clicking on this icon opens a dialog box that shows you the variables in the data editor on the left and summary information about the selected variable on the right. Figure 4.5 shows the dialog box for the same data file that we discussed for the previous icon. I have selected the first variable in the list on the left, and on the right we see the variable name (Question_01), the label (Statistics makes me cry), the measurement level (ordinal) and the value labels (e.g., the number 1 represents the response of 'Strongly agree').

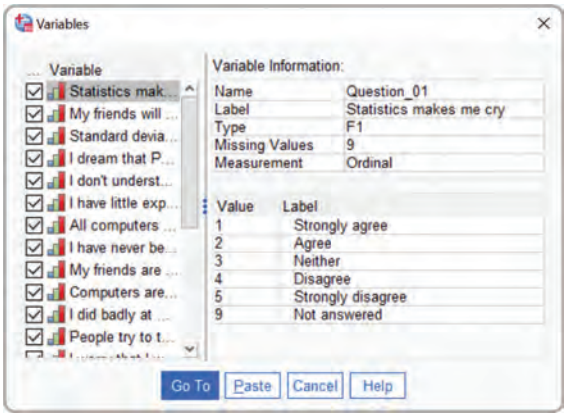


Figure 4.5 Dialog box for the *Variables* icon



If you select a variable (column) in the data editor by clicking on the name of the variable (at the top of the column) so that the column is highlighted, then clicking this icon will produce a table of descriptive statistics for that variable in the viewer window. To get descriptive statistics for multiple variables hold down *Ctrl* as you click at the top of the columns you want to summarize to highlight them, then click the icon.



I initially thought that this icon would allow me to spy on my neighbours, but this shining diamond of excitement was snatched cruelly from me as I discovered that it enables me to search for words or numbers in the data editor or viewer. In the data editor, clicking this icon initiates a search within the variable (column) that is currently active. This shortcut is useful if you realize from plotting the data that you have made an error, for example typed 20.02 instead of 2.02 (see Section 6.9.3), and you need to find the error – in this case by searching for 20.02 within the relevant variable and replacing it with 2.02 (Figure 4.6).

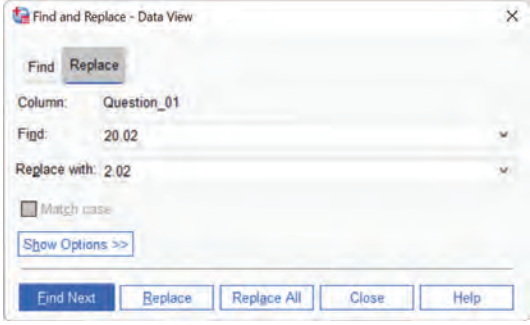


Figure 4.6 The *Find and Replace* dialog box



Clicking on this icon is a shortcut to the Data > Split File... dialog box (see Section 7.5.2). In SPSS, we differentiate groups of cases by using a coding variable (see Section 4.6.5), and this function runs any analyses separately for groups coded with such a variable. For example, imagine we test dogs and cats on how quickly they can be trained to fetch a ball. We would code each participant with a number that represents their species (e.g., 1 = dog, 0 = cat). If we then want to know the mean statistical ability for dogs and cats separately, we ask SPSS to split the file by the variable **animal** and then run descriptive statistics.



This icon is a shortcut to the Data > Select Cases... dialog box, which can be used if you want to analyse only a portion of your data. This function allows you to specify what cases you want to include in the analysis.



Clicking on this icon either displays or hides the value labels of any coding variables in the data editor. We use a coding variable to input information about category or group membership. We discuss this in Section 4.6.5. when we input some data about the rock band Metallica. We want to record which instrument each band member plays, which is a categorical variable (with categories of guitar, bass and drums). We create a variable called **instrument** and assign numbers to each category (1 as guitar, 2 as bass and 3 as drums). We then assign a *value label* (e.g. 'guitar') to the number assigned to the category (e.g., 1). In the data editor, we'd enter a number 1 for any guitarists, 2 for any bass players and 3 for the drummer. Clicking this icon toggles between the numbers you entered (you'd see a column of 1s, 2s and 3s) and the value labels you assigned to those numbers (you'd see a column displaying the word 'guitar', 'bass' or 'drums' in each cell).



This icon is a shortcut to the *Utilities* > Use Variable Sets... dialog box, which allows you to select which 'variable sets' are displayed in the data editor and any dialog boxes. This facility is useful when you have a large data set. You can use *Utilities* > Define Variable Sets... to select variables and collect them in a named 'set'. Having done that, use this icon to opt to see only the sets of variables that you want to see. In other words, you can hide variables that you're not currently using to focus on the more manageable subset (or subsets) of variables that you are currently using.



Clicking on this icon opens a dialog box where you can choose to switch off the toolbars that are currently active. You're probably not going to want to do that, but the dialog box also allows you to add a custom toolbar. That is, you create a toolbar of icons that act as shortcuts to features of SPSS to which you want quick access. This is quite handy if there are things you tend to do very frequently in SPSS: a custom toolbar will save you the time of scrolling through the menus.

4.6 Entering data into IBM SPSS Statistics Ⓐ

4.6.1 Data formats Ⓐ

There are two common data entry formats, which are sometimes referred to as **wide format data** and **long format data**. Most of the time we enter data into SPSS in wide format, although you can switch between wide and long formats using the *Data* > Restructure... menu. In the wide format *each row represents data from one entity and each column represents a variable*. There is no discrimination between predictor (independent) and outcome (dependent) variables: both appear in a separate column. The key point is that each row represents one entity's data (be that entity a human, mouse, tulip, business, or water sample) and any information about that entity should be entered across the data editor. Contrast this with long format in which scores on an outcome variable appear in a single column and rows represent a combination of the attributes of those scores. In long format data, scores from a single entity can appear over multiple rows, where each row represents a combination of the attributes of the score (the entity from which the score came, to which level of an independent variable the score belongs, the time point at which the score was recorded etc.). We use the long format in Chapter 21, but for everything else in this book we use the wide format.

Let's look at an example. Imagine you were interested in how perceptions of pain created by hot and cold stimuli were influenced by whether or not you swore while in contact with the stimulus (Stephens et al., 2009). You time how many seconds each of six people can keep their hand in a bucket of ice cold water. You then time for how many seconds they can hold a hot potato. Half the participants

are encouraged to shout profanities during the experiences. Figure 4.7 shows the data for our six fictional participants. The first person who was encouraged to swear held her hand in the ice water for 86 s and held the hot potato for 67 s. In wide format every participant's information is stored in a single row. There is a column representing the group to which they were assigned. This column contains what's known as a grouping variable, coding variable or, in experimental designs, an independent measure: an entity can belong to either the group that could swear or the group that was forbidden, but not both (see Section 4.6.5). The two timings make up a repeated measure because all participants were timed while in contact with a hot and with a cold stimulus. In wide format repeated measures are spread across columns in SPSS. So, for this example, in wide format the data for each participant occupies a single row, with their group membership, time with their hand in ice water and time holding a hot potato spread across three columns (see SPSS Tip 4.1).

Contrast this format with long format. In long format, every *observation* occupies a row in the data editor and columns are used to code information about that observation (e.g., the participant from which it came, whether they could swear, and whether the score relates to the ice water or the potato). In this format, participants scores are spread across multiple rows (in this specific example each participant occupies two rows, one containing their score for the ice water task and the other for the hot potato task).

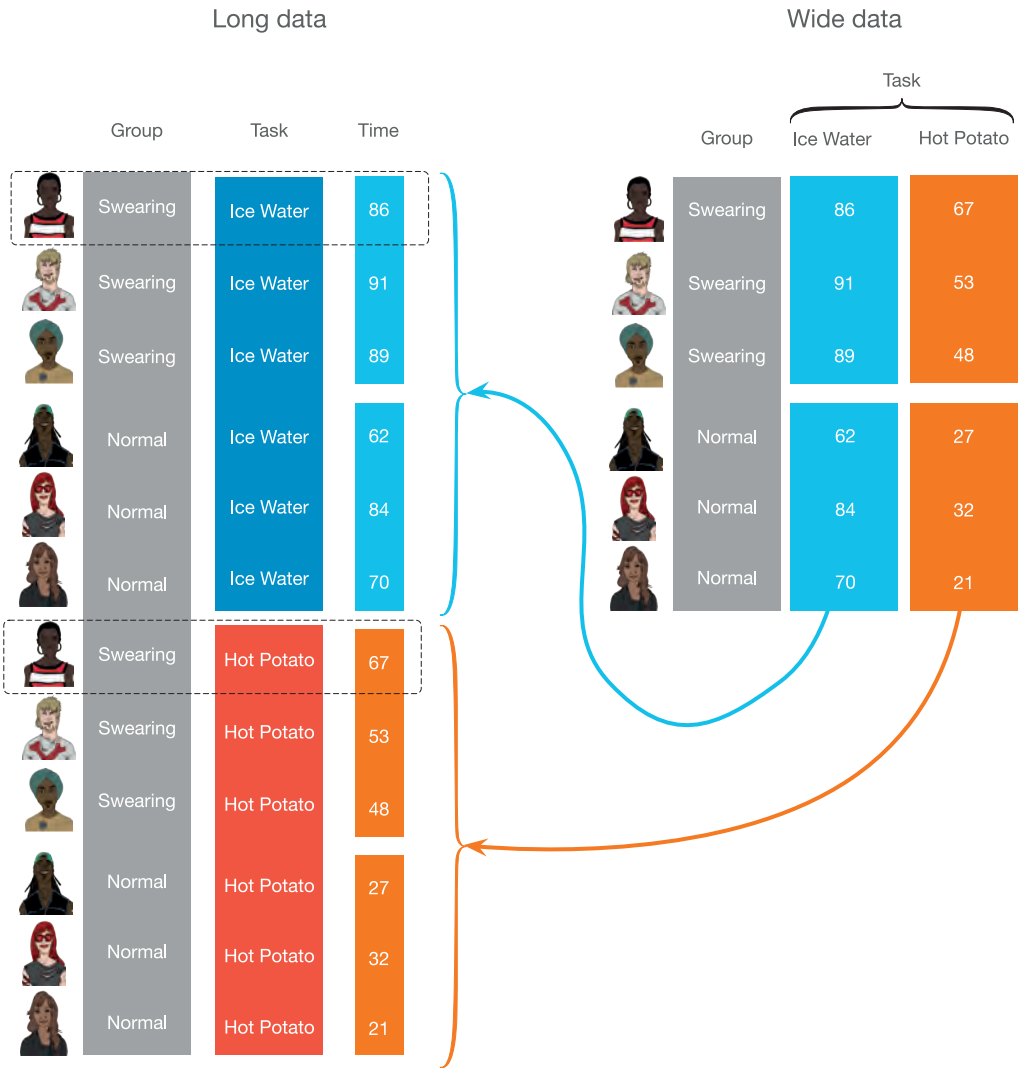
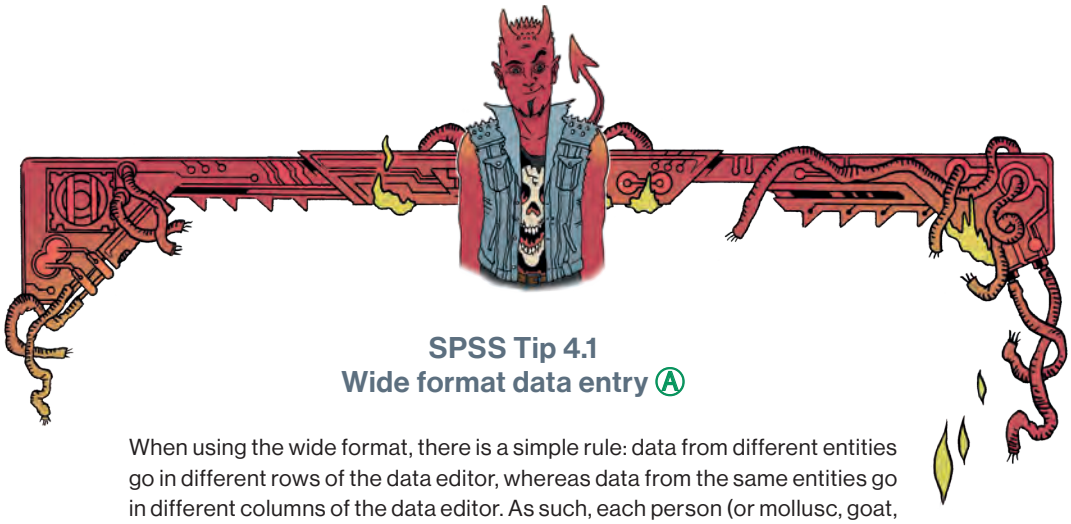


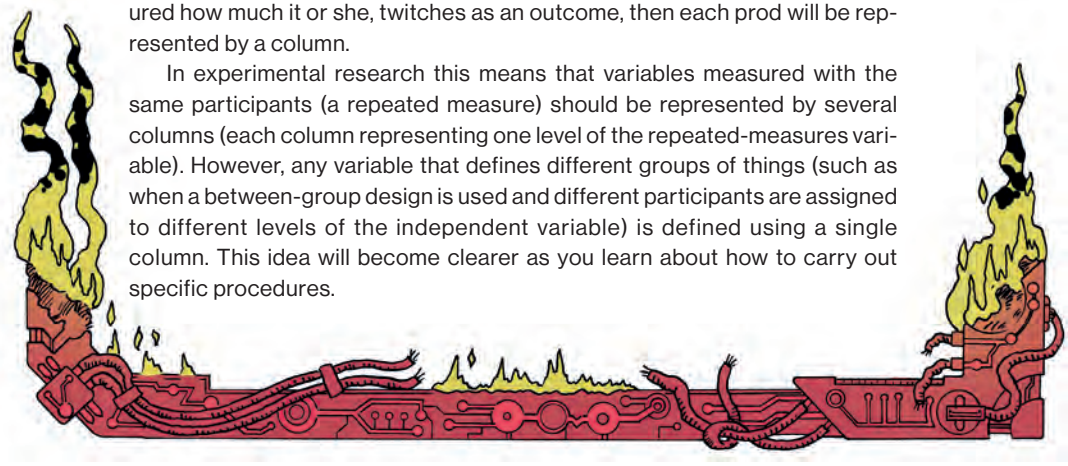
Figure 4.7 Wide and long data formats



SPSS Tip 4.1
Wide format data entry Ⓐ

When using the wide format, there is a simple rule: data from different entities go in different rows of the data editor, whereas data from the same entities go in different columns of the data editor. As such, each person (or mollusc, goat, organization or whatever you have measured) is represented in a different row. Data within each person (or mollusc, etc.) go in different columns. So, if you've prodded your mollusc, or human, several times with a pencil and measured how much it or she, twitches as an outcome, then each prod will be represented by a column.

In experimental research this means that variables measured with the same participants (a repeated measure) should be represented by several columns (each column representing one level of the repeated-measures variable). However, any variable that defines different groups of things (such as when a between-group design is used and different participants are assigned to different levels of the independent variable) is defined using a single column. This idea will become clearer as you learn about how to carry out specific procedures.



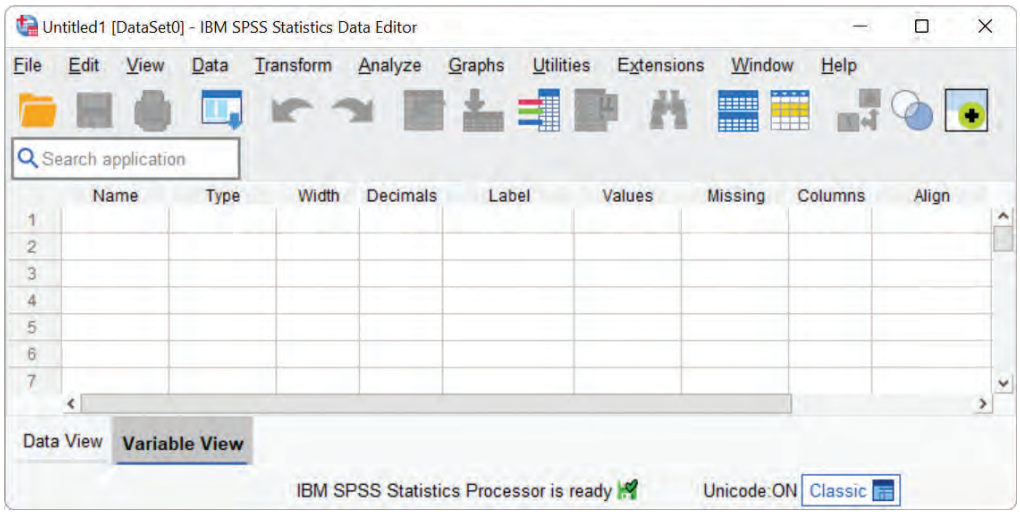


Figure 4.8 The variable view of the SPSS data editor

4.6.2 The variable view Ⓐ

Before we input data into the data editor, we need to create the variables using the variable view. To access this view click the *Variable View* tab at the bottom of the data editor (Data View | Variable View); the contents of the window will change (see Figure 4.8).

Every row of the variable view represents a variable, and you set characteristics of each variable by entering information into the following labelled columns (play around, you'll get the hang of it):

Name — Enter a name in this column for each variable. This name will appear at the top of the corresponding column in the data view, and helps you to identify variables. You can more or less write what you like, but there are certain symbols you can't use (mainly those that have other uses in SPSS such as +, –, &), and you can't use spaces. (I find 'hard spaces' useful, where you use an underscore instead of a space; for example, Andy_Field instead of Andy Field.) If you use a character that SPSS doesn't like, you'll get an error message saying that the variable name is invalid when you try to move out of the cell.

Type — Variables can contain different types of data. Mostly you will use the default **numeric variables** (the variable contains numbers). Other types are **string variables** (strings of letters), **currency variables** (i.e., £s, $s, €s) and **date variables** (e.g., 21-06-1973). So, for example, if you want to type in a person's name, you would need to change the variable type from numeric to string.

Width — By default, SPSS sets up new variables as *numeric* and storing 8 digits/characters. For numeric variables 8 digits is usually fine (unless you have very large numbers) but for string variables you can't write a lot in only 8 characters so you might want to increase the width to accommodate the length of text you intend to use. Width differs from Columns (see below) in that it affects what is stored in the variable rather than what is displayed in the data editor.

Decimals — By default, 2 decimal places are displayed. (If you don't change this option then when you enter whole numbers SPSS adds a decimal place with two zeros after it, which can be disconcerting.) If you want to change the number of decimal places for a given variable, replace the 2 with a new value or increase or decrease the values using .

Label — The variable Name (see above) has some restrictions on characters, and you also wouldn't want to use huge long names at the top of your columns because they become hard to read. Therefore, you can write a longer variable description in this column. This may seem pointless, but it is a good habit to get into (see SPSS Tip 4.2).

Values — This column is for assigning numbers to represent groups of people (see Section **4.6.5**).

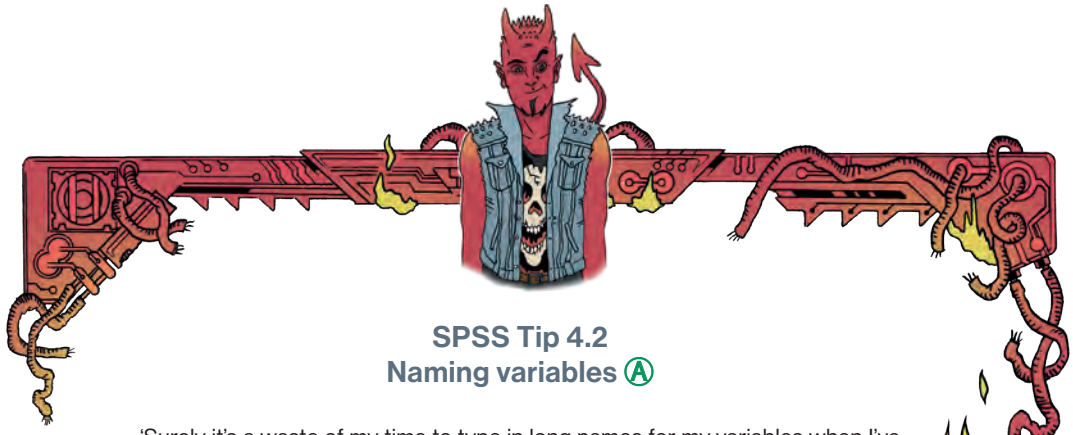
Missing — This column is for assigning numbers to missing data (see Section 4.6.7).

Columns — Enter a number into this column to determine the width of the column (i.e., how many characters are displayed in the column). The information entered under Columns differs from that for Width (see above) which determines the width of the variable itself. So, you could have a variable with a width of 10 characters but by setting the *column* width to 8 you would see only 8 of the 10 characters in the data editor. It can be useful to increase the column width if you have a string variable (Section 4.6.3) or a variable with value labels (Section 4.6.5) that exceed 8 characters.

Align — You can use this column to select the alignment of the data in the corresponding column of the data editor. You can choose to align the data to the Left, Right or Center.

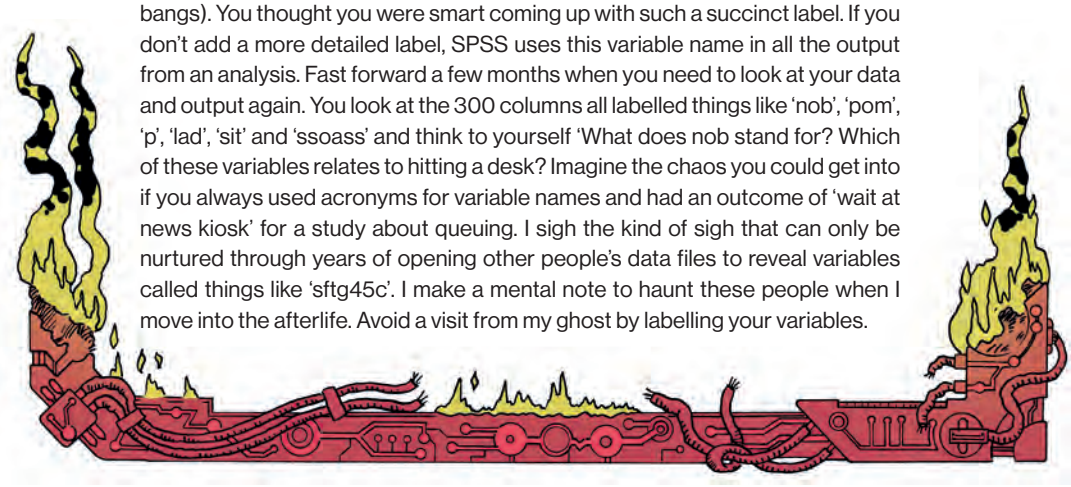
Measure — Use this column to define the scale of measurement for the variable as Nominal, Ordinal or Scale (Section 1.5.2).

Role — SPSS Statistics has some procedures that attempt to run analyses automatically without you needing to think about what you're doing (one example is *Analyze*>*Regression*>Automatic Linear Modeling...). To think on your behalf, SPSS needs to know whether a variable is a predictor (Input), an outcome (Target), both (Both, although I'm not sure how that works out in practice), a variable that splits the analysis by different groups (Split), a variable that selects part of the data (Partition) or a variable that has no predefined role (None). It's rarely a good idea to let a computer do your thinking for you, so I'm not a fan of the sorts of automated procedures that take advantage of these role assignments (they have their place, but that place is not in this book). Therefore, I'm not going to mention roles again.



SPSS Tip 4.2
Naming variables Ⓐ

'Surely it's a waste of my time to type in long names for my variables when I've already given them a short one?' I hear you ask. I can understand why it would seem so, but as you go through university or your career accumulating data files, you will be grateful that you did. Imagine you had a variable called 'number of times I wanted to bang the desk with my face during Andy Field's statistics lecture'; then you might have named the column in SPSS 'nob' (short for number of bangs). You thought you were smart coming up with such a succinct label. If you don't add a more detailed label, SPSS uses this variable name in all the output from an analysis. Fast forward a few months when you need to look at your data and output again. You look at the 300 columns all labelled things like 'nob', 'pom', 'p', 'lad', 'sit' and 'ssoass' and think to yourself 'What does nob stand for? Which of these variables relates to hitting a desk? Imagine the chaos you could get into if you always used acronyms for variable names and had an outcome of 'wait at news kiosk' for a study about queuing. I sigh the kind of sigh that can only be nurtured through years of opening other people's data files to reveal variables called things like 'sftg45c'. I make a mental note to haunt these people when I move into the afterlife. Avoid a visit from my ghost by labelling your variables.



Let's use the variable view to create some variables. We're going to input some data about the rock band Metallica. The complete data are in Table 4.1.

Table 4.1 Some Metalli-data

Name	Birth date	Death date	Instrument	Current member	Songs	Net worth ($ million)	Albums recorded	Net worth per song ($ million)
Lars Ulrich	26-Dec-1963		Drums	True	111	$350	10	$3.15
James Hetfield	03-Aug-1963		Guitar	True	112	$300	10	$2.68
Kirk Hammett	18-Nov-1962		Guitar	True	56	$200	10	$3.57
Rob Trujillo	23-Oct-1964		Bass	True	16	$40	2	$2.50

(Continued)

Table 4.1 (Continued)

Jason Newsted	04-Mar-1963		Bass	False	3	$60	4	$20.00
Cliff Burton	10-Feb-1962	27-Sep-1986	Bass	False	11	$1	3	$0.09
Dave Mustaine	13-Sep-1961		Guitar	False	6	$14	0	$2.33

4.6.3 Creating a string variable Ⓐ

The first variable in Table 4.1 is the name of the band member. This variable is a *string variable* because it consists of names (which are strings of letters). To create this variable in the variable view:

1. Click in the first white cell in the column labelled *Name*.
2. Type the word 'Name'.
3. Move from this cell using the arrow keys on the keyboard (you can also just click in a different cell, but this is a very slow way of doing it).

Well done, you've just created your first variable. Notice that once you've typed a name, SPSS creates default settings for the variable (such as assuming it's numeric and assigning 2 decimal places). However, we don't want a numeric variable (i.e., numbers), we want to enter people's names, so we need a *string* variable and need to change the variable type. Move into the column labelled Type using the arrow keys on the keyboard. The cell will now look like this Numeric. Click to activate the *Variable Type* dialog box. By default, the numeric variable type is selected (Numeric) – see the top of Figure 4.9. To change the variable to a string variable, click String (bottom left of Figure 4.9). Next, if you need to enter text of more than 8 characters (the default width) then change this default value to a number reflecting the maximum number of characters that you will use for a given case of data. Click OK to return to the variable view.

Next, because I want you to get into good habits, move to the cell in the Label column and type a description of the variable, such as 'Band member's name'. Finally, we can specify the scale of measurement for the variable (see Section 1.5.2) by going to the column labelled *Measure* and selecting Nominal, Ordinal or Scale from the drop-down list. In the case of a string variable, it represents a description of the case and provides no information about the order of cases, or the magnitude of one case compared to another. Therefore, select Nominal.

Once the variable has been created, return to the data view by clicking on the *Data View* tab at the bottom of the data editor (Data View Variable View). The contents of the window will change, and notice that the first column now has the label *Name*. We can enter the data for this variable in the column underneath. Click the white cell at the top of the column labelled *Name* and type the first name, 'Lars Ulrich'. To register this value in this cell, move to a different cell and because we are entering data down a column, the most sensible way to do this is to press the ↓ key on the keyboard. This action moves you down to the next cell, and the word 'Lars Ulrich' should appear in the cell above. Enter the next name, 'James Hetfield', and then press ↓ to move down to the next cell, and so on.

4.6.4 Creating a date variable Ⓐ

The second and third columns in our table contains dates (birth dates and death dates). To create a date variable we more or less repeat what we've just done. First, move back to the variable view using the tab at the bottom of the data editor (Data View Variable View). Move to the cell in row 2 of the column labelled *Name* (under the previous variable you created). Type the word 'birth_date' (note that I have used a hard space to separate the words). Move into the column labelled Type using the → key on the

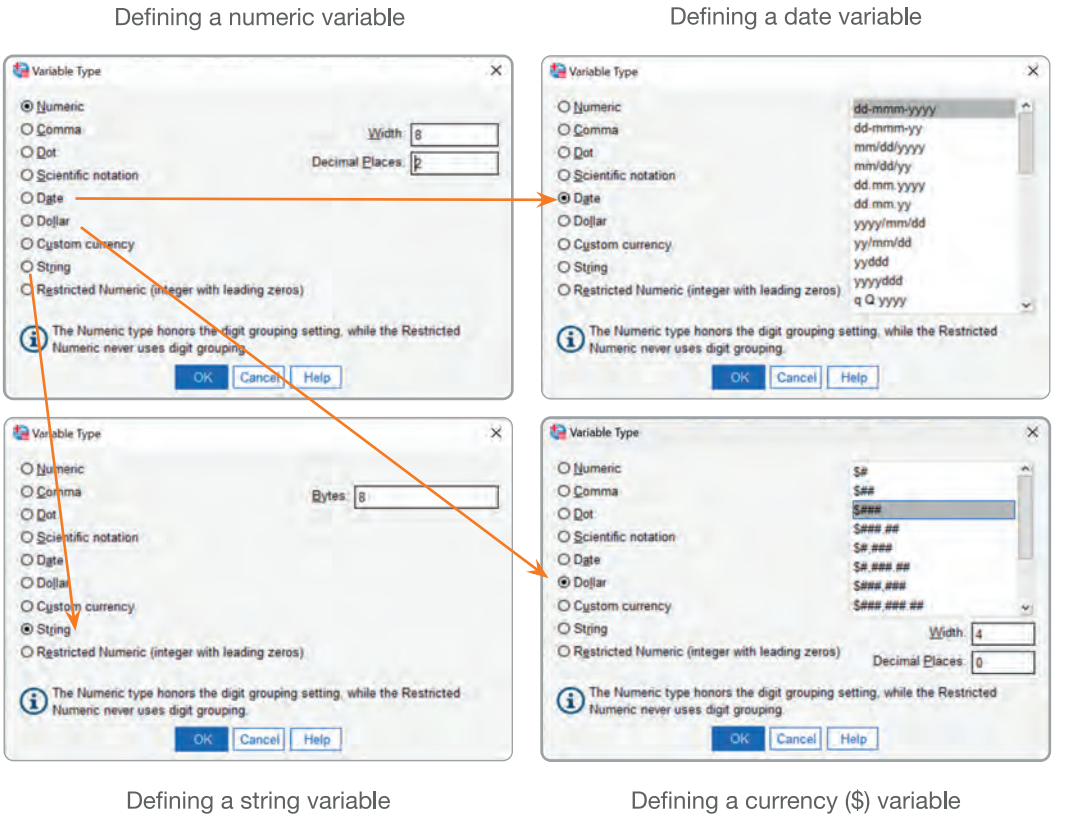


Figure 4.9 Defining numeric, string, date and currency ($) variables

keyboard (doing so creates default settings in the other columns). As before, the cell you have moved into will indicate the default of Numeric, and to change this we click to activate the *Variable Type* dialog box, and click Date (top right of Figure 4.9). On the right of the dialog box is a list of date formats, from which you can choose your preference; being British, I am used to the day coming before the month and have chosen dd-mmm-yyyy (i.e., 26-Dec-1963), but Americans, for example, more often put the month before the date so might select mm/dd/yyyy (12/26/1963). When you have selected a date format, click OK to return to the variable view. Finally, move to the cell in the column labelled *Label* and type 'Date of birth'.

Once the variable has been created, return to the data view by clicking on the 'Data View' tab (Data View Variable View). The second column now has the label *birth_date*; click the white cell at the top of this column and type the first value, 26-12-1963. To register this value in this cell, move down to the next cell by pressing the ↓ key. Note that SPSS automatically changes the 12 to 'Dec'. Now enter the next date, and so on.



4.6.5 Creating coding variables Ⓐ

Coding or grouping variables use numbers to represent different groups or categories of data. As such, a coding variable is *numeric*, but because the numbers represent names its variable type is Nominal.

The groups of data represented by coding variables could be levels of a treatment variable in an experiment (an experimental group or a control group), different naturally occurring groups (ethnic groups, gender identity, socioeconomic categories, marital status, etc.), different geographic locations (countries, states, cities, etc.) or different organizations (different hospitals within a healthcare trust, different schools in a study, different companies).

In experiments that use an independent design, coding variables represent predictor (independent) variables that have been measured between groups (i.e., different entities were assigned to different groups). We do not, generally, use this kind of coding variable for experimental designs where the independent variable was manipulated using repeated measures (i.e., participants take part in all experimental conditions). For repeated-measures designs we typically use different columns to represent different experimental conditions.

Think back to our swearing and pain experiment. This *was* an independent design because we had two groups representing the two levels of our independent variable: one group could swear during the pain tasks, the other could not. Therefore, we can use a coding variable. We might assign the experimental group (swearing) a code of 1 and the control group (no swearing) a code of 0. To input these data you would create a variable (which you might call **group**) and type the value 1 for any participants in the experimental group, and 0 for any participant in the control group. These codes tell SPSS that the cases that have been assigned the value 1 should be treated as belonging to the same group, and likewise for the cases assigned the value 0. The codes you use are arbitrary because the numbers themselves won't be analysed, so although people typically use 0, 1, 2, 3, etc., if you're a particularly arbitrary person feel free to code one group as 616 and another as 11 and so on.

We have a coding variable in our data that codes which instrument the band member plays (guitar, bass or drums) and another that categorizes band members according to whether they are current members. To create the **instrument** coding variable, we follow the same steps as before, but we will also have to record which numeric codes are assigned to which instrument. First, return to the variable view (Data View Variable View) if you're not already in it and move to the cell in the third row under the column labelled *Name*. Type a name (let's call it **instrument**). I'm still trying to instil good habits, so move

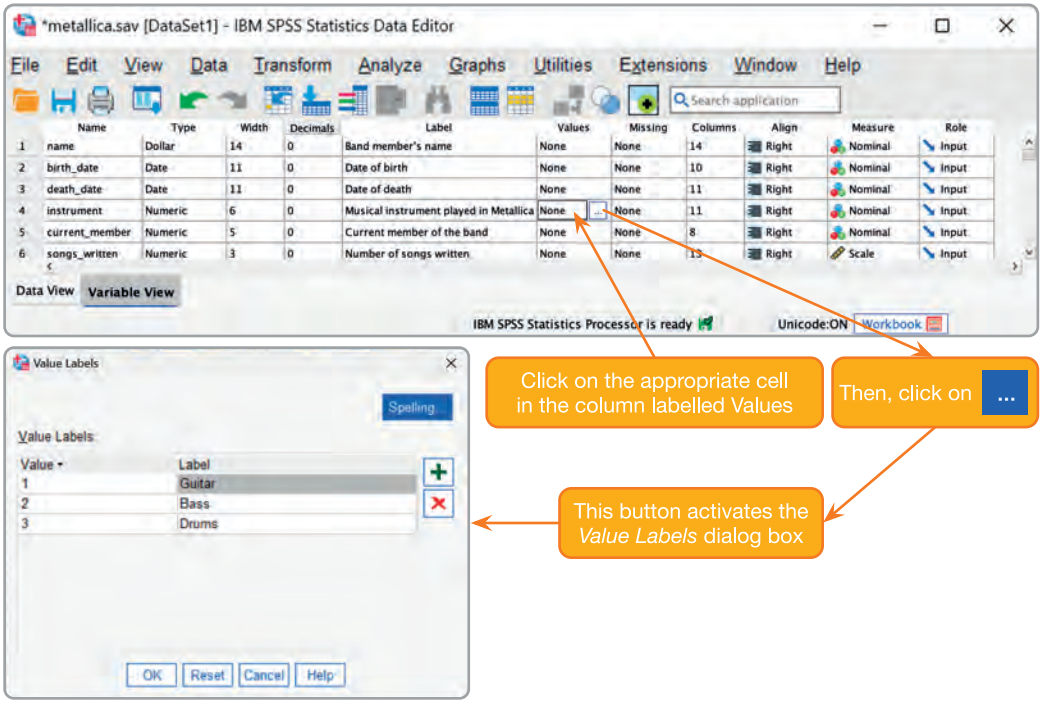


Figure 4.10 Defining coding variables and their values

along the third row to the column called *Label* and give the variable a full description such as 'Musical instrument played in Metallica'. To define the group codes, move along the row to the column labelled Values. The cell will indicate the default of None. Click ... to access the *Value Labels* dialog box (see Figure 4.10).

The *Value Labels* dialog box is used to specify group codes. To add a category and attach a value and label click +. A new row will be created in the table in the dialog box with an empty cell under the column labelled *Value* and another under the column labelled *Label*. We want to create three categories (guitar, bass, drums) so we can click + three times to create the necessary rows. If you accidentally create too many rows then click in one of the cells to make it active and click ×, which will delete the row of the currently active cell. Having created three blank rows, the second step is to fill in the cells of the table. You can navigate the cells of the table using the ← ↑ ↓ → keys on your keyboard (or selecting cells with your mouse). Click on the first empty cell in the *Value* column and type a code (in this case 1), then press ↓ to move to the cell underneath and type the next code (in this case 2), then press ↓ to move to the cell underneath and type the final code (in this case 3). Having set the numeric values for the categories, we can assign each category a label by filling in the cells of the *Label* column. Click on the first empty cell in this column and type a descriptive label (in this case 'Guitar'), then press ↓ to move to the cell underneath and type the next label ('Bass'), then press ↓ to move to the cell underneath and type the final label ('Drums'). When you have defined all your coding values you might want to check for spelling mistakes in the value labels by clicking Spelling... To finish, click OK; if you do this while any cells in the dialog box are empty SPSS displays a warning that any 'One or more value label definitions are incomplete. Would you like to discard them?'. If you click Yes then SPSS will discard any value labels that are incomplete, and if you click No it will take you back to the dialog box to complete the empty cells. Once you're back in the variable view, set the level of measurement for the variable. Coding variables represent categories and so the scale of measurement is nominal (or ordinal if the categories have a meaningful order). To specify this level of measurement, go to the column labelled *Measure* and select Nominal (or Ordinal if the groups have a meaningful order) from the drop-down list.

Having defined your codes, switch to the data view and for each participant type the numeric value that represents the instrument they play into the column labelled *instrument*. In our example, if a person was a guitarist, type '1', but if they played bass then type '2' (see SPSS Tip 4.3). SPSS can display either the numeric codes, or the value labels that you assigned and you can toggle between the two states by clicking (see Figure 4.11). Figure 4.11 shows how the data should be arranged. Remember that each row of the data editor represents data from one entity: the first member was a drummer (I use the past tense deliberately), the next two are guitarists, the next three are bassists and the final member is a guitarist.

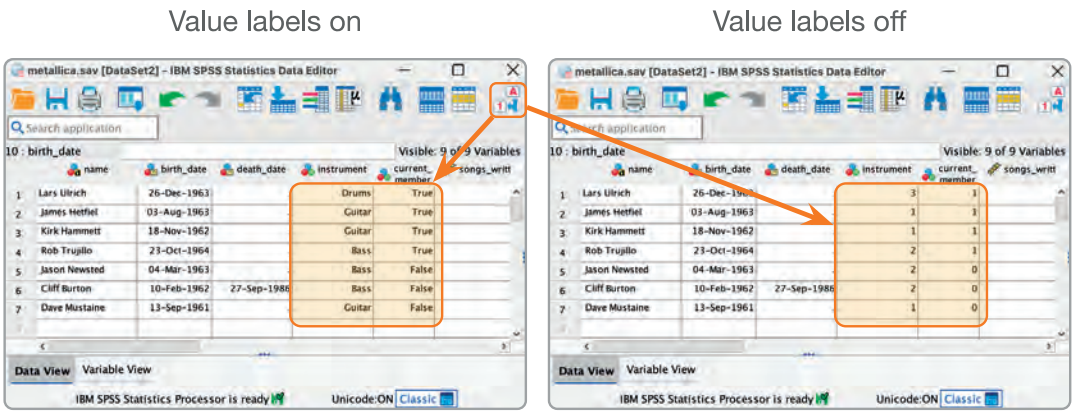
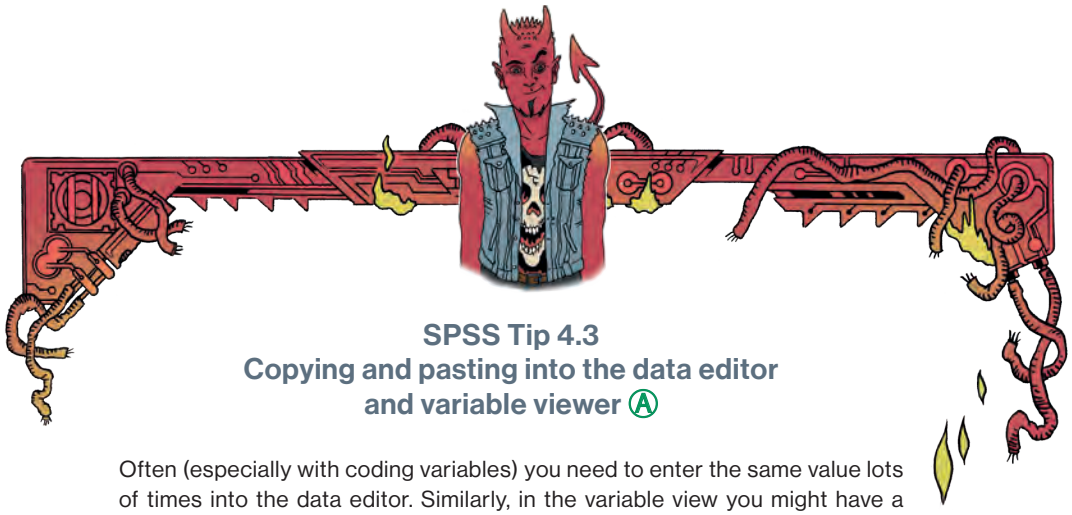


Figure 4.11 Coding values in the data editor with the value labels switched off and on



SPSS Tip 4.3
Copying and pasting into the data editor and variable viewer Ⓐ

Often (especially with coding variables) you need to enter the same value lots of times into the data editor. Similarly, in the variable view you might have a series of variables that all have the same value labels (e.g., variables representing questions on a questionnaire might all have value labels of 0 = never, 1 = sometimes, 2 = always to represent responses to those questions). Rather than typing the same number lots of times, or entering the same value labels

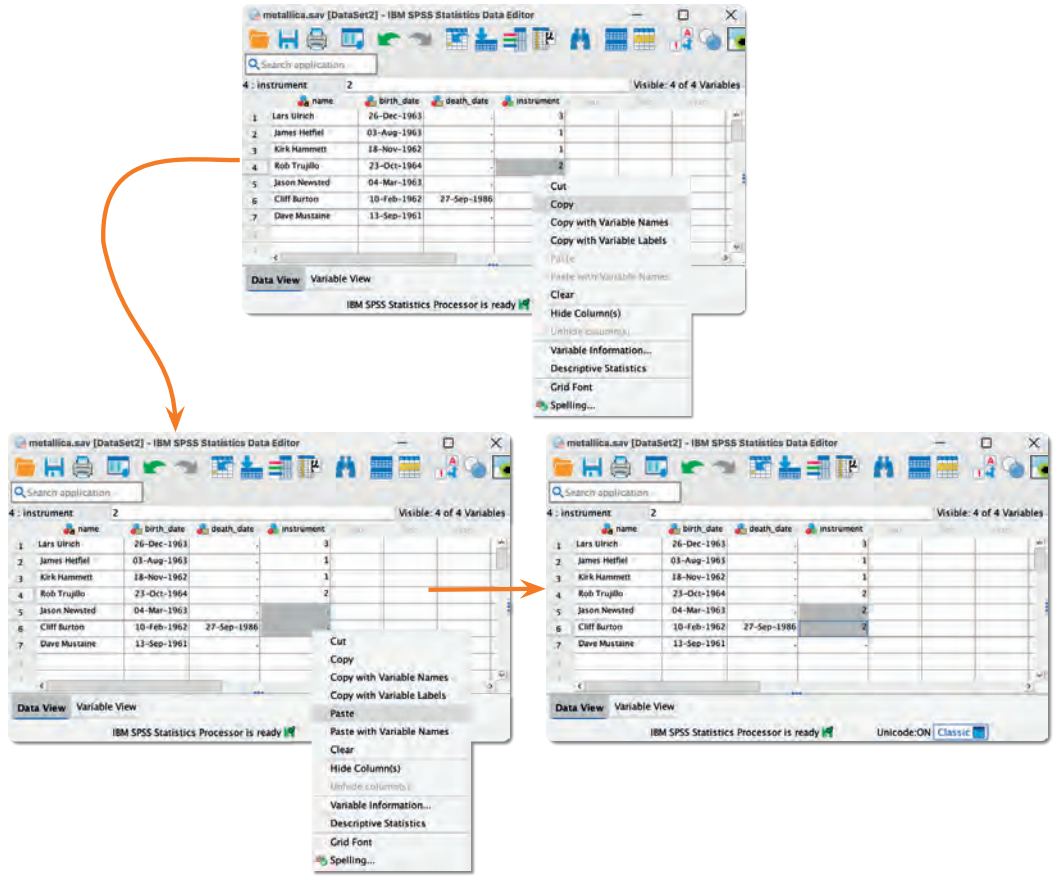
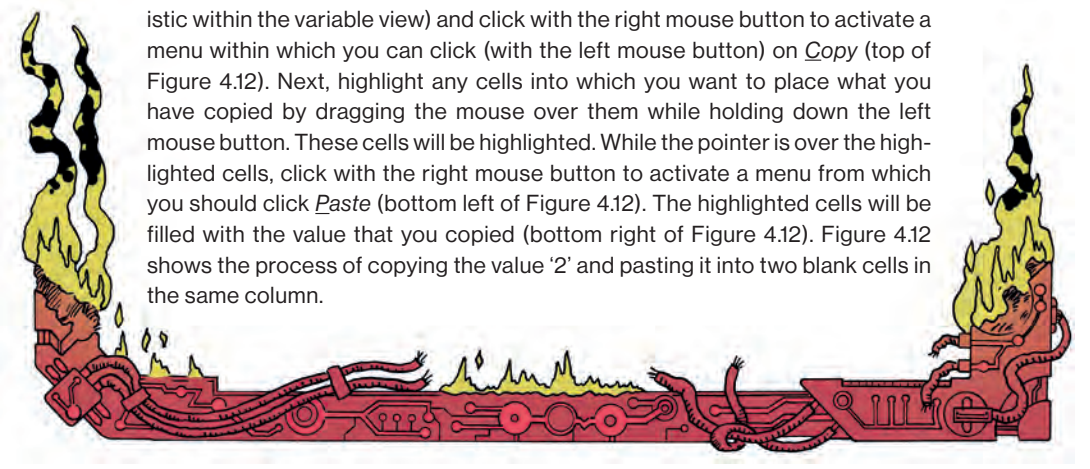
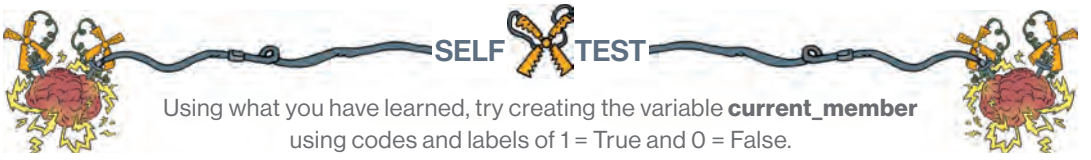


Figure 4.12 Copying and pasting into empty cells

multiple times, you can use the copy and paste functions to speed things up. Select the cell containing the information that you want to copy (whether that is a number or text in the data view, or a set of value labels or another characteristic within the variable view) and click with the right mouse button to activate a menu within which you can click (with the left mouse button) on *Copy* (top of Figure 4.12). Next, highlight any cells into which you want to place what you have copied by dragging the mouse over them while holding down the left mouse button. These cells will be highlighted. While the pointer is over the highlighted cells, click with the right mouse button to activate a menu from which you should click *Paste* (bottom left of Figure 4.12). The highlighted cells will be filled with the value that you copied (bottom right of Figure 4.12). Figure 4.12 shows the process of copying the value '2' and pasting it into two blank cells in the same column.





Using what you have learned, try creating the variable **current_member** using codes and labels of 1 = True and 0 = False.

4.6.6 Creating a numeric variable Ⓐ

Our next variable is **songs_written**, which is a count of the number of songs that the person has contributed to writing.[3] This variable is numeric. Numeric variables are the easiest ones to create because they are the default format. Move back to the variable view using the tab at the bottom of the data editor (Data View Variable View). Go to the cell in row 6 of the column labelled *Name* (under the previous variable you created). Type 'songs_written'. Move into the column labelled Type using the → key on the keyboard. As with the previous variables we have created, SPSS has assumed that our new variable is Numeric , and because our variable *is* numeric we don't need to change this setting.

The scores for the number of songs are integers (have no decimal places), so in the interests of satisfying my anal retentiveness move to the Decimals column and type '0' (or decrease the value from 2 to 0 using) to tell SPSS that you don't want to display decimal places.

Let's continue our good habit of naming variables and move to the cell in the column labelled *Label* and type 'Number of song writing credits'. Finally, the number of songs written is measured on the ratio scale of measurement (see Section 1.5.2) and we can specify this by going to the column labelled *Measure* and selecting Scale from the drop-down list (this will have been done automatically, but it's worth checking).



Why is the **songs_written** variable a 'scale' variable?

3 At the time of writing, the new album *72 Seasons* hasn't been released and the songwriting credits are unknown. So, I'm not including that album. My expectations are incredibly low for this album, but I hope to be proved wrong.



Labcoat Leni's Real Research 4.1
Gonna be a rock 'n' roll singer Ⓐ

Oxoby, R. J. (2008). *Economic Enquiry, 47*(3), 598–602.

AC/DC are one of the best-selling hard rock bands in history, with around 100 million certified sales, and an estimated 200 million actual sales. In 1980 their original singer Bon Scott died. He was replaced by Brian Johnson, who has been their singer ever since.[4] Debate rages with unerring frequency within the rock music press over who is the better frontman. The conventional wisdom is that Bon Scott was better, although personally, and I seem to be somewhat in the minority here, I prefer Brian Johnson. Anyway, Robert Oxoby in a playful paper decided to put this argument to bed once and for all (Oxoby, 2008).

Using a task from experimental economics called the ultimatum game, individuals are assigned the role of either proposer or responder and paired randomly. Proposers are allocated \$10 from which they have to make a financial offer to the responder (i.e., \$2). The responder can accept or reject this offer. If the offer is rejected neither party gets any money, but if the offer is accepted the responder keeps the offered amount (e.g., \$2), and the proposer keeps the original amount minus what they offered (e.g., \$8). For half of the participants the song 'It's a long way to the top' sung by Bon Scott was playing in the background, for the remainder 'Shoot to thrill' sung by Brian Johnson was playing. Oxoby measured the offers made by proposers, and the minimum offers that responders accepted (called the minimum acceptable offer). He reasoned that people would accept lower offers and propose higher offers when listening to something they like (because of the 'feel-good factor' the music creates). Therefore, by comparing the value of offers made and the minimum acceptable offers in the two groups he could see whether people have more of a feel-good factor when listening to Bon or Brian. The offers made (in \$) are[5] as follows (there were 18 people per group):

- Bon Scott group: 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5
- Brian Johnson group: 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5

Enter these data into the SPSS data editor, remembering to include value labels, to set the *measure* property, to give each variable a proper label and to set the appropriate number of decimal places. Answers are on the companion website and my version of how this file should look can be found in **acdc.sav**.

4 Well, until all that weird stuff with W. Axl Rose in 2016, which I'm trying to pretend didn't happen.

5 These data are estimated from Figures 1 and 2 in the paper because I couldn't get hold of the author to get the original data files.

Once the variable has been created, you can return to the data view by clicking on the *Data View* tab at the bottom of the data editor (Data View Variable View). The contents of the window will change, and you'll notice that the fourth column now has the label *songs_written*. To enter the data, click the white cell at the top of the column and type the first value, 111. Because we're entering scores down the column the most sensible way to record this value in this cell is to press the ↓ key on the keyboard. This action moves you down to the next cell, and the number 111 is stored in the cell above. Enter the next number, 112, and then press ↓ to move down to the next cell, and so on.

4.6.7 Creating currency variables Ⓐ

The next variable in the data is the net worth of each band member. This variable is a currency variable, and is recorded in US dollars. It is also likely to be highly inaccurate because the data were gathered from www.celebritynetworth.com (in February 2023), a site that more than likely doesn't know the actual net worth of anyone. However, Lars has blocked my number so I can't get more accurate estimates until the restraining order is lifted. If you happen to be from a country that uses dollars as a currency then SPSS has your back, because there is an option to set a variable to be a currency variable ($) by clicking Dollar when you set the variable type. If, like me, you live somewhere that uses a different currency then you have to set up a custom currency. Let's get the net worth variable set up and then talk about custom currencies.

Making sure you're in the variable view using the tab at the bottom of the data editor (Data View Variable View), move to the cell in row 7 of the column labelled *Name* (under the previous variable you created) and type 'net_worth'. Move into the column labelled Type using the → key on the keyboard, click ... to activate the *Variable Type* dialog box, and select Dollar (bottom right of Figure 4.9). On the right of the dialog box is a list of convenient formats that are a shortcut for setting the width and decimal places of the variable. In this list the $ symbol is the currency, the # symbol represents a number and a period represents a decimal place. For example, there are formats representing a variable with single-digit amounts ($#), two-digit amounts ($##), three-digit amounts ($###), three-digit amounts with two decimal places ($###.##) and so on. As you select these options the values in the boxes labelled *Width* and *Decimal Places* change. For example, if you select $## then *Width* becomes 3 (to accommodate the dollar sign and two digits) and *Decimal Places* becomes zero, whereas if you select $###.## *Width* becomes 7 (to accommodate the dollar sign, the decimal place and five digits) and *Decimal Places* becomes 2. Therefore, you have the choice of selecting a convenience format from the list, or using a custom format by entering values into *Width* and *Decimal Places*. For net worth we are entering values in millions of dollars because the data are approximate and rounded to at least the nearest million; we can save ourselves typing a lot of zeros if we express wealth in millions of dollars (as in Table 4.1). The largest values are in the hundreds of millions (e.g., $350 million) so we can select $### because we want three digits and no decimal places (see bottom right of Figure 4.9).

We can go a step further than this and set up a custom currency that displays the fact that we're working in millions of dollars. To do this, select *Edit* > Options... to open the options dialog box and then select the currency tab (Currency) to get the options for setting up custom currencies (top left of Figure 4.13). We can in SPSS store up to five custom currencies labelled CCA, CCB, CCC, CCD and CCE.[6] These five placeholders are listed in the box labelled *Custom Output Formats:* and to store a currency in one of these five slots select it from this list. In Figure 4.13 we have selected CCA, so whatever we specify in this dialog box will be saved as custom currency A when we click OK. There is a drop-down box of currency symbols, select *Dollar($)* from that list and doing so will automatically fill the box labelled *Prefix:* with a dollar symbol. This means that all values with this currency will have a $ placed in front of them in the data editor and output. There is also a box labelled *Suffix:* and whatever

6 I assume theses codes consist of the acronym 'CC' for custom currency and the letters A to E.

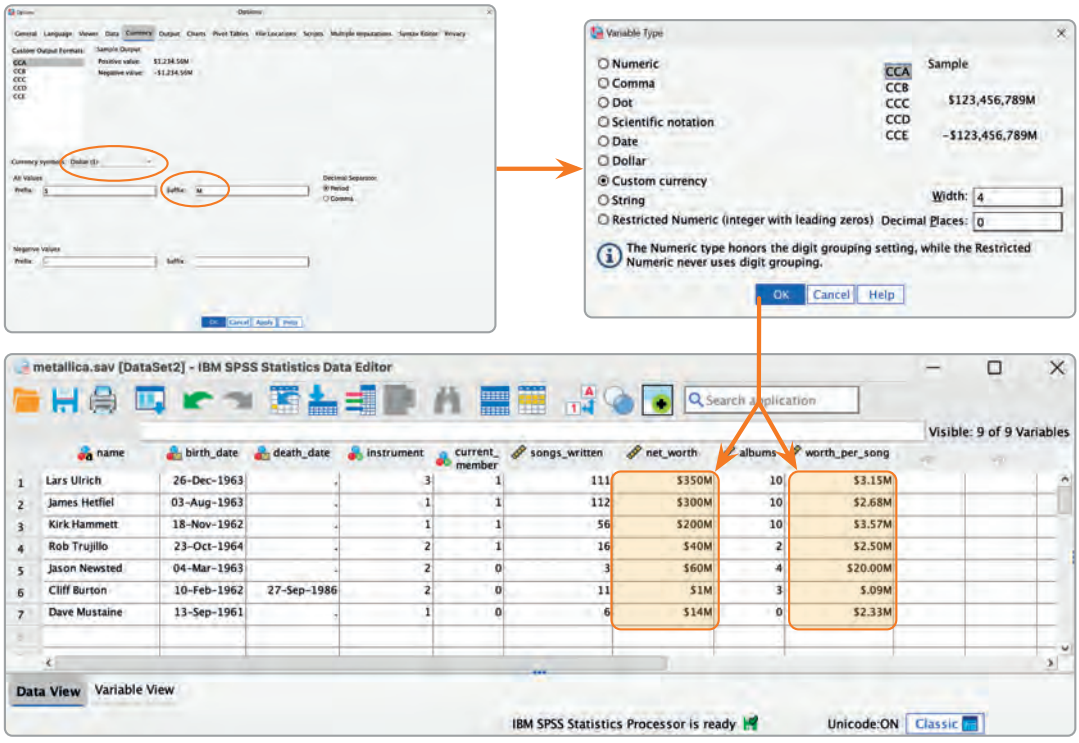
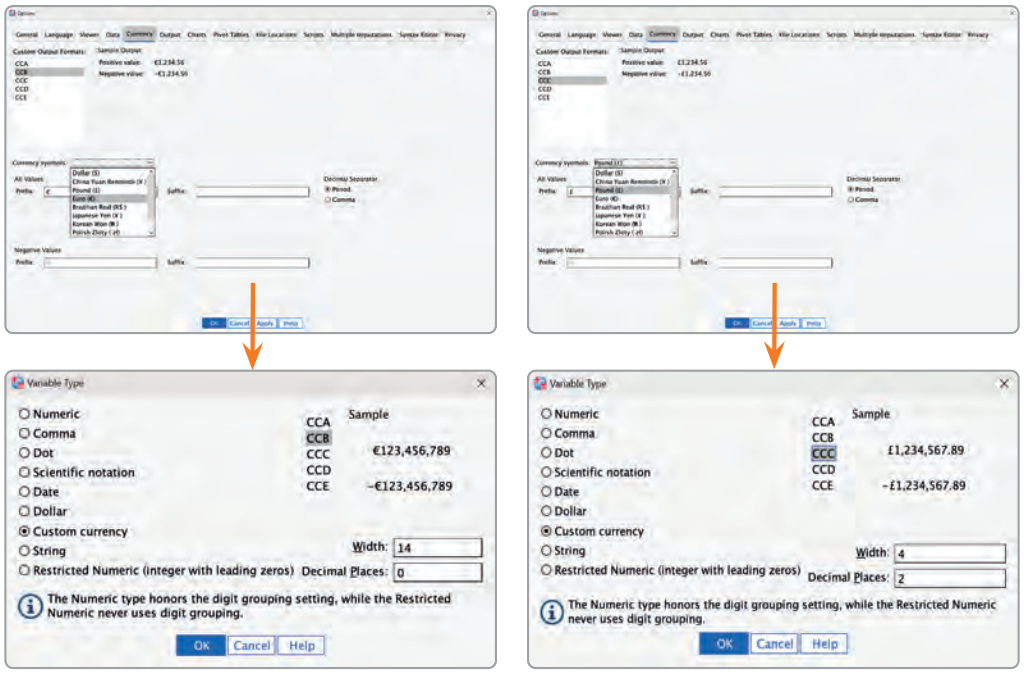


Figure 4.13 Setting a custom currency

we type in this box will be appended to any data values that have this currency. Let's type an uppercase M (for Millions) into that box. Now, if a variable is set up to have custom currency A as its type, then if we enter a value such as 350 it will appear in the data editor as $350M. This is handy for reminding is we're working in millions of dollars. Click OK to store this custom currency, then go back to the data editor, go into the variable view(Data View Variable View) and activate the variable type dialog box for net worth (as in Section 4.6.3). Now select ⦿ Custom currency. Notice that a list appears with CCA to CCE listed. If we select CCA (as in the top right Figure 4.13) then net worth will be set to the custom currency that we just created, and the area labelled *Sample* shows us an example of what this currency format looks like. It confirms that values will be prefixed with $ and have a suffix of M. Click OK to save the change, and then repeat the process for **worth_per_song**. Notice that in the data editor the values now appear with M after them to remind us that we are working in millions of dollars (bottom of Figure 4.13).

The other use of custom currencies is if you live in a country that has a currency that doesn't use a dollar symbol, or will work with data from countries that use these currencies. If you never intend to work with any currencies other than dollars then skip this next bit. Setting up a custom currency is mainly about displaying currency symbols in the data editor. The symbols available are $ (dollars), £ (pounds), € (euros), ¥ (yen, yuan renminbi), ₽ (ruble), ₩ (won) and ₹ (rupee), which cover a wide range of currencies and countries. The list doesn't cover every currency symbol (e.g., dinar, rial and dirham are missing). I am English (£) but identify as a European (€) so setting up custom currencies for these regions might be useful for me. Figure 4.14 illustrates the process. Again, we begin by opening the currency options by selecting *Edit* > Options... and the currency tab (Currency). Let's store euros as a custom currency B by selecting CCB in the box labelled *Custom Output Formats:* and selecting *Euro (€)* in the drop-down box of currency symbols. Note that the *Prefix:* box now contains €, which means that all values with this currency will have € placed in front of them in the data editor and output. We don't need to add a suffix. Click OK to store this custom currency. Once in the



Setting up euros (€) Setting up pounds (£)

Figure 4.14 Setting more custom currencies

variable view of the data editor we could select a variable, activate the variable type dialog box and select ⦿ Custom currency. If we select CCB (as in the bottom left of Figure 4.14) then the variable will be in the custom currency of euros area labelled *Sample* shows us an example this currency format, which confirms that values will be prefixed with €. Click OK to save the change. The right-hand side of Figure 4.14 shows the process for setting up a current currency of pounds (£), which is the same as for euros except select CCC and select *Pound (£)* in the drop-down currency symbols in the options dialog box, and set the ⦿ Custom currency as CCC when we set the variable type.



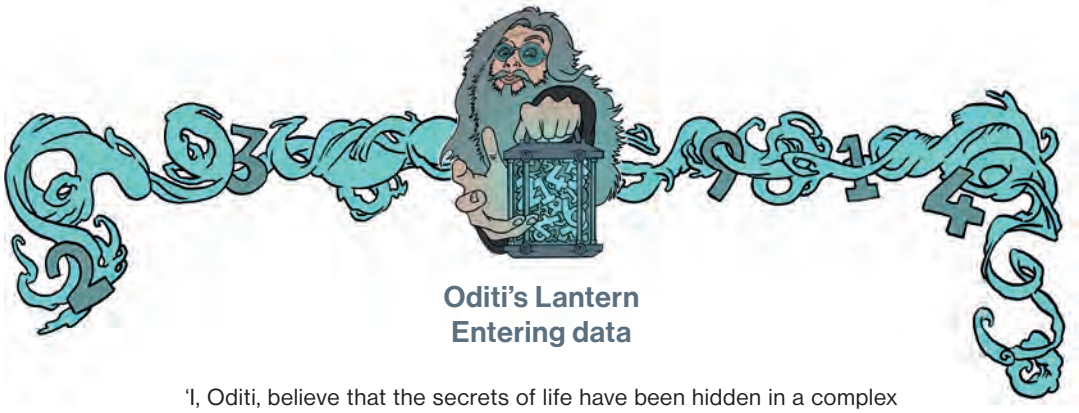
Having created the first four variables with a bit of guidance, try to enter the rest of the variables in Table 4.1 yourself.

4.6.8 Missing values Ⓐ

Although we strive to collect complete sets of data, often scores are missing. Missing data can occur for a variety of reasons: in long questionnaires participants accidentally (or, depending on how paranoid you're feeling, deliberately to irritate you) miss out questions; in experimental procedures mechanical faults can lead to a score not being recorded; and in research on delicate topics (e.g., sexual behaviour) participants may exercise their right not to answer a question. However, just because we have missed out on some data for a participant, we don't have to ignore the data we do have (although it creates statistical difficulties). The simplest way to record a missing score is to leave the cell in the data editor empty (which is what we have done for **death_date**), but it can be helpful to tell SPSS explicitly that a score is missing. We do this, much like a coding variable, by choosing a

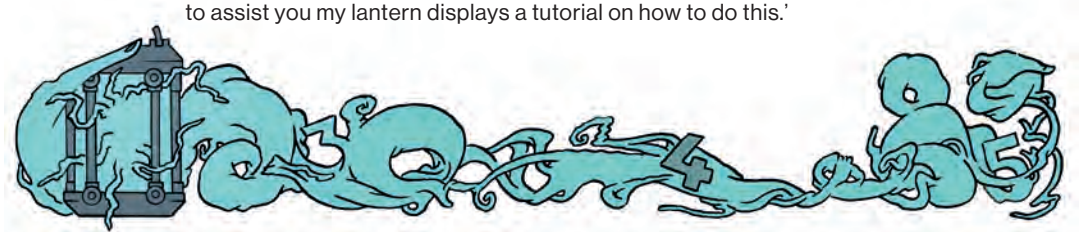
number to represent the missing data point. You then tell SPSS to treat that number as missing. For obvious reasons, it is important to choose a code that cannot naturally occur in the data. For example, if we use the value 9 to code missing values and several participants genuinely scored 9, then SPSS will wrongly treat those scores as missing. You need an 'impossible' value, so people usually pick a score greater than the maximum possible score on the measure. For example, in an experiment in which attitudes are measured on a 100-point scale (so scores vary from 1 to 100) a reasonable code for missing values might be 101, 999 or, my personal favourite, 666 (because missing values *are* the devil's work).

To specify missing values click in the column labelled Missing in the variable view (Data View Variable View) and then click ⊡ to activate the *Missing Values* dialog box in Figure 4.15. By default SPSS assumes that no missing values exist, but you can define them in one of two ways. The first is to select discrete values (by clicking on the radio button next to where it says *Discrete missing values*), which are single values that represent missing data. SPSS allows you to specify up to three values to represent missing data. The reason why you might choose to have several numbers to represent missing values is that you can assign a different meaning to each discrete value. For example, you could have the number 8 representing a response of 'not applicable', a code of 9 representing a 'don't know' response and a code of 99 meaning that the participant failed to give any response. SPSS treats these values in the same way (it ignores them), but different codes can be helpful to remind you of why a particular score is missing. The second option is to select a range of values to represent missing data, which is useful in situations in which it is necessary to exclude data falling between two points. So, we could exclude all scores between 5 and 10. With this last option you can also (but don't have to) specify one discrete value.



Oditi's Lantern
Entering data

'I, Oditi, believe that the secrets of life have been hidden in a complex numeric code. Only by "analysing" these sacred numbers can we reach true enlightenment. To crack the code I must assemble thousands of followers to analyse and interpret these numbers (it's a bit like the chimps and typewriters theory). I need you to follow me. To spread the numbers to other followers you must store them in an easily distributable format called a "data file". You, my follower, are loyal and loved, and to assist you my lantern displays a tutorial on how to do this.'



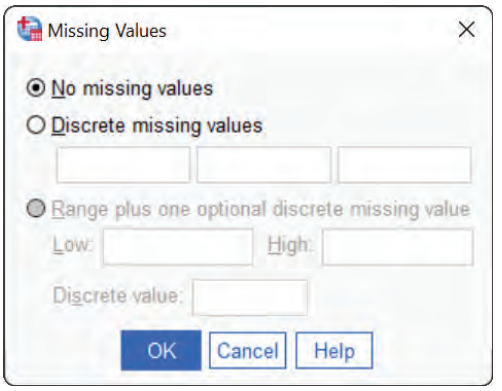
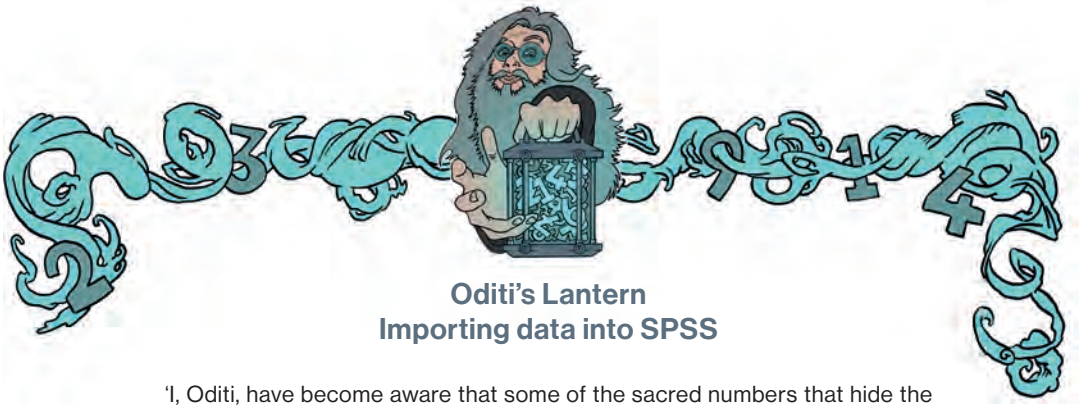


Figure 4.15 Defining missing values

4.6.9 Importing data Ⓐ

We can import data into SPSS from other software packages such as Microsoft Excel, R, SAS, and Systat by using the *File* > *Import Data* menu and selecting the corresponding software from the list. If you want to import from a package that isn't listed (e.g., R or Systat) then export the data from these packages as tab-delimited text data (*.txt* or *.dat*) or comma-separated values (*.csv*) and select the *Text Data* or *CSV Data* options in the menu.



Oditi's Lantern
Importing data into SPSS

'I, Oditi, have become aware that some of the sacred numbers that hide the secrets of life are contained within files other than those of my own design. We cannot afford to miss vital clues that lurk among these rogue files. Like all good cults, we must convert all to our cause, even data files. Should you encounter one of these files, you must convert it to the SPSS format. My lantern shows you how.'



4.7 SPSS syntax Ⓐ

Once your data are in, you'll want to do stuff with it. For most people 'doing stuff' involves navigating through the menus of SPSS and selecting the procedures you want to carry out using dialog boxes. This book assumes that's what you'll do. However, you can also get SPSS to do stuff using a language of commands known as SPSS syntax. When you use syntax you're typing instructions for SPSS to follow instead of generating those instructions by clicking on the screen. Most beginners ignore syntax, but using syntax opens up additional functions and can save time in the long run. Here are some good reasons to consider using syntax.

1. **Reproducibility**: The major problem with using dialog boxes is that your work is not reproducible. That is, unless you keep very detailed and public notes of what you clicked on in which dialog boxes it's going to be almost impossible for you or anyone else to exactly reproduce your SPSS output. Using syntax means that you have a written record of everything you did, which is an important part of engaging in open science practices (Section 3.6).
2. **Added functionality**: There are things you can do with syntax that you can't do through dialog boxes (admittedly most of these things are advanced, but I will periodically show you some nice tricks using syntax).
3. **Time saving**: If you carry out very similar analyses on different data sets, then it takes ages to keep clicking through the dialog boxes every time. If you run the analysis once and save the syntax as you go along then you can quickly adapt the syntax to new data sets, which is usually quicker than going through dialog boxes.

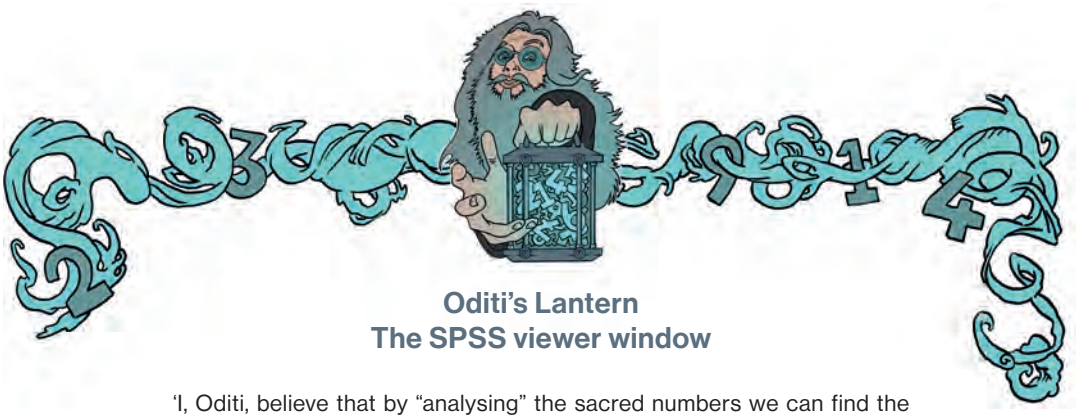
Like grammatical rules when we write, there are rules that ensure SPSS 'understands' the syntax. For example, each line must end with a full stop. If you make a syntax error (i.e., break one of the rules), SPSS produces an error message in the viewer window. The messages can be indecipherable until you gain experience of translating them, but they helpfully identify the line in the syntax window in which the error occurred. Each line in the syntax window is numbered so you can easily find the line in which the error occurred, even if you don't understand what the error is! Learning SPSS syntax is time-consuming, so in the beginning the easiest way to generate syntax is to use dialog boxes to specify the analysis you want to do and then click 

(many dialog boxes have this button). Clicking this button converts the analysis you specified in the dialog box into syntax and pastes it into your SPSS workbook or a syntax window (more on those soon). Using dialog boxes in this way is a good way to get a feel for syntax.

4.8 The SPSS viewer Ⓐ

The *SPSS viewer* is a separate window that displays the output of any procedures in SPSS: tables of results, plots, error messages and pretty much everything you could want, except for photos of your cat. Although the *SPSS viewer* is all-singing and all-dancing, my prediction in previous editions of this book that it will one-day include a tea-making facility have not come to fruition (IBM take note☺). You can have the viewer in classic mode (files are saved with the extension *.spv*) or workbook mode (which uses *.spwb* as the file extension).[7] In the data editor (look back at Figures 4.3 and 4.8) you can switch modes by clicking  (to switch out of classic mode into workbook mode) or  (to switch out of workbook mode into classic mode). The main difference between the formats is that the newer workbook format integrates SPSS syntax (explained in Section 4.7) and output into a single

7 Version 28 onwards.



Oditi's Lantern
The SPSS viewer window

'I, Oditi, believe that by "analysing" the sacred numbers we can find the answers to life. I have given you the tools to spread these numbers far and wide, but to interpret these numbers we need "the viewer". The viewer is like an X-ray that reveals what is beneath the raw numbers. Use the viewer wisely, my friends, because if you stare long enough you will see your very soul. Stare into my lantern and see a tutorial on the viewer.'





SPSS Tip 4.4
Funny numbers Ⓐ

SPSS sometimes reports numbers with the letter 'E' placed in the mix just to confuse you. For example, you might see a value such as 9.612E−02. Many students find this notation confusing. This notation means 9.612×10^{-2} , which might be a more familiar notation or could be even more confusing. Think of E−02 as meaning 'move the decimal place 2 places to the left', so 9.612E−02 becomes 0.09612. If the notation reads 9.612E−01, then that would be 0.9612, and 9.612E−03 would be 0.009612. Conversely, E+02 (notice the minus sign has changed) means 'move the decimal place 2 places to the right', so, 9.612E+02 becomes 961.2.

notebook-style document, whereas in classic mode SPSS syntax is run in a separate window to the viewer. For the reasons explained in Section 4.7, consider using syntax. If you decide to use syntax, the workbook format makes for a more integrated and satisfying experience. This book doesn't assume you use a particular viewer mode.

4.8.1 Classic mode Ⓐ

In both classic (Figure 4.16) and workbook modes the window layout has a left-hand pane containing a tree diagram of the output, which helps you to navigate parts of the output (useful when you have conducted tonnes of analyses). The tree structure is self-explanatory: every time you do something in SPSS (such as plot data or run a statistical procedure), this procedure is listed as a main heading. The right-hand pane has a large space in which all output is displayed. Plots and tables can be edited by double-clicking on them. For example, in Figure 4.16 I ran a graphing procedure followed by a univariate analysis of variance (ANOVA), and these names appear as main headings in the tree diagram. For each procedure there are subheadings that represent different parts of the analysis. For example, in the ANOVA procedure, which you'll learn more about later in the book, there are sections such as *Tests of Between-Subjects Effects* (this is the table containing the main results). You can skip to any one of these sub-components by clicking on the appropriate branch of the tree diagram. So, if you wanted to skip to the between-group effects you would move the on-screen arrow to the left-hand portion of the window and click where it says *Tests of Between-Subjects Effects*. This action will highlight this part of the output in the main part of the viewer (see SPSS Tip 4.5). You can rename a heading by clicking on it (e.g., select the text part of Univariate Analysis of Variance in Figure 4.16), which should highlight the text, and typing something more informative.



Oditi's Lantern
Editing tables

'I, Oditi, impart to you my loyal posse the knowledge that SPSS will conceal the secrets of life within tables of output. Like the personality of the author of this book, these tables appear flat and lifeless; however, if you give them a poke they have hidden depths. Often you will need to seek out the hidden codes within the tables. To do this double-click on them. This will reveal the "layers" of the table. Stare into my lantern and find out how.'



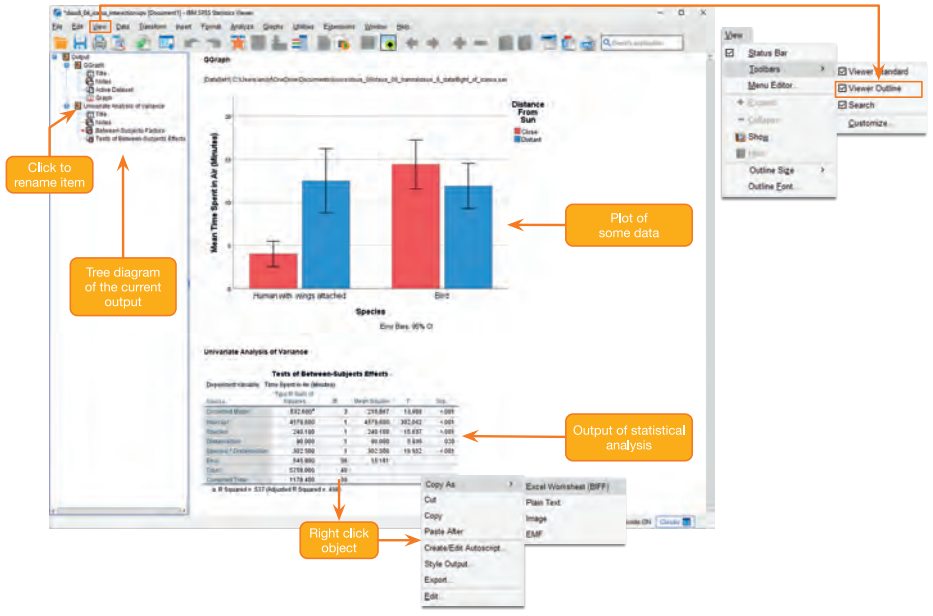


Figure 4.16 The SPSS viewer in classic mode

The classic and workbook modes share some icons with the data editor such as those to open ( ) and save ( ) files, print ( ), list recently used dialog boxes ( ), and summarize variables ( ). In classic mode you'll also see icons to go to cases ( ) and variables ( ) and to customize the toolbar ( ).



This icon is also common to both modes and it returns you to the data editor. I'm not sure what the big red star is all about.

In classic mode, there are some other icons, but a useful set of icons for navigating the output must be activated by selecting *View* > *Toolbars* > ☑ Viewer Outline (as in Figure 4.16).



This icon displays a print preview in a new window.



This icon opens a dialog box for exporting the contents of the viewer (see Section 4.9).



This icon *promotes* the currently active part of the tree structure to a higher branch of the tree. For example, in Figure 4.16 the *Tests of Between-Subjects Effects* are a sub-component under the heading of *Univariate Analysis of Variance*. If we wanted to promote this part of the output to a higher level (i.e., to make it a main heading) then we'd click this icon.



This icon is the opposite of the above: it *demotes* parts of the tree structure. For example, in Figure 4.16 if we didn't want the *Univariate Analysis of Variance* to be a unique section we could select this heading and click this icon to demote it so that it becomes part of the previous heading (the *Graph* heading). This icon is useful for combining parts of the output relating to a specific research question.



This icon collapses parts of the tree structure: it hides the sub-components under a selected heading. For example, in Figure 4.16 if we selected the heading *Univariate Analysis of Variance* and clicked this icon, all the subheadings would disappear. The sections that disappear from the tree structure don't disappear from the output itself; the tree structure is merely condensed. This option is useful when you have conducted lots of analyses and the tree diagram has become very complex.



This icon expands any collapsed sections. By default, main headings are displayed in the tree diagram in their expanded form. If you have opted to collapse part of the tree diagram (using the icon above) then use this icon to undo your dirty work.



This icon and the following one show or hide parts of the output. If you select part of the output in the tree diagram and click this icon, that part of the output disappears. Some procedures produce copious amounts of output, so this icon can be useful to hide the less relevant parts so you can differentiate the wood from the trees. The output isn't erased, just hidden from view. This icon is like the collapse icon above except that it affects the output rather than the tree structure.



This icon undoes the work of the previous one: if you have hidden part of the output from view then clicking this icon makes it reappear. By default, this icon is not active because none of the output is hidden, but it becomes active if you use the icon above to hide stuff.



Given this icon looks like it has a slot in which to insert a CD, I had high hopes for it activating some blistering thrash metal, which, as everyone knows, is the best music to have on when doing stats. I was disappointed to discover that rather than blasting out Anthrax at high volume it instead inserts a new heading into the tree diagram. For example, if you had several statistical tests that related to one of many research questions you could insert a main heading and demote the headings of the relevant analyses to fall under this new heading.

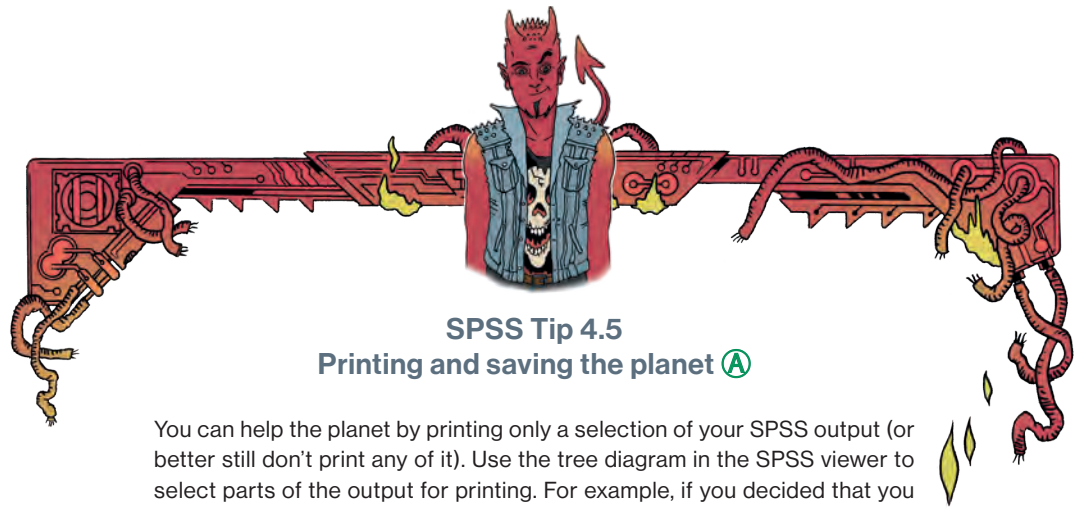


Assuming you had done the above, you can use this icon to provide your new heading with a title. The title you type appears in the output. So, you might have a heading like 'Hypothesis 1: thrash metal music helps you to understand stats' which tells you that the analyses under this heading relate to your first hypothesis.



This final icon is used to place a text box in the output window. You can type anything into this box. In the context of the previous two icons, you might use a text box to explain your hypothesis. For example, 'Hypothesis 1 is that thrash metal music helps you to understand stats. This hypothesis stems from research showing that a 10-minute blast of classical music can enhance learning and memory (Rauscher et al., 1993). Thrash metal music has similar rhythmic and melodic complexity to classical music so why not?'

The final thing to note is that if you right-click on any part of the output, it opens a menu with various actions that you can perform on that part of the output (see Figure 4.16). For example, right-clicking on a table or plot gives you the option to copy it to the clipboard in various formats, to open the export dialog box (see Section 4.9) or to open it in a new window for editing (which can also be achieved by double-clicking on things in the viewer).



SPSS Tip 4.5
Printing and saving the planet Ⓐ

You can help the planet by printing only a selection of your SPSS output (or better still don't print any of it). Use the tree diagram in the SPSS viewer to select parts of the output for printing. For example, if you decided that you wanted to print a particular plot, click on the word *GGraph* in the tree structure to highlight the plot in the output. Then, in the *Print* menu you can print only the selected part of the output (Figure 4.17). Note that if you click a main heading (such as *Univariate Analysis of Variance*) SPSS will highlight all the subheadings under that heading, which is useful for printing all the output from a single statistical procedure.

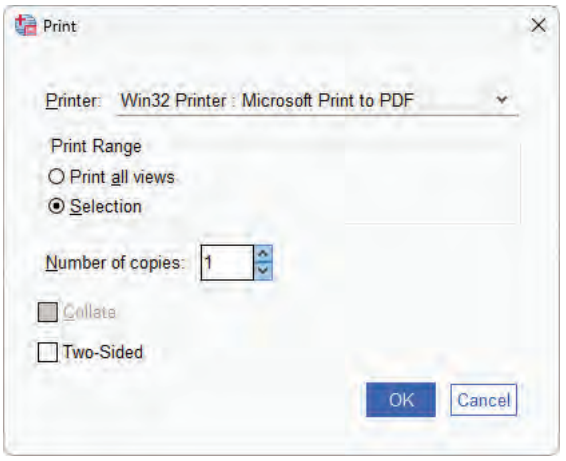
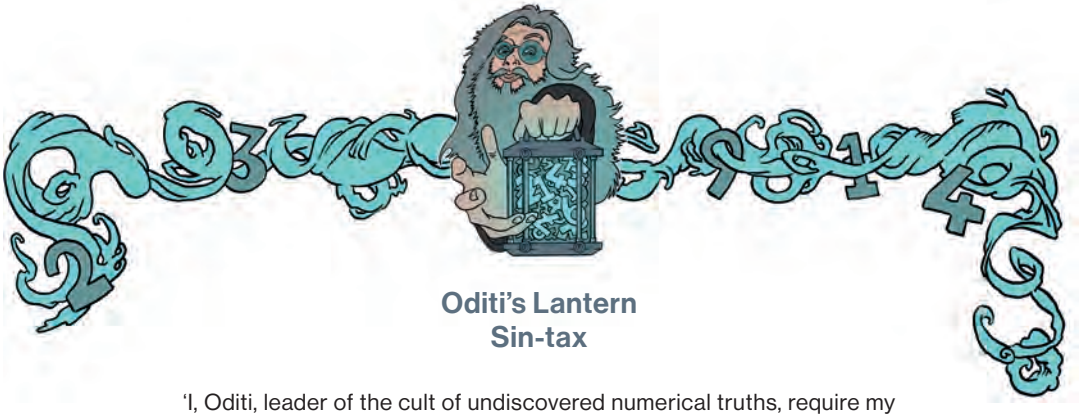


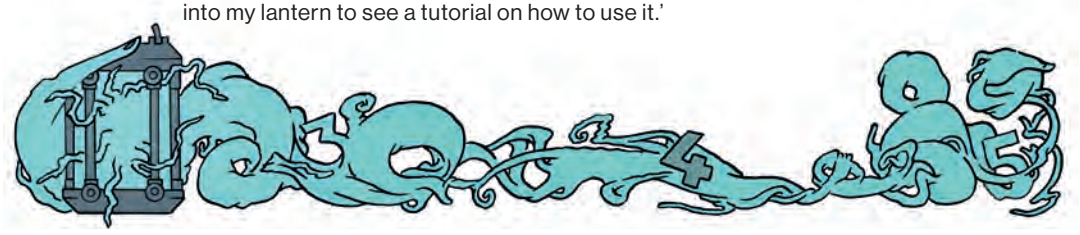
Figure 4.17 Printing only the selected parts of SPSS output

If you want to use syntax in classic mode, you need to open a **syntax editor** window, like the one in Figure 4.18, using *File* > *New* > Syntax. The area on the right (the *command area*) is where you type syntax commands, and on the left is a navigation pane (like the viewer window). When you have a large file of syntax commands the navigation area helps you find the bit of syntax that you need.



Oditi's Lantern
Sin-tax

'I, Oditi, leader of the cult of undiscovered numerical truths, require my brethren to focus only on the discovery of those truths. To focus their minds I shall impose a tax on sinful acts. Sinful acts (such as dichotomizing a continuous variable) can distract from the pursuit of truth. To implement this tax, followers will need to use the sin-tax window. Stare into my lantern to see a tutorial on how to use it.'



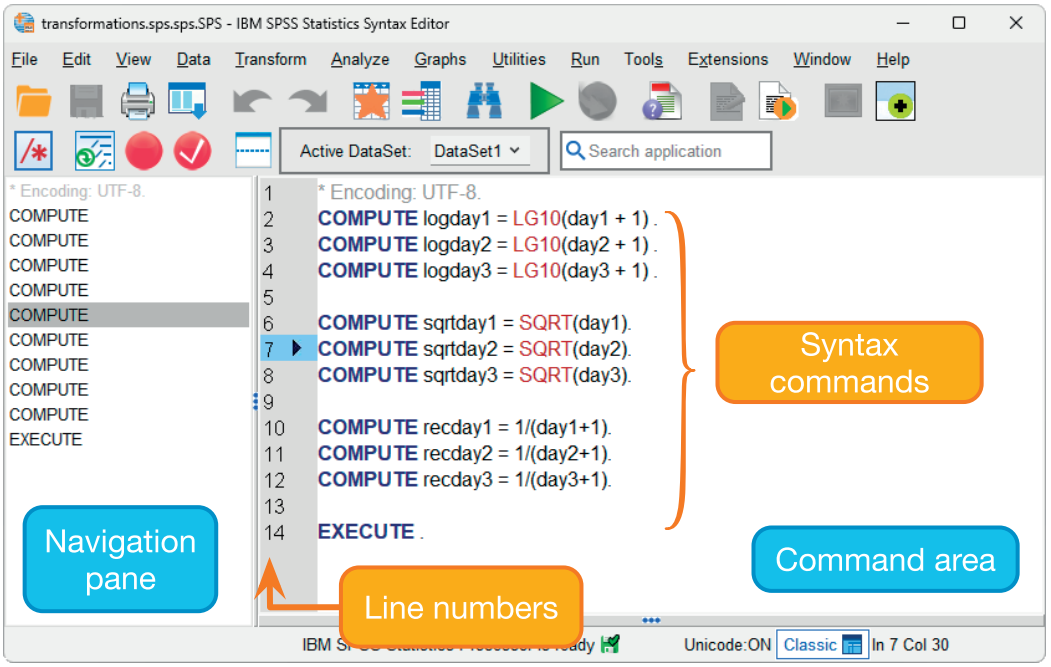


Figure 4.18 A syntax window with some syntax in it

Once you've typed in your syntax you run it using the *Run* menu. *Run* > All will run all the syntax in the window, or you can highlight a selection of syntax using the mouse and select *Run* > Selection (or click ▶ in the syntax window) to process the selected syntax. You can also run the syntax a command at a time from either the current command (*Run* > *Step Through* > *From Current*), or the beginning (*Run* > *Step Through* > *From Start*). You can also process the syntax from the cursor to the end of the syntax window by selecting *Run* > To End.

4.8.2 Workbook mode Ⓐ

In workbook mode, the viewer looks largely the same as classic mode (compare Figure 4.16 with Figure 4.19). The big difference is that in workbook mode there are editable sections for typing SPSS syntax, whereas in classic mode syntax is used through a separate window. A workbook is, therefore, made up of three components: syntax paragraphs (where you write syntax), rich text paragraphs (where you can write notes and format the text) and output from your analyses/syntax. You can use a workbook like a classic viewer window, but also insert syntax paragraphs. You would type or paste syntax commands into these and run them by clicking Run Syntax. Because SPSS syntax is embedded into the workbook it has some icons not seen in classic mode:



This icon inserts a new syntax paragraph.



This icon inserts a new text paragraph. By default the new paragraph will be given a heading of Paragraph. As with any heading, you can rename it by clicking on Paragraph in the tree diagram; the word 'Paragraph' should become highlighted and you can replace it with a more informative label. In Figure 4.19 I have renamed the paragraph as Summary.



This icon runs all the syntax in the workbook.



This icon hides all syntax paragraphs in the workbook. To view them click this icon again.

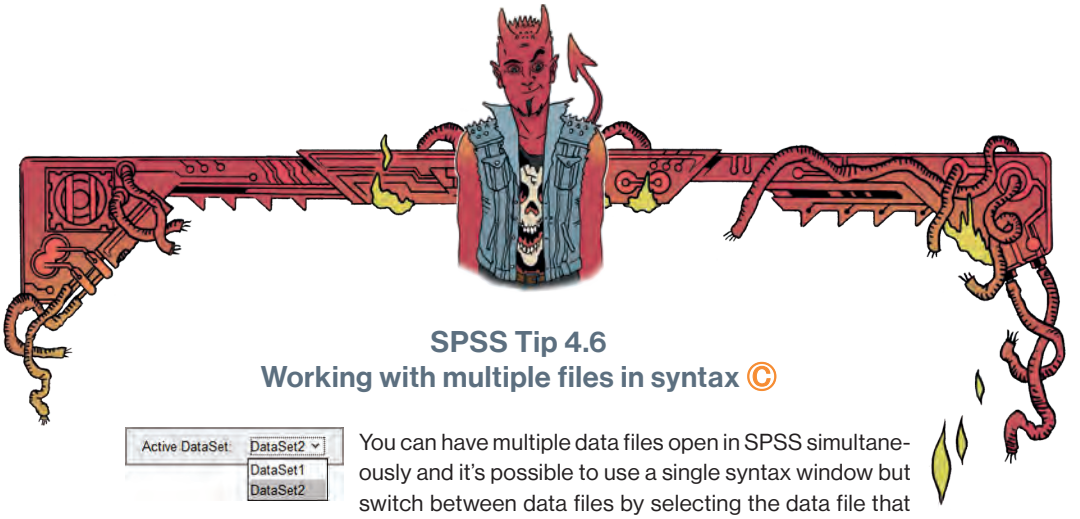


This icon clears all output in the workbook.



This icon takes you to the IBM SPSS Statistics syntax help files online.

As in classic mode, if you right-click on anything in workbook it opens a menu that allows you to copy the selected object to your clipboard or export it into several formats. In the workbook, this menu also gives you the option to show or hide the selected object.

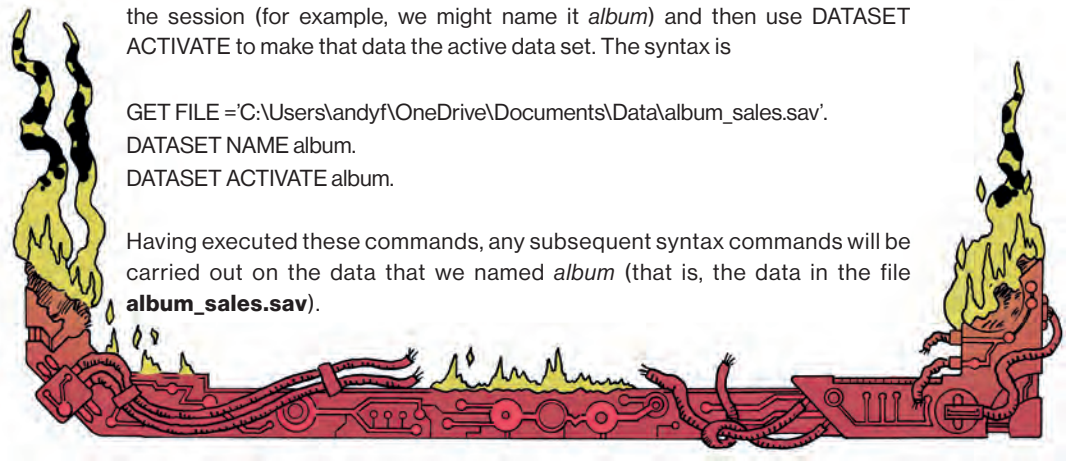


SPSS Tip 4.6
Working with multiple files in syntax ©

You can have multiple data files open in SPSS simultaneously and it's possible to use a single syntax window but switch between data files by selecting the data file that you want to run the syntax commands on before you run them using the drop-down list. The problem with this method is that data files are labelled only as *DataSet1*, *DataSet2*, etc. in the order you open the files. To run the syntax again in a new session you'd need to open these files in the same order for everything to work. You can instead hard-code the data file for a particular block of syntax. For example, let's assume we want to use the file **album_sales.sav** (from Chapter 9) and we have stored it in a folder called *Data* in our *Documents* folder on OneDrive. We can get SPSS to load this file using the GET FILE command and specifying the file path. We can then use DATASET NAME to give the data set a name within the session (for example, we might name it *album*) and then use DATASET ACTIVATE to make that data the active data set. The syntax is

```
GET FILE ='C:\Users\andyf\OneDrive\Documents\Data\album_sales.sav'.
DATASET NAME album.
DATASET ACTIVATE album.
```

Having executed these commands, any subsequent syntax commands will be carried out on the data that we named *album* (that is, the data in the file **album_sales.sav**).



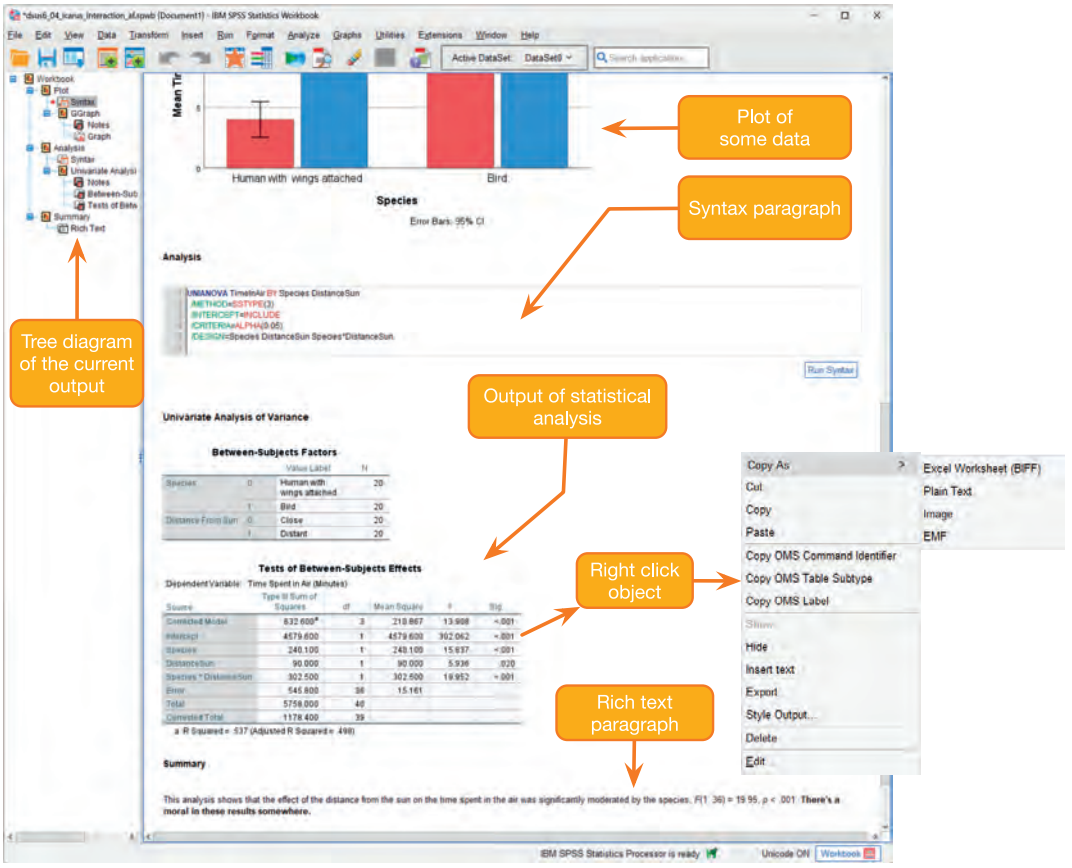


Figure 4.19 The SPSS viewer in workbook mode

4.9 Exporting SPSS output Ⓐ

If you want to share your SPSS output with other people who don't have access to SPSS you need to export the output into a format that they can open on their computer. There are various options such as exporting in portable document format (PDF), which can be read by various free apps, HTML, which can be opened in any web browser, or XML, which can be opened in Microsoft Word. There used to be a standalone app to read SPSS output files (IBM SPSS Smartreader), but with the introduction of workbooks in version 28 of SPSS, this app appears to be discontinued (at least for now).

4.10 Saving files and restore points Ⓐ

Most of you should be familiar with how to save files. Like most software, SPSS has a save icon and you can use *File* > Save or *File* > *Save as* … or Ctrl + S (⌘ + S on Mac OS). If the file hasn't been saved previously then initiating a save will open the *Save As* dialog box (see Figure 4.20). SPSS will save whatever is in the window that was active when you initiated the save; for example, if you are in the data editor when you initiate the save then SPSS will save the data file (not the output or syntax). You use this dialog box as you would in any other software: type a name in the space next to where it says *File name*. If you have sensitive data, you can password encrypt it by selecting Encrypt file with password. By default, the file will be saved in an SPSS format, which has a *.sav* file extension for data files, *.zsav* for



Oditi's Lantern
Exporting SPSS output

'That I, the almighty Oditi, can discover the secrets within the numbers, they must spread around the world. But non-believers do not have SPSS, so we must send them output in a format that they can access. I give to you, my subservient brethren, a tutorial on how to export SPSS output. These are the tools you need to spread the numbers. Go forth and stare into my lantern.'



compressed data files, *.spv* for classic viewer documents, *.spwb* for workbook documents and *.sps* for syntax files. Once a file has been previously saved, it can be saved again (updated) by clicking on  or pressing Ctrl + S (⌘ + S on Mac OS).

You can save data in formats other than SPSS. Three of the most useful are Microsoft Excel files (*.xls*, *.xlsx*), comma-separated values (*.csv*) and tab-delimited text (*.dat*). The latter two file types are plain text, which means that they can be opened by virtually any data analysis software you can think of (including Excel, OpenOffice, Numbers, R, SAS and Systat). To save your data file in one of these formats (and others), click and select a format from the drop-down list (Figure 4.20).

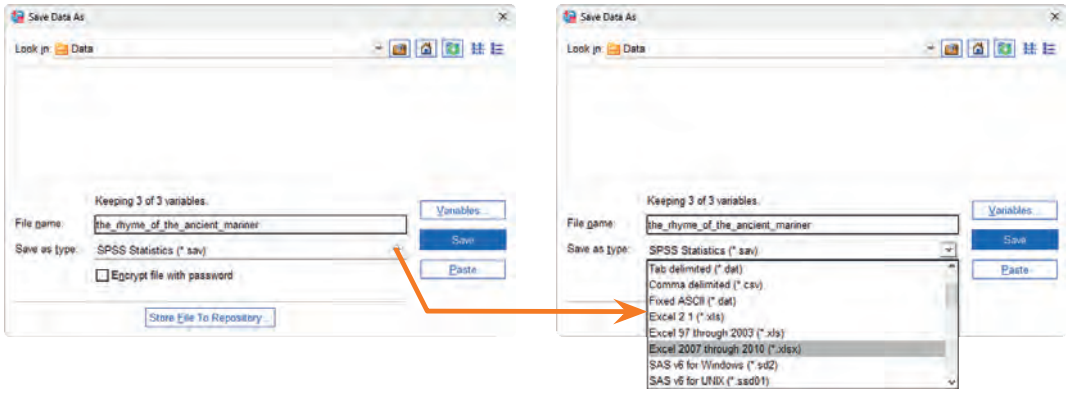


Figure 4.20 The *Save Data As* dialog box

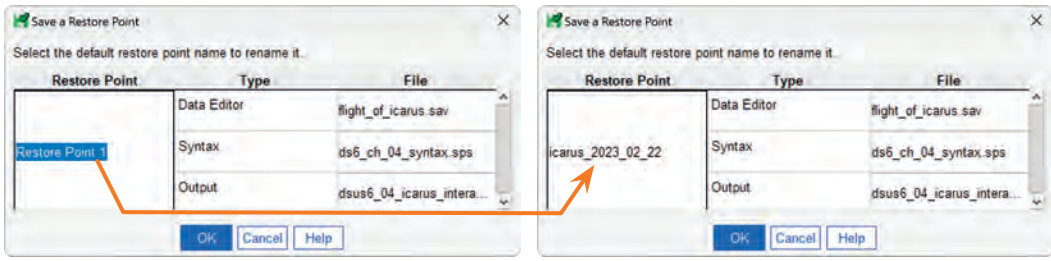


Figure 4.21 Saving a restore point

If SPSS quits unexpectedly, it creates a restore point, which is a snapshot of your session that contains information about the data, syntax and output that was active when SPSS crashed. Rather than waiting for SPSS to crash, you can create manual restore points, which is like saving your entire SPSS workspace at a point in time. To do this choose *File* > Save Restore Point... to activate the dialog box in Figure 4.21. Note that all the currently active files are listed along with the type of file. To change the default name (*Restore Point 1* in Figure 4.21) for the restore point, click on the text to highlight it, then type a more informative name. I recommend a word or two relating to the project and then a date. In Figure 4.21 I have renamed the restore point by the project name (*icarus*) and the year, month and day. Using the date helps you to keep track of different restore points.

4.11 Opening files and restore points Ⓐ

This book relies on your working with data files that you can download from the companion website. You probably don't need me to tell you how to open a file, but just in case … To load a file into SPSS use the icon or select *File* > *Open* > and then Data... to open a data file, Output... to open a classic viewer file, Workbook... to open a workbook file or Syntax... to open a syntax file. This process opens a (probably)

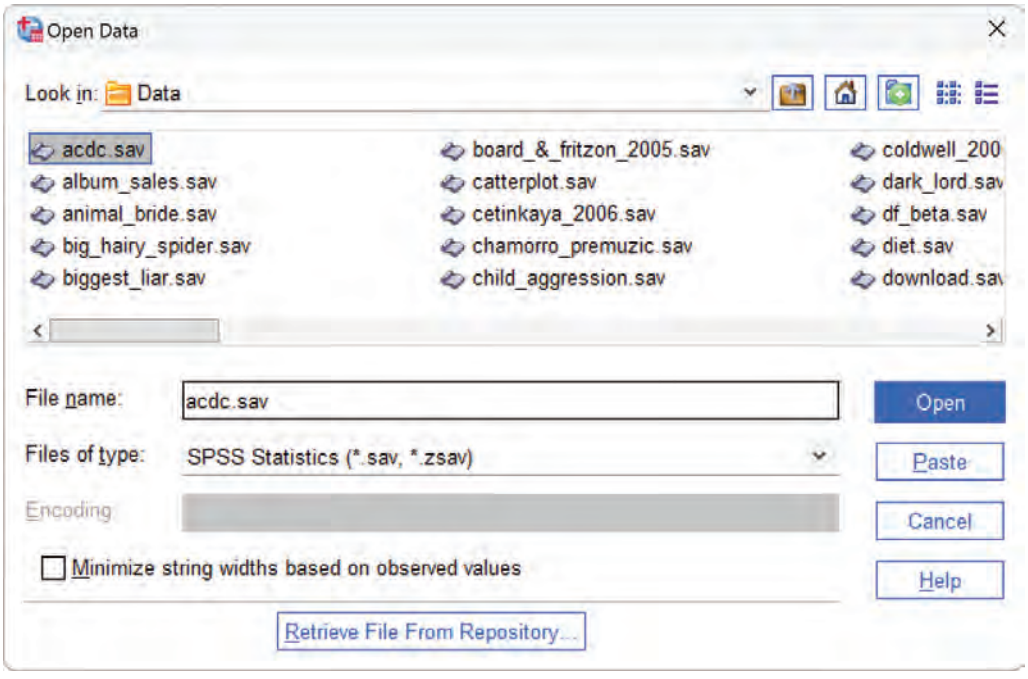


Figure 4.22 Dialog box to open a file

familiar dialog box (Figure 4.22). Navigate to wherever you saved the file that you need. SPSS will list the files of the type you asked to open (so, data files if you selected Data...). Open the file you want by either selecting it and clicking Open, or double-clicking on the icon next to the file you want (e.g., double-clicking on). If you want to open data in a format other than SPSS (*.sav*), then click SPSS Statistics (*.sav) and to display a list of alternative file formats. Click the appropriate file type (Microsoft Excel file (**.xls*, **.xlsx*), text file (**.dat, *.txt* etc.) to list files of that type in the dialog box. To open a restore point select *File* > Open Restore Point... to see a dialog box with restore points listed in chronological order (most recent first). Scroll down the list to the restore point you want and click on OK. You can also delete a restore point by clicking on .

4.12 A few useful options ©

You can set default options for SPSS by selecting *Edit* > Options... to activate the dialog box in Figure 4.23. There are a lot of settings, and most people won't want to change any of them, but we'll run through a few. As we have seen, you can run the viewer in classic mode or workbook mode. By default it will run in classic mode, but if you'd rather use workbooks as your default then select Workbook in the general options. By default, SPSS will display variable labels rather than variable names in dialog boxes (and any lists of variables). It also orders variables as they appear in the data editor. For example, if we open a dialog box to run an analysis on the data about Metallica that we entered earlier in this chapter, the variables will be listed as

- Band member's name
- Date of birth
- Date of death
- Musical instrument played in Metallica
- Current member of the band
- Number of song writing credits
- Net worth ($million)
- Albums played on
- Worth per song ($million)

Note that the labels we gave the variables have been used and the variables are listed in the same order as the data editor. This default is sensible because variable labels are typically more informative than variable names. However, if you have particularly long variable labels, or an unwieldy data set, you might prefer the dialog boxes to list variables by their name, in which case select Display names and they will appear as

- name
- birth_date
- death_date
- instrument
- current_member
- songs_written
- net_worth
- albums
- worth_per_song

You can also choose to list them alphabetically (Alphabetical) rather than as they appear in the data editor, or to group them by their measurement level (Measurement level), in which case all the variables set as nominal would be listed together before the ordinal and then scale variables. Finally (for us), you can

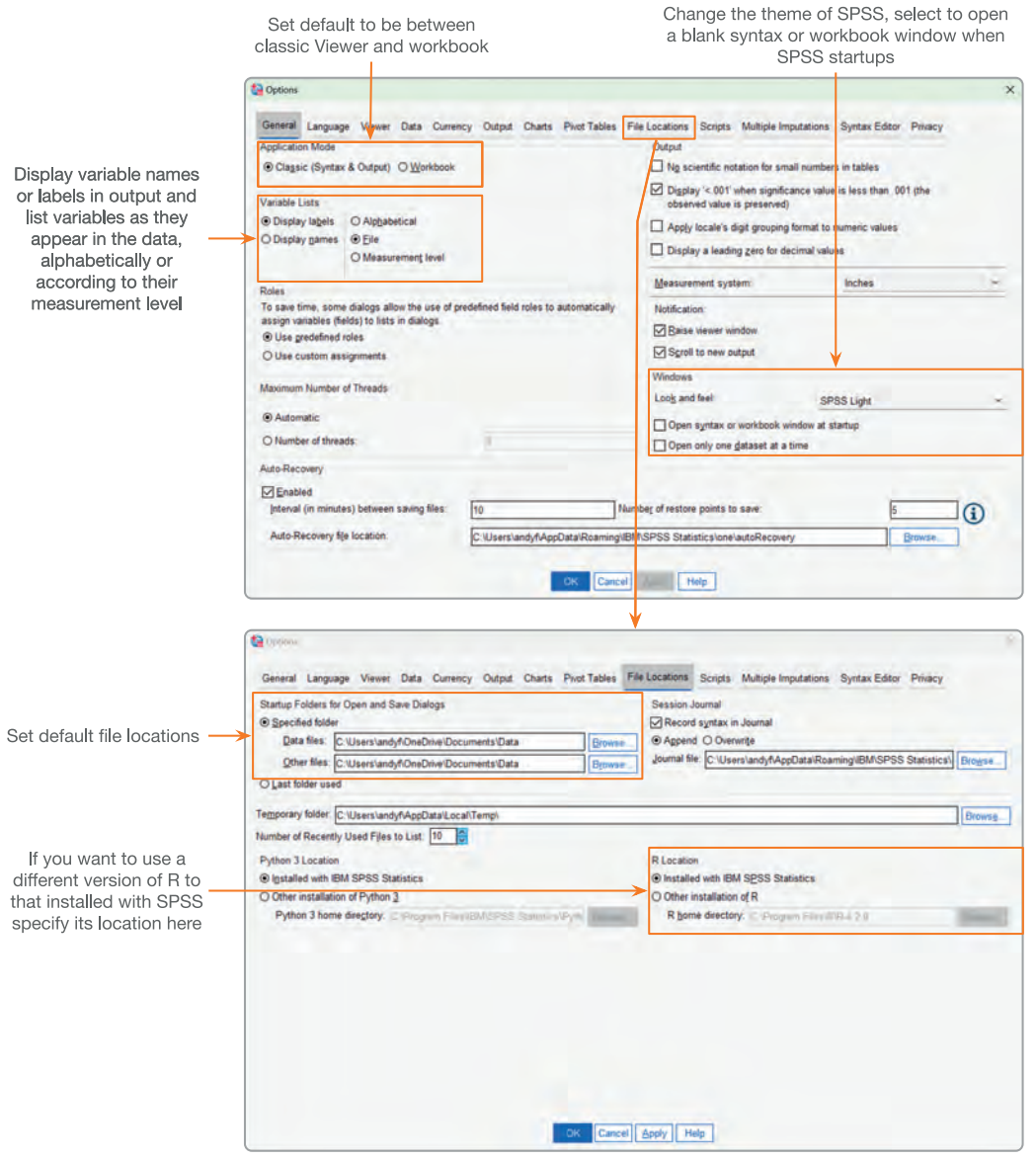


Figure 4.23 The *Options* dialog box

switch the theme of SPSS if you want to try a different look. You can also choose to open a new workbook (or syntax window) when SPSS starts up and to restrict SPSS to only open one data file at a time (by default you can have multiple data editors open simultaneously).

There are lots of tabs in the options which reveal options for other aspects of SPSS. Feel free to explore these.☺ For now, select File Locations. By default, when you go to open a file SPSS looks in the last folder that you used, which makes sense but is often not where the current file you want to use is located. So, you waste time navigating your computer trying to find your data. If you use SPSS as much as I do then this has two consequences: (1) all those seconds have added up to weeks navigating my computer when I could have been doing something useful like playing my drum kit; (2) I have increased my chances of getting RSI in my wrists, and if I'm going to get RSI in my wrists I can think of more enjoyable ways to achieve it than navigating my computer (drumming again, obviously). Luckily, we can avoid wrist death by selecting a default data folder in the options. Imagine that I keep my data files in a single folder called, rather unimaginatively, 'Data'. In the dialog box in Figure 4.23.

I have clicked on Browse... and then navigated to my data folder. SPSS will now use this as the default location when I open files and my wrists are spared the indignity of RSI. You can also set a default folder for other types of files (which can be a different location than the default folder for data). SPSS 29 installs a data analysis app called R for you (we'll talk about this in the next section). By default SPSS uses the version of R that it installed, and for the vast majority of you that will be fine. However, if you're someone who uses R and perhaps have a newer version of it running on your machine than the one SPSS installed, then you can tell SPSS where to find your version of R and it will use that instead when you run R commands through SPSS.

4.13 Extending IBM SPSS Statistics Ⓑ

IBM SPSS Statistics has some powerful tools for users to build their own functionality. For example, you can create your own dialog boxes and menus to run syntax that you may have written. SPSS also interfaces with a powerful open source statistical computing language called *R* (R Core Team, 2016). There are two extensions to SPSS that we use in this book. One is a tool called PROCESS and the other is the *Essentials for R for Statistics* plug-in, which will give us access to R so that we can implement robust models using the *WRS2* package (Mair & Wilcox, 2019).

4.13.1 The PROCESS tool Ⓑ

The PROCESS tool (Hayes, 2022) wraps up a range of functions to do moderation and mediation analyses based on work by Andrew Hayes and Kristopher Preacher (Hayes & Matthes, 2009; Preacher & Hayes, 2004, 2008). We will use this tool in Chapter 11. While using these tools, spare a thought of gratitude to Hayes and Preacher for using their spare time to do cool stuff like this that makes it possible for you to analyse your data without having a nervous breakdown. Even if you still feel think you are having a nervous breakdown, trust me, it's not as bad as the one you'd be having if PROCESS didn't exist. The easiest way to use the PROCESS tool is through what's known as a custom dialog box, which can be installed in four steps:

1. *Download the install file*: Go to https://haskayne.ucalgary.ca/CCRAM/resource-hub and download the file **processvXX.zip** where the XX refers to the current version of PROCESS. At the time of writing the file is **processv43.zip**. It should be saved to your Downloads folder by default.
2. *Extract the zip file*: the file you have downloaded is what's known as a zip file. You need to unzip it. This should be as simple as double-clicking on it. Once unzipped, it turns into a folder. Within this folder are various useful documents, so consider moving the folder out of your downloads folder to somewhere on your computer or the cloud where that makes sense to you.
3. *Start up SPSS*: On Mac OS make sure you have SPSS running and move to the next step. On Windows you need to start SPSS *as an administrator*. So, if SPSS is already open, close it. Then click the *Start* menu (⊞) or type 'SPSS' in the search bar to open a search (Figure 4.25). Once the dialog box is open make sure 'SPSS' is typed in the search bar and you should see the app (IBM SPSS Statistics) listed under *Best match*. Select this app and then on the right-hand side you'll see options associated with the app, one of which is Run as administrator. Click on this option and SPSS will start but with the ability to make changes to your computer. A dialog box will appear that asks you whether you want to let SPSS make changes to your computer and you should reply 'yes'.
4. *Install PROCESS*: Once SPSS has loaded, navigate to the extracted zip folder from step 2 (which will be in your Downloads folder or wherever you moved it to). Within that folder is a folder called something like 'PROCESS vx.y for SPSS' (where *x* and *y* will be numbers referring to the version. At the time of writing the folder is called 'PROCESS v4.3 for SPSS'. Within that folder is another called 'Custom dialog builder file (vx.y)' (again the *x* and *y* will be numbers referring to

the version, at the time of writing the folder is called 'Custom dialog builder file (v4.3)'. Within this folder is a file called **process.spd**. Double-click on this file, which should automatically install PROCESS. You will likely get a dialog box asking you to confirm whether you want to install it, and obviously you do!

Once the PROCESS tool has been added to SPSS it appears in the *Analyze* > *Regression* menu structure. If you can't see it, then the install hasn't worked and you'll need to work through this section again. On Windows, if you get an error message, the most likely explanation is that you haven't opened SPSS as an administrator (see step 3).

At the time of writing, in version 29 of SPSS (and presumably later versions too) you also need to tweak how the output of PROCESS is displayed. This is because the default output font changed in SPSS 29 and this has messed up the formatting of the output from PROCESS. It may be the case that future versions of PROCESS adapt to SPSS's new defaults, but let's assume not. Therefore, within SPSS, you need to go to *Edit* > Options... to open the dialog box in Figure 4.26. Within that dialog box select the Viewer tab and in the area labelled *Text Output*, select Courier New from the drop-down list as shown in the figure. Click OK to make the change.

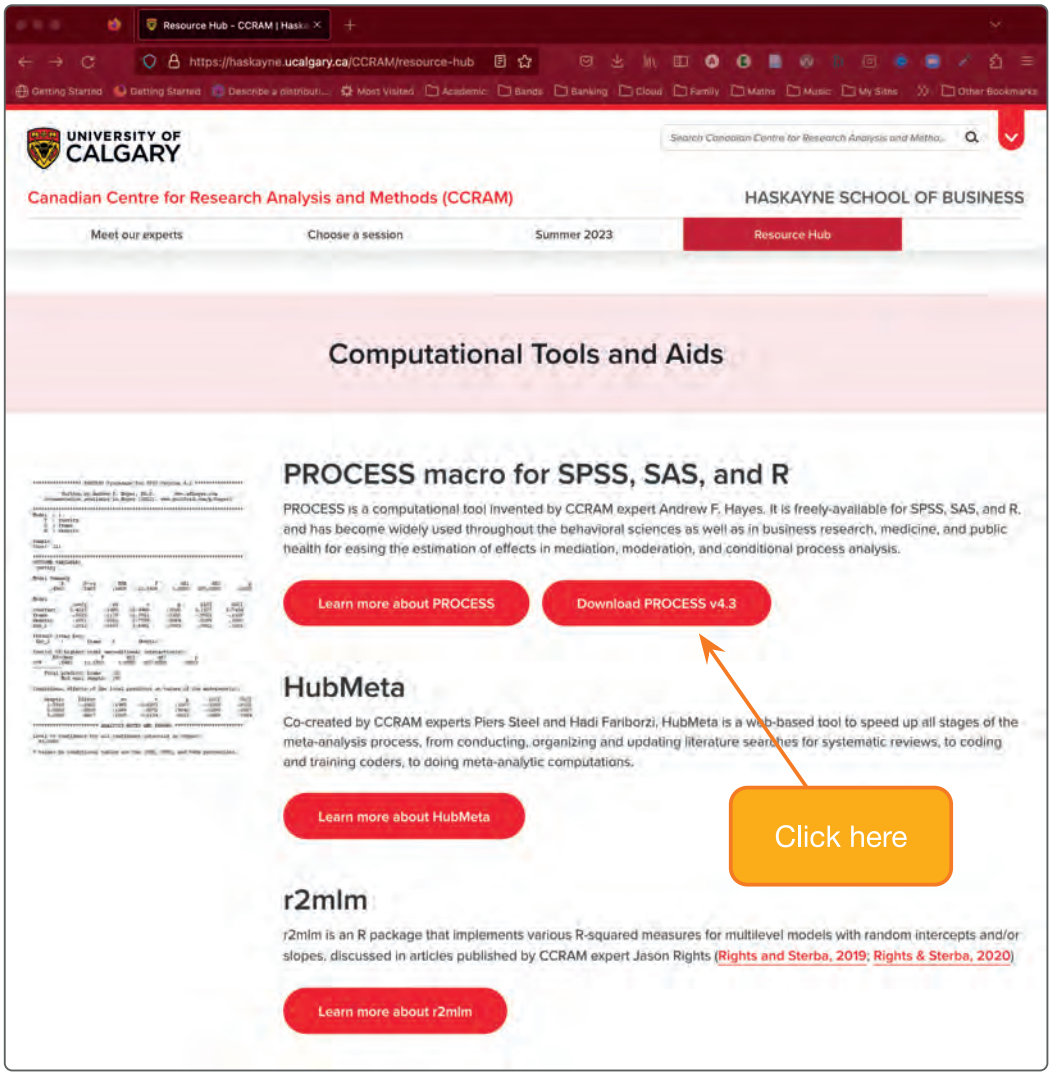


Figure 4.24 Downloading PROCESS

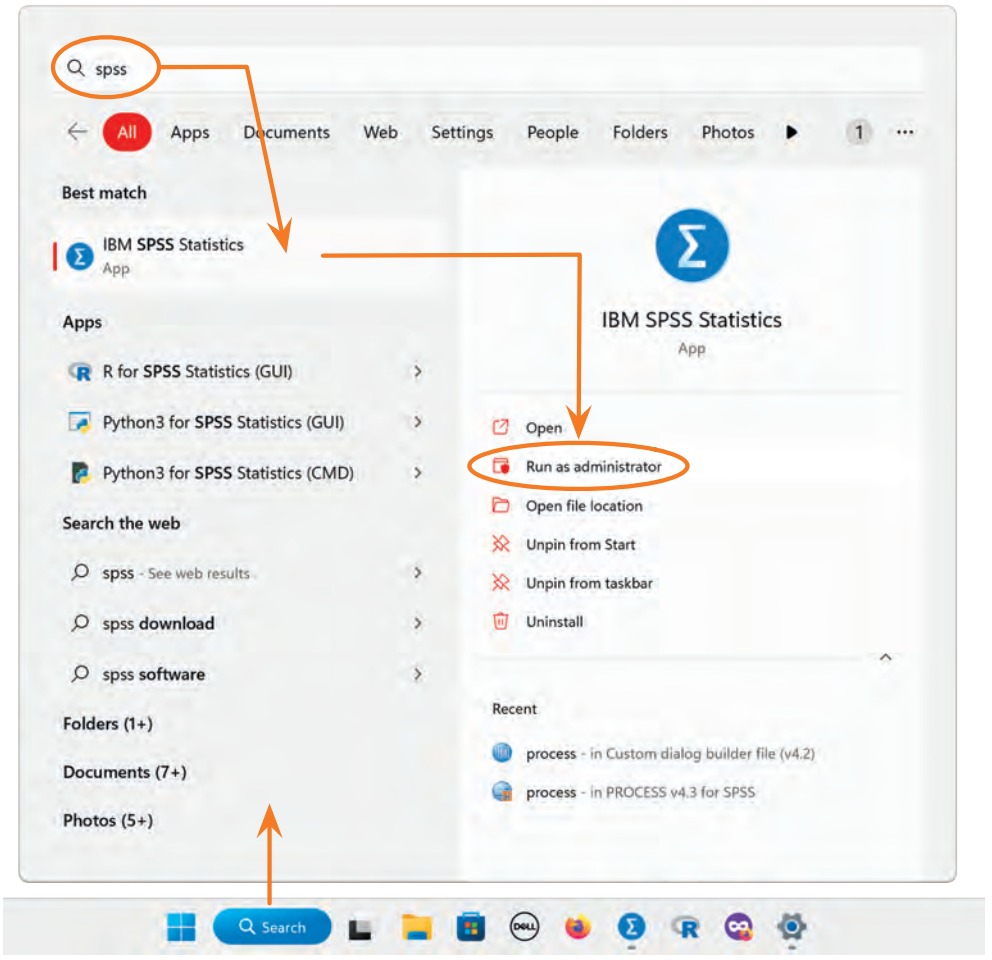


Figure 4.25 Running SPSS as an administrator (Windows)

4.13.2 Integration plug-in for R Ⓑ

At various points in the book we're going to use robust tests which use R. To get SPSS to interface with R we need to have the IBM® SPSS® Statistics – Integration plug-in for R. Prior to version 28 of SPSS this involved an untold world of pain (see the previous edition of this book). Luckily from version 28 this plug-in is automatically installed, as well as a version of R itself (at the time of writing R 4.2). This makes it very easy to work with R from within SPSS. You have two choices for running R commands. You can activate a command-line console by activating the R for SPSS Statistics app. On windows use the Start menu or search 'R for SPSS Statistics' to find it. On Mac OS go to the applications folder, find the IBM SPSS Statistics folder and the app is within that, or press ⌘ + space to open a spotlight search window and search 'R for SPSS Statistics'. Alternatively, we can run R scripts from with the SPSS syntax window by enclosing the R commands within:

```
BEGIN PROGRAM R.
Some R commands
END PROGRAM.
```

The first and last lines (remember the full stops because they matter) tell SPSS to talk to R and then to stop. Any stuff you type in between should be instructions that R understands.

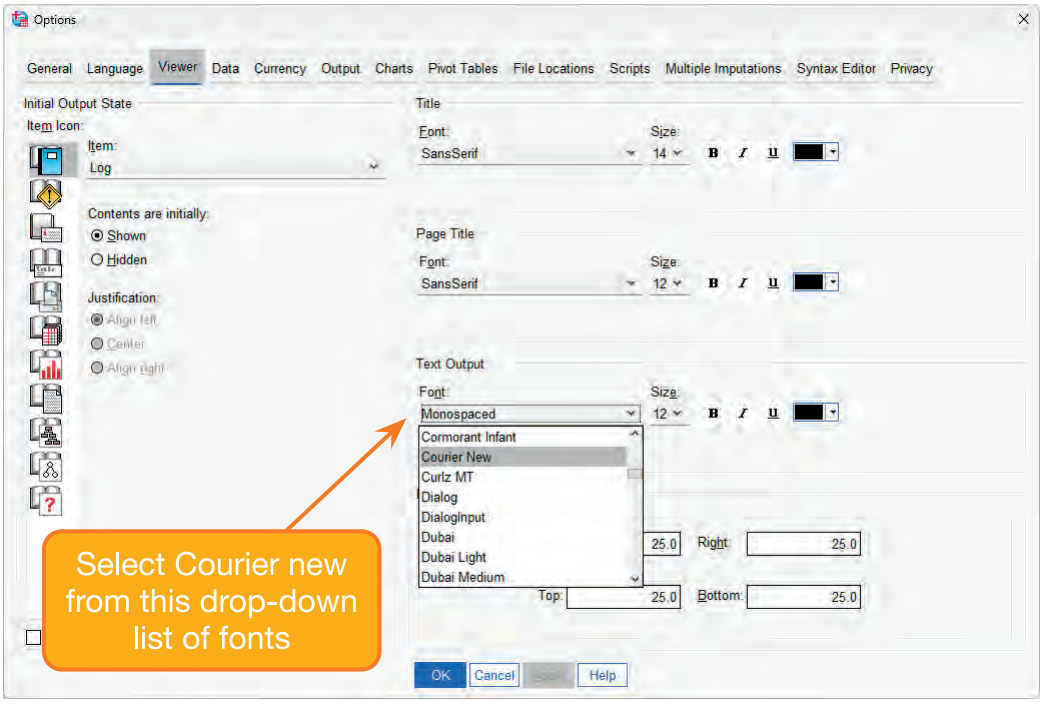


Figure 4.26 Changing the default font for text output

4.13.3 The *WRS2* package Ⓑ

To access the WRS2 package for R (Mair & Wilcox, 2019) open a syntax window and type and run the following syntax:

```
BEGIN PROGRAM R.
install.packages("WRS2", repos = "https://cloud.r-project.org")
END PROGRAM.
```

As mentioned in the previous section, the first and last lines tell SPSS to talk to R and then to stop. It's worth reiterating that the full stops at the end of these lines need to be there! The middle line uses an R function called *install.packages()*, which, as the name suggests, tells R to install a package. Specifically, we tell it to install *WRS2* (which needs to be in quotes) from the URL of the main R repository (this URL also needs to be in quotes).

Throughout the book I supply various workbook files for robust analyses in R and at the top of each one I include this program (for those who skipped this section). However, you only need to execute this program once, not every time you run an analysis. The only times you'd need to re-execute this program would be: (1) if you change computers; (2) if you upgrade SPSS or R itself, for some reason; (3) something goes wrong and you think it might help to reinstall *WRS2*.

4.14 Jane and Brian's story Ⓐ

Brian had been stung by Jane's comment. He was many things, but he didn't think he had his head up his own backside. He retreated from Jane to get on with his single life. He listened to music, met his friends and played *Uncharted 4*. Truthfully, he mainly played *Uncharted 4*. The more he played, the more he thought of

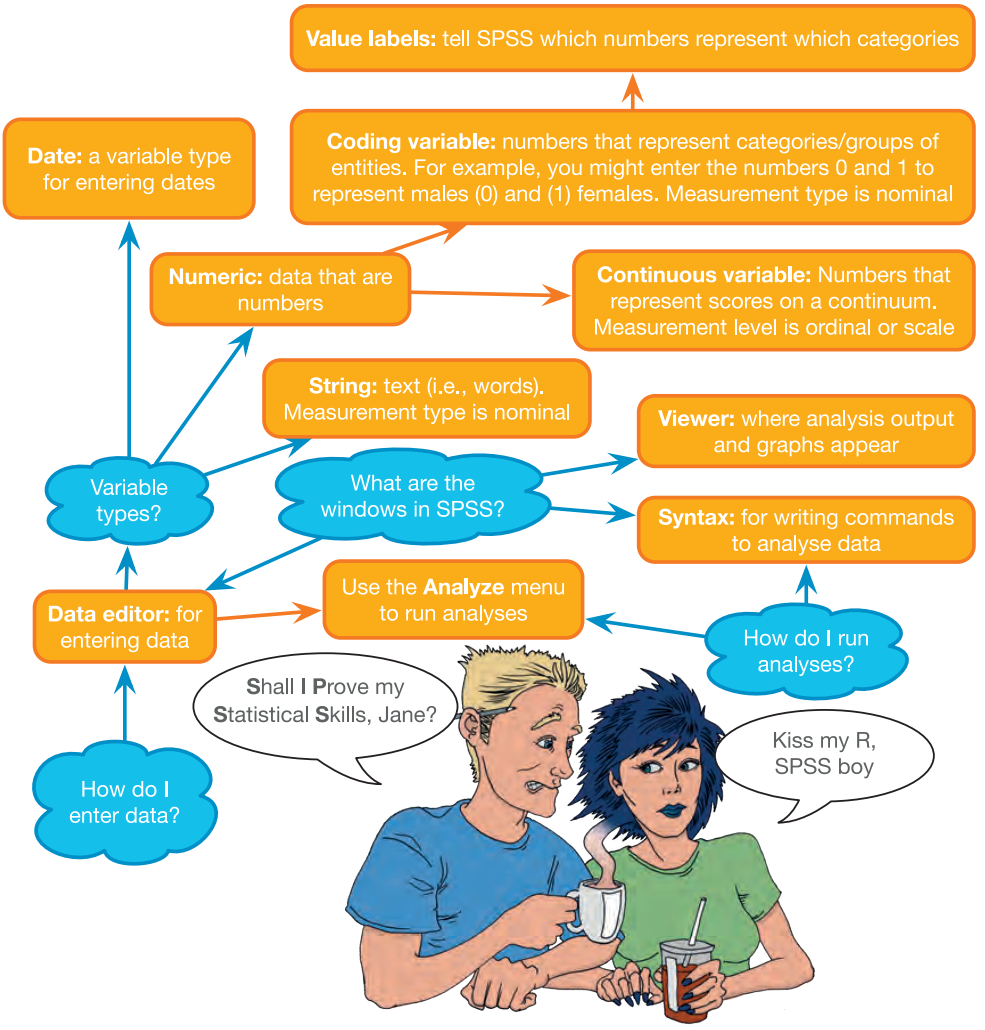


Figure 4.27 What Brian learnt from this chapter

Jane, and the more he thought of Jane, the more convinced he became that she'd be the sort of person who was into video games. When he next saw her he tried to start a conversation about games, but it went nowhere. She said computers were good only for analysing data. The seed was sewn, and Brian went about researching statistics packages. There were a lot of them. Too many. After hours on Google, he decided that one called SPSS looked the easiest to learn. He would learn it, and it would give him something to talk about with Jane. Over the following week he read books, blogs, watched tutorials on YouTube, bugged his lecturers and practised his new skills. He was ready to chew the statistical software fat with Jane.

He searched around campus for her: the library, numerous cafés, the quadrangle – she was nowhere. Finally, he found her in the obvious place: one of the computer rooms at the back of campus called the Euphoria cluster. Jane was studying numbers on the screen, but it didn't look like SPSS. 'What the hell …', Brian thought to himself as he sat next to her and asked …

4.15 What next? Ⓐ

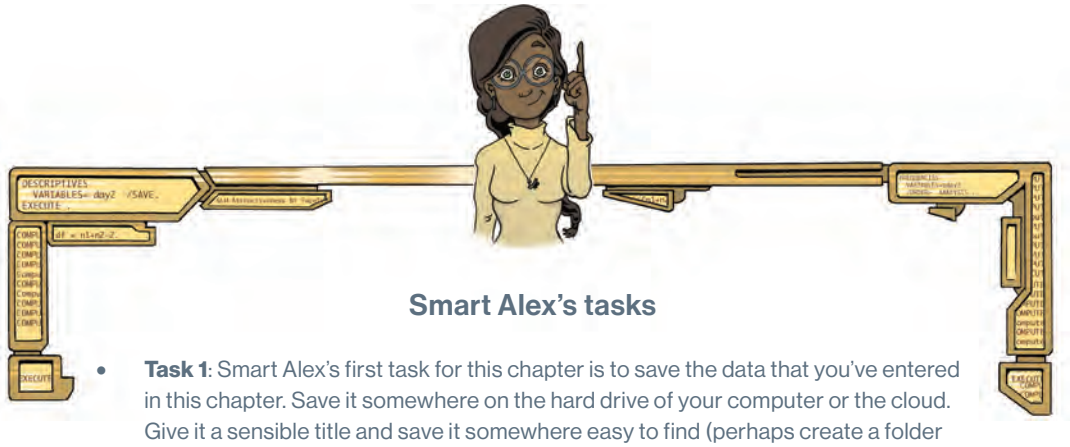
At the start of this chapter we discovered that I yearned for a world full of obedient robots to carry out my wishes. If these robots had existed their first instruction would have been to zap the school's daily

milk delivery with their obliteratronic laser fingers. Let me explain. During the time I was growing up in England, a politician had decided that all school children had to drink a small bottle of milk at the start of the day. The government supplied the milk, I think, for free, but most free things come at some kind of price. The price of free milk turned out to be a lifetime of unpleasant milk-related memories. The milk was usually delivered early in the morning and then left in the hottest place someone could find until we innocent children hopped and skipped into the playground oblivious to the gastric turmoil that awaited. We were greeted with one of these bottles of warm milk and a very small straw. We were then forced to drink it through grimacing faces. The straw was a blessing because it filtered out the lumps formed in the gently curdling milk. Politicians take note: if you want children to enjoy school, don't force feed them warm, lumpy milk.

But despite the warm milk every morning and the lack of robot slaves, primary school was a very happy time for me. With the help of Jonathan Land (Chapter 2) my confidence grew. With this new confidence I began to feel comfortable not just at school but in the world more generally. It was time to explore.

4.16 Key terms that I've discovered

Currency variable
Data editor
Data view
Date variable
Long format data
Numeric variable
String variable
Syntax editor
Variable view
Viewer
Wide format data



Smart Alex's tasks

- **Task 1**: Smart Alex's first task for this chapter is to save the data that you've entered in this chapter. Save it somewhere on the hard drive of your computer or the cloud. Give it a sensible title and save it somewhere easy to find (perhaps create a folder called 'My Data Files' where you can save all of your files when working through this book). Ⓐ
- **Task 2**: What are the following icons shortcuts to? Ⓐ



- **Task 3**: The data below show the score (out of 20) for 20 different students, some of whom are male and some female, and some of whom were taught using positive reinforcement (being nice) and others who were taught using punishment (electric shock). Enter these data into SPSS and save the file as **teach_method.sav**. (Hint: the data should not be entered in the same way that they are laid out below.) Ⓐ

Male		Female	
Electric Shock	Being Nice	Electric Shock	Being Nice
15	10	6	12
14	9	7	10
20	8	5	7
13	8	4	8
13	7	8	13

- **Task 4**: Thinking back to Labcoat Leni's Real Research 4.1, Oxoby also measured the minimum acceptable offer; these MAOs (in dollars) are below (again, they are approximations based on the graphs in the paper). Enter these data into the SPSS data editor and save this file as **acdc.sav**. Ⓐ
  - Bon Scott group: 2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5
  - Brian Johnson group: 0, 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 1
- **Task 5**: According to some highly unscientific research done by a UK department store chain and reported in *Marie Claire* magazine (https://tinyurl.com/mcsgh), shopping is good for you. They found that the average woman spends 150 minutes and walks 2.6 miles when she shops, burning off around 385 calories. In contrast, men spend only about 50 minutes shopping, covering 1.5 miles. This was based on strapping a pedometer on a mere 10 participants. Although I don't have the actual data, some simulated data based on these means are below. Enter these data into SPSS and save them as **shopping.sav**. Ⓐ

Male		Female	
Distance	Time	Distance	Time
0.16	15	1.40	22
0.40	30	1.81	140
1.36	37	1.96	160
1.99	65	3.02	183
3.61	103	4.82	245

- **Task 6**: I wondered whether a fish or cat made a better pet. I found some people who had either fish or cats as pets and measured their life satisfaction and how much they like animals. Enter these data into SPSS and save as **pets.sav**. Ⓐ

Fish		Cat	
Animal Liking	Life Satisfaction	Animal Liking	Life Satisfaction
69	47	16	52
25	6	65	66
31	47	39	65
29	33	35	61
12	13	19	60
49	56	53	68
25	42	27	37
35	51	44	72
51	42		
40	46		
23	27		
37	48		

- **Task 7**: One of my favourite activities, especially when trying to do brain-melting things like writing statistics books, is drinking tea. I am English, after all. Fortunately, tea improves your cognitive function – well, it does in older Chinese people at any rate (Feng et al., 2010). I may not be Chinese and I'm not *that* old, but I nevertheless enjoy the idea that tea might help me think. Here are some data based on Feng et al.'s study that measured the number of cups of tea drunk and cognitive functioning in 15 people. Enter these data into SPSS and save the file as **tea_15.sav**. Ⓐ

Cups of Tea	Cognitive Functioning
2	60
4	47
3	31
4	62
2	44
3	41
5	49
5	56
2	45
5	56
1	57
3	40
3	54
4	34
1	46

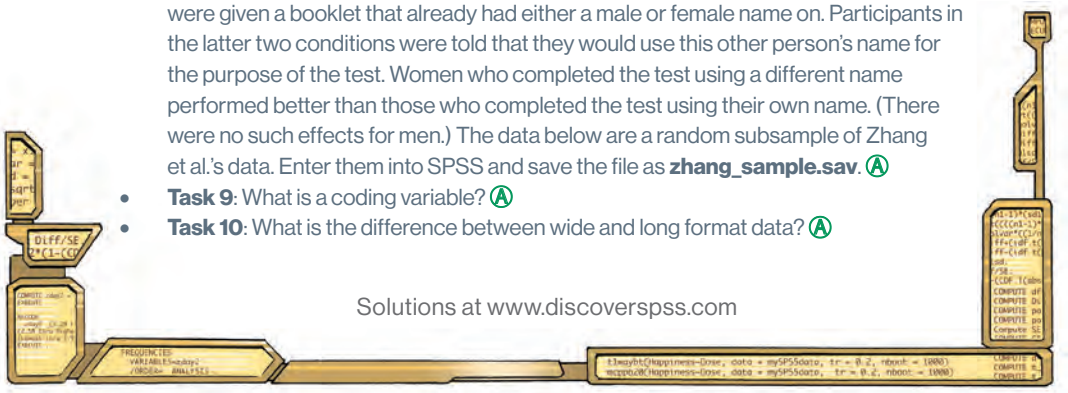
- **Task 8**: Statistics and maths anxiety are common and affect people's performance on maths and stats assignments; women in particular can lack confidence in mathematics (Field, 2010, 2014). Zhang et al. (2013) did an intriguing study in which students completed a maths test and some put their own name on the test booklet, while others

Male			Female		
Female Fake Name	Male Fake Name	Own Name	Female Fake Name	Male Fake Name	Own Name
33	69	75	53	31	70
22	60	33	47	63	57
46	82	83	87	34	33
53	78	42	41	40	83
14	38	10	62	22	86
27	63	44	67	17	65
64	46	27	57	60	64
62	27			47	37
75	61			57	80
50	29				

were given a booklet that already had either a male or female name on. Participants in the latter two conditions were told that they would use this other person's name for the purpose of the test. Women who completed the test using a different name performed better than those who completed the test using their own name. (There were no such effects for men.) The data below are a random subsample of Zhang et al.'s data. Enter them into SPSS and save the file as **zhang_sample.sav**. Ⓐ

- **Task 9**: What is a coding variable? Ⓐ
- **Task 10**: What is the difference between wide and long format data? Ⓐ

Solutions at www.discoverspss.com





VISUALIZING DATA

5.1 What will this chapter tell me? 198
5.2 The art of visualizing data 199
5.3 The SPSS Chart Builder 202
5.4 Histograms 205
5.5 Boxplots (box–whisker diagrams) 211
5.6 Visualizing means: bar charts and error bars 214
5.7 Line charts 226
5.8 Visualizing relationships: the scatterplot 227
5.9 Editing plots 236
5.10 Brian and Jane's story 239
5.11 What next? 240
5.12 Key terms that I've discovered 240
Smart Alex's tasks 240

Letters A to D indicate increasing difficulty

5.1 What will this chapter tell me? Ⓐ

As I got older I found the joy of exploring. Many a happy holiday was spent clambering over the rocky Cornish coastline with my dad. At school they taught us about maps and the importance of knowing where you are going and what you are doing. I had a more relaxed view of exploration and there is a little bit of a lifelong theme of me wandering off to whatever looks most exciting at the time and assuming I know where I am going.[1] I got lost at a holiday camp once when I was about 3 or 4. I remember nothing about it, but my parents tell a tale of frantically running around trying to find me while I was happily entertaining myself (probably by throwing myself out of a tree or something). My older brother, 'the clever one', apparently wasn't 'the observant one' and got a bit of flak for neglecting his duty of watching me. In his defence he was probably mentally deriving equations to bend time and space at the time. He did that a lot when he was 7. The careless explorer in me hasn't really gone away: in new cities I tend to just wander off and hope for the best, usually get lost and so far haven't managed to die (although I tested my luck once by unknowingly wandering through a part of New Orleans where tourists get attacked a lot – it seemed fine to me). To explore data in the way that the 6-year-old me used to explore the world is to spin around 8000 times while drunk and then run along the edge of a cliff. With data you can't be as careless as I am in new cities. To negotiate your way around your data you need a map – maps of data are called plots, and it is into this tranquil and tropical ocean that we now dive (with a compass and ample supply of oxygen, obviously).



Figure 5.1 Explorer Field borrows a bike and gets ready to ride it recklessly around a caravan site

1 It terrifies me that my sons might have inherited this characteristic.

5.2 The art of visualizing data Ⓐ

Wright (2003) adopts Rosenthal's view that researchers should 'make friends with their data'. Although it is true that statisticians need all the friends they can get, Rosenthal didn't mean that: he was urging researchers not to rush data analysis. Wright uses the analogy of a fine wine: you should savour the bouquet and delicate flavours to truly enjoy the experience. He's considerably overstating the joys of data analysis, but rushing your analysis is, I suppose, a bit like downing a bottle of wine: the consequences are messy and incoherent. So, how do we make friends with data? The first thing is to plot the data; for data this is the equivalent of a profile picture. Although it is definitely wrong to judge people based on their appearance, with data the opposite is true.

5.2.1 What is good data visualization? Ⓐ

What makes a good profile picture? Lots of people seem to think that it's best to jazz it up: have some impressive background location, strike a stylish pose, mislead people by inserting some status symbols that you've borrowed, adorn yourself with eye-catching accessories, wear your best clothes to conceal the fact you usually wear a onesie, look like you're having the most fun ever so that people think your life is perfect. This is OK for a person's profile picture, but it is not OK for data: you should avoid impressive backgrounds, eye-catching accessories or symbols that distract the eye, don't fit models that are unrepresentative of reality, and definitely don't mislead anyone into thinking the data matches your predictions perfectly.

Unfortunately all stats software enables you to do all of things that I just told you not to. You may find yourself losing consciousness at the excitement of colouring your plot bright pink (really, it's amazing how excited my students get at the prospect of bright pink plots – personally I'm not a fan). Much as pink plots might send a twinge of delight down your spine, remember why you're plotting the data – it's not to make yourself (or others) purr with delight at the pinkness, it's to present information (dull, but true).

Tufte (2001) points out that plots should do the following, among other things:

- ✓ Show the data.
- ✓ Induce the reader to think about the data being presented (rather than some other aspect of the plot, like how pink it is).
- ✓ Avoid distorting the data.
- ✓ Present many numbers with minimum ink.
- ✓ Make large data sets (assuming you have one) coherent.
- ✓ Encourage the reader to compare different pieces of data.
- ✓ Reveal the underlying message of the data.

Plots often don't do these things (see Wainer, 1984, for some examples), and there is a great example of how not to plot data in the first edition of this book (Field, 2000). Overexcited by IBM SPSS Statistics software's ability to add distracting fluff to plots (like 3-D effects, fill effects and so on – Tufte calls these **chartjunk**) I went into some weird ecstatic state and produced the absolute abomination in Figure 5.2. I literally don't know what I was thinking. Data visualization pioneer, Florence Nightingale, also wouldn't have known what I was thinking.[2] The only positive is that it's not pink! What do you think is wrong with this plot?

2 You may be more familiar with Florence Nightingale as a pioneer of modern nursing, but she also pioneered the visualization of data, not least by inventing the pie chart. Pretty amazing woman.

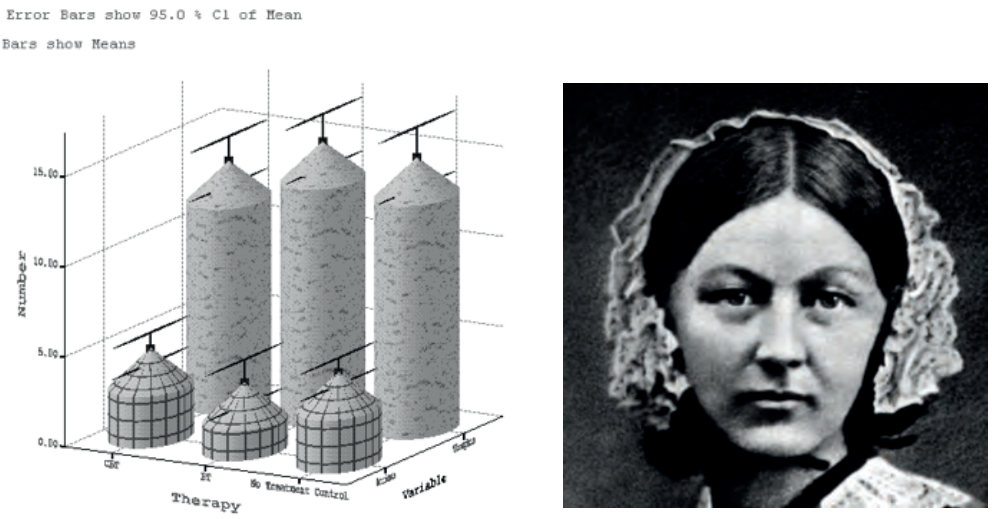


Figure 5.2 A cringingly bad example of a plot from the first edition of this book (left) and Florence Nightingale (right) giving my plot the 'stare of death'

- The bars have a 3-D effect: Never use 3-D plots for a plotting two variables because it obscures the data.[3] In particular, 3-D effects make it hard to see the values of the bars: in Figure 5.2, for example, the 3-D effect makes the error bars almost impossible to read.
- Patterns: The bars also have patterns, which, although very pretty, distract the eye from what matters (namely the data). These are completely unnecessary.
- Cylindrical bars: Were my data so sewage-like that I wanted to put them in silos? The cylinder effect muddies the data and distracts the eye from what is important.
- Badly labelled *y*-axis: 'Number' of what? Delusions? Fish? Cabbage-eating sea lizards from the eighth dimension? Fools who don't know how to plot data?

Now take a look at the alternative version of this plot (Figure 5.3). Can you see what improvements have been made?

- A 2-D plot: The completely unnecessary third dimension is gone, making it much easier to compare the values across therapies and thoughts/behaviours.
- I have superimposed the summary statistics (means and confidence intervals) over the raw data so readers get a full sense of the data (without it being overwhelming)
- The *y*-axis has a more informative label: We now know that it was the number of obsessions per day that was measured. I've also added a legend to inform readers that obsessive thoughts and actions are differentiated by colour.
- Distractions: There are fewer distractions like patterns, cylindrical bars and so on.
- Minimum ink: I've got rid of superfluous ink by getting rid of the axis lines and by using subtle gridlines to make it easy to read values from the *y*-axis. Tufte would be pleased.

3 If you do 3-D plots when you're plotting only two variables then a bearded statistician will come to your house, lock you in a room and make you write I μυστ νοτ δο 3–Δ γραπησ 75,172 times on the blackboard. Seriously.

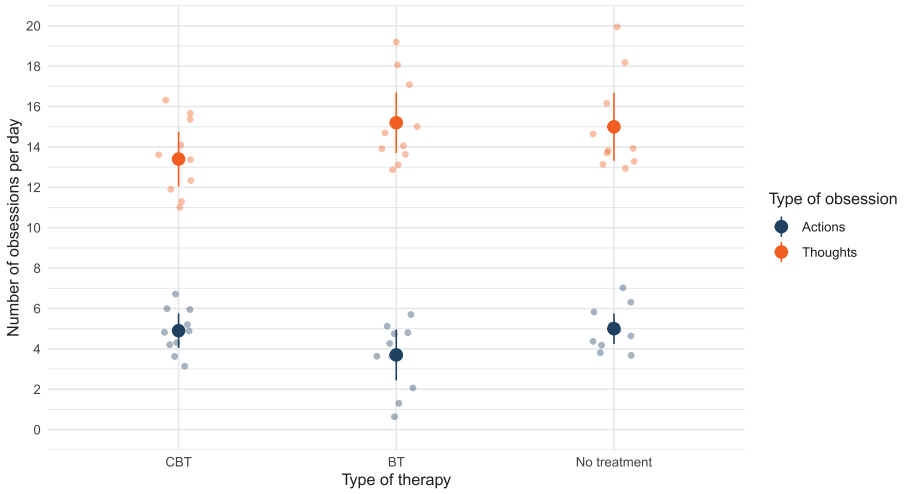


Figure 5.3 Figure 5.2 plotted properly



Cramming Sam's Tips
Plots

- The vertical axis of a plot is known as the *y*-axis (or ordinate).
- The horizontal axis of a plot is known as the *x*-axis (or abscissa).

If you want to draw a good plot follow the cult of Tufte:

- ✓ Don't create false impressions of what the data show (likewise, don't hide effects) by scaling the *y*-axis in some weird way.
- ✓ Avoid chartjunk: Don't use patterns, 3-D effects, shadows, pictures of spleens, photos of your Uncle Fred, pink cats or anything else.
- ✓ Avoid excess ink: a bit radical, but if you don't need the axes, then get rid of them.



5.2.2 Lies, damned lies, and ... erm ... plots Ⓐ

Governments lie with statistics, but scientists shouldn't. How you present your data makes a huge difference to the message conveyed to the audience. As a big fan of cheese, I'm often curious about

whether the urban myth that it gives you nightmares is true. Shee (1964) reported the case of a man who had nightmares about his workmates: 'He dreamt of one, terribly mutilated, hanging from a meat-hook.[4] Another he dreamt of falling into a bottomless abyss. When cheese was withdrawn from his diet the nightmares ceased.' This would not be good news if you were the minister for cheese in your country.

Figure 5.4 shows two plots that, believe it or not, display the same data: the number of nightmares had after eating cheese. The plot on the left shows how the plot should probably be scaled. The *y*-axis reflects the maximum of the scale, and this creates the correct impression: that people have more nightmares about colleagues hanging from meat-hooks if they eat cheese before bed. However, as minister for cheese, you want people to think that cheese isn't so bad, so you rescale the plot (by extending the *y*-axis way beyond the average number of nightmares) and the difference in nightmares appears to have diminished considerably. Tempting as it is, don't do this (unless, of course, you plan to be a politician at some point in your life).

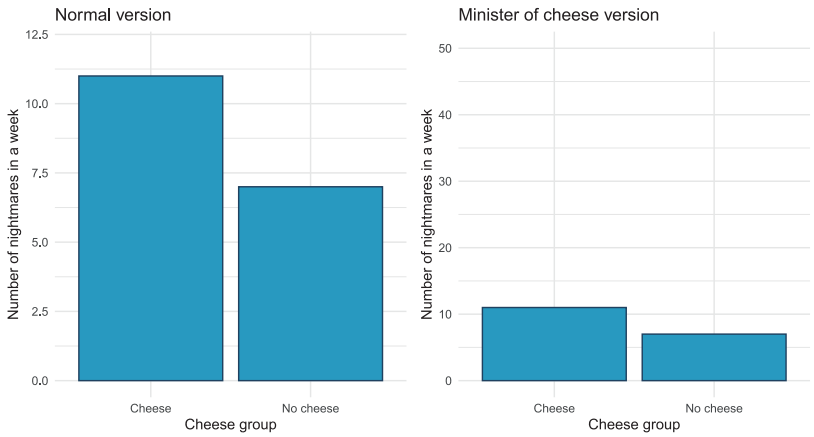
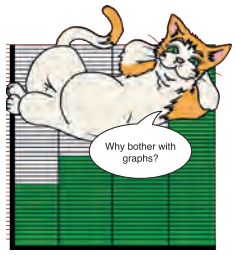


Figure 5.4 Two plots about cheese

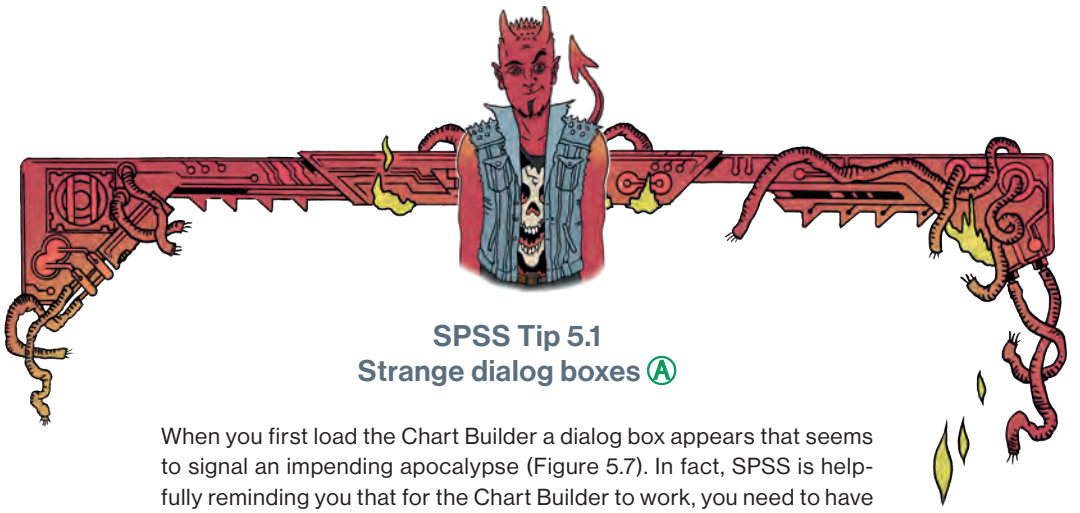
5.3 The SPSS Chart Builder Ⓐ



You are probably drooling like one of Pavlov's dogs to get into the statistics and to discover the answer to your fascinating research question – surely plots are a waste of your precious time? Data analysis is a bit like Internet dating (it's not, but bear with me). You can scan through the vital statistics and find a perfect match (good IQ, tall, physically fit, likes arty French films, etc.) and you'll think you have found the perfect answer to your question. However, if you haven't looked at a picture, then you don't really know how to interpret this information – your perfect match might turn out to be Rimibald the Poisonous, King of the Colorado River Toads, who has genetically combined himself with a human to further his plan to start up a lucrative rodent farm (they like to eat small rodents).[5] Data analysis is much the same: inspect your data with a picture, see how it looks and only then can you interpret the more vital statistics.

4 I have similar dreams but that has more to do with some of my colleagues than cheese.

5 On the plus side, he would have a long sticky tongue and if you smoke his venom (which, incidentally, can kill a dog) you'll hallucinate (if you're lucky, you'll hallucinate that you weren't on a date with a Colorado river toad–human hybrid).



SPSS Tip 5.1
Strange dialog boxes Ⓐ

When you first load the Chart Builder a dialog box appears that seems to signal an impending apocalypse (Figure 5.7). In fact, SPSS is helpfully reminding you that for the Chart Builder to work, you need to have set the level of measurement correctly for each variable. That is, when you defined each variable you must have set them correctly to be scale, ordinal or nominal (see Section 4.6.2). This is because SPSS needs to know whether variables are categorical (nominal) or continuous (scale) when it creates the plots. If you have been diligent and set these properties when you entered the data then click OK to make the dialog disappear. If you forgot to set the level of measurement for any variables then click Define Variable Properties... to go to a new dialog box in which you can change the properties of the variables in the data editor.

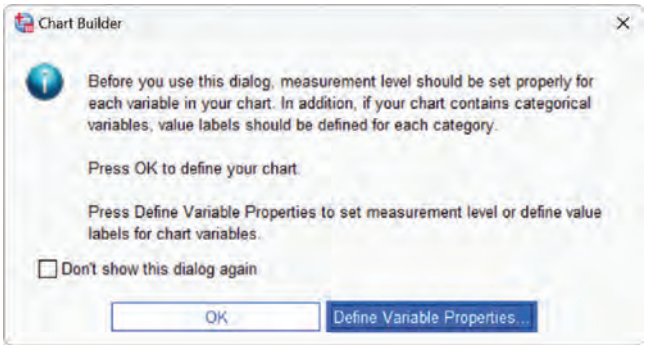


Figure 5.7 Initial dialog box when the Chart Builder is opened

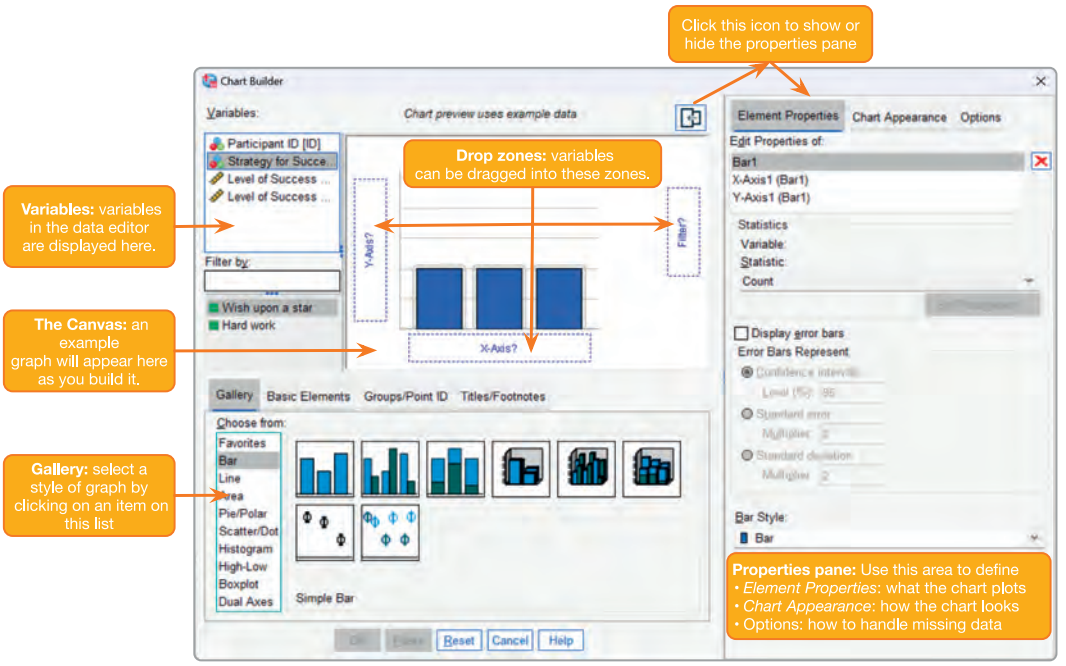


Figure 5.5 The SPSS Chart Builder

Although SPSS's plotting facilities are quite versatile (you can edit most things – see Section 5.9), they are still quite limited for repeated-measures data.[6] To plot data in SPSS we use the all-singing and all-dancing **Chart Builder**.[7]

Figure 5.5 shows the basic *Chart Builder* dialog box, which is accessed through the *Graphs* ❯ Chart Builder... menu. There are some important parts of this dialog box:

- *Gallery*: For each type of plot, a gallery of possible variants is shown. Double-click an icon to select a particular type of plot.
- *Variables*: The variables in the data editor are listed here. These can be dragged into *drop zones* to specify what is displayed on the plot.
- *The canvas*: This is the main area in the dialog box and is where the plot is previewed as you build it.
- *Drop zones*: You can drag variables from the variable list into zones designated with blue dotted lines, called drop zones.
- *Properties panel*: the right-hand panel is where you determine what the plot displays, its appearance, and how to handle missing values.

There are two ways to build a plot: the first is by using the gallery of predefined plots and the second is by building a plot on an element-by-element basis. The gallery is the default option and this tab (Gallery Basic Elements Groups/Point ID Titles/Footnotes) is automatically selected; however, if you want to build your plot from basic elements then click the *Basic Elements* tab (Gallery Basic Elements Groups/Point ID Titles/Footnotes) to change the bottom of the dialog box in Figure 5.5 to look like Figure 5.6.

6 For this reason, some of the plots in this book were created using a package called *ggplot2* for the software R, in case you're wondering why you can't replicate them in SPSS.

7 Unfortunately it's dancing like an academic at a conference disco and singing 'I will always love you' in the wrong key after 34 pints of beer.

We will have a look at building various plots throughout this chapter rather than trying to explain everything in this introductory section (see also SPSS Tip 5.1). Most plots that you are likely to need can be obtained using the gallery view, so I will tend to stick with this method.

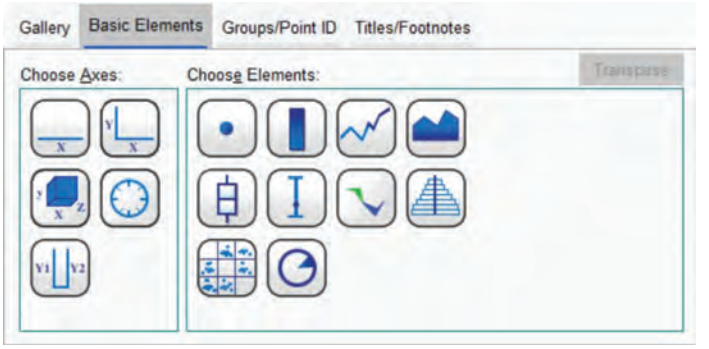


Figure 5.6 Building a plot from basic elements

5.4 Histograms Ⓐ

We encountered histograms (frequency distributions) in Chapter 1; they're a useful way to look at the shape of your data and spot problems (more on that in the next chapter). We will now learn how to create one in SPSS. My wife and I spent our honeymoon at Disney in Orlando. It was two of the best weeks of my life. Although some people find the Disney experience a bit nauseating, in my view there is absolutely nothing wrong with spending time around people who are constantly nice and congratulate you on your marriage. The world could do with more 'nice' in it. The one blip in my tolerance of Disney was the obsession with dreams coming true and wishing upon a star. Don't misunderstand me, I love the idea of having dreams (I haven't yet given up on the idea that one day Steve Harris from Iron Maiden might call requiring my drumming services for their next world tour, nor have I stopped thinking, despite all the physical evidence to the contrary, that I could step in and help my favourite soccer team at their time of need). Dreams are good, but a completely blinkered view that they'll come true without any work on your part is not. My chances of playing drums for Iron Maiden will be greatly enhanced by my practising, forging some kind of name for myself as a professional drummer and incapacitating their current drummer (sorry, Nicko). I think it highly unlikely that merely 'wishing upon a star' will make my dream come true. I wonder if the seismic increase in youth internalizing disorders (Twenge, 2000) is, in part, caused by millions of Disney children reaching the rather depressing realization that 'wishing upon a star' didn't work.

Sorry, I started that paragraph in the happy glow of honeymoon memories but somewhere in the middle it took a downward turn. Anyway, I collected some data from 250 people on their level of success using a composite measure involving their salary, quality of life and how closely their life matches their aspirations. This gave me a score from 0 (complete failure) to 100 (complete success). I then implemented an intervention: I told people that for the next 5 years they should either wish upon a star for their dreams to come true or work as hard as they could to make their dreams come true. I measured their success again 5 years later. People were randomly allocated to these two instructions. The data are in **jiminy_cricket.sav**. The variables are **strategy** (hard work or wishing upon a star), **success_pre** (their baseline level of success) and **success_post** (their level of success 5 years later).

SELF TEST

What does a histogram show?

First, access the Chart Builder as in Figure 5.5 and select *Histogram* in the list labelled *Choose from* to bring up the gallery shown in Figure 5.8. This gallery has four icons representing different types of histogram. Select the appropriate one either by double-clicking on it, or by dragging it onto the canvas:

- ⇒ *Simple histogram*: Use this option to visualize frequencies of scores for a single variable.
- ⇒ *Stacked histogram*: If you have a grouping variable (e.g., whether people worked hard or wished upon a star) you can produce a histogram in which each bar is split by group. In this example, each bar would have two colours, one representing people who worked hard and the other those who wished upon a star. This option is a good way to compare the relative frequency of scores across groups (e.g., were those who worked hard more successful than those who wished upon a star?).
- ⇒ *Frequency polygon*: This option displays the same data as the simple histogram, except that it uses a line instead of bars to show the frequency, and the area below the line is shaded.
- ⇒ *Population pyramid*: Like a stacked histogram, this shows the relative frequency of scores in two populations. It plots the variable (e.g., success after 5 years) on the vertical axis and the frequencies for each population on the horizontal: the populations appear back to back on the plot. This option is useful for comparing distributions across groups.

Let's do a simple histogram. Double-click the icon for a simple histogram (Figure 5.8). The *Chart Builder* dialog box will show a preview of the plot on the canvas. At the moment it's not very exciting (top of Figure 5.9) because we haven't told SPSS which variables to plot. The variables in the data editor are listed on the left-hand side of the Chart Builder, and any of these variables can be dragged into any of the spaces surrounded by blue dotted lines (the *drop zones*).

A histogram plots a single variable (*x*-axis) against the frequency of scores (*y*-axis), so all we need to do is select a variable from the list and drag it into X-Axis?. Let's do this for the post-intervention success scores. Drag **Level of success (post)** from the variable list to X-Axis? as shown in Figure 5.9; you will now find the histogram previewed on the canvas. (It's not a preview of the actual data – it displays the general form of the plot, not what your specific data will look like.) You can edit what the histogram displays using the properties panel (SPSS Tip 5.2). To plot the histogram click OK.

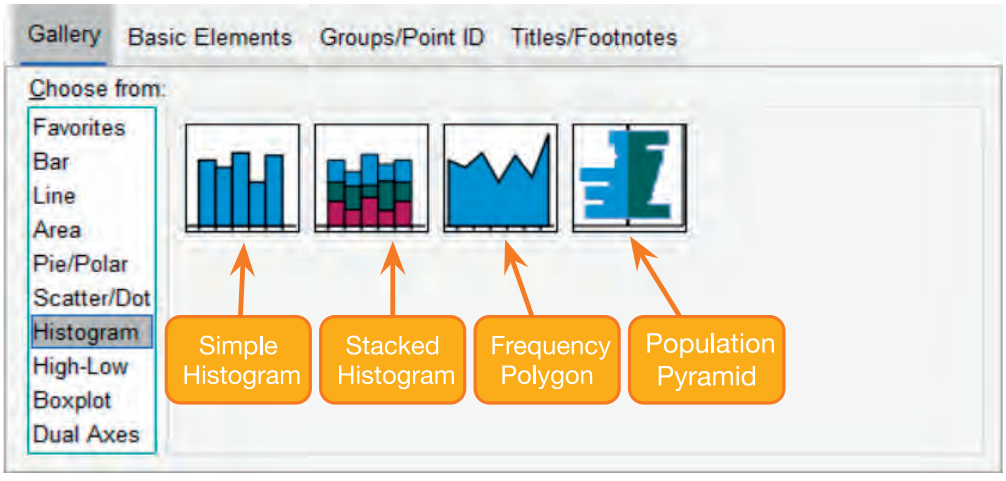


Figure 5.8 The histogram gallery

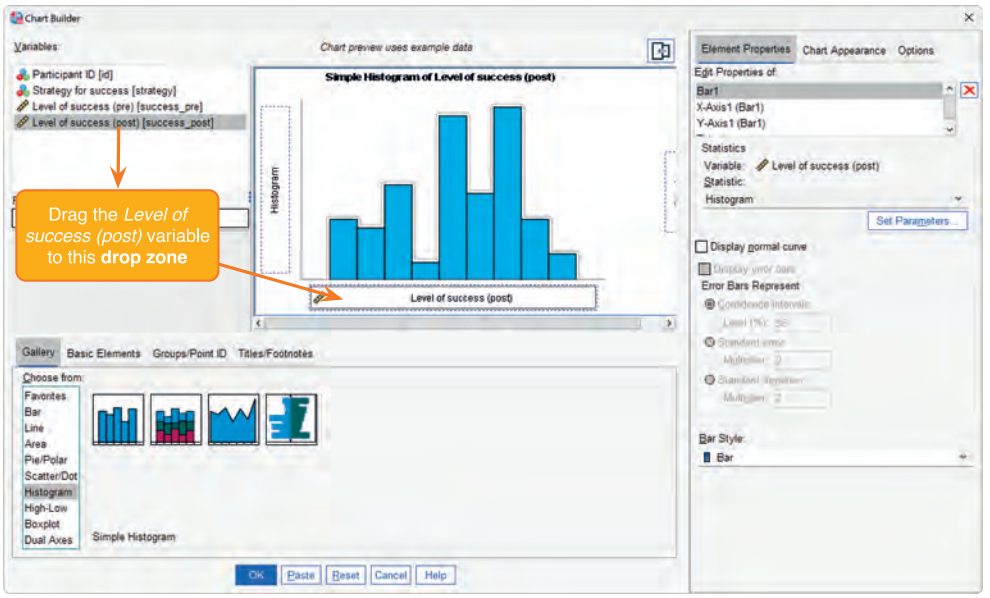
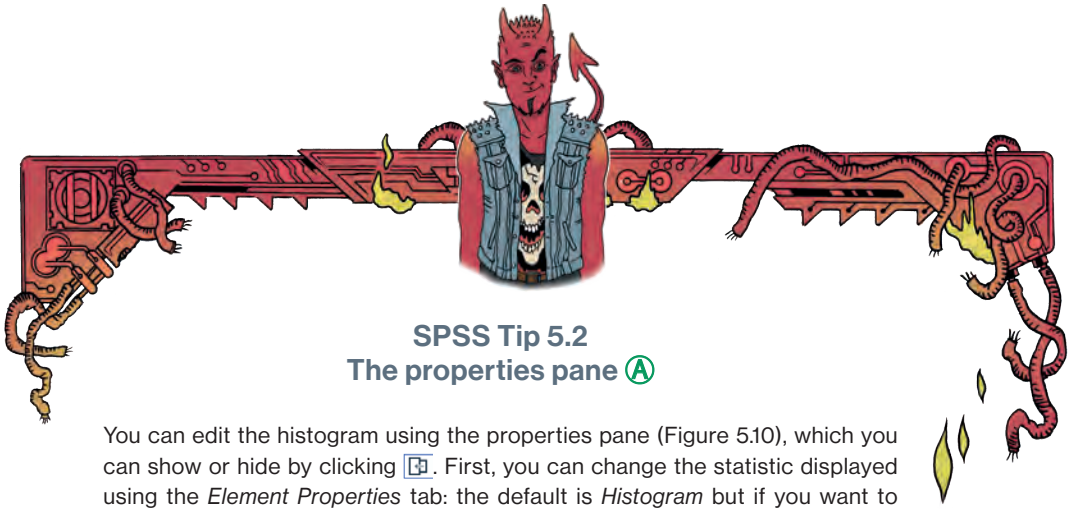


Figure 5.9 Defining a histogram in the Chart Builder



SPSS Tip 5.2
The properties pane Ⓐ

You can edit the histogram using the properties pane (Figure 5.10), which you can show or hide by clicking . First, you can change the statistic displayed using the *Element Properties* tab: the default is *Histogram* but if you want to express values as a percentage rather than a frequency, select *Histogram Percent*. You can also decide manually how you want to divide up your data to compute frequencies by clicking Set Parameters... . In the resulting dialog box you can determine properties of the 'bins' used to make the histogram. Think of a bin as, well, a rubbish bin (this is a pleasing analogy as you will see): on each rubbish bin you write a score (e.g., 3), or a range of scores (e.g., 1–3), then you go through each score in your data set and throw it into the rubbish bin with the appropriate label on it (so a score of 2 gets thrown into the bin labelled 1–3). When you have finished throwing your data into these rubbish bins, you count how many scores are in each bin. A histogram is created in much the same way; either SPSS decides how the bins are labelled (the default), or you decide. Our success scores range from 0 to 100, therefore we might decide that our bins should begin with 0 and we could set the ◉ Custom value for anchor property to 0. We might also decide that we want each bin to contain scores between whole numbers

(i.e., 0–1, 1–2, 2–3, etc.), in which case we could set the ◉ Interval width to be 1. This what I've done in Figure 5.9, but for the time being leave the default settings (i.e., everything set to ◉ Automatic).

In the *Chart Appearance* tab we can change the default colour scheme (the bars will be coloured blue, but we can change this by selecting *Category 1* and choosing a different colour. We can also choose to have an inner or outer frame (I wouldn't bother) and switch off the grid lines (I'd leave them). It's also possible to apply templates so you can create colour schemes, save them as templates and apply them to subsequent charts.

Finally, the *Options* tab enables us to determine how we deal with user-defined missing values (the defaults are fine for our purposes) and whether to wrap panels on charts where we have panels representing different categories (this is useful when you have a variable containing a lot of categories).

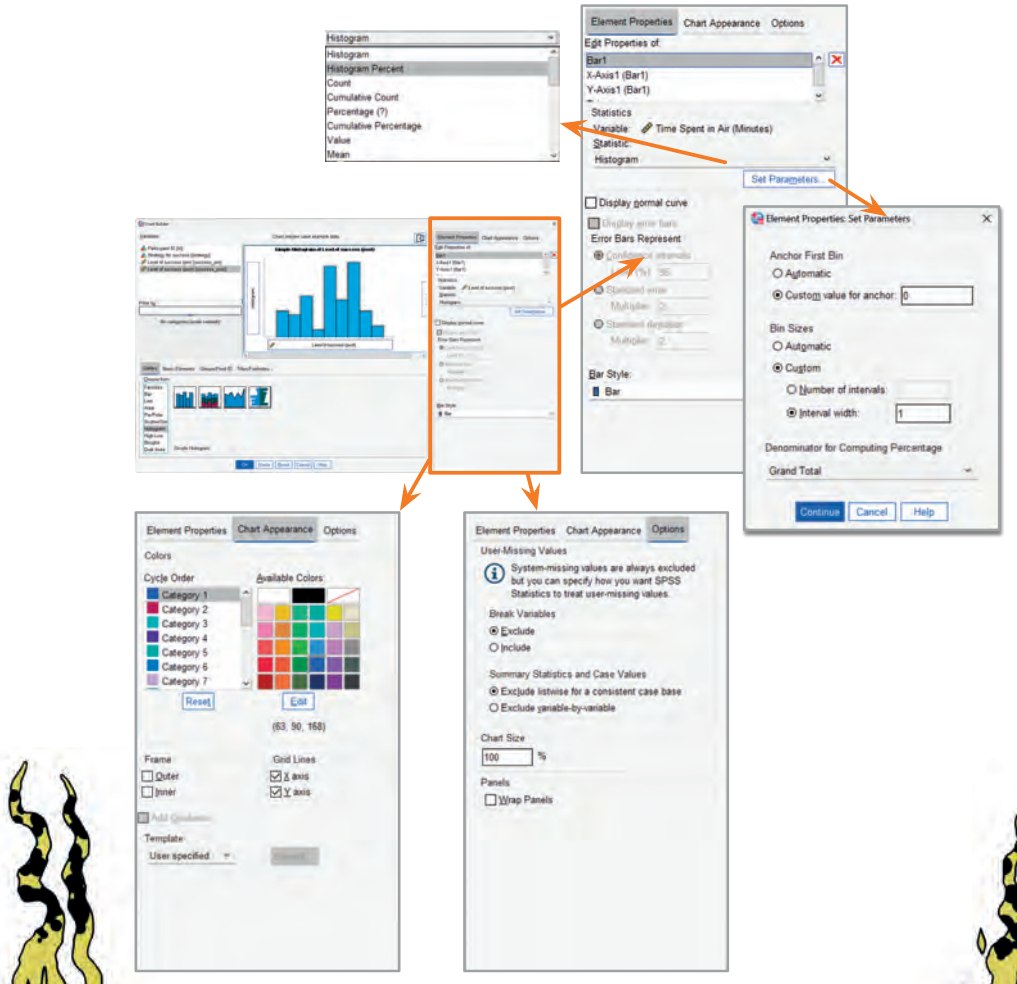


Figure 5.10 The properties pane

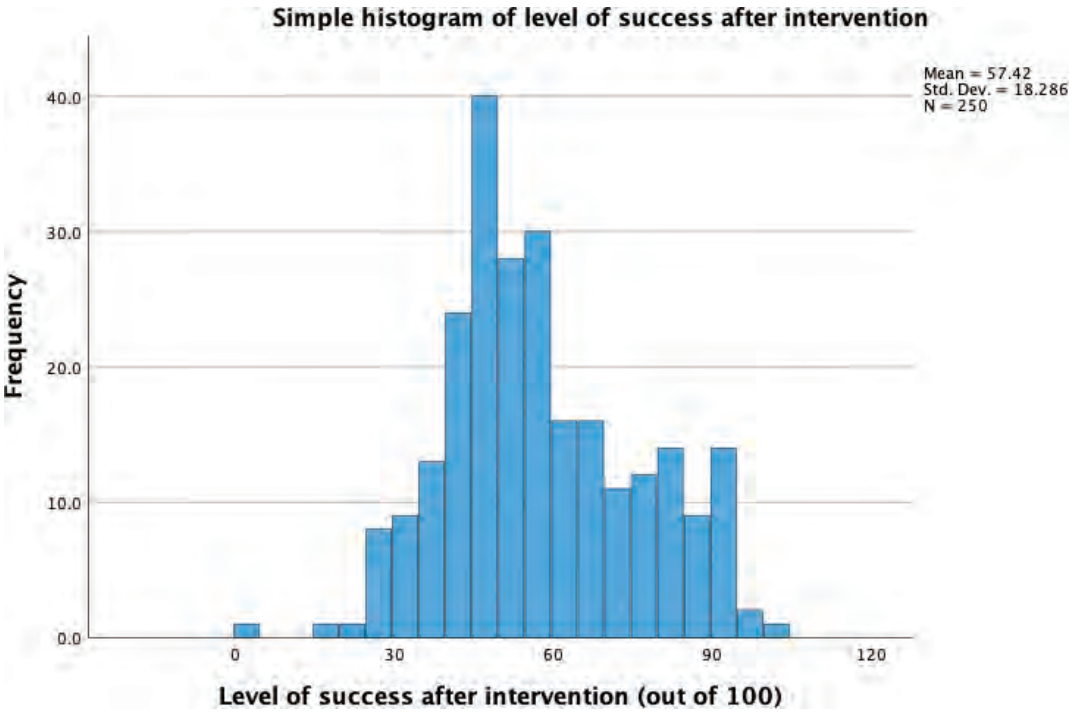


Figure 5.11 Histogram of the post-intervention success scores



Figure 5.12 Defining a population pyramid in the Chart Builder

The resulting histogram is shown in Figure 5.11. The distribution is quite lumpy: although there is a peak of scores around 50 (the midpoint of the scale), there are quite a few scores at the high end, and fewer at the low end. This creates the impression of negative skew, but it's not quite as simple as that. To help us to dig a bit deeper it might be helpful to plot the histogram separately for those who wished upon a star and those who worked hard: after all, if the intervention was a success then their distributions should be from different populations.

To compare frequency distributions of several groups simultaneously we can use a population pyramid. Click the population pyramid icon (Figure 5.8) to display the template for this plot on the canvas. Then from the variable list select the variable representing the success scores after the intervention and drag it into Distribution Variable? to set it as the variable that you want to plot. Then drag **Strategy for success** to Split Variable? to set it as the variable for which you want to plot different distributions. The dialog should now look like Figure 5.12 (note I have hidden the properties pane) – the variable names are displayed in the drop zones, and the canvas now displays a preview of our plot (e.g., there are two histograms representing each strategy for success). Click OK to produce the plot.

The resulting population pyramid is show in Figure 5.13. It shows that for those who wished upon a star there is a fairly normal distribution centred at about the midpoint of the success scale (50%). A small minority manage to become successful just by wishing, but most just end up sort of averagely successful. Those who work hard show a skewed distribution, where a large proportion of people (relative to those wishing) become very successful, and fewer people are around or below the midpoint of the success scale. Hopefully, this example shows how a population pyramid can be a very good way to visualize differences in distributions in different groups (or populations).

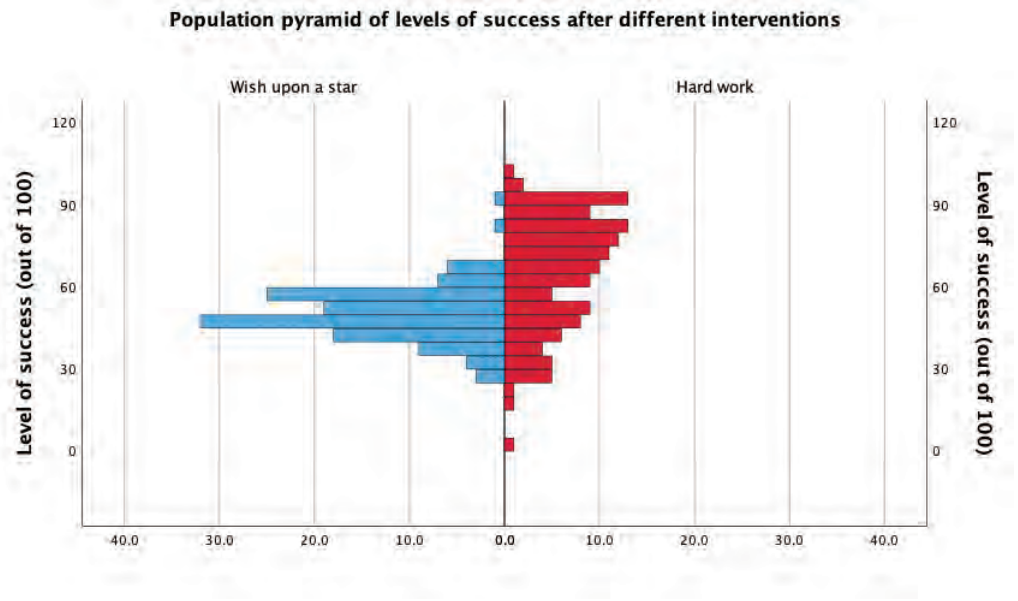


Figure 5.13 Population pyramid of success scores (5 years after different strategies were implemented)



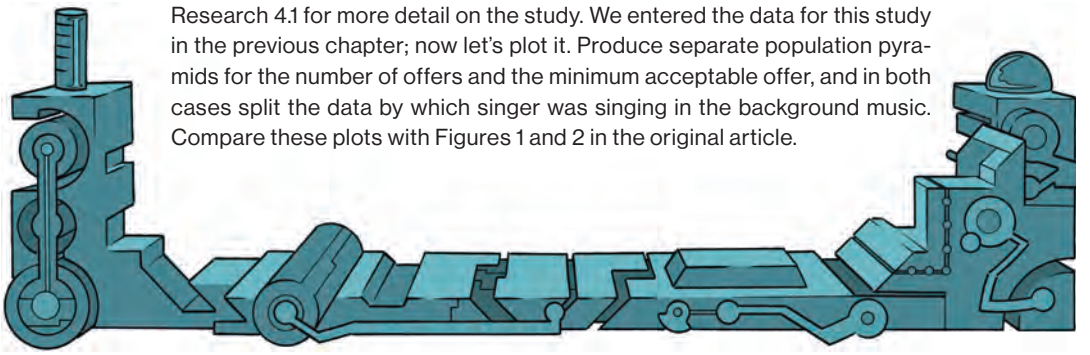
Produce a histogram and population pyramid for the success scores *before* the intervention.



Labcoat Leni's Real Research 5.1
Gonna be a rock 'n' roll singer (again) Ⓐ

Oxoby, R. J. (2008). *Economic Enquiry, 47*(3), 598–602.

In Labcoat Leni's Real Research 4.1 we came across a study that compared economic behaviour while different music by AC/DC played in the background. Specifically, Oxoby (2008) manipulated whether the background song was sung by AC/DC's original singer (Bon Scott) or his replacement (Brian Johnson). He measured how many offers participants accepted and the minimum offer they would accept (**acdc.sav**). See Labcoat Leni's Real Research 4.1 for more detail on the study. We entered the data for this study in the previous chapter; now let's plot it. Produce separate population pyramids for the number of offers and the minimum acceptable offer, and in both cases split the data by which singer was singing in the background music. Compare these plots with Figures 1 and 2 in the original article.



5.5 Boxplots (box–whisker diagrams) Ⓐ

A **boxplot** or box–whisker diagram is one of the best ways to display your data. At the centre of the plot is the *median*. This statistic is surrounded by a box the top and bottom of which are the limits within which the middle 50% of observations fall (the interquartile range, IQR). Sticking out of the top and bottom of the box are two whiskers which show the top and bottom 25% of scores (approximately). First, we will plot some boxplots using the Chart Builder, and then we'll look at what they tell us in more detail.



In the Chart Builder (Figure 5.5) select *Boxplot* in the list labelled *Choose from* to bring up the gallery in Figure 5.14. Boxplots display a summary for a single outcome variable; for example, we might choose level of success after 5 years for our 'wishing upon a star' example. There are three types of boxplot you can produce:

⇒ *1-D boxplot*: This option produces a single boxplot of all scores for the chosen outcome (e.g., level of success after 5 years).

⇒ *Simple boxplot*: This option produces multiple boxplots for the chosen outcome by splitting the data by a categorical variable. For example, we had two groups: wishers and workers. It would

be useful to use this option to display different boxplots (on the same plot) for these groups (unlike the 1-D boxplot, which lumps the data from these groups together).

⇒ *Clustered boxplot*: This option is the same as the simple boxplot except that it splits the data by a second categorical variable. Boxplots for this second variable are produced in different colours. For example, imagine we had also measured whether our participants *believed* in the power of wishing. We could produce boxplots not just for the wishers and workers, but within these groups have different-coloured boxplots for those who believe in the power of wishing and those who do not.

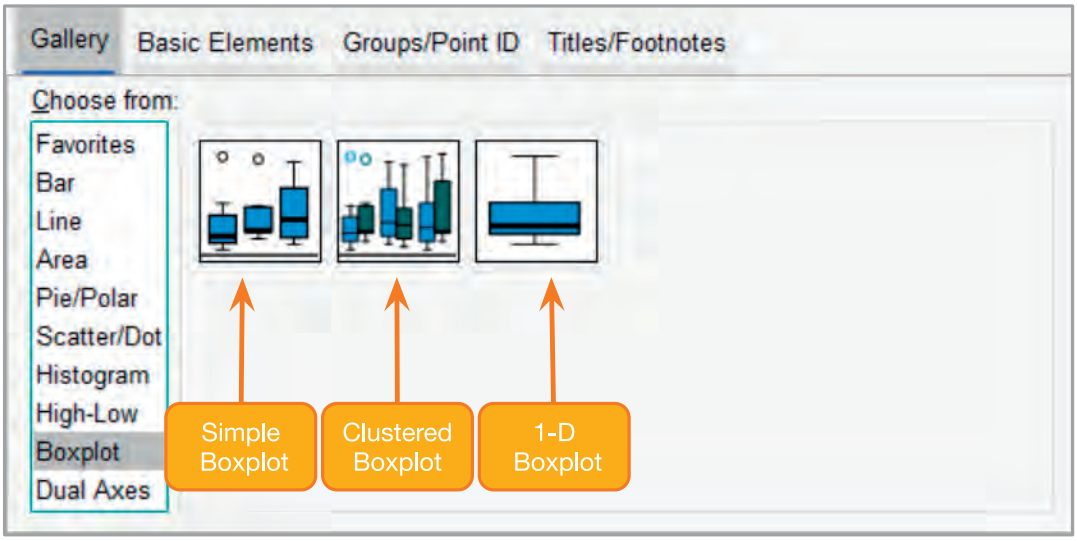


Figure 5.14 The boxplot gallery

In the data file of the success scores we have information about whether people worked hard or wished upon a star. Let's plot this information. To make a boxplot of the post-intervention success scores for our two groups, double-click the *simple boxplot* icon (Figure 5.14), then drag **success_post** from the variable list into Y-Axis? and drag **strategy** from the variable list into X-Axis?. The dialog box should now look like Figure 5.15 – note that the variable names are displayed in the drop zones, and the canvas displays a preview of our plot (there are two boxplots: one for wishers and one for hard workers). Click OK to produce the plot.

Figure 5.16 shows the boxplots for the success data. The blue box represents the IQR (i.e., the middle 50% of scores). The box is much longer in the hard-work group than for those who wished upon a star, which means that the middle 50% of scores are more spread out in the hard-work group. Within the box, the thick horizontal line shows the median. The workers had a higher median than the wishers, indicating greater success overall. The top and bottom of the blue box represent the upper and lower quartile, respectively (see Section 1.7.5). The distance between the top of the box and the top of the whisker shows the range of the top 25% of scores (approximately); similarly, the distance between the bottom of the box and the end of the bottom whisker shows the range of the lowest 25% of scores (approximately). I say 'approximately' because SPSS looks for unusual cases before creating the whiskers: any score greater than the upper quartile plus 1.5 times the IQR is deemed to be an 'outlier' (more on those in Chapter 6), and any case greater than the upper quartile plus 3 times the IQR is an 'extreme case'. The same rules are applied to cases below the lower quartile. When there are no unusual cases, the whiskers show the top and bottom 25% of scores exactly, but when there are unusual cases they show the top and bottom 25% of scores only approximately because the unusual cases are excluded. The whiskers also tell us about the range of scores because the top and bottom of the whiskers show the lowest and highest scores *excluding unusual cases*.

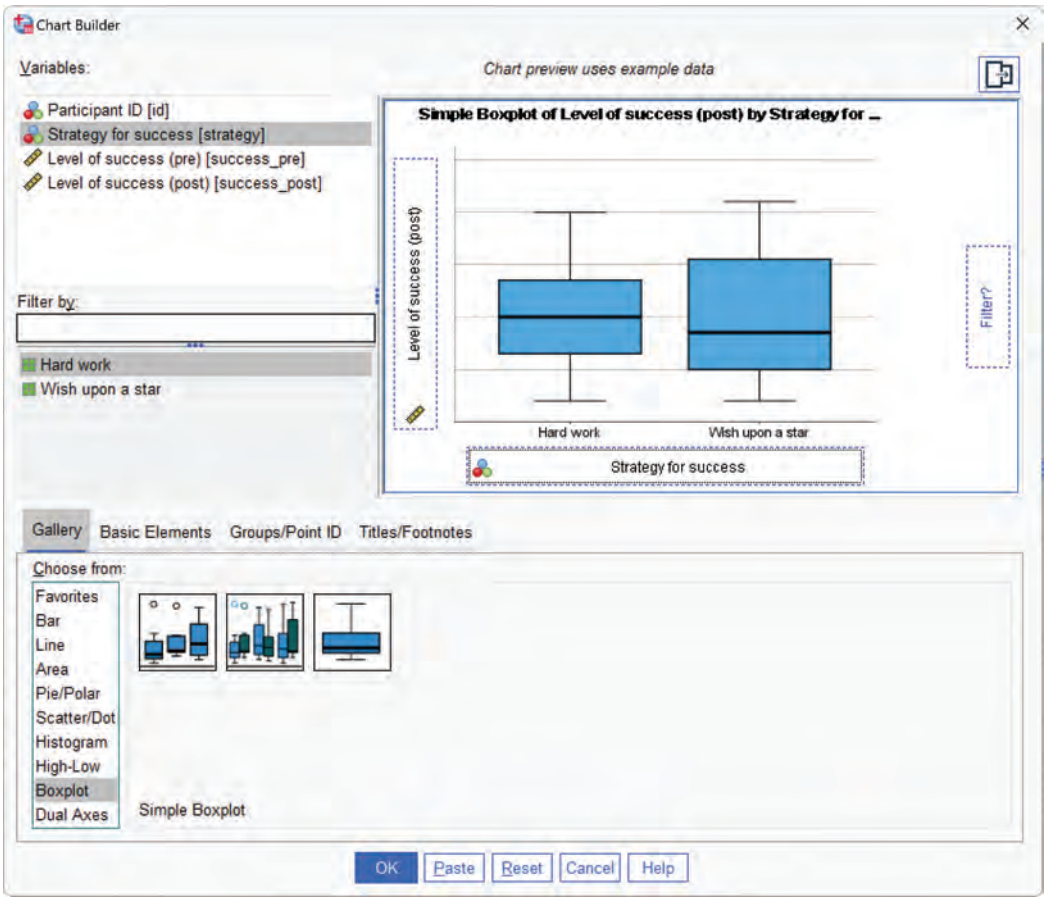


Figure 5.15 Completed dialog box for a simple boxplot

In terms of the success scores, the range of scores was much wider for the workers than the wishers, but the wishers contained an outlier (which SPSS shows as a circle) and an extreme score (which SPSS shows as an asterisk). SPSS labels these cases with the row number form the data editor (in this case rows 204 and 229), so that we can identify these scores in the data, check that they were entered correctly, or look for reasons why they might have been unusual. Like histograms, boxplots also tell us whether the distribution is symmetrical or skewed. If the whiskers are the same length then the distribution is symmetrical (the range of the top and bottom 25% of scores is the same); however, if the top or bottom whisker is much longer than the opposite whisker then the distribution is asymmetrical (the range of the top and bottom 25% of scores is different). The scores from those wishing upon a star look symmetrical because the two whiskers are of similar lengths, but the hard-work group show signs of skew because the lower whisker is longer than the upper one.



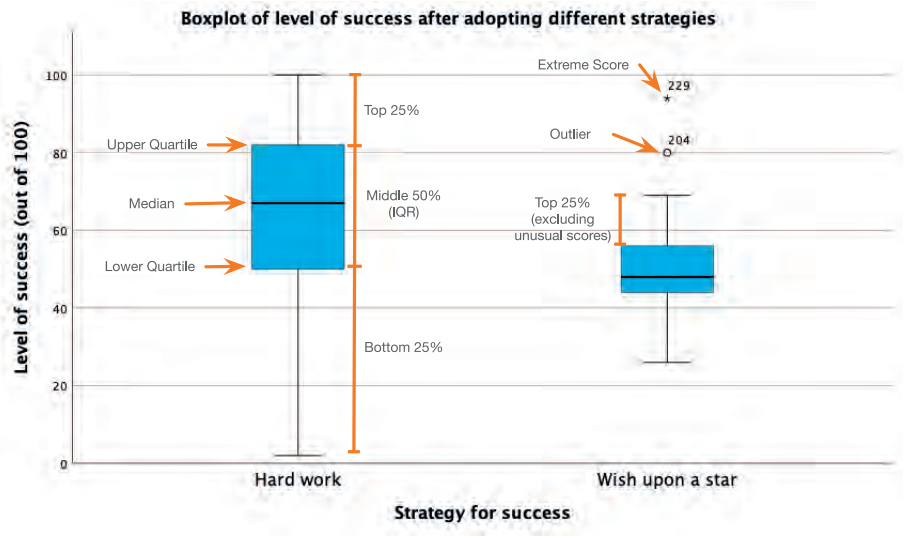


Figure 5.16 Boxplot of success scores 5 years after implementing a strategy of working hard or wishing upon a star

5.6 Visualizing means: bar charts and error bars Ⓐ

Bar charts are a common way for scientists to display means, although they probably shouldn't be because they use a lot of ink to display only one piece of information. How you create bar charts in SPSS depends on whether the means come from independent cases and so are, therefore, independent, or come from the same cases and so are related. We'll look at both situations. Our starting point is always the Chart Builder (Figure 5.5). In this dialog box select *Bar* in the list labelled *Choose from* to bring up the gallery shown in Figure 5.17. This gallery has eight icons representing different types of bar chart that you can select by double-clicking one, or by dragging it onto the canvas.

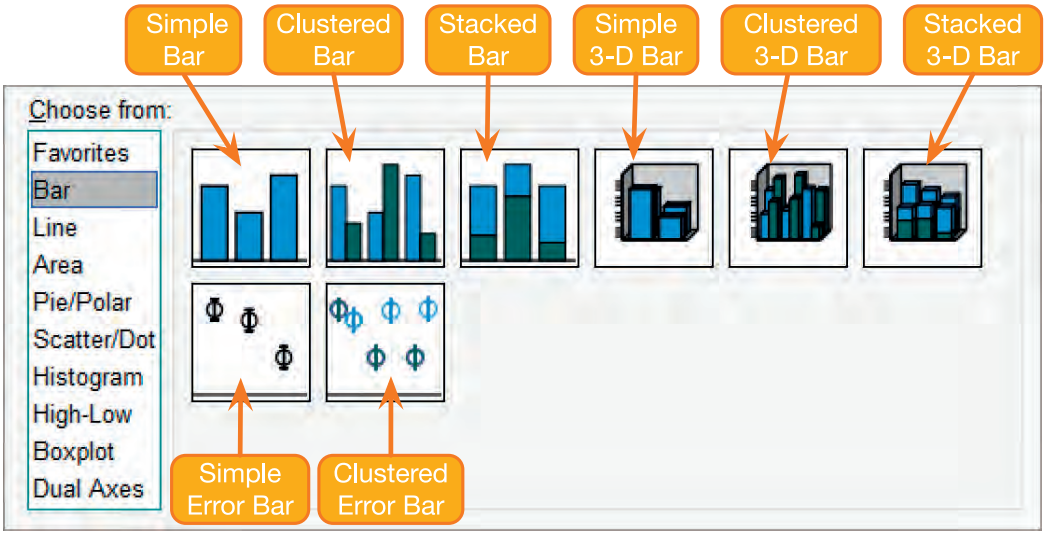


Figure 5.17 The bar chart gallery

- ⇒ *Simple bar*: Use this option to display the means of scores across different groups or categories of cases. For example, you might want to plot the mean ratings of two films.
- ⇒ *Clustered bar*: If you have a second grouping variable you can produce a simple bar chart (as above) but with different coloured bars to represent levels of a second grouping variable. For example, you could have ratings of the two films, but for each film have a bar representing ratings of 'excitement' and another bar showing ratings of 'enjoyment'.
- ⇒ *Stacked bar*: This is like the clustered bar except that the different-coloured bars are stacked on top of each other rather than placed side by side.
- ⇒ *Simple 3-D bar*: This is also like the clustered bar except that the second grouping variable is displayed not by different-coloured bars, but by an additional axis. Given what I said in Section 5.2.1 about 3-D effects obscuring the data, my advice is to stick to a clustered bar chart and not use this option.
- ⇒ *Clustered 3-D bar*: This is like the clustered bar chart above except that you can add a third categorical variable on an extra axis. The means will almost certainly be impossible for anyone to read on this type of plot so don't use it.
- ⇒ *Stacked 3-D bar*: This plot is the same as the clustered 3-D plot except the different-coloured bars are stacked on top of each other instead of standing side by side. Again, this is not a good type of plot for presenting data clearly.
- ⇒ *Simple error bar*: This is the same as the simple bar chart except that instead of bars, the mean is represented by a dot, and a line represents the precision of the estimate of the mean (usually the 95% confidence interval is plotted, but you can plot the standard deviation or standard error of the mean instead). You can add these error bars to a bar chart anyway, so really the choice between this type of plot and a bar chart with error bars is largely down to personal preference. (Including the bar adds a lot of superfluous ink, so if you want to be Tuftian about it you'd probably use this option over a bar chart.)
- ⇒ *Clustered error bar*: This is the same as the clustered bar chart except that the mean is displayed as a dot with an error bar around it. These error bars can also be added to a clustered bar chart.

5.6.1 Simple bar charts for independent means Ⓐ

Let's switch example. Imagine that a film company director was interested in whether there was really such a thing as a 'chick flick' (a film that has the stereotype of appealing to people who identify as women more than to people who identify as men). He recruited 20 people who identify as male and another 20 who identify as female. Half of each sample viewed a film that was supposed to be a 'chick flick' (*The Notebook*). The other half watched a documentary about notebook computers as a (poor) control. In all cases the company director measured participants' arousal[9] as an indicator of how much they enjoyed the film. Load the data in a file called **notebook.sav** from the companion website.

Let's plot the mean rating of the two films. To do this, double-click the icon for a simple bar chart in the Chart Builder (Figure 5.17). On the canvas you will see a plot and two drop zones: one for the *y*-axis and one for the *x*-axis. The *y*-axis needs to be the outcome variable, the thing you've measured, or more simply the thing for which you want to display the mean. In this case it would be **arousal**, so drag this variable from the variable list into the *y*-axis drop zone (

). The *x*-axis should be the variable by which we want to split the arousal data. To plot the means for the two films, drag the variable **film** from the variable list into the drop zone for the *x*-axis (

).

9 I had an email from someone expressing their 'disgust' at measuring arousal while watching a film. This reaction surprised me because to a psychologist (like me) 'arousal' means a heightened emotional response – the sort of heightened emotional response you might get from watching a film you like. Apparently if you're the sort of person who complains about the contents of textbooks then 'arousal' means something different. I can't think what.

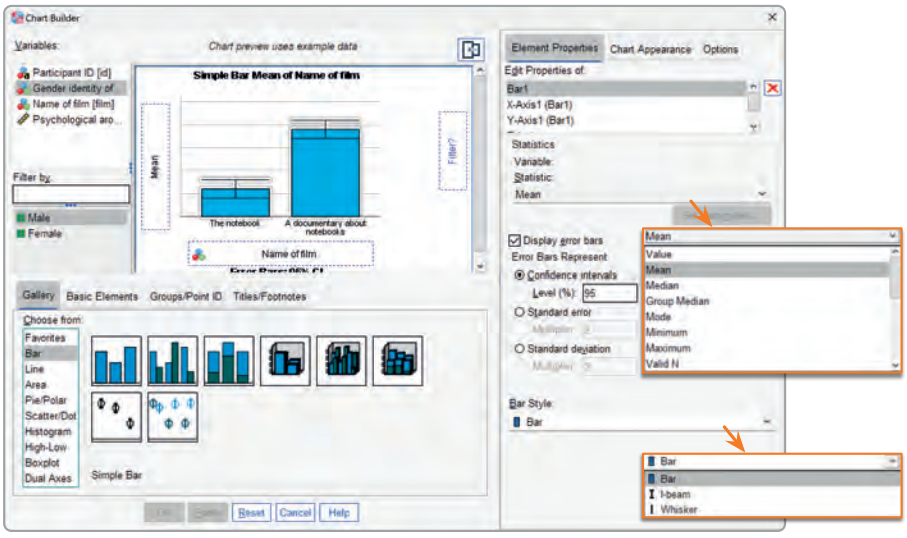


Figure 5.18 Dialog box for a simple bar chart with error bar

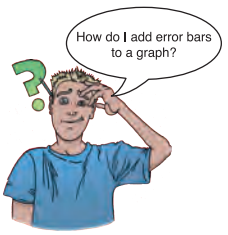


Figure 5.18 shows some other useful options in the *Element Properties* tab (if this isn't visible, click). There are three important features of this tab. The first is that, by default, the bars will display the mean value, which is fine, but note that you can plot other summary statistics such as the median or mode. Second, just because you've selected a simple bar chart, that doesn't mean that you *have* to have a bar chart. You can select to show an I-beam (the bar is reduced to a line with horizontal bars at the top and bottom), or just a whisker (the bar is reduced to a vertical line). The I-beam and whisker options might be useful when you're not planning on adding error bars, but because we are going to show error bars we should stick with a bar. Finally, you can add error bars to your chart to create an **error bar chart** by selecting Display error bars. You have a choice of what your error bars represent. Normally, error bars show the 95% confidence interval (see Section 2.9), and I have selected this option (Confidence intervals).[10] You can change the width of the confidence interval displayed by changing the '95' to a different value. You can also change the options so that instead of the confidence interval the error bars display the standard error or standard deviation. SPSS will plot 2 standard errors or standard deviations around the mean, but you can override this default by replacing 2 with a different value; for example, some people plot 1 standard error/deviation. The completed dialog box is in Figure 5.18. Click OK to produce the plot.

Figure 5.19 shows the resulting bar chart. This plot displays the mean (and the 95% confidence intervals of those means) and shows us that on average, people were more aroused by *The Notebook* than by a documentary about notebooks. However, we originally wanted to look for differences between genders, so this plot isn't telling us what we need to know. We need a *clustered plot*.[11]

5.6.2 Clustered bar charts for independent means Ⓐ

To do a clustered bar chart for means that are independent (i.e., have come from different groups) we need to double-click the clustered bar chart icon in the Chart Builder (Figure 5.17). On the canvas you

10 It's also worth mentioning at this point that because confidence intervals are constructed assuming a normal distribution, you should plot them only when this is a reasonable assumption (see Section 2.8).

11 You can also use a drop-line plot, which is described in Section 5.8.6.

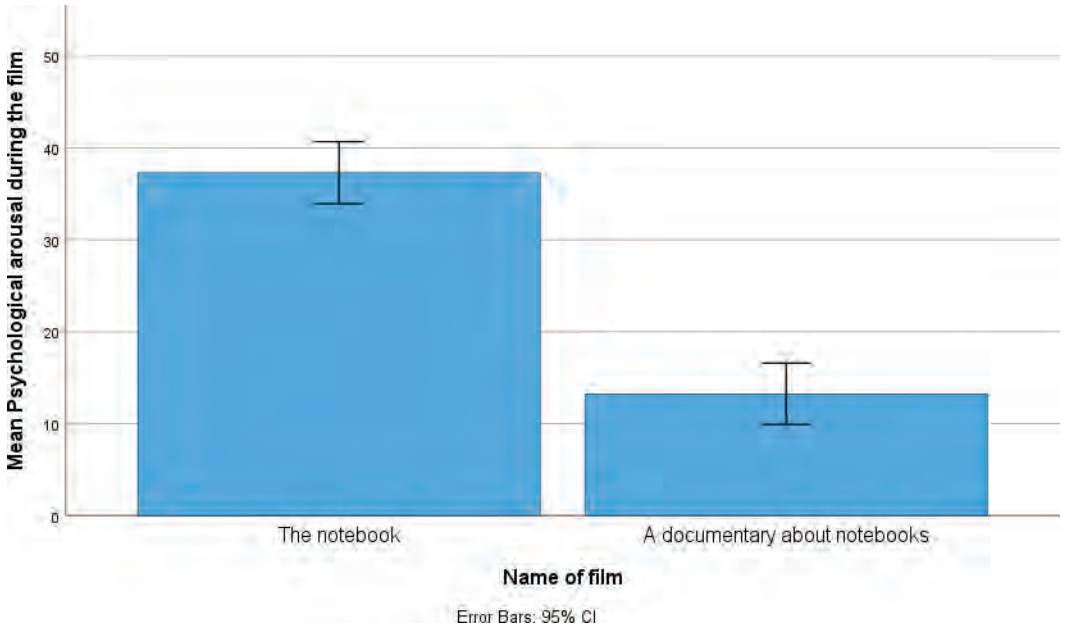


Figure 5.19 Bar chart of the mean arousal for each of the two films.

will see a plot similar to the simple bar chart but with an extra drop zone: Cluster on X: set color. We need to drag our second grouping variable into this drop zone. As with the previous example, drag **arousal** from the variable list into Y-Axis?, then drag **film** from the variable list into X-Axis?. Dragging the variable **gender_identity** into Cluster on X: set color will result in different-coloured bars representing males and females (but see SPSS Tip 5.3). As in the previous section, select error bars in the properties dialog box. Figure 5.20 shows the completed Chart Builder. Click OK to produce the plot.



Figure 5.20 Dialog box for a clustered bar chart with error bar

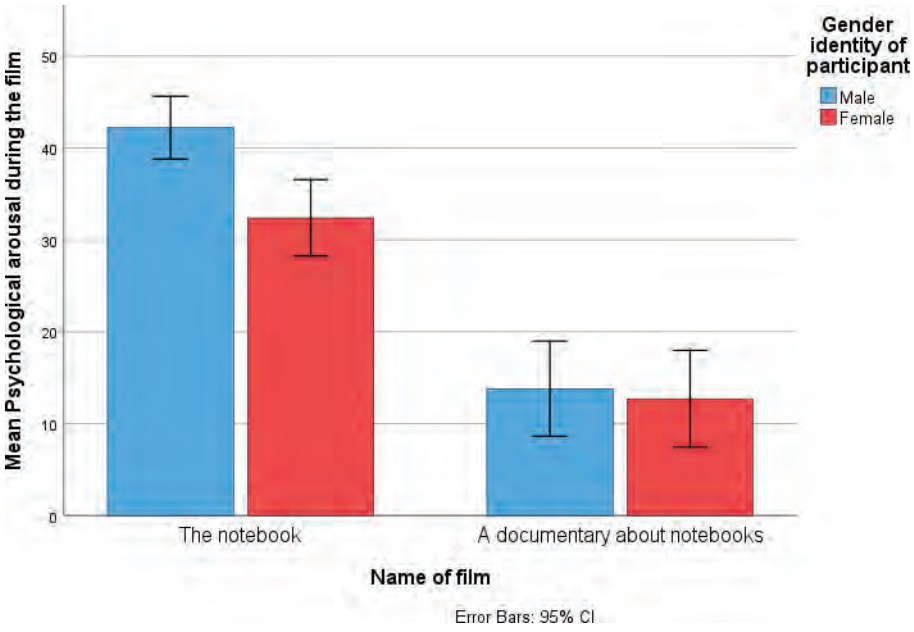
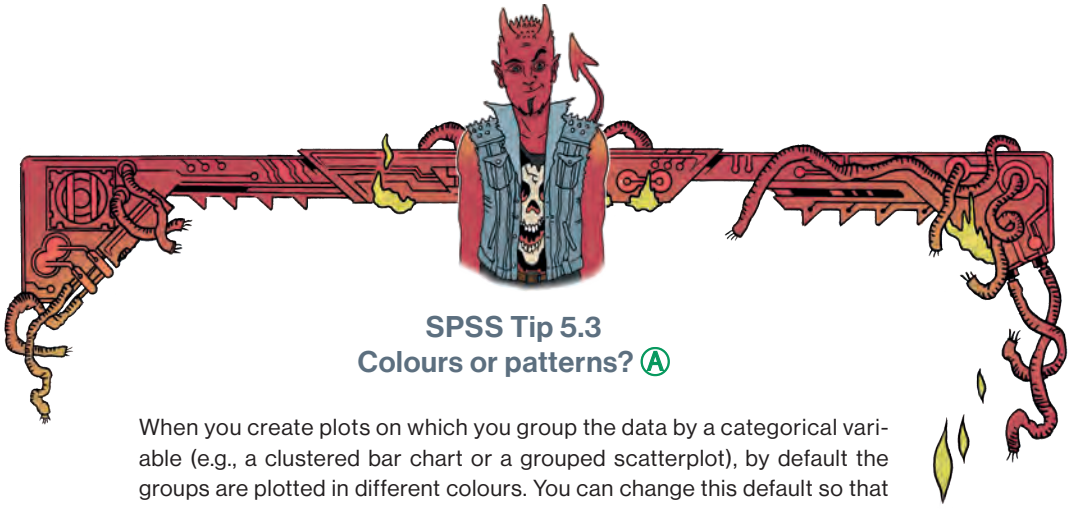


Figure 5.21 Bar chart of the mean arousal for each of the two films

Figure 5.21 shows the resulting bar chart. Like the simple bar chart, this plot tells us that arousal was overall higher for *The Notebook* than for the documentary about notebooks, but it splits this information by gender identity. The mean arousal for *The Notebook* shows that those identifying as male were more aroused during this film than those identifying as female, suggesting that they enjoyed the film more. Contrast this with the documentary, for which arousal levels are comparable in the two gender identity groups. On the face of it, this result contradicts the idea of a 'chick flick': it seems that people who identify as male enjoy these films more than those who identify as women (deep down, males are romantics at heart …).



SPSS Tip 5.3
Colours or patterns? Ⓐ

When you create plots on which you group the data by a categorical variable (e.g., a clustered bar chart or a grouped scatterplot), by default the groups are plotted in different colours. You can change this default so that the groups are plotted using different patterns. In a bar chart the result is bars filled with different patterns, not different colours. With a scatterplot (see later) different symbols are used to show data from different groups (rather than colours). To make this change, double-click in the

Cluster on X: set color drop zone (bar chart) or Set color? (scatterplot) to bring up a new dialog box. Within this dialog box there is a drop-down list labelled *Distinguish Groups by* within which you can select *Color* or *Pattern*. To change the default select *Pattern* and then click OK. You can switch back to displaying groups in different colours in the same way.

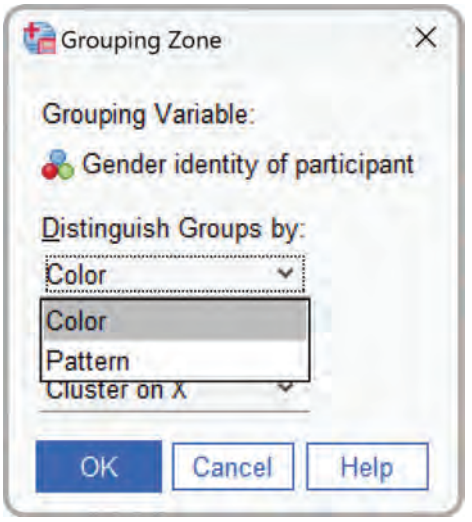
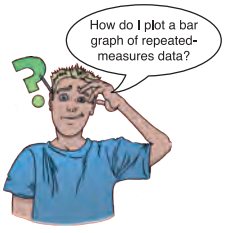


Figure 5.22 Dialog box to define whether groups are displayed in different colours or patterns

5.6.3 Simple bar charts for related means Ⓐ



Visualizing means from the same entities is trickier, but as they say, if you're going to die, die with your boots on. So, let's put our boots on and hopefully not die. Hiccups can be a serious problem: Charles Osborne apparently got a case of hiccups that lasted 67 years while slaughtering a hog (well, who wouldn't?). People have many methods for stopping hiccups (a surprise, holding your breath), and medical science has put its collective mind to the task too. The official treatment methods include tongue-pulling manoeuvres, massage of the carotid artery, and, believe it or not, digital rectal massage (Fesmire, 1988). I don't know the details of what the digital rectal massage involved, but I can imagine. Let's say we wanted to put digital rectal massage to the test (erm, as a cure for hiccups). We took 15 hiccup sufferers, and during a bout of hiccups administered each of the three procedures (in random order and at intervals of 5 minutes) after taking a baseline of how many hiccups they had per minute. We counted the number of hiccups in the minute after each procedure. Load the file **hiccups.sav**. Note that these data are laid out in different columns; there is no grouping variable that specifies the interventions, because each patient experienced all interventions. In the previous two examples, we used grouping variables to specify aspects of the plot (e.g., we used the grouping variable **film** to specify

the *x*-axis). For repeated-measures data we don't have these grouping variables and so the process of building a plot is a little more complicated (but only a little).

To plot the mean number of hiccups, go to the Chart Builder and double-click the icon for a simple bar chart (Figure 5.17). As before, you will see a plot on the canvas with drop zones for the *x*- and *y*-axes. Previously we specified the column in our data that contained data from our outcome measure on the *y*-axis, but for these data our outcome variable (number of hiccups) is spread over four columns. We need to drag all four of these variables from the variable list into the *y-axis* drop zone simultaneously. To do this, we first select multiple items in the variable list by clicking on the first variable that we want (which becomes highlighted), then holding down the *Ctrl* key (or *Cmd* if you're on a Mac) while we click any others. Each variable you click will become highlighted to indicate that it has been selected. Sometimes (as is the case here) you want to select a list of consecutive variables, in which case you can click the first variable that you want to select (in this case **Baseline**), hold down the *Shift* key (also on a Mac) and then click the last variable that you want to select (in this case **Digital rectal massage**); this action selects these two variables and any in between. Once you have selected the four variables, click any one of them (while still pressing *Cmd* or *Shift* if you're on a Mac) and drag to the Y-Axis? drop zone. This action transfers all selected variables to that drop zone (see Figure 5.23).

Once you have dragged the four variables onto the *y*-axis drop zone a new dialog box appears (Figure 5.24). This box tells us that SPSS is creating two temporary variables. One is called **Summary**, which is going to be the outcome variable (i.e., what we measured – in this case the number of hiccups per minute). The other is called **Index**, which will represent our independent variable (i.e., what we manipulated – in this case the type of intervention). SPSS uses these temporary names because it doesn't know what our variables represent, but we will change them to be something more helpful. First, click OK to get rid of this dialog box.

To edit the names of the **Summary** and **Index** variables, we use the *Element Properties* tab, which we have used before; if you can't see it then click . Figure 5.25 shows the options that need to be set.

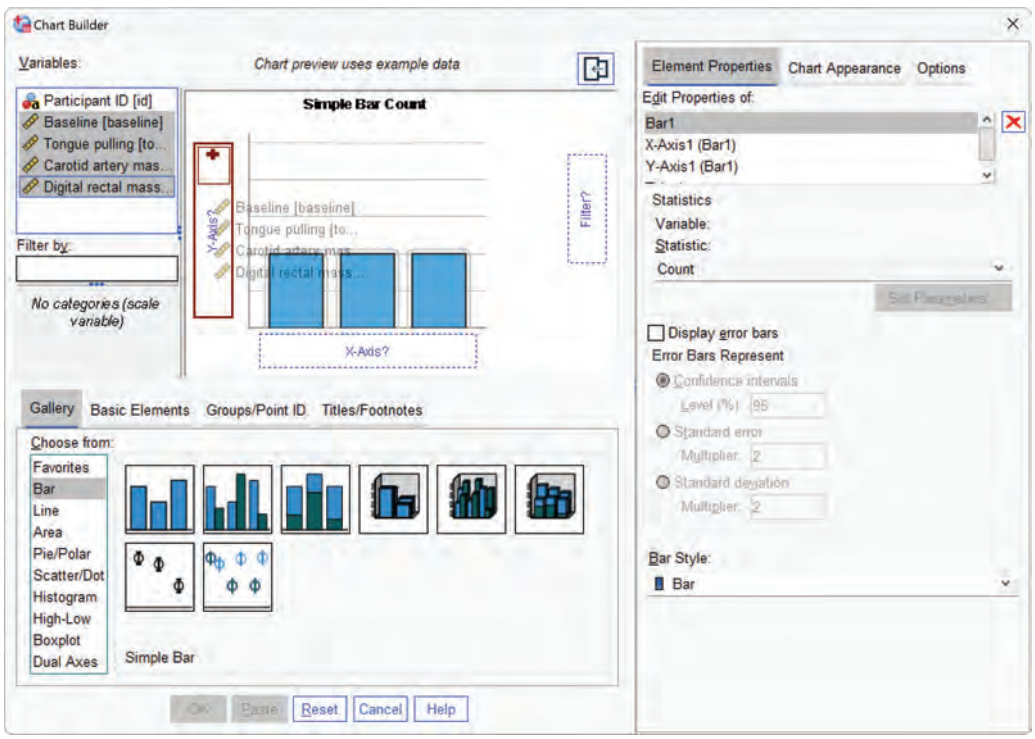


Figure 5.23 Specifying a simple bar chart for repeated-measures data

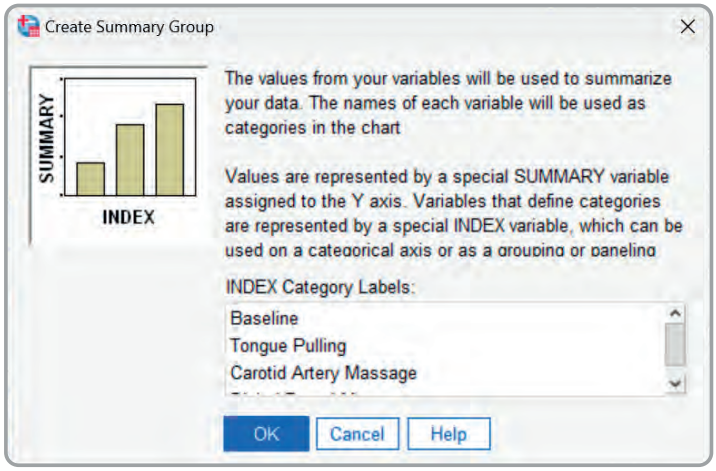


Figure 5.24 The *Create Summary Group* dialog box

In the left panel note that I have selected to display error bars (see the previous two sections for more information). The middle panel is accessed by clicking on *X-Axis1 (Bar1)* in the list labelled *Edit Properties of*, which allows us to edit properties of the horizontal axis. First, we'll give the axis a sensible title. I have typed *Intervention* in the space labelled *Axis Label*, which will now be the *x*-axis label on the plot. We can also change the order of our variables by selecting a variable in the list labelled *Order* and moving it up or down using and . This is useful if the levels of our predictor variable have a meaningful order that is not reflected in the order of variables in the data editor. If we change our mind about displaying one of our variables then we can remove it from the list by selecting it and clicking on .

The right panel of Figure 5.25 is accessed by clicking on *Y-Axis1 (Bar1)* in the list labelled *Edit Properties of*, which allows us to edit properties of the vertical axis. The main change that I have made here is to give the axis a label so that the final plot has a useful description on the axis (by default it will display 'Mean', which is too vague). I have typed 'Mean number of hiccups per minute' in the box labelled *Axis Label*. Also note that you can use this dialog box to set the scale of the vertical axis (the minimum value, maximum value and the major increment, which is how often a mark appears on the axis).

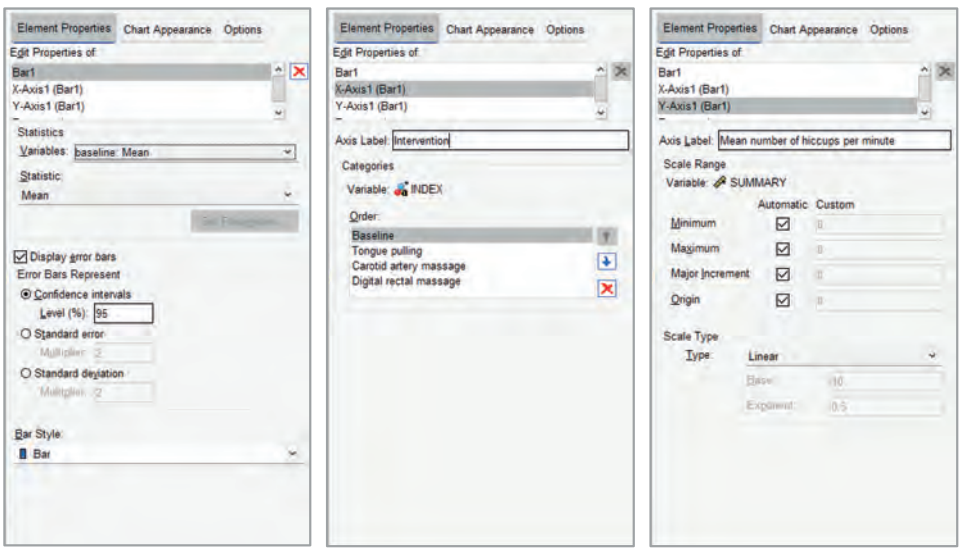


Figure 5.25 Setting *Element Properties* for a repeated-measures plot

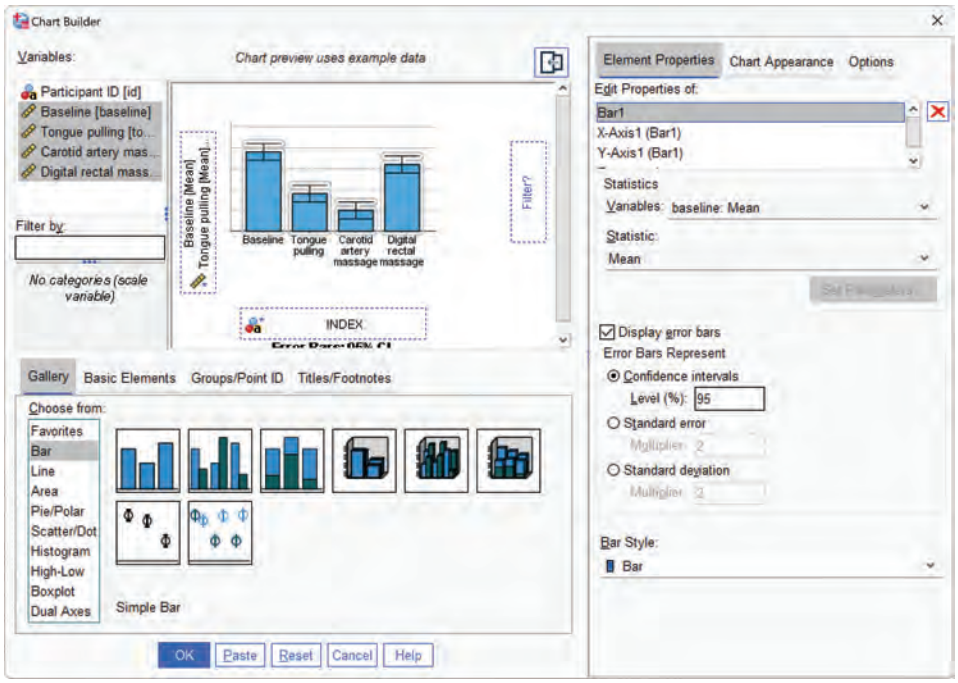


Figure 5.26 Completed Chart Builder for a repeated-measures plot

Mostly you can let SPSS construct the scale automatically – if it doesn't do it sensibly you can edit it later. Figure 5.26 shows the completed Chart Builder. Click OK to produce the plot.

The resulting bar chart in Figure 5.27 displays the mean number of hiccups (and associated 95% confidence interval)[12] at baseline and after the three interventions. Note that the axis labels that I typed in have appeared on the plot. We can conclude that the amount of hiccups after tongue pulling was about the same as at baseline; however, carotid artery massage reduced hiccups, but not by as much as digital rectal massage. The moral? If you have hiccups, find some digital amusement for a few minutes. Lock the door first.

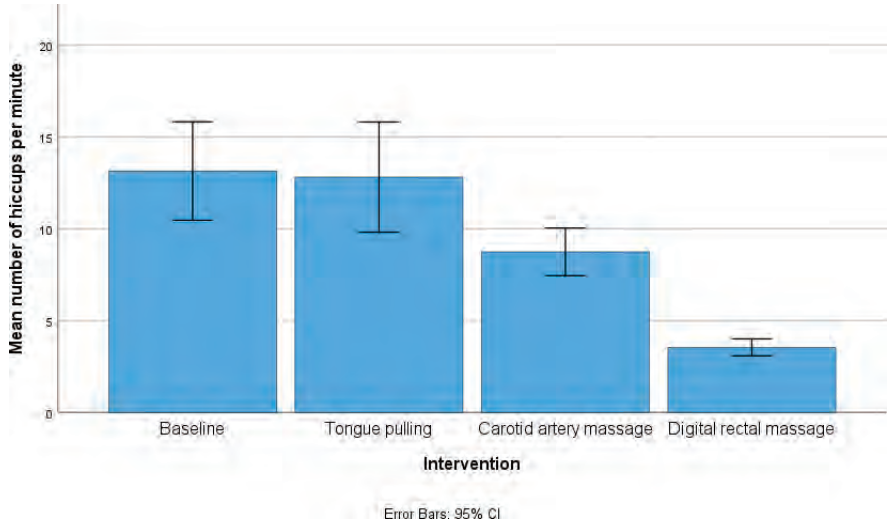


Figure 5.27 Bar chart of the mean number of hiccups at baseline and after various interventions

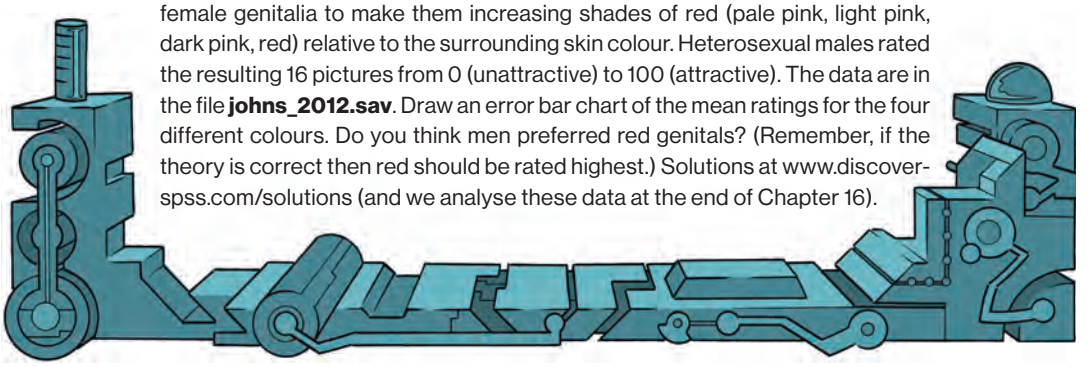
12 The error bars on plots of repeated-measures designs should be adjusted as we will see in Chapter 10; so if you're plotting your own data have a look at Section 10.9.2 before you do.



Labcoat Leni's Real Research 5.2
Seeing red Ⓐ

Johns, S. E. et al. (2012). *PLoS One, 7*(4), e34669.

Certain theories suggest that heterosexual males have evolved a predisposition towards the colour red because it is sexually salient. The theory also suggests that women use the colour red as a proxy signal for genital colour to indicate ovulation and sexual proceptivity. Decide for yourself how plausible this theory seems, but a hypothesis to test it is that using the colour red in this way ought to attract men (otherwise it's a pointless strategy). In a novel study, Johns et al. (2012) tested this idea by manipulating the colour of four pictures of female genitalia to make them increasing shades of red (pale pink, light pink, dark pink, red) relative to the surrounding skin colour. Heterosexual males rated the resulting 16 pictures from 0 (unattractive) to 100 (attractive). The data are in the file **johns_2012.sav**. Draw an error bar chart of the mean ratings for the four different colours. Do you think men preferred red genitals? (Remember, if the theory is correct then red should be rated highest.) Solutions at www.discoverspss.com/solutions (and we analyse these data at the end of Chapter 16).



5.6.4 Clustered bar charts for related means Ⓐ

Now we have seen how to plot means that are related (i.e., display scores from the same group of cases in different conditions), you might well wonder what you do if you have a second independent variable that had been measured in the same sample. You'd do a clustered bar chart, right? Unfortunately the SPSS Chart Builder doesn't appear to be able to cope with this situation at all – at least not that I can work out from playing about with it. (Cue a deluge of emails along the general theme of 'Dear Professor Field, I was recently looking through my FEI Titan 80-300 scanning transmission electron microscope and I think I may have found your brain. I have enclosed it for you – good luck finding it in the envelope. May I suggest that you take better care next time there is a slight gust of wind or else, I fear, it might blow out of your head again. Yours, Professor Enormobrain. PS Doing clustered charts for related means in SPSS is simple for anyone whose mental acumen can raise itself above that of a louse.')

5.6.5 Clustered bar charts for 'mixed' designs Ⓐ

The Chart Builder *can* produce plots of a mixed design (see Chapter 14). A mixed design has one or more independent variables measured using different groups, and one or more independent variables measured using the same entities. The Chart Builder can produce a plot provided you have only one repeated-measures variable.

My students like to message on their phones during my lectures (I assume they text the person next to them things like 'This loser is so boring – need to poke my eyes out to relieve tedium. LOL.', or tweeting 'In lecture of @profandyfield #WillThePainNeverEnd). With all this typing on phones, though, what will become of humanity? Maybe we'll evolve miniature thumbs, or lose the ability to write correct English. Imagine we conducted an experiment in which a group of 25 people were encouraged to message their friends and post on social media using their mobiles over a 6-month period. A second group of 25 people were banned from messaging and social media for the same period by being given armbands that administered painful shocks in the presence of microwaves (like those emitted from phones).[13] The outcome was a percentage score on a grammatical test that was administered both before and after the intervention. The first independent variable was, therefore, social media use (encouraged or banned) and the second was the time at which grammatical ability was assessed (baseline or after 6 months). The data are in the file **social_media.sav**.

To plot these data we begin as though we are creating a clustered bar chart (Section 5.6.2). However, because one of our independent variables was a repeated measure, we specify the outcome variable as we did for a bar chart of related means (Section 5.6.3). Our repeated-measures variable is time (whether grammatical ability was measured at baseline or after 6 months) and is represented in the data file by two columns, one for the baseline data and the other for the follow-up data. In the Chart Builder select these two variables simultaneously by clicking on one and then holding down the *Ctrl* key (*Cmd* on a Mac) on the keyboard and clicking on the other. When they are both highlighted drag them (keep *Cmd* pressed on a Mac) into Y-Axis? as shown in Figure 5.28.

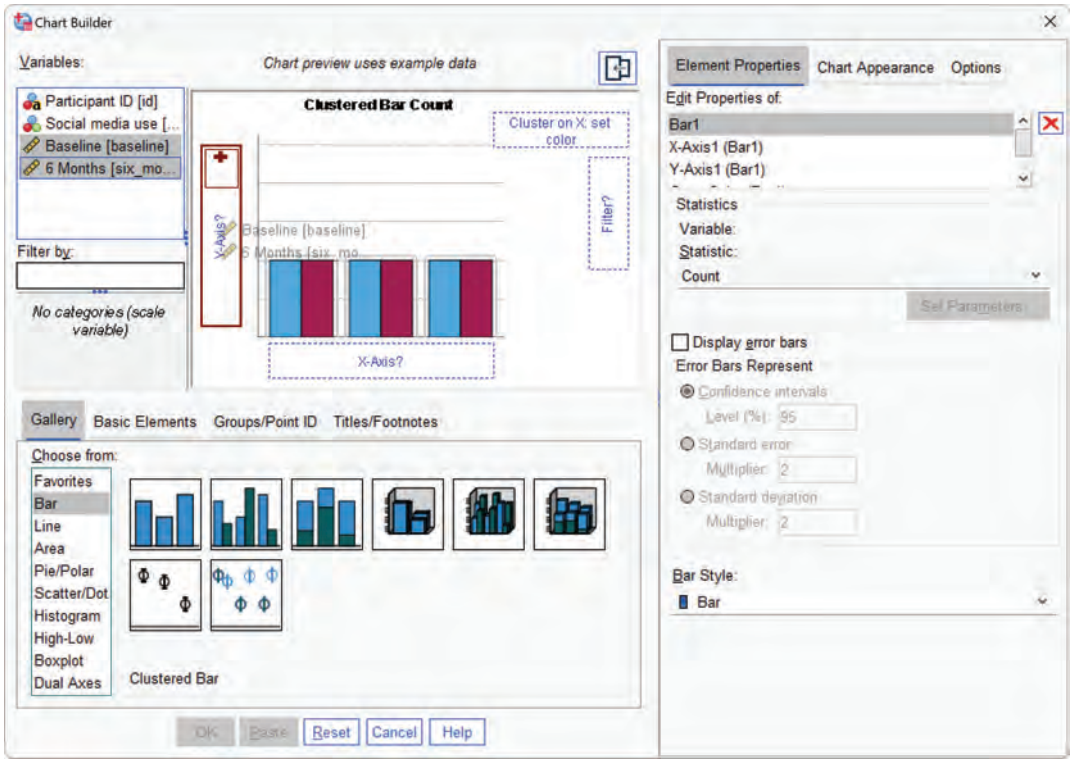
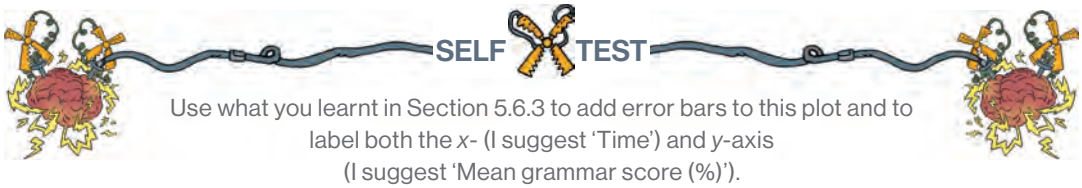


Figure 5.28 Selecting the repeated-measures variable in the Chart Builder

13 This turned out to be a very bad idea because other people's phones also emit microwaves. Let's just say there are now 25 people battling chronic learned helplessness.

The second variable (whether people were encouraged to use social media or were banned) was measured using different participants and is represented in the data file by a grouping variable (**Social media use**). Drag this variable from the variable list into Cluster on X: set color. The two groups will be displayed as different-coloured bars. The finished Chart Builder is in Figure 5.29. Click OK to produce the plot.



Use what you learnt in Section 5.6.3 to add error bars to this plot and to label both the *x*- (I suggest 'Time') and *y*-axis (I suggest 'Mean grammar score (%)').

Figure 5.30 shows the resulting bar chart. It shows that at baseline (before the intervention) the grammar scores were comparable in our two groups; however, after the intervention, the grammar scores were lower in those encouraged to use social media than in those banned from using it. If you compare the two blue bars you can see that social media users' grammar scores have fallen over the 6 months; compare this to the controls (green bars) whose grammar scores are similar over time. We might, therefore, conclude that social media use has a detrimental effect on people's understanding of English grammar. Consequently, civilization will crumble and Abaddon will rise cackling from his bottomless pit to claim our wretched souls. Maybe.

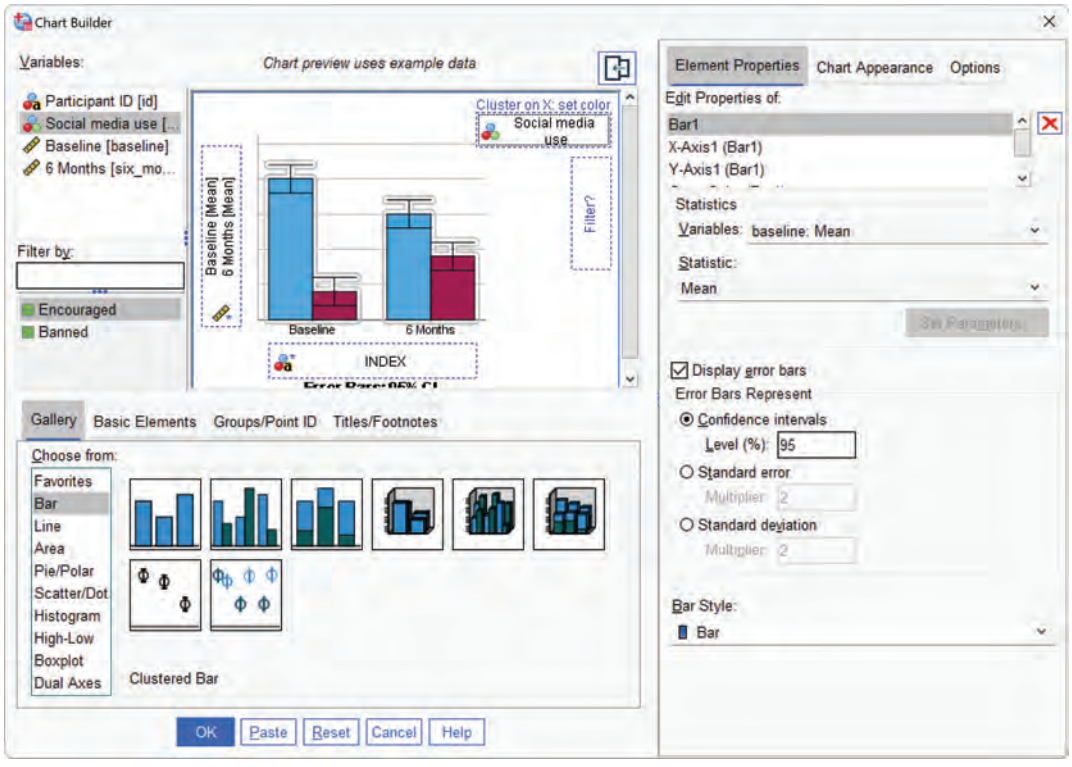


Figure 5.29 Completed dialog box for an error bar plot of a mixed design

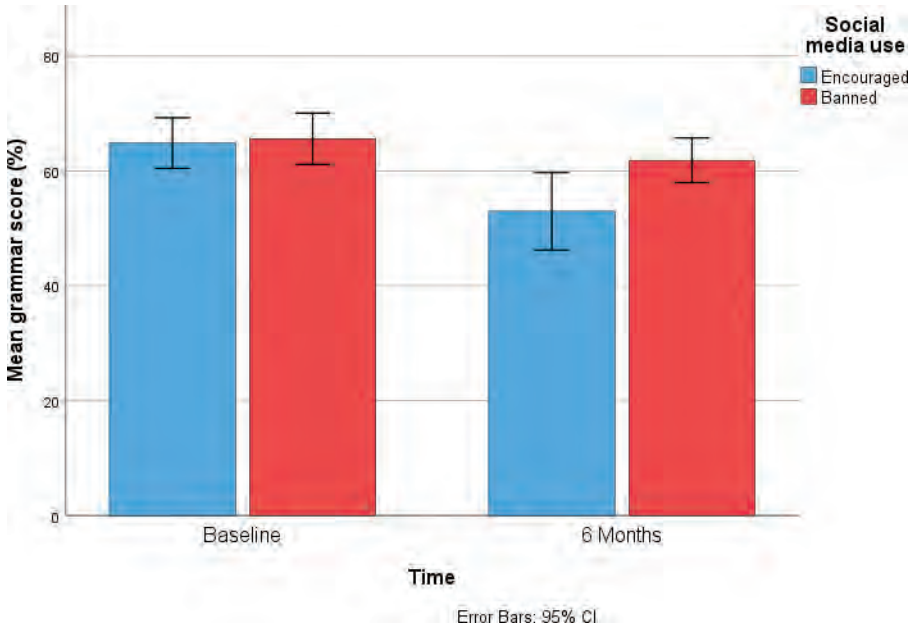


Figure 5.30 Error bar plot of the mean grammar score over 6 months in people who were encouraged to use social media compared to those who were banned

5.7 Line charts Ⓐ

Line charts are bar charts but with lines instead of bars. Therefore, everything we have just done with bar charts we can do with line charts instead. As ever, our starting point is the Chart Builder (Figure 5.5). In this dialog box select *Line* in the list labelled *Choose from* to bring up the gallery shown in Figure 5.31. This gallery has two icons which you can double-click or drag into the canvas to initialize a plot.

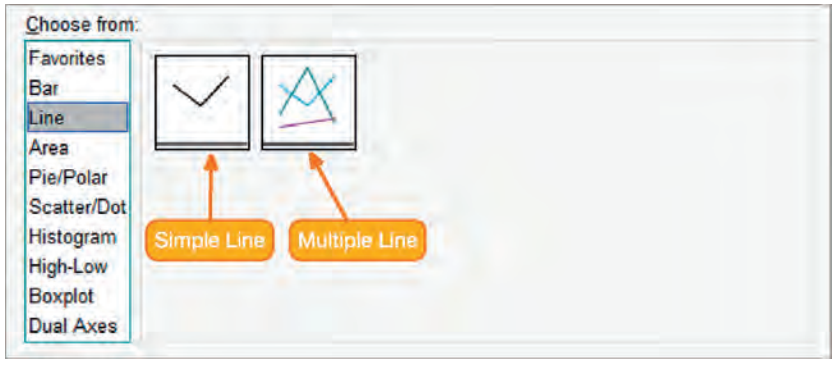


Figure 5.31 The line chart gallery

⇒ *Simple line*: Use this option to display the means of scores across different groups of cases.
⇒ *Multiple line*: This option is equivalent to the clustered bar chart: it will plot means of an outcome variable for different categories/groups of a predictor variable and produce different-coloured lines for each category/group of a second predictor variable.

SELF TEST

The procedure for producing line charts is basically the same as for bar charts. Follow the previous sections for bar charts but selecting a simple line chart instead of a simple bar chart, and a multiple line chart instead of a clustered bar chart. Produce line chart equivalents of each of the bar charts in the previous section. If you get stuck, the self-test answers will help (www.discoverspss.com/solutions).

5.8 Visualizing relationships: the scatterplot Ⓐ



Sometimes we need to look at the relationships between variables (rather than their means or frequencies). A **scatterplot** plots each entity's score on one variable against their score on another. It visualizes the relationship between the variables, but also helps us to identify unusual cases that might bias that relationship. In fact, we encountered a scatterplot when we discussed effect sizes (see Section 3.7.4).

As ever, open the Chart Builder dialog box (Figure 5.5). Select *Scatter/Dot* in the list labelled *Choose from* to bring up the gallery shown in Figure 5.32. This gallery has seven icons representing different types of scatterplot. Select one by double-clicking on it or dragging it to the canvas.

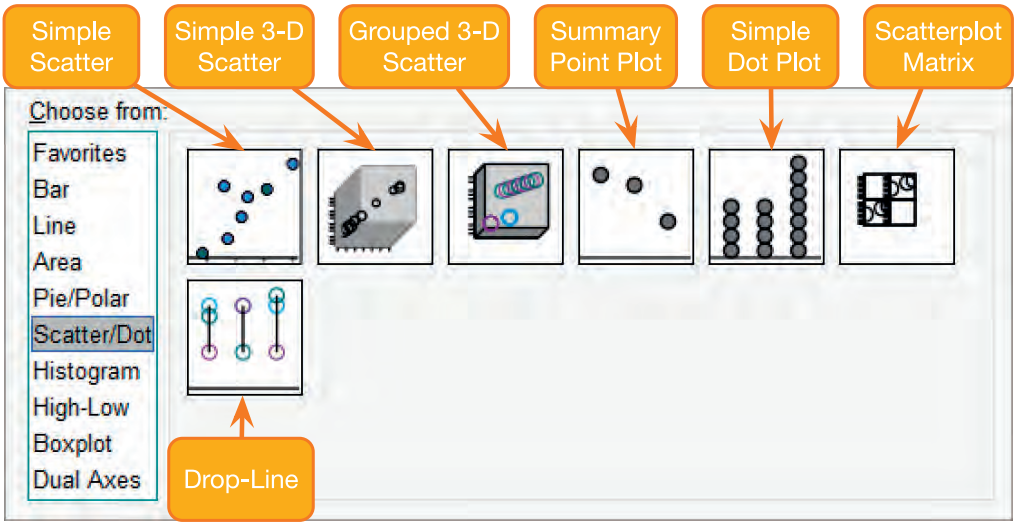


Figure 5.32 The scatter/dot gallery

⇒ *Simple scatter*: Use this option to plot values of one continuous variable against another. You can also choose to display points belonging to different groups in different colours (or symbols).

⇒ *Simple 3-D scatter*: Use this option (or preferably don't because these plots are typically hard to interpret) to plot values of one continuous variable against values of two others.

⇒ *Grouped 3-D scatter*: As above, but with the option to differentiate groups of cases with different-coloured dots.

⇒ *Summary point plot*: This plot is the same as a bar chart (see Section 5.6), except that a dot is used instead of a bar.

⇒ *Simple dot plot*: Otherwise known as a **density plot**, this plot is like a histogram (see Section 5.4), except that, rather than having a summary bar representing the frequency of scores, individual scores are displayed as dots. Like histograms, they are useful for looking at the shape of the distribution of scores.

⇒ *Scatterplot matrix*: This option produces a grid of scatterplots showing the relationships between multiple pairs of variables in each cell of the grid. These plots are much easier to interpret than 3-D scatterplots.

⇒ *Drop-line:* This option produces a plot similar to a clustered bar chart (see, for example, Section 5.6.2) but with a dot representing a summary statistic (e.g., the mean) instead of a bar, and with a line connecting the 'summary' (e.g. mean) of each group. These plots are useful for comparing statistics, such as the mean, across groups or categories.

5.8.1 Simple scatterplot Ⓐ

This type of scatterplot is for looking at just two variables. For example, a psychologist was interested in the effects of exam stress on exam performance. She devised and validated a questionnaire to assess state anxiety relating to exams (called the Exam Anxiety Questionnaire, or EAQ). This scale produced a measure of anxiety scored out of 100. Anxiety was measured before an exam, and the percentage mark of each student on the exam was used to assess the exam performance. The first thing that the psychologist should do is draw a scatterplot of the two variables (her data are in the file **exam_anxiety.sav**, so load this file into SPSS).

In the Chart Builder double-click the icon for a simple scatterplot (Figure 5.33). On the canvas you will see a plot and two drop zones: one for the *y*-axis and one for the *x*-axis. Typically the *y*-axis

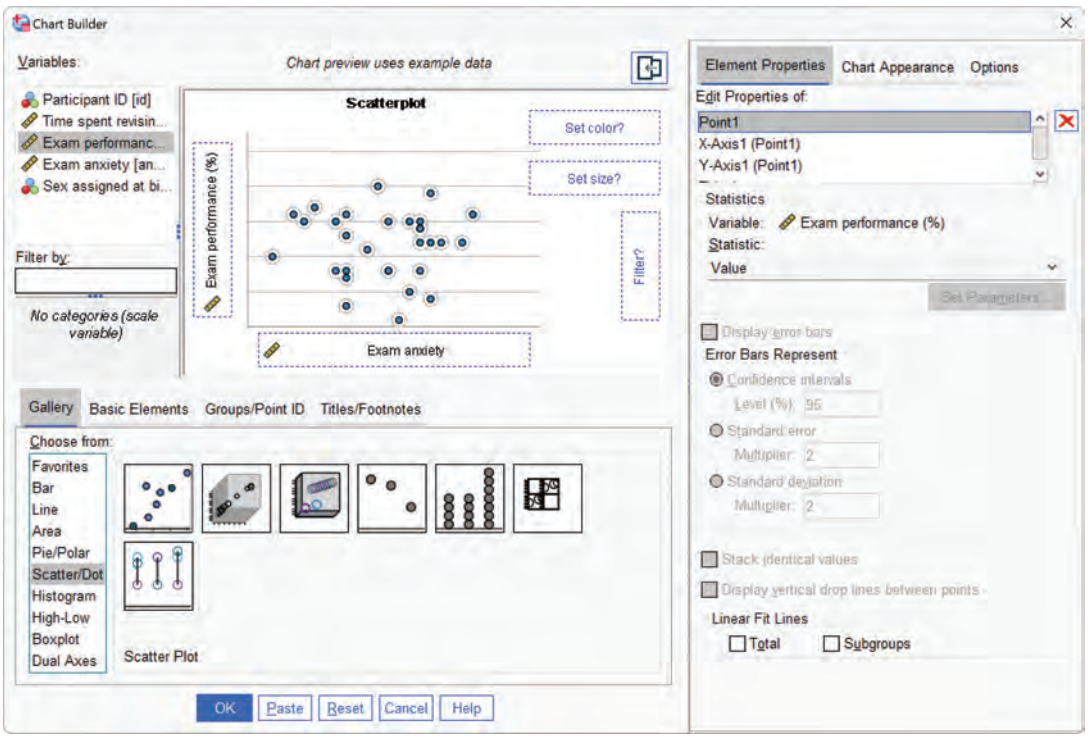


Figure 5.33 Completed *Chart Builder* dialog box for a simple scatterplot

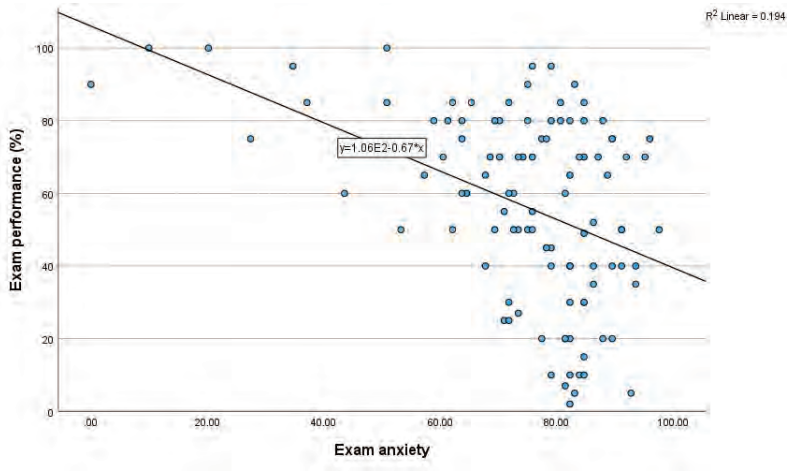


Figure 5.34 Scatterplot of exam anxiety and exam performance

displays the outcome variable and the *x*-axis the predictor.[14] In this case the outcome is exam performance (**Exam performance (%)**), so drag it from the variable list into Y-Axis?, and the predictor is exam anxiety (**Exam anxiety**), so drag it into X-Axis?. Figure 5.33 shows the completed Chart Builder. Click OK to produce the plot.

Figure 5.34 shows the resulting scatterplot; yours won't have a funky line on it yet, but don't get too despondent about it because I'm going to show you how to add one very soon. The scatterplot tells us that the majority of students suffered from high levels of anxiety (there are very few cases that had anxiety levels below 60). Also, there are no obvious outliers in that most points seem to fall within the vicinity of other points. There also seems to be some general trend in the data, shown by the line, such that higher levels of anxiety are associated with lower exam scores and low levels of anxiety are almost always associated with high examination marks. Another noticeable trend in these data is that there were no cases having low anxiety and low exam performance – in fact, most of the data are clustered in the upper region of the anxiety scale.



Often it is useful to plot a line that summarizes the relationship between variables on a scatterplot (this is called a **regression line** and we will discover more about them in Chapter 9). Figure 5.35 shows the process of adding a regression line to the scatterplot. First, any plot in the SPSS viewer can be edited by double-clicking on it to open the SPSS **Chart Editor** (we explore this window in detail in Section 5.9). Once in the *Chart Editor*, click to open the *Properties* dialog box. Using this dialog box, we can add a line to the plot that represents the overall mean of all data, a linear (straight line) model, a quadratic model, a cubic model and so on (these trends are described in Section 12.5.5). We'll add a linear regression line, so select Linear. By default, SPSS attaches a label to the line (Attach label to line) containing the equation for the line (more on this in Chapter 9). Often this label obscures the data so I tend to switch this option off. Click Apply to register any property changes to the scatterplot. To exit the *Chart Editor* close the window ( ). The scatterplot should now look like Figure 5.34. A variation on the scatterplot is the catterplot, which is useful for plotting unpredictable data files that view you as their human servant (Jane Superbrain Box 5.1).

14 This makes sense in experimental research because changes in the independent variable (the variable that the experimenter has manipulated) cause changes in the dependent variable (outcome). In English we read from left to right so by having the causal variable on the horizontal we naturally scan across the changes in the 'cause' and see the effect on the vertical plane as we do. In correlational research, variables are measured simultaneously and so no cause-and-effect relationship can be established, so although we can still talk about *predictors* and *outcomes* these terms do not imply causal relationships.

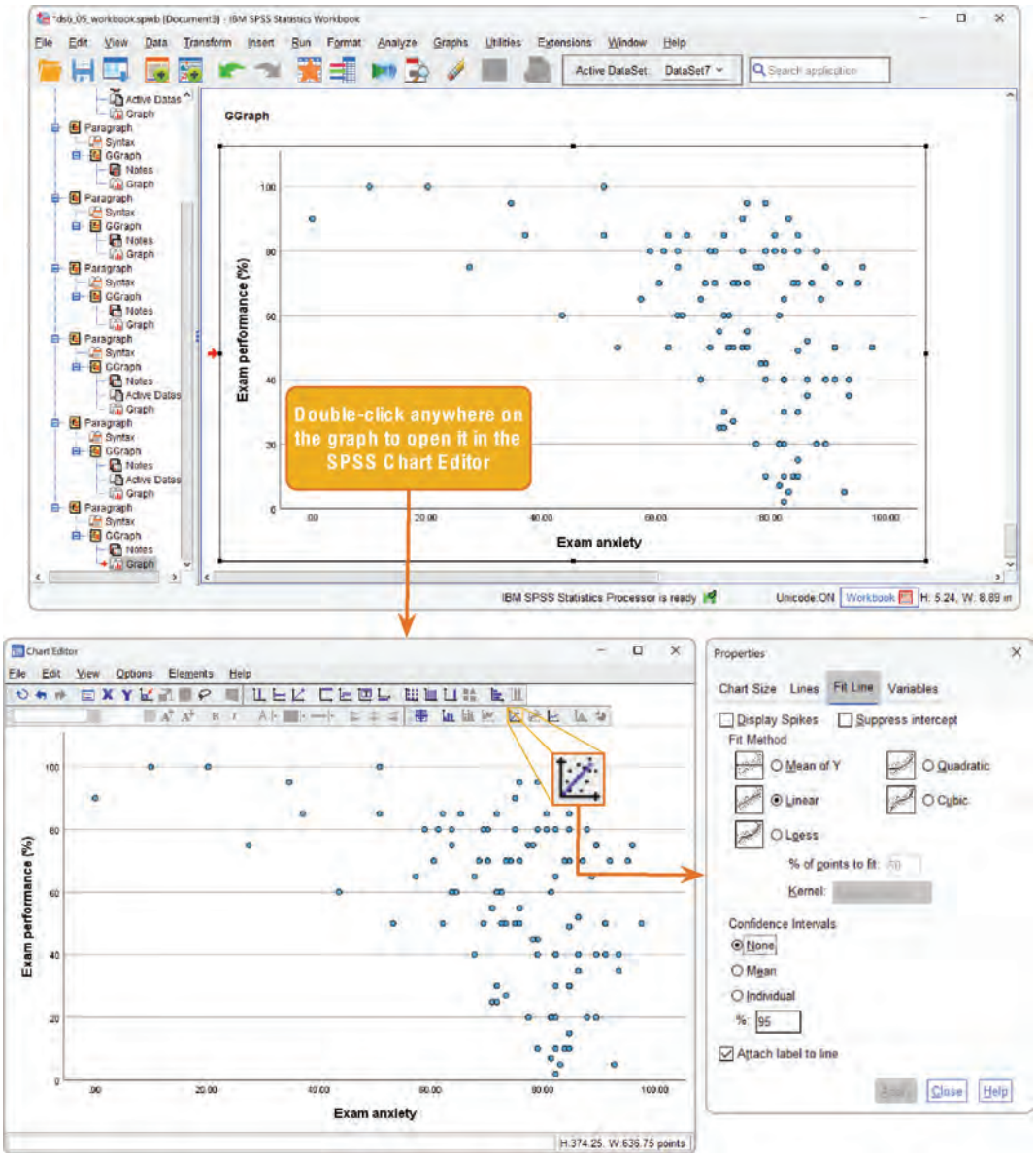
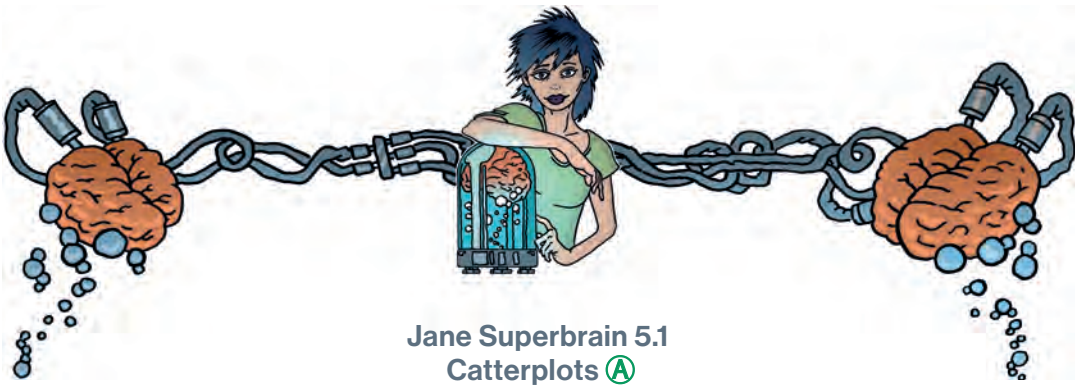


Figure 5.35 Opening the Chart Editor and *Properties* dialog box for a simple scatterplot

5.8.2 Grouped scatterplot Ⓐ

Imagine that we want to see whether male and female students had different reactions to exam anxiety. We can visualize this with a grouped scatterplot, which displays scores on two continuous variables, but colours the data points by a third categorical variable. To create this plot for the exam anxiety data, double-click the grouped scatter icon in the Chart Builder (Figure 5.32). As in the previous example, drag **Exam Performance (%)** from the variable list into Y-Axis?, and drag **Exam Anxiety** into X-Axis?. There is an additional drop zone (Set color?) into which we can drop any categorical variable. If we want to visualize the relationship between exam anxiety and performance separately for students assigned male or female at birth then we drag **Sex assigned at birth** into Set color?. (To display the sexes using different symbols rather than colours see SPSS Tip 5.3.) Figure 5.37 shows the completed Chart Builder. Click OK to produce the plot.



Jane Superbrain 5.1
Catterplots Ⓐ

The catterplot is a variation on the scatterplot that was designed by Herman Garfield to overcome the difficulty that sometimes emerges when plotting very unpredictable data. He named it the catterplot because of all the things he could think of that were unpredictable, cat behaviour topped his list. To illustrate the catterplot open the data in the file **catterplot.sav**. These data measure two variables: the time since last feeding a cat (**dinner_time**), and how loud their purr is (**meow**). In SPSS, to create a catterplot you follow the same procedure as a simple scatterplot: drag **dinner_time** into the drop zone for the *x*-axis (X-Axis?), then drag **meow** into the *y*-axis drop zone (Y-Axis?). Click OK to produce the plot.

The catterplot is shown in Figure 5.36. You might expect that there is a positive relationship between the variables: the longer the time since being fed, the more vocal the cat becomes. However, the plot shows something quite different: there doesn't seem to be a consistent relationship.[15]

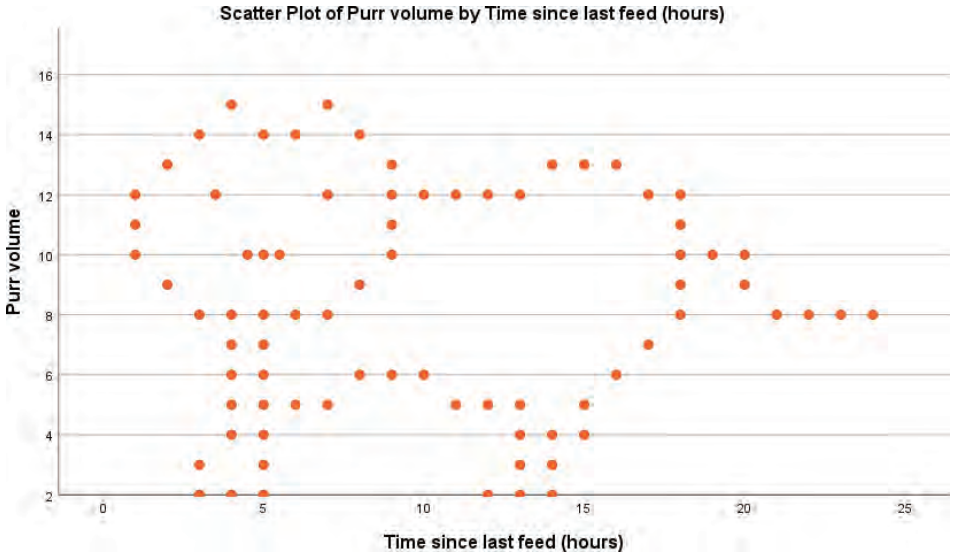


Figure 5.36 A catterplot



15 I'm hugely grateful to Lea Raemaekers for sending me these data.

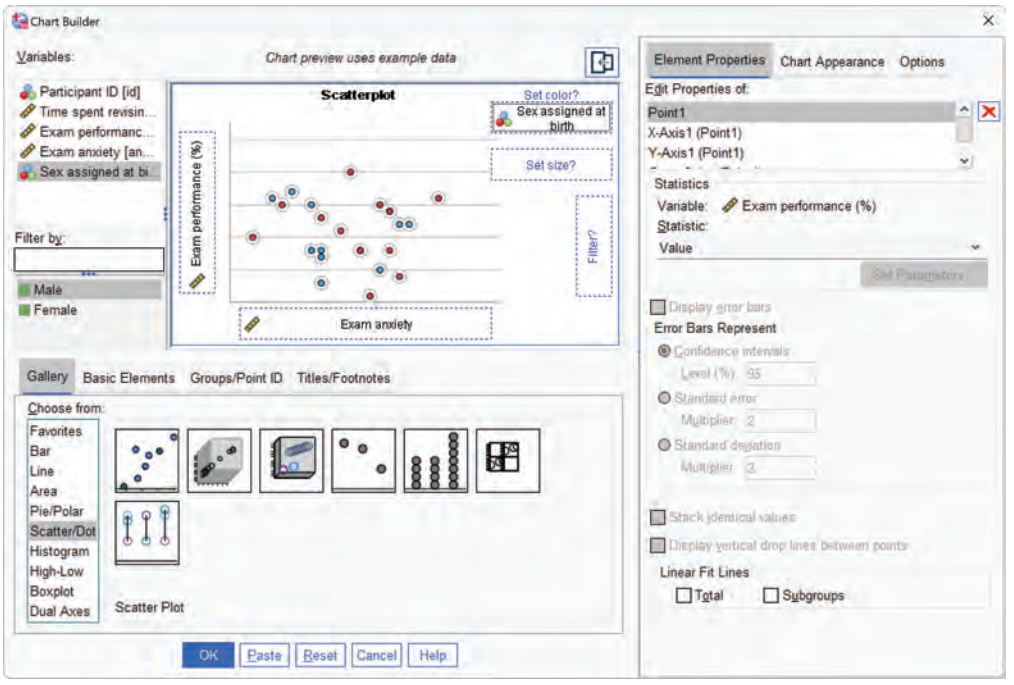


Figure 5.37 Completed *Chart Builder* dialog box for a grouped scatterplot

Figure 5.38 shows the resulting scatterplot; as before, I have added regression lines, but this time I have added different lines for each group. We saw in the previous section that plots can be edited using the Chart Editor (Figure 5.35), and that we could fit a regression line that summarized the whole data set by clicking on . We could do this again, but having split the data by sex it might be more useful to fit separate lines for our two groups. To this, double-click the plot to open the Chart Editor, then click to open the *Properties* dialog box (Figure 5.35), and again select Linear. Click Apply, and close the Chart Editor window to return to the viewer. Note that SPSS has plotted separate lines for the men and women (Figure 5.38). These lines tell us that the relationship between exam anxiety and exam performance was slightly stronger in males (the line is steeper), indicating that men's exam performance was more adversely affected by anxiety than women's exam anxiety. (Whether this difference is significant is another issue – see Section 8.6.1.)

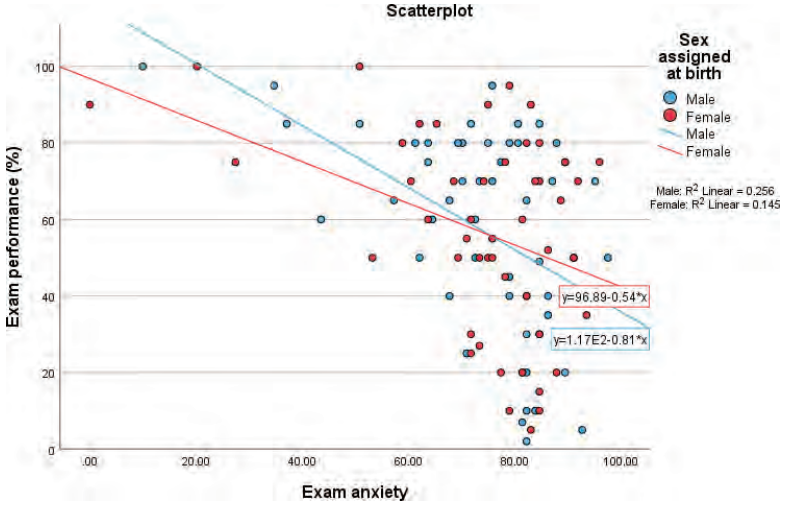


Figure 5.38 Scatterplot of exam anxiety and exam performance split by sex



Oliver Twisted
Please, Sir, can I have some more ... dimensions?

'I need to discover how to bend space and time so that I can escape from Dickensian London and enter the twenty-first century, where when you pick a pocket or two you get an iPhone rather than a snotty hanky. To do this I need extra dimensions – preferably fourth ones,' says Oliver. At present SPSS won't let you manipulate the space-time continuum, but it will let you add an extra dimension to a scatterplot. To find out how look at the additional material.

5.8.3 Simple and grouped 3-D scatterplots Ⓐ

One of the few times you can use a 3-D plot without a statistician locking you in a room and whipping you with a broom is a scatterplot. A 3-D scatterplot displays the relationship between three variables, and the reason why it's sometimes OK to use a 3-D plot in this context is that the third dimension tells us something useful (it isn't there to look pretty). As an example, imagine our researcher decided that exam anxiety might not be the only factor contributing to exam performance. So, she also asked participants to keep a revision diary from which she calculated the number of hours spent revising for the exam. She might want to look at the relationships between these variables simultaneously and she could do this using a 3-D scatterplot. Personally, I don't think a 3-D scatterplot is a clear way to present data – a matrix scatterplot is better – but if you want to do one see Oliver Twisted.

5.8.4 Matrix scatterplot Ⓐ

Instead of plotting several variables on the same axes on a 3-D scatterplot (which can be difficult to interpret), I think it's better to plot a matrix of 2-D scatterplots. This type of plot allows you to see the relationship between all combinations of many different pairs of variables. Let's continue with the example of the relationships between exam performance, exam anxiety and time spent revising. First, access the Chart Builder and double-click the icon for a scatterplot matrix (Figure 5.32). A different type of plot than you'll have seen before will appear on the canvas, and it has only one drop zone (Scattermatrix?). We need to drag all the variables that we want to see plotted against each other into this drop zone. We have dragged multiple variables into a drop zone in previous sections. To recap, select the first variable (**Time spent revising**), then hold down the *Ctrl* key (*Cmd* on a Mac) and click the

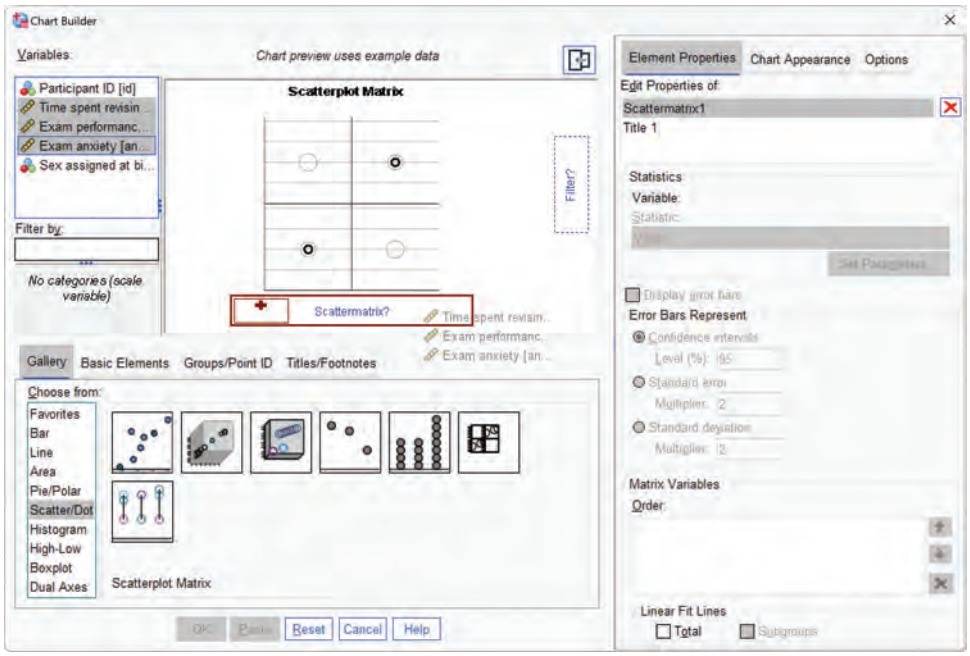


Figure 5.39 *Chart Builder* dialog box for a matrix scatterplot

other two variables (**Exam performance (%)** and **Exam anxiety**).[16] The three variables should now be highlighted and can be dragged into Scattermatrix? as shown in Figure 5.39 (you need to keep *Cmd* held down as you drag on a Mac). Click OK to produce the plot.

The six scatterplots in Figure 5.40 represent the various combinations of each variable plotted against each other variable. Using the grid references to help locate specific plots, we have:

- **A2**: revision time (*Y*) against exam performance (*X*)
- **A3**: revision time (*Y*) against anxiety (*X*)
- **B1**: exam performance (*Y*) against revision time (*X*)
- **B3**: exam performance (*Y*) against anxiety (*X*)
- **C1**: anxiety (*Y*) against revision time (*X*)
- **C2**: anxiety (*Y*) against exam performance (*X*)

Notice that the three scatterplots below the diagonal of the matrix are the same plots as the ones above the diagonal, but with the axes reversed. From this matrix we can see that revision time and anxiety are inversely related (i.e., the more time spent revising, the less exam anxiety the person experienced). Also, in the scatterplot of revision time against anxiety (grids A3 and C1) it looks as though there is one possible unusual case – a single participant who spent very little time revising yet suffered very little anxiety about the exam. Because all participants who had low anxiety scored highly on the exam (grid C2), we can deduce that this person probably also did well on the exam (it was probably Smart Alex). We could examine this case more closely if we believed that their behaviour was caused by some external factor (such as a dose of *antiSTATic*).[17] Matrix scatterplots are very convenient for examining pairs of relationships between variables (see SPSS Tip 5.4). However, they can become very confusing indeed if you plot them for more than about three or four variables.

16 We could also have clicked on **Time spent revising**, then held down the *Shift* key and clicked on **Exam anxiety**.

17 If this joke isn't funny it's because you didn't pay attention to Chapter 3, or it's just not funny.

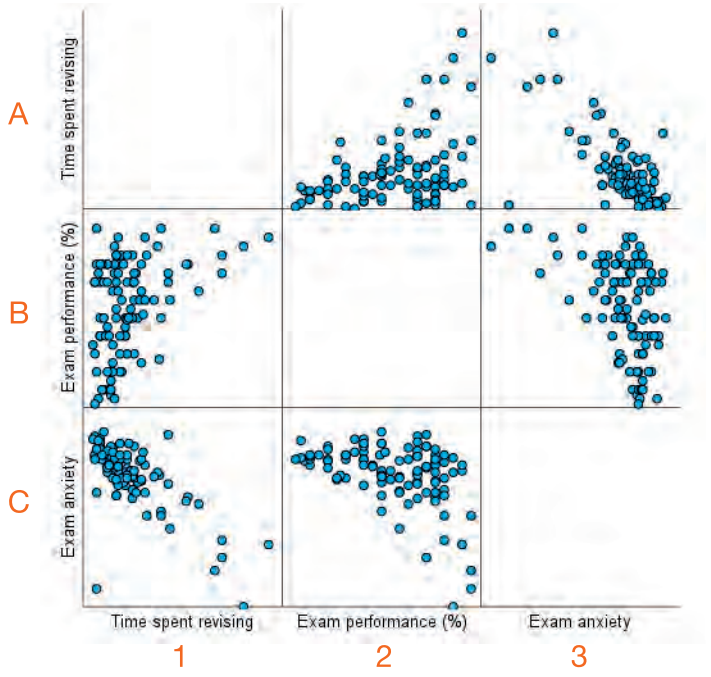
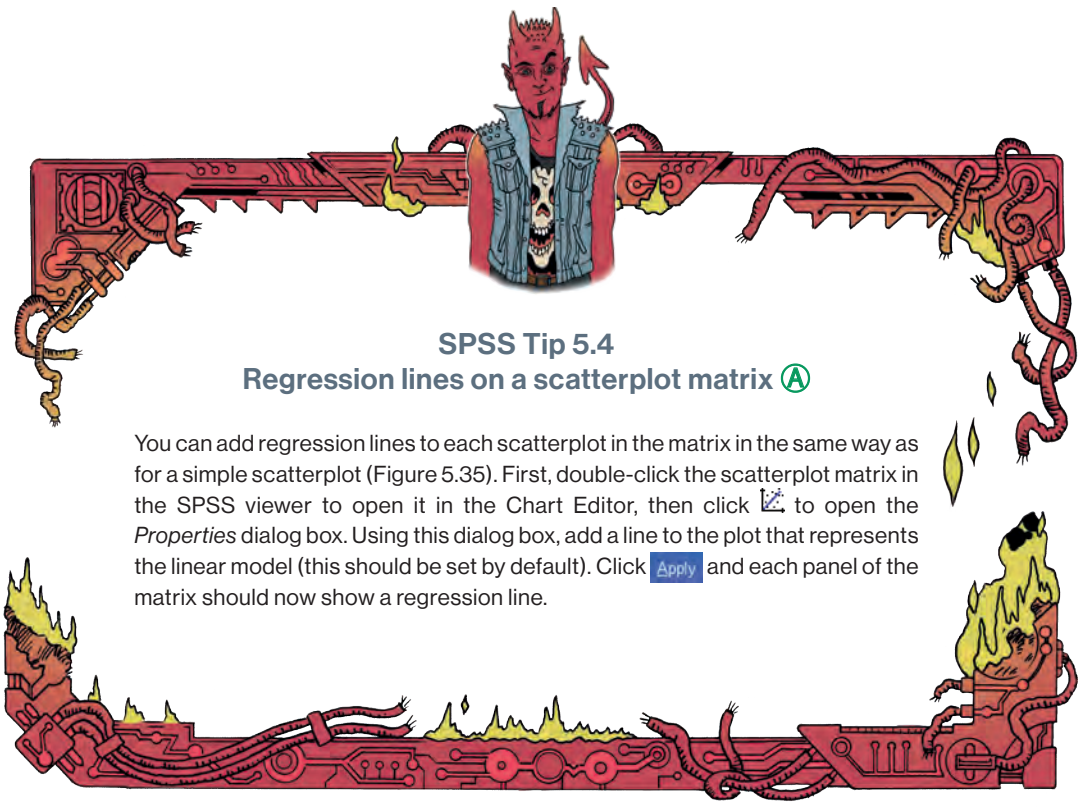


Figure 5.40 Matrix scatterplot of exam performance, exam anxiety and revision time. Grid references have been added for clarity.



SPSS Tip 5.4
Regression lines on a scatterplot matrix Ⓐ

You can add regression lines to each scatterplot in the matrix in the same way as for a simple scatterplot (Figure 5.35). First, double-click the scatterplot matrix in the SPSS viewer to open it in the Chart Editor, then click to open the *Properties* dialog box. Using this dialog box, add a line to the plot that represents the linear model (this should be set by default). Click Apply and each panel of the matrix should now show a regression line.

5.8.5 Simple dot plot or density plot Ⓐ

I mentioned earlier that the simple dot plot, or density plot as it is also known, is a histogram except that each data point is plotted (rather than using a single summary bar to show each frequency). Like a histogram, the data are still placed into bins (SPSS Tip 5.2), but a dot is used to represent each data point. You should be able to follow the instructions for a histogram to draw one.



Doing a simple dot plot in the Chart Builder is quite similar to drawing a histogram. Reload the **jiminy_cricket.sav** data and see if you can produce a simple dot plot of the success scores after the intervention. Compare the resulting plot to the earlier histogram of the same data (Figure 5.11). Remember that your starting point is to double-click the icon for a simple dot plot in the Chart Builder (Figure 5.32), then use the instructions for plotting a histogram (Section 5.4) – there is guidance on the companion website.

5.8.6 Drop-line plot Ⓐ

I also mentioned earlier that the drop-line plot is fairly similar to a clustered bar chart (or line chart), except that each mean is represented by a dot (rather than a bar), and within groups these dots are linked by a line (contrast this with a line plot, where dots are joined across groups, rather than within groups). The best way to see the difference is to plot one, and to do this you can apply what you were told about clustered line plots (Section 5.6.2) to this new situation.



Doing a drop-line plot in the Chart Builder is quite similar to drawing a clustered bar chart. Reload the **notebook.sav** data and see if you can produce a drop-line plot of the arousal scores. Compare the resulting plot to the earlier clustered bar chart of the same data (Figure 5.21). The instructions in Section 5.6.2 should help. Now see if you can produce a drop-line plot of the **social_media.sav** data from earlier in this chapter. Compare the resulting plot to the earlier clustered bar chart of the same data (Figure 5.30). The instructions in Section 5.6.5 should help.

Remember that your starting point for both tasks is to double-click the icon for a drop-line plot in the Chart Builder (Figure 5.32). There is full guidance for both examples in the additional material on the companion website.

5.9 Editing plots Ⓐ

We have already seen how to add regression lines to scatterplots using the Chart Editor (Section 5.8.1). We'll now look in more detail at the Chart Editor window. Remember that to open this window you double-click the plot you want to edit in the viewer window (Figure 5.35). You can edit almost every aspect of the plot: in the Chart Editor you can click virtually anything that you want to change and change it. Once in the Chart Editor (Figure 5.41) there are several icons that you can click to change aspects of the plot. Whether a particular icon is active depends on the

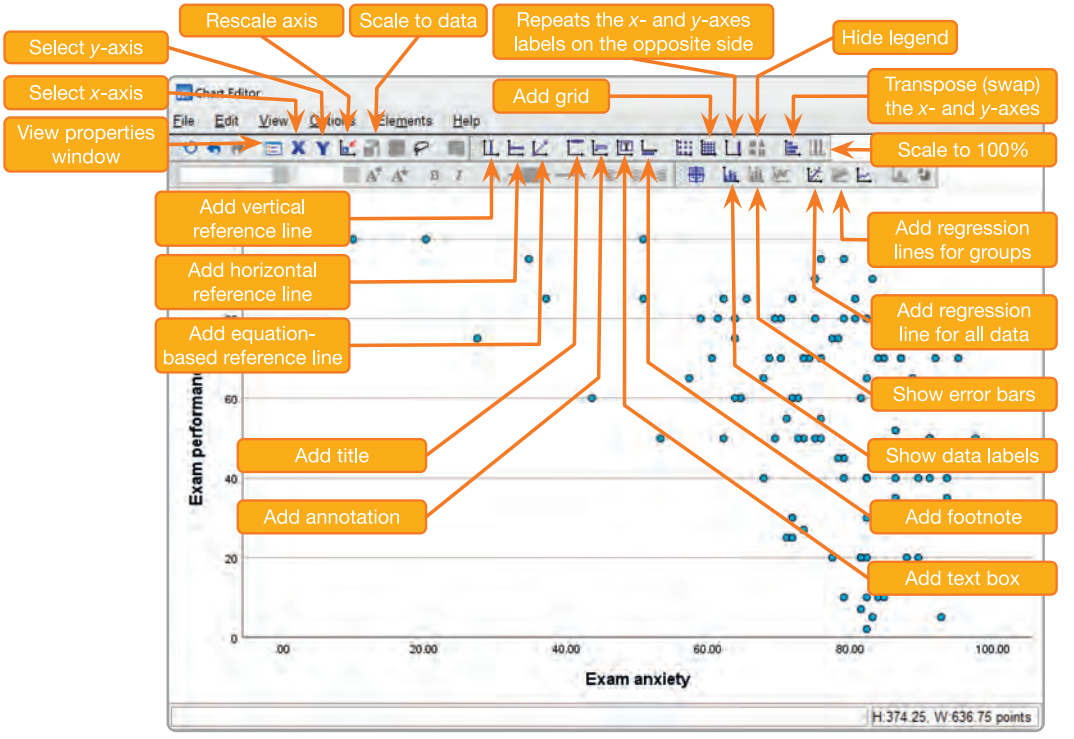


Figure 5.41 The Chart Editor

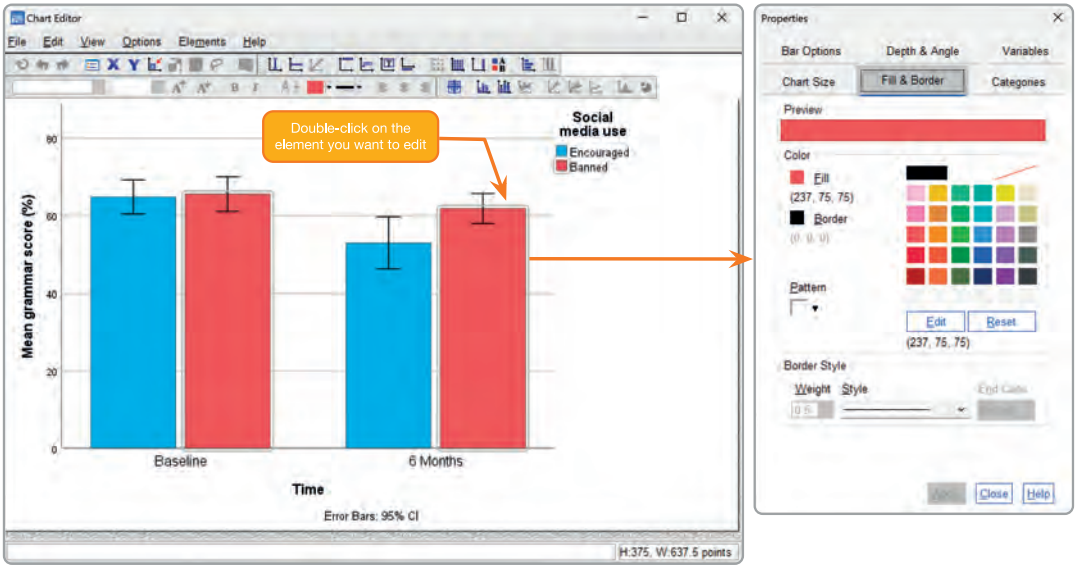


Figure 5.42 To select an element in the plot simply double-click it and its *Properties* dialog box will appear

type of chart that you are editing (e.g., the icon to fit a regression line will not be active for a bar chart). The figure tells you what some of the icons do, but most of them are self-explanatory. Play around and you'll find icons that add elements to the plot (such as grid lines, regression lines, data labels).



Oliver Twisted
Please, Sir, can I have some more ... plots?

'Blue and green should never be seen!', shrieks Oliver with so much force that his throat starts to hurt. 'This plot offends my delicate artistic sensibilities. It *must* be changed immediately!' Never fear, Oliver. Using the editing functions in SPSS, it's possible to create some very tasteful plots. These facilities are so extensive that I could probably write a whole book on them. In the interests of saving trees, I have prepared an online tutorial. We look at an example of how to edit an error bar chart to make it conform to some of the guidelines that I talked about at the beginning of this chapter. In doing so we will look at how to edit the axes, add grid lines, change the bar colours, change the background and borders. It's a very extensive tutorial.



Oditi's Lantern
Editing plots

'I, Oditi, have been dazzled and confused by the pinkness of many a plot. Those who seek to prevent our worthy mission do bedazzle us with their pink and lime green monstrosities. These colours burn our retinas until we can no longer see the data within the sacred drawings of truth. To complete our mission to find the secret of life we must make the sacred drawings palatable to the human eye. Stare into my lantern to find out how.'



You can also edit parts of the plot by selecting them and changing their properties using the *Properties* dialog box. To select part of the plot double-click it; it will become highlighted and a new dialog box will appear (Figure 5.42). This *Properties* dialog box enables you to change virtually anything about the item that you have selected. You can change the bar colours, the axis titles, the scale of each axis, and so on. You can also do things like make the bars three-dimensional and pink, but we know better than to do things like that. There are both written (see Oliver Twisted) and video (see Oditi's Lantern) tutorials on the companion website.

5.10 Brian and Jane's story Ⓐ

During their brief encounter in the Euphoria cluster on campus, Brian had noticed some pictures on Jane's screen. He knew that these pictures were plots – he wasn't that stupid – but he didn't understand what they showed, or how to create them. He began to wonder whether he could create one for Jane, sort of like painting her a picture with numbers. She had seemed entranced by the minimalist, stylish images on her screen. 'If she was enthralled by those plain images,' Brian considered, 'imagine how impressed she'd be by some 3-D effects.' He made a mental note that she'd probably love it if he coloured the bars pink.

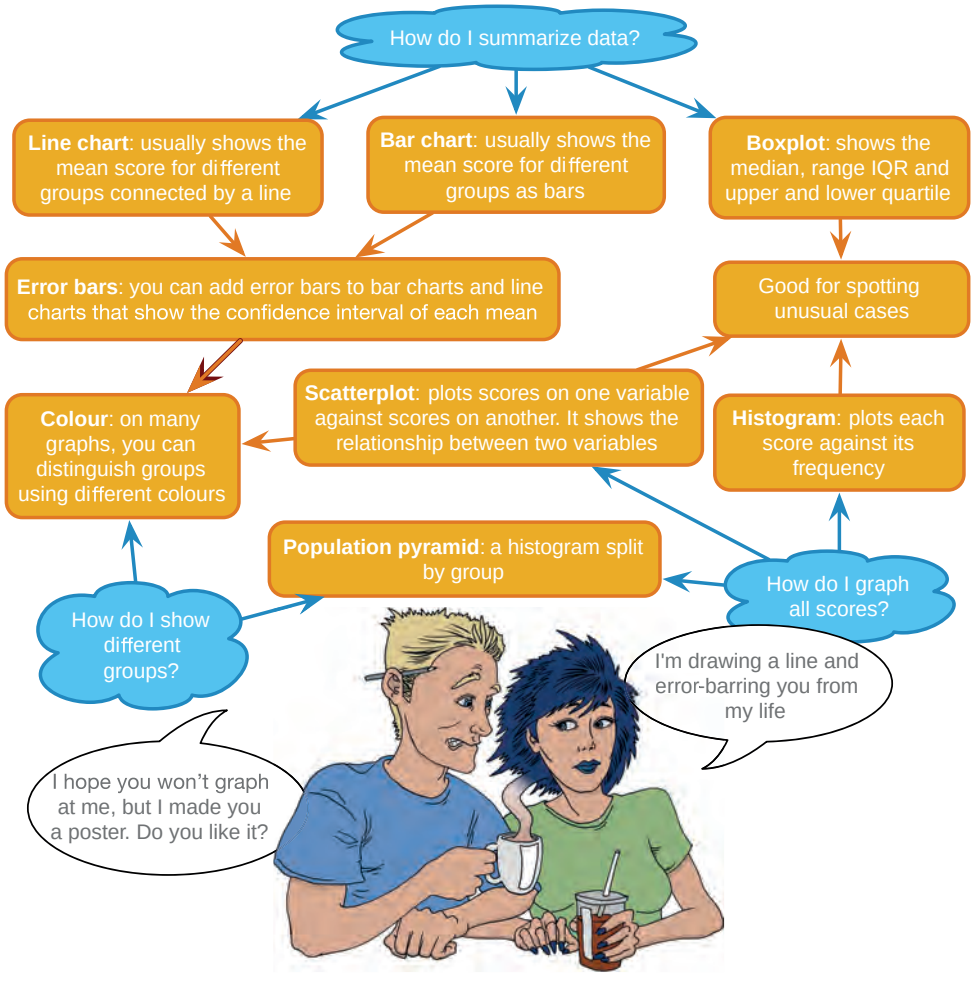


Figure 5.43 What Brian learnt from this chapter

Across town Jane wandered the long corridors of a labyrinthine basement at the University. The walls were made from hundreds of jars. Each contained a brain suspended in lightly glowing green fluid. It was like an ossuary of minds. There was no natural light, but the glow of the jars created enough ambient light to see. Jane loved it down here. The jars were beautiful, elegant designs with brass bases and lids etched with green lines that made an elaborate external circuit board peppered with glistening lights. Jane wondered whether anyone had noticed that she had cracked the electronic lock, or that she spent so many evenings down here? Her parents and teachers had never appreciated how clever she was. At school they'd mocked her 'human flatworm' theory of learning. She would prove them wrong. She stopped and turned to a jar labelled 'Florence'. She took a small electronic device from her pocket and pressed it against the jar. With a quiet hum the jar moved forward from the wall, and the lid slowly opened. Jane took a knife and fork from her pocket.

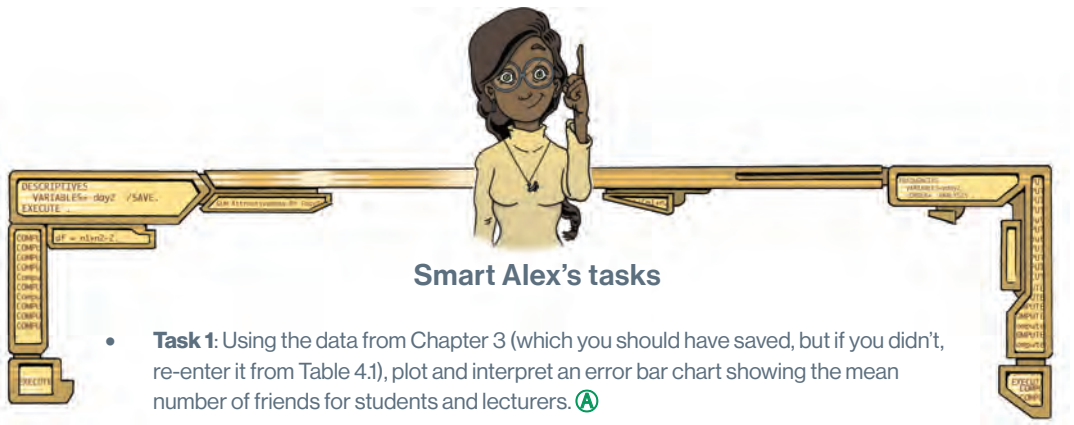
Brian admired his poster. It was the most magnificent pink 3-D plot he had ever seen. He turned it to show the dude behind the counter at the printers. The guy smiled at Brian; it was a patronizing smile. Brian rushed to campus, poster tube under his arm. He had a spring in his step, and a plan in his mind.

5.11 What next? Ⓐ

We have discovered that when it comes to plots, minimal is best: no pink, no 3-D effects, no pictures of Errol your pet ferret superimposed on the data – oh, and did I mention no pink? Plots are a useful way to visualize life. At the age of about 5 I was trying to visualize my future and my favoured career choice at that age was to become a famous sports person. Like many a UK-born child, my sport of choice was football (or soccer as people outside of the UK sometimes like to call it to avoid confusion with a game in which a ball is predominantly passed through the hands, and not the feet). In the next chapter we learn what became of my football career.

5.12 Key terms that I've discovered

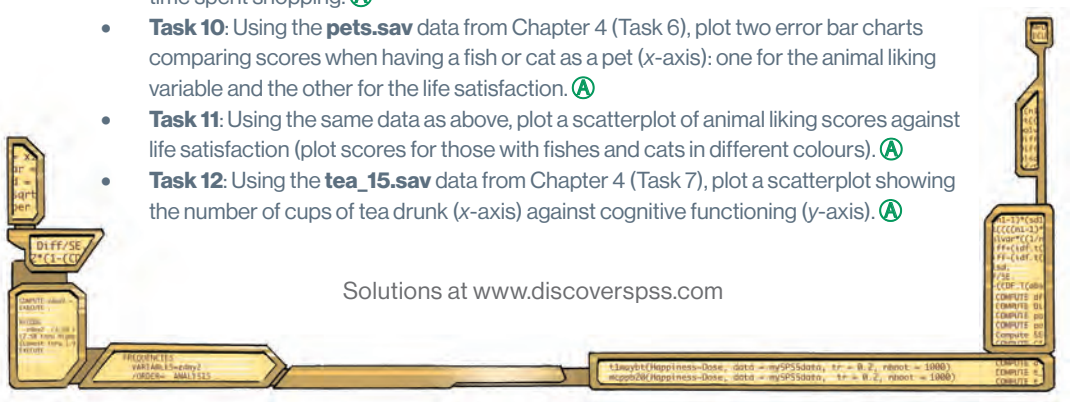
Bar chart
Boxplot (box–whisker diagram)
Chart Builder
Chart Editor
Chartjunk
Density plot
Error bar chart
Line chart
Regression line
Scatterplot



Smart Alex's tasks

- **Task 1**: Using the data from Chapter 3 (which you should have saved, but if you didn't, re-enter it from Table 4.1), plot and interpret an error bar chart showing the mean number of friends for students and lecturers. Ⓐ

- **Task 2**: Using the same data, plot and interpret an error bar chart showing the mean alcohol consumption for students and lecturers. Ⓐ
- **Task 3**: Using the same data, plot and interpret an error line chart showing the mean income for students and lecturers. Ⓐ
- **Task 4**: Using the same data, plot and interpret error a line chart showing the mean neuroticism for students and lecturers. Ⓐ
- **Task 5**: Using the same data, plot and interpret a scatterplot with regression lines of alcohol consumption and neuroticism grouped by lecturer/student. Ⓐ
- **Task 6**: Using the same data, plot and interpret a scatterplot matrix with regression lines of alcohol consumption, neuroticism and number of friends. Ⓐ
- **Task 7**: Using the **zhang_sample.sav** data from Chapter 4 (Task 8) plot a clustered error bar chart of the mean test accuracy as a function of the type of name participants completed the test under (*x*-axis) and whether they identified as male or female (different coloured bars). Ⓐ
- **Task 8**: Using the **teach_method.sav** data from Chapter 4 (Task 3), plot a clustered error line chart of the mean score when electric shocks were used compared to being nice, and plot males and females as different coloured lines. Ⓐ
- **Task 9**: Using the **shopping.sav** data from Chapter 4 (Task 5), plot two error bar charts comparing men and women (*x*-axis): one for the distance walked, and the other for the time spent shopping. Ⓐ
- **Task 10**: Using the **pets.sav** data from Chapter 4 (Task 6), plot two error bar charts comparing scores when having a fish or cat as a pet (*x*-axis): one for the animal liking variable and the other for the life satisfaction. Ⓐ
- **Task 11**: Using the same data as above, plot a scatterplot of animal liking scores against life satisfaction (plot scores for those with fishes and cats in different colours). Ⓐ
- **Task 12**: Using the **tea_15.sav** data from Chapter 4 (Task 7), plot a scatterplot showing the number of cups of tea drunk (*x*-axis) against cognitive functioning (*y*-axis). Ⓐ



Solutions at www.discoverspss.com



THE BEAST OF BIAS

6.1 What will this chapter tell me? 244
6.2 Descent into statistics hell 245
6.3 What is bias? 265
6.4 Outliers 266
6.5 Overview of assumptions 269
6.6 Linearity and additivity 271
6.7 Spherical errors 271
6.8 Normally distributed something or other 275
6.9 Checking for bias and describing data 279
6.10 Reducing bias with robust methods 303
6.11 A final note 316
6.12 Jane and Brian's story 317
6.13 What next? 317
6.14 Key terms that I've discovered 319
Smart Alex's tasks 319

Letters A to D indicate increasing difficulty

6.1 What will this chapter tell me? Ⓐ

My first career choice was to become a soccer star. My granddad (Harry) had been something of a local soccer hero in his day and I wanted nothing more than to emulate him. Harry had a huge influence on me: he had been a goalkeeper, and consequently I became a goalkeeper too. The model I fitted to my life data predicted great goalkeeping ability from natural ability (based on the goalkeeping genes I assumed I'd have from my granddad). Unfortunately, this prediction had error. Quite a bit of error. I was quite short for my age, which meant that I got overlooked to play in goal for my school in favour of a taller boy. I think I was the better goalkeeper technically, but I did have an Achilles heel fatal to my goalkeeping career: the opposition could lob the ball over my head. Instead of goal, I typically got played at left back ('left back in the changing room', as the joke used to go) because, despite being right-footed, I could kick with my left one too. The trouble was, having spent years learning my granddad's goalkeeping skills, I didn't have a clue what a left back was supposed to do.[1] Consequently, I didn't shine in the role, and for many years that put an end to my belief that I could play soccer. This example shows that a highly influential thing (like your granddad) can bias the conclusions you from your model. The same thing happens in data analysis: sources of influence and bias lurk within the data, and unless we identify and correct for them we'll end up becoming goalkeepers despite being too short for the job. Or something like that. This chapter introduces the linear model and the sources of bias within it.



Figure 6.1 My first failed career choice was to become soccer star

1 In the 1970s at primary school, 'teaching' soccer involved shoving 11 boys onto a pitch and watching them chase the ball. It didn't occur to teachers to develop your technique, tactical acumen, or even tell you the rules.

6.2 Descent into statistics hell ©

This chapter is all about things that bias linear models. People fit linear models for different reasons: some want to estimate its parameters, others want to use those parameter estimates to test hypotheses, and some people's focus will be on both of those things. When we think about bias we need to think about our purpose in fitting the model. If we've fitted a model to estimate parameters then we care about things that bias parameter estimates and their confidence intervals, but if we want to use our parameter estimates to test hypotheses then we also need to care about things that bias significance tests. To understand exactly what things bias which parts of the model requires a fairly unpleasant detour into the underbelly of the linear model. This section is that detour. I do think it's worth trying to get to grips with a more technical understanding of what's going on, but you can't really do that without deep-diving into matrix algebra, which won't be pleasant for either of us. However, with a few grams of matrices and a pinch of taking liberties with definitions a beautiful cake of understanding will emerge. Or not. Although I encourage you read this section so that you share the pain I felt in writing it, if maths isn't your thing you can skip this section and still make sense of what's to come. Later sections attempt to give you a non-mathematical guide to what things matter and how to detect them. See you on the other side.

6.2.1 Revision of the linear model ©

We saw in Sections 2.5 and 2.6 that having collected data, we fit a model representing the hypothesis that we want to test. A common model is the linear model, which in a simple form looks like this:

$$\text{outcome}_i = b_0 + b_1 X_i + error_i \tag{6.1}$$

More formally, we can denote the outcome with the letter Y, and error with the Greek letter epsilon (ε):

$$Y_i = b_0 + b_1 X_i + \varepsilon_i \tag{6.2}$$

This model can be summed up as 'we want to predict the values of an outcome variable (Y_i) from a predictor variable (X_i)'. If we want to predict an outcome variable from more than one predictor the model expands to:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni} + \varepsilon_i \tag{6.3}$$

In this model, we're predicting the value of the outcome for a particular entity (i) from the value of the outcome when the predictors have a value of zero (b_0) and the entity's score on each of the predictor variables (X_{1i}, X_{2i}, up to and including the last predictor, X_{ni}). Each predictor variable has a parameter (b_1, b_2, b_n) attached to it, which tells us something about the relationship between that predictor and the outcome.

Let's look at an example with a single predictor. Imagine that we're interested in predicting hearing loss (having seen well over 1000 bands in concert, I *am* concerned about this). We want to predict the number of minutes a person experiences ringing in their ears after a gig (Y_i) from how loud the gig was (X_i). We can represent this model as:

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{Ringing}_i &= b_0 + b_1 \text{Volume}_i + \varepsilon_i \end{aligned} \tag{6.4}$$

If we know the values of the parameters, then we can estimate the time a person's ears will ring given the volume of a concert. In particular, if we know the value of b_1 then we'll know something about

the strength and direction of the relationship between concert volume and how long a person might experience ringing in their ears (which we might be using as a proxy for hearing damage). If we know b_0 then we'll know for how long your ears will ring if you went to a concert that had a volume of 0 dB (that'd be a weird gig).

We also saw in Section 2.7 that the values of the parameters (i.e., the value(s) of b) have to be estimated from data that we have collected in a sample. Recapping Jane Superbrain Box 2.4, when referring to parameter estimates we use hats to remind us that they are estimates, so we might write:

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + e_i \\ \text{Ringing}_i &= \hat{b}_0 + \hat{b}_1 \text{Volume}_i + e_i \end{aligned} \tag{6.5}$$

Similarly when we predict the value of the outcome (in this case ringing), this predicted value is an estimate and we likewise give it a hat. Predicted values also have no error because they are not observed values, so there is no error term when referring to predicted values:

$$\begin{aligned} \hat{Y}_i &= b_0 + b_1 X_i \\ \widehat{\text{Ringing}}_i &= \hat{b}_0 + \hat{b}_1 \text{Volume}_i \end{aligned} \tag{6.6}$$

Having had a brief recap of what we learnt in Chapter 1, we will explore the linear model in more depth.

6.2.2 Populations, samples and error in prediction ©

In an ideal world, we'd have access to data on gig volume and ear ringing from every person attending every concert ever. We could regard this as the population of scores. This situation is represented by the top plot in Figure 6.2 in which each dot shows the peak gig volume and the minutes of ringing for a given person. The red line is the population model (see Jane Superbrain Box 2.2). It is described by a version of Eq. 6.4, in which the bs are replaced with their population values:

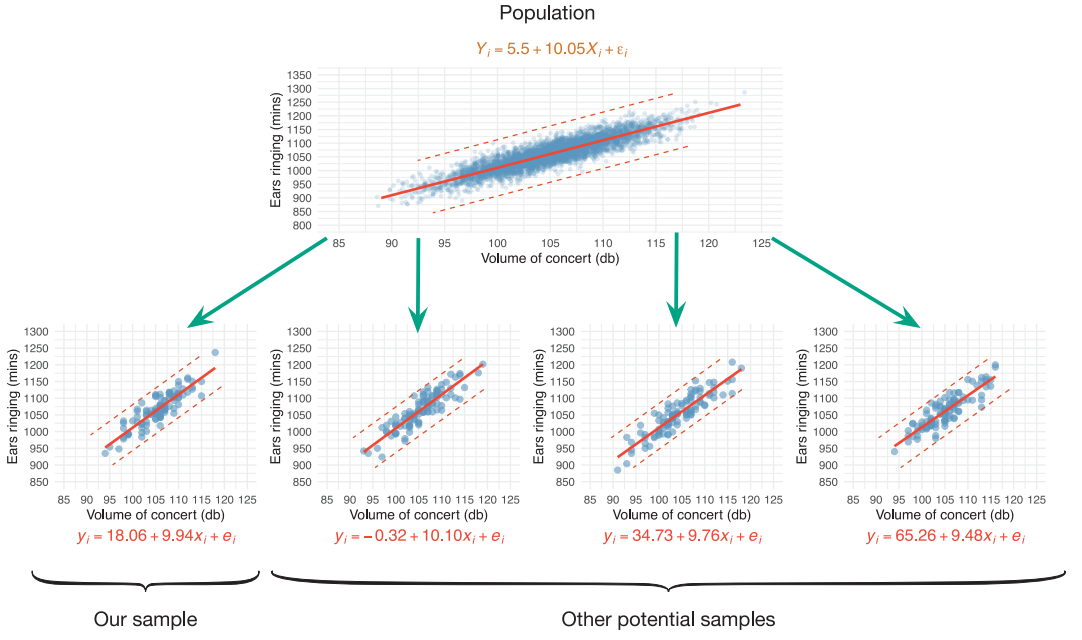


Figure 6.2 We use samples to estimate the values of parameters

$$\text{Ringing}_i = 5.5 + 10.05\text{Volume}_i + \varepsilon_i \quad (6.7)$$

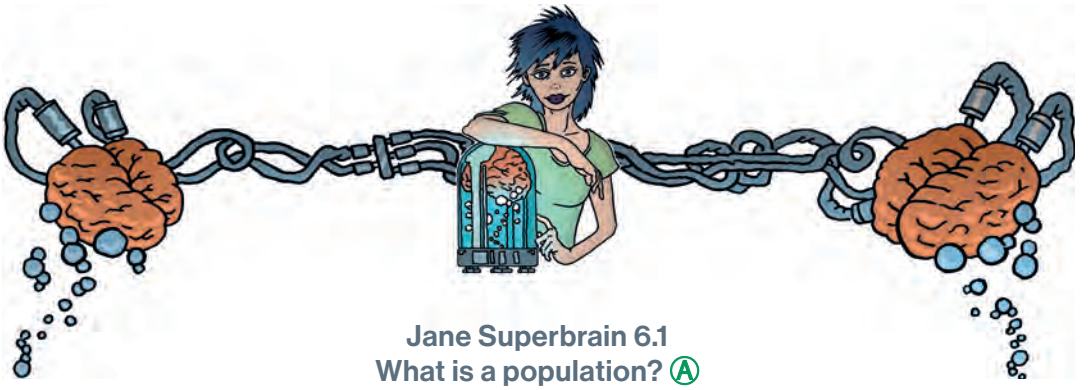
The model (red line) shows a positive relationship: the louder the concert, the longer your ears ring. We could regard this model as representing the 'true relationship' between gig volume and hearing damage. We'd like to know what this 'true relationship' is. Unfortunately, we rarely, if ever, have access to the population data. This is where sampling comes in, which we discussed in Chapter 1.

Imagine we take 100 people who plan to attend different concerts. We measure how loud the concert is, and for how long they subsequently experienced ringing in their ears. We can use these scores to estimate the parameters (the *b*s) in the model. In other words, we use scores that we can observe to estimate the true relationship between concert volume and ear ringing, which we can't observe. Figure 6.2 illustrates this idea of using samples to estimate things in a population. The bottom of Figure 6.2 shows four different random samples of people from the population. One of the images shows the sample we actually have, the other three show some (but not all) of the different random samples that we might have taken but didn't. There are an infinite number of samples we could have taken from the population, but we took only one. We're hoping that the one we took is representative of the population, which will be the case if samples from this population are, in general, representative. A few key points that this figure illustrates:

- **The form of the model is similar in the samples to the population**. In all samples the general form of the model is the same as both the population model and the other samples. Even with a random sample of 100 people, we get a pretty good idea of the 'true relationship' between gig volume and hearing damage. The relationship is consistent across samples: the red line slopes upward from the horizontal at a similar angle.
- **Parameter estimates vary across samples**. The sample models are not perfect representations of the population model. The population model is defined by $b_0 = 5.5$ and $b_1 = 10.05$, but none of the samples have values that match these exactly. For example, the values of b_1 in the samples range from 9.48 to 10.10 but none of them are 10.05, the population value. Put another way, there is sampling variation: estimates of the model parameters (the *b*s) vary from sample to sample, but *on average should reflect the true population value*. The *b*s, therefore, have standard errors associated with them that tell us the degree to which they vary from sample to sample (if you need to, revisit Chapter 1 to revise sampling variation).
- **The spread of scores around the population model is consistent**. The dotted red lines mark out the limits between which (hypothetical) data points around the population model fall. Note these lines are parallel to the model itself, showing that at all values of the predictor variable (all values along the *x*-axis), the spread of scores around the model is assumed to be similar. This fact is true also in the samples: that is, the data in the samples reflect the assumed pattern in the population. We'll return to this point in Section 6.7.1.
- **The population model contains error**. The population model does not perfectly predict all cases in the population – if we could observe all cases in the population (which we can't) then there is a difference between what the population model predicts and the value of the outcome for a given member of the population. In this case, for a given volume (say 100 dB) the population model predicts a specific value of ear ringing, however, in reality when the volume is 100 dB there are a range of possible durations for ear ringing, so although on average the predicted value is correct, there is generally some difference between this predicted value and what might actually be observed.

6.2.3 Errors in models ©

At the end of the previous section, I mentioned that the population model contains error in prediction. That is, if the population data were observable and we use the model to predict the time a given individual's ears will ring for (the outcome) from the volume of the concert they attended (predictor)



Jane Superbrain 6.1
What is a population? Ⓐ

In Jane Superbrain Box 2.2 we asked the question, 'What is a population?' As I mentioned, some people find it useful to think of the population as a non-mathematical entity, something that exists in the physical world. I think there is some value to talking about a population model as though it is 'what we would find if only we had collected data from the entire population'. Figure 6.2 is one of many examples where I conceptualize the population as though it is somehow possible to observe data from every entity. However, the 'population' is in fact a mathematical idea rather than a physical entity. To recap Jane Superbrain Box 2.2, when we use a sample to estimate something in the population what we're doing is testing (in the sample) whether it's reasonable to describe a process in a particular way. In this example, we're asking, 'Is it reasonable to model the relationship between volume and ringing in the ears with an equation of the form

$$\text{Ringing}_i = b_0 + b_1\text{Volume}_i + \varepsilon_i$$

and, if so, what are the values of b_0 and b_1 based on the data?' The 'population' is this equation. Therefore, in Figure 6.2 it would be somewhat more accurate to think of the population as in Figure 6.3, that is, it is the red line that is the population and in reality there are no actual data points!

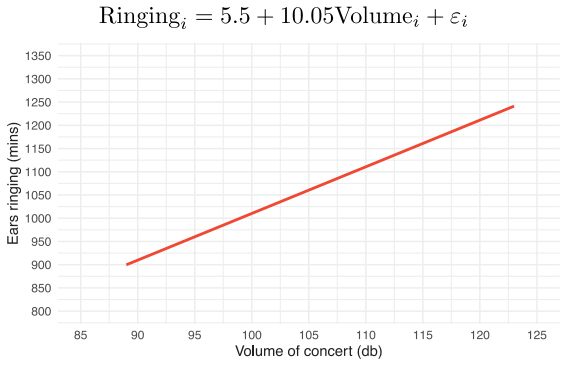


Figure 6.3 The population model



there may be a discrepancy between the value that the model predicts and the actual amount of time the person experienced ringing. This discrepancy is the error in prediction for that person. To illustrate, let's pretend that the population data are known (but see Jane Superbrain Box 6.1). Figure 6.4 (left) shows the population model (red line) and data, but I have zoomed in on part of it so that we can see the errors that I referred to at the end of the previous section. For a given observation (the blue circles), you can trace a vertical line to the model (the red diagonal line) to find the corresponding value of the outcome predicted by the model. In Figure 6.5 I have highlighted one case as a blue dot labelled *actual value*. Now, trace a vertical line down from this point and stop when you reach the model. I have marked this point with a red dot. This point is the predicted value of the case that I have highlighted. Because these points represent the same case, they have the same value for the predictor variable (both are 119 dB) but differ in the value of the outcome variable. If we trace a horizontal line across from these points to the *y*-axis (the red dotted lines in the figure) we find that the actual amount of ear ringing for this case was 1250 minutes, but the amount predicted by the model is only 1200 minutes. In other words, there is error in prediction: the model *underestimates* the number of hours of ringing ears for this case by 50 minutes. This difference between the predicted value and an observed value is the error, denoted by ε_i in Eq. 2.3. The subscript i reminds us that there is an error associated with every (hypothetical) case. For some cases the error in prediction, ε_i, is large (such as the aforementioned case in Figure 6.5), but for other cases the error might be small or even zero, such as the case highlighted in green in Figure 6.5. In fact, on average, the error *will* be zero.

We've just discovered what we mean by an error, but we did so by pretending that the population data are known and that we can calculate actual distances between predicted and observed cases for each member of the population. However, the population data are *not* known. Therefore, we can't know the individual values of the population errors, but we can model their distribution. In doing so we're able to say something about how model errors (in general) will behave. We could choose to model them with any known distribution, but for most of the models in this book the errors, for each possible value of the predictor, are represented by a random variable (Jane Superbrain Box 6.2) that is normally distributed with a mean of 0 and a constant variance (σ^2). To reiterate, this is like saying 'we don't know the values of the individual errors, but we will treat them as though they are normally distributed with a mean of 0 and a constant variance'.

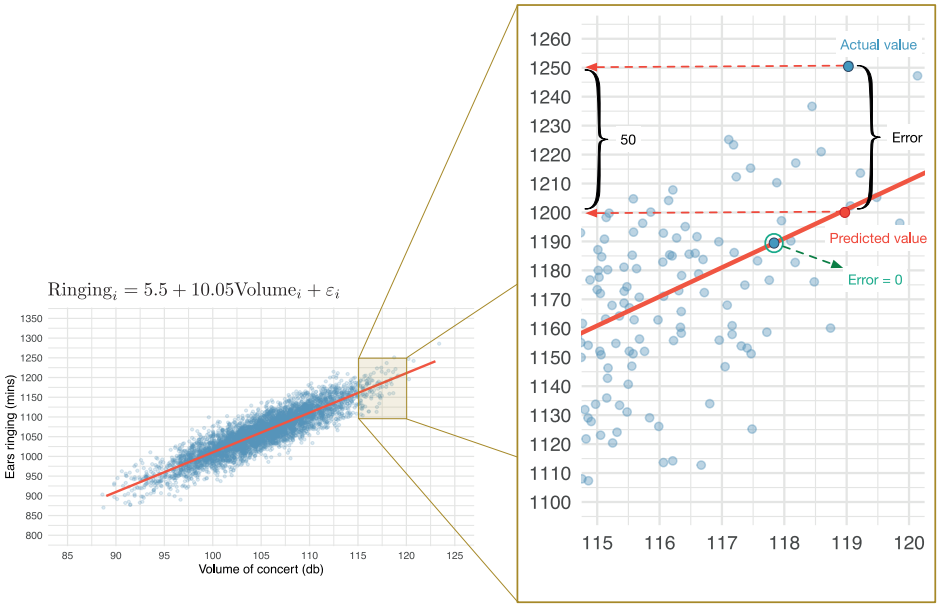


Figure 6.4 Error in prediction

Figure 6.5 illustrates this process. Figure 6.5 (left) shows the population model, and again I have included data points, but bear in mind that these are purely illustrative. The right-hand side of Figure 6.5 zooms in on a small part of the population model and simplifies it by ignoring all values of volume that are not whole numbers. For each value of volume (99, 100, 101, etc.) there is a predicted value from the model (red line) that is represented by the red dot. For each predicted value (red dot) the actual values of the outcome (if we could observe them) cluster above and below, indicating that sometimes the predicted value overestimates but other times it underestimates and, therefore, some errors will be positive and others negative. Also, for each predicted value (red dot) the pattern of observations is the same: most data points are very close to the predicted value and as we move further away from the predicted value we see fewer and fewer data points. This pattern tells us that most errors in prediction will be close to zero (because the predicted value of the outcome is close to the actual value) and as errors become larger in magnitude they occur with reducing frequency (the further from the predicted value we get, the fewer data points we see). The curves superimposed over the data points represent the assumed distribution of the errors and illustrate the points just made: for each value of the predictor, the errors are modelled with a normal distribution with a mean of zero (i.e., on average the error in prediction is zero) and a fixed variance (on the diagram the normal curves all have the same spread and look identical).

We've just discovered that when we fit a liner model, for each possible value of the predictor the errors are represented by a random variable that is normally distributed with a mean of 0 and a



Jane Superbrain 6.2
Random variables Ⓑ

The term 'random variable' is confusing because it's tempting to interpret it as 'a variable selected randomly', and that's not what it means. In statistics a random variable is a variable that quantifies a random process. For example, if you have ever played a national lottery or bingo you might be familiar with the idea of a bunch of balls spinning around a container and every so often one of them being spat out. Eventually, let's say, six balls are spat out of the machine. The process of selecting balls is random because there is no systematic mechanism that determines which ball is spat out of the machine. If (like in the lottery) we attach a number to each ball, then these numbers quantify this random process. The collection of numbers attached to the balls that were spat out are, therefore, a random variable because they are values that quantify a random process. The fact that model errors are represented by random variables means that the values of the errors are assumed to have been generated by a random process; that is, nothing systematic influences them.



constant variance (σ^2). It follows that across all values of the predictor, that is, if we were to pool the errors across the entire range of the model, the errors would also be normally distributed with a mean of 0 and variance, σ^2. Figure 6.6 illustrates this idea. and it is important when considering how we might test whether modelling errors as normally distributed is reasonable.

To sum up, then, any model we fit contains error in prediction for each case. So, when we write

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{Ringing}_i &= 5.5 + 10.05\text{Volume}_i + \varepsilon_i \end{aligned} \tag{6.8}$$

we're saying that we can predict the duration of ear ringing for case i (Ringing_i) from the volume of the concert that person attended (Volume_i) but there will be some error in that prediction (ε_i). To stress again, the population data (and associated errors) are *not* known, so instead we assume that the errors are independent random variables that are normally distributed with a mean of zero and a constant variance. We can make this assumption explicit by writing the linear model as

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{6.9}$$

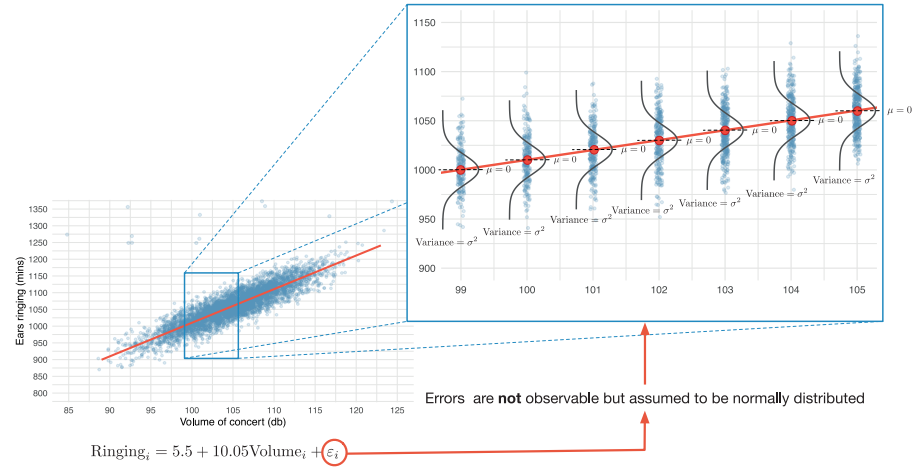


Figure 6.5 Errors in prediction are assumed to be normally distributed

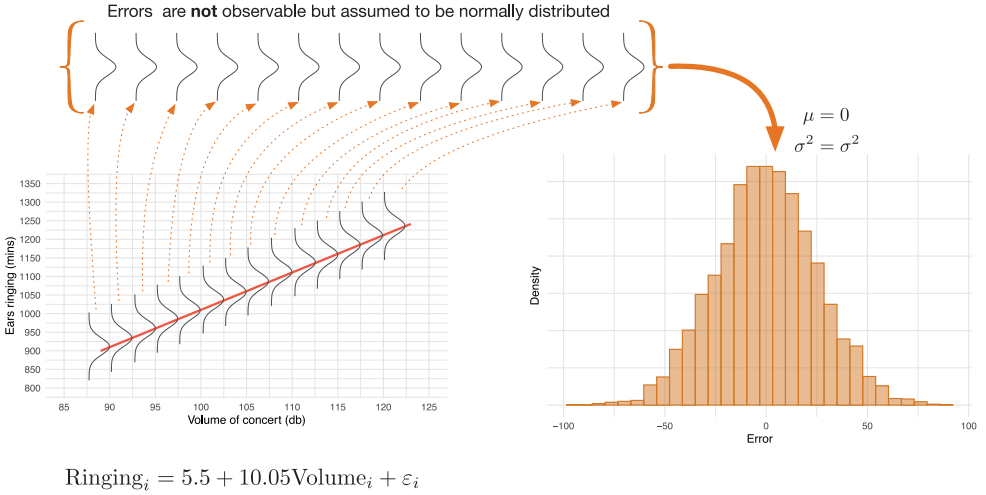


Figure 6.6 Errors are assumed to be normally distributed

The last line states explicitly that the model errors follow a normal distribution (that's what the $\mathcal{N}$ means) with a mean of 0 and a variance of σ^2. I tend to omit the final line when I specify models in this book, for reasons explained in Jane Superbrain Box 6.3.



Jane Superbrain 6.3
Different ways to express the linear model Ⓑ

Throughout this book I use equations for linear models that I think give you the best chance of having an intuitive grasp of what the models represent. An alternative way to represent linear models is to make the distribution of the errors explicit. For example, we can express a model with one predictor as

$$Y_i = b_0 + b_1X_i + \varepsilon_i$$
$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

The first line is the same as Eq. 6.10, but I've added a second line that explicitly states (in a mathsy way) that the model errors follow a normal distribution (that's what the $\mathcal{N}$ means) with a mean of 0 and a variance of σ^2. This second line is often omitted because when estimating the model using ordinary least squares it is not a requirement that errors are normally distributed (although it's nice if they are).

When using maximum likelihood estimation (rather than OLS) there *is* a requirement that the distribution of the outcome is known. So, we would *have* to express the model as above, although it is more common to use this mathematically equivalent form because it generalizes easily to other types of models:

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$
$$\mu_i = b_0 + b_1X_i$$

The first line says that the outcome, Y, is predicted from a normal distribution with expected value (i.e., mean) that is equal to μ_i (which is defined in the second row as the linear combination of predictors) and variance σ^2. In other words, the predicted values of Y are normally distributed with a mean determined by the linear combination of the predictors and constant variance. An equivalent, but more succinct, expression is

$$Y_i \sim \mathcal{N}(b_0 + b_1X_i, \sigma^2)$$

Although I have used examples with one predictor, with multiple predictors all that changes is that $b_0 + b_1X_i$ is replaced by a linear term involving more predictors, such as $b_0 + b_1X_{1i} + b_2X_{2i}$.

These forms of the linear model make explicit a difference between OLS and ML estimation: OLS does not *require* a particular distribution of the outcome or residuals, whereas to calculate the likelihood in ML an assumed distribution for the outcome *is* required. In most of the models in this book, the assumed distribution is normal, but as we shall see in Chapter 20 it can be some other distribution.



6.2.4 Errors vs. residuals ©

We've assumed that population errors are normally distributed, so it would be good to know whether this assumption is reasonable. How can we do this when population errors are unobservable? It turns out that we can use the sample data (which have been observed) because, as we saw in Figure 6.2, the sample model should have similar characteristics to the population model (more so in large samples). Like the population model, the model estimated from the sample will have error in prediction, usually denoted by e_i. Again the subscript reminds us that for every case there is error. We could write the model estimated from the sample as:

$$\begin{aligned} Y_i &= \hat{b}_0 + \hat{b}_1 X_i + e_i \\ \text{Ringing}_i &= 18.06 + 9.94\text{Volume}_i + e_i \end{aligned} \tag{6.10}$$

It is basically the same as the population model except that we have added some hats. Not everyone can pull off a hat, but *b*s can. The population *b*s mock the sample ones because they are merely estimates, but the sample ones don't care because they have badass hats. The hats remind us that we're dealing with *estimates*. So, the *b*s in the sample are not the population values but are *estimates* of the population values. The only other difference between Eq. 6.4 and Eq. 6.10 is that the error in prediction for each case is denoted with e_i rather than ε_i. This is the sort of difference that confuses people because both letters represent error in prediction. However, the different letters remind us that the error in prediction in the sample model is not the same as in the population model. They are conceptually similar though.

At this point, it's also useful to remind ourselves that we can also express the model estimated from the sample as:

$$\begin{aligned} \hat{Y}_i &= \hat{b}_0 + \hat{b}_1 X_i \\ \widehat{\text{Ringing}}_i &= 18.06 + 9.94\text{Volume}_i \end{aligned} \tag{6.11}$$

Compare this equation with Eq. 6.10. Note that the outcome ($\hat{Y}_i$) has had an outing to the local milliner and acquired a fetching hat, and the error term has vanished. WTF? The hat on the outcome means 'the predicted value of …'; again it reminds us that we are estimating something, in this case the value of the outcome. Fine, but why is there no error? Think back to Figure 6.5, which shows that for any predicted value (i.e. the values along the red line) the error associated with that predicted value is, on average, zero. Therefore, the predicted values are assumed to have no error and the error term drops out of the equation. I probably have a nasty habit of using the forms of the model in Eq. 6.10 and Eq. 6.11 somewhat interchangeably, switching without warning – sorry about that.

Given that the model estimated from the sample is supposed to be representative of the population model, it probably won't come as a surprise that the error in prediction in the population is conceptually similar to the error in prediction in the sample model. Figure 6.7 shows the sample of data from the bottom left of Figure 6.2. I have zoomed in on a section of the data so that we can see some data points (the blue dots) and the estimated model in the sample (the red line). Compare this figure with Figure 6.4. It's fairly similar, with the main difference being that the dots in Figure 6.4 are hypothetical values that we can't observe, whereas Figure 6.7 shows actual observed data (from the sample). For every observed case, we can draw a vertical line to the model that represents the difference between the observed value (the value of the outcome for the blue dot) and the predicted value (the value of the outcome for the corresponding point on the red line). The difference between the observed and predicted value of the outcome for each case is known as the **residual**. We can write this idea formally as:

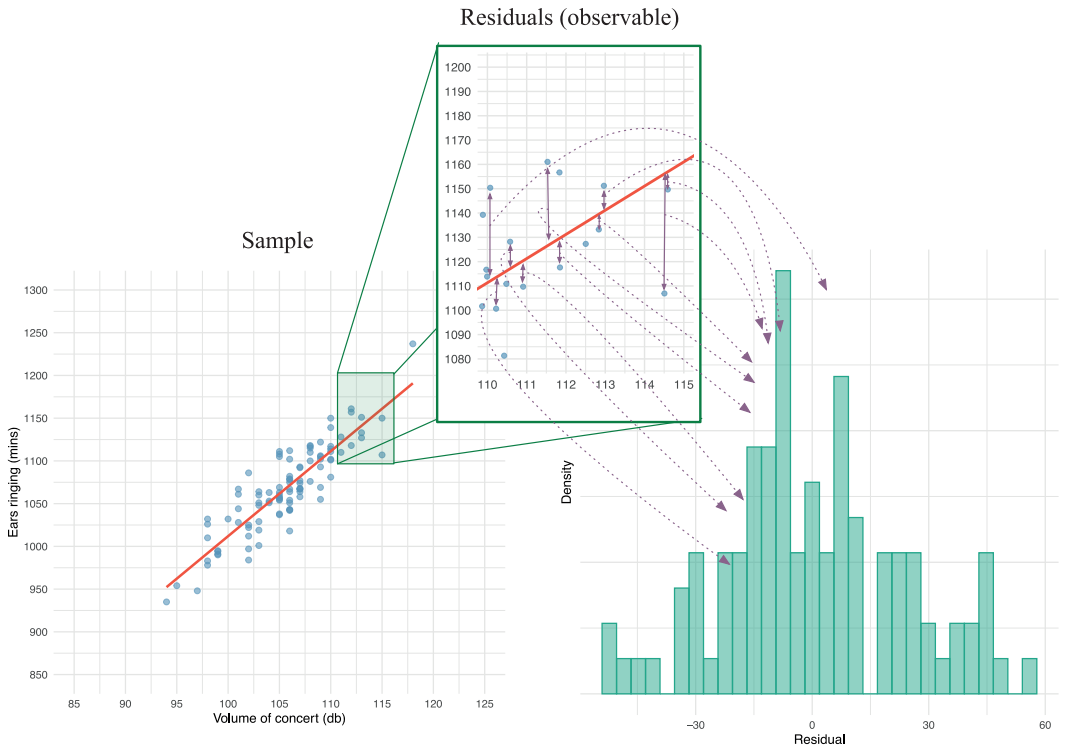


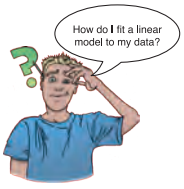
Figure 6.7 Sample residuals are observable and should follow a normal distribution centred on zero

$$\text{residual}_i = \text{observed}_i - \text{model}_i$$
$$e_i = Y_i - \hat{Y}_i \tag{6.12}$$

This equation is a formal way of saying that the residual for case i (denoted by e_i), is the predicted value of the outcome for that case ($\hat{Y}_i$) subtracted from the value of the outcome that was actually observed for that case ($\hat{Y}_i$). If you look back to Eq. 1.7 you'll see the equation is identical to that for the deviance – all that's changed is that in this context we use the word 'residual'.

If the sample model is what we use to estimate the population model, then it seems reasonable that the residuals from the sample model might be good proxies for the population errors. In fact, this is the case. If we plotted the frequency of the sample residuals we would hope to find a vaguely normal distribution centred on zero if the population errors are normally distributed and centred on zero. The spread (variance) of residuals is different from that in the population but as sample sizes increase they become increasingly similar. Put simply, we can use residuals (which we can observe) to infer things about errors in prediction for the population model (which we can't observe).

6.2.5 Estimation ©



Residuals also have a role to play in estimating the model parameters (the *b*s). When we use the sample data to estimate the values of the *b*s for the model, we typically use the method of least squares that was described in Section 2.7. We saw then that we could assess the fit of a model (the example we used was the mean) by looking at the 'deviations' between the model and the data collected. These 'deviations' were the vertical distances between what the model predicted and

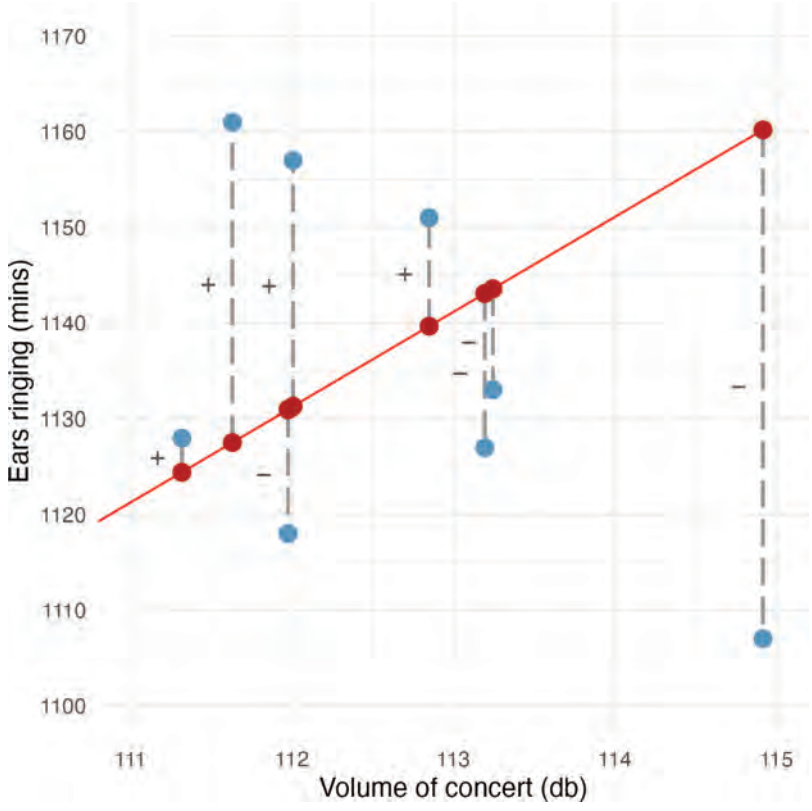


Figure 6.8 The vertical lines (dotted) represent the differences (or residuals) between the predicted values from the model (red dots) and the observed data (blue dots)

each data point that was observed. In other words, we use what we've just described as the residuals, it's just that in the context of using the mean as a model, we called these residuals 'deviations'. Despite the different terminology, the same principles apply to assessing the fit of the mean as a model, and the fit of the linear model.

Figure 6.8 again zooms in on some of the sample data that we have been using so that we can see the residuals for some of the data. The blue circles are the observed data. The line is the model (as we've seen in previous figures) and the red dots are the predicted value from the model for each observed data point. If the model is a perfect fit to the data then for a given value of the predictor(s) the model will predict the same value of the outcome as was observed. In terms of Figure 6.8 this would mean that the red and blue dots fall in the same locations. This is not what happens because the model is not a perfect fit of the data (and it never will be): sometimes it overestimates the observed value of the outcome and sometimes it underestimates it. Remember that the differences between values of the outcome predicted by the model and the observed values (the vertical dashed lines in Figure 6.8) are the residuals we have just discussed (and they are conceptually the same as the *deviations* we discussed when we looked at the mean).

We saw in Chapter 1, Eq. 2.17, that to calculate the total error in a model we square the differences between the observed values of the outcome, and the predicted values that come from the model:

$$\text{Total error} = \sum_{i=1}^{n}\left(\text{observed}_i - \text{model}_i\right)^2$$

$$\text{RSS} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \qquad (6.13)$$

As shown in Figure 6.8, sometimes the predicted value of the outcome is less than the actual value and sometimes it is greater. Consequently, some residuals are positive but others are negative, and if we summed them they would cancel out. The solution is to square them before we add them up (this idea should be familiar from Section 2.6.2). Therefore, to assess the error in a linear model, just like when we assessed the fit of the mean using the variance, we use a sum of squared errors, and because we call these errors residuals, this total is called the *sum of squared residuals* (SS_R) or **residual sum of squares** (RSS). The residual sum of squares is a gauge of how well a model fits the data: if the squared differences are large, the model is not representative of the data (there is a lot of error in prediction); if the squared differences are small, the model is representative.

Let's get back to how we estimate the *b* values. If you were particularly bored, you could draw every possible straight line (linear model) through your data and calculate the residual sum of squares for each one. You could then compare these 'goodness-of-fit' measures and keep the line with the smallest SS_R because it would be the best-fitting model. We have better things to do, so like when we estimate the mean, we use the method of least squares to estimate the parameters (*b*) that define the linear model for which the sum of squared errors is the minimum it can be (given the data). This method is known as **ordinary least squares (OLS)** regression. The method of least squares uses mathematical techniques for finding minima to find the *b*-values that minimize the sum of squared residuals for the model. The main point is that the residuals are part of the process of estimating the *b*s.

Many of the models in this book use OLS estimation (pretty much everything up to Chapter 16) but as we saw in Section 2.7 it is not the only way to estimate model parameters. Maximum likelihood (ML) estimation is a more general estimation method, which, because it is more versatile, is used to estimate some models later in the book. To recap, for models based on the normal distribution, maximum likelihood also uses the residual (see Eq. 2.15) to compute the likelihood of a value, and then selects the parameter value that maximizes the likelihood. Again, residuals are part of the process of estimating the *b*s. OLS and ML are very different processes but, for linear models, converge on the same estimates when model errors are normally distributed. That is, when model errors are normally distributed the OLS estimate *is* the ML estimate (Jane Superbrain Box 6.3).

We've mentioned several times that the model errors have 'constant variance' (σ^2), but how can we know what this variance is when the model errors are unobservable? We can't, but we can estimate it using the sample residuals. Therefore, as well as estimating the model parameters from the data, behind the scenes we also estimate the variance of the model errors. Given the general principle of sticking hats on things that are estimates, you might expect this estimate to be denoted by $\hat{\sigma}^2$, and you be correct when using ML estimation, but for some reason when using OLS it is typically denoted by s^2. In Chapter 1 we saw that we can convert a sum of squared errors to a variance by dividing by the degrees of freedom, and Eq. 6.14 shows that in OLS we do exactly this – the estimate of the variance of errors is the sum of squared residuals divided by the degrees of freedom for the model, $n - p$ (where n is the sample size and p is the number of parameters (*b*s) estimated):

$$s^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-p} = \frac{\text{RSS}}{n-p} \tag{6.14}$$

When ML estimation is used you get a slightly different estimate ($\hat{\sigma}^2$) but it is related to the estimate from OLS,

$$\hat{\sigma}^2 = \frac{n-p}{n}s^2 \tag{6.15}$$

and as samples get larger they converge on the same value (because as n increases the fraction $(n - p) / n$ get closer to 1).

6.2.6 Confidence intervals and significance tests Ⓓ

Having estimated the model parameters ($\hat{b}$s) and error variance ($\hat{\sigma}^2$) using OLS or ML, we can use these estimates to construct confidence intervals around the parameter estimates and test hypotheses about them. I will try to give you a flavour of some key points about what's going on under the hood, but you'll have to consult a more mathematical book for proofs and detail. In Section 2.9 we discovered that to construct confidence intervals around a parameter estimate (we used the mean, but the principles apply to any parameter estimate) we need to know its standard error. Two quick reminders:

1 The sampling distribution is the distribution of parameter estimates across samples. Its width reflects the variability of parameter estimates across samples (i.e. the variability in sampling error).
2 The width of the sampling distribution is quantified using its variance, the square root of which is its standard deviation. The standard deviation of a sampling distribution is called the standard error. The standard error is, therefore, the square root of the variance of the sampling distribution.

Therefore, to establish the standard error of a parameter estimate we need to know the variance of its sampling distribution. In a linear model with one predictor, *when residuals are normally distributed*, the sampling distribution of the parameter estimate for that predictor can be shown to be distributed as:

$$\hat{b} \sim \mathcal{N}\left(b, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \tag{6.16}$$

In plain language Eq. 6.16 means that, *when residuals are normally distributed*, the parameter estimate for the predictor is distributed as a normal distribution with a mean equal to the population value (b) and a variance that is a function of the error variance (σ^2) and the sum of squared errors for the predictor variable itself (the bottom of the fraction). In other words, it is the ratio of the variance in the error in prediction to the total spread of the predictor variable. With more than one predictor, you get a version of Eq. 6.16 involving matrices (see the material on matrices at the start of the book), but it is conceptually the same:

$$\hat{\mathbf{b}} \sim \mathcal{N}(\mathbf{b}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \tag{6.17}$$

Instead of a single parameter estimate $\hat{b}$ we collect of all of the parameter estimates into a vector ($\hat{\mathbf{b}}$) and all these parameter estimates are normally distributed with a mean equal to the corresponding population parameter (collected together in $\mathbf{b}$) and a *variance* equal to the error variance divided by the sum of squared errors for the corresponding predictor variable.[2] Remember that we are talking about the distribution of parameter estimates so these are sampling distributions and, therefore, the variance in Eq. 6.16 and Eq. 6.17 is the standard error squared:

$$SE(\hat{\mathbf{b}})^2 = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{6.18}$$

2 The matrix $\mathbf{X}$ contains a column of 1s to represent the intercept and the remaining columns contain the values of each predictor variable (i.e. the raw data). The $\mathbf{X}^T\mathbf{X}$ is conceptually the matrix $\mathbf{X}$ multiplied by itself but because it contains a column of 1s it also contains information about the means of each predictor variable. You can't divide matrices, but the equivalent process is to multiple by the inverse of the matrix, so $(\mathbf{X}^T\mathbf{X})^{-1}$ is conceptually the same as dividing by the sum of squared errors for the predictors as we did in Eq. 6.16.

As we discovered in the previous section, we don't know the value of σ^2 so we estimate it using the estimate from the data s^2, which gives

$$\widehat{SE}\left(\hat{\mathbf{b}}\right)^2 = s^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$
$$\widehat{SE}\left(\hat{\mathbf{b}}\right) = \sqrt{s^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}} \quad (6.19)$$

The standard error is an estimate and so is, itself, a random variable that has a distribution. Remembering from Eq. 6.14 that s^2 is based on the RSS, it turns out that the distribution of RSS is a chi-square distribution with $n - p$ degrees of freedom scaled by a factor of σ^2:

$$\text{RSS} \sim \sigma^2 \chi^2_{n-p} \quad (6.20)$$

You will have to take on trust that because $\widehat{SE}(\hat{b})$ is related to s^2, which in turn is based on the RSS, which has a scaled chi-square distribution, the sampling distribution of $\widehat{SE}(\hat{b})$ also follows a scaled chi-square distribution. This fact becomes important later.

What about hypothesis tests? We saw in Section 2.10 that bs represent hypotheses. One way that we can frame the null and alternative hypothesis is:

- $H_0 : b = 0$. If there is no association between a predictor and outcome variable we'd expect the parameter for that predictor to be zero. Therefore, a common null hypothesis is that $b = 0$.
- $H_1 : b \neq 0$. If there is a non-zero association between a predictor and outcome variable we'd expect the parameter for that predictor to be non-zero. Therefore, a common alternative hypothesis is that $b \neq 0$.

We also saw in Section 2.10 that we can test hypotheses (Eq. 2.33) using a test statistic that is the ratio of the size of the parameter to its sampling variation. To show how we get to this conclusion, Eq. 6.21 shows that to test an effect we compare the effect to the error in measuring that effect. In this case the 'effect' we want to test is the difference between the observed parameter estimate ($\hat{b}$) and the value we'd expect if the null hypothesis were true ($\hat{b}_{\text{Expected}}$). If the null is true, we expect the parameter value to be 0 ($\hat{b}_{\text{Expected}} = 0$) so the top half of the ratio reduces to the value of the observed $\hat{b}$. The error in measuring this parameter estimate is its sampling variation (i.e., how much estimates vary across samples), which is quantified by its standard error. The resulting test statistic for each parameter estimate is therefore its value ($\hat{b}$) divided by its standard error:

$$\begin{aligned}\text{Test statistic} &= \frac{\text{effect}}{\text{error}} \\ t_{n-p} &= \frac{\hat{b} - \hat{b}_{\text{Expected}}}{\widehat{SE}\left(\hat{b}\right)} \\ &= \frac{\hat{b}}{\widehat{SE}\left(\hat{b}\right)} \\ &= \frac{\text{size of parameter}}{\text{sampling variation in the parameter}}\end{aligned} \quad (6.21)$$

Notice that in Eq. 6.21 I end up replacing the phrase 'test statistic' with t_{n-p}. The top half of the equation for the test statistic is a parameter estimate that, as we discovered earlier (Eq. 6.16), has a normal sampling distribution. The bottom half of the equation is an estimate of the standard error that, as we saw earlier (Eq. 6.20), follows a scaled chi-square distribution. The resulting test statistic is, therefore, the result of dividing a normally distributed random variable with one that is a scaled chi-square

distribution with $n - p$ degrees of freedom. You can look elsewhere for the frankly terrifying mathematic proof, but when you divide a normal distribution by a scaled chi-square distribution the result is a *t*-distribution.[3]

Like we did in Section 2.9 we can also get the upper and lower boundary of a $100(1 - \alpha)\%$ confidence interval for $\hat{b}$ by rearranging a version of Eq. 6.21 in which $\hat{b}_{\text{Expected}}$ is the observed estimate of the population value from the data ($\hat{b}$), and t is replaced with the $(1 - \alpha/2)$ -quantile of the *t*-distribution (with $n - p$ degrees of freedom). For example, with the standard $\alpha = 0.05$, we want a $1 - 0.05 = 0.95 = 95\%$ confidence interval, which means that we want the value of t that cuts off the top and bottom $\alpha/2 = 0.025$ of the distribution, which will be $\pm t_{0.975}$. In general, then,

$$\frac{\hat{b}_{\text{CI boundary}} - \hat{b}}{\widehat{SE}\left(\hat{b}\right)} = \pm t_{1-(\alpha/2)}$$
$$\hat{b}_{CI\ boundary} = \hat{b} \pm t_{1-(\alpha/2)} \widehat{SE}\left(\hat{b}\right) \tag{6.22}$$

6.2.7 The central limit theorem ©

After that fun trip down standard error lane, you'll be looking forward to a nice non-technical section on how to overcome maths-induced trauma. Think again. In the previous section we saw that *when model errors are normally distributed*, the sampling distribution of $\hat{b}$ is also normal, and knowing that allows us to estimate the standard error of $\hat{b}$, which in turn allows us to construct confidence intervals and test statistics (based on a *t*-distribution). However, there is another situation in which we know that the sampling distribution of $\hat{b}$ is normal: *in large samples*.

The central limit theorem (CLT)[4] is a fundamental theorem in statistics. Let's get the maths out of the way. Imagine that $X_1, X_2, \ldots X_n$ is a sequence of independent random variables taken from a population with mean μ and variance σ^2 and we define the mean in the usual way as

$$\bar{X} = \frac{X_1 + X_2, + \ldots + X_n}{n} \tag{6.23}$$

Then the central limit theorem shows that

$$n \to \infty, \quad \sqrt{n}\left(\bar{X} - \mu\right) \approx \mathcal{N}\left(0, \sigma^2\right) \tag{6.24}$$

Breaking apart Eq. 6.24 and phrasing in a (hopefully) understandable way:

- $n \to \infty$ means 'as n tends to infinity'. In plain language, as sample sizes become larger.
- $\sqrt{n}(\bar{X} - \mu)$ means 'the difference between the sample mean and the population mean (factoring in the sample size)'.

3 There are a couple of simplifications here that you almost certainly won't care about. First, technically I should have written 'when you divide a normal distribution by *the square root of* a scaled chi-square distribution the result is a *t*-distribution'. However, the route to how we end up at a *t*-distribution is immeasurably painful and you'll thank me for my simplified version. Second, the resulting distribution is *t when the null hypothesis is true*, but when it's not true it's actually a non-central *t*-distribution. Knowing these things will add precisely no value to your life.

4 The 'central' in the name refers to the theorem being important and far-reaching and has nothing to do with centres of distributions. Also, there are several variants of the CLT rather than a single theorem, but I focus on the classic variant.

- $\approx \mathcal{N}(0,\sigma^2)$ means 'is approximately distributed as a normal distribution with a mean of 0 and variance of σ^2' (i.e., the variance of the population from which samples came).

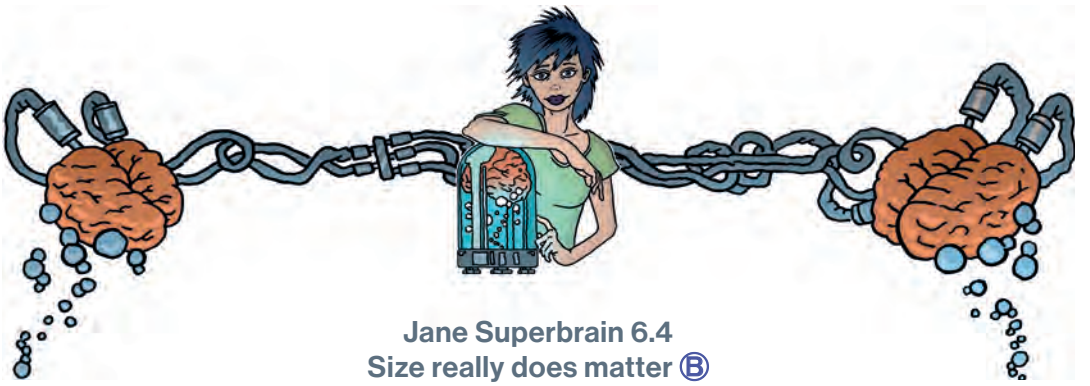
Eq. 6.24 can be expressed equivalently as

$$n \to \infty, \quad \overline{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{6.25}$$

Eq. 6.24 which we can break down as follows:

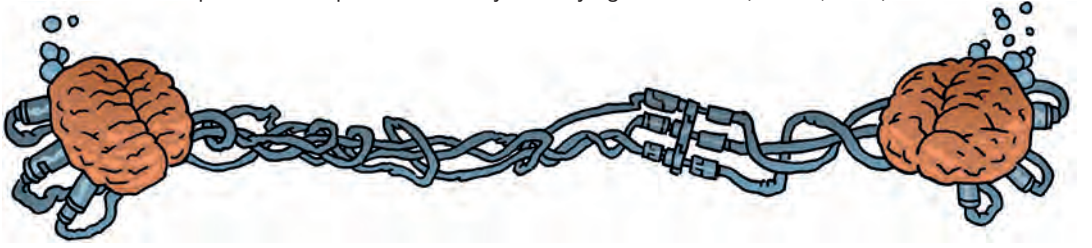
- $n \to \infty$ again means 'as sample sizes become larger'.
- $\overline{X}$ means 'the sample mean'.
- $\approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ means 'is approximately normally distributed with a mean of μ and a variance of σ^2/n'.

Therefore, an equivalent interpretation is that as samples get larger, the distribution of sample means converges to a normal distribution around the population mean (μ) with variance σ^2/n. That is, as samples get larger sample means, on average, tend to be the same as the population mean and are normally distributed with variance equal to the population variance divided by the sample size. We're discussing the mean as the parameter of interest, but the theorem applies to any parameter (and remember that the mean is $\hat{b}_0$ in a linear model with no predictors).



Jane Superbrain 6.4
Size really does matter Ⓑ

How big is 'big enough' for the central limit theorem to kick in? The widely accepted theoretical value is a sample size of 30, and we saw in Figure 6.9 that with samples of this size we got a sampling distribution that approximated normal. We also saw that samples of 100 yielded a better approximation of normal. There isn't a simple answer to how big is 'big enough': it depends on the population distribution. In light-tailed distributions (where outliers are rare) an N as small as 20 can be 'big enough', but in heavy-tailed distributions (where outliers are common) then up to 100 or even 160 might be necessary. If the distribution has a lot of skew and kurtosis you might need a very large sample indeed for the central limit theorem to work. It also depends on the parameter that you're trying to estimate (Wilcox, 2010).



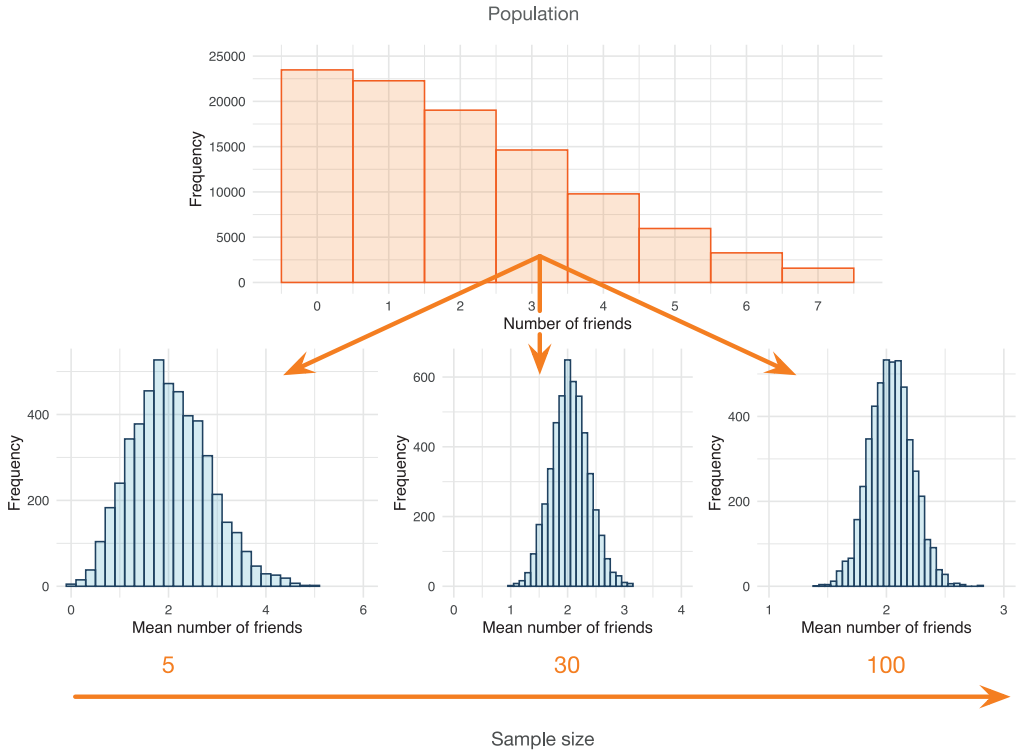


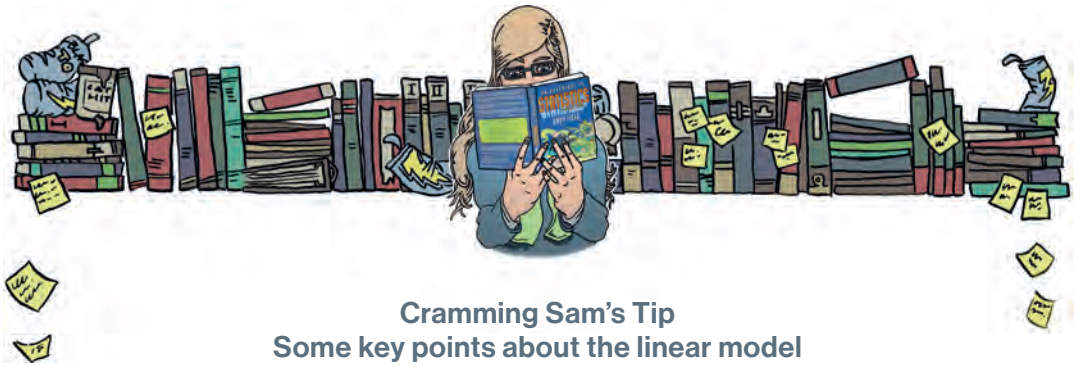
Figure 6.9 Parameter estimates sampled from a non-normal population. As the sample size increases, the distribution of those parameters becomes increasingly normal

To get an intuition about this theorem, imagine we have a population of scores that is not normally distributed. Figure 6.9 shows such a population containing scores of how many friends statistics lecturers have. It is very skewed, with most lecturers having no friends, and the frequencies declining as the number of friends increases to the maximum score of 7 friends. I'm not tricking you, this population is as far removed from the bell-shaped normal curve as it looks. Imagine that I take samples of five scores from this population and in each sample I estimate a parameter (let's say I compute the mean) and then replace the scores. In total, I take 5000 samples, which gives me 5000 values of the parameter estimate (one from each sample). The frequency distribution of the 5000 parameter estimates from the 5000 samples is on the far left of Figure 6.9. This is the sampling distribution of the parameter. Note that it is a bit skewed, but not nearly as skewed as the population. Imagine that I repeat this sampling process, but this time my samples each contain 30 scores instead of five. The resulting distribution of the 5000 parameter estimates is in the centre of Figure 6.9. The skew is gone and the distribution looks normal. Finally, I repeat the whole process but this time take samples of 100 scores rather than 30. Again, the resulting distribution is basically normal (right of Figure 6.9). As the sample sizes get bigger the sampling distributions become more normal, until a point at which the sample is big enough that the sampling distribution *is* normal – even though the population of scores is very non-normal indeed. This is the central limit theorem: regardless of the shape of the population, parameter estimates from that population will have a normal distribution provided the samples are 'big enough' (see Jane Superbrain Box 6.4).

In the context of the linear model the central limit theorem has an important implication, which is that *in large samples* the sampling distribution of $\hat{b}$ will be normal. This fact means that we can rephrase our statement at the beginning of this section as:

- *In large samples and when model errors are normally distributed*, we know that the sampling distribution of $\hat{b}$ is normal, and knowing that allows us to estimate the standard error of $\hat{b}$, which in turn allows us to construct confidence intervals and test statistics (based on a *t*-distribution).

This section has been a quite mind-blowing and cursory overview of some key theoretical points about the linear model, which Cramming Sam will summarize in an attempt to restore your brain to some kind of pain-free state.



Cramming Sam's Tip
Some key points about the linear model

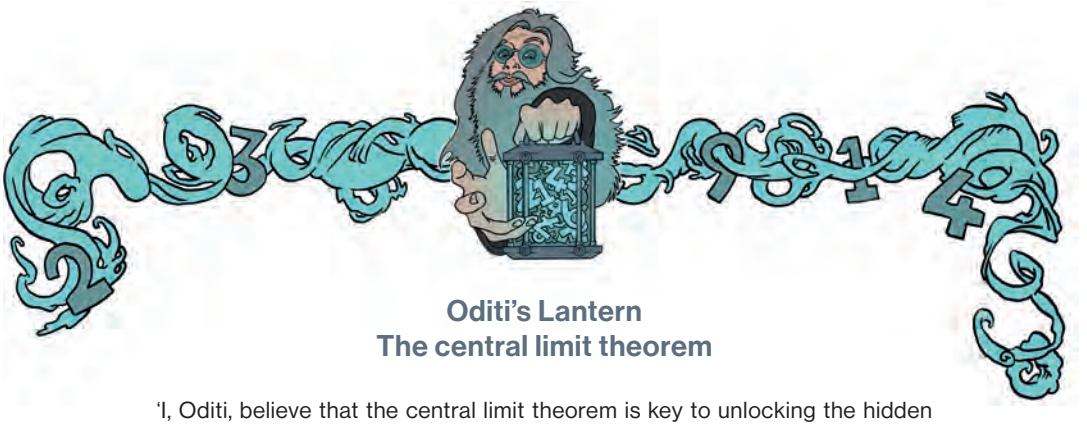
The key points in this section:

- *In large samples and when model errors are normally distributed*, the sampling distribution for $\hat{b}$ is normal.
- When the sampling distribution for $\hat{b}$ is normal, the equation for its variance (and hence standard error) is known.
- This equation uses the model error variance (σ^2), which is unobservable. Instead, we use an estimate from the data (s^2) to get an estimate of the standard error, $\widehat{SE}(\hat{b})$.
- The sampling distribution of $\widehat{SE}(\hat{b})$ is a scaled chi-square distribution with $n - p$ degrees of freedom.
- Knowing $\widehat{SE}(\hat{b})$ enables us to construct a confidence interval and hypothesis test for $\hat{b}$.
- We can test whether each $\hat{b}$ differs from zero by diving it by its estimated standard error.
- The resulting test statistic has a *t*-distribution with $n - p$ degrees of freedom.
- We can also construct confidence intervals using the *t*-distribution with $n - p$ degrees of freedom



6.2.8 The Gauss–Markov theorem Ⓓ

The previous sections have discussed the model errors at length, and in doing so we have touched implicitly upon three assumptions, which are at the heart of something called the Gauss–Markov theorem. This theorem says that when certain conditions are met the OLS estimator has the smallest sampling variance of any linear unbiased estimator. Put more simply, the OLS estimate is the 'best' of the relevant estimators when these conditions are met:



Oditi's Lantern
The central limit theorem

'I, Oditi, believe that the central limit theorem is key to unlocking the hidden truths that the cult strives to find. The true wonder of the CLT cannot be understood by a static diagram and the ramblings of a damaged mind. Only by staring into my lantern can you see the CLT at work in all its wonder. Go forth and look into the abyss.'



1 Model errors are, on average, zero. More formally, we can state that the expected value of model errors is zero, which is written as $E[\varepsilon_i] = 0$
2 We have mentioned many times that model errors have constant variance, and this condition is referred to as **homoscedasticity** (also known as **homogeneity of variance**). More formally, errors are assumed to have a fixed finite (i.e. less than infinity) variance, which can be written as $\text{Var}(\varepsilon_i) = \sigma^2 < \infty$.
3 We haven't really mentioned this until now, but model errors are assumed to be independent (i.e. uncorrelated), which can be written as $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$. This condition is referred to as **independent errors**.

The last two conditions are often lumped together under the term *spherical errors*.

What does it mean to have spherical errors? Imagine we could actually observe population errors and we put them into a vector (basically a column):

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{6.26}$$

We've called the vector $\boldsymbol{\varepsilon}$ because it collects together all of the εs for the model. Note, for example, that the first element in the vector is the error in prediction for case 1 (ε_1), the second element is the error associated with case 2 (ε_2) and you need to imagine this column continuing down until we reach the final case, n, whose error in prediction is represented by ε_n.

Now imagine we wanted to square these errors (like we've done with other errors because some of them are positive and others negative). You can't multiply a single column vector by itself,[5] but you can multiple it by what's known as its transpose ($\boldsymbol{\varepsilon}'$) which has the same values but arranged in a single row (instead of a column):

$$\boldsymbol{\varepsilon}' = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3 \quad \cdots \quad \varepsilon_n] \tag{6.27}$$

When you multiple these vectors together each element in the column vector ($\boldsymbol{\varepsilon}$) gets multiplied by every element in the row vector ($\boldsymbol{\varepsilon}'$). The result is a square matrix:

$$\begin{aligned}
\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' &= \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \cdots & \varepsilon_n \end{bmatrix} \\
&= \begin{bmatrix} \varepsilon_1\varepsilon_1 & \varepsilon_1\varepsilon_2 & \varepsilon_1\varepsilon_3 & \cdots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2\varepsilon_2 & \varepsilon_2\varepsilon_3 & \cdots & \varepsilon_2\varepsilon_n \\ \varepsilon_3\varepsilon_1 & \varepsilon_3\varepsilon_2 & \varepsilon_3\varepsilon_3 & \cdots & \varepsilon_3\varepsilon_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \varepsilon_n\varepsilon_3 & \cdots & \varepsilon_n\varepsilon_n \end{bmatrix} \\
&= \begin{bmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \varepsilon_1\varepsilon_3 & \cdots & \varepsilon_1\varepsilon_n \\ \varepsilon_1\varepsilon_2 & \varepsilon_2^2 & \varepsilon_2\varepsilon_3 & \cdots & \varepsilon_2\varepsilon_n \\ \varepsilon_1\varepsilon_3 & \varepsilon_2\varepsilon_3 & \varepsilon_3^2 & \cdots & \varepsilon_3\varepsilon_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varepsilon_1\varepsilon_n & \varepsilon_2\varepsilon_n & \varepsilon_3\varepsilon_n & \cdots & \varepsilon_n^2 \end{bmatrix}
\end{aligned} \tag{6.28}$$

Notice that (1) the elements along the diagonal are errors multiplied by themselves, (2) the off-diagonal elements are pairs of errors multiplied together; and (3) elements below the diagonal are a mirror image of those above. In fact, this oversimplifies things a bit because we're actually interested in the squared errors given the observed data (i.e., given the values of the predictor(s)). We'll denote the values of the predictors by $\mathbf{X}$.[6] Therefore, we can represent the *expected* errors given the values of the predictors as:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}) = \boldsymbol{\Phi} = \begin{bmatrix} E(\varepsilon_1^2 \mid \mathbf{X}) & E(\varepsilon_1\varepsilon_2 \mid \mathbf{X}) & E(\varepsilon_1\varepsilon_3 \mid \mathbf{X}) & \cdots & E(\varepsilon_1\varepsilon_n \mid \mathbf{X}) \\ E(\varepsilon_1\varepsilon_2 \mid \mathbf{X}) & E(\boldsymbol{\varepsilon}_2^2 \mid \mathbf{X}) & E(\varepsilon_2\varepsilon_3 \mid \mathbf{X}) & \cdots & E(\varepsilon_2\varepsilon_n \mid \mathbf{X}) \\ E(\varepsilon_1\varepsilon_3 \mid \mathbf{X}) & E(\varepsilon_2\varepsilon_3 \mid \mathbf{X}) & E(\varepsilon_3^2 \mid \mathbf{X}) & \cdots & E(\varepsilon_3\varepsilon_n \mid \mathbf{X}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E(\varepsilon_1\varepsilon_n \mid \mathbf{X}) & E(\varepsilon_2\varepsilon_n \mid \mathbf{X}) & E(\varepsilon_3\varepsilon_n \mid \mathbf{X}) & \cdots & E(\varepsilon_n^2 \mid \mathbf{X}) \end{bmatrix} \tag{6.29}$$

This matrix, denoted by $\boldsymbol{\Phi}$, is called the is the variance–covariance matrix for the errors in the model. You can think of it as the expected squared error in the outcome for each case given the values of the predictors. The diagonal elements of this matrix represent the variance in the error around a given value of the predictors, and the off-diagonal elements represent the relationships between errors (for different values of the predictors). That's why it's called a variance–covariance matrix. The assumption of spherical errors relates to this matrix and it basically says that the matrix has the following form:

5 You'll need to look up some matrix theory to find out why.

6 We met $\mathbf{X}$ in an earlier footnote, it is a matrix containing a column of 1s for the intercept term, and then columns containing the observed values of each predictor variable.

$$\Phi = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \tag{6.30}$$

The diagonal elements relate to homoscedasticity, and are assumed to be the same and equal to the overall population variance in errors (σ^2). Put another way, the variance in the population errors does not depend on the values of the predictors and the spread of errors around the model is constant as the values of predictors change. This formalizes the idea that errors have constant variance. The off-diagonal elements relate to independence of the errors, because they represent the covariances between model errors. Note that all of the off-diagonal elements are 0, meaning that the covariance between all pairs of errors is zero, and no error is affected by any other error. Put another way, all model errors are independent.

You might recall that I said that we couldn't observe the population errors. You might reasonably wonder how then we can put them into a matrix. Essentially, the matrix I have just described, $\mathbf{\Phi}$, is involved in the equations to estimate the variances of the model parameters and the covariances between model parameters. However, if $\mathbf{\Phi}$ has the form in Eq. 6.29 then we don't need to know all of the elements of $\mathbf{\Phi}$, we only need to know the variance in population errors (σ^2), the term on the diagonal. As we have seen, this value can be estimated from the residuals in the sample model (Eq. 6.14). That's why we need to assume that $\mathbf{\Phi}$ has this form, because when it does the only thing that needs to be estimated from the data (other than the parameters) is σ^2.

6.3 What is bias? Ⓐ

If you skipped the hellish mathsy section then welcome back; if you read it, treat yourself to a break with some tea and chocolate. With all the technical stuff out of the way we can begin to think about what we mean by bias. If you support a sports team then at some point in your life you've probably accused a referee of being 'biased' (or worse). If not, perhaps you've watched a TV show like *The Voice* and felt that one of the judges was 'biased' towards the singers that they mentored. In these contexts, bias means that the summary information from the person ('Jasmin's singing was note perfect throughout') is at odds with the objective truth (pitch analysis shows that 33% of Jasmin's notes were sharp or flat). Similarly, in statistics the parameters that we estimate can be at odds with the true values, and statistical tests can make errors.

At the start of this chapter, I mentioned that people fit models with different aims: some want to estimate parameters, some want to test hypotheses, and some want to do both. In terms of bias, if we've fitted a model to estimate parameters then we care about things that bias parameter estimates and their confidence intervals, but if we want to use those parameter estimates to test hypotheses then we also need to care about things that bias significance tests.

We can think about parameter estimates as 'at odds with the true values' in terms of these properties:

- An **unbiased estimator** is one that yields an expected value that is the same as the one it is trying to estimate.[7] Informally, we're saying that *on average* the estimate in the sample will match the population value. Formally, we can write this property as the expected value of the

7 You might recall that when estimating the population variance we divide by $N - 1$ instead of N (see Section 3.6.2). This has the effect of turning the estimate from a biased one (using N) to an unbiased one (using $N - 1$).

parameter (e.g., the mean of its sampling distribution) being equal to the population value, $E[\hat{b}] = b$. For example, for the OLS estimator of $\hat{b}$ to be unbiased, it needs to be the case that, on average, it yields estimates that match the population value. (Although that doesn't mean that they will match in your particular sample.)

- A **consistent estimator** is one that produces estimates that tends to the population value as the sample size tends to infinity, $\lim_{N\to\infty} \hat{b} = b$. For example, for the OLS estimator of $\hat{b}$ to be consistent, it needs to be the case that its estimates converge on the population value as the sample size increases.
- An **efficient estimator** is one that produces estimates that are in some sense 'the best' of the available methods of estimation. Often 'best' is quantified by 'lowest variance' because smaller variance implies that estimates are distributed more tightly around the population, that is, estimates tend to be similar to each other and to the population value.

For confidence intervals we can consider this property:

- The **coverage** of a confidence interval is the proportion of time it contains the population value of the thing it is estimating. For example, for 95% confidence intervals the coverage should be 0.95, and if it is not then it has been biased.

For significance tests we need to consider:

- **Type I error rate**. An unbiased significance test is one that controls the Type I error rate to the degree it is supposed to. If we set the Type I error rate to $\alpha = 0.05$ (or 5%), then a significance test should falsely reject the null hypothesis 5% of the time. If the Type I error rate of a test is greater or less than the designated α then it is biased.
- **Type II error rate**: An unbiased significance test is one that controls the Type II error rate to the degree it is supposed to. If we set the Type II error rate to $\beta = 0.2$ (or 20%), then a significance test should falsely accept the null hypothesis 20% of the time. If the Type II error rate of a test is greater or less than the designated β then it is biased.

Needless to say, if the statistical information we use to infer things about the world is biased, then our inferences will be too. Throughout the rest of the chapter we will look at different sources of bias, being mindful of *what* they bias ($\hat{b}$s and their standard errors, confidence intervals or associated *p*-values). The beast of bias is a two-headed, fire-breathing, green-scaled beast that jumps out from behind a mound of blood-soaked moss to try to eat us alive. One head goes by the name of *outliers*, whereas the other is called *violations of assumptions*. With names like that the beast probably hated its parents and got teased at school, but it could breathe fire from both heads, so it could handle it. Onward into battle …

6.4 Outliers Ⓐ

Before we get to the assumptions of the linear model, we'll look at the first head of the beast of bias: outliers. An **outlier** is a score very different from the rest of the data. Let's look at an example. When I published my first book (the first edition of this book), I was very excited and I wanted everyone in the world to love my new creation and me. Consequently, I obsessively checked the book's ratings on amazon.co.uk. Customer ratings can range from 1 to 5 stars, where 5 is the best. Back in around 2002, my first book had seven ratings (in the order given) of 2, 5, 4, 5, 5, 5, and 5. All but one of these ratings are similar (mainly 5 and 4) but the first rating was quite different from the rest – it was a rating of 2 (a mean and horrible rating). Figure 6.10 plots the seven reviewers on the horizontal axis and their ratings on the vertical axis. The blue horizontal line shows the mean rating (4.43 as it happens).

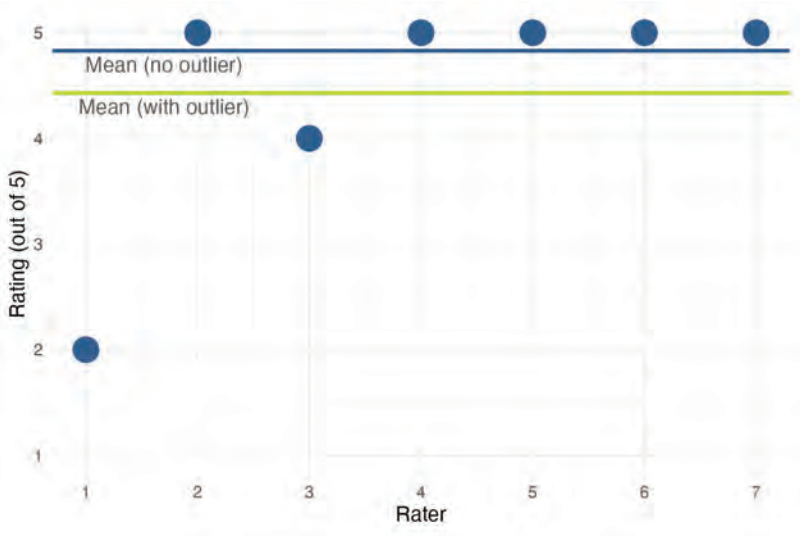


Figure 6.10 The first seven customer ratings of the first edition of this book on www.amazon.co.uk (in about 2002). The first score biases the mean

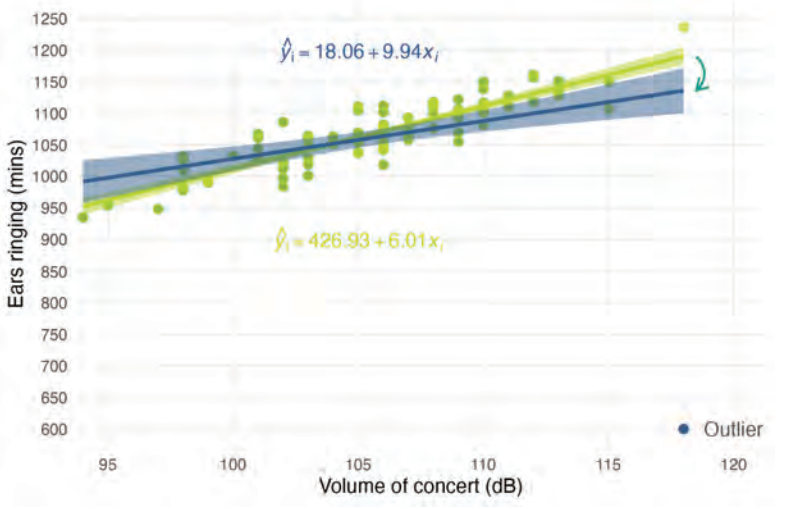


Figure 6.11 Graph demonstrating the effect of an outlier. The blue line represents the original regression line for these data, whereas the orange line represents the regression line when an outlier is present

All of the scores except one lie close to this line. The rating of 2 lies way below the mean and is an example of an outlier – a weird and unusual person (I mean, score) that deviates from the rest of humanity (I mean, data set). The green horizontal line shows the mean excluding the outlier (4.83). This line is higher than the original mean, indicating that by ignoring this score the mean increases (by 0.4). This example shows how a single score, from some mean-spirited badger turd, can bias a parameter estimate such as the mean: the first rating of 2 drags the average down. Based on this biased estimate, new customers might erroneously conclude that my book is worse than the population thinks it is. An important point though is that much as I like to remove this troll from the data, the fact the score is an outlier isn't a reason to do so unless we have strong evidence that this person isn't from the population of interest. For example, if I am interested in the population of people who are not arseholes, then maybe I could exclude this case. If my population of interest is the general population then I definitely can't. I'm not bitter.

Outliers can affect the estimates in more complex models too. Figure 6.11 shows the sample data we have been using throughout this chapter (e.g., in Figure 6.7), but also includes the data with one score changed to be an outlier (in this case a concert that was very loud, but resulted in a relatively low 600 minutes of ears ringing). The blue line shows the original model, and the orange line shows the model with the outlier included. The outlier makes the line flatter and the equations for the two models show that the outlier makes b_1 smaller (it decreases from 9.94 to 6.01) and although you can't see it on the plot, b_0 increases (from 18.06 to 426.93). This example shows how outliers bias model parameters (the *b*s).

Outliers bias parameter estimates, but they have an even greater impact on the error associated with that estimate. The two models in Figure 6.11 have shaded areas around them depicting the standard error around the model. Think of these areas as the uncertainty we have around the predicted values from the model: the wider the area, the more the predicted values in the population could plausibly differ from the values in our sample. Notice that with the outlier included this area (blue) is wider than the model without the outlier (green): there is more uncertainty around the model with the outlier than the model with no outlier.

We can demonstrate why outliers have an impact on the standard error by going back to the example of the mean as a model. Back in Section 2.6.1 we looked at an example showing of the number of friends that five statistics lecturers had. The data were 1, 3, 4, 3, 2, the mean was 2.6 and the sum of squared error was 5.2. Let's replace one of the scores with an outlier by changing the 4 to a 10. The data are now: 1, 3, 10, 3 and 2.



If you did the self-test, you should find that the mean of the data with the outlier is 3.8 and the sum of squared error is 50.8. Figure 6.12 shows these values. Like Figure 2.8, it shows the sum of squared error (*y*-axis) associated with different potential values of the mean (the parameter we're estimating, *b*). For both the original data set and the one with the outlier the estimate for the mean is the optimal estimate: it is the one with the least error, which you can tell because the curve converges on the values of the mean (2.6 and 3.8). The presence of the outlier, however, pushes the curve to the right (it makes the mean higher) and upwards (it makes the sum of squared error larger). By comparing the horizontal to vertical shift in the curve you should see that the outlier affects the sum of squared error more dramatically than the parameter estimate itself (the curve shifts vertically more than it shifts horizontally). The reason is that the sum of squared errors uses *squared* errors, so any bias created by the outlier is magnified by the fact that deviations are squared.[8]

The dramatic effect of outliers on the sum of squared errors is important because it is used to compute the standard deviation, which in turn is used to estimate the standard error, which itself is used to calculate confidence intervals around the parameter estimate and test statistics. If the sum of squared errors is biased, the associated standard error, confidence interval and test statistic will be too. Remember that in the linear model the standard error for a parameter is based on a squared error (RSS in Eq. 6.13) so the same logic applies: outliers will bias the standard error of parameter estimates.

8 In this example, the difference between the outlier and the mean (the deviance) is 10 – 3.8 = 6.2. The deviance squared is $6.2^2 = 38.44$. Therefore, of the 50.8 units of error that we have, a whopping 38.44 are attributable to the outlier.

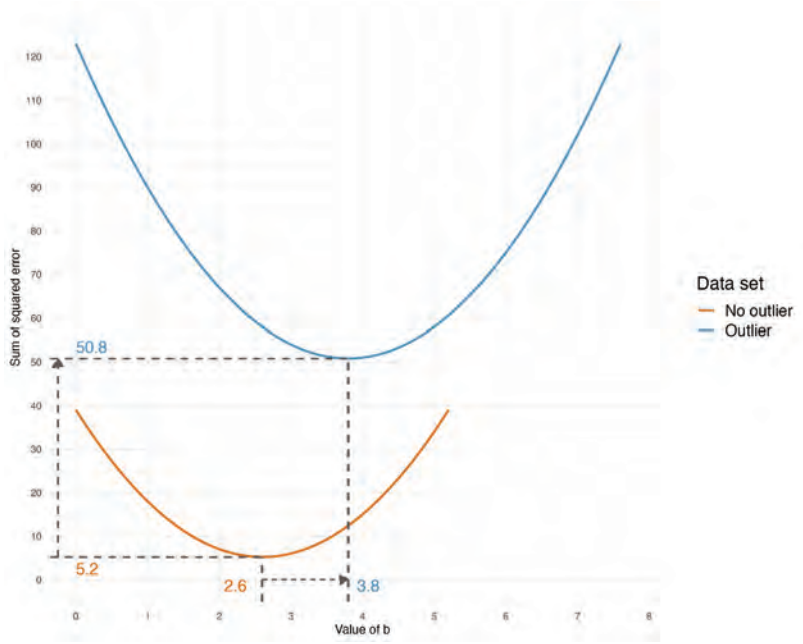
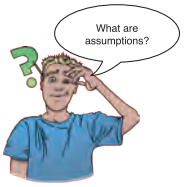


Figure 6.12 The effect of an outlier on a parameter estimate (the mean) and its associated estimate of error (the sum of squared errors)

To end, I'd reiterate that even though outliers can bias models they are, in general, best retained in the data (although you might need to adjust the model you fit). Unless you have concrete evidence that an outlier doesn't reflect the population of interest, you retain it. In the current example, our outlier was a concert that was very loud, but resulted in a relatively short duration of ringing in the ears. Assuming the data have been entered correctly, the question to ask is why the duration of ringing ears was so short. The default assumption should be that it reflects natural variation and you'd want to retain this information in your data. However, let's say that you measured whether people wore earplugs at the concerts and it turned out that this person was the only person who did. This not only explains the duration of ringing ears, but also puts this person in a different population (assuming your interest is in concert goers who don't protect their hearing). In this situation you could reasonably exclude them from the model, provided that you report that you have and the reason why. Ideally you'd pre-register these sorts of exclusion criteria. This outlier might also prompt a future study that compares the populations of concert goers who wear earplugs to those that don't.

6.5 Overview of assumptions Ⓐ



The second head of the beast of bias is called 'violation of assumptions'. An assumption is a condition that ensures that what you're attempting to do works as it should. The statistical procedures common in the social sciences are often presented as unique tests with idiosyncratic assumptions, which can be confusing. However, because most of these procedures are variations of the linear model (see Section 2.3) estimated using OLS they share a common set of assumptions. We discussed these assumptions in Section 6.2.8 and we've also at various points mentioned normality of errors. However, by far the most important assumption is that of additivity and linearity, which can be summed up as the assumption that the process we're trying to model can be described by a linear model. This assumption trumps all others because if it is incorrect then we're

simply describing the world with the wrong model and need to fit a different one. We can list the assumptions in order or importance as:

- linearity and additivity
- the expected value of errors is zero
- spherical errors
  - homoscedasticity (homogeneity of variance)
  - independence of errors
- normality of something or other (the **assumption of normality**).

We need to think about each of these assumptions with respect to what we're trying to achieve with our model. Table 6.1 summarizes some of the things we might want from a model, and the assumptions required for them. Note that linearity and additivity are required by everything! For our $\hat{b}$s to be unbiased we only need the errors to be, on average, zero and this will almost always be the case so I won't mention this assumption any more. If we want our $\hat{b}$s to be unbiased and also the best (in terms of minimizing variance) then we additionally need spherical errors. For the standard error of $\hat{b}$ to be unbiased we need spherical errors (see Section 6.2.8) but also normal errors or a large sample size because the equation for the standard error (Eq. 6.16) holds only when residuals are normal (Section 6.2.6) or the sample size is large (Section 6.2.7). Bias in the standard error will create bias in confidence intervals and significance tests (because they both use the standard error in their calculation) so outcomes involving confidence intervals and significance tests have the same assumptions as the standard error. We'll now look at each assumption in detail.

Table 6.1 Summary of the linear model assumptions

Desirable outcome	Required conditions	Conditions not met
Unbiased estimates	• Linearity and additivity • Errors are, on average, zero • No outliers/influential cases	• Robust estimation
Unbiased and efficient (BLUE) estimates	• Linearity and additivity • Errors are, on average, zero • Spherical errors (homoscedastic and independent) • No outliers/influential cases	• Weighted least squares estimation
The OLS estimate to equal the ML estimate	• Linearity and additivity • Errors are, on average, zero • Spherical errors (homoscedastic and independent) • Normal residuals	• Stop wanting the OLS and ML estimates to be equal!
Unbiased standard error	• Linearity and additivity • Errors are, on average, zero • Spherical errors (homoscedastic and independent) • Normal residuals (or a large sample) • No outliers/influential cases	• Sandwich estimate (e.g., HC4) • Bootstrap estimate
Unbiased confidence intervals	• Unbiased standard error (see above)	• Sandwich estimate of standard error • Bootstrap estimate of standard error
Unbiased significance tests	• Unbiased standard error (see above)	• Sandwich estimate of standard error • Bootstrap estimate of standard error

6.6 Linearity and additivity Ⓐ

The vast majority of statistical models in this book are based on the linear model, which we reviewed a few pages back. It means that scores on the outcome variable are, in reality, linearly related to any predictors, and that if you have several predictors then their combined effect is best described by adding their effects together. This assumption is the most important because if it is not true then, even if all other assumptions are met, your model is invalid because your description of the process you want to model is wrong. If the relationship between variables is curvilinear, then describing it with a linear model is wrong (think back to Jane Superbrain Box 2.1). It's a bit like calling your pet cat a dog: you can try to get it to go into a kennel, or fetch a ball, or sit when you tell it to, but don't be surprised when it coughs up a hairball because no matter how often you describe it as a dog, it is in fact a cat. Similarly, if you describe your statistical model inaccurately it won't behave itself and there's no point in interpreting its parameter estimates or worrying about significance tests of confidence intervals: the model is wrong.

6.7 Spherical errors Ⓒ

The next set of assumptions are conditions for the Gauss–Markov theorem (Section 6.2.8). The first condition is that model errors are, on average, zero. This condition is almost always met, so we'll focus on the other two conditions, which fall under the banner of 'spherical errors'. The two conditions are:

1. Population errors are assumed to have homoscedasticity (also spelt as homoskedasticity and sometimes referred to as homogeneity of variance).
2. Population errors are assumed to be independent or uncorrelated.

Section 6.2.8 was a dip into the acid bath of trying to understand what homoscedasticity means for estimating the model. Having done the somewhat technical explanation there, we now turn to a more intuitive one.

6.7.1 What is homoscedasticity? Ⓑ

Back in Section 6.2 we looked at the process of sampling from a population and, referring to Figure 6.2, I wrote:

> The dotted red lines mark out the limits between which (hypothetical) data points around the population model fall. Note these lines are parallel to the model itself, showing that at all values of the predictor variable (all values along the *x*-axis), the spread of scores around the model is assumed to be similar.

Think about what this means for the errors. The errors are the vertical distances between each (hypothetical) observation and the line. Therefore, at a given point along the line the largest error is the distance between the red dashed line (which marks out the far edge of the data cloud) and the line representing the model. The spread of the errors at any point on the model will be the distance between the two red dashed lines. The top of Figure 6.2 is reproduced in Figure 6.13 (left) but I have used arrows to marked the spread of the data around the model (and therefore the spread of the errors) at different values of the predictor (concert volumes of 95, 100, 105, 110, and 115 dB) as the distance between the two red dashed lines. Notice that this distance is the same at all values of the predictor. The spread, or variance, of errors is consistent no matter what the value of the predictor. This is what the assumption of homoscedasticity means: the variance in errors is the same at all values of the predictor(s).

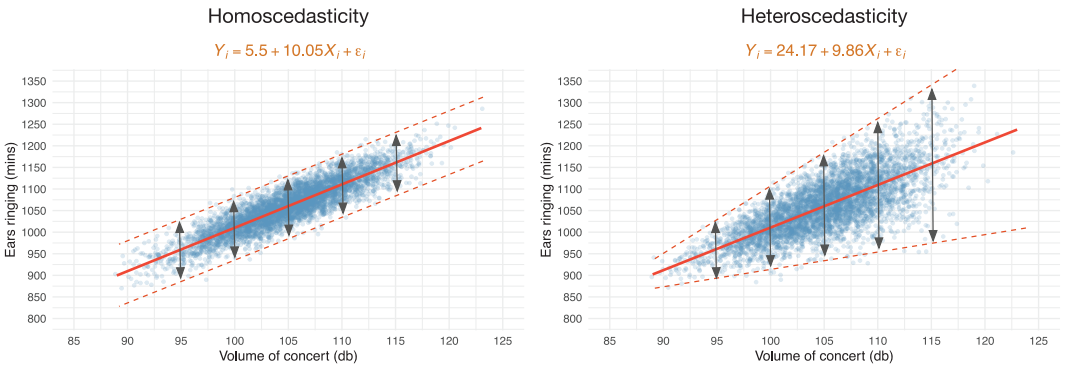


Figure 6.13 Populations showing characteristics of homoscedasticity and heteroscedasticity

Contrast this situation with Figure 6.13 (right) where at relatively low volumes (e.g., 95 dB) the spread of hypothetical observations (and therefore errors) is much smaller than at higher volumes (e.g., 115 dB). In fact, the spread of errors increases as a function of the predictor: as concert volumes increase so does the spread of population values (and, therefore, the variance of the associated errors in prediction). In this situation the red dashed lines that trace the outline of the data cloud funnel out. This pattern is typical of when the assumption of homoscedasticity is violated and is known as **heteroscedasticity** or **heterogeneity of variance**. It's important to realize that Figure 6.13 (right) shows just one of many possible patterns of heteroscedasticity. The funnelling wouldn't have to be in the direction depicted, variances in errors might get smaller as the predictor increases, or be large at the extreme values of the predictor but small for middle values of the predictor, or indeed small at the extremes and large for middle values. The key thing is that the variance associated with errors is different for different values of the predictor(s).

What if we're predicting an outcome based on a categorical variable? Sticking with our concert theme, an audiologist was interested in the effects of loud concerts on people's hearing. She recruited people attending concerts of the loudest band in history, Manowar,[9] in London, Brighton, Bristol, Edinburgh, Newcastle, Cardiff and Dublin and measured for how many hours after the concert they had ringing in their ears. The top of Figure 6.14 shows the number of hours that each person (represented by a circle) had ringing in his or her ears after each concert. The asterisks show the predicted number of hours of ringing from the model (it's actually the mean for the concert). The plots on the left and right show similar predicted values but different *spreads* of the population data (circles) around their mean. To make this difference clearer, the bottom of Figure 6.14 removes the hypothetical population data and replaces them with a bar that shows the range of the dots from the top figure. In the left-hand plots these bars are similar lengths, indicating that the spread of population data (and, therefore, errors in prediction) around the model is roughly the same at each concert. This is not the case on the right-hand side of Figure 6.14: the spread of population data around the model (and therefore errors) is different at each concert. For example, the spread of population data after the London concert is small (the vertical distance from the lowest score to the highest score is small), but the population data for the Dublin show are very spread out around the mean (the vertical distance from the lowest score to the highest score is large). The uneven spread of scores is easiest to see if we look at the bars in the lower right-hand plot. Basically, even with categories we're looking for a similar pattern: we want the spread of errors around the model to be roughly the same for all values

9 Before they realized that it's a bad idea to encourage bands to be loud, the *Guinness Book of World Records* cited a 1984 Manowar concert as the loudest. Before that Deep Purple held the honour for a 1972 concert of such fearsome volume that it rendered three members of the audience unconscious.

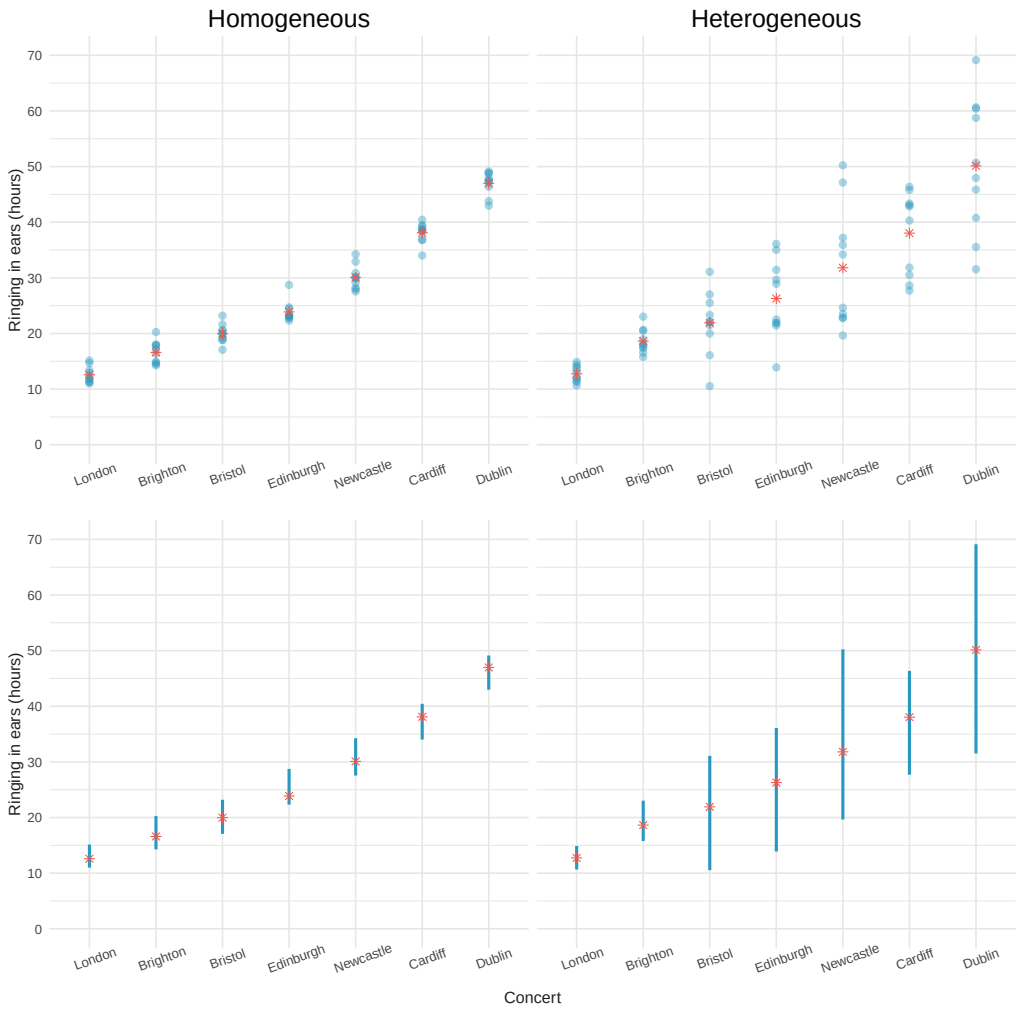


Figure 6.14 Plots illustrating group data with homogeneous (left) and heterogeneous (right) variances

of the predictor. When the predictor is made up of categories this means we want the spread to be the same in each category.

Finally, remember that the assumption applies to *population errors*, not to your sample data, or indeed the residuals of the sample model. However, the population errors are unobservable, which leaves us with a dilemma over how to know whether the assumption is met. Luckily, as we saw from Section 6.2.4 the form of the model is similar in samples to the population. In Figure 6.2, for example, when the spread of scores (and, therefore, errors) around the population model was consistent for all values of the predictor (i.e. homogeneity), this was also true for the samples. Figure 6.15 replicates this diagram but for a population that exhibits heteroscedasticity. Again, I have taken four random samples of 100 scores and fitted the models. Note that in the population the red dashed lines that mark out the edges of the data funnel out, and that this pattern is replicated in the samples. Therefore, if the sample residuals exhibits the characteristics of homogeneity or heteroscedasticity it's likely that the population errors do too. Therefore, we can use the residuals in the sample as an indicator of whether the assumption is met, even though the assumption itself relates to population errors.

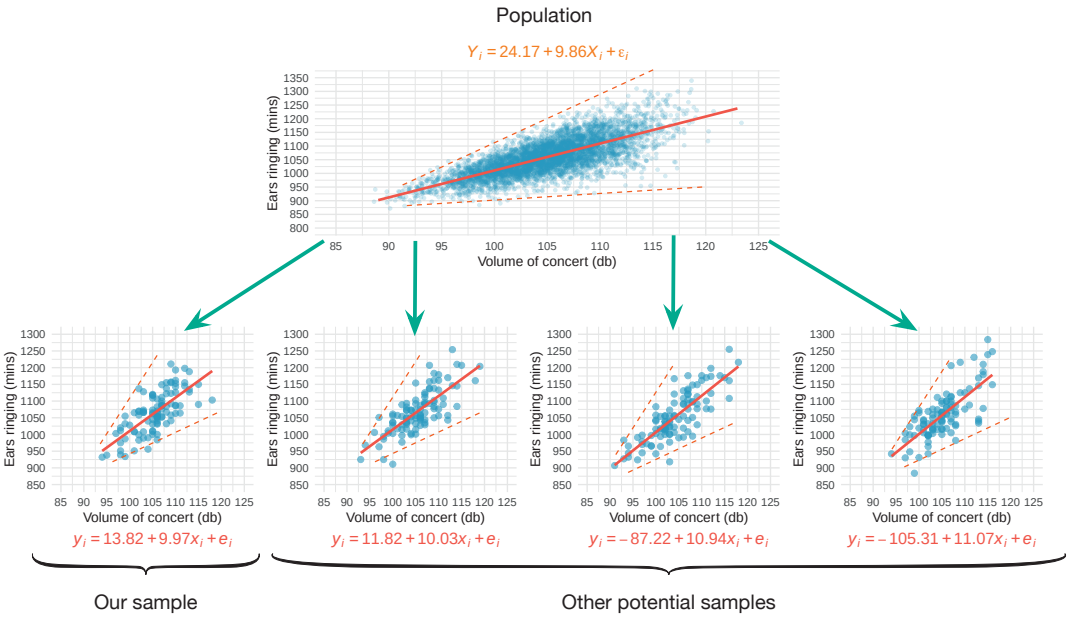
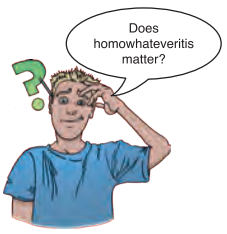


Figure 6.15 We can use sample data to gain insight into population heteroscedasticity

6.7.2 When does homoscedasticity matter? Ⓑ



In Section 6.2.8 I mentioned that the matrix $\mathbf{\Phi}$ in Eq. 6.29 is used to calculate the sampling variances (and covariances) of the model parameters (the *b*s), but we can do this only when the assumption of homoscedasticity holds. If the assumption is not true then we shouldn't use this matrix, but it gets used anyway meaning that we end up basing the variance of the sampling distribution associated with the *b*s in the model on incorrect values. The standard errors associated with the *b*s are based on these variances, which means that these will be incorrect too. Confidence intervals and test statistics use the standard error in their computation so these will be wrong also. In short, then, if the assumption is violated then the standard error, confidence interval and significance test associated with a given parameter estimate (*b*) will be inaccurate (see Table 6.1). Confidence intervals can be 'extremely inaccurate' when homogeneity of variance/homoscedasticity cannot be assumed (Wilcox, 2010). Therefore, if you want to look at the confidence intervals around your model parameter estimates or to test the significance of the model or its parameter estimates then homogeneity of variance matters. Some test statistics are designed to be accurate even when this assumption is violated, and we'll discuss these in the appropriate chapters.

With respect to estimating the parameters themselves (the *b*s), the matrix $\mathbf{\Phi}$ is not used to compute these, so even when the assumption is violated the method of least squares will produce 'unbiased' estimates. However, these estimates are not efficient, which means there are other ways to estimate the $\hat{b}$s that produce estimates with less error – in particular, by using **weighted least squares** (Hayes & Cai, 2007) in which each case is weighted by a function of its variance. If all you care about is estimating the parameters of the model in your sample then you don't need to worry about homogeneity of variance in most cases: the method of least squares will produce unbiased estimates (Hayes & Cai, 2007).

- **Key point:** In the absence of homoscedasticity (homogeneity of variance) standard errors, confidence intervals and significance tests will be inaccurate.

- **Key point:** In the absence of homoscedasticity (homogeneity of variance) the method of least squares produces unbiased estimates of the model parameters b, but alternative methods will produce estimates with less error.

6.7.3 Independent errors Ⓑ

The technical explanation in Section 6.2.8 highlighted that this condition relates to population errors being independent (uncorrelated). To get some intuition about when this assumption might be violated in practice, imagine Paul and Julie were participants in an experiment where they had to indicate whether they remembered having seen particular photos. If Paul and Julie were to confer about whether they'd seen certain photos then their answers would *not* be independent: Paul's response to a given question would have plausibly been influenced (if only to a tiny degree) by Julie's answer. We know already that if we estimate a model to predict their responses, there will be error in those predictions, and because Paul and Julie's scores are not independent, the errors associated with these predicted values will also not be independent. If Paul and Julie were unable to confer (if they were locked in different rooms) then the error terms should be independent (unless they're telepathic): the error in predicting Paul's response should not be influenced by the error in predicting Julie's response. Therefore, this assumption will always be violated in situations where you collect data from the same entities over time (i.e. time series, repeated-measures designs, longitudinal designs, growth models, etc.) because an entity's response at one point in time should correlate with their response at a different point in time.

The consequences of violating this assumption are the same as for homoscedasticity because both assumptions relate to the matrix $\mathbf{\Phi}$ (see above). To solution is to fit statistical models that factor in dependence between errors (see Chapter 21) rather than using OLS estimation.

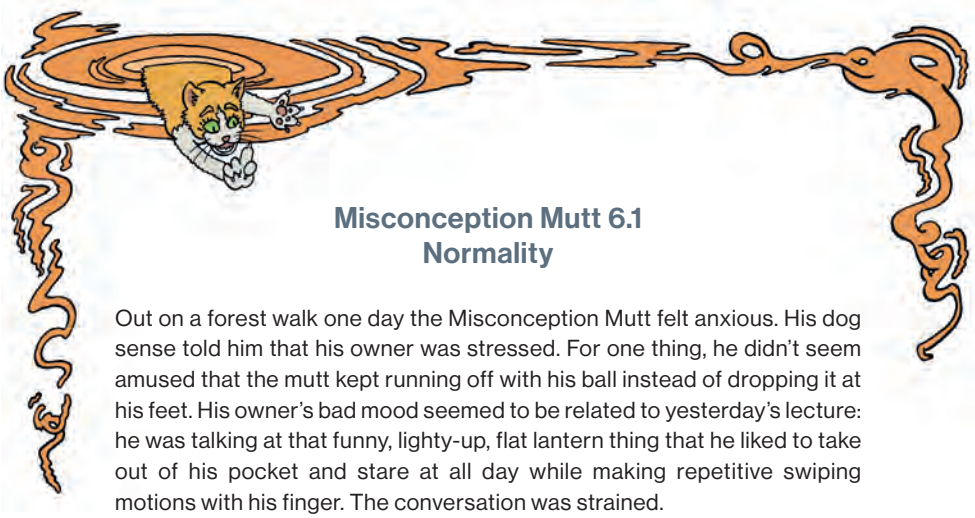
6.8 Normally distributed something or other Ⓑ

The least important condition relates to the normal distribution. This assumption might be one of the most misunderstood things in statistics. It's plausible that this misunderstanding is the result of fools like me explaining it badly. Many people wrongly take the 'assumption of normality' to mean that the data need to be normally distributed (Misconception Mutt 6.1). This misunderstanding probably comes from the fact that if the outcome is normally distributed then the model errors will be too, so the outcome variable can be a proxy of the thing that matters. Again, the technical detail is in the first few sections of this chapter, so here we'll try to develop your intuition about the two ways that normality comes into play.

6.8.1 Normality, sampling distributions, the standard error, confidence intervals and hypothesis tests Ⓑ

We know that parameter estimates vary from sample to sample, which creates uncertainty about the particular estimates we obtain because they would be slightly different had we taken a different sample. It is important to quantify this uncertainty, and we do this using the standard error, which tells us *how* variable parameter estimates are across different samples. Essentially, the standard error tells us the width of the sampling distribution: if it's wide (large standard error) estimates vary a lot across samples, if it's narrow (small standard error) estimates vary less across samples. To know the standard error of the parameter estimates from the linear model we need to know how they are distributed. If we can assume that parameter estimates follow a normal sampling distribution then we can use an established equation for the variance of this distribution, and hence know the standard error (which is the square root of variance of the sampling distribution). In Section 6.2.6 we saw that knowing an

equation for the standard error allows us to estimate it using the observed data, and this estimate is then used to compute confidence intervals and test statistics. If the sampling distribution is not normal, then the standard error and resulting confidence intervals and test statistics will be inaccurate because we've estimated them using an equation based on a normal sampling distribution.



Misconception Mutt 6.1
Normality

Out on a forest walk one day the Misconception Mutt felt anxious. His dog sense told him that his owner was stressed. For one thing, he didn't seem amused that the mutt kept running off with his ball instead of dropping it at his feet. His owner's bad mood seemed to be related to yesterday's lecture: he was talking at that funny, lighty-up, flat lantern thing that he liked to take out of his pocket and stare at all day while making repetitive swiping motions with his finger. The conversation was strained.

'I don't get it though,' his owner said into the lantern. 'What is the assumption of normality?'

The mutt wanted to help. He had enjoyed his owner taking him along to statistics lectures: he got a lot of strokes, but also seemed to be learning statistics too. He wanted to help.

'It means that your data need to be normally distributed', the mutt said. His owner stopped shouting at the lantern briefly to wonder why his dog was suddenly whining.

A nearby puddle started to ripple. The mutt turned in time to see some ginger ears appearing from the water. He sighed a depressed sigh.

Having emerged from the puddle, the Correcting Cat sauntered over. He quietly biffed the dog's nose.

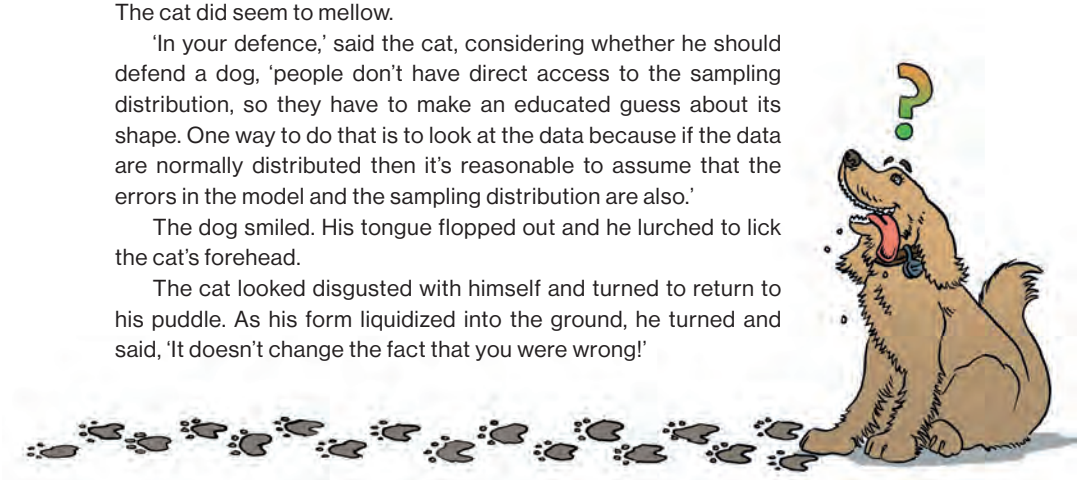
'No,' the cat purred. 'The assumption of normality refers to the errors of the model being normally distributed, or the sampling distribution of the parameter, not the data themselves.'

The dog offered his paw – maybe that would appease his persecutor. The cat did seem to mellow.

'In your defence,' said the cat, considering whether he should defend a dog, 'people don't have direct access to the sampling distribution, so they have to make an educated guess about its shape. One way to do that is to look at the data because if the data are normally distributed then it's reasonable to assume that the errors in the model and the sampling distribution are also.'

The dog smiled. His tongue flopped out and he lurched to lick the cat's forehead.

The cat looked disgusted with himself and turned to return to his puddle. As his form liquidized into the ground, he turned and said, 'It doesn't change the fact that you were wrong!'



1 **Key point**: For standard errors, confidence intervals and test statistics associated with a parameter estimate (whether that's a mean, or a $\hat{b}$ in a linear model) to be accurate, the parameter estimate must come from a normal sampling distribution.

It turns out that linear model parameters follow a normal sampling distribution in two situations: (1) when model errors are normally distributed (Section 6.2.6); and (2) in large samples (Section 6.2.7). These observations lead to two other key points.

2 **Key point**: If we have evidence that sample residuals are normally distributed, we can assume that population errors are also normally distributed, and that the sampling distribution of model parameter estimates will also be normal**.**

3 **Key point**: In large samples the sampling distribution of model parameter estimates will be normal so we can ignore this assumption (Section 6.2.7).

6.8.2 Normality and parameter estimates Ⓑ

Focusing on the population errors again, when they are normally distributed the OLS estimate of a parameter is said to be *efficient*, which in statistical terms means that it is the estimate with the minimum variance among all estimators (see Table 6.1). Basically, there are ways other than the method of least squares to estimate parameters. When the population errors are normally distributed, the OLS estimate is the method with the least 'error' given the data. In this situation it is also equivalent to the maximum likelihood estimate.

4 **Key point**: When the population errors are normally distributed, the method of least squares produces parameter estimates that are optimal. When the population errors are not normally distributed the OLS estimates are still valid, but there will be other estimation methods that produce parameter estimates with less variance (i.e. less error given the data).

6.8.3 When does the assumption of normality matter? Ⓑ

I've said that assumptions involving normality are the least important. Let's sum up why.



1 **Parameter estimation does not require normality**: For the OLS estimates of model parameters (the $\hat{b}$s) to be optimal using the method of least squares, the errors in the population must be normally distributed. However, even when population errors are not normally distributed the method of least squares will always give you an estimate of the model parameters that minimizes error and is unbiased. In this sense, you don't need to assume normality of anything to fit a linear model and estimate the parameters that define it (Gelman & Hill, 2007).

2 **Normality doesn't matter in large samples**: Estimation of the standard error requires that the $\hat{b}$s have a normal sampling distribution. By extension, confidence intervals and test statistics (which use the standard error) also require that the $\hat{b}$s have a normal sampling distribution. The central limit theorem tells us that this will be the case in large samples (Section 6.2.7). There is a question around how large is 'large'. The extent to which test statistics perform as they should do in large samples varies across different test statistics, and we will deal with these idiosyncratic issues as they arise.

3 **Normality doesn't matter when residuals are normally distributed**: In small samples we can't be sure that the $\hat{b}$s have a normal sampling distribution. However, the model residuals act as a reasonable proxy for the model errors, and if the model errors are normally distributed

then the $\hat{b}$s have a normal sampling distribution. So, if the sample residuals are normal, then we can move on with our lives.

To sum up, then, if all you want to do is estimate the parameters of your model then normality of population errors matters mainly in deciding how best to estimate them. For practical purposes this is largely irrelevant because OLS estimates are still valid even when population errors are not normally distributed. If you want to construct confidence intervals around those parameters, or compute significance tests relating to those parameters, then the assumption of normality matters in small samples, but because of the central limit theorem we don't really need to worry about this assumption in larger samples (but see Jane Superbrain Box 6.4). In smaller samples, we can check the normality of model residuals (as a proxy for the population errors) because if these are normal it implies that the sampling distribution of the model parameters is normal and that we can trust the confidence intervals and significance tests.

In practical terms, provided your sample is large, outliers are a more pressing concern than normality. Although we tend to think of outliers as isolated very extreme cases, you can have outliers that are less extreme but are not isolated cases. These outliers can dramatically reduce the power of significance tests (Jane Superbrain Box 6.5).



Jane Superbrain 6.5
Stealth outliers ©

We tend to think of outliers as one or two very extreme scores, but sometimes they soak themselves in radar-absorbent paint and contort themselves into strange shapes to avoid detection. These 'stealth outliers' (that's my name for them, no one else calls them that) hide undetected in data sets, radically affecting analyses. Imagine you collected happiness scores, and when you plotted the frequency distribution it looked like Figure 6.16 (left). You might decide that this distribution is normal, because it has the characteristic bell-shaped curve. However, it is not: it is a **mixed normal distribution** or **contaminated normal distribution** (Tukey, 1960). The happiness scores on the left of Figure 6.16 are made up of two distinct populations: 90% of scores are from humans, but 10% are from muppets (note: they live amongst us). Figure 6.16 (right) reproduces this overall distribution (the blue one), but also shows the unique distributions for the humans (red) and muppets (Kermit green) that contribute to it.

The human distribution is a perfect normal distribution, but the curve for the muppets is flatter and heavier in the tails, showing that muppets are more likely than humans to be extremely happy (like Kermit) or extremely miserable (like Statler and Waldorf). When these populations combine, the muppets contaminate the perfectly normal distribution of humans: the combined distribution (blue) has slightly more scores in the extremes than a perfect normal distribution (orange). The muppet scores have affected the overall distribution even though (1) they make up only 10% of the scores; and (2) their scores are more frequent at the extremes of 'normal' and not radically different like you might expect an outlier to be. These extreme scores inflate estimates of the population variance (think back to Jane Superbrain Box 1.5). Mixed normal distributions are very common and they reduce the power of significance tests – see Wilcox (2010) for a thorough account of the problems associated with these distributions.

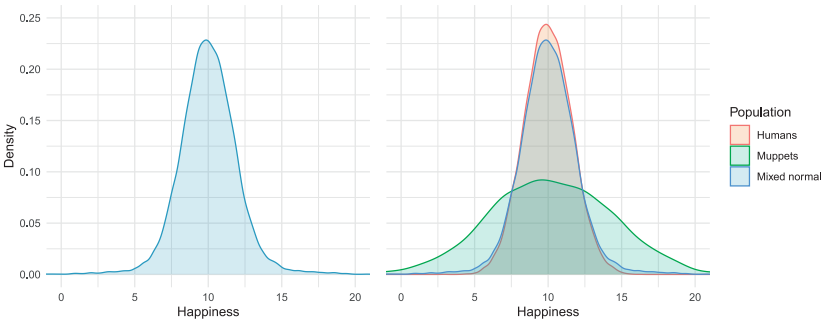


Figure 6.16 An apparently normal distribution (left), which is actually a mixed normal distribution made up of two populations (right)



6.9 Checking for bias and describing data Ⓑ

The sources of bias that we have discussed can be spotted in the data (outliers) or in the model residuals (assumptions). However, we have also seen that there is a relationship between the outcome variable and the model residuals. In Jane Superbrain Box 6.3 we noted two mathematically equivalent ways to express the linear model

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$
$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

and

$$Y_i \sim \mathcal{N}(b_0 + b_1 X_i, \sigma^2).$$

The first way says that the outcome can, on average, be predicted from a linear combination of predictor variables but there will be normally distributed error around that predicted value. The second way of expressing the model says that the predictions about the outcome are normally distributed such that, on average, predicted values are a linear combination of predictor variables but there is variance in those predicted values. They both say the same thing, but in the first version the errors are made implicit, whereas in the second they are implicit within the variance term (σ^2). The point is, if the outcome variable is normally distributed then the errors ought to be too, and by implication we can check this in the sample by looking at whether our observed values of the outcome are normally distributed (and by implication the model residuals). This wouldn't stop us from also looking at the model residuals but is a useful first step because if the outcome variable is not normal there are better models to fit than the OLS linear model. For example, if the outcome is heavily positively skewed we can model it with an exponential distribution using something called a generalized linear model, which we don't explore in this book.

So, we will look specifically at how to look for outliers, influential cases and violations of assumptions using the model residuals in Chapter 8 (after we have learnt about fitting models), but for now we'll look at methods to describe and summarize the variables in your model that can also alert us to sources of bias.

6.9.1 An insanitary example Ⓐ

A biologist was worried about the potential health effects of music festivals. She went to the Download Music Festival[10] (those of you outside the UK and not into metal music can pretend it is whatever festival you like to visit) and measured the hygiene of 810 randomly selected concert-goers over the three days of the festival. She tried to measure every person on every day but, because it was difficult to track people down, there were missing data on days 2 and 3. Hygiene was measured using the standardized technique of armpit licking. I joke, it was some sophisticated sci-fi gadget involving counting bacteria and allocating a score ranging between 0 (you smell like flatulent skunk) and 4 (you smell of sweet roses on a fresh spring day). I know from bitter experience that sanitation is not always great at these places (the Reading Festival seems particularly bad) and so the biologist predicted that personal hygiene would go down dramatically over the three days of the festival. The data can be found in **download.sav**.

6.9.2 Exploring data Ⓑ

A good way to quickly get a lot of information about your data is to select *Analyze* > *Descriptive Statistics* > Explore... to access the dialog boxes in Figure 6.17. Select and drag (or click ▸) the variables you want to summarize to the box labelled *Dependent List*. For this example, select the hygiene scores for the three days. Click Statistics... and select ☑ Descriptives and ☑ Outliers in the resulting dialog box. Click Plots... and select ☑ Histogram to produce histograms of each variable, and select ☑ Normality plots with tests to produce something called a Q-Q plot (Section 6.9.4). By default, IBM SPSS Statistics software will produce boxplots (split according to group if a factor has been specified), which is great. We need to click Options... to tell SPSS how to deal with missing values. This option is important because although we start off with 810 scores on day 1, by day two we have only 264 and by day 3 only 123. By default SPSS will use only cases for which there are valid scores on all selected variables. This would mean that for day 1, even though we have 810 scores, SPSS will use only the 123 cases for which there are scores on all three days. This is known as excluding cases *listwise*. However, we want it to use all of the scores it has on a given day, which is known as *pairwise* (SPSS Tip 6.1). To do this select ◉ Exclude cases pairwise. Having selected these options, click Continue to return to the main dialog box and OK to run the analysis. We will look at the resulting output over the next few sections.

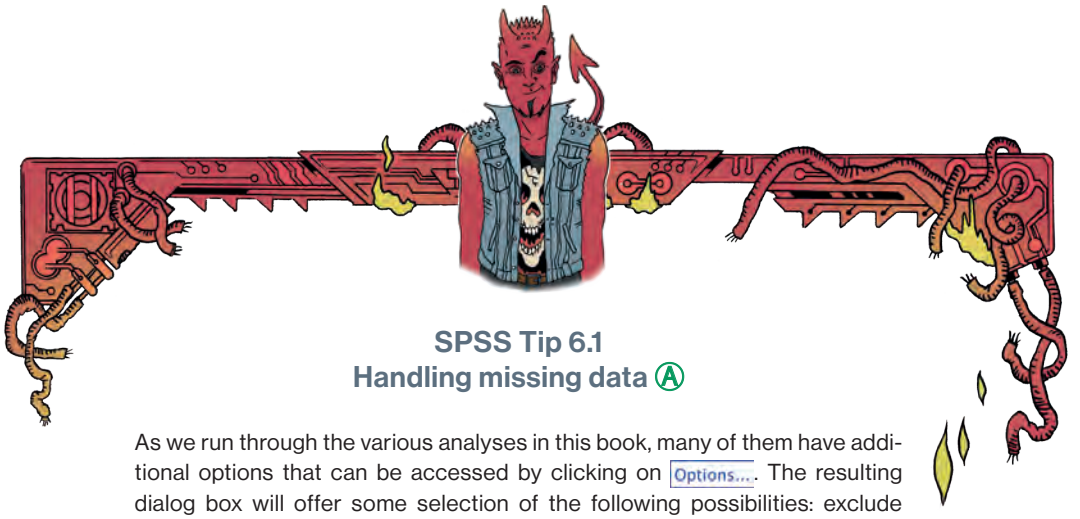
6.9.3 Spotting outliers Ⓑ

When they are isolated, extreme cases and outliers are fairly easy to spot using plots such as histograms, boxplots, and scatterplots; it is considerably trickier when outliers are more subtle (using *z*-scores may be useful – Jane Superbrain Box 6.6). The output from the *explore* command provides a lot of useful information for spotting outliers, but let's also plot a scatterplot of two of the variables.



Using what you learnt in Chapter 5, plot a scatterplot of the day 2 scores against the day 1 scores.

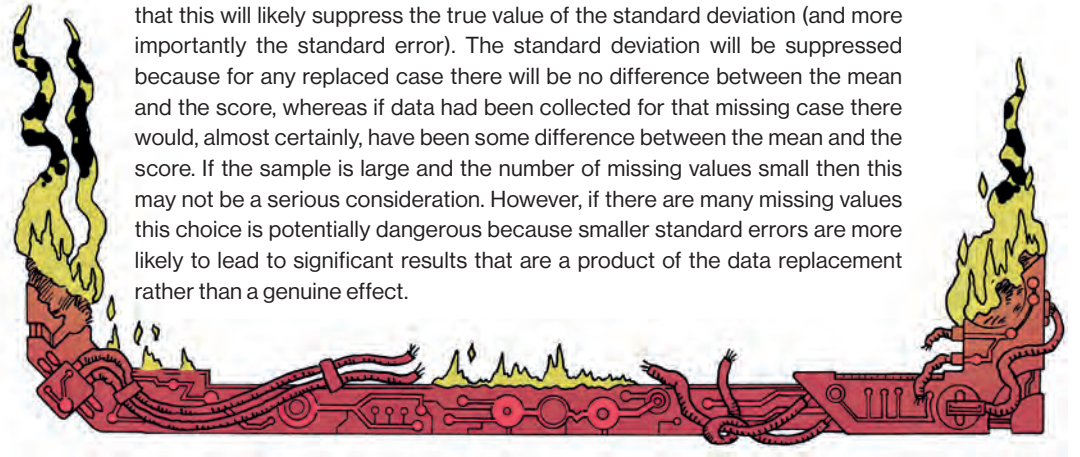
10 www.downloadfestival.co.uk



SPSS Tip 6.1
Handling missing data Ⓐ

As we run through the various analyses in this book, many of them have additional options that can be accessed by clicking on Options.... The resulting dialog box will offer some selection of the following possibilities: exclude cases 'pairwise', 'analysis by analysis' or 'listwise', and sometimes 'replace with mean'. Let's imagine we wanted to use our hygiene scores to compare mean scores on days 1 and 2, days 1 and 3, and days 2 and 3. First, we can exclude cases listwise, which means that if a case has a missing value for any variable, then the case is excluded from the whole analysis. So, for example, if we had the hygiene score for a person (let's call her Melody) at the festival on days 1 and 2, but not day 3, then Melody's data will be excluded for all of the comparisons mentioned above. Even though we have her data for days 1 and 2, we won't use them for that comparison – *they would be completely excluded from the analysis.* Another option is to excluded cases on a *pairwise* (a.k.a. *analysis-by-analysis* or *test-by-test*) basis, which means that Melody's data will be excluded only for analyses for which she has missing data: so her data would be used to compare days 1 and 2, but would be excluded for the other comparisons (because we don't have her score on day 3).

Sometimes SPSS will offer to replace the missing score with the average score for this variable and then include that case in the analysis. The problem is that this will likely suppress the true value of the standard deviation (and more importantly the standard error). The standard deviation will be suppressed because for any replaced case there will be no difference between the mean and the score, whereas if data had been collected for that missing case there would, almost certainly, have been some difference between the mean and the score. If the sample is large and the number of missing values small then this may not be a serious consideration. However, if there are many missing values this choice is potentially dangerous because smaller standard errors are more likely to lead to significant results that are a product of the data replacement rather than a genuine effect.



The resulting scatterplot is shown in Figure 6.18 (top). The first thing that should leap out at you is that there is one case that is very different from the others. All the scores are squashed up in a cloud at one end of the distribution except for one, which is floating about to the far right. This very different data point should make our outlier senses tingle. This strange score is someone whose hygiene score was over 20 on day one, but about 2 on day 2. Hang on, a score of over 20 on day 1? Didn't the hygiene rating scale range from 0 to 4?

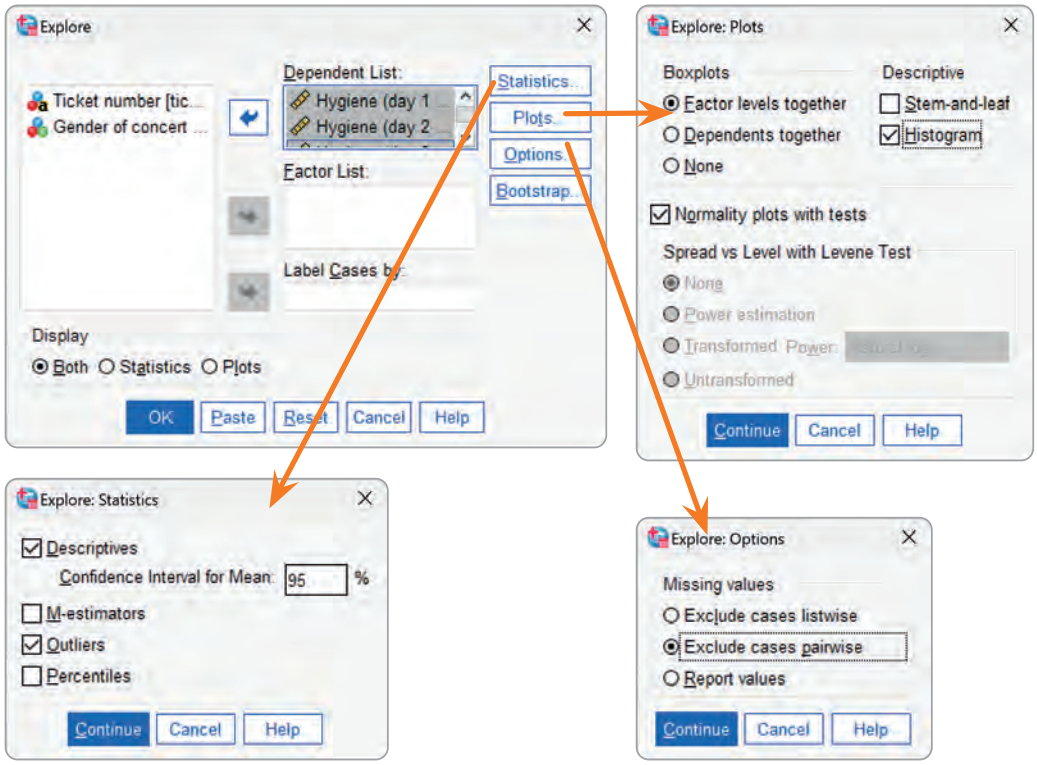


Figure 6.17 Dialog boxes for the *explore* command

The output from the *explore* command will include a histogram (Figure 6.18, middle) and boxplot (Figure 6.18, bottom). Confirming what we noticed on the scatterplot, it's obvious that one case is very different from the others. On the histogram all the scores are squashed up at one end of the distribution because they are less than 5 (yielding a very pointy distribution) except for one, which has a value of 20. On the boxplot, the outlier shows up as an extreme score (*). SPSS helpfully tells us the row number (611) of this outlier. Output 6.1 also lists the top and bottom five most extreme scores for each variable. This tells us that row 611 in the data has a score of 20.02. We go back and check the printouts for case 611 from our hygiene measuring device, and sure enough it says 2.02 for day 1. Whoever entered the data has mistyped 2.02 as 20.02. Having checked the raw data, we should replace the value 20.02 with 2.02 before continuing. To make this change click  in the data editor and type 611 in the resulting dialog box, which will transport you to row 611, where you can change the value. No other values in Output 6.1 seem extreme enough to warrant further scrutiny.



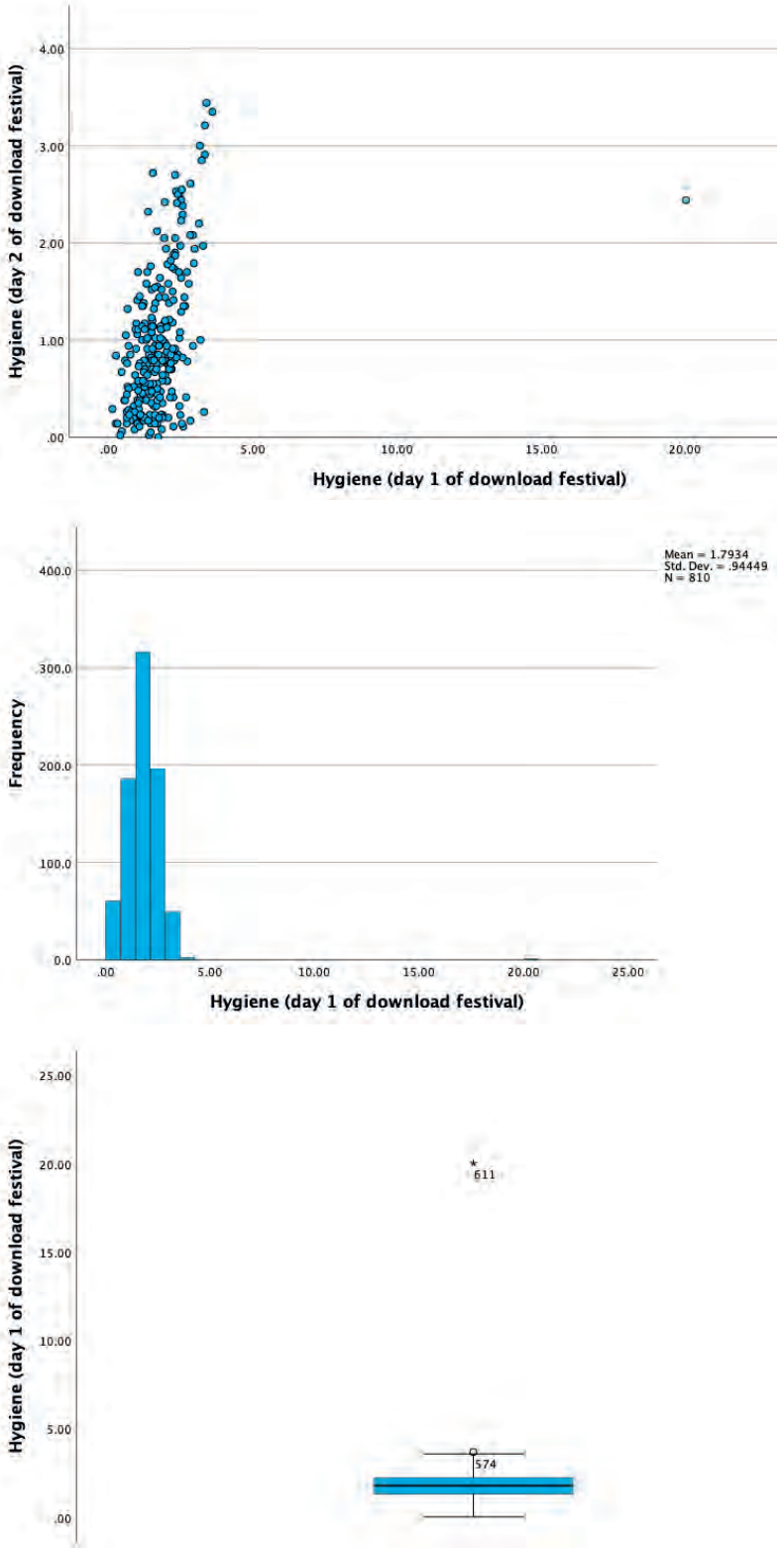


Figure 6.18 Scatterplot (top), histogram (middle) and boxplot (bottom) of hygiene scores on day 1 of the Download Festival

Extreme Values

			Case Number	Value
Hygiene (day 1 of download festival)	Highest	1	611	20.02
		2	574	3.69
		3	303	3.58
		4	657	3.44
		5	300	3.41
	Lowest	1	595	.02
		2	366	.05
		3	177	.11
		4	41	.11
		5	569	.23[a]
Hygiene (day 2 of download festival)	Highest	1	774	3.44
		2	303	3.35
		3	488	3.21
		4	476	3.00
		5	623	2.91
	Lowest	1	734	.00
		2	613	.02
		3	15	.02
		4	191	.05
		5	423	.06
Hygiene (day 3 of download festival)	Highest	1	774	3.41
		2	618	3.02
		3	623	3.02
		4	153	2.70
		5	785	2.55
	Lowest	1	631	.02
		2	31	.02
		3	694	.08
		4	164	.11
		5	377	.14[b]

a. Only a partial list of cases with the value .23 are shown in the table of lower extremes.

b. Only a partial list of cases with the value .14 are shown in the table of lower extremes.

Output 6.1

Figure 6.19 shows the scatterplot, histogram and boxplot for the data after the extreme case has been corrected. The scatterplot (top) looks a bit more like we'd expect, with dots scattered the length of *x*- and *y*-axes. It also looks like it might be reasonable to summarize the relationship between day 1 and day 2 hygiene scores with a linear model because there isn't any curvature to the pattern of dots. Focusing on the day 1 scores, the histogram (middle) shows that the day 1 scores now look normal: the plot is symmetrical and doesn't seem too pointy or flat. The boxplot (bottom) shows this symmetry too because the whiskers at the top and bottom are similar lengths, and the median (the dark horizontal line) sits about halfway up the box. None of these plots suggest further extreme scores: the boxplot suggests that case 574 is a mild outlier, but the other plots suggest this case isn't a concern because they do not indicate any unusual cases. We can also identify outliers numerically using *z*-scores, as described in Jane Superbrain Box 6.6.

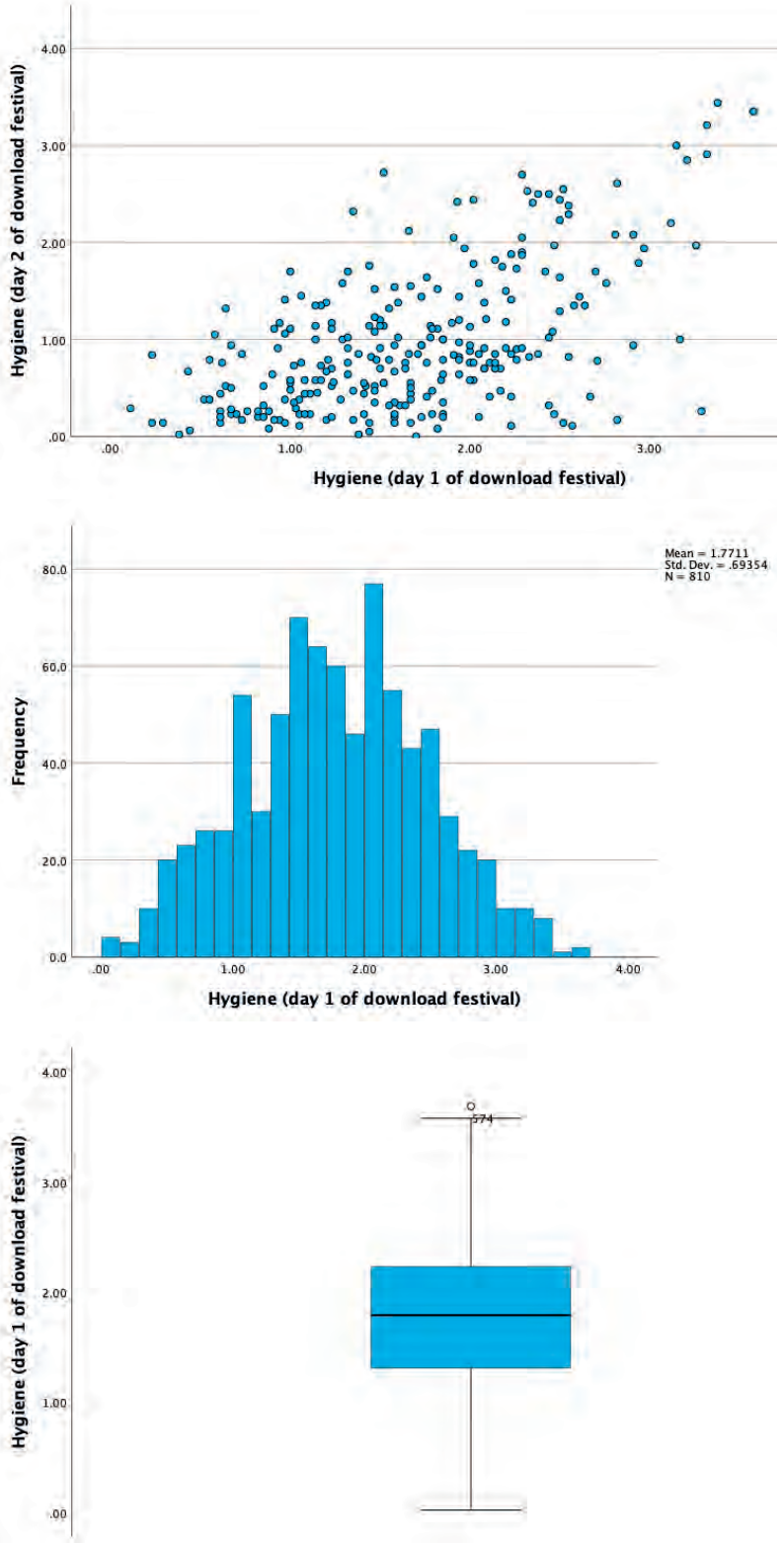


Figure 6.19 Scatterplot (top), histogram (middle) and boxplot (bottom) of hygiene scores on day 1 of the Download Festival after removing the extreme scores



Jane Superbrain 6.6
Using *z*-scores to find outliers ©

We saw in Section 1.7.6 that *z*-scores express scores in terms of a distribution with a mean of 0 and a standard deviation of 1. By converting our data to *z*-scores we can use benchmarks that we can apply to any data set to search for outliers. Activate the *Analyze* > *Descriptive Statistics* > Descriptives... dialog box, select the variable(s) to convert (such as day 2 of the hygiene data) and tick ☑ Save standardized values as variables (Figure 6.20). SPSS will create a new variable in the data editor (with the same name as the selected variable, but prefixed with *Z*).

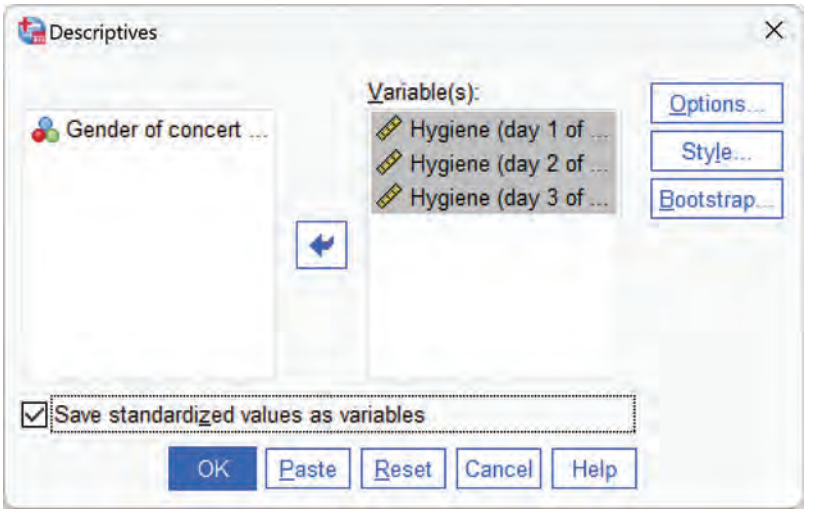


Figure 6.20 Saving *z*-scores

To look for outliers we can count how many *z*-scores fall within certain important limits. If we ignore whether the *z*-score is positive or negative (called the 'absolute value'), then in a normal distribution we'd expect about 5% to be greater than 1.96 (we often use 2 for convenience), 1% to have absolute values greater than 2.58, and none to be greater than about 3.29. To get SPSS to do the counting for you, use the syntax in the workbook **outliers.spwb** (on the companion website), which will produce a table for day 2 of the Download Festival hygiene data. Load this file and run the syntax. The first three lines use the *descriptives* function on the variable **day_2** to save the *z*-scores in the data editor (as a variable called **Zday_2**).

```
DESCRIPTIVES
VARIABLES= day2/SAVE.
EXECUTE.
```

Next, we use the *compute* command to change **Zday_2** so that it contains the absolute values (i.e. converts all minus values to plus values).

```
COMPUTE Zday_2 = abs(Zday_2).
EXECUTE.
```

The next commands recode the variable **Zday_2** so that if a value is greater than 3.29 it's assigned a code of 1, if it's greater than 2.58 it's assigned a code of 2, if it's greater than 1.96 it's assigned a code of 3 and if it's less than 1.95 it gets a code of 4.

```
RECODE
Zday_2 (3.29 thru highest = 1)(2.58 thru highest = 2)(1.96 thru highest = 3)
(Lowest thru 1.95 = 4).
EXECUTE.
```

We then use the value labels command to assign helpful labels to the codes we defined above.

```
VALUE LABELS Zday_2
4 'Normal range' 3 'Potential Outliers (z > 1.96)' 2 'Probable Outliers
(z > 2.58)' 1 'Extreme (z-score > 3.29)'.
```

Finally, we use the *frequencies* command to produce a table (Output 6.2) telling us the percentage of 1s, 2s, 3s and 4s found in the variable **Zday_2**.

```
FREQUENCIES
VARIABLES= Zday_2
/ORDER=ANALYSIS.
```

Thinking about what we know about the absolute values of *z*-scores, we would expect to see only 5% (or less) with values greater than 1.96, 1% (or less) with values greater than 2.58, and very few cases above 3.29. The column labelled *Cumulative Percent* tells us the corresponding percentages for the hygiene scores on day 2: 0.8% of cases were above 3.29 (extreme cases), 2.3% (compared to the 1% we'd expect) had values greater than 2.58, and 6.8% (compared to the 5% we would expect) had values greater than 1.96. The remaining cases (which, if you look at the *Valid Percent*, constitute 93.2%) were in the normal range. These percentages are broadly consistent with what we'd expect in a normal distribution (around 95% were in the normal range).

Zscore: Hygiene (day 2 of download festival)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Extreme (z-score > 3.29)	2	.2	.8	.8
	Probable Outliers (z > 2.58)	4	.5	1.5	2.3
	Potential Outliers (z > 1.96)	12	1.5	4.5	6.8
	Normal range	246	30.4	93.2	100.0
	Total	264	32.6	100.0	
Missing	System	546	67.4		
Total		810	100.0		

Output 6.2



6.9.4 Q-Q and P-P plots Ⓐ

The *explore* command will also produce something called a **Q-Q plot** (quantile–quantile plot), which is a useful way to look at the distributional properties of the variables we have measured. The Q-Q plot displays the quantiles of the observed data against the quantiles you'd expect to get from a particular distribution (e.g., a normal distribution). There is a similar plot known as a **P-P plot** (probability–probability plot). These plots are sometimes confused, but that's OK because they are interpreted in the same way. Before we look at the Q-Q plots produced by the *explore* command for our hygiene scores, we'll dig into what a Q-Q plot is.

Let's look at an example. A study investigating mate choice used an innovative methodology in which participants engaged in a 'choose your own adventure' style activity where they read a story in which they were entering a new romantic relationship (Joel & Charlot, 2022). At 17 points in the narrative a potential relationship dealbreaker was introduced and the participant indicated whether they would continue with the relationship. Dealbreakers were information around common sources of relationship incompatibility (e.g., desire to have children, hygiene, dependability, exclusivity, political orientation, sex drive). The researchers were interested in how many dealbreakers it took for participants to choose not to pursue the relationship. Figure 6.21 shows two sets of 12 (fictional) scores from a study such as this. Both data sets range from 1 to 17 (i.e., at least one person quit the relationship at the first dealbreaker and at least one person pursued the relationship until the bitter end). The scores in between have very different distributions. The first data set looks about as normal as you might expect 12 scores to be (it's vaguely symmetrical around a central value), but the second looks heavily skewed and non-normal. How can we use quantiles to assess whether these scores are likely to be normally distributed?

Remember that quantiles are values that cut a distribution into equal portions (Section 1.7.5). To try to keep things straightforward, we'll use quantiles with which you are familiar: quartiles. A quick reminder that quartiles are quantiles that split the data into four equal parts; that is, they are values that split the data into four portions each of which contains 25% of the scores. The lowest 25% of scores fall below the value of the lower quartile ($Q_{0.25}$), the next largest 25% of scores fall between the lower quartile and the median ($Q_{0.50}$), the next 25% of scores fall between the median and the value of the upper quartile ($Q_{0.75}$) and the top 25% of scores fall above the upper quartile. We have two data sets each containing 12 scores. To divide 12 scores into four equal portions, the quartiles would be the values that divide the data into groups of three scores when the scores are arranged in ascending order. Figure 6.21 shows the quartiles for the two data sets. For the first data set the lower quartile sits between the scores of 6 and 8 so we take its value to be the midpoint between these scores[11] (i.e., $X_{Q_{0.25}} = 7$), the median sits between two scores of 10 and so is also $X_{Q_{0.5}} = 10$ and the upper quartile sits between two scores of 13 and so is also $X_{Q_{0.75}} = 13$. For the second data set we get quartiles of X_Q =2, 4 and 10. My fictional data have deliberately straightforward values.

For each data set we have three quantiles ($Q_{0.25}$, $Q_{0.50}$ and $Q_{0.75}$). If we want to see whether our distributions match a normal distribution we could ask the question 'what are the equivalent quantiles from a standard normal distribution?' We can find this out by looking at the value of z that cuts off the lower 25% of scores, the lower 50% of scores and the lower 75% of scores in the table in Appendix A.1. This table includes only positive values of z (the values above the centre of the distribution). First off, the value that cuts the distribution into two equal parts is $z_{Q_{0.5}} = 0$ (the centre). We can see this by looking at the first row of the table and seeing that $z = 0$ corresponds to both the larger and smaller portion under the curve being equal to 0.5 (or 50%). If we look down the column labelled *Larger Portion* until we find 0.75, then the corresponding value of z will be the value that separates the

11 There are at least nine different ways to compute a quantile, but I've chosen the simplest and most intuitive one!

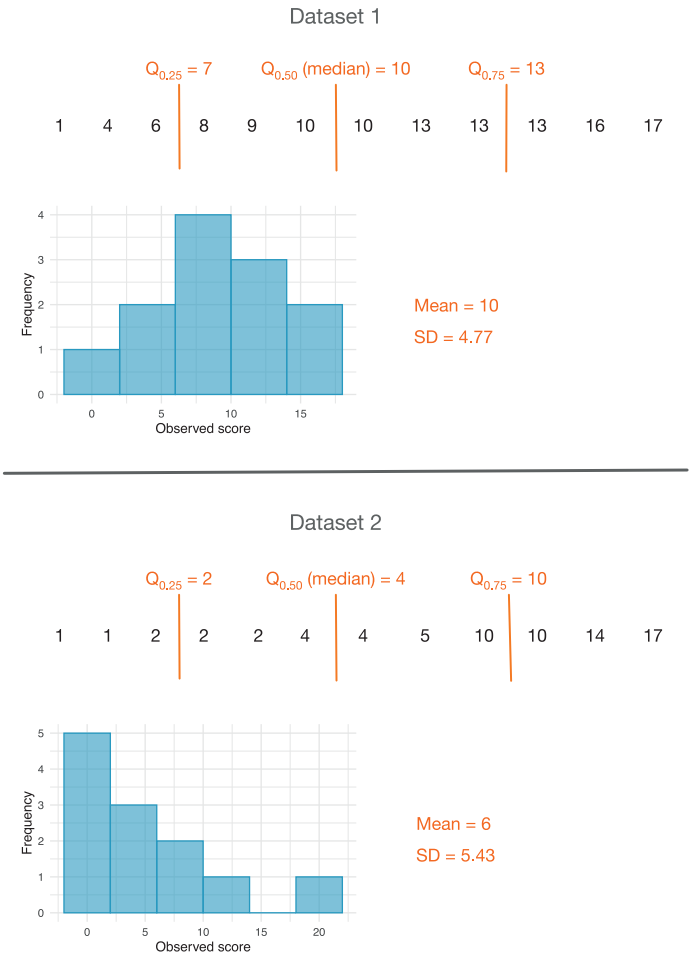


Figure 6.21 Quartiles of two data sets

bottom 75% of scores from the top 25%; in other words it gives us the upper quartile ($Q_{0.75}$). The closest value (which will do for our purposes) is when the larger portion = 0.75175, and we get $z_{Q_{0.75}} = 0.68$. The value of z that does the reverse and cuts off the bottom 25% from the top 75% is the corresponding negative value, so the lower quartile is $z_{Q_{0.25}} = -0.68$.

We know what the quartiles are for a standard normal distribution, and we know what the quartiles are for our data, the final thing we need to know is the values of the quartiles that we *would* get if the data were normal. To do this we need to convert our z-score quantiles into the same scale as our raw data by rearranging an equation we met in Chapter 1:

$$z = \frac{X - \bar{X}}{s}$$
$$X = \bar{X} + (s \times z) \qquad (6.31)$$

In which $\bar{X}$ and s are the mean and standard deviation of a set of observations, and z is a value from the standard normal distribution. Using this equation allows us to express the quartiles from a standard normal distribution in the units of measurement of our data (rather than as z-scores). We'll denote these values by X_{z_Q}. Figure 6.21 shows that the first data set has a mean of $\bar{X} = 10$ and standard deviation of $s = 4.77$, whereas the second data set has a mean of $\bar{X} = 6$ and standard deviation of $s = 5.43$. Therefore, the lower quartile of the first data set expressed in the same units as the observations is

$$\begin{aligned} X_{Z_Q} &= \bar{X} + \left(s \times z_{Q_{0.25}}\right) \\ &= 10 + \left(4.77 \times -0.68\right) \\ &= 6.76. \end{aligned} \tag{6.32}$$

Similarly, the upper quartile of the second data set expressed in the same units as the observations is

$$\begin{aligned} X_{Z_Q} &= \bar{X} + \left(s \times z_{Q_{0.75}}\right) \\ &= 6 + \left(5.43 \times 0.68\right) \\ &= 9.69. \end{aligned} \tag{6.33}$$

Table 6.2 completes the calculations, and shows the quartiles from the standard normal distribution (Z_Q), those quartiles expressed in the units of measurement of the raw data (X_{z_Q}) and the quartiles from the observed scores (X_Q). Figure 6.22 shows what happens if we now plot the values from Table 6.2. The red symbols plot the quartiles from the standard normal distribution (Z_Q) against those quartiles expressed in the units of measurement of the raw data (X_{z_Q}), and because these represent the same values but expressed as either *z*-scores or unstandardized units, they correlate perfectly. This perfect association is shown by the diagonal line, which connects the three points. What we can take from this is that if we plot the quantiles from a normal distribution against themselves (but expressed in unstandardized units) we'd get a perfect diagonal line. This line therefore represents the expected quantiles from a normal distribution. The blue symbols plot the quartiles from the standard normal distribution (Z_Q) against the values of the quartiles that we observed in the data (X_Q). In data set 1 (Figure 6.22, left) the blue symbols sit close to the diagonal, in fact the median sits on the line. The fact the blue points are close to the diagonal line (and the corresponding red points) shows that the quantile values in the data (blue points) are almost the same as the quantile values we'd expect to get from a normal distribution (red dots and the diagonal line). In other words, the observed quantiles are consistent with those from a normal distribution. In data set 2 (Figure 6.22, right) not all of the blue symbols sit close to the diagonal; the median in particular is much further left (on the *x*-axis) than we'd expect from a normal distribution. (The median in a normal distribution gives as a raw score of 6 but we observe a median of 4.) The fact that some blue points deviate substantially from the diagonal line (and the corresponding red points) shows that the quantile values in the data are dissimilar to the quantile values we'd expect to get from a normal distribution (the diagonal line). In other words, the observed quantiles are inconsistent with those from a normal distribution. In short, data set 1 seems consistent with a normal distribution whereas data set 2 does not, which confirms what we suspected from the histograms in Figure 6.21.

Table 6.2 Quartiles from a normal distribution and two data sets

Quartile		Data set 1		Data set 2	
	Z_Q	X_Q	$X_{z_Q} = \bar{X} + (s \times z_Q)$	X_Q	$X_{z_Q} = \bar{X} + (s \times z_Q)$
$Q_{0.25}$	−0.68	7	6.76	2	2.31
$Q_{0.50}$	0.00	10	10.00	4	6.00
$Q_{0.75}$	0.68	13	13.24	10	9.69

We looked at plotting the quartiles (just three values) to keep things straightforward, but imagine that instead of plotting only quartiles, we converted every observed value to a quantile and from this obtained the corresponding value from a standard normal distribution. We'd still get a diagonal line representing the expected quantiles, but we'd get a lot more dots representing the observed quantiles.

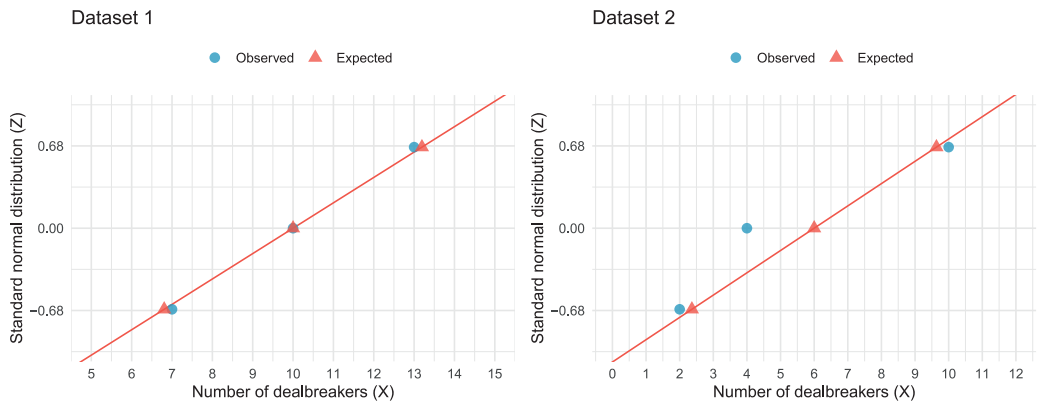


Figure 6.22 Observed and expected quartiles of two data sets

By looking at the extent to which these dots deviate from the diagonal line we can ascertain the extent to which the observed quantiles deviate from those expected from a normal distribution. If the dots fall along the diagonal line it suggests the observed quantiles are consistent with those from a normal distribution. A P-P plot uses a similar idea but instead of plotting quantiles it plots the cumulative probability of each value against the corresponding cumulative probability from a normal distribution. The cumulative probability is related to the quantile so it's essentially a different way of expressing the same information, but if you have a lot of scores Q-Q plots can be easier to interpret than P-P plots because they display fewer values.

Figure 6.23 shows a normal distribution and some characteristic deviations from normal alongside the corresponding Q-Q plots.[12] You can see from the top plots that when scores are normally distributed the dots on the Q-Q plot follow the diagonal line that represents perfect normality. If scores are positively skewed then the points of the Q-Q plot point upwards in the middle, and for positive skew the opposite is true: points sag downwards in the middle. When there are too many scores in the tails (a heavy-tailed distribution) the points form an S with points rising above the diagonal line at the left of the *x*-axis and falling below it at the right. The points also form an S when there are too few scores in the tails (a light-tailed distribution) but the pattern is reversed with dots falling below the diagonal at the left of the *x*-axis and rising above it at the right.

6.9.5 Spotting normality Ⓐ

One way we can assess the normality of our variables is, as we have just seen, a Q-Q plot. Now we know what a Q-Q plot is, we can look at the plots produced by the *explore* command (Section 6.9.2), which are shown in Figure 6.24 alongside the corresponding histograms. The day 1 scores, which we looked at a few sections back, looked normally distributed on the histogram, and the Q-Q plot echoes this conclusion because the observed data fall very close to the 'ideal' diagonal line. In contrast, the distributions for days 2 and 3 look positively skewed. This can be seen in the Q-Q plots by the observed data showing the characteristic upward bend illustrated in Figure 6.23. These plots suggest that relative to day 1, hygiene scores on days 2 and 3 were more clustered around the low end of the scale (more people had low hygiene scores). Basically, people became smellier as the festival progressed. The skew on days 2 and 3 probably occurs because a minority were able to uphold their hygiene over the course of the festival (baby wet-wipes are indispensable, I find).

12 Be careful because in software such as R the default Q-Q plots flip the axes so that the theoretical quantiles are on the *x*-axis, instead of the *y*-axis, which inverts the patterns in this figure!

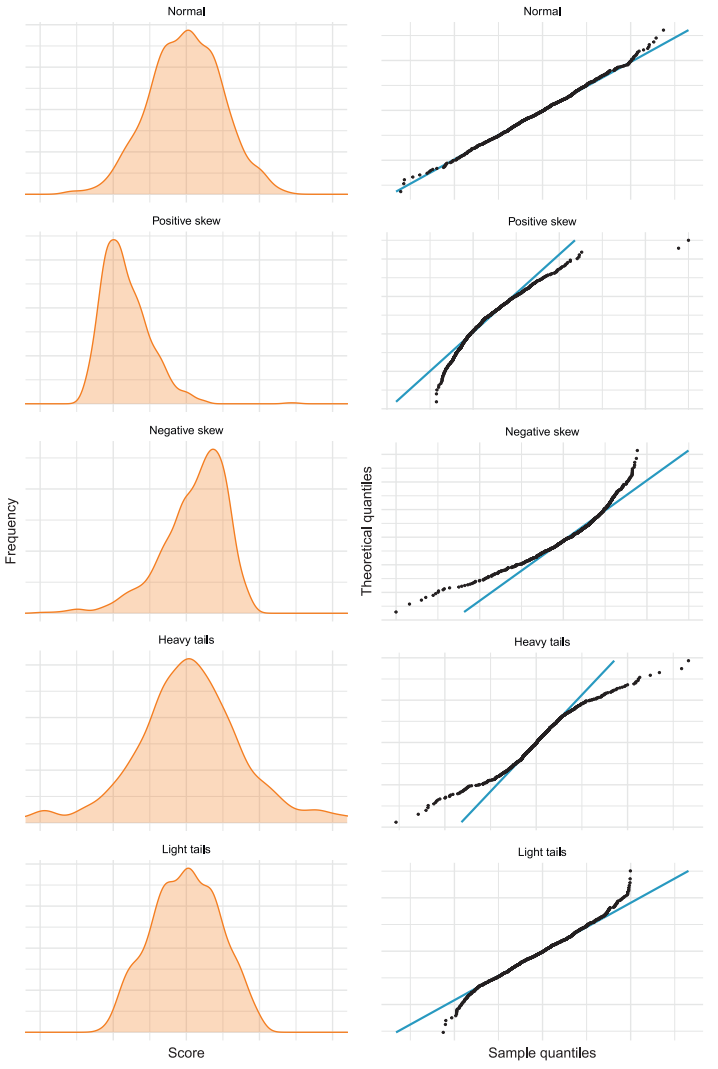


Figure 6.23 Characteristic shapes of Q-Q and P-P plots

Descriptives

		Hygiene (day 1 of download festival)		Hygiene (day 2 of download festival)		Hygiene (day 3 of download festival)	
		Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Mean		1.7711	.02437	.9609	.04436	.9765	.06404
95% Confidence Interval for Mean	Lower Bound	1.7233		.8736		.8497	
	Upper Bound	1.8190		1.0483		1.1033	
5% Trimmed Mean		1.7699		.9059		.9238	
Median		1.7900		.7900		.7600	
Variance		.481		.520		.504	
Std. Deviation		.69354		.72078		.71028	
Minimum		.02		.00		.02	
Maximum		3.69		3.44		3.41	
Range		3.67		3.44		3.39	
Interquartile Range		.92		.94		1.11	
Skewness		-.004	.086	1.095	.150	1.033	.218
Kurtosis		-.410	.172	.822	.299	.732	.433

Output 6.3

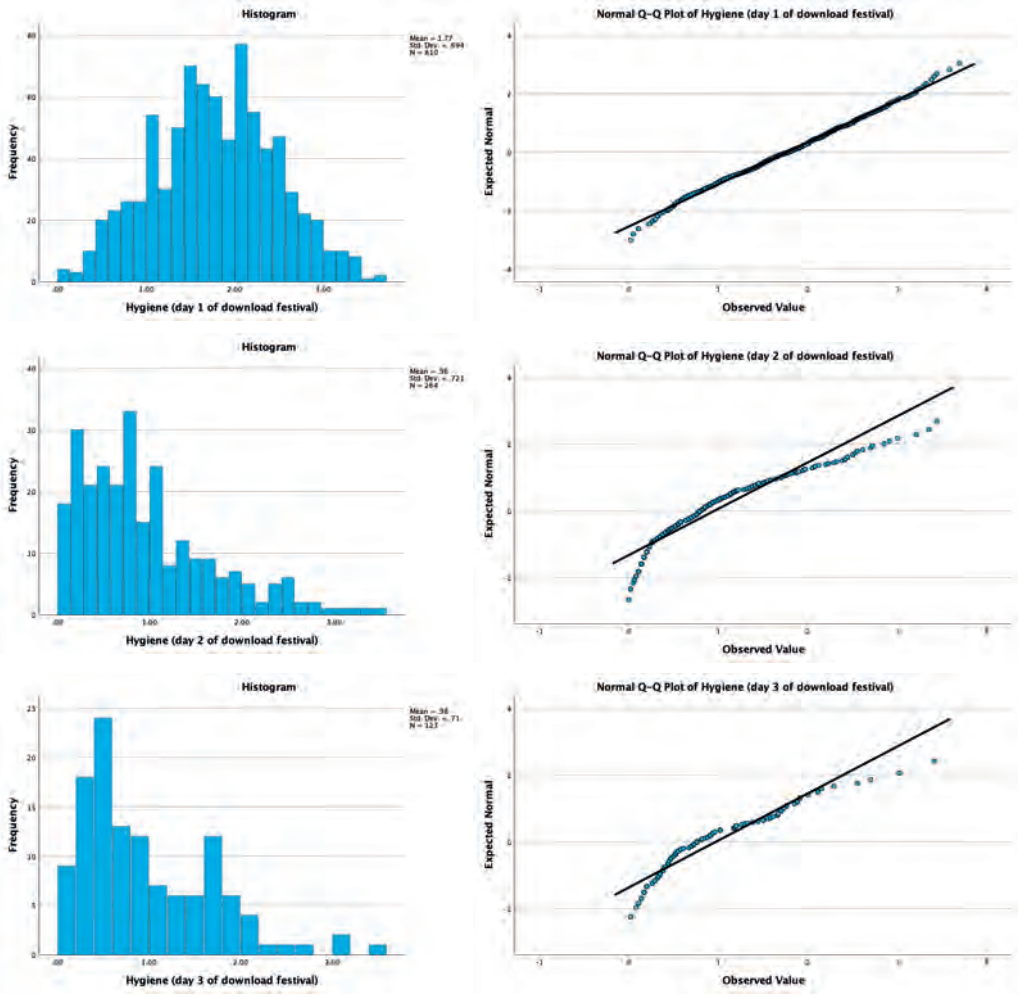
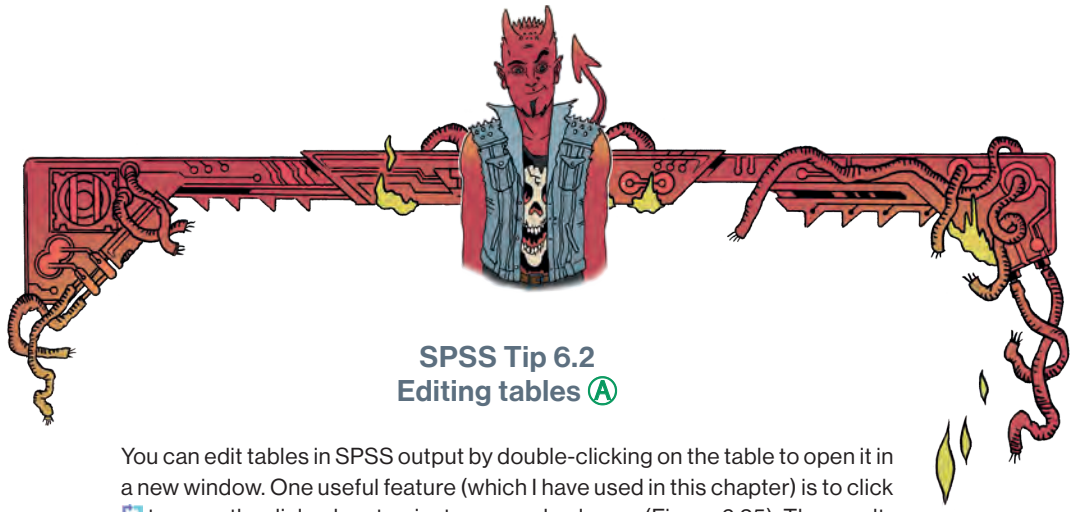


Figure 6.24 Histograms (left) and Q-Q plots (right) of the hygiene scores over the three days of the Download Festival

The *explore* command will also produce a table of descriptive statistics that we can use to assess the distributional properties of the data. Output 6.3 shows the table of descriptive statistics for the three variables in this example. To get the table in the same format as me, follow SPSS Tip 6.2. On average, hygiene scores were 1.77 (out of 5) on day 1 of the festival, but went down to 0.96 and 0.98 on days 2 and 3, respectively. The range and variances of scores stay similar across the festival. However, the skew on days 2 and 3 is evident from the fact that the mean and median differ. In a symmetrical distribution the mean and median will be approximately equal (as is the case on day 1), but positive skew tends to drag up the mean (because of the presence of too many large scores) but not the median, so if the mean is larger than the median you likely have positive skew.[13] The output also gives us a direct measure of skew and kurtosis and their associated standard errors. There are different ways to calculate skewness and kurtosis, but SPSS uses methods that give values of zero for a normal distribution, positive values for positive skew and negative values for negative skew. The further the value is from zero, the more severe the skew. Note that on day 1 skewness is basically 0, indicating a symmetric distribution,

13 Conversely, negative skew tends to drag down the mean (because of the presence of too many small scores) but not the median so if the mean is smaller than the median you likely have negative skew.



SPSS Tip 6.2
Editing tables Ⓐ

You can edit tables in SPSS output by double-clicking on the table to open it in a new window. One useful feature (which I have used in this chapter) is to click to open the dialog box to pivot rows and columns (Figure 6.25). The resulting dialog box lists the variables that make up rows, which in Figure 6.25 are the dependent variables (days of the festival) and statistics (a list of statistics such as the mean, median and so on). The resulting dialog box also lists the variables that make up columns, which in Figure 6.25 is something called *Stat Type*, which is whether the column shows the statistic or its standard error. At the moment the table is long and narrow because the days of the festival are stacked underneath each other. We could place the information for the different days of the festival in different columns by dragging the icon labelled *Dependent ...* from the row part of the table to the column part (above *Stat Type*). In doing so, each day of the festival will be represented by a column under which there are two sub-columns representing the statistic and standard error. You can see the result in Output 6.3. Compare this layout with what you get in your output.

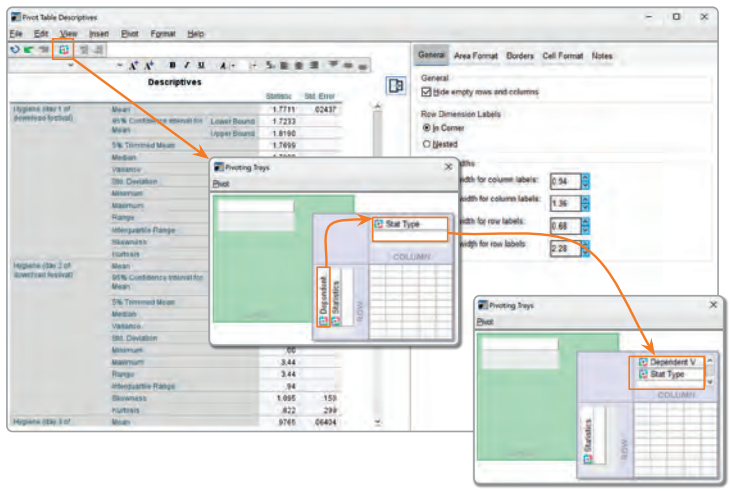
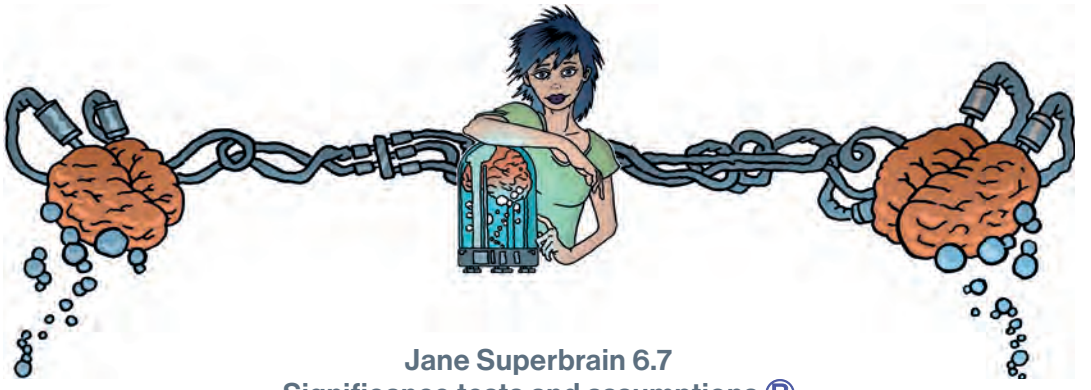


Figure 6.25 Using pivot controls

but day 2 and day 3 have skewness values of around 1, indicating positive skew. For kurtosis, positive values indicate a heavy-tailed distribution, whereas negative values indicate a light-tailed distribution. Again, the further the value is from zero, the heavier or lighter the tails. For day 1 kurtosis indicates slightly light tails, and for days 2 and 3 slightly heavy tails (but not excessively so).

A final way to assess normality is using significance tests, but this practice should be approached *with extreme caution* (see Jane Superbrain Box 6.7). In larger samples *don't do these tests*; instead, look at the shape of the distribution visually, interpret the value of the skewness and kurtosis statistics, and, because of the central limit theorem (Section 6.2.7), arguably don't worry about normality at all. I mainly include this section because you will see researchers use these tests in research papers and it's good for you to get that smug sensation of knowing better than they do.



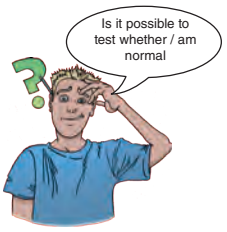
Jane Superbrain 6.7
Significance tests and assumptions Ⓑ

In this chapter we look at various significance tests that have been devised to tell us whether assumptions are violated. These include tests of whether a distribution is normal (the Kolmogorov–Smirnov and Shapiro–Wilk tests), tests of homogeneity of variances (Levene's test) and tests of significance for skew and kurtosis. I cover these tests mainly because people expect to see these sorts of things in introductory statistics books, and not because they are a good idea. All these tests are based on null hypothesis significance testing and this means that (1) in large samples they can be significant even for small and unimportant effects, and (2) in small samples they will lack power to detect violations of assumptions.

We have also seen in this chapter that the central limit theorem means that as sample sizes get larger, the assumption of normality matters less because the sampling distribution will be normal regardless of what our population (or indeed sample) data look like. So, in large samples, where normality matters less (or not at all), a test of normality is more likely to be significant and make us worry about and correct for something that doesn't need to be worried about or corrected for. Conversely, in small samples, where we should worry about normality, a significance test won't have the power to detect non-normality and so is likely to encourage us not to worry about something that we probably ought to.

The best advice is that if your sample is large then don't use significance tests of normality, in fact don't worry too much about normality at all. In small samples pay attention if your significance tests are significant but resist being lulled into a false sense of security if they are not.





A simple test of skew and kurtosis can be done by converting these statistics to *z*-statistics and comparing the resulting values to the corresponding probability value from the standard normal distribution. For example, an absolute value greater than 1.96 is significant at $p \leq 0.05$, above 2.58 is significant at $p \leq 0.01$ and above 3.29 is significant at $p \leq 0.001$. Using the same logic as we did when we developed a test statistic for our model parameters (Eq. 6.21), we can look at the difference between each estimate of each of skew ($\hat{S}$) and kurtosis ($\hat{K}$) and its expected value in a normal distribution (which in both cases is zero) and divide by the estimated standard error:

$$Z_S = \frac{\hat{S} - \hat{S}_{Expected}}{\widehat{SE}(\hat{S})} = \frac{\hat{S}}{\widehat{SE}(\hat{S})} \qquad Z_K = \frac{\hat{K} - \hat{K}_{Expected}}{\widehat{SE}(\hat{K})} = \frac{\hat{K}}{\widehat{SE}(\hat{K})} \tag{6.34}$$



For the hygiene scores, the test statistics for skewness are $Z_S = -0.004/0.086 = 0.047$ on day 1, 1.095/0.150 = 7.300 on day 2 and 1.033/0.218 = 4.739 on day 3. It is clear, then, that although on day 1 scores are not at all skewed, on days 2 and 3 there is a very significant positive skew (as was evident from the Q-Q plots). For kurtosis we get $Z_k = -0.410/0.172 = -2.38$ on day 1, 0.822/0.299 = 2.75 on day 2 and 0.732/0.433 = 1.69 on day 3. These values indicate significant kurtosis for all three days; however, because of the large sample we need to take these findings with a pinch of salt and take comfort from the central limit theorem.

You can get more elaborate tests of normality such as the **Kolmogorov–Smirnov test** and **Shapiro–Wilk test**, which compare the scores in the sample to a normally distributed set of scores with the same mean and standard deviation. If the test is non-significant ($p > 0.05$) it tells us that the distribution of the sample is not significantly different from a normal distribution (i.e., it is probably normal). If, however, the test is significant ($p \leq 0.05$) then the distribution in question is significantly different from a normal distribution (i.e., it is non-normal). These tests are tempting: they lure you with an easy way to decide whether scores are normally distributed. However, Jane Superbrain Box 6.7 explains excellent reasons not to use them. If you insist on using them, bear Jane's advice in mind and always plot your data as well and try to make an informed decision about the extent of non-normality based on converging evidence.

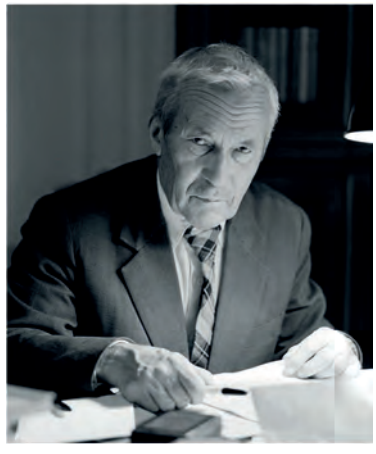


Figure 6.26 Andrei Kolmogorov, wishing he had a Smirnov

The *explore* command will produce the K-S test if you select ☑ Normality plots with tests (Section 6.9.2). Output 6.4 includes the test statistic, the degrees of freedom (which should equal the sample size)[14] and the significance value of this test. For day 1 the K-S test is just about not significant ($p = 0.097$), albeit surprisingly close to significant given how normal the day 1 scores looked in the histogram (Figure 6.24). This has occurred because

14 It is not $N - 1$ because the test compares the sample to an idealized normal, so the sample mean isn't used as an estimate of the population mean, which means that all scores are free to vary.

the sample size on day 1 is very large (N = 810) so the test is highly powered; it shows how in large samples even small and unimportant deviations from normality might be deemed significant by this test (Jane Superbrain Box 6.7). For days 2 and 3 the test is highly significant, indicating that these distributions are not normal, which is likely to reflect the skew seen in the histograms for these data (Figure 6.24).

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Hygiene (day 1 of download festival)	.029	810	.097	.996	810	.032
Hygiene (day 2 of download festival)	.121	264	<.001	.908	264	<.001
Hygiene (day 3 of download festival)	.140	123	<.001	.908	123	<.001

a. Lilliefors Significance Correction

Output 6.4



Cramming Sam's Tip
Skewness and kurtosis

- To check that the distribution of scores is approximately normal, look at Q-Q plots, histograms and the values of skewness and kurtosis.
- In a normal distribution the dots on a Q-Q plot will fall along the diagonal line.
- Positive values of skewness indicate too many low scores in the distribution, whereas negative values indicate a build-up of high scores.
- Positive values of kurtosis indicate a heavy-tailed distribution, whereas negative values indicate a light-tailed distribution.
- The further the value is from zero, the more likely it is that the data are not normally distributed.
- Significance tests of normality, skew and kurtosis should, in general, not be used. This advice especially applies to large samples, where the tests are likely to be significant even when the scores are not too different from normal.





Oliver Twisted
Please, Sir, can I have some more ... normality tests?

'There is another test reported in the table (the Shapiro–Wilk test),' whispers Oliver as he creeps up behind you, 'and a footnote saying that the "Lilliefors significance correction" has been applied. What the hell is going on?' Well, Oliver, all will be revealed in the additional material for this chapter on the companion website: you can find out more about the K-S test, and information about the Lilliefors correction and Shapiro–Wilk test. What are you waiting for?

6.9.6 Spotting linearity and heteroscedasticity Ⓑ

The reason for looking at the assumption of linearity and homoscedasticity together is that we can check both using the same plots. Both assumptions relate to the errors in the model (but we use model residuals as a proxy) and we can plot the values of these residuals against the corresponding values of the outcome predicted by our model in a scatterplot. The resulting plot shows whether there is a systematic relationship between what comes out of the model (the predicted values) and the errors in the model. Normally we convert the predicted values and errors to *z*-scores,[15] so this plot is sometimes referred to as *zpred vs. zresid*. If linearity and homoscedasticity hold true then there should be no systematic relationship between the errors in the model and what the model predicts. If this graph funnels out, then the chances are that there is heteroscedasticity in the data. If there is any sort of curve in this graph then the assumption of linearity is likely to be suspect.

Figure 6.27 shows examples of the plot of standardized residuals against standardized predicted values. The top left-hand graph shows a situation in which the assumptions of linearity and homoscedasticity have been met. The top right-hand graph shows a similar plot for a data set that violates the assumption of homoscedasticity. Note that the points form a funnel: they become more spread out across the graph. This funnel shape is typical of heteroscedasticity and indicates increasing variance across the residuals. The bottom left-hand graph shows a plot of some data in which there is a non-linear relationship between the outcome and the predictor: there is a clear curve in the residuals. Finally, the bottom right-hand graph illustrates data that not only have a non-linear relationship but also show heteroscedasticity: there is a curved trend in the residuals *and* at one end of the plot the spread of residuals is very small, whereas at the other end the residuals are widely dispersed. When

15 These standardized errors are called standardized residuals, which we'll discuss in Chapter 9.

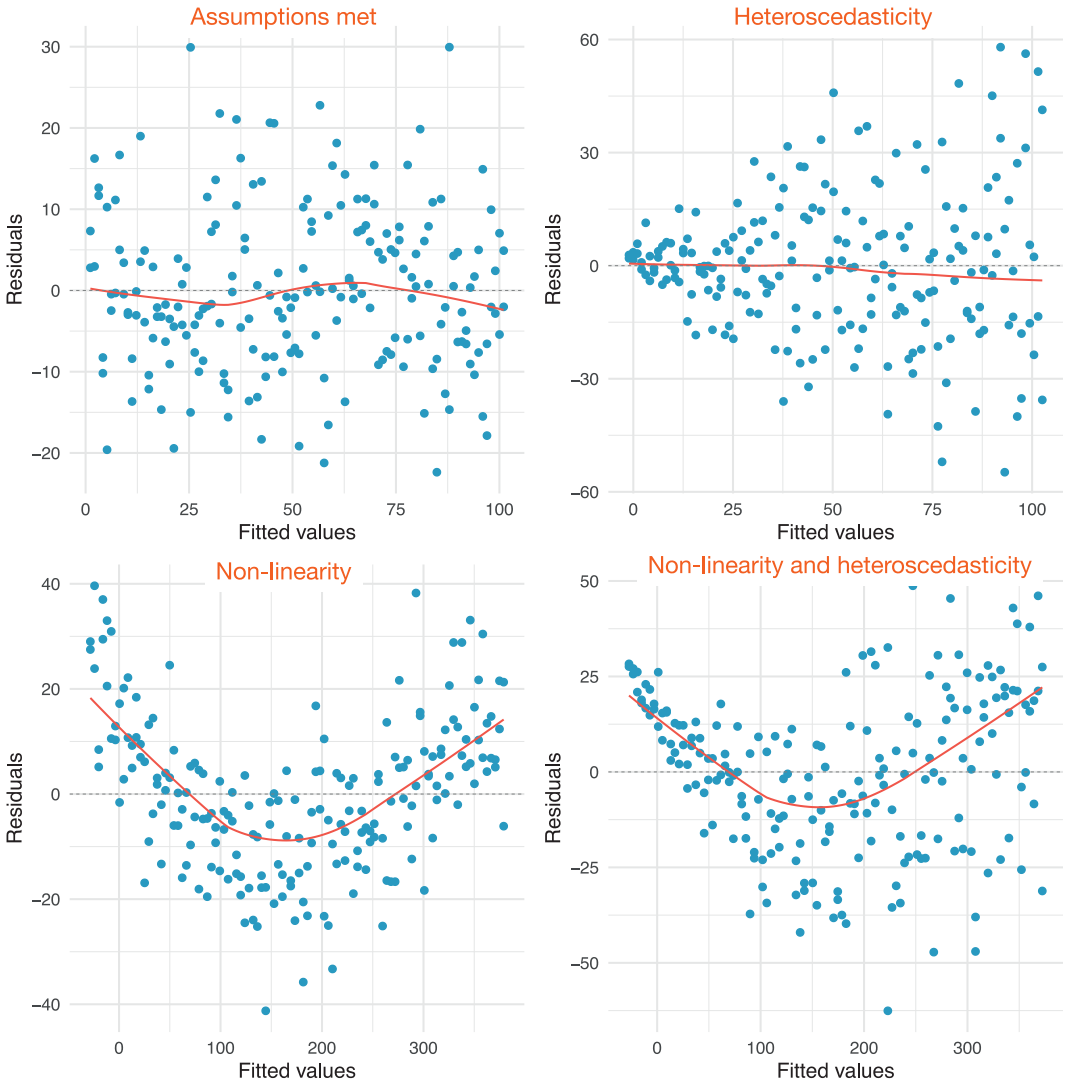


Figure 6.27 Plots of standardized residuals against predicted (fitted) values

these assumptions have been violated you won't see these exact patterns, but hopefully these plots will help you to understand the general anomalies to look out for.

We'll look at an example of how to use this plot in Chapter 9, but to give you a flavour, we will look at a scatterplot of the day 1 and day 2 hygiene scores. When a model has only one predictor then a scatterplot between the two variables will be a reasonable proxy for the residual plot. So, imagine we predicted day 2 hygiene from day 1. The scatterplot of these variables, which we first saw in Figure 6.19, is annotated in Figure 6.28 to show that although it seems reasonable to model the relationship as linear, the spread of observations around the model isn't constant, suggesting likely heteroscedasticity in the residuals.

When you have groups of cases, you're essentially trying to assess whether the variance within each group is equal. Again, we'll look at how to do this in subsequent chapters, but we can get some basic output using the *explore* command that we saw in Section 6.9.2. We'll use the *explore* command to look at hygiene scores on day 1 across the gender groups. Select *Analyze* > *Descriptive Statistics* > Explore... and then select and drag (or click ▸) the day 1 hygiene scores to the box labelled *Dependent List*). To get output for the different groups select and drag (or click ▸) the variable representing

gender to the area labelled *Factor List*. The completed dialog box will look like Figure 6.29. Click Plots... and notice that now we have placed a variable in the *Factor List* the area labelled *Spread vs. Level with Levene Test* is active. Leave the default selection of None and click Continue to return to the main dialog box and OK to generate the output.

Similarly to when we ran the *explore* command before, we will get Q-Q plots, but because we specified gender as a factor we will get separate plots for self-identifying males, females and non-binary festival-goers. From the point of view of assessing variance of hygiene scores across the groups the boxplot in Figure 6.30 shows that the spread of scores in each group is comparable (the length from the bottom to top of the boxplot is similar in the three groups), and the interquartile range (the length of the blue box) is similar in the three groups. This suggests the variances of hygiene scores are likely to be similar across the groups. We can also look at the values of the variances in Output 6.5,[16] which confirm what the boxplots show: the variances are very similar in the three groups. Remember that it's the variance of model errors that matter, so we are just using the observed scores as a proxy, and in subsequent chapters we will revisit this topic.

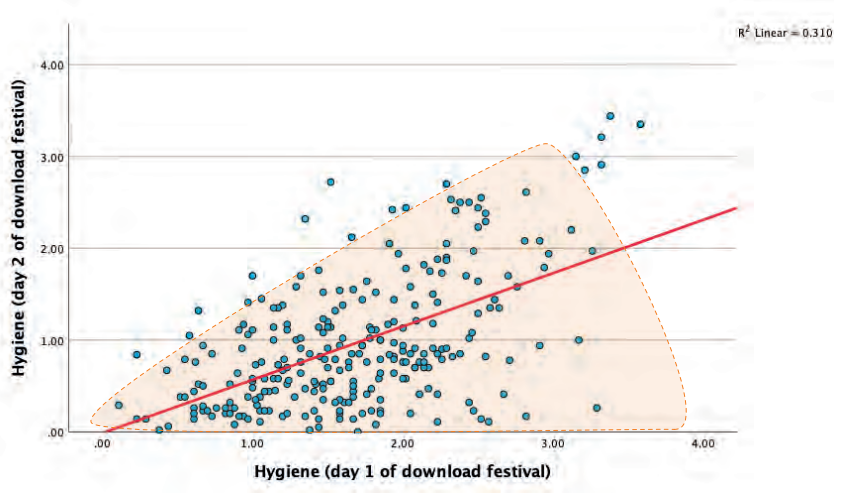


Figure 6.28 Plot of the day 2 against the day 1 hygiene scores

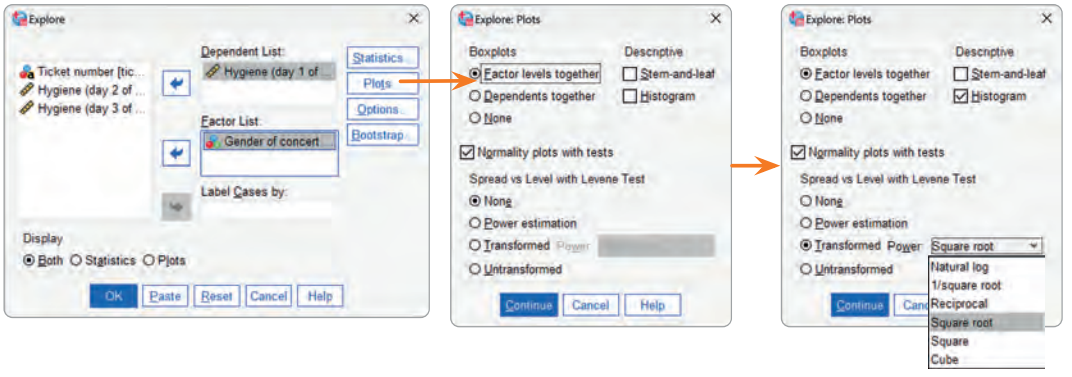


Figure 6.29 Dialog boxes for the *explore* command

16 I have changed the layout in a similar way as described in SPSS Tip 1.2.

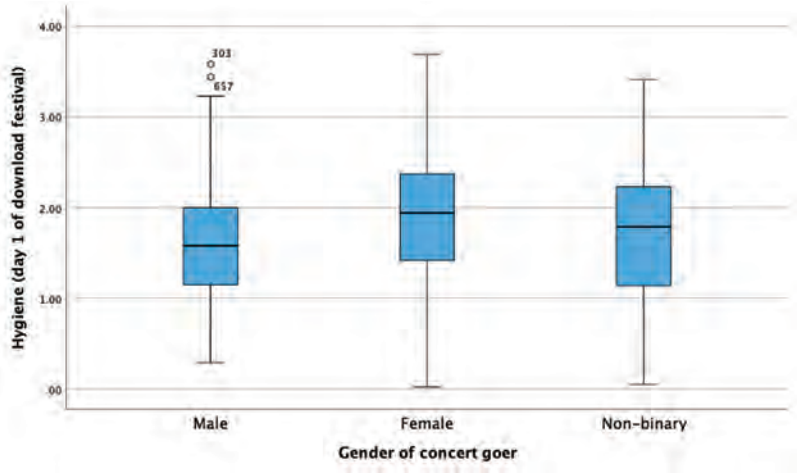


Figure 6.30 Boxplot of hygiene scores on day 1 of the festival across gender groups

Descriptives

			Gender of concert goer: Male		Female		Non-binary	
			Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Hygiene (day 1 of download festival)	Mean		1.6072	.03807	1.8842	.03315	1.7191	.07886
	95% Confidence Interval for Mean	Lower Bound	1.5322		1.8191		1.5624	
		Upper Bound	1.6821		1.9494		1.8758	
	5% Trimmed Mean		1.5965		1.8920		1.7156	
	Median		1.5800		1.9400		1.7900	
	Variance		.401		.487		.560	
	Std. Deviation		.63361		.69768		.74817	
	Minimum		.29		.02		.05	
	Maximum		3.58		3.69		3.41	
	Range		3.29		3.67		3.36	
	Interquartile Range		.86		.96		1.09	
	Skewness		.205	.146	-.168	.116	-.017	.254
	Kurtosis		-.112	.292	-.428	.231	-.289	.503

Output 6.5

A lot of people test the homoscedasticity assumption in groups by conducting something called **Levene's test** (Levene, 1960) on the observed scores (rather than the residuals). If Levene's test is significant at $p \leq 0.05$ then people tend to conclude that the null hypothesis is incorrect and that the variances are significantly different – therefore, the assumption of homogeneity of variances has been violated. If Levene's test is non-significant (i.e., $p > 0.05$) people take this to mean that the variances are roughly equal and the assumption is tenable.

The *explore* command spews out Levene's test, but you probably should ignore it for the same reasons that you shouldn't use the K-S test (Jane Superbrain Box 6.7): in large samples trivial differences in group variances can produce a Levene's test that is significant, and in small samples the test will only pick up on big differences. There are other considerations too. First, violating this assumption only really matters only if you have unequal group sizes; if group sizes are equal this assumption is pretty much irrelevant and can be ignored. Second, tests of homogeneity of variance work best when you have equal group sizes and large samples (when it doesn't matter if you have violated the assumption) and are less effective with unequal group sizes and smaller samples – which is exactly when the assumption matters. Plus, there are adjustments to correct for violations of this assumption that can be applied (as we shall see): typically, a correction is applied to offset whatever degree of heterogeneity is in the data (no heterogeneity = no correction). You might as well always apply the correction and forget about the assumption. If you're really interested in this issue, I like the article by

Zimmerman (2004). As if to prove my point, Output 6.6 shows the results of Levene's test for the day 1 hygiene scores and it is telling us that the variances are significantly different, $F(2, 807) = 3.17$, $p = 0.043$. According to the test we have violated the homoscedasticity assumption and should run around shrieking like a goat. However, we noted from the boxplots and descriptive statistics that the variances in the three groups are very similar. Despite the variances being similar, Levene's test is significant (because we have a large sample) and is misleading us into panic.

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Hygiene (day 1 of download festival)	Based on Mean	3.168	2	807	.043
	Based on Median	2.987	2	807	.051
	Based on Median and with adjusted df	2.987	2	800.400	.051
	Based on trimmed mean	3.167	2	807	.043

Output 6.6

Some people also look at **Hartley's F_{Max}** also known as the **variance ratio** (Pearson & Hartley, 1954). This is the ratio of the variances between the group with the biggest variance and the group with the smallest. This ratio was compared to critical values in a table published by Hartley. Again, though, it's unwise to use significance as a guide. However, as a rule of thumb if the variance ratio is less than 2 then we probably don't need to worry. The variance ratio for the day 1 hygiene scores is 0.56/0.40 = 1.4 (Output 6.5).



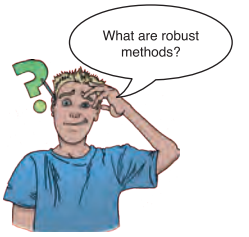
Cramming Sam's Tip
Homogeneity of variance

- Homoscedasticity is the assumption that the spread of population model errors is roughly equal at different points on the predictor variable.
- We can't directly observe population residuals, so the assumption can be evaluated by looking at a plot of the predicted values from your sample model against the residuals.
- You're looking for a random spread of dots on the plot. Anything resembling a funnel, a sausage or a sausagey funnel (a sausnel? a funnage?) and you've got problems.



6.10 Reducing bias with robust methods Ⓑ

This section looks at methods to mitigate bias. Historically, scientists have used tests that do not rely on the assumption of a normally distributed sampling distribution, known as non-parametric or assumption-free tests.[17] These tests are widely taught (see Chapter 7) and used, but they work by transforming the raw data (which, as we shall see, is problematic). Specifically, raw scores are converted to ranks, meaning that you lose information. Also, these tests have been developed for a limited range of situations; happy days if you want to compare two means, but sad and lonely days listening to Joy Division if you have a complex experimental design. Non-parametric tests were born in an era before computing, but with the advent of sophisticated computers came new possibilities for robust statistical models, meaning that there are a wide range of techniques, not based on ranks, that we can use.



By far the best option if you have irksome data (other than sticking a big samurai sword through your head) is to estimate parameters, standard errors, confidence intervals and *p*-values using methods that are unbiased in the presence of outliers and when the model assumptions discussed in this chapter are not met. In other words, use methods that are relatively unaffected by irksome data. These methods are collectively known as **robust methods** (see Field & Wilcox, 2017, for a gentle introduction). Because the phrase 'robust models' covers such a huge amount of ground, you need a separate book to get into the theory (I recommend Wilcox, 2017). Here, we'll overview some general principles that are commonly used in robust methods but not get into the nitty-gritty. We'll look at trimming data, Winsorizing, heteroscedasticity-consistent standard errors (as fun as they sound) and the bootstrap. Before that, let's talk about a widely used approach that is probably best avoided.

6.10.1 Transforming data Ⓑ



Historically, transforming data has been a popular way to deal with assumption violations. As I already mentioned, non-parametric tests all take advantage of transforming the raw scores into ranks. The idea behind **transformations** is that you do something to every score to correct for distributional problems, outliers, lack of linearity or unequal variances. For example, imagine you have a set of scores ranging from 1 to 100 that are positively skewed as in Figure 6.31 (left). Think what happens when you take the square root of each score. The highest score of 100 becomes 10 (a reduction of 90), a score of 25 becomes 5 (a reduction of 20), whereas the lowest score of 1 becomes 1 (no change). The point is that by taking the square root of each score, large raw scores are more affected by the transformation than small scores. The result of taking the square root of all scores is shown on the right of Figure 6.31. I've deliberately kept the scale on the plots the same so that you can see that large scores close to 100 have ended up as scores of around 10. Note that the skew has disappeared because large values (the values in the skewed tail) are reduced more than values in the main body of the distribution. It's like someone takes the right side of the plot and pushes it left, squashing the scores in the skewed area towards the main body of scores (think Luke Skywalker and Hans Solo trapped in the garbage compactor …). What I have just described is a square-root transformation.

17 Neither of these terms is particularly accurate. For example, none of these tests are assumption-free.

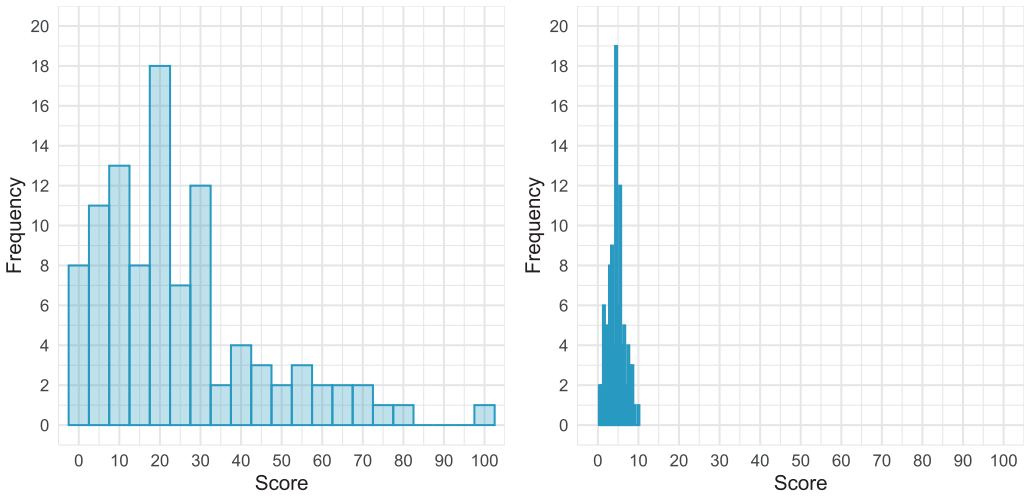


Figure 6.31 The effect of a square-root transformation

Various transformations exist: the most common are taking the square root of scores (as just discussed), taking the log of scores (or if the scores contain zeros taking the log of the scores plus 1) and taking the reciprocal of scores. Transformations are used to achieve one or more of these two aims:

1. Transform *predictor variables* to make their relationship to the outcome variable linear;
2. Transform the *outcome variable* to address issues with the assumptions of the model errors (e.g. to correct heteroscedasticity, lack of normality).

I will try to convince you that transforming predictor variables to make their relationship to the outcome variable linear is useful, whereas transformations to address heteroscedasticity and lack of normality are not.

Over the next couple of sections, we'll use an example related to a study by Meston and Frohlich (2003) that showed that heterosexual people rate a picture of someone of the opposite sex as more attractive after riding a roller-coaster compared to before. Imagine we took 20 people as they came off the Rockit roller-coaster at Universal Studios in Orlando[18] and asked them to rate the attractiveness of people in a series of photographs on a scale of 0 (looks like Jabba the Hutt) to 10 (looks like Princess Leia or Han Solo). The mean of their attractiveness ratings was the outcome. We also recorded their fear during the ride using a device that collates various indicators of physiological arousal and returns a value from 0 (chill) to 10 (terrified). This variable is the predictor. First, let's look at how transformations change the relationship between two variables. Figure 6.32 plots the attractiveness scores against the fear scores, summarizing their relationship (top left) and showing the means of the two variables (bottom left). Let's take the square root of the attractiveness ratings (but not the fear scores). The form of the relationship between attractiveness and fear (top middle) is changed (the slope is less steep), but the relative position of scores is unchanged (the relationship is still positive; the line still slopes up). If we look at the means (lower panels of Figure 6.32) the transformation creates a difference between means (middle) that didn't exist before the transformation (left). If we transform *both* variables (right-hand panels) then the relationship remains intact but the similarity between means is restored also.

Therefore, if you are looking at relationships between variables you can transform only the problematic variable, but if you are looking at differences between variables (e.g., changes in a variable over

18 I have a video of my wife and me on this roller-coaster during our honeymoon. I swear quite a lot on it, but I might stick it on my YouTube channel so you can see me cacking my pants.

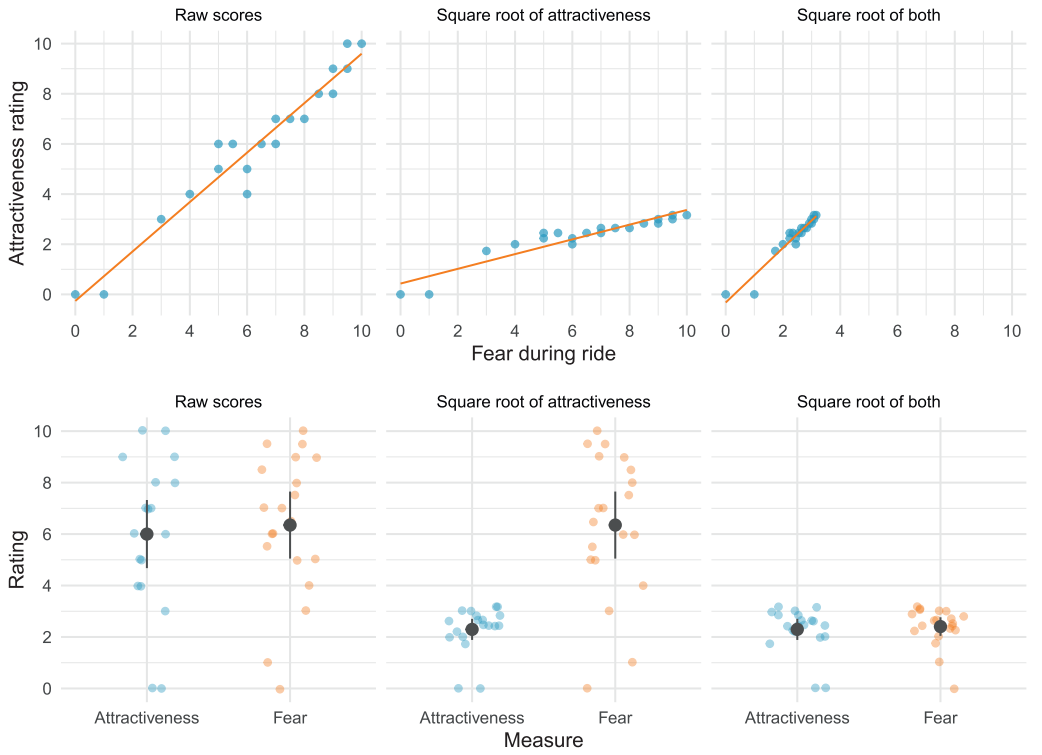


Figure 6.32 The effect of transforming attractiveness scores on its relationship to fear (top) and its mean relative to fear (bottom)

time) you *must* transform all the relevant variables. For example, the festival hygiene data were not normally distributed on days 2 and 3 of the festival, so we decide to transform them. If we want to look at how hygiene levels changed across the three days by comparing the mean hygiene score on day 1 to the corresponding means on days 2 and 3, we *must also transform the day 1 data* (even though scores were not skewed). If we don't transform the day 1 scores then differences in hygiene from day 1 to day 2 or 3 will be due to our transforming scores for days 2 and 3 but not day 1. If, however, we want to quantify the relationship between day 1 and day 2 scores (not the difference between them) we could transform the day 2 scores and leave the day 1 scores alone.

6.10.2 Transforming predictor variables Ⓑ

Linearity is the most important assumption of linear models, so if we believe the relationship between a predictor (or combination of predictors) and outcome is not linear then we can either fit a non-linear model or transform the predictor variable to try to make the relationship linear. Going back to our roller-coaster example, imagine we collected a new sample. These data are in **rollercoaster.sav**. This time when we plot the relationship between fear and attractiveness ratings (Figure 6.33, left) we notice that the relationship is not linear: the green curved (non-linear) line is a closer fit of the observed scores than the straight red line. Figure 6.33 (middle) plots the square root of fear scores against attractiveness. Notice that the red (linear) and green (curvilinear) relationships are very similar now; transforming the predictor variable (**fear**) has 'straightened out' the relationship between fear and attractiveness. Figure 6.33 (right) shows a similar effect from taking the natural logarithm of the fear. The red (linear) and green (curvilinear) relationships are even closer than for the square-root

transformation. Again, transforming the predictor variable (**fear**) has straightened out the relationship between fear and attractiveness. This example illustrates how transforming predictor variables can be useful. However, we are making the interpretation of the model more complicated because the parameter estimate for **fear** now represents the change in **attractiveness** from a unit change in the square root or log of **fear**.

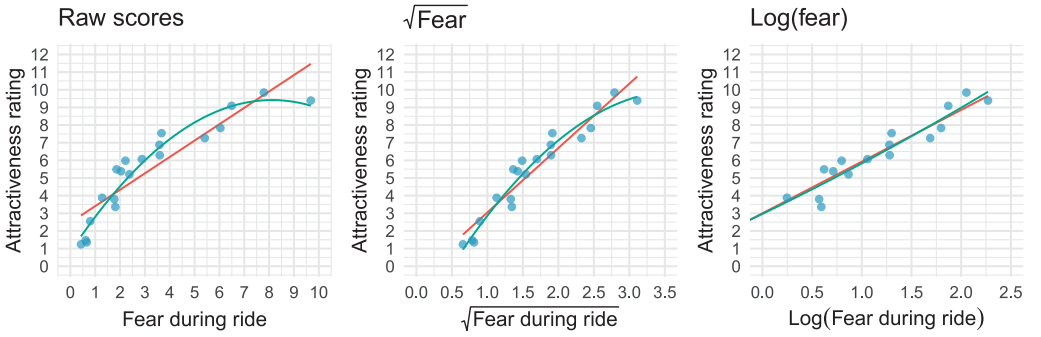


Figure 6.33 The effect of transforming fear scores on its relationship to attractiveness

6.10.3 Don't transform the outcome variable ... don't do it Ⓑ

The other context in which people apply transformations is to correct for lack of normality or homoscedasticity in model residuals by transforming the outcome variable. However, lack of normality doesn't matter in large samples, there are robust ways to deal with heteroscedasticity (as we will discuss) and transforming the outcome variable creates more problems than it solves. As Glass et al. (1972, p. 241) put it, 'the payoff of normalizing transformations in terms of more valid probability statements is low, and they are seldom considered to be worth the effort'. The issue is complicated, but the core question is whether a statistical model performs better when applied to transformed data or data that violate the assumption that the transformation corrects. The answer will depend on which 'model' you're applying and how robust it is.

For example, the *F*-test (see Chapter 12) is often claimed to be robust. Early findings suggested that *F* performed as it should when applied to skewed distributions and that transforming the data helped as often as it hindered the accuracy of *F* (Games & Lucas, 1966). However, in a lively but informative exchange Levine and Dunlap (1982) showed that transformations of skew did improve the performance of *F*; Games (1983) argued that this conclusion was incorrect; and Levine and Dunlap (1983) disagreed in a response to the response. In a response to the response to the response, Games (1984) raised four important issues:

1. The central limit theorem (Section 6.2.7) tells us that in samples larger than about 30 the sampling distribution will be normal regardless (see Section 6.2.7).
2. Transforming the data changes the hypothesis being tested. For example, when comparing means, converting from raw scores to log scores means that you're now comparing geometric means rather than arithmetic means. Transformation also means that you're addressing a different construct than the one originally measured, and this has obvious implications for interpreting the data (Grayson, 2004).
3. It is tricky to determine normality one way or another in small samples (see Jane Superbrain Box 6.7).
4. The consequences for the statistical model of applying the 'wrong' transformation could be worse than the consequences of analysing the untransformed scores.

Given these issues, unless you're correcting for a lack of linearity I would use robust models, where possible, in preference to transforming the data.

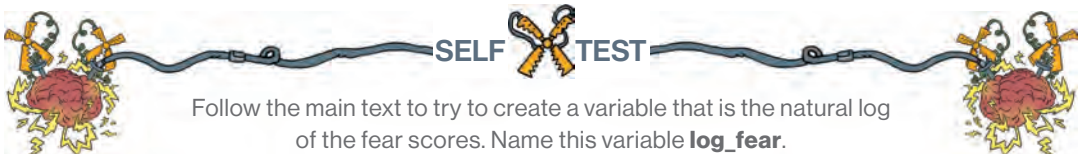
6.10.4 The *compute* command Ⓑ

If you do need to 'straighten' a curvilinear relationship you can apply transformations by selecting *Transform* > Compute Variable... Let's open the data in **rollercoaster.sav** that are shown in Figure 6.33 (left) and look at performing transformations on the **fear** variable as in the middle and right panels of that plot. Figure 6.34 shows the *Compute Variable* dialog box; it has a list of functions on the right-hand side, a calculator-like keyboard in the centre and a blank space that I've labelled the command area. You type a name for a new variable in the area labelled *Target Variable* and then you use the command area to tell SPSS how to create this new variable. You can:

- *Create new variables from existing variables*: For example, you could use it like a calculator to add variables (i.e., add two columns in the data editor to make a third), or to apply a function to an existing variable (e.g., take the square root).
- *Create new variables from functions*: There are hundreds of built-in functions that SPSS has grouped together. In the dialog box these groups are listed in the area labelled *Function group*. Upon selecting a function group, a list of available functions within that group will appear in the box labelled *Functions and Special Variables*. When you select a function, a description of it appears in the white box indicated in Figure 6.34.

You can enter a variable into the command area by selecting it required from the variables list and dragging it into the command area (or click). Likewise, you can select a function from the list of available functions and enter it into the command area by clicking . In Figure 6.34 I am creating a new variable called **log_fear** which is the natural logarithm of the variable **fear**. To create this variable we first type a name for it ('log_fear') in the box labelled *Target Variable*, then click Type & Label... to activate a dialog box in which we can add a more descriptive label and specify whether it is a numeric or string variable. Click on Continue to return to the main dialog box. We now need to tell SPSS *how* to create this new variable. In the box labelled *Function group* select *Arithmetic* and then in the box labelled *Functions and Special Variables* select *Ln* (this function takes the natural logarithm, whereas Lg10 takes the logarithm to base 10) and transfer it to the command area by clicking . The command will appear as 'LN(?)' and the question mark needs to be replaced with a variable name; replace it with the variable **fear** by selecting the variable in the list and dragging it across, clicking or typing 'fear' where the question mark is. In fact, rather than selecting functions as just described, you can just type the instruction into the command area, in this case you'd type 'LN(fear)'.

When you have written your command, execute it by clicking OK. If you have specified a variable name that already exists, SPSS will tell you so and ask you whether you want to replace this existing variable. If you respond with *Yes*, SPSS will replace the data in the existing column with the result of the *compute* command; if you respond with *No*, then nothing will happen and you will need to rename the target variable. If you're computing a lot of new variables it can be quicker to use syntax (see SPSS Tip 6.3).



Follow the main text to try to create a variable that is the natural log of the fear scores. Name this variable **log_fear**.

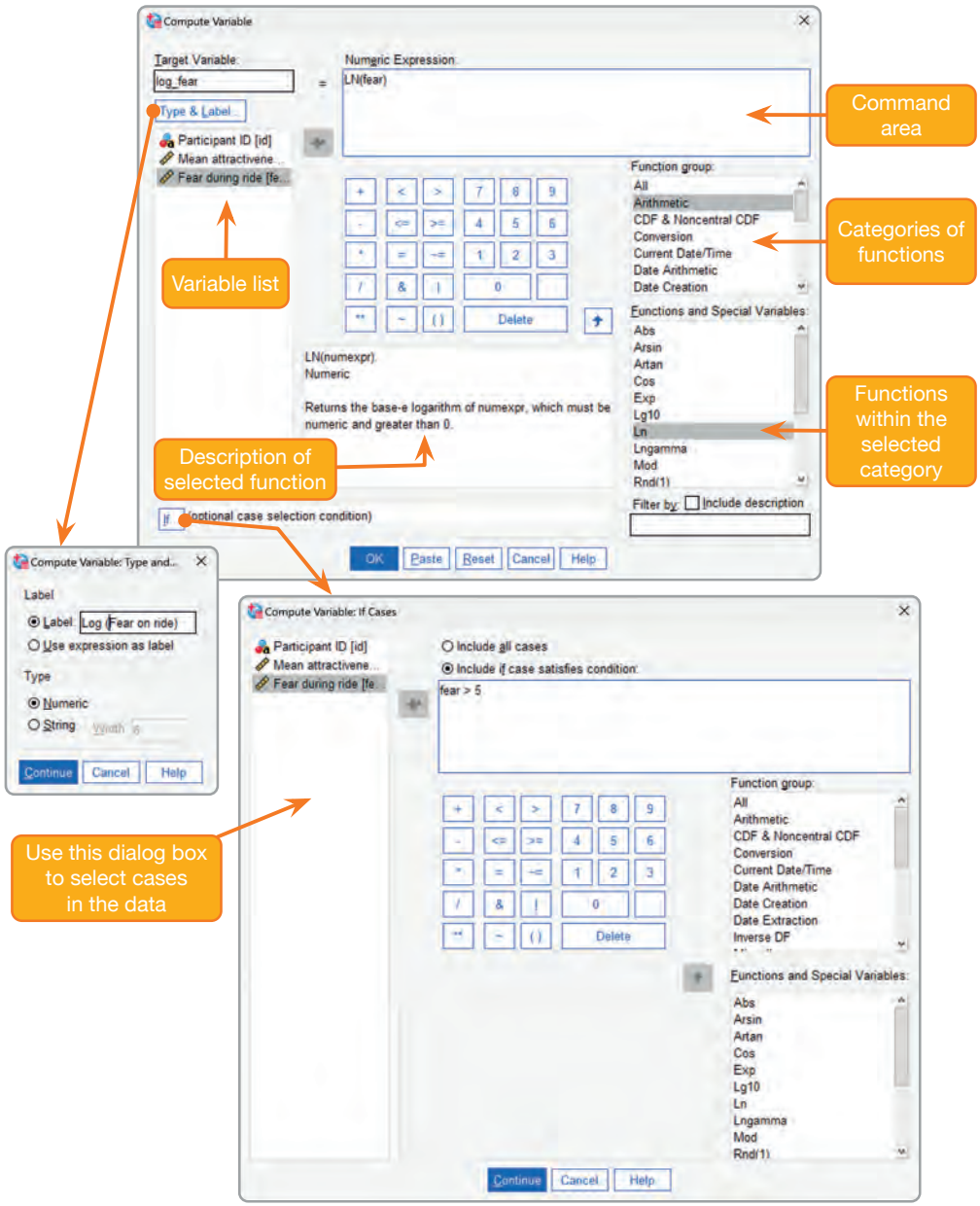


Figure 6.34 Dialog box for the *compute* command

To orient you to things you can do to compute new variables, let's look at some of the buttons in Figure 6.34.

+

Addition: This button places a plus sign in the command area. For example, with our roller-coaster data, 'attractiveness + fear' creates a column in which each row contains the score from the column labelled *attractiveness* added to the score from the column labelled *fear* (e.g., for participant id *cnob*, 5.98 + 2.22 = 8.20).

–

Subtraction: This button places a minus sign in the command area. For example, if we typed 'attractiveness – fear', *compute* would create a column in which each row contains the score from the column labelled *fear* subtracted from the score from the column labelled *attractiveness* (e.g., for participant id *cnob*, 5.98 – 2.22 = 3.76).

* **Multiply**: This button places a multiplication sign in the command area. For example, 'attractiveness*fear' creates a column that contains the score from the column labelled *attractiveness* multiplied by the score from the column labelled *fear* (e.g., for participant id *cnob*, 5.98 × 2.22 = 13.27).

/ **Divide**: This button places a division sign in the command area. For example, 'attractiveness/fear' creates a column that contains the score from the column labelled *attractiveness* divided by the score from the column labelled *fear* (e.g., for participant id *cnob*, 5.98/2.22 = 2.69).

** **Exponentiation**: This button raises the preceding term by the power of the succeeding term. So, 'fear**2' creates a column that contains the scores in the *fear* column raised to the power of 2 (i.e., the square of each number in the *fear* column: for participant id *cnob*, $2.22^2 = 4.93$). Likewise, 'fear**3' creates a column with values of **fear** cubed.

> **More than**: This operation is useful for creating variables to label or exclude certain cases that have values above a certain threshold. So, if you click on If... and typed 'fear > 5' as in Figure 6.34 then SPSS would carry out any *compute* commands only on cases for which **fear** scores are greater than 5. This could be used, for example, to create a variable that labels cases as 'high fear'.

>= **More than or equal to**: This operation is the same as above but will include cases that are exactly the value (in the example above exactly 5) as well.

< **Less than**: This is the reverse of the more than button, so is similarly helpful for creating variables to label or exclude certain cases but when you want to select cases below a certain threshold. Again, if you clicked on If... and typed 'fear < 5' then SPSS would carry out any compute commands only on cases for which fear scores are less than 5. This could be used, for example, to create a variable that labels cases as 'low fear'.

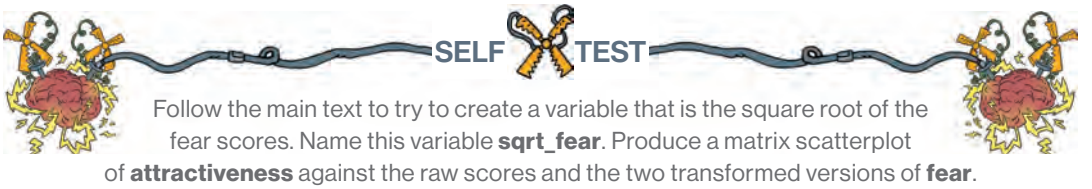
<= **Less than or equal to**: This operation is the same as above except that in the example above, cases that are exactly 5 would be included as well.

= **Equal to**: You can use this operation to include cases for which participants have a specific value. So, if you clicked on If... and typed 'fear = 2.22' then only cases that have a value of exactly 2.22 for the **fear** variable are included. This operation is most useful when you have a coding variable and you want to apply a function to only one of the groups. For example, if you had a data set with a variable called **animal** that is coded 1 = dog, 2 = cat, 3 = lizard, and you wanted to compute a new variable but only for dogs, you could type 'animal = 1'.

~= **Not equal to**: This operation will include all cases except those with a specific value. To use the example above, typing 'animal ~= 1' will result in the *compute* command being applied only on the cats and lizards and not dogs (because cats are coded as 2, and lizards as 3 so their values for **animal** is not equal to 1).

Some useful functions are listed in Table 6.3, which shows the standard form of the function, the name of the function, an example of how the function can be used and what the resulting variable would contain if that command were executed. If you want to know more the SPSS help files have details of all the functions available through the *Compute Variable* dialog box (click Help when you're in the dialog box).

In Figure 6.33 we looked at how the natural logarithm and square-root transformation could be used to straighten up the relationship between fear and attractiveness ratings. We've created a variable containing the natural log of **fear**, so let's now create one containing the square root of **fear** scores. Referring to Figure 6.34 enter the name **sqrt_fear** in the box labelled *Target Variable* (and click Type & Label... to give the variable a descriptive name). In the list box labelled *Function group* select *Arithmetic*, select *Sqrt* in the box labelled *Functions and Special Variables* and drag it to the command area (or click ↑). The command appears as SQRT(?); replace the question mark with the variable **fear** by selecting the variable in the list and dragging it, clicking ↵, or typing 'fear' where the question mark is. The final expression will read SQRT(**fear**). Click OK to create the variable.



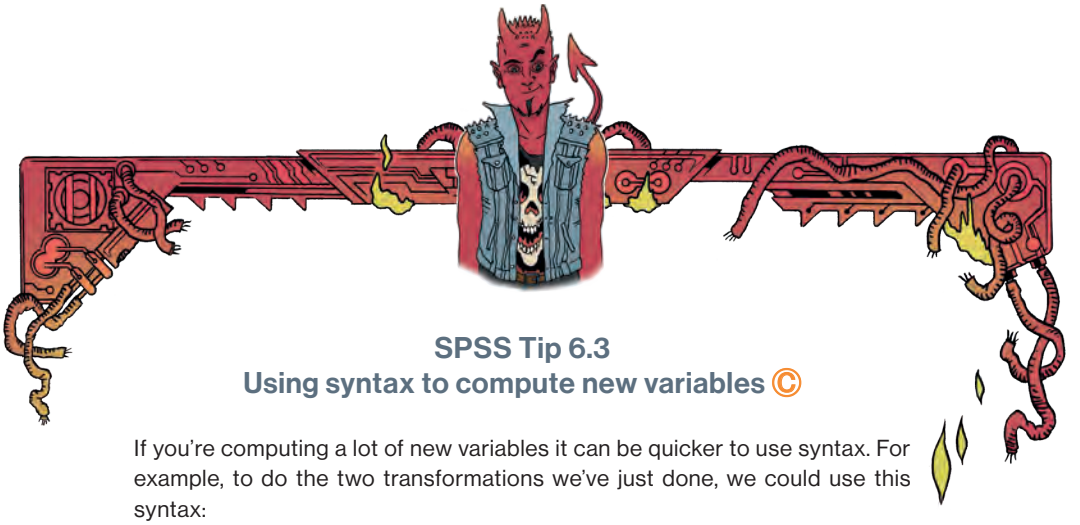
Follow the main text to try to create a variable that is the square root of the fear scores. Name this variable **sqrt_fear**. Produce a matrix scatterplot of **attractiveness** against the raw scores and the two transformed versions of **fear**.

Table 6.3 Some useful *compute* functions

Function	Name	Example Input	Output
MEAN(?,?, ...)	Mean	MEAN(attractiveness, fear)	For each row, SPSS calculates the average of the attractiveness and fear scores.
SD(?,?, ...)	Standard deviation	SD(attractiveness, fear)	Across each row, SPSS calculates the standard deviation of the values in the columns labelled *attractiveness* and *fear*.
SUM(?,?, ...)	Sum	SUM(attractiveness, fear)	For each row, SPSS adds the values in the columns labelled *attractiveness* and *fear*.
SQRT(?)	Square root	SQRT(fear)	Produces a column containing the square root of each value in the column labelled *fear*.
ABS(?)	Absolute value	ABS(attractiveness)	Produces a variable that contains the absolute value of the values in the column labelled *attractiveness* (absolute values are ones where the signs are ignored: so −5 becomes +5 and +5 stays as +5).
LN(?)	Natural logarithm	LN(fear)	Produces a variable that contains the logarithmic values (to base *e*) of the variable **fear**.
LG10(?)	Base 10 logarithm	LG10(fear)	Produces a variable that contains the logarithmic values (to base 10) of the variable **fear**.
RV.NORMAL(mean, stddev)	Normal random numbers	RV.NORMAL(20, 5)	Produces a variable of pseudo-random numbers from a normal distribution with a mean of 20 and a standard deviation of 5

6.10.5 Trimming the data Ⓑ

Having purged our souls of the urge to transform scores, we will look at some common techniques that robust models use. The first is trimming data, which involves removing extreme scores using one of two rules: (1) a percentage based rule; and (2) a standard deviation based rule. One of these is a good idea and the other not so much (Jane Superbrain Box 6.8). A percentage-based rule would be, for example, deleting the 10% of highest and lowest scores. Think back to the first set of data from the roller-coaster example where we have 20 fear scores and 20 attractiveness scores (plotted in Figure 6.32). Figure 6.35 shows the attractiveness scores. Most people gave ratings above the midpoint of the scale: they were reasonably positive in their ratings. However, there were two people who gave zeros. If we were to trim 5% of the data from either end, this would mean deleting one score at each extreme (there are 20 scores and 5% of 20 is 1). Figure 6.35 shows that this involves deleting a 0 and a 10. We could compute a 5% trimmed mean by working out the mean for this trimmed data set. Similarly, Figure 6.35 shows that with 20 scores, a 10% trim would mean deleting two scores from each extreme, and a 20% trim would entail deleting four scores from each extreme. If you take trimming to its extreme then you get the median, which is the value left when you have trimmed all but the middle score. If we calculate the mean in a sample that has been trimmed in this way, it is called (unsurprisingly) a **trimmed mean**. A similar robust measure of location is an **M-estimator**, which



SPSS Tip 6.3
Using syntax to compute new variables ©

If you're computing a lot of new variables it can be quicker to use syntax. For example, to do the two transformations we've just done, we could use this syntax:

```
COMPUTE log_fear = LN(fear).
COMPUTE sqrt_fear = SQRT(fear).
EXECUTE.
```

Each *compute* command above does the equivalent of what you'd do using the *Compute Variable* dialog box in Figure 6.34. So, the first line creates a new variable (**log_fear**), which is the natural log of the variable **fear**. The next line creates a new variable (**sqrt_fear**) by using the *SQRT* function to take the square root of **fear** scores. The final line has the command *execute* without which none of the *compute* commands beforehand will run. Note also that every line ends with a full stop.

differs from a trimmed mean in that the amount of trimming is determined empirically. In other words, rather than the researcher deciding before the analysis how much of the data to trim, an M-estimator determines the optimal amount of trimming necessary to give a robust estimate of the parameter of interest (e.g. the mean). This method has the obvious advantage that you never over- or under-trim your data; however, the disadvantage is that it is not always possible to reach a solution.



Compute the mean and variance of the attractiveness ratings. Now compute them for the 5%, 10% and 20% trimmed data.

If you do the self-test you should find that the mean rating was 6. The 5% trimmed mean is 6.11, and the 10% and 20% trimmed means are both 6.25. The means get higher in this case because the extreme scores were both low (two people who gave ratings of 0) and trimming reduces their impact (which would have been to lower the mean). For the overall sample the variance was 8, for the 5%, 10% and 20% trimmed data you get 5.87, 3.13 and 1.48, respectively. The variances get smaller (and

more stable) because, again, scores at the extreme have no impact (because they are trimmed). We saw earlier that the accuracy of the mean and variance depends on a symmetrical distribution, but a trimmed mean (and variance) will be relatively accurate even when the distribution is not symmetrical, because by trimming the ends of the distribution we remove outliers and skew that bias the mean. Many robust methods work by taking advantage of the properties of the trimmed mean, the median and M-estimators.

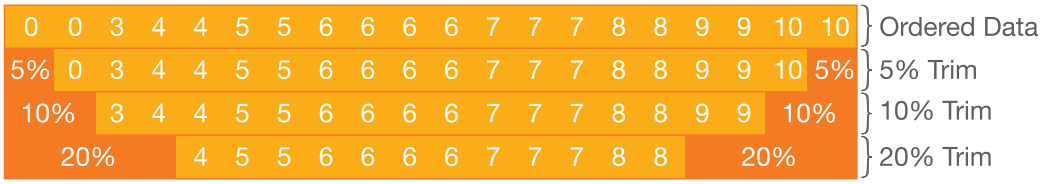
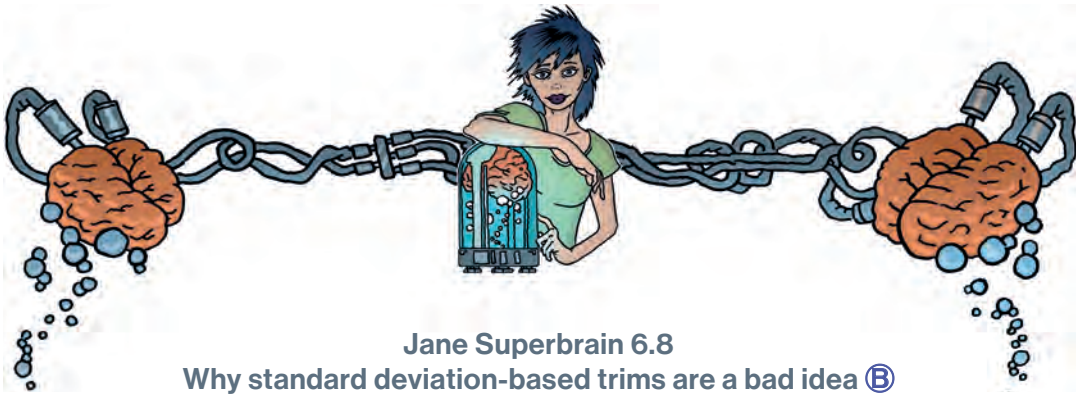


Figure 6.35 Illustration of trimming data



Jane Superbrain 6.8
Why standard deviation-based trims are a bad idea Ⓑ

Standard deviation-based trimming involves calculating the mean and standard deviation of a set of scores, and then removing values that are a certain number of standard deviations greater than the mean. This technique is really popular for dealing with reaction time data, which are notoriously messy and often skewed. A typical approach is to remove reaction times greater than (or below) 2.5 standard deviations from the mean (Ratcliff, 1993). For the roller-coaster data (main text) the variance was 8 and the standard deviation is the square root of this value, 2.83. If we wanted to trim 2.5 times the standard deviation, then we'd use 2.5 × 2.83 = 7.08 as a criterion. The mean was 6, therefore, we would delete scores greater than 6 + 7.08 = 13.08, of which there were none (it was only a 10-point scale); we would also delete scores less than 6 − 7.08 = −1.08, which again means deleting no scores (because the lowest score was zero). In short, applying this rule wouldn't affect the mean or standard deviation, which is odd, isn't it? The example illustrates the fundamental problem with standard deviation based trimming, which is that the mean and standard deviation are both highly influenced by outliers (see Section 6.4); therefore, the outliers in the data bias the criterion that you use to reduce their impact. In this case, the inflated standard deviation also inflates the trimming rule to beyond the limits of the data. It's probably not wise to correct bias using a criterion that is biased by the problem for which you're trying to correct.



6.10.6 Winsorizing Ⓐ

Winsorizing the data is a bit like trimming except that, rather than discarding scores outside of a certain threshold, they are replaced. Figure 6.36 shows the idea (compare this figure with Figure 6.35). We set a threshold, such as 10%, and above (and below) this threshold we do something to scores. With trimming the 'something' we do is to discard the scores, whereas with winsorizing the 'something we do' is to replace them with the value of the score at the threshold. For example, if we set the threshold as 0.1 (or 10%) then we order the scores in ascending order, and replace the bottom 10% of scores with the value of the score at the threshold (i.e., at 10% along the distribution). In Figure 6.36 the score 10% along the distribution is 4, so all scores up to this point are replaced with 4. Similarly, at the top end of the distribution of scores we replace the largest 10% of scores with the value of the score at the threshold (i.e., 90% along the distribution). In Figure 6.36 the score 90% along the distribution is 8, so all scores above this point are replaced with 8.

It's perfectly natural to feel uncomfortable at the idea of changing the scores you collected to different values. It feels a bit like cheating. Bear in mind, though, that if the score you're changing is very unrepresentative of the sample and biases your statistical model then it's better than reporting and interpreting a biased model.[19] What would be cheating is not dealing with extreme cases because they bias the results in favour of your hypothesis, or changing scores in a systematic way other than to reduce bias (again, perhaps to support your hypothesis).



Oditi's Lantern
Select cases

'I, Oditi, believe that those who would try to prevent our cult from discovering the truths behind the numbers have placed dead herrings within the data. These rotting numerical fish permeate our models and infect the nostrils of understanding with their putrid stench. We must banish them, we must select only the good data, the pure data, the data uncontaminated by piscine putrefaction. You, the trooper of truth, must stare into my lantern to discover how to select cases using SPSS.'



19 It is worth making the point that having outliers is interesting in itself, and if you don't think they represent the population then you need to ask yourself why they are different. The answer to the question might be a fruitful topic of more research.

There are some variations on winsorizing, such as replacing extreme scores with a score 3 standard deviations from the mean. A *z*-score of 3.29 constitutes an outlier (see Section 6.9), so we can calculate what score would give rise to a *z*-score of 3.29 (or perhaps 3) by rearranging the *z*-score equation: $X = (z \times s) + \bar{X}$. All we're doing is calculating the mean ($\bar{X}$) and standard deviation (s) of the data and, knowing that z is 3 (or 3.29 if you want to be exact), adding 3 (or 3.29) times the standard deviation to the mean and replacing our outliers with that score.

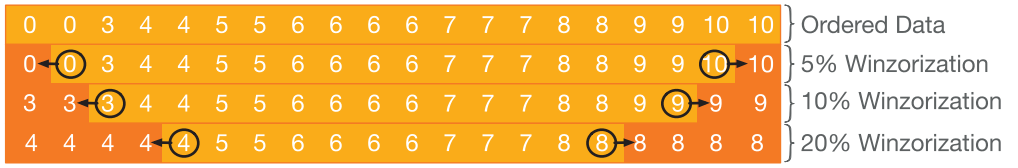


Figure 6.36 Illustration of winsorizing data

6.10.7 Heteroscedasticity consistent standard errors ©

If heteroscedasticity is the problem then we know that this mainly impacts the standard errors in the model (and hence the confidence intervals and significance tests). However, there are ways to estimate the standard error without needing to assume homoscedasticity. We saw in Section 6.7 that when computing the variance (and hence standard error) of the model parameters, a matrix ($\boldsymbol{\Phi}$) is used that represents the variances and covariances of the population model errors. We can't observe these errors, and so can't know the elements of this matrix, so instead it is assumed to have the form in Eq. 6.30. The assumption of spherical errors means that we can know (or at least estimate) the elements of this matrix. If the assumption is not met it essentially means that we need a different way to estimate the values in $\boldsymbol{\Phi}$ based on the sample data. A range of methods have been suggested, all based on what came to be known as sandwich estimators (Eicker, 1963; Huber, 1967; White, 1980). The most basic of these is the HC0 estimator, which simply replaces the diagonal elements of the matrix in Eq. 6.30 with the squared residuals from the sample model (which, unlike the population errors, are observed):

$$\widehat{\boldsymbol{\Phi}} = \begin{bmatrix} e_1^2 & 0 & 0 & \cdots & 0 \\ 0 & e_2^2 & 0 & \cdots & 0 \\ 0 & 0 & e_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & e_n^2 \end{bmatrix} \tag{6.35}$$

Using the squared residuals gives us an estimate of $\boldsymbol{\Phi}$ (hence we put a hat on it, because estimates love hats), meaning that we no longer need to assume homoscedasticity. Note, though, that the off-diagonal elements remain zero so we do still assume independent errors. It turns out that in large samples this estimate is consistent under conditions of heteroscedasticity, but it's not so good in small samples. Subsequent tweaks have been made (imaginatively named the HC1, HC2, HC3 and HC4 estimators). We don't need to know how they are calculated – they all adjust the HC0, so are based on the idea of using the sample residuals to estimate $\boldsymbol{\Phi}$. They all, in different ways, attempt to improve the estimate in the presence of what are known as influential cases (more on that in Chapter 8). In short, HC3 has been shown to outperform HC0 to HC2 (Long & Ervin, 2000), and consequently the HC3 estimator is widely used in the presence of heteroscedasticity. However, the more recent HC4 estimator (Cribari-Neto, 2004) appears to be more robust than HC3 when there are influential cases and non-normal errors. The bottom line is use the HC3 or HC4 method (see Hayes & Cai, 2007, for an excellent review).

6.10.8 The bootstrap ©

Another way to get robust standard errors is the **bootstrap** (Efron & Tibshirani, 1994). The problem we have is that we don't know the shape of the sampling distribution, but normality in our errors allows us to infer that the sampling distribution is normal. Lack of normality prevents us from inferring the shape of the sampling distribution unless we have big samples. Bootstrapping gets around this problem by estimating the properties of the sampling distribution *empirically* from the sample data. Figure 6.37 illustrates the process: in effect, the sample data are treated as a population from which smaller samples (called bootstrap samples) are taken (putting each score back before a new one is drawn from the sample). Each bootstrap sample is the same size as the observed sample, but it does not contain the same scores because each observed score could be sampled more than once. For example, if your sample scores were 2, 4, 5 and 7, a bootstrap sample might be 2, 5, 5, 5 (the value 5 has been sampled 3 times, while the values 4 and 7 haven't been sampled at all), or 4, 5, 7, 7 (the value 7 has been sampled twice, but the value 2 has not been sampled). You get the idea.

The parameter of interest (e.g., the mean) is estimated in each bootstrap sample. This process is repeated perhaps 2000 times. The result is 2000 parameter estimates, one from each bootstrap sample. There are two things we can do with these estimates: the first is to order them and work out the limits within which 95% of them fall. For example, in Figure 6.37, 95% of bootstrap sample means fall between 5.15 and 6.80. We can use these values as estimate of the limits of the 95% confidence interval of the parameter. The result is known as a percentile bootstrap confidence interval (because it is based on the values between which 95% of bootstrap sample estimates fall). The second thing we can do is to calculate the standard deviation of the parameter estimates from the bootstrap samples and use it as the standard error of the parameter estimates. When we use bootstrapping, we're effectively getting the computer to use our sample data to mimic the sampling process described in Section 2.8. An important

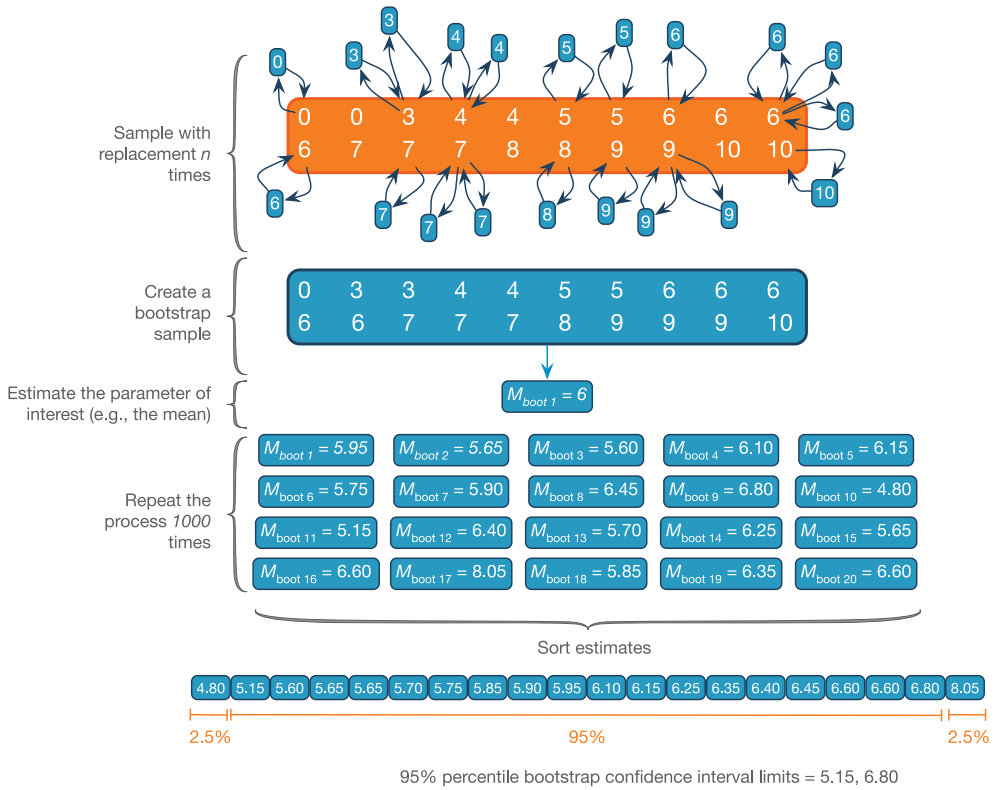


Figure 6.37 Illustration of the bootstrap

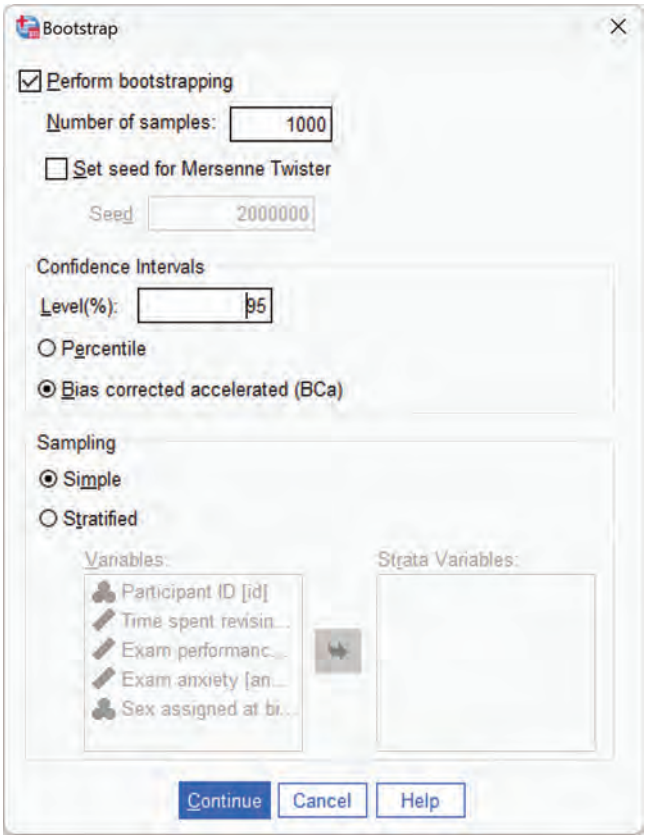


Figure 6.38 The standard *Bootstrap* dialog box

point to remember is that because bootstrapping is based on taking random samples from the data you've collected, the estimates you get will be slightly different each time. This is nothing to worry about. For a fairly gentle introduction to the bootstrap see Wright et al. (2011).

Some procedures in SPSS have a bootstrap option, which can be accessed by clicking Bootstrap... to activate the dialog box in Figure 6.38 (see Oditi's Lantern). Select ☑ Perform bootstrapping to activate bootstrapping for the procedure you're currently doing. In terms of the options, SPSS will compute a 95% percentile confidence interval (◉ Percentile), but you can change the method to a slightly more accurate one called a bias corrected and accelerated confidence interval (Efron & Tibshirani, 1993) by selecting ◉ Bias corrected accelerated (BCa). You can also change the confidence level by typing a number other than 95 in the box labelled *Level(%)*. By default, SPSS uses 1000 bootstrap samples, which is a reasonable number, and you certainly wouldn't need to use more than 2000.

6.11 A final note Ⓑ

There are robust versions of the general linear model, and its special cases such as ANOVA and ANCOVA, that enable you to ignore everything we have discussed about bias in this chapter. That's a happy story, but one with a tragic ending because you can't implement many of them directly in SPSS. The definitive guide to these tests is Wilcox's (2017) outstanding book. Thanks to Wilcox, these tests can be implemented using a free statistics program called R (www.r-project.org). You can access these tests in SPSS Statistics using R and the WRS2 package (Section 4.13.3) and I'll describe some of these tests as we go along. If you want more detail and fancy getting into R (shameless plug alert) try my R



Oditi's Lantern
Bootstrapping

'I, Oditi, believe that R is so-called because it makes you shout "Arrghhh!!?" You, my followers, are precious to me and I would not want you to place your sensitive body parts into that guillotine. Instead, stare into my lantern to see how we can use bootstrapping within SPSS.'



version of this textbook (Field, 2024). For those who don't want to get bogged down with R, SPSS often has options for bootstrapping and heteroscedastic consistent standard errors.

6.12 Jane and Brian's story Ⓐ

Jane had been thinking about Brian. She didn't want to– it was a distraction from work– but he kept returning to her thoughts. Guys didn't interest her, or relationships at all. She didn't understand other people, they were so … unpredictable. Jane liked certainty, it made her feel safe. Like all people, Brian freaked her out with his random invasions of her space, but his determined efforts to impress her were becoming predictable, in a good way. Was it possible that she was starting to like the routine of seeing him? She waved her hand across the electronic lock, placed a small metal box in her pocket, walked away from the door and up the stairs back to ground level. As she emerged from the Pleiades building the light made her recoil. How long had she been down there? As her eyes adjusted, a fuzzy Brian in the distance gained focus. He smiled and waved at her. Usually seeing him skewed her, made her agitated, but today it made her feel safe. This was an interesting development. Was she transforming?

6.13 What next? Ⓐ

This chapter has taught us how to identify bias. Had I read this chapter I might have avoided being influenced by my idolization of my granddad[20] and realized that I could be a useful midfield player

20 Despite worshipping the ground my granddad walked on, I ended up supporting the local rivals of the north London team that he supported.

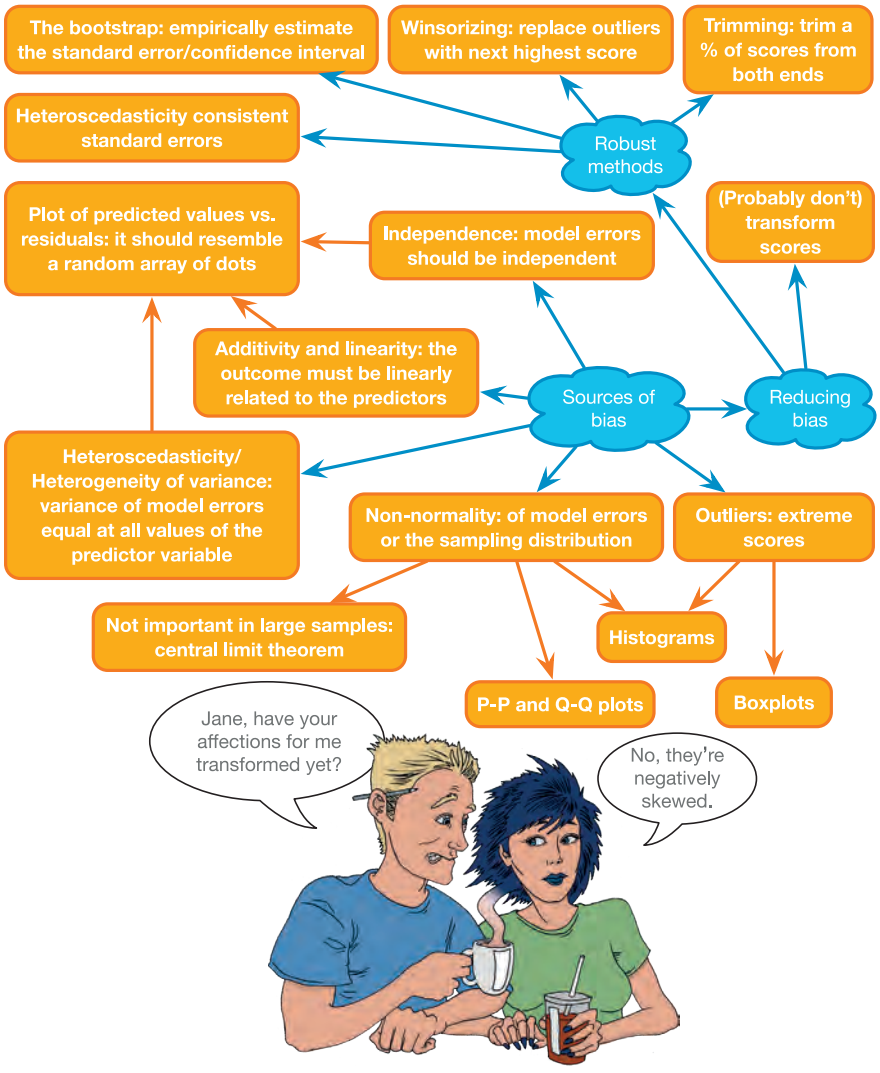
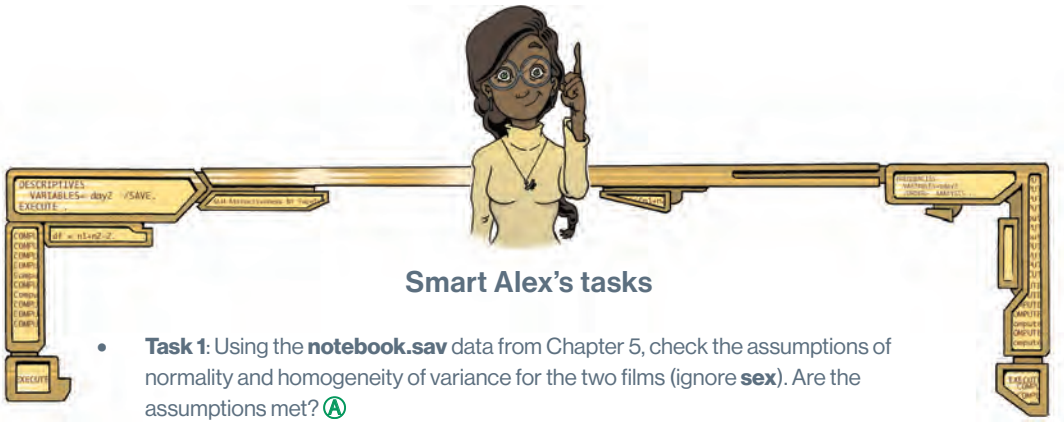


Figure 6.39 What Brian learnt from this chapter

instead of fruitlessly throwing my miniature body around a goalmouth. Had I played midfield a successful career in soccer would undoubtedly have unfolded in front of me. Or, as anyone who has seen me play will realize, perhaps not. I sort of had the last laugh on the goalkeeping front. At the end of my time at primary school we had a five-a-side tournament between local schools so that kids from different schools could get to know each other before going to secondary school together. My goalkeeping nemesis was, of course, chosen to play and I was the substitute. In the first game he had a shocker, and I was called up to play in the second game during which I made a series of dramatic and acrobatic saves (at least that's how I remember them). I did likewise in the next game, and my nemesis had to sit out the whole of the rest of the tournament. Of course, five-a-side goals are shorter than normal goals so I didn't have my usual handicap. When I arrived at secondary school I didn't really know a position other than goalkeeper, and I was still short, so I gave up football. Years later when I started playing again, I regretted the years I didn't spend honing ball skills (my teammates regretted it too). During my non-sporting years I read books and immersed myself in music. Unlike 'the clever one' who was reading Albert Einstein's papers (well, Isaac Asimov) as an embryo, my literary preferences were more in keeping with my intellect ...

6.14 Key terms that I've discovered

Assumption of normality
Bootstrap
Consistent estimator
Contaminated normal distribution
Coverage
Efficient estimator
Hartley's F_{max}
Heterogeneity of variance
Heteroscedasticity
Homogeneity of variance
Homoscedasticity
Independent errors
Kolmogorov–Smirnov test
Levene's test
M-estimator
Mixed normal distribution
Normally distributed data
Outlier
P-P plot
Parametric test
Q-Q plot
Residual
Robust methods
Shapiro–Wilk test
Transformation
Trimmed mean
Unbiased estimator
Variance ratio
Weighted least squares
Winsorizing



Smart Alex's tasks

- **Task 1**: Using the **notebook.sav** data from Chapter 5, check the assumptions of normality and homogeneity of variance for the two films (ignore **sex**). Are the assumptions met? Ⓐ
- **Task 2**: The file **spss_exam.sav** contains data on students' performance on an SPSS exam. Four variables were measured: **exam** (first-year SPSS exam scores as a percentage), **computer** (measure of computer literacy as a percentage), **lecture** (percentage of SPSS lectures attended) and **numeracy** (a measure of numerical ability out of 15). There is a variable called **uni** indicating whether the student attended Sussex University (where I work) or Duncetown University. Compute and interpret descriptive statistics for **exam**, **computer**, **lecture** and **numeracy** for the sample as a whole. Ⓐ
- **Task 3**: Calculate and interpret the *z*-scores for skewness for all variables. Ⓐ
- **Task 4:** Calculate and interpret the *z*-scores for kurtosis for all variables. Ⓐ
- **Task 5**: Use the *split file* command to look at and interpret the descriptive statistics for **numeracy** and **exam**. Ⓐ
- **Task 6**: Repeat Task 5 but for the computer literacy and percentage of lectures attended. Ⓐ
- **Task 7**: Conduct and interpret a K-S test for **numeracy** and **exam**. Ⓐ
- **Task 8**: Conduct and interpret a Levene's test for **numeracy** and **exam**. Ⓐ
- **Task 9**: Transform the **numeracy** scores (which are positively skewed) using one of the transformations described in this chapter. Do the data become normal? Ⓑ
- **Task 10:** Use the *explore* command to see what effect a natural log transformation would have on the four variables measured in **spss_exam.sav**. Ⓑ

View the solutions at www.discoverspss.com



NON-PARAMETRIC MODELS

7.1 What will this chapter tell me? 322
7.2 When to use non-parametric tests 323
7.3 General procedure of non-parametric tests using SPSS 324
7.4 Comparing two independent conditions: the Wilcoxon rank-sum test and Mann–Whitney test 325
7.5 Comparing two related conditions: the Wilcoxon signed-rank test 337
7.6 Differences between several independent groups: the Kruskal–Wallis test 346
7.7 Differences between several related groups: Friedman's ANOVA 359
7.8 Jane and Brian's story 368
7.9 What next? 370
7.10 Key terms that I've discovered 370
Smart Alex's tasks 370

Letters A to D indicate increasing difficulty

7.1 What will this chapter tell me? Ⓐ

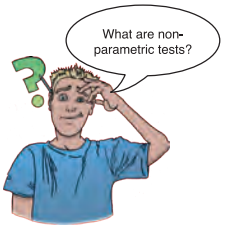
When we were learning to read at primary school, we used to read versions of stories by the famous storyteller Hans Christian Andersen. One of my favourites was the story of the ugly duckling. This duckling was a big ugly grey bird, so ugly that even a dog would not bite him. The poor duckling was ridiculed, ostracized and pecked by the other ducks. Eventually, it became too much for him and he flew to the swans, the royal birds, hoping that they would end his misery by killing him because he was so ugly. Still, life sometimes throws up surprises, and as he stared into the water, he saw not an ugly grey bird but a beautiful swan. Data are much the same. Sometimes they're just big, grey and ugly and don't do any of the things that they're supposed to do. When we get data like these, we swear at them, curse them, peck them and hope that they'll fly away and be killed by the swans. Alternatively, we can try to force our data into becoming beautiful swans. That's what this chapter is all about: trying to make an ugly duckling of a data set turn into a swan. Be careful what you wish your data to be, though: a swan can break your arm.[1]



Figure 7.1 I came first in the competition for who has the smallest brain

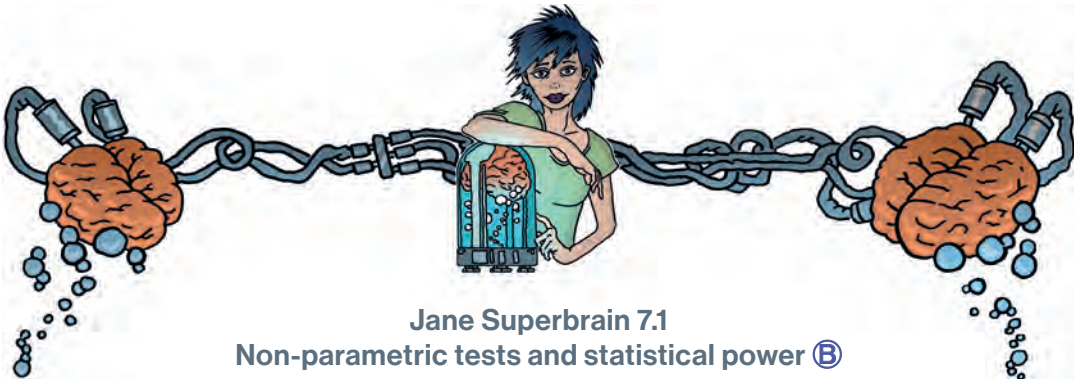
1 Although it is theoretically possible, apparently you'd have to be weak-boned, and swans are nice and wouldn't do that sort of thing.

7.2 When to use non-parametric tests Ⓐ



In the previous chapter we looked at several ways to reduce bias. Sometimes, however, no matter how hard you try, you will find that you can't correct the problems in your data. This is especially irksome if you have a small sample and can't rely on the central limit theorem to get you out of trouble. The historical solution is a small family of models called **non-parametric tests** or 'assumption-free tests' that make fewer assumptions than the linear model that we looked at in the previous chapter.[2] Robust methods have superseded non-parametric tests, but we'll look at them anyway because (1) the range of robust methods in IBM SPSS Statistics software is limited; and (2) non-parametric tests act as a gentle introduction to using a statistical test to evaluate a hypothesis. Some people believe that non-parametric tests have less power than their parametric counterparts, but this is not always true (Jane Superbrain Box 7.1).

In this chapter, we'll explore four of the most common non-parametric procedures: the Mann–Whitney test, the Wilcoxon signed-rank test, Friedman's test and the Kruskal–Wallis test. All four tests overcome distributional problems by **ranking** the data: that is, finding the lowest score and giving it a rank of 1, then finding the next highest score and giving it a rank of 2, and so on. This process results in high scores being represented by large ranks, and low scores being represented by small ranks. The model is then fitted to the ranks and not the raw scores. By using ranks we eliminate the effect of outliers. Imagine you have 20 data points and the two highest scores are 30 and 60 (a difference of 30); these scores will become ranks of 19 and 20 (a difference of 1). In much the same way, ranking irons out problems with skew.



Jane Superbrain 7.1
Non-parametric tests and statistical power Ⓑ

Ranking the data reduces the impact of outliers and weird distributions, but the price you pay is to lose information about the magnitude of differences between scores. Consequently, non-parametric tests can be less powerful than their parametric counterparts. Remember that statistical power (Section 2.10.8) is the ability of a test to find an effect that genuinely exists, so we're saying that if there is a genuine effect then a non-parametric test is less likely to detect it than a parametric one. This statement is true only *if the assumptions described in Chapter 6 are met*. If we use a parametric test and a non-parametric test on the same data, and those data meet the appropriate assumptions, then the parametric test will have greater power to detect the effect than the non-parametric test.

The problem is that to define the power of a test we need to be sure that it controls the Type I error rate (the proportion of times a test will find a significant effect when there is no effect to find – see Section 2.10.6).

2 Some people describe non-parametric tests as 'distribution-free tests' and claim that they make no distributional assumptions. In fact they do, they just don't assume a normal distribution: the ones in this chapter, for example, all assume a continuous distribution.

We saw in Chapter 1 that this error rate is normally set at 5%. When the sampling distribution is normally distributed then the Type I error rate of tests based on this distribution is indeed 5%, and so we can work out the power. However, when the sampling distribution is not normal the Type I error rate of tests based on this distribution won't be 5% (in fact we don't know what it is because it will depend on the shape of the distribution) and so we have no way of calculating power (because it is linked to the Type I error rate – see Section 2.10.8). So, if someone tells you that non-parametric tests have less power than parametric tests, tell them that this is true only if the sampling distribution is normally distributed.



7.3 General procedure of non-parametric tests using SPSS Ⓐ

The tests in this chapter use a common set of dialog boxes, which I'll describe here before we look at the specific tests. If you're comparing groups containing different entities select *Analyze* > *Nonparametric Tests* > Independent Samples... . But if you're comparing scores that were taken from the same entities under different conditions, select *Analyze* > *Nonparametric Tests* > Related Samples... . Both menus take you to a similar dialog box that has three tabs:



As Figure 7.2 shows, regardless of whether you have scores from the same entities or different ones, this tab offers the choice to compare scores automatically (i.e., SPSS selects a test for you, which I don't recommend because it's not a good idea to let a computer think on your behalf), or to select the analysis yourself (Customize analysis).



Selecting this tab activates a screen in which you select the variables that you want to analyse. Within this screen, if you set roles for your variables when you entered the data (Section 4.6.2) then SPSS uses these roles to guess what analysis you want to do (Use predefined roles). If you did not set roles, or if you don't think it's wise to let SPSS guess what you want to do, then you can specify the variables within the analysis yourself (Use custom field assignments). This tab changes depending on whether you have independent or related samples, but in both cases your variables are listed in the left-hand box labelled *Fields* (e.g., Figure 7.7). By default, all variables are listed (All), but you can filter this list to show only nominal/categorical variables ( ) or only scale variables ( ). This button is helpful in big data sets to find your outcome (which will typically be scale,) and predictor variable (which is likely to be nominal,). You can toggle between showing the variable name and the variable label in the list by clicking . The right-hand box labelled *Test Fields* is where you place outcome variables within an analysis, and for tests of independent samples there will be a box labelled *Groups* to place categorical predictors. We'll look at the exact configuration of this tab within each analysis. For now, just note that it's similar regardless of which non-parametric test you do.



Selecting this tab activates the options for what test to perform. You can let SPSS pick a test for you (Automatically choose the tests based on the data), but I'd advise you to make the decisions yourself (Customize tests). Regardless of the test you're doing, you can set the significance level (the default is 0.05), the confidence interval level (the default is 95%) and whether to exclude cases listwise or test by test (see **SPSS Tip 6.1**) by clicking on *Test Options* (see Figure 7.3). Similarly, if you have categorical variables and missing values you can choose to exclude or include these missing values by selecting *User-Missing Values* and checking the appropriate option (see Figure 7.3). The default option is to exclude them, which makes sense a lot of the time.

The general process for any non-parametric analysis, then, is as follows:

1. Choose ◉ Customize analysis in the Objective Fields Settings tab (Figure 7.2), because I don't think you should trust a computer to make your decisions.
2. In the Objective Fields Settings tab, if SPSS fails to correctly guess what analysis you want to do (◉ Use predefined roles) then select ◉ Use custom field assignments and specify your predictor and outcome variables.
3. In the Objective Fields Settings tab you can let SPSS pick a test for you (◉ Automatically choose the tests based on the data), but you have more options if you select ◉ Customize tests. I recommend the latter option. Change the test or missing-values options if necessary, although the defaults are fine (Figure 7.3).

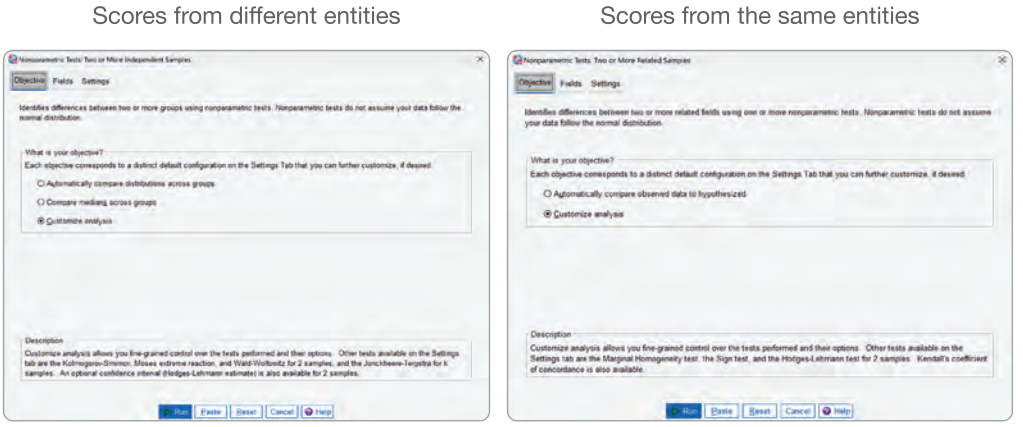


Figure 7.2 Dialog boxes for the *Objective* tab of the *Nonparametric Tests* menu

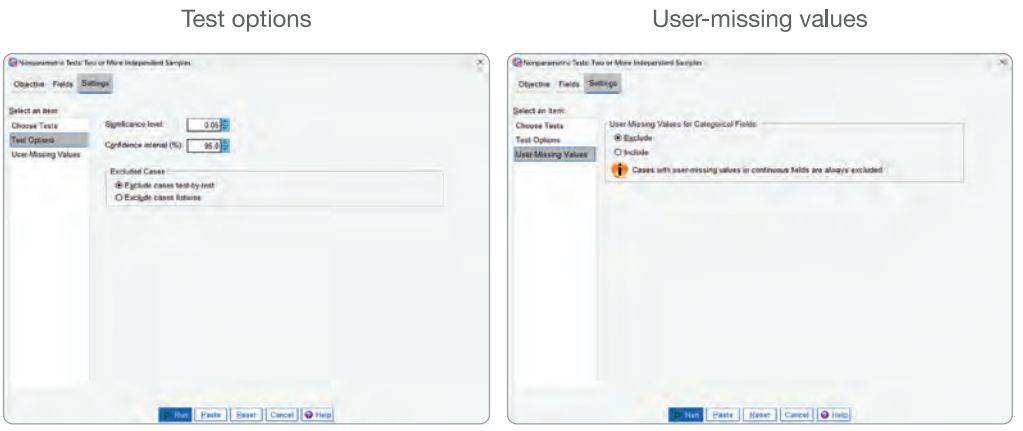
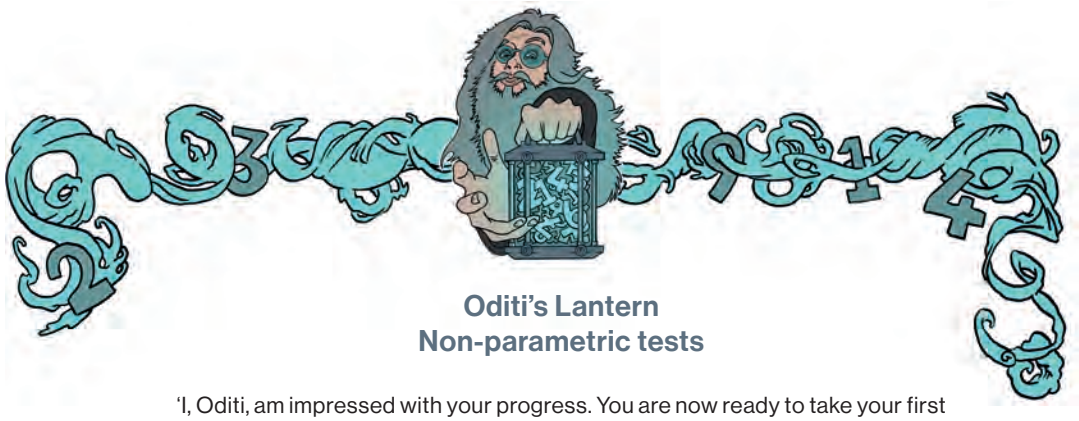


Figure 7.3 Dialog box for the *Settings* tab when choosing *Test Options* and *User-Missing Values*

7.4 Comparing two independent conditions: the Wilcoxon rank-sum test and Mann–Whitney test Ⓐ

Imagine that you have a hypothesis that two groups of different entities will differ from each other on some variable. For example, a psychologist collects data to investigate the depressant effects of two recreational drugs. At a club on a Saturday night she recruited 10 people who had taken an ecstasy



Oditi's Lantern
Non-parametric tests

'I, Oditi, am impressed with your progress. You are now ready to take your first steps towards understanding the hidden meanings behind the data. However, I love and value your precious brains, and do not want them to end up like a fly on a windshield. Stare into my lantern to discover how to test hypotheses with all of the non-parametric tests covered in this chapter.'



tablet but not consumed alcohol and 10 who had drunk alcohol but not taken ecstasy. Levels of depression were measured using the Beck Depression Inventory (BDI) the day after (Sunday) and mid-week (Wednesday). The data are in Table 7.1. She had two hypotheses: between those who took alcohol and those who took ecstasy, depression levels will be different the day after (hypothesis 1) and mid-week (hypothesis 2). To test these hypotheses, we need to fit a model that compares the distribution in the alcohol group to that in the ecstasy group.



There are two choices to compare the distributions in two conditions containing scores from different entities: the **Mann–Whitney test** (Mann & Whitney, 1947) and the **Wilcoxon rank-sum test** (Wilcoxon, 1945). Both tests are equivalent, and to add to the confusion there's a second Wilcoxon test that tests something different.

7.4.1 Theory Ⓑ

The logic behind the Wilcoxon rank-sum and Mann–Whitney tests is incredibly elegant. First, let's imagine a scenario in which there is no difference in depression levels between ecstasy and alcohol users. If you were to rank the data *ignoring the group to which a person belonged* from lowest to highest (i.e., give the lowest score a rank of 1 and the next lowest a rank of 2, etc.), if there's no difference between the groups then you should find a similar number of high and low ranks in each group; specifically, if you added up the ranks, then you'd expect the summed total of ranks in each group to

Table 7.1 Data for drug experiment

Participant	Drug	BDI (Sunday)	BDI (Wednesday)
1	Ecstasy	15	28
2	Ecstasy	35	35
3	Ecstasy	16	35
4	Ecstasy	18	24
5	Ecstasy	19	39
6	Ecstasy	17	32
7	Ecstasy	27	27
8	Ecstasy	16	29
9	Ecstasy	13	36
10	Ecstasy	20	35
11	Alcohol	16	5
12	Alcohol	15	6
13	Alcohol	20	30
14	Alcohol	15	8
15	Alcohol	16	9
16	Alcohol	13	7
17	Alcohol	14	6
18	Alcohol	19	17
19	Alcohol	18	3
20	Alcohol	18	10

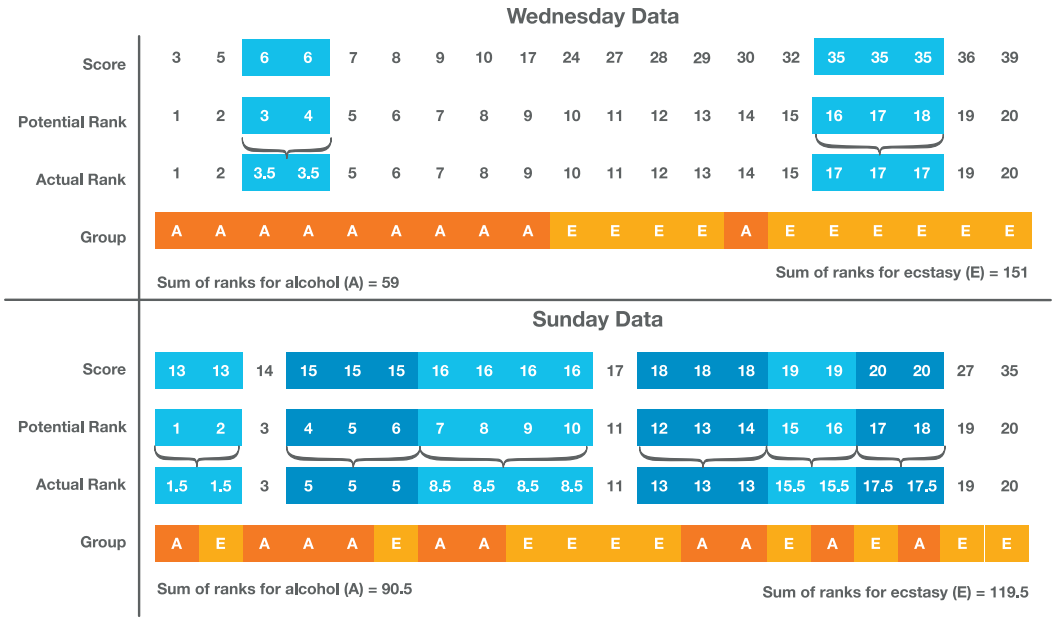


Figure 7.4 Ranking the depression scores

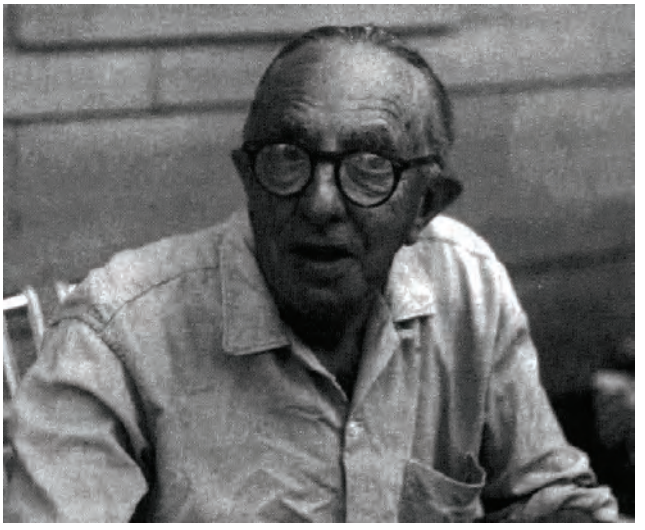
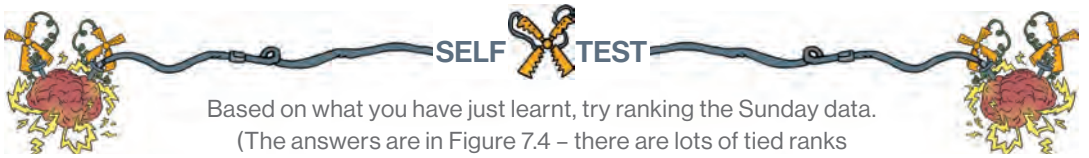


Figure 7.5 Frank Wilcoxon

be about the same. Now let's imagine that the ecstasy group is more depressed than the alcohol group. What do you think would happen to the ranks? If you rank the scores as before, then you would expect more higher ranks to be in the ecstasy group and more lower ranks to be in the alcohol group. Again, if we summed the ranks in each group, we'd expect the sum of ranks to be higher in the ecstasy group than in the alcohol group. The Mann–Whitney and Wilcoxon rank-sum tests use this principle. In fact, when the groups have unequal numbers of participants in them, the test statistic (W_s) for the Wilcoxon rank-sum test is the sum of ranks in the group that contains the fewer people; when the group sizes are equal it's the value of the smaller summed rank.



Let's look at how ranking works. Figure 7.4 shows the ranking process for both the Wednesday and Sunday data. To begin with, let's focus on Wednesday, because the ranking is more straightforward. First, we arrange the scores in ascending order and attach a label to remind us from which group each score came (I've used A for alcohol and E for ecstasy). Starting at the lowest score, we assign potential ranks starting with 1 and going up to the number of scores we have. I've called these 'potential ranks' because sometimes the same score occurs more than once in a data set (e.g., in these data a score of 6 occurs twice, and a score of 35 occurs three times). These are called *tied ranks*, and we rank them with the value of the average potential rank for those scores. For example, our two scores of 6 would've been ranked as 3 and 4, so we assign a rank of 3.5, the average of these values. Likewise, the three scores of 35 have potential ranks of 16, 17 and 18, so we assign a rank of 17, the average of these three values ((16 + 17 + 18)/3 = 17). Once we've ranked the data, we add the ranks for the two groups. First, we add the ranks of the scores from the alcohol group (you should find the sum is 59) and then add the ranks of the scores from the ecstasy group (this value is 151). Our test statistic is the lower of these sums, which for these data is the sum for the alcohol group, $W_s = 59$.



Based on what you have just learnt, try ranking the Sunday data. (The answers are in Figure 7.4 – there are lots of tied ranks and the data are generally horrible.)

Having done the self-test, you should find that the sum of ranks is 90.5 for the alcohol group and 119.5 for the ecstasy group. The test statistic is the lower of these sums, which is the sum for the alcohol group, $W_s = 90.5$.

How do we determine whether this test statistic is significant? It turns out that the mean ($\bar{W}_s$) and standard error ($SE_{\bar{W}_s}$) of this test statistic can be calculated from the sample sizes of each group (n_1 is the sample size of group 1 and n_2 is the sample size of group 2):

$$\bar{W}_s = \frac{n_1(n_1 + n_2 + 1)}{2} \tag{7.1}$$

$$SE_{\bar{W}_s} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{7.2}$$

We have equal-sized groups with 10 people in each, so n_1 and n_2 are both 10. Therefore, the mean and standard deviation are:

$$\bar{W}_s = \frac{10(10 + 10 + 1)}{2} = 105 \tag{7.3}$$

$$SE_{\bar{W}_s} = \sqrt{\frac{(10 \times 10)(10 + 10 + 1)}{12}} = 13.23 \tag{7.4}$$

If we know the test statistic, the mean test statistic and the standard error, then we can convert the test statistic to a *z*-score using the equation that we came across in Chapter 1 (Eq. 1.11):

$$z = \frac{X - \bar{X}}{s} = \frac{W_s - \bar{W}_s}{SE_{\bar{W}_s}} \tag{7.5}$$

We also know that we can use Table A.1 in the Appendix to ascertain a *p*-value for a *z*-score (and more generally that a *z* greater than 1.96 or smaller than –1.96 is significant at $p \leq 0.05$). The *z*-scores for the Sunday and Wednesday depression scores are:

$$z_{\text{Sunday}} = \frac{W_s - \bar{W}_s}{SE_{\bar{W}_s}} = \frac{90.5 - 105}{13.23} = -1.10 \tag{7.6}$$

$$z_{\text{Wednesday}} = \frac{W_s - \bar{W}_s}{SE_{\bar{W}_s}} = \frac{59 - 105}{13.23} = -3.48 \tag{7.7}$$

So, there is a significant difference between the groups on Wednesday, but not on Sunday.

The procedure I've described is the Wilcoxon rank-sum test. The Mann–Whitney test is basically the same, but uses a test statistic *U*, which has a direct relationship with the Wilcoxon test statistic. SPSS produces both statistics in the output. If you're interested, *U* is calculated using an equation in which n_1 and n_2 are the sample sizes of groups 1 and 2 respectively, and R_1 is the sum of ranks for group 1:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{7.8}$$

For our data we'd get the following (remember we have 10 people in each group and the sum of ranks for group 1, the ecstasy group, was 119.5 for the Sunday data and 151 for the Wednesday data):

$$U_{\text{Sunday}} = (10 \times 10) + \frac{10(11)}{2} - 119.50 = 35.50 \tag{7.9}$$

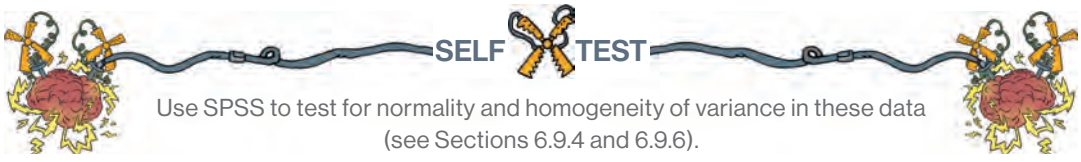
$$U_{\text{Wednesday}} = (10 \times 10) + \frac{10(11)}{2} - 151.00 = 4.00 \tag{7.10}$$

7.4.2 Inputting data and provisional analysis Ⓐ



The data editor will have three columns. The first column is a coding variable (called something like **Drug**), which will have two codes (for convenience I suggest 1 = ecstasy group and 2 = alcohol group). When you enter this variable into SPSS remember to enter value labels to set the codes as discussed in Section 4.6.5). The second column will have values for the dependent variable (BDI) measured the day after (call this variable **sunday_bdi**) and the third will have the midweek scores on the same questionnaire (call this variable **wednesday_bdi**). You can, if you like, add a fourth column that is a variable to identify the participant (with a code or number). Save the file as **drug.sav**.

Your first step should always be plots and exploratory analysis. Given that we have a small sample (10 per group), we know that tests of normality and homogeneity of variance will be underpowered (Jane Superbrain Box 6.7), and so if they are significant we can assume there's a problem. If they are not significant it will tell us very little other than the tests are underpowered. In any case, the plots will be informative. For normality, because we're looking for group differences, we need to run the analyses separately for each group.



The results of our exploratory analysis are shown in Output 7.1 and Figure 7.6. The normal Q-Q plots show quite clear deviations from normality for ecstasy on Sunday and alcohol on Wednesday because the dots deviate from the diagonal line. The tables in Output 7.1 confirm these observations: in the ecstasy group the Sunday BDI scores, $D(10) = 0.28$, $p = 0.03$, and in the alcohol group the Wednesday BDI scores, $D(10) = 0.31$, $p = 0.009$, significantly deviate from normal even using underpowered tests. In the alcohol group the Sunday BDI scores, $D(10) = 0.17$, $p = 0.20$, and in the ecstasy group the Wednesday BDI scores, $D(10) = 0.24$, $p = 0.13$, do not significantly deviate from a normal distribution. We can't read too much into the non-significance of these two tests (because they are underpowered) but in the corresponding Q-Q plots (Figure 7.6) the observed data (dots) do hug the diagonal line, suggesting they are approximately normally distributed. In any case, a non-parametric test seems appropriate. Levene's test shows that the variances are not significantly different between the drug groups on Sunday, $F(1, 18) = 3.64$, $p = 0.072$, and Wednesday, $F(1, 18) = 0.51$, $p = 0.485$ (Output 7.1, bottom table), suggesting that the assumption of homogeneity has been met. However, this test is again hugely underpowered so the $p = 0.072$ for the Sunday BDI scores is not reassuring.

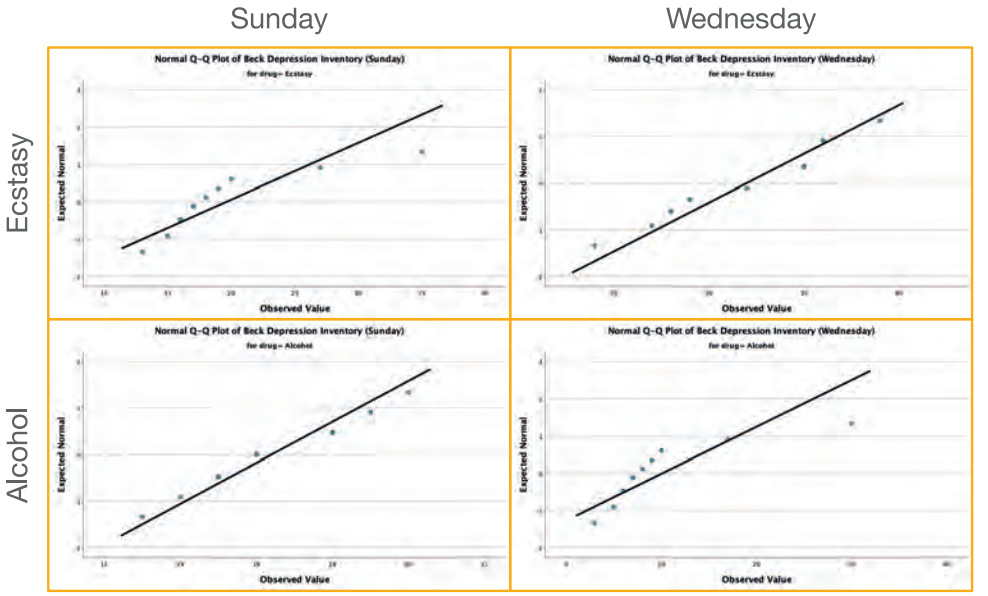


Figure 7.6 Normal Q-Q plots of depression scores after ecstasy and alcohol on Sunday and Wednesday

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Type of drug	Statistic	df	Sig.	Statistic	df	Sig.
Beck Depression Inventory (Sunday)	Ecstasy	.276	10	.030	.811	10	.020
	Alcohol	.170	10	.200*	.959	10	.780
Beck Depression Inventory (Wednesday)	Ecstasy	.235	10	.126	.941	10	.566
	Alcohol	.305	10	.009	.753	10	.004

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

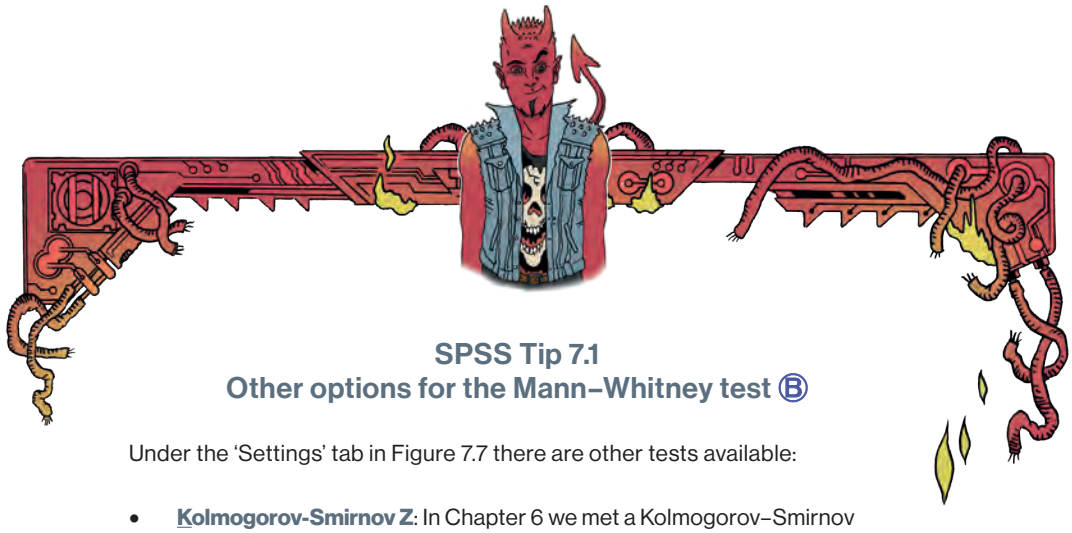
Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Beck Depression Inventory (Sunday)	Based on Mean	3.644	1	18	.072
	Based on Median	1.880	1	18	.187
	Based on Median and with adjusted df	1.880	1	10.076	.200
	Based on trimmed mean	2.845	1	18	.109
Beck Depression Inventory (Wednesday)	Based on Mean	.508	1	18	.485
	Based on Median	.091	1	18	.766
	Based on Median and with adjusted df	.091	1	11.888	.768
	Based on trimmed mean	.275	1	18	.606

Output 7.1

7.4.3 The Mann–Whitney test using SPSS Statistics Ⓐ

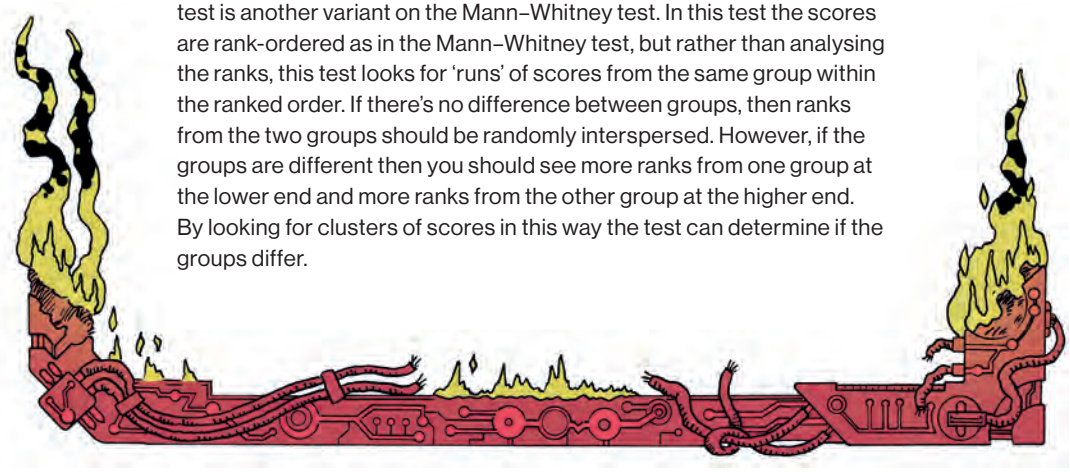
To run a Mann–Whitney test follow the general procedure outlined in Section 7.3 by first selecting *Analyze* > *Nonparametric Tests* > Independent Samples... . When you reach the Objective Fields Settings tab, if you need to assign variables then select ◉ Use custom field assignments and specify the model (Figure 7.7 top) by dragging **Beck Depression Inventory (Sunday)** and **Beck Depression Inventory (Wednesday)** into the box labelled *Test Fields* (or select them in the box labelled *Fields* and click ▸). Next, transfer **Type of drug** to the box labelled *Groups*. Activate the Objective Fields Settings tab, select ◉ Customize tests and check ☑ Mann-Whitney U (2 samples) (Figure 7.7, bottom). The dialog box lists tests other than the Mann–Whitney test, which are explained in SPSS Tip 7.1. Click ▶ Run to run the analysis.



SPSS Tip 7.1
Other options for the Mann–Whitney test Ⓑ

Under the 'Settings' tab in Figure 7.7 there are other tests available:

- **Kolmogorov-Smirnov Z**: In Chapter 6 we met a Kolmogorov–Smirnov test that tested whether a sample was from a normally distributed population. This test is different: it tests whether two groups have been drawn from the same population (regardless of what that population may be). In effect, this test does much the same as the Mann–Whitney test, but it tends to have more power when sample sizes are less than about 25 per group, and so is worth selecting if that's the case.
- **Moses Extreme Reactions**: This test makes me think of a bearded man standing on Mount Sinai reading a stone tablet and then suddenly bursting into a wild rage, smashing the tablet and screaming, 'What do you mean, do not worship any other God?' This test is nowhere near as exciting as my mental image. It's a bit like a non-parametric Levene's test (Section 6.9.6) that compares the variability of scores across two groups.
- **Wald-Wolfowitz runs**: Despite sounding like a bad case of diarrhoea, this test is another variant on the Mann–Whitney test. In this test the scores are rank-ordered as in the Mann–Whitney test, but rather than analysing the ranks, this test looks for 'runs' of scores from the same group within the ranked order. If there's no difference between groups, then ranks from the two groups should be randomly interspersed. However, if the groups are different then you should see more ranks from one group at the lower end and more ranks from the other group at the higher end. By looking for clusters of scores in this way the test can determine if the groups differ.



7.4.4 Output from the Mann–Whitney test Ⓐ

I explained earlier that the Mann–Whitney test works by looking at differences in the ranked positions of scores in different groups. The first part of the output is a plot summarizing the data after they have been ranked. SPSS shows us the distribution of ranks in the two groups (alcohol and ecstasy) and the mean rank in each condition (see Output 7.2). Remember that the Mann–Whitney test relies on scores

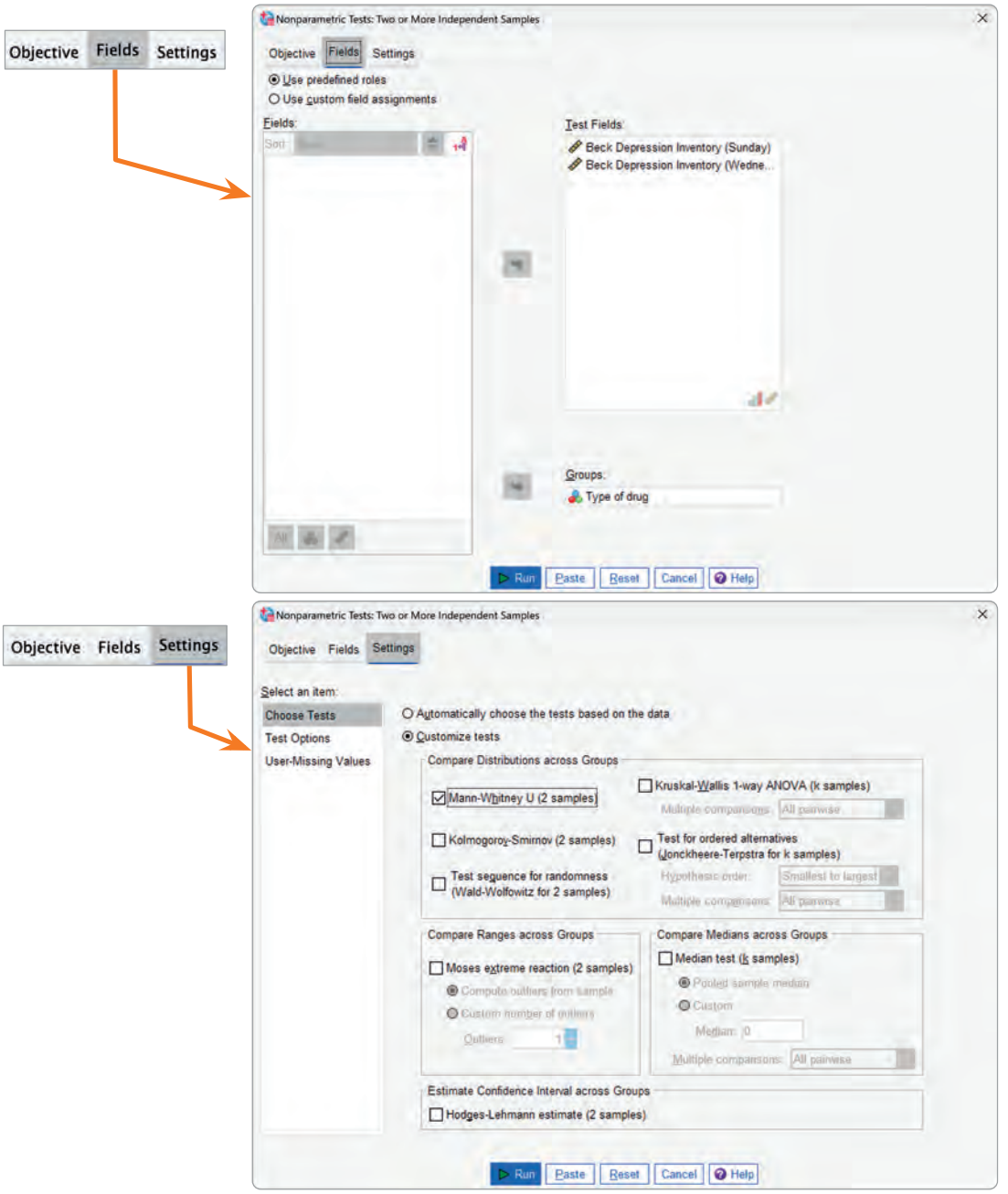
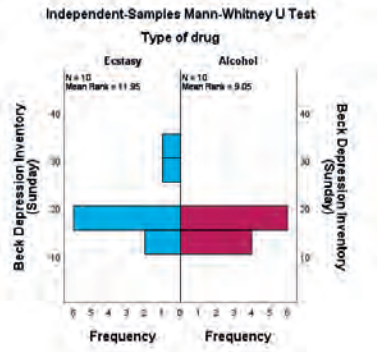


Figure 7.7 Dialog boxes for the Mann–Whitney test

being ranked from lowest to highest; therefore, the group with the lowest mean rank is the group with the greatest number of lower scores in it. Conversely, the group that has the highest mean rank should have a greater number of high scores within it. Therefore, this plot can be used to ascertain which group had the highest scores, which is useful for interpretation. For example, we can see for the Sunday data that the distributions in the two groups are almost identical (the ecstasy has a couple of higher ranks but otherwise the bars look the same) and the mean ranks are similar (11.95 and 9.05); on Wednesday, however, the distribution of ranks is shifted upwards in the ecstasy group compared to the alcohol group, which is reflected in a much larger mean rank (15.10 compared to 5.90).

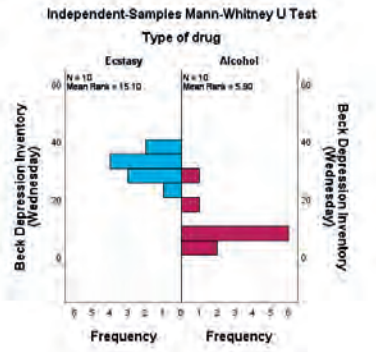
Sunday



Independent-Samples Mann-Whitney U Test Summary

Total N	20
Mann-Whitney U	35.500
Wilcoxon W	90.500
Test Statistic	35.500
Standard Error	13.124
Standardized Test Statistic	-1.105
Asymptotic Sig.(2-sided test)	.269
Exact Sig.(2-sided test)	.280

Wednesday



Independent-Samples Mann-Whitney U Test Summary

Total N	20
Mann-Whitney U	4.000
Wilcoxon W	59.000
Test Statistic	4.000
Standard Error	13.204
Standardized Test Statistic	-3.484
Asymptotic Sig.(2-sided test)	<.001
Exact Sig.(2-sided test)	.000

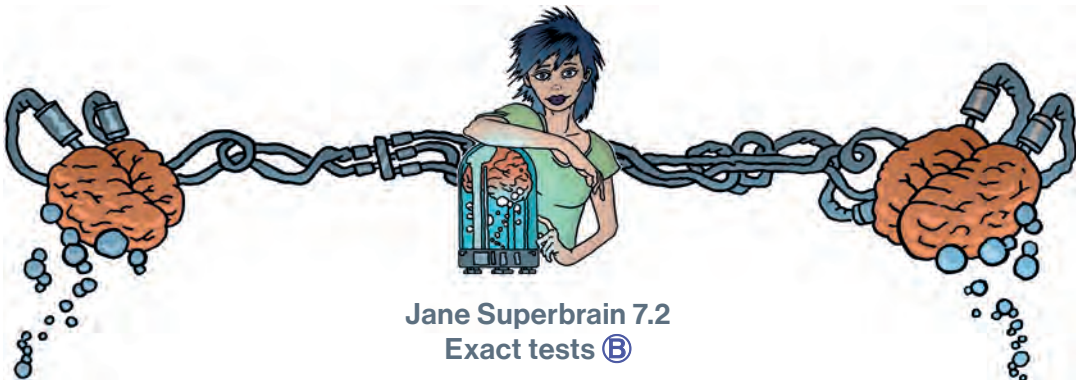
Output 7.2

Underneath the plot a table shows the test statistics for the Mann–Whitney test, the Wilcoxon procedure and the corresponding *z*-score. Note that the values of U, W_s and the *z*-score are the same as we calculated in Section 7.4.1 (phew!). The rows labelled *Asymptotic Sig.* and *Exact Sig.* tell us the probability that a test statistic of at least that magnitude would occur if there were no difference between groups. The two *p*-values are interpreted in the same way but are computed differently: our sample is small so we'll use the exact method (see Jane Superbrain Box 7.2). For these data, the Mann–Whitney test is non-significant for the depression scores taken on the Sunday because the *p*-value of 0.280 is greater than the critical value of 0.05. This finding indicates that ecstasy is no more of a depressant, the day after taking it, than alcohol: both groups report comparable levels of depression. This confirms what we concluded from the mean ranks and distribution of ranks. For the midweek measures the results are highly significant because the exact *p*-value of 0.000 is less than the critical value of 0.05. In this case we write $p < 0.001$ because the observed *p* is very small indeed. This finding also confirms what we suspected based on the distribution of ranks and mean ranks: the ecstasy group (mean rank = 15.10) had significantly higher levels of depression midweek than the alcohol group (mean rank = 5.90).

7.4.5 Calculating an effect size Ⓑ

SPSS doesn't calculate an effect size for us, but we can calculate something called the **rank-biserial correlation** (Cureton, 1956) which can be interpreted like any other correlation coefficient (see Kerby, 2014, for a lovely intuitive interpretation of this particular effect size). There are various ways to compute this correlation (Kerby, 2014) but for our purposes it makes sense to use a method the directly involves the *U*-statistic devised by Wendt (1972):

$$r_{rb} = 1 - \frac{2U}{n_1 n_2} \tag{7.11}$$



Jane Superbrain 7.2
Exact tests Ⓑ

If you have the premium version of IBM SPSS Statistics, the *p*-value for non-parametric tests is computed in two ways. The *asymptotic method* gives you a sort of approximation that in large samples will be a perfectly serviceable answer. However, when samples are small, or the data are particularly poorly distributed, it doesn't give you a good answer. The *exact method* is more computationally difficult (but we don't care because our computer is doing the computations for us) and gives us an exact significance value. You should use this exact significance in small samples (by which I mean anything under 50 really). There is a third method, which isn't available through the non-parametric menus that we're using, but is available for some other tests, so we might as well learn about it now. The **Monte Carlo method**[3] is slightly less labour-intensive method than computing an exact *p*-value. This method is like bootstrapping (Section 6.10.8) and involves creating a distribution similar to that found in the sample and then taking several samples (the default is 10,000) from this distribution. From those samples the mean significance value and the confidence interval around it can be created.



Output 7.2 tells us that U is 35.5 for the Sunday data and 4 for the Wednesday data and we had 10 people per group so $n_1 = n_2 = 10$. This gives us

$$\begin{aligned} r_{rb(\text{Sunday})} &= 1 - \frac{2 \times 35.5}{10 \times 10} \\ &= 1 - \frac{71}{100} \\ &= 1 - 0.71 \\ &= 0.29 \end{aligned} \tag{7.12}$$

3 It's called the Monte Carlo method because back in the late nineteenth century when Karl Pearson was trying to simulate data he didn't have a computer to do it for him. So he used to toss coins. A lot. That is, until a friend suggested that roulette wheels, if unbiased, were excellent random number generators. Rather than trying to persuade the Royal Society to fund trips to Monte Carlo casinos to collect data from their roulette wheels, he purchased copies of *Le Monaco*, a weekly Paris periodical that published exactly the data that he required, at the cost of 1 franc (Pearson, 1894; Plackett, 1983). When simulated data are used to test a statistical method, or to estimate a statistic, it is known as the Monte Carlo method even though we use computers now and not roulette wheels.

$$r_{rb(\text{Wednesday})} = 1 - \frac{2 \times 4}{10 \times 10} = 1 - \frac{8}{100} = 1 - 0.08 = 0.92 \quad (7.13)$$

Bearing in mind that when we ranked, group 1 was the ecstasy group (i.e. R_1 in Eq. 7.8 represented the ecstasy group ranks) we can interpret these effects as follows (Kerby, 2014). On Sunday about 29% of the ranked data support the idea that depression was higher after ecstasy than after alcohol, but on Wednesday about 92% of the ranked data support the idea that depression was higher after ecstasy than after alcohol. In other words, this is very strong support for the idea that depression is higher after ecstasy than after alcohol 4 days after consuming it, but rather weak evidence that depression is higher after ecstasy than after alcohol 1 day after consuming it.

7.4.6 Writing the results Ⓐ

For the Mann–Whitney test, report only the test statistic (denoted by U) and its significance. In keeping with good practice (Section 3.8), include the effect size and report exact values of p (rather than summary values such as $p \leq 0.05$). We could report something like:



Cramming Sam's Tips
Mann–Whitney Test

- The Mann–Whitney test and Wilcoxon rank-sum test compare two conditions when different participants take part in each condition and the resulting data have unusual cases or violate any assumption in Chapter 6.
- Look at the row labelled *Asymptotic Sig.* or *Exact Sig.* (if your sample is small). If the value is less than .05 then the two groups are significantly different.
- The values of the mean ranks tell you how the groups differ (the group with the highest scores will have the highest mean rank).
- Report the U-statistic (or W_s if you prefer), the corresponding z and the significance value. Also report the medians and their corresponding ranges (or draw a boxplot).
- Calculate the rank-biserial correlation as an effect size and report this statistic too.



- Depression levels in ecstasy users (*Mdn* = 17.50) did not differ significantly from alcohol users (*Mdn* = 16.00) the day after the drugs were taken, $U = 35.50$, $z = -1.11$, $p = 0.280$, $r_{rb} = 0.29$. However, by Wednesday, ecstasy users (*Mdn* = 33.50) were significantly more depressed than alcohol users (*Mdn* = 7.50), $U = 4.00$, $z = -3.48$, $p < .001$, $r_{rb} = 0.92$.

I've reported the median for each condition because this statistic is more appropriate than the mean for non-parametric tests. You can get these values by running descriptive statistics, or you could report the mean ranks instead of the median. If you want to report Wilcoxon's test rather than Mann–Whitney's *U* statistic you could write:

- Depression levels in ecstasy users (*Mdn* = 17.50) did not significantly differ from alcohol users (*Mdn* = 16.00) the day after the drugs were taken, $W_s = 90.50$, $z = -1.11$, $p = 0.280$, $r_{rb} = 0.29$. However, by Wednesday, ecstasy users (*Mdn* = 33.50) were significantly more depressed than alcohol users (*Mdn* = 7.50), $W_s = 59.00$, $z = -3.48$, $p < 0.001$, $r_{rb} = 0.92$.

7.5 Comparing two related conditions: the Wilcoxon signed-rank test Ⓐ

The **Wilcoxon signed-rank test** (Wilcoxon, 1945), not to be confused with the rank-sum test in the previous section, is used in situations where you want to compare two sets of scores that are related in some way (e.g., they come from the same entities). Imagine the psychologist in the previous section was interested in the *change* in depression levels, within people, for each of the two drugs. She now wants to compare the BDI scores on Sunday to those on Wednesday. Remember that the distributions of scores for both drugs were non-normal on one of the two days, implying (because the sample is small) that the sampling distribution will be non-normal too (see Output 7.1), so the psychologist would have to use a non-parametric test.

7.5.1 Theory of the Wilcoxon signed-rank test Ⓑ

The Wilcoxon signed-rank test is based on ranking the differences between scores in the two conditions you're comparing. Once these differences have been ranked (just like in Section 7.4.1), the sign of the difference (positive or negative) is assigned to the rank.

Table 7.2 shows the ranking for comparing depression scores on Sunday to those on Wednesday for the two drugs separately. First, we calculate the difference between scores on Sunday and Wednesday (that's just Sunday's score subtracted from Wednesday's). If the difference is zero (i.e., the scores are the same on Sunday and Wednesday) we exclude this score from the ranking. We make a note of the sign of the difference (positive or negative) and then rank the differences (starting with the smallest), ignoring whether they are positive or negative. The ranking process is the same as in Section 7.4.1, and we deal with tied scores in the same way. Finally, we collect together the ranks that came from a positive difference between the conditions and add them up to get the sum of positive ranks (T_+). We also add up the ranks that came from negative differences between the conditions to get the sum of negative ranks (T_-). For ecstasy, $T_+ = 36$ and $T_- = 0$ (in fact there were no negative ranks), and for alcohol, $T_+ = 8$ and $T_- = 47$. The test statistic is T_+, and so it is 36 for ecstasy and 8 for alcohol.

To calculate the significance of the test statistic (T), we again look at the mean ($\bar{T}$) and standard error ($SE_{\bar{T}}$), which, like the Mann–Whitney and rank-sum test in the previous section, are functions of the sample size, n (because data for the two days come from the same participants, there is only one sample size):

$$\bar{T} = \frac{n(n+1)}{4} \tag{7.14}$$

Table 7.2 Ranking data in the Wilcoxon signed-rank test

BDI Sunday	BDI Wednesday	Difference	Sign	Rank	Positive Ranks	Negative Ranks
Ecstasy						
15	28	13	+	2.5	2.5	
35	35	0	Exclude			
16	35	19	+	6	6	
18	24	6	+	1	1	
19	39	20	+	7	7	
17	32	15	+	4.5	4.5	
27	27	0	Exclude			
16	29	13	+	2.5	2.5	
13	36	23	+	8	8	
20	35	15	+	4.5	4.5	
				Total =	36	0
Alcohol						
16	5	−11	−	9		9
15	6	−9	−	7		7
20	30	10	+	8	8	
15	8	−7	−	3.5		3.5
16	9	−7	−	3.5		3.5
13	7	−6	−	2		2
14	6	−8	−	5.5		5.5
19	17	−2	−	1		1
18	3	−15	−	10		10
18	10	−8	−	5.5		5.5
				Total =	8	47

$$SE_{\bar{T}} = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{7.15}$$

In both groups, n is 10 (the number of participants), but two people were excluded in the ecstasy group because they had differences of zero, so for that group we actually use $n = 8$, which gives

$$\bar{T}_{\text{Ecstasy}} = \frac{8(8+1)}{4} = 18 \tag{7.16}$$

$$SE_{\bar{T}_{\text{Ecstasy}}} = \sqrt{\frac{8(8+1)(16+1)}{24}} = 7.14 \tag{7.17}$$

For the alcohol group there were no exclusions, so we get:

$$\bar{T}_{\text{Alcohol}} = \frac{10(10+1)}{4} = 27.50 \tag{7.18}$$

$$SE_{\bar{T}_{\text{Alcohol}}} = \sqrt{\frac{10(10+1)(20+1)}{24}} = 9.81 \tag{7.19}$$

As with the Mann–Whitney test, if we know the test statistic, the mean of test statistics and the standard error, then we can convert the test statistic to a *z*-score using the standard equation that we revisited in the previous section:

$$z = \frac{X - \bar{X}}{s} = \frac{T - \bar{T}}{SE_{\bar{T}}} \tag{7.20}$$

If we calculate the z for the ecstasy and alcohol depression scores we get:

$$z_{\text{Ecstasy}} = \frac{T - \bar{T}}{SE_{\bar{T}}} = \frac{36 - 18}{7.14} = 2.52 \tag{7.21}$$

$$z_{\text{Alcohol}} = \frac{T - \bar{T}}{SE_{\bar{T}}} = \frac{8 - 27.5}{9.81} = -1.99 \tag{7.22}$$

If these values are greater than 1.96 (ignoring the minus sign) then the test is significant at $p \leq 0.05$. So, it looks as though there is a significant difference between depression scores on Wednesday and Sunday for both ecstasy and alcohol.

7.5.2 The *split file* command Ⓐ



We want to look at the change for each drug *separately*, which means performing the Wilcoxon signed-rank test on BDI scores for the alcohol group and then repeating the analysis for those in the ecstasy group. The *split file* command allows you to specify a coding variable that SPSS uses to carry out separate analyses on each category of cases. Therefore, to perform the Wilcoxon signed-rank test on BDI scores for the alcohol and ecstasy groups separately we would specify the variable **drug** as the variable to split the file by and then specify the Wilcoxon test (which we'll get to in the next section). The Wilcoxon test will then be performed separately at each level of **drug**.

To split the file, select *Data* > Split or click . In the resulting dialog box (Figure 7.8) select Organize output by groups. Once this option is selected, the *Groups Based on* dialog box will activate. Select the variable containing the group codes by which you wish to repeat the analysis (in this example, select **drug**) and drag it to the box or click . Finally, click on OK. By default, SPSS will sort the file by these groups (i.e., it will list one category followed by the other in the data editor).

7.5.3 Doing the Wilcoxon signed-rank test using SPSS Ⓐ

Having split the file, we can run the Wilcoxon tests by following the general procedure outlined in Section 7.3. Select *Analyze* > *Nonparametric Tests* > Related Samples... and when you reach the Objective Fields Settings tab (if SPSS hasn't correctly guessed what you want to do) select Use custom field assignments and drag **Beck**

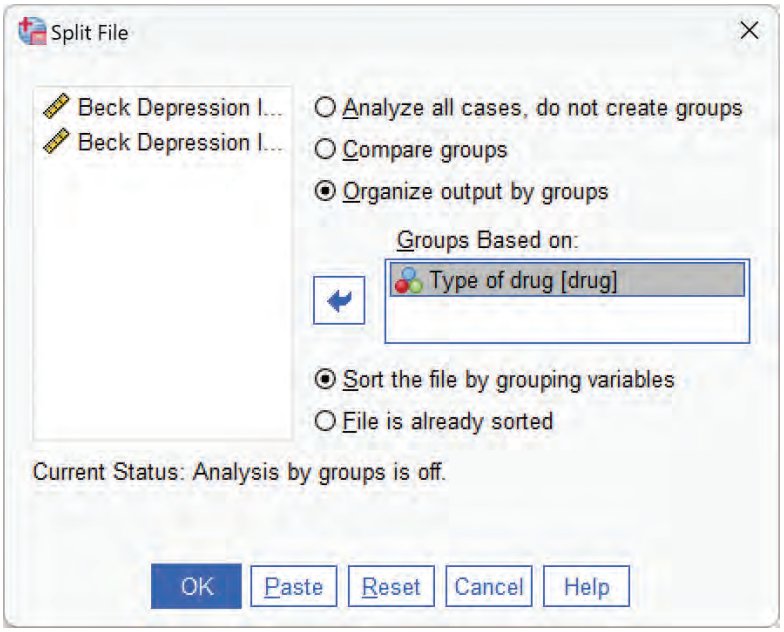
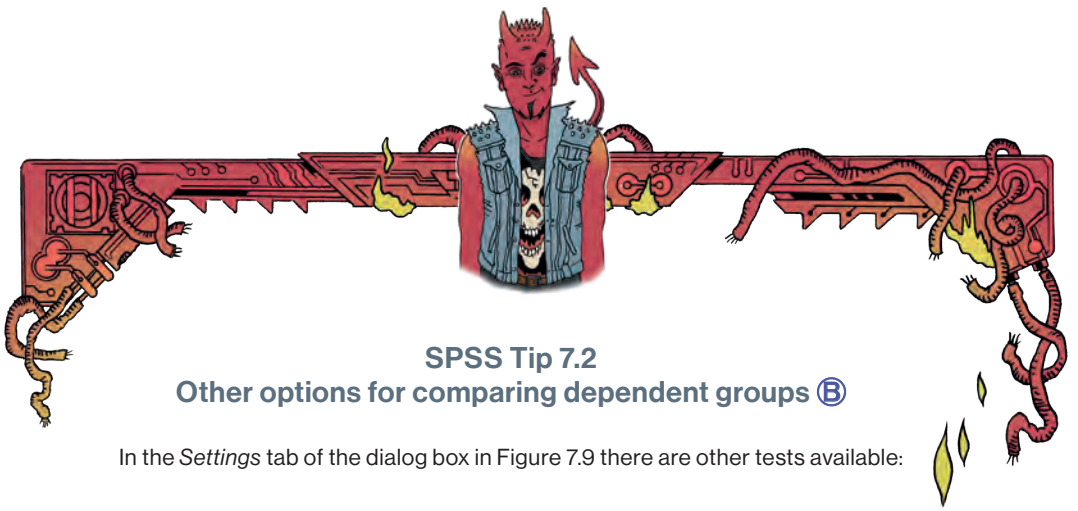


Figure 7.8 Dialog box for the *split file* command

Depression Inventory (Sunday) and **Beck Depression Inventory (Wednesday)** to the box labelled *Test Fields* (or select them in the box labelled *Fields* and click) as in Figure 7.9. Activate the Objective Fields Settings tab, select Customize tests, check Wilcoxon matched-pair signed-rank (2 samples) (Figure 7.9, bottom) and click Run. Other options are explained in SPSS Tip 7.2.



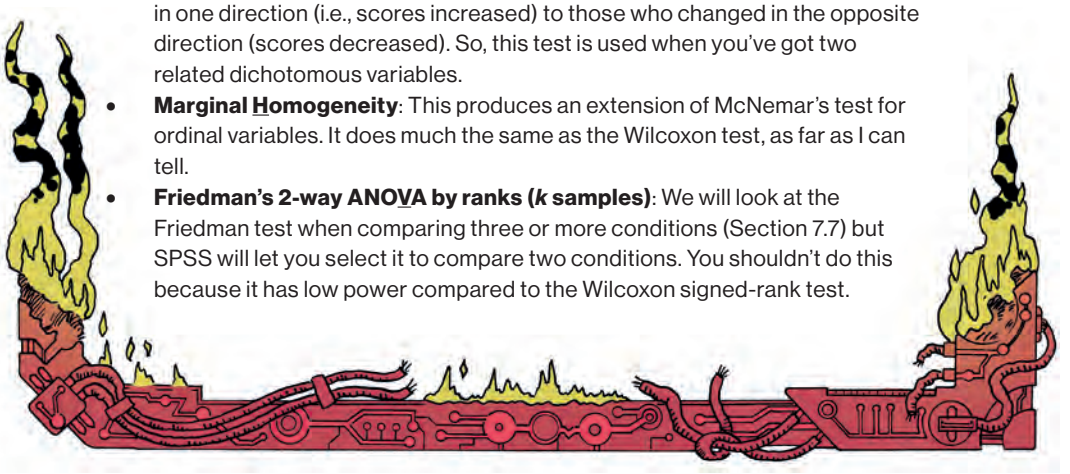
SPSS Tip 7.2
Other options for comparing dependent groups Ⓑ

In the *Settings* tab of the dialog box in Figure 7.9 there are other tests available:

- **Sign**: The sign test does the same thing as the Wilcoxon signed-rank test, except that it is based only on the direction of difference (positive or negative). The magnitude of change is completely ignored (unlike in the Wilcoxon test where the rank tells us something about the relative magnitude of change). For these reasons the sign test lacks power (it's not very good at detecting effects) unless sample sizes are very small (six or less). I wouldn't use it.
- **McNemar**: This test is useful when you have nominal rather than ordinal data. It's typically used when you're looking for changes in people's

scores, and it compares the number of people who changed their response in one direction (i.e., scores increased) to those who changed in the opposite direction (scores decreased). So, this test is used when you've got two related dichotomous variables.

- **Marginal Homogeneity**: This produces an extension of McNemar's test for ordinal variables. It does much the same as the Wilcoxon test, as far as I can tell.
- **Friedman's 2-way ANOVA by ranks (*k* samples)**: We will look at the Friedman test when comparing three or more conditions (Section 7.7) but SPSS will let you select it to compare two conditions. You shouldn't do this because it has low power compared to the Wilcoxon signed-rank test.



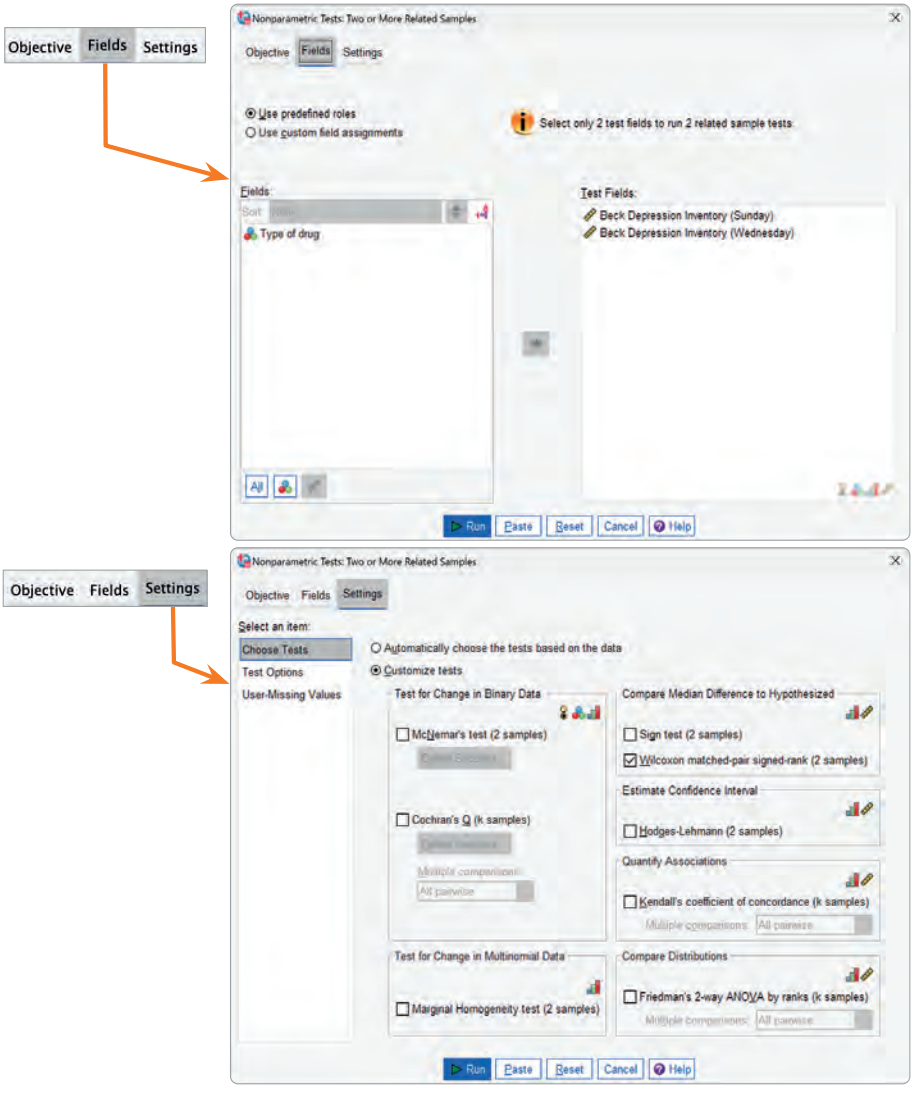
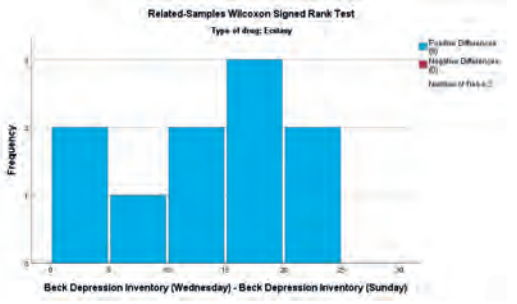


Figure 7.9 Dialog boxes for the Wilcoxon signed-rank test

7.5.4 Output for the ecstasy group Ⓐ

If you have split the file, then the first set of results obtained will be for the ecstasy group (Output 7.3). The histogram of the distribution of differences shows the Sunday scores subtracted from the Wednesday scores (which we're told underneath the histogram) corresponding to the values in the *Difference* column in Table 7.2. A positive difference means more depression on Wednesday than Sunday, a negative difference means more depression on Sunday than Wednesday, and a difference of zero means that depression levels were identical on Sunday and Wednesday. The histogram is colour-coded based on whether ranks are positive or negative: positive ranks appear as blue bars, and negative ranks as a colour I find quite hard to describe – claret red? You might notice that there are no claret red bars, which tells us that there were no negative ranks. Therefore, the histogram is a very quick indication of the ratio of positive to negative ranks: in this case all ranks are positive (or tied) and none are negative. We are told the same in the legend to the histogram: there were 8 positive differences, 0 negative differences and 2 ties.



Related-Samples Wilcoxon Signed Rank Test Summary

Total N	10
Test Statistic	36.000
Standard Error	7.124
Standardized Test Statistic	2.527
Asymptotic Sig.(2-sided test)	.012

Output 7.3

In Section 7.5.1 I explained that the test statistic, T, is the sum of positive ranks, so our test value here is 36. I also showed how this value can be converted to a z-score, and in doing so we can compute exact significance values based on the normal distribution. The table in Output 7.3 tells us the test statistic (36), its standard error (7.12) and the z-score (2.53), which all correspond (more or less) to the values we computed by hand in Section 7.5.1. This z-score has a significance value of $p = 0.012$. This value is less than the standard critical value of 0.05, so we conclude that there is a significant change in depression scores from Sunday to Wednesday (i.e., we reject the null hypothesis). From the histogram we know that this test statistic is based on there being many more positive differences (i.e., scores being higher on Wednesday than Sunday), so we can conclude that when taking ecstasy there was a significant increase in depression (as measured by the BDI) from the morning after to midweek.

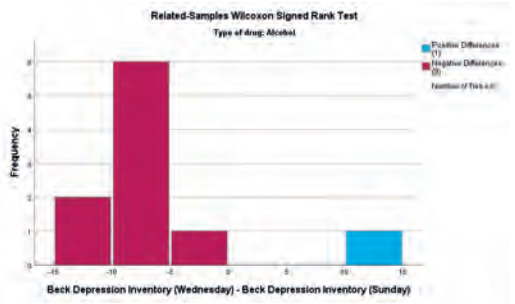
7.5.5 Output for the alcohol group Ⓐ

For the alcohol group the histogram (Output 7.4) *does* contain different-coloured bars: the blue bars represent positive differences and the claret red bars negative differences. For the ecstasy group we saw only blue bars, but for the alcohol group we see the opposite: the bars are predominantly claret red. This pattern indicates that differences between Wednesday and Sunday were generally negative. In other words, scores were generally higher on Sunday than they were on Wednesday. Again, these differences are the same as those in the *Difference* column in Table 7.2. The legend of the plot confirms that there was only 1 positive difference, 9 negative differences and 0 ties.

The table in Output 7.4 tells us the test statistic (8), its standard error (9.80), and the corresponding z-score (–1.99). (These are the values we calculated in Section 7.5.1.) The p-value associated with the z-score is 0.047, which means that there's a probability of 0.047 that we would get a value of z at least as large as the one we have if there were no effect in the population; because this value is less than the critical value of 0.05, people would typically conclude that there is a significant difference in

depression scores. We know that because the histogram showed predominantly negative differences (i.e., scores higher on Sunday than on Wednesday) there was a significant *decline* in depression (as measured by the BDI) from the morning after to midweek in the alcohol group.

The results of the ecstasy and alcohol groups show an opposite effect of alcohol and ecstasy on depression. After taking alcohol depression is higher the morning after than midweek, whereas after taking ecstasy, depression increases from the morning after to the middle of the week. A different effect across different groups or conditions is known as *moderation* (i.e., the effect you get under certain circumstances and is different than the effect under different circumstances). You can't look at moderation effects directly using non-parametric tests, but we will look at these effects in detail in due course (see Chapters 11 and 14).



Related-Samples Wilcoxon Signed Rank Test Summary

Total N	10
Test Statistic	8.000
Standard Error	9.798
Standardized Test Statistic	-1.990
Asymptotic Sig.(2-sided test)	.047

Output 7.4

7.5.6 Calculating an effect size Ⓑ

We can again compute a rank-biserial correlation for these effects, but using a different formula to reflect the fact that data come from the same entities. The formula comes from King et al. (2011), but I present an equivalent, but simpler, formula from Kerby (2014) but using symbols consistent with this chapter

$$r_{rb} = \frac{T_+ - T_-}{T_+ + T_-} \tag{7.23}$$

Here T_+ is the sum of positive ranks and T_- is the sum of negative ranks. This equation essentially takes the difference between the positive and negative ranks (i.e. the effect) and expresses it as a proportion of the total of all ranks. It is literally the effect expressed as a proportion. The direction of the correlation reflects the direction of the effect: a positive correlation expresses the proportion of data in the direction of T_+, and a negative correlation expresses the proportion of data in the direction of T_-.

For the ecstasy group we have $T_+ = 36$ and $T_- = 0$. Substituting these values into Eq. 7.23 gives us

$$\begin{aligned} r_{rb} &= \frac{36 - 0}{36 + 0} \\ &= \frac{36}{36} \\ &= 1 \end{aligned} \tag{7.24}$$

There is a perfect correlation. This value of 1 is a positive value and so is the proportion of ranks in the direction of T_+, that is, in the direction of depression being higher on Wednesday than Sunday. The magnitude of 1 is a proportion but can be converted to a percentage by multiplying by 100. Therefore, we can say that 100% of the ranked data are in the direction that supports the idea that depression was higher on Wednesday than Sunday. Conversely, 0% of the ranked data went against this idea. This result makes intuitive sense because there were no negative ranks (i.e. ranks reflecting higher depression levels on a Sunday). This example illustrates the beauty of the rank-biserial correlation because it very directly maps onto the proportion of ranks that are consistent with a hypothesis.

For the alcohol group $T_+ = 8$ and $T_- = 47$. Substituting these values into Eq. 7.23 gives us

$$\begin{aligned} r_{rb} &= \frac{8-47}{8+47} \\ &= -\frac{39}{55} \\ &= -0.71 \end{aligned} \tag{7.25}$$

This is a very strong correlation. However, because the correlation is a minus value it tells us the proportion of ranks in the direction of T_-, that is, in the direction of depression being higher on Sunday than Wednesday. This value suggests that 71% of the ranked data support the idea that depression was higher on Sunday than Wednesday and 29% of the ranked data went against this idea. We could flip this interpretation (for comparison with the ecstasy result) and say that for alcohol 29% of the ranked data are in the direction that supports the idea that depression was higher on Wednesday than Sunday.

7.5.7 Writing the results Ⓐ

For the Wilcoxon test, we report the test statistic (denoted by the letter T), its exact significance and an effect size (see Section 3.8). So, we could report something like:

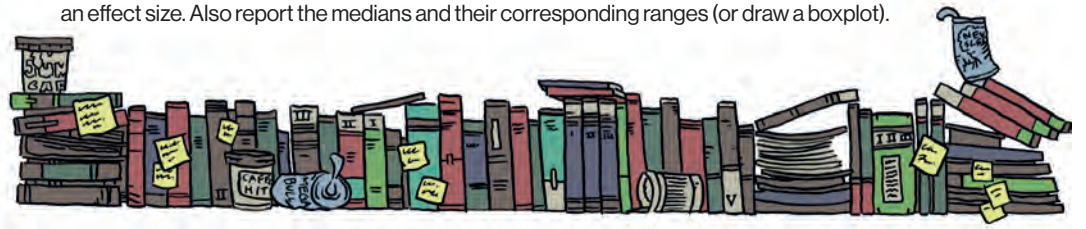
✓ For ecstasy users, depression levels were significantly higher on Wednesday (*Mdn* = 33.50) than on Sunday (*Mdn* = 17.50), $T = 36$, $p = 0.012$, $r_{rb} = 1$. However, for alcohol users the opposite was true: depression levels were significantly lower on Wednesday (*Mdn* = 7.50) than on Sunday (*Mdn* = 16.0), $T = 8$, $p = 0.047$, $r_{rb} = -0.71$.

You can get the median values by running descriptive statistics.



Cramming Sam's Tips

- The Wilcoxon signed-rank test compares two conditions when the scores are related (e.g., scores come from the same participants) and the resulting data have unusual cases or violate any assumption in Chapter 6.
- Look at the row labelled *Asymptotic Sig. (2-sided test)*. If the value is less than or equal to 0.05 then the two conditions are significantly different.
- Look at the histogram and numbers of positive or negative differences to tell you how the groups differ (the greater number of differences in a particular direction tells you the direction of the result).
- Report the *T*-statistic, the corresponding *z*, the exact significance value and the rank-biserial correlation as an effect size. Also report the medians and their corresponding ranges (or draw a boxplot).



Alternatively, we could report the values of z:

✓ For ecstasy users, depression levels were significantly higher on Wednesday (*Mdn* = 33.50) than on Sunday (*Mdn* = 17.50), $z = 2.53$, $p = 0.012$, $r_{rb} = 1$. However, for alcohol users the opposite was true: depression levels were significantly lower on Wednesday (*Mdn* = 7.50) than on Sunday (*Mdn* = 16.0), $z = -1.99$, $p = 0.047$, $r_{rb} = -0.71$.



Labcoat Leni's Real Research 7.1
Having a quail of a time? Ⓐ

Matthews, R. C. et al. (2007). *Psychological Science*, *18*(9), 758–762.

We encountered some research in Chapter 2 in which we discovered that you can influence aspects of male quail's sperm production through 'conditioning'. The basic idea is that the male is granted access to a female for copulation in a certain chamber (e.g., one that is coloured green) but gains no access to a female in a different context (e.g., a chamber with a tilted floor). The male, therefore, learns that when he is in the green chamber his luck is in, but if the floor is tilted then frustration awaits. For other males the chambers will be reversed. The human equivalent (well, sort of) would be if you always managed to pull at Coalition but never at Chalk.[4] During the test phase, males get to mate in both chambers. The question is: after the males have learnt that they will get a mating opportunity in a certain context, do they produce more sperm or better-quality sperm when mating in that context compared to the control context? (That is, are you more of a stud at Coalition? OK, I'm going to stop this analogy now.)

Mike Domjan and his colleagues predicted that if conditioning evolved because it increases reproductive fitness then males who mated in the context that had previously signalled a mating opportunity would fertilize a significantly greater number of eggs than quails that mated in their control context (Matthews et al., 2007). They put this hypothesis to the test in an experiment that is utter genius. After training, they allowed 14 females to copulate with two males (counterbalanced): one male copulated with the female in the chamber that had previously signalled a reproductive opportunity **(signalled)**, whereas the second male copulated with the same female but in the chamber that had not previously signalled a mating opportunity **(control)**. Eggs were collected from the females for 10 days after the mating and a genetic analysis was used to determine the father of any fertilized eggs.

The data from this study are in the file **matthews_2007.sav.** Labcoat Leni wants you to carry out a Wilcoxon signed-rank test to see whether more eggs were fertilized by males mating in their signalled context compared to males in their control context.

Answers are on the companion website (or look at page 760 in the original article).

4 These are both clubs in Brighton that I don't go to because I'm too old for that sort of thing, as opposed to in my youth when I too socially inept for that sort of thing.

7.6 Differences between several independent groups: the Kruskal–Wallis test Ⓐ

Having looked at models that compare two groups or conditions, we now move onto models that can compare more than two conditions: the Kruskal–Wallis test compares groups or conditions containing independent scores, whereas the Friedman test is used use when scores are related. Let's look at the **Kruskal–Wallis test** first (Kruskal & Wallis, 1952), which assesses the hypothesis that multiple independent groups come from different populations. If you'd like to know a bit more about William Kruskal (Figure 7.10) there is a lovely biography by Fienberget al. (2007).



Figure 7.10 William Kruskal

I read a news story claiming that the chemical genistein, which is naturally occurring in soya, was linked to lowered sperm counts in Western males. When you read the actual study, it had been conducted on rats, it found no link to lowered sperm counts, but there was evidence of abnormal sexual development in male rats (probably because genistein acts like oestrogen). A study showing no link between soya and sperm counts was used as the scientific basis for an article about soya being the cause of declining sperm counts in Western males (never trust what you read). Imagine the rat study was enough for us to want to test this idea in humans. We recruit 80 males and split them into four groups that vary in the number of soya 'meals' they ate per week over a year-long period (a 'meal' would be a dinner containing 75g of soya). The first group was a control and ate no soya meals (i.e., none in the whole year); the second group had one soya meal per week (that's 52 over the year); the third group had four soya meals per week (208 over the year); and the final group had seven soya meals a week (364 over the year). At the end of the year, the participants were sent away to produce some sperm that I could count (when I say 'I', I mean someone else in a laboratory as far away from me as humanly possible).[5]

7.6.1 Theory of the Kruskal–Wallis test Ⓑ

Like the other tests we've seen, the Kruskal–Wallis test is used with ranked data. To begin with, scores are ordered from lowest to highest, ignoring the group to which the score belongs. The lowest score is assigned a rank of 1, the next highest a rank of 2 and so on (see Section 7.4.1 for more detail). Once ranked, the scores are collected back into their groups and their ranks are added within each group. The sum of ranks within each group is denoted by R_i (where i denotes the group). Table 7.3 shows the raw data for this example along with the ranks.



5 In case any medics are reading this chapter, these data are made up and, because I have absolutely no idea what a typical sperm count is, they're probably ridiculous. Feel free to laugh at my ignorance.

Table 7.3 Data for the soya example with ranks

No Soya		1 Soya Meal		4 Soya Meals		7 Soya Meals	
Sperm (million/ml)	Rank	Sperm (million/ml)	Rank	Sperm (million/ml)	Rank	Sperm (million/ml)	Rank
3.51	4	3.26	3	4.03	6	3.10	1
5.76	9	3.64	5	5.98	10	3.20	2
8.84	17	6.29	11	9.59	19	5.60	7
9.23	18	6.36	12	12.03	21	5.70	8
12.17	22	7.66	14	13.13	24	7.09	13
15.10	30	15.33	32	13.54	27	8.09	15
15.17	31	16.22	34	16.81	35	8.71	16
15.74	33	17.06	36	18.28	37	11.80	20
24.29	41	19.40	38	20.98	40	12.50	23
27.90	46	24.80	42	29.27	48	13.25	25
34.01	55	27.10	44	29.59	49	13.40	26
45.15	59	41.16	57	29.95	50	14.90	28
47.20	60	56.51	61	30.87	52	15.02	29
69.05	65	67.60	64	33.64	54	20.90	39
75.78	68	70.79	66	43.37	58	27.00	43
77.77	69	72.64	67	58.07	62	27.48	45
96.19	72	79.15	70	59.38	63	28.30	47
100.48	73	80.44	71	101.58	74	30.67	51
103.23	75	120.95	77	109.83	76	32.78	53
210.80	80	184.70	79	182.10	78	41.10	56
Total (R_i)	927		883		883		547
Average ($\bar{R}_i$)	46.35		44.15		44.15		27.35

Once the sum of ranks has been calculated within each group, the test statistic, H, is calculated as follows:

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1) \tag{7.26}$$

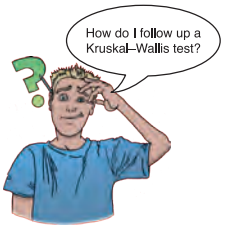
In this equation R_i is the sum of ranks for each group, N is the total sample size (in this case 80) and n_i is the sample size of a particular group (in this example group sample sizes are all 20). The middle part of the equation means that for each group we sum the squares of ranks, divide this value by the sample size for that group, then add up these values. The rest of the equation involves calculating values based on the total sample size. For these data we get:

$$\begin{aligned} H &= \frac{12}{80(81)}\left(\frac{927^2}{20}+\frac{883^2}{20}+\frac{883^2}{20}+\frac{547^2}{20}\right)-3(81) \\ &= \frac{12}{6480}(42966.45+38984.45+38984.45+14960.45)-243 \\ &= 0.0019(135895.8)-243 \\ &= 251.66-243 \\ &= 8.659 \end{aligned} \tag{7.27}$$

This test statistic has a distribution from the family of chi-square distributions. Whereas the standard normal distribution is defined by two parameters (a mean of 0 and a standard deviation of 1), the chi-square distribution is defined by a single parameter, the degrees of freedom, which is one less than the number of groups (i.e., k–1), in this case 3.

7.6.2 Follow-up analysis Ⓑ

The Kruskal–Wallis test tells us that, overall, groups come from different populations. However, it doesn't tell us which groups differ. Are all groups different, or only some of them? The simplest way to break down the overall effect is to compare all pairs of groups (known as **pairwise comparisons**). In our current example, this would entail six tests: none vs. 1 meal; none vs. 4 meals; none vs. 7 meals; 1 vs. 4 meals; 1 vs. 7 meals; and 4 vs. 7 meals. A very simple approach would be to perform six Mann–Whitney tests, one for each of these comparisons. However, we saw in Section 2.10.8 that when we do lots of tests on the same data we inflate the familywise error rate: we'll make at least one Type I error more than 5% of the time. Ideally, we want a 5% chance of making a Type I error over *all the tests* we do, and we've seen that one method to achieve this is to use a lower probability as the threshold for significance. Therefore, we *could* perform six tests (one to compare each pair of groups) if we adjust the p-value so that overall, across all the tests, the Type I error rate remains at 5%. This is what a pairwise comparison does.



By being stricter about the p-value you deem to be significant you reduce the power of the tests – you might throw the baby out with the bathwater (Section 2.10.9). An alternative is to use a stepped procedure. The one SPSS uses begins by ordering the groups based on the sum of ranks from smallest to largest (if there are ties, the median decides the order rather than the sum).[6] For our data the rank sums were 7 meals (rank sum = 547, median = 13.33), 4 meals (rank sum = 883, median = 29.43), 1 meal (rank sum = 883, median = 25.95), no meals (rank sum = 927, median = 30.96), resulting in the order 7 meals, 1 meal, 4 meals and no meals. Figure 7.11 shows how the step-down process works. Step 1 is to see whether the first ordered group is the same as the second (i.e., is there a non-significant difference?). If they are equivalent, you put in the third ordered group and see if all three are equivalent. If they are, you put in the fourth group and see if all four are equivalent. If at any point you find a significant difference (i.e., the groups are not equivalent) then you stop, carry the group that you included last into the next step and consider the groups you don't carry forward as a subset (i.e., they are equivalent). In step 2 you repeat the same process.

In Example 1 in Figure 7.11, we start with the first two groups in our ordered list (7 meals and 1 meal). They are equivalent (not significantly different) so we introduce the third ordered group (4 meals). Doing so makes the groups significantly different (they are not equivalent), so we carry 4 meals into the second step, and conclude that 7 meals and 1 meal are equivalent (they are said to be *homogeneous groups*). In step 2, we compare 4 meals to the one remaining group (0 meals). These groups are equivalent (not significantly different), so we put them together in a different subgroup of 'homogeneous groups' and stop the process. In Example 2, we start with the first two groups in our ordered list (7 meals and 1 meal). They are significantly different (i.e. not equivalent) so we carry 1 meal into the second step, and conclude that 7 meals is a group on its own. In step 2, we compare 1 meal to 4 meals. They are equivalent (not significantly different), so we introduce 0 meals, but doing so makes the groups significantly different (not equivalent), so we conclude that 4 meals and 1 meal are a homogeneous groups, and (because there are no other groups to compare it with) remove 0 meals into a subgroup on its own. These follow-up procedures are quite complicated, so don't worry if you don't fully understand them – we will look later at how to interpret them.

6 Each group has 20 scores, so the median will be the average of the 10th and 11th scores when the scores are in ascending order. The data in Table 7.3 are presented in ascending order in each group, so we can see that the medians are: (27.90 + 34.01)/2 = 30.96 (0 meals); (24.80 + 27.10)/2 = 25.95 (1 meal); (29.27 + 29.59)/2 = 29.43 (4 meals); (13.25 + 13.40)/2 = 13.33 (7 meals).

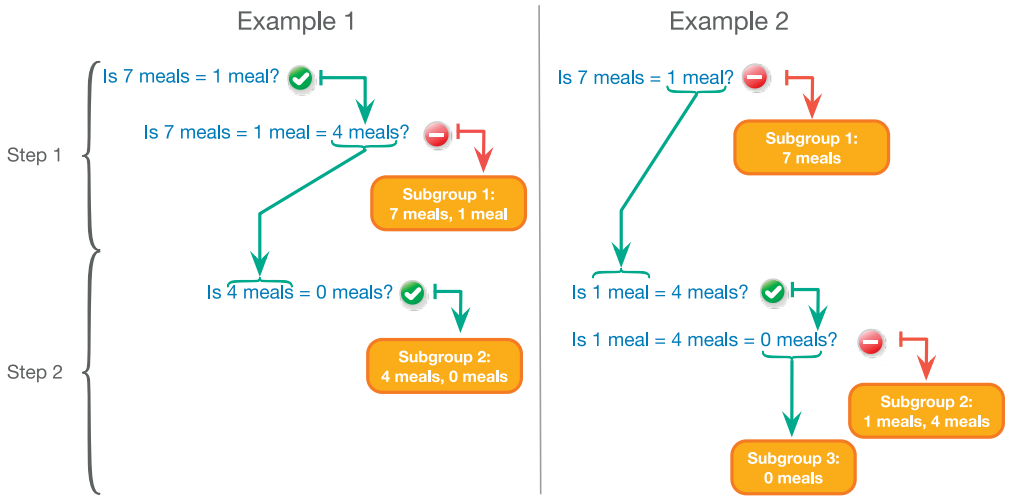


Figure 7.11 The non-parametric step-down procedure

7.6.3 Inputting data and provisional analysis Ⓐ



See whether you can enter the data in Table 7.3 into SPSS (you don't need to enter the ranks). Then conduct some exploratory analyses on the data (see Sections 6.9.4 and 6.9.6).

As a minimum, the data editor will have two columns. One column should be a coding variable (called something like **soya**) containing four numeric codes that define the groups (for convenience I suggest 1 = no soya, 2 = one soya meal per week, 3 = four soya meals per week and 4 = seven soya meals per week). Remember to define values so that you know which group is represented by which code (see Section 4.6.5). Another column should contain the sperm counts measured at the end of the year (call this variable **sperm**). The data can be found in the file **soya.sav** (which also contains a participant **id** variable).

The results of our exploratory analysis are shown in Figure 7.12 and Output 7.4. The normal Q-Q plots show quite clear deviations from normality for all four groups because the dots deviate from the diagonal line. We don't really need to do anything more than look at these plots – the evidence of non-normality is plain to see, and formal tests can be problematic (see Jane Superbrain Box 6.7). However, given that within each group the sample is quite small (n = 20), if tests of normality are significant then this can be informative (because if the test has detected a deviation in such a small sample, then it's probably a fairly substantial deviation). If you do these tests (Output 7.4) you'll find that the more accurate Shapiro–Wilk test is significant ($p \leq 0.05$) for all groups except the 7 meals (but even for that group it is close to being significant). Notwithstanding the pointlessness of Levene's test, its output would favour the conclusion that homogeneity of variance cannot be assumed, $F(3, 76) = 5.12$, $p = 0.003$, because the p-value is less than 0.05. As such, the information converges on a sad story: that data are probably not normally distributed, and the groups probably have heterogeneous variances.

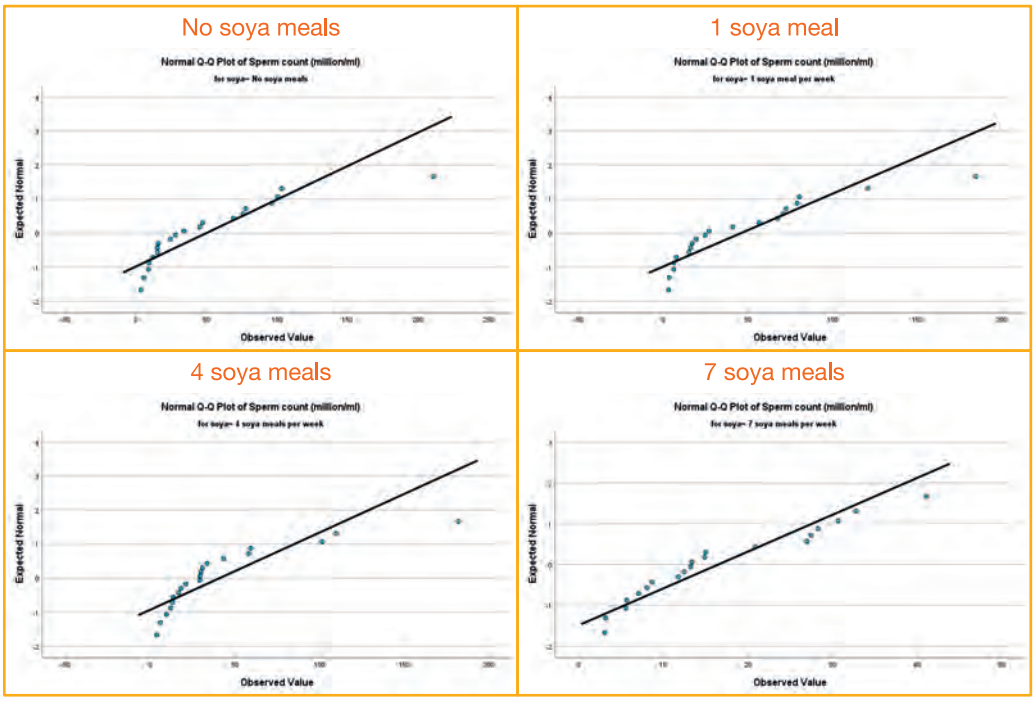


Figure 7.12 Normal Q-Q plots of sperm counts after different doses of soya meals per week

Tests of Normality

	Number of soya meals per week	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Sperm count (million/ml)	No soya meals	.181	20	.085	.805	20	.001
	1 soya meal per week	.207	20	.024	.826	20	.002
	4 soya meals per week	.267	20	<.001	.743	20	<.001
	7 soya meals per week	.204	20	.028	.912	20	.071

a. Lilliefors Significance Correction

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Sperm count (million/ml)	Based on Mean	5.117	3	76	.003
	Based on Median	2.860	3	76	.042
	Based on Median and with adjusted df	2.860	3	58.107	.045
	Based on trimmed mean	4.070	3	76	.010

Output 7.5

7.6.4 Doing the Kruskal–Wallis test using SPSS Ⓐ

To run a Kruskal–Wallis test, follow the general procedure outlined in Section 7.3, by first selecting *Analyze* > *Nonparametric Tests* > Independent Samples... When you reach the Objective Fields Settings tab, if you need to assign variables then select ◉ Use custom field assignments and specify the model by dragging **Sperm Count (Millions)** into the box labelled *Test Fields* (or select it in the box labelled *Fields* and click ↳). Next, drag **Soya** to the box labelled *Groups* (Figure 7.13, top). Activate the Objective Fields Settings tab, select ◉ Customize tests (SPSS Tip 7.3) and check ☑ Kruskal-Wallis 1-way ANOVA (k samples) (Figure 7.13, bottom). Next to this option there is a drop-down list labelled *Multiple comparisons*. Within this list are two options, which we discussed earlier: to compare every group against every other group (*All pairwise*)

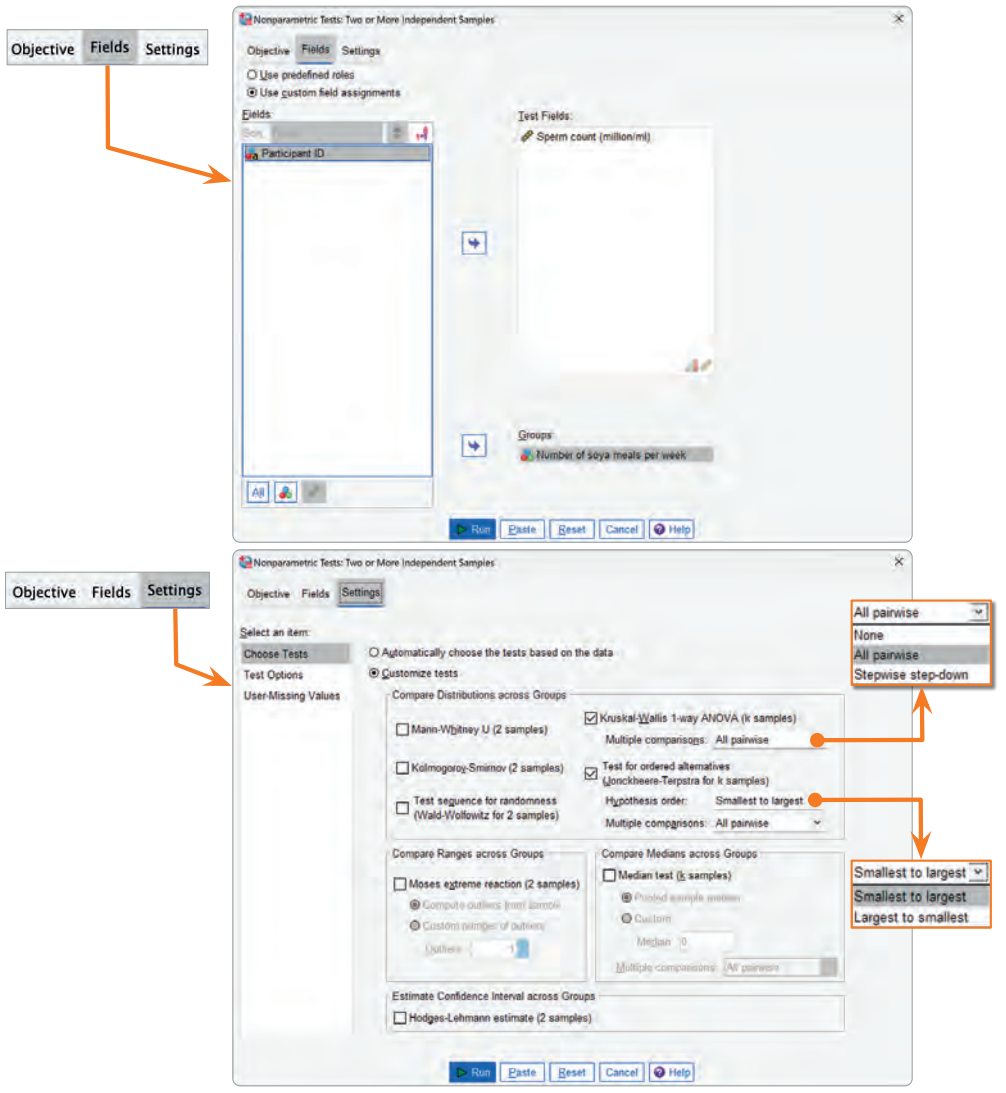
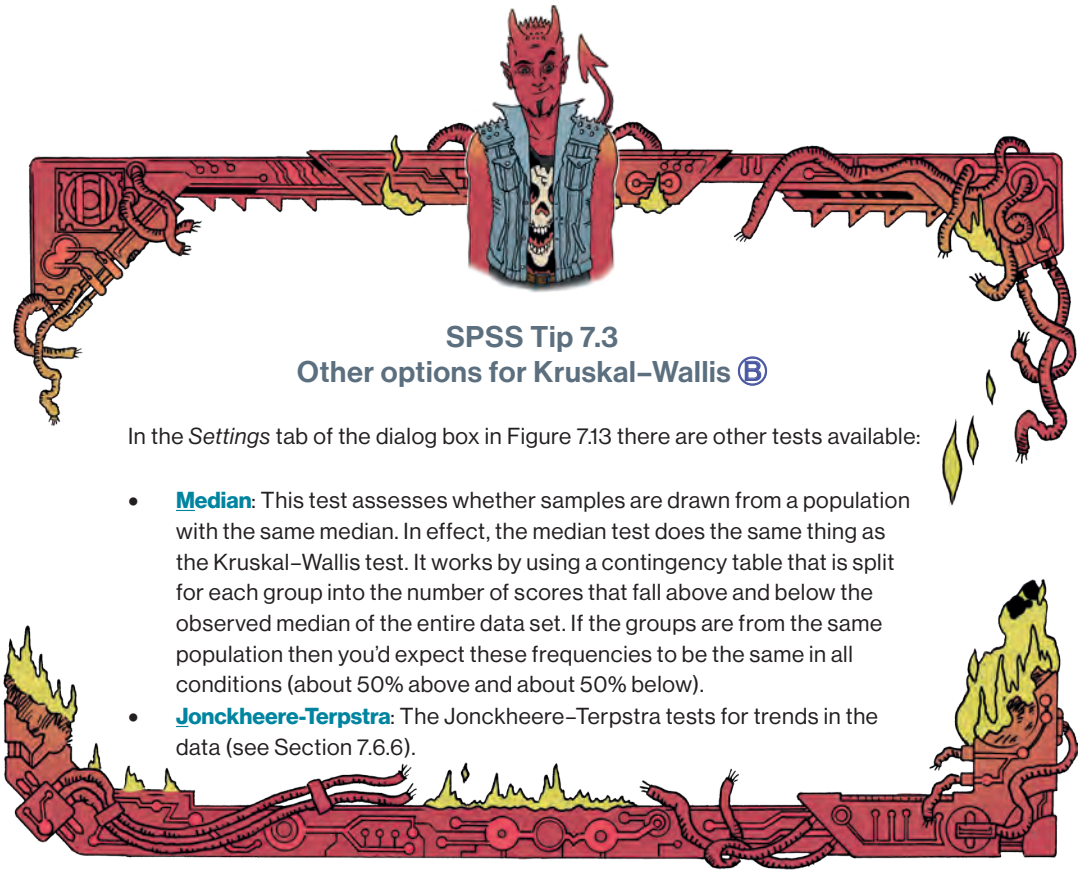


Figure 7.13 Dialog boxes for the Kruskal–Wallis test

or to use a step-down method (*Stepwise step-down*). You can also ask for the Jonckheere–Terpstra trend test, which is useful to see whether the medians of the groups increase or decrease in a linear way. For the time being don't select this option, but we will look at this test in due course. To run the analysis click Run.

7.6.5 Output from the Kruskal–Wallis test Ⓐ

Output 7.6 shows a boxplot of the data, and a table containing the Kruskal–Wallis test statistic, H (8.659, the same value that we calculated earlier), its associated degrees of freedom (we had 4 groups, so we have 3 degrees of freedom) and the significance. The significance value is 0.034; because this value is less than 0.05, people would typically conclude that the amount of soya meals eaten per week significantly affects sperm counts. As we discussed earlier, the overall effect tells us that sperm counts were different in some of the groups, but we don't know specifically which groups differed. The boxplots (Output 7.6) can help here. The first thing to note is that there are a few men who

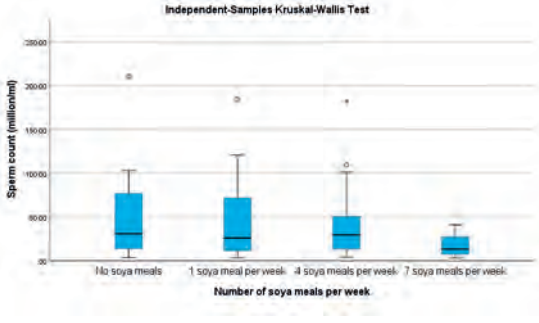


SPSS Tip 7.3
Other options for Kruskal–Wallis Ⓑ

In the *Settings* tab of the dialog box in Figure 7.13 there are other tests available:

- **Median**: This test assesses whether samples are drawn from a population with the same median. In effect, the median test does the same thing as the Kruskal–Wallis test. It works by using a contingency table that is split for each group into the number of scores that fall above and below the observed median of the entire data set. If the groups are from the same population then you'd expect these frequencies to be the same in all conditions (about 50% above and about 50% below).
- **Jonckheere-Terpstra**: The Jonckheere–Terpstra tests for trends in the data (see Section 7.6.6).

produced a particularly rampant quantity of sperm (the circles and asterisks that lie above the top whiskers). These are outliers. The scores are generally more compressed after 7 soya meals per week than in the other groups. For the 7 meal group the vertical height of the box is smaller and the whiskers are shorter than the other groups, suggesting that sperm counts are less variable in this group than the others. The median of the 7 soya meals per week group also seems a little lower than the other groups (and the medians of the other groups seem quite similar); perhaps the difference lies between the 7-meal group and the others? We can find out using follow-up analyses like those we discussed in Section 7.6.2.



Independent-Samples Kruskal-Wallis Test Summary

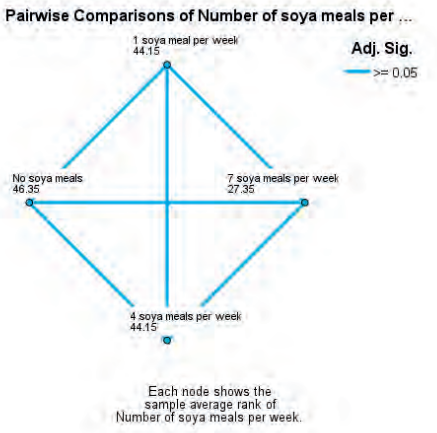
Total N	80
Test Statistic	8.659[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	.034

a. The test statistic is adjusted for ties.

Output 7.6

The output you see will depend on whether you selected *All pairwise* or *Stepwise step-down* in the drop-down list labelled *Multiple comparisons* when you ran the analysis (Figure 7.13). Let's look at the pairwise comparisons first (Output 7.7). The diagram at the top shows the average rank within each group: so, for example, the average rank in the 7 meal group was 27.35, and for the no soya meals group was 46.35. This diagram highlights differences between groups using different-coloured lines to connect them (in the current example, there are no significant differences between groups, so all the connecting lines are blue). The table underneath shows the comparisons between all possible pairs of groups. In each case the test statistic is the difference between the mean ranks of those groups. For 7 vs. 1 soya meal, this will be 44.15 – 27.35 = 16.80, for 0 vs. 4 soya meals it will be 46.35 – 44.15 = 2.20, and so on. These test statistics are converted into *z*-scores (the column labelled *Std. Test Statistic*) by dividing by their standard errors, from which a *p*-value can be ascertained. For example, the 7 vs. 1 soya meal comparison has a *z*-score of 2.286 and the exact *p*-value for this *z* is 0.022. As I mentioned in Section 7.6.2, to control the Type I error rate the *p*-value is adjusted for the number of tests done. The column labelled *Adj. Sig.* contains these adjusted *p*-values and we interpret the values in this column (no matter how tempted we are to interpret the column labelled *Sig.*). Looking at the *Adj. Sig.* column, none of the values fall below the criterion of 0.05 (although the comparison between 7 soya meals and no soya meals comes close with a $p = 0.058$, which reminds us that significance testing encourages black-and-white thinking and effect sizes might be useful).

To sum up, despite the significant overall effect, none of the specific comparisons between groups indicates a significant difference in sperm counts due to differing amounts of soya consumption. The effect we got seems to mainly reflect the fact that eating soya seven times per week lowers (I know this from the mean ranks) sperm counts compared to eating no soya, although even this comparison was just non-significant.



Pairwise Comparisons of Number of soya meals per week

Sample 1–Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
7 soya meals per week–1 soya meal per week	16.800	7.348	2.286	.022	.133
7 soya meals per week–4 soya meals per week	16.800	7.348	2.286	.022	.133
7 soya meals per week–No soya meals	19.000	7.348	2.586	.010	.058
1 soya meal per week–No soya meals	2.200	7.348	.299	.765	1.000
4 soya meals per week–No soya meals	2.200	7.348	.299	.765	1.000
1 soya meal per week–4 soya meals per week	.000	7.348	.000	1.000	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .050.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Output 7.7

If you choose the *Stepwise step-down* procedure to follow up the Kruskal–Wallis test then you'll see the output in Output 7.8. The step-down method doesn't compare every group with every other group, which means that the *p*-values are not adjusted so strictly (because we're not doing as many significance tests on the same data). Output 7.8 clusters equivalent (homogeneous) groups in the same columns. From column 1, we can see that the group that ate 7 soya meals a week clusters on its own. In other words, comparing it with the next highest-ranking group (the 1 soya meal group) produced a significant difference. Consequently, the 1 soya meal group was moved into a different subset in column 2 and was then compared to the next highest-ranking group (4 soya meals), which did not lead to a significant difference, so they were both compared to the no soya meals group, which also produced no significant difference (think back to Figure 7.11). The fact that these three groups (1, 4 and no soya meals) are clustered within the same column tells us that they are equivalent (i.e., homogeneous). The *Adjusted Sig.* tells us that the *p*-value associated with comparing the 1, 4 and no soya meals groups was 0.943, which means not at all significant. In short, having 7 soya meals per week seemed to lower sperm counts significantly compared to all other groups, but all other doses of soya had no significant effect on sperm counts.

Homogeneous Subsets based on Number of soya meals per week

		Subset 1	Subset 2
Sample[a]	7 soya meals per week	27.350	
	1 soya meal per week		44.150
	4 soya meals per week		44.150
	No soya meals		46.350
Test Statistic		.[b]	.118
Sig. (2-sided test)		.	.943
Adjusted Sig. (2-sided test)		.	.943

Homogeneous subsets are based on asymptotic significances. The significance level is .050.

a. Each cell shows the sample average rank of Number of soya meals per week.

b. Unable to compute because the subset contains only one sample.

Output 7.8

7.6.6 Testing for trends: The Jonckheere–Terpstra test Ⓑ

Back in Section 7.6.4 I mentioned the Jonckheere–Terpstra test, ☑ Test for ordered alternatives (Jonckheere-Terpstra for k samples) (Jonckheere, 1954; Terpstra, 1952), which tests for an ordered pattern to the medians of the groups you're comparing. It does the same thing as the Kruskal–Wallis test (i.e., test for a difference between the medians of the groups) but it incorporates information about whether the order of the groups is meaningful. As such, you should use this test when you expect the groups you're comparing to produce a meaningful order of medians. In the current example we expect that the more soya a person eats, the lower their sperm count. The control group should have the highest sperm count, those having one soya meal per week should have a lower sperm count, the sperm count in the four meals per week group should be lower still, and the seven meals per week group should have the lowest sperm count. Therefore, there is an order to our medians: they should decrease across the groups.

Conversely there might be situations where you expect your medians to increase. For example, there's a phenomenon in psychology known as the 'mere exposure effect': the more you're exposed to something, the more you'll like it. Back in the days when people paid money for music, record companies put this effect to use by making sure songs were played on radio for about 2 months prior to their release, so that by the day of release, everyone loved the song and rushed out to buy it.[7] Anyway, if you took three groups of people, exposed them to a song 10 times, 20 times and 30 times respectively and measured how much they liked the song, you'd expect the medians to increase. Those who heard it 10 times would like it less than those who heard it 20 times who in turn would like it less than those who heard it 30 times. The Jonckheere–Terpstra test was designed for these situations, and there are two options (Figure 7.13):

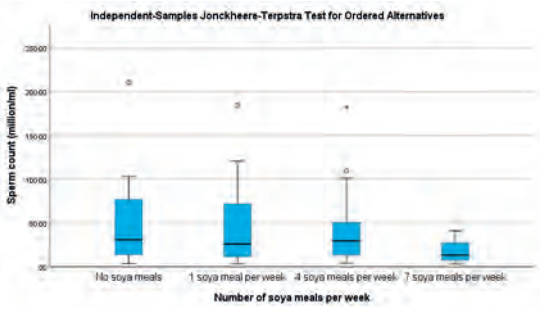
- *Smallest to largest*: This option tests whether the first group differs from the second group, which in turn differs from the third group, which in turn differs from the fourth and so on until the last group.
- *Largest to smallest*: This option tests whether the last group differs from the group before, which in turn differs from the group before and so on until the first group.

In both cases the test looks at differences across ordered groups, it does not distinguish whether medians increase or decrease. The test determines whether the medians of the groups *ascend or descend* in the order specified by the coding variable; therefore, the coding variable must code groups *in the order that you expect the medians to change* (to reiterate, it doesn't matter whether you expect them to increase or decrease). For our soya example, we coded our groups as 1 = no soya, 2 = one soya meal per week, 3 = four soya meals per week and 4 = seven soya meals per week, so the Jonckheere–Terpstra test would assess whether the median sperm count increases or decreases across the groups when they're ordered in that way. If we wanted to test a different order then we'd have to specify a different coding variable that reflected this different order. Figure 7.13 shows how to specify the test, so rerun the analysis (as in Section 7.6.4) but select ☑ Test for ordered alternatives (Jonckheere-Terpstra for k samples) (*Smallest to largest*) instead of the Kruskal–Wallis test.

Output 7.9 shows the output from the Jonckheere–Terpstra test for the soya data. The output tells us the value of test statistic, *J*, which is 912. In large samples (more than about eight per group) this test statistic can be converted to a *z*-score, which for these data is –2.476. As with any *z*-score, we can ascertain a *p*-value, which in this case is 0.013, which indicates a significant trend in the medians because it is lower than the typical critical value of 0.05. The sign of the *z*-value tells us the direction of the trend: a positive value indicates ascending medians (i.e., the medians get bigger as the values of the coding variable get bigger), whereas a negative value (as we have here) indicates descending medians (the medians get smaller as the value of the coding variable gets bigger). In this example, because we set the test option to be *Smallest to largest* (Figure 7.13) and we coded the variables as 1 = no soya, 2 = one soya meal per week, 3 = four soya meals per week and 4 = seven soya meals per week, the negative value of *z* means that the medians get smaller as we go from no soya, to one soya meal, to four soya meals and on to seven soya meals.[8]

7 For most popular music it had the opposite effect on me.

8 If you're bored, rerun the test but specify *Largest to smallest*. The results will be identical, except that the *z* will be 2.476 rather than –2.476. This positive value shows an ascending trend, rather than a descending one. This will happen because by selecting *Largest to smallest* we would be looking at the medians in the opposite direction (i.e., from 7 to 4 to 1 to 0 meals) compared to selecting *Smallest to largest* (i.e., 0 to 1 to 4 to 7 meals).



Independent-Samples Jonckheere-Terpstra Test for Ordered Alternatives Summary

Total N	80
Test Statistic	912.000
Standard Error	116.333
Standardized Test Statistic	-2.476
Asymptotic Sig.(2-sided test)	.013

Output 7.9



Oliver Twisted
Please Sir, can I have some more ... Jonck?

'I want to know how the Jonckheere–Terpstra Test works,' complains Oliver. Of course you do, Oliver. Sleep is hard to come by these days. I am only too happy to oblige, my friend. The material for this chapter on the companion website explains how to do the test by hand. I bet you're glad you asked.

7.6.7 Calculating an effect size Ⓑ

There isn't an easy way to convert a Kruskal–Wallis test statistic that has more than 1 degree of freedom to an effect size, *r*. However, this kind of effect size is rarely useful anyway (because it's summarizing a general effect). In most cases it's more interesting to know the effect size for a focused comparison (such as when comparing two things). For this reason, I'd suggest calculating effect sizes for the pairwise tests we used to follow up the main analysis. We can do this using Eq. 7.11 but to get the *U*-statistic for each of the six comparisons in Output 7.7 we'd need to select the cases for the two relevant groups, conduct a Mann–Whitney test, select the cases for the next two relevant groups, conduct a Mann–Whitney test, and repeat until we'd died of boredom. Luckily syntax is your friend here. If you use a workbook (Section 4.8.2) then create a new syntax paragraph and use the following generic syntax:

```
NPAR TESTS /M-W= outcome_variable BY group_variable(first_group_code second_
group_code).
```

Replace *outcome_variable* with the name of your outcome variable (in this case **sperm**), *group_variable* with the name of your coding variable (in this case **soya**) and *first_group_code* and *second_group_code*

with the numbers that represent the levels of the groups you want to compare. For example, to conduct the six Mann–Whitney tests that we need we'd type the following syntax into a syntax paragraph and click Run Syntax.

```
NPAR TESTS /M-W= sperm BY soya(1 2).
NPAR TESTS /M-W= sperm BY soya(1 3).
NPAR TESTS /M-W= sperm BY soya(1 4).
NPAR TESTS /M-W= sperm BY soya(2 3).
NPAR TESTS /M-W= sperm BY soya(2 4).
NPAR TESTS /M-W= sperm BY soya(3 4).
```

The first line of syntax conducts a Mann–Whitney test of the sperm counts between level 1 (no soya) and level 2 (1 soya meal), whereas the second line compares level 1 (no soya) with level 3 (4 soya meals) and so on. Every group had 20 participants, so in Eq. 7.11 we get $n_1 n_2 = 400$.

The U-statistics can be taken from each Mann-Whitney test. Table 7.4 shows the effect sizes for each comparison. The effect sizes are substantial (and similar in size) for 7 soya meals compared to all other groups (r_{rb}), in the region of 0.5, but the effect sizes for comparisons of the other groups are trivial ($r_{rb} < 0.1$).

Table 7.4 Calculating effect sizes for pairwise comparisons

Comparison	U	$\frac{2U}{n_1 n_2}$	$r_{rb} = 1 - \frac{2U}{n_1 n_2}$
None vs. 1 meal	191	0.96	0.04
None vs. 4 meals	188	0.94	0.06
None vs. 7 meals	104	0.52	0.48
1 vs. 4 meals	196	0.98	0.02
1 vs. 7 meal	122	0.61	0.39
4 vs. 7 meals	111	0.56	0.44

7.6.8 Writing and interpreting the results Ⓐ

For the Kruskal–Wallis test, report the test statistic (which we saw earlier is denoted by H), its degrees of freedom and its significance. So, we could report something like:

✓ Sperm counts were significantly affected by eating soya meals, $H(3) = 8.66$, $p = 0.034$.

However, we need to report the follow-up tests as well (including their effect sizes):

✓ Sperm counts were significantly affected by eating soya meals, $H(3) = 8.66$, $p = 0.034$. Pairwise comparisons with adjusted p-values showed that there were no significant differences between sperm counts when people ate 7 soya meals per week compared to 4 meals ($p = 0.133$, $r_{rb} = 0.44$), 1 meal ($p = 0.133$, $r_{rb} = 0.39$) or no meals ($p = 0.058$, $r_{rb} = 0.48$). There were also no significant differences in sperm counts between those eating 4 soya meals per week and those eating 1 meal ($p = 1.00$, $r_{rb} = 0.02$) and no meals ($p = 1.00$, $r_{rb} = 0.06$). Finally, there were no significant differences in sperm counts between those eating 1 soya meal per week and those eating none ($p = 1.00$, $r_{rb} = 0.04$).

✓ Sperm counts were significantly affected by eating soya meals, $H(3) = 8.66$, $p = 0.034$. Step-down follow-up analysis showed that if soya is eaten every day it significantly reduces sperm counts compared to eating none; however, eating soya less often than every day has no significant effect on sperm counts, $p = 0.943$ ('phew!' says the vegetarian book author).

Alternatively, we might want to report our trend:

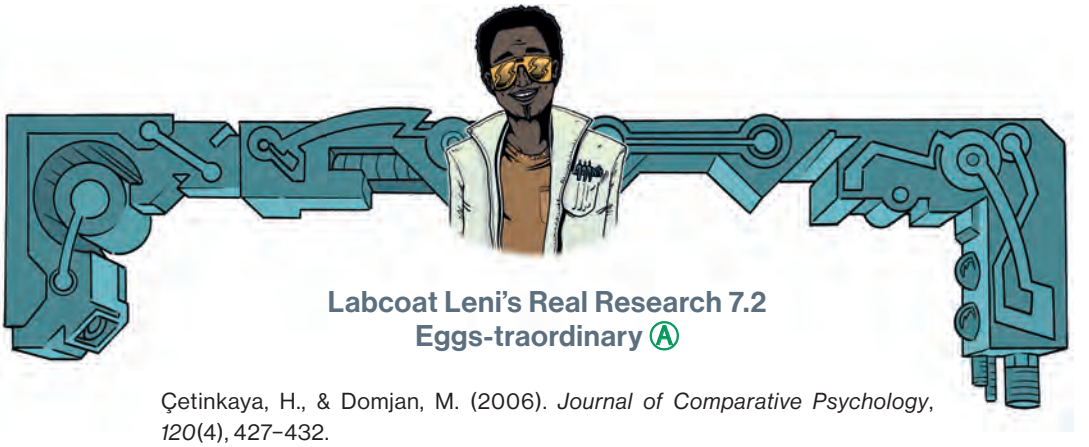
✓ Sperm counts were significantly affected by eating soya meals, $H(3)= 8.66$, $p = 0.034$. A Jonckheere–Terpstra test revealed that as more soya was eaten, the median sperm count significantly decreased, $J = 912$, $z = -2.48$, $p = 0.013$.



Cramming Sam's Tips

- The Kruskal–Wallis test compares several conditions when different participants take part in each condition and the resulting data have unusual cases or violate any assumption in Chapter 6.
- Look at the row labelled *Asymptotic Sig.* A value less than or equal to 0.05 is typically taken to mean that the groups are significantly different.
- Pairwise comparisons compare all possible pairs of groups with a *p*-value that is corrected so that the error rate across all tests remains at 5%.
- If you predict that the medians will increase or decrease across your groups in a specific order then test this with the Jonckheere–Terpstra test.
- Report the *H*-statistic, the degrees of freedom and the significance value for the main analysis. For any follow-up tests, report an effect size, the corresponding *z* and the significance value. Also report the medians and their corresponding ranges (or draw a boxplot).





Labcoat Leni's Real Research 7.2
Eggs-traordinary Ⓐ

Çetinkaya, H., & Domjan, M. (2006). *Journal of Comparative Psychology*, *120*(4), 427–432.

There is a lot of sperm in this book (not literally I hope, stats is not *that* exciting). We saw in Labcoat Leni's Real Research 7.1 that male quail fertilized

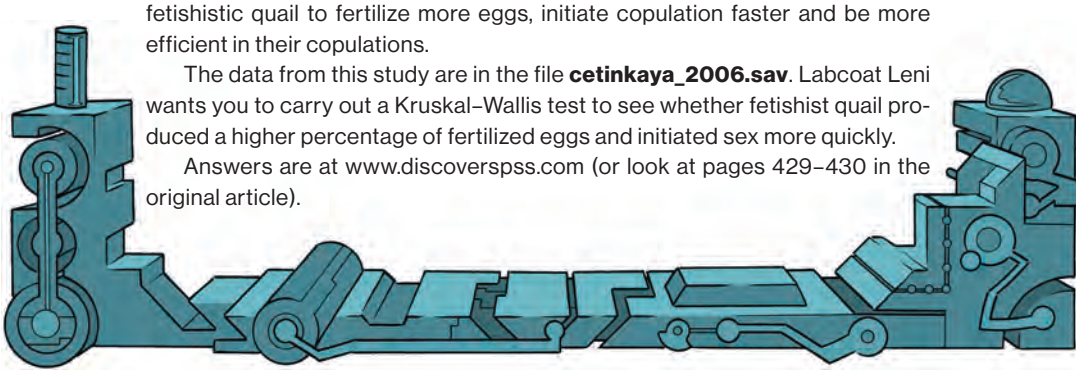
more eggs if they had been trained to be able to predict when a mating opportunity would arise. Some quail develop fetishes. Really. In studies where a terrycloth object acts as a sign that a mate will shortly become available, some quail start to direct their sexual behaviour towards the terrycloth object. (I will regret this analogy, but in human terms if every time you were going to have sex with your boyfriend you gave him a green towel a few moments before seducing him, then after enough seductions he would start getting really excited by green towels. If you're planning to dump your boyfriend a towel fetish could be an entertaining parting gift.) In evolutionary terms, this fetishistic behaviour seems counterproductive because sexual behaviour becomes directed towards something that cannot provide reproductive success. However, perhaps this behaviour serves to prepare the organism for the 'real' mating behaviour.

Hakan Çetinkaya and Mike Domjan sexually conditioned male quail (Çetinkaya & Domjan, 2006). All quail experienced the terrycloth stimulus and an opportunity to mate, but for some the terrycloth stimulus immediately preceded the mating opportunity (paired group) whereas others experienced a 2-hour delay (this acted as a control group because the terrycloth stimulus did not predict a mating opportunity). In the paired group, quail were classified as fetishistic or not depending on whether they engaged in sexual behaviour with the terrycloth object.

During a test trial the quail mated with a female and the researchers measured the percentage of eggs fertilized, the time spent near the terrycloth object, the latency to initiate copulation, and copulatory efficiency. If this fetishistic behaviour provides an evolutionary advantage then we would expect the fetishistic quail to fertilize more eggs, initiate copulation faster and be more efficient in their copulations.

The data from this study are in the file **cetinkaya_2006.sav**. Labcoat Leni wants you to carry out a Kruskal–Wallis test to see whether fetishist quail produced a higher percentage of fertilized eggs and initiated sex more quickly.

Answers are at www.discoverspss.com (or look at pages 429–430 in the original article).



7.7 Differences between several related groups: Friedman's ANOVA Ⓐ

The final test we'll look at is **Friedman's ANOVA** (Friedman, 1937), which tests differences between three or more conditions when the scores across conditions are related (usually because the same entities have provided scores in all conditions). As with all the tests in this chapter, Friedman's ANOVA is used to counteract the presence of unusual cases or when one of the assumptions from Chapter 6 has been violated.

In the Western world we are influenced by the media into believing that only certain body types are attractive. We can end up terribly depressed that we're not perfect because we don't have someone to photoshop out our imperfections, our teeth are not white enough, and we have jobs so we can't spend 8 hours a day at the gym working on a six-pack (not bitter at all …). Parasites exploit our vulnerability

to make loads of money on diets and exercise regimes that will help us attain the body beautiful, which we think will fill the emotional void in our lives. Not wishing to miss this great opportunity, I developed the Andikins diet.[9] The principle is to follow my exemplary (ahem) lifestyle: eat no meat, drink lots of Darjeeling tea, fill up on truckloads of lovely European cheese, fresh crusty bread, pasta and chocolate at every available opportunity (especially when writing books), enjoy an occasional beer, play soccer and drums as much as humanly possible (not simultaneously). To test the efficacy of my new diet, I took 10 people who thought that they needed to lose weight and put them on this diet for 2 months. Their weight was measured in kilograms at the start of the diet and then after 1 month and 2 months.

7.7.1 Theory of Friedman's ANOVA Ⓑ

As with all the tests in this chapter, Friedman's ANOVA works on ranked data. To begin with, we'll place the data for different conditions into different columns (in this case there were three conditions, so we have three columns) – see Table 7.5. Each row represents the weight of a different person, and each column represents their weight at a different point in time. Next, we rank the data *for each person*. So, we start with person 1, we look at their scores (in this case person 1 weighed 63.75 kg at the start, 65.38 kg after one month on the diet and 81.34 kg after two months on the diet) and then we give the lowest one a rank of 1, the next highest a rank of 2 and so on (see Section 7.4.1 for more detail). When you've ranked the data for the first person, you move onto the next person, and starting at 1, rank their lowest score, then rank the next highest as 2 and so on. You do this for every row, and then add up the ranks for each condition (R_i, where i denotes the particular condition).



Table 7.5 Data for the diet example with ranks

	Weight				Weight		
	Start	Month 1	Month 2		Start (Ranks)	Month 1 (Ranks)	Month 2 (Ranks)
Person 1	63.75	65.38	81.34		1	2	3
Person 2	62.98	66.24	69.31		1	2	3
Person 3	65.98	67.70	77.89		1	2	3
Person 4	107.27	102.72	91.33		3	2	1
Person 5	66.58	69.45	72.87		1	2	3
Person 6	120.46	119.96	114.26		3	2	1
Person 7	62.01	66.09	68.01		1	2	3
Person 8	71.87	73.62	55.43		2	3	1
Person 9	83.01	75.81	71.63		3	2	1
Person 10	76.62	67.66	68.60		3	1	2
				R_i	19	20	21

9 Not to be confused with the Atkins diet, obviously.☺

Once the sum of ranks has been calculated for each group, the test statistic, F_r, is calculated as follows:

$$F_r = \left[\frac{12}{Nk(k+1)}\sum_{i=1}^{k} R_i^2\right] - 3N(k+1) \tag{7.28}$$

in which R_i is the sum of ranks for each group, N is the total sample size (in this case 10) and k is the number of conditions (in this case 3). The equation is similar to the one for the Kruskal–Wallis test (Eq. 7.26). The middle part of the equation tells us to square the sum of ranks for each condition and add up these values. The rest of the equation calculates constants based on the total sample size and the number of conditions. For these data we get:

$$\begin{aligned} F_r &= \left[\frac{12}{(10\times 3)(3+1)}\left(19^2 + 20^2 + 21^2\right)\right] - (3\times 10)(3+1) \\ &= \frac{12}{120}(361 + 400 + 441) - 120 \\ &= 0.1(1202) - 120 \\ &= 120.2 - 120 \\ &= 0.2 \end{aligned} \tag{7.29}$$

When the number of people tested is greater than about 10 this test statistic, like the Kruskal–Wallis test in the previous section, has a chi-square distribution with degrees of freedom that are one fewer than the number of groups (i.e., k–1), in this case 2.

7.7.2 Inputting data and provisional analysis Ⓐ



Using what you know about inputting data, enter these data into SPSS and run exploratory analyses (see Chapter 6).

When the data are collected using the same participants in each condition, scores are entered in different columns. The data editor will have at least three columns of data. One column is for the data from the start of the diet (called something like **start**), another will have values for the weights after one month (called **month1**) and the final column will have the weights at the end of the diet (called **month2**). The data can be found in the file **diet.sav**.

An exploratory analysis is shown in Figure 7.14 and Output 7.10. The normal Q-Q plots show quite clear deviations from normality at all three time points because the dots deviate from the diagonal line. These plots are evidence enough that our data are not normal, and because our sample size is small we can't rely on the central limit theorem to get us out of trouble. If you're keen on normality tests, then p-values less than or equal to 0.05 (or whatever threshold you choose) in these tests would support the belief of a lack of normality because the small sample size would mean that these tests would only have power to detect severe deviations from normal. (It's worth reminding you that non-significance in this context tells us nothing useful because our sample size is so small.) If you do these tests (Output 7.10), you'll find that the more accurate Shapiro–Wilk test is significant at the start of the diet (p = 0.009), after 1 month (p = 0.001), but not at the end of the diet (p = 0.121). The tests and Q-Q plots converge on a belief in non-normal data or unusual cases at all time points.

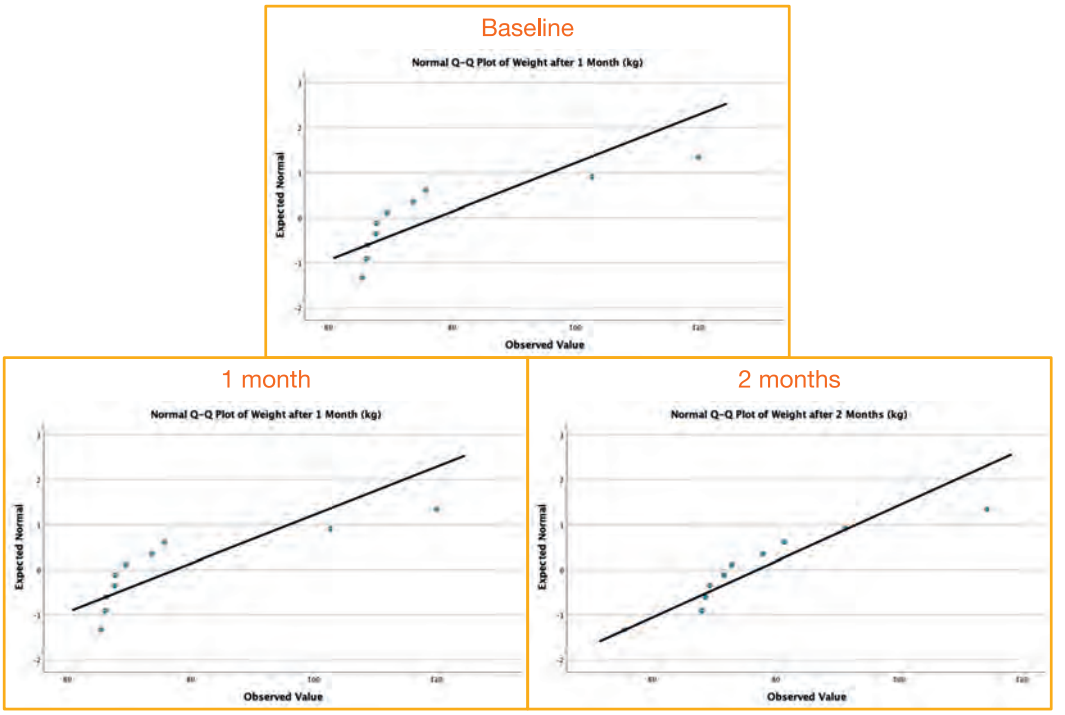


Figure 7.14 Q-Q plots of the diet data

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Weight at Start (kg)	.228	10	.149	.784	10	.009
Weight after 1 Month (kg)	.335	10	.002	.685	10	<.001
Weight after 2 Months (kg)	.203	10	.200*	.877	10	.121

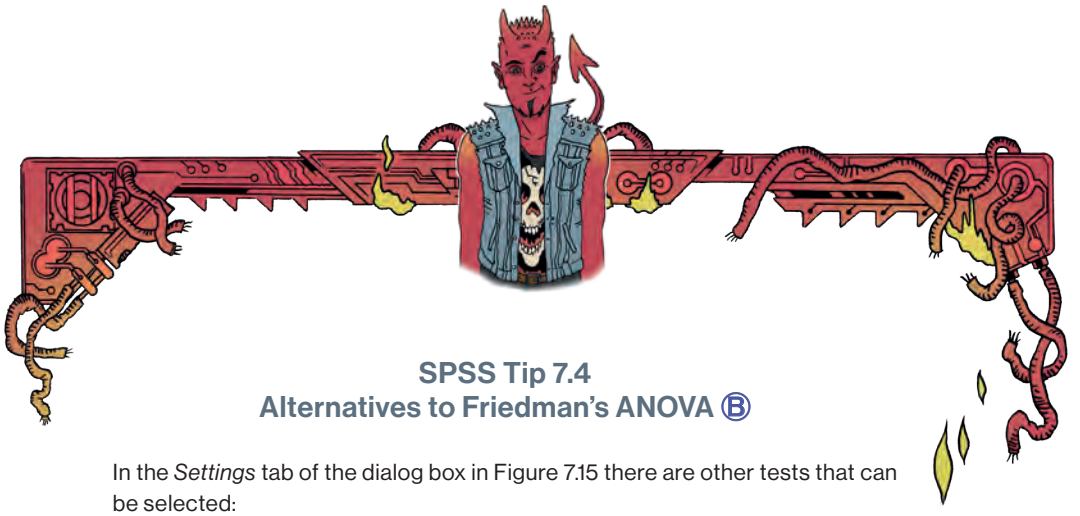
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Output 7.10

7.7.3 Doing Friedman's ANOVA using SPSS Ⓐ

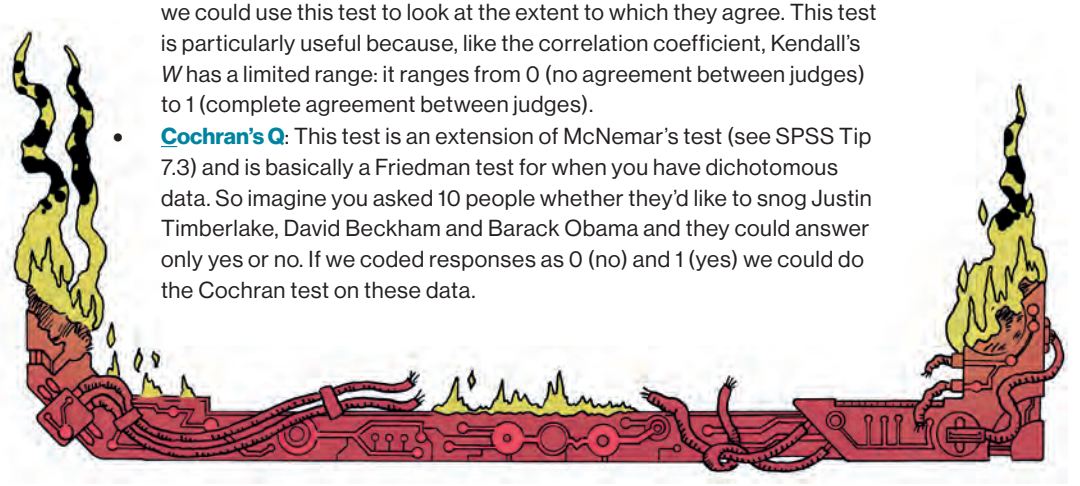
Follow the general procedure outlined in Section 7.3 by first selecting *Analyze* > *Nonparametric Tests* > Related Samples... . When you reach the Objective Fields Settings tab, if you need to assign variables then select ◉ Use custom field assignments and specify the model (Figure 7.15, top) by dragging **Start, Month1** and **Month2** into the box labelled *Test Fields* (or select them in the box labelled *Fields* and click ➧). Activate the Objective Fields Settings tab, select ◉ Customize tests (see SPSS Tip 7.4) and check ☑ Friedman's 2-way ANOVA by ranks (k samples). Next to this option is a drop-down list labelled *Multiple comparisons* (Figure 7.15, bottom), just as there was for the Kruskal–Wallis test. Within this list are the two options we discussed earlier: compare every group against every other group (*All pairwise*) or use a step-down method (*Stepwise step-down*). To run the analysis click Run.



SPSS Tip 7.4
Alternatives to Friedman's ANOVA Ⓑ

In the *Settings* tab of the dialog box in Figure 7.15 there are other tests that can be selected:

- **Kendall's W** (coefficient of concordance): This is similar to Friedman's ANOVA but is used specifically for looking at the agreement between raters. If, for example, we asked 10 different people to rate the attractiveness of Justin Timberlake, David Beckham, and Barack Obama we could use this test to look at the extent to which they agree. This test is particularly useful because, like the correlation coefficient, Kendall's *W* has a limited range: it ranges from 0 (no agreement between judges) to 1 (complete agreement between judges).
- **Cochran's Q**: This test is an extension of McNemar's test (see SPSS Tip 7.3) and is basically a Friedman test for when you have dichotomous data. So imagine you asked 10 people whether they'd like to snog Justin Timberlake, David Beckham and Barack Obama and they could answer only yes or no. If we coded responses as 0 (no) and 1 (yes) we could do the Cochran test on these data.



7.7.4 Output from Friedman's ANOVA Ⓐ

Output 7.11 has a table containing the test statistic, F_r, for the Friedman test (0.2, which we calculated earlier), its degrees of freedom (in this case we had 3 groups so they are 3 – 1, or 2) and the associated *p*-value (significance). The significance value is 0.905, which is well above 0.05, and would typically lead to a conclusion that the weights didn't significantly change over the course of the diet. The histograms in the output show the distribution of ranks across the three groups. It's clear that the distributions do not change over time, and in particular their centres (the mean ranks) change very little over time: weights are centred at 1.90 (baseline), 2.00 (1 month) and 2.10 (2 months). The lack of significance of the test statistic reflects the fact that the mean rank doesn't change over time.

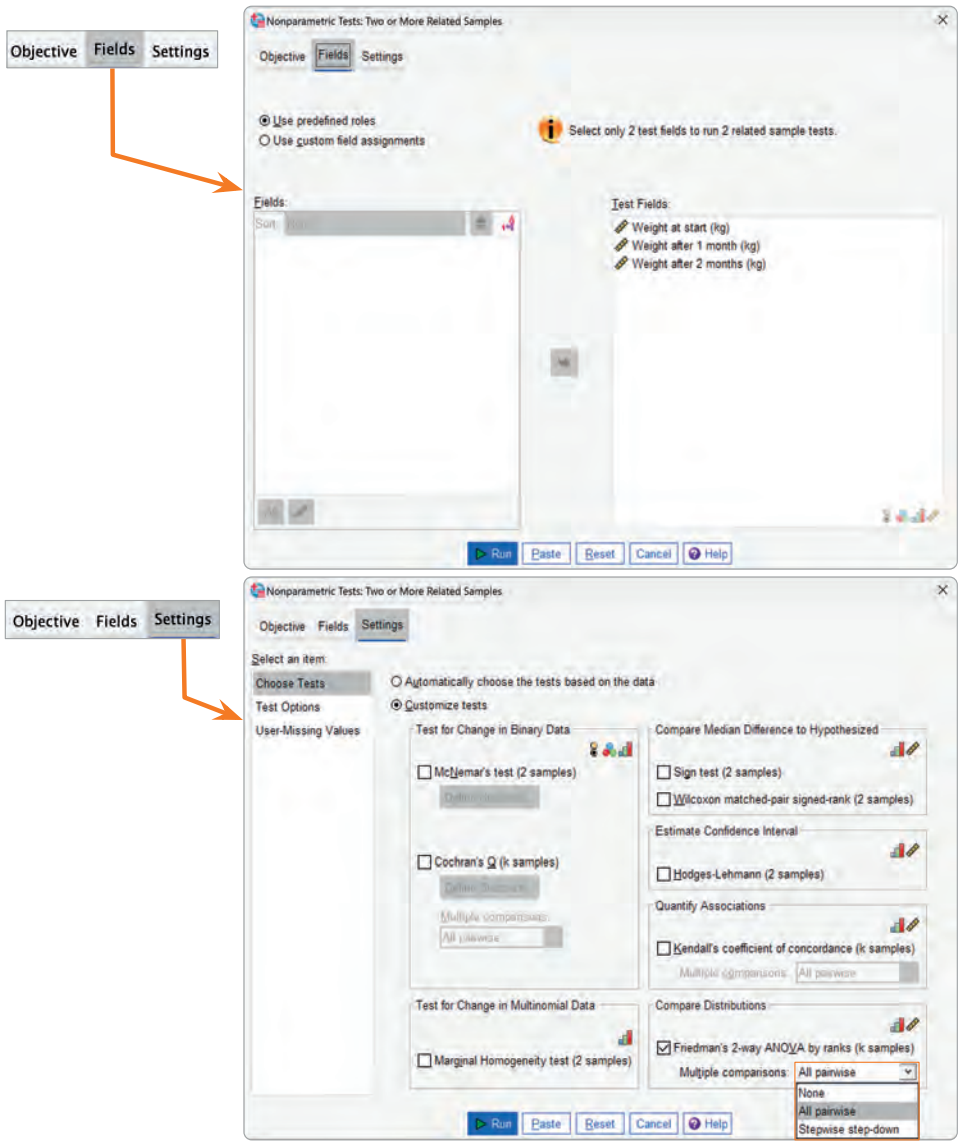
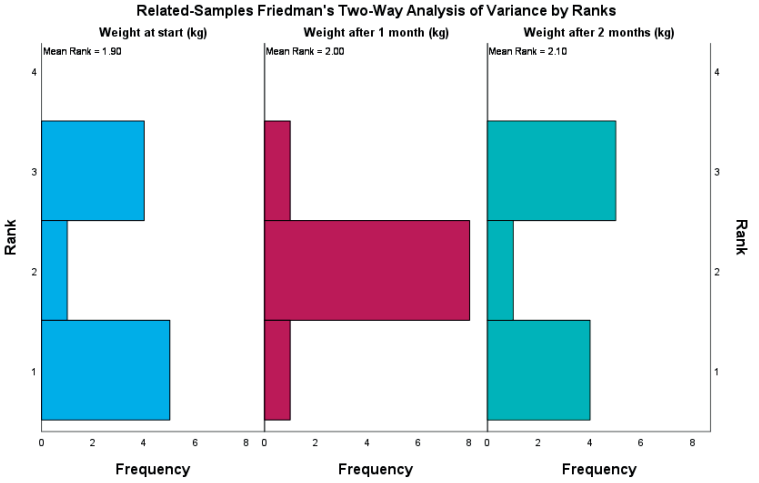


Figure 7.15 Dialog boxes for Friedman's ANOVA

7.7.5 Following up Friedman's ANOVA Ⓑ

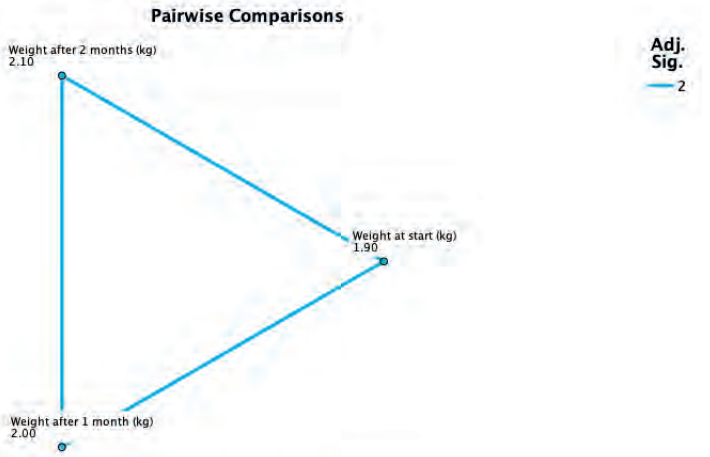
As with the Kruskal–Wallis test, we can follow up a Friedman test by comparing all groups, or using a step-down procedure (Section 7.6.2). The output you see depends on whether you selected *All pairwise* or *Stepwise step-down* in the drop-down list labelled *Multiple comparisons* when you ran the analysis (Figure 7.15). Given that the Friedman test isn't significant, we shouldn't look at these *post hoc* tests, but for the record Output 7.12 shows this part of the output. First there is a diagram showing the mean rank of each condition and linking them with lines that differ in colour depending on whether the conditions that the line links are significantly different or not. Our diagram only has blue lines, and based on the legend of the plot tells us that all paths are non-significant. We can also see this fact from the table of *post hoc* tests showing that for all comparisons the adjusted *p*-value is 1 (i.e. as non-significant as it's possible to be).



Related-Samples Friedman's Two-Way Analysis of Variance by Ranks Summary

Total N	10
Test Statistic	.200
Degree Of Freedom	2
Asymptotic Sig.(2-sided test)	.905

Output 7.11



Pairwise Comparisons

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Weight at start (kg)-Weight after 1 month (kg)	-.100	.447	-.224	.823	1.000
Weight at start (kg)-Weight after 2 months (kg)	-.200	.447	-.447	.655	1.000
Weight after 1 month (kg)-Weight after 2 months (kg)	-.100	.447	-.224	.823	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .050.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

Output 7.12

7.7.6 Calculating an effect size Ⓑ

It's most sensible (in my opinion) to calculate effect sizes for any comparisons you've done after the main test. In this example we didn't have any follow-up analyses because the overall effect was non-significant. However, effect sizes for these comparisons might still be useful so that people can see the magnitude of group differences. Like when we followed up the Kruskal–Wallis test, we'd need to conduct a series of Wilcoxon tests from which we can extract T_+ and T_- to plug into Eq. 7.23. In this example, we have three groups, which we can compare with three tests:

- Test 1: Weight at the start of the diet compared to at 1 month.
- Test 2: Weight at the start of the diet compared to at 2 months.
- Test 3: Weight at 1 month compared to at 2 months.

One issue is that the Wilcoxon test output gives us the sum of positive ranks (the tests statistic, T_+) but not the sum of negative ranks (T_-). Again, syntax comes to the rescue. In your workbook (Section 4.8.2) create a new syntax paragraph and use the following generic syntax:

```
NPAR TESTS
  /WILCOXON= list_of_variables WITH list_of_variables (PAIRED)
  /STATISTICS DESCRIPTIVES.
```

Replace *list_of_variables* with the names of the variables that you want to compare against each other. For example, to conduct the three Wilcoxon tests between all combinations of the variables **start**, **month1** and **month2** type the following syntax into a syntax paragraph and click Run Syntax:

```
NPAR TESTS
  /WILCOXON=start start month1 WITH month1 month2 month2 (PAIRED)
  /STATISTICS DESCRIPTIVES
```

The first two lines run the Wilcoxon tests but the last line is crucial because it will produce a table of descriptive statistics containing the sums of positive (T_+) and negative (T_-) ranks. We need these values to compute the effect sizes.

Output 7.13 shows the descriptive statistics for the three Wilcoxon signed-rank tests. We need the values in the *Sum of Ranks* column and the rows labelled *Negative Ranks* and *Positive Ranks*. When comparing baseline to 1 month we get $T_+ = 27$ and $T_- = 28$, and a tiny effect size (using Eq. 7.23) of

$$r_{rb(\text{Start–1 Month})} = \frac{T_+ - T_-}{T_+ + T_-} = \frac{27 - 28}{27 + 28} = -0.02 \tag{7.30}$$

For the second comparison (start weight vs. 2 months), $T_+ = 25$ and $T_- = 30$, again yielding a tiny effect size of

$$r_{rb(\text{Start–2 Months})} = \frac{25 - 30}{25 + 30} = -0.09 \tag{7.31}$$

For the final comparison (1 month vs. 2 months), $T_+ = 26$ and $T_- = 29$, yielding yet another miniscule effect size of

$$r_{rb(\text{1 Month–2 Months})} = \frac{26 - 29}{26 + 29} = -0.05 \tag{7.32}$$

Unsurprisingly, given the lack of significance of the Friedman test, these effect sizes are all very close to zero, indicating virtually non-existent effects.

Ranks

		N	Mean Rank	Sum of Ranks
Weight after 1 month (kg) – Weight at start (kg)	Negative Ranks	4[a]	7.00	28.00
	Positive Ranks	6[b]	4.50	27.00
	Ties	0[c]		
	Total	10		
Weight after 2 months (kg) – Weight at start (kg)	Negative Ranks	5[d]	6.00	30.00
	Positive Ranks	5[e]	5.00	25.00
	Ties	0[f]		
	Total	10		
Weight after 2 months (kg) – Weight after 1 month (kg)	Negative Ranks	4[g]	7.25	29.00
	Positive Ranks	6[h]	4.33	26.00
	Ties	0[i]		
	Total	10		

a. Weight after 1 month (kg) < Weight at start (kg)
b. Weight after 1 month (kg) > Weight at start (kg)
c. Weight after 1 month (kg) = Weight at start (kg)
d. Weight after 2 months (kg) < Weight at start (kg)
e. Weight after 2 months (kg) > Weight at start (kg)
f. Weight after 2 months (kg) = Weight at start (kg)
g. Weight after 2 months (kg) < Weight after 1 month (kg)
h. Weight after 2 months (kg) > Weight after 1 month (kg)
i. Weight after 2 months (kg) = Weight after 1 month (kg)

Output 7.13

7.7.7 Writing and interpreting the results Ⓐ

For Friedman's ANOVA we report the test statistic, denoted by χ^2_F, its degrees of freedom and its significance.[10] So, we could report something like:

✓ The weight of participants did not significantly change over the two months of the diet, $\chi^2(2) = 0.20$, $p = 0.91$.

Although with no significant initial analysis we wouldn't report follow-up tests for these data, in case you need to, you should write something like this:

✓ The weight of participants did not significantly change over the two months of the diet, $\chi^2(2) = 0.20$, $p = 0.91$. Wilcoxon tests were used to follow up this finding. It appeared that weight didn't significantly change from the start of the diet to one month, $T = 27$, $r_{rb} = -0.02$, from the start of the diet to two months, $T = 25$, $r_{rb} = -0.09$, or from one month to two months, $T = 26$, $r_{rb} = -0.05$. We can conclude that the Andikins diet, like its creator, is a complete failure.

10 The test statistic is sometimes written without the F as χ^2.



Cramming Sam's Tips
Friedman's ANOVA

- Friedman's ANOVA compares several conditions when the data are related (usually because the same participants take part in each condition) and the resulting data have unusual cases or violate any assumption in Chapter 6.
- Look at the row labelled *Asymptotic Sig.* If the value is less than or equal to 0.05 then typically people conclude that the conditions are significantly different.
- You can follow up the main analysis with pairwise comparisons. These tests compare all possible pairs of conditions using a *p*-value that is adjusted such that the overall Type I error rate remains at 5%.
- Report the χ^2 statistic, the degrees of freedom and the significance value for the main analysis. For any follow-up tests, report the rank-biserial correlation as an effect size, the corresponding *z* and the significance value.
- Report the medians and their ranges (or draw a boxplot).



7.8 Jane and Brian's story Ⓐ

'Jane is an anomaly,' Brian thought, 'an awkward, mysterious, complicated set of observations that don't conform to my conventional assumptions.' Maybe that was why he couldn't shake her from his mind. She had a brilliant mind that compelled her to him. He'd never felt that comfortable around girls. He'd had a few relationships, but they all ended the same way: he'd get dumped for some more exciting guy. His dad said he was too nice, and maybe he was – he'd certainly had his fair share of being a doormat. He'd never lost hope – he dreamed of a Disney fairy tale, where the last person he expected turned out to be his soulmate. He probably needed to get real. Maybe with Jane the writing was on the wall. He should forget about her and knuckle down to his studies. The trouble was, he got excited every time he understood something new about statistics, and Jane was the person with whom he wanted to share his excitement.

'The campus guy is an anomaly,' Jane thought, 'an awkward, mysterious, complicated set of observations that don't conform to my conventional assumptions.' Maybe that was why she couldn't shake him from her mind. His efforts to impress her were sweet. He certainly didn't fit into her mental model of men. She'd got into the habit of being curt to men; it was what they deserved. She'd liked a guy, Josh, at school. She was crazy about him, but he was one of the popular kids and she was not on

his radar. Not until he asked her to the cinema. She could hardly breathe when he asked. She bought clothes, spent hours getting ready, and became the stereotype that she'd fought against being. She arrived and waited. Most of her class were in the foyer. She hadn't expected that and it made her feel conspicuous and awkward. He was late and every minute increased her embarrassment and stress. She wanted to run. She should have run: it would have spared her. When he paraded confidently through the doors like a film star with Eliza Hamilton on his arm she knew instantly that she'd been set up. They turned to Jane, placed their fingers in an L on their foreheads, and mouthed 'loser', causing the foyer to erupt in laughter. Jane felt a rage that she hadn't known she was capable of, but she didn't run. She put her headphones in, calmly walked past Josh and smiled. It was a smile that wiped his own one from his face.

The funny thing was that Josh wasn't popular for long. First a bunch of emails leaked in which he tried to sell information about the school football team, then he posted highly personal information about Eliza all over social media. He became an outcast. He lost his place in the team, his grades plummeted and no one spoke to him. He protested his innocence, but only a genius could have hacked his online life so comprehensively. Jane smiled at the memory. Campus guy seemed different. He was sweet … but not enough to let her guard down.



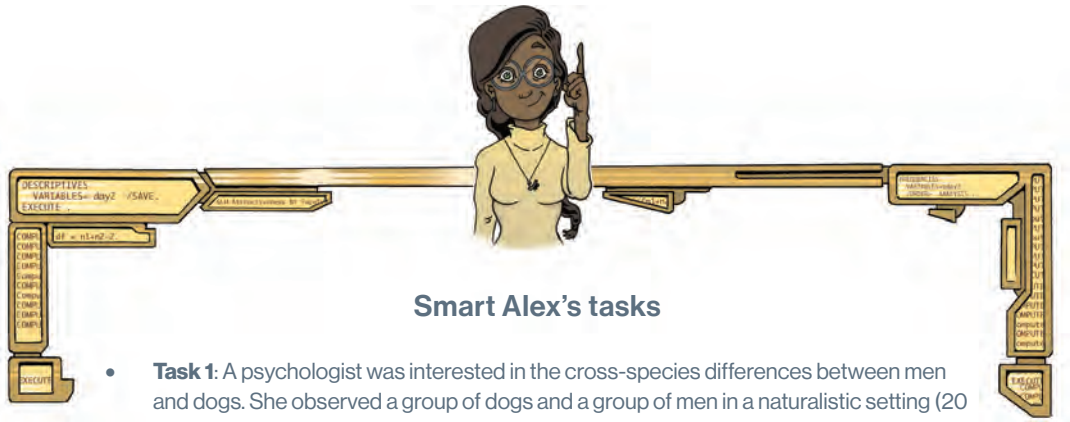
Figure 7.16 What Brian learnt from this chapter

7.9 What next? Ⓐ

'You promised us swans,' I hear you cry, 'and all we got was Kruskal this and Wilcoxon that. Where were the swans?!' Well, the King owns them all so I wasn't allowed to have them. Nevertheless, this chapter did negotiate Dante's eighth circle of hell (Malebolge), where data of deliberate and knowing evil dwell. That is, data don't always behave themselves. Unlike the data in this chapter, my formative years at school were spent being very well behaved and uninteresting. However, a mischievous and rebellious streak was growing inside. Perhaps the earliest signs were my taste in music. Even from about the age of 3 music was my real passion: one of my earliest memories is of listening to my dad's rock and soul records while waiting for my older brother to come home from school. I still have a nostalgic obsession with vinyl, which finally the world has caught up with; wait long enough and you become a hipster again (sort of). The first record I asked my parents to buy me was 'Take on the world' by heavy metal band Judas Priest, which I'd heard on *Top of the Pops* (a now defunct UK TV show) and liked. Watching the Priest on *Top of the Pops* was such a musical epiphany that the memory was etched into my brain with (British) steel. This record came out in 1978 when I was 5. Some people think that this sort of music corrupts young minds. Let's see if it did …

7.10 Key terms that I've discovered

Cochran's Q
Friedman's ANOVA
Jonckheere–Terpstra test
Kendall's *W*
Kolmogorov–Smirnov *Z*
Kruskal–Wallis test
Mann–Whitney test
McNemar's test
Median test
Monte Carlo method
Moses extreme reactions
Non-parametric tests
Pairwise comparisons
Rank-biserial correlation
Ranking
Sign test
Wald–Wolfowitz runs
Wilcoxon rank-sum test
Wilcoxon signed-rank test



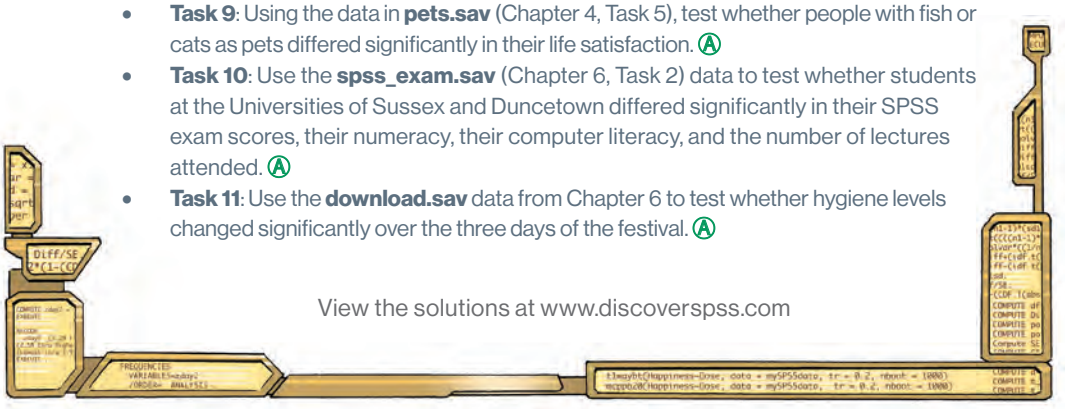
Smart Alex's tasks

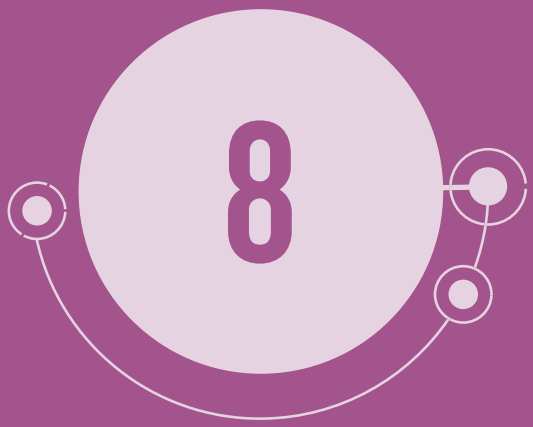
- **Task 1**: A psychologist was interested in the cross-species differences between men and dogs. She observed a group of dogs and a group of men in a naturalistic setting (20 of each). She classified several behaviours as being dog-like (urinating against trees and lampposts, attempts to copulate and attempts to lick their own genitals). For each man and dog she counted the number of dog-like behaviours displayed in a 24-hour period. It was hypothesized that dogs would display more dog-like behaviours than men. Analyse the data in **men_dogs.sav** with a Mann–Whitney test. Ⓐ
- **Task 2**: Both Ozzy Osbourne and Judas Priest have been accused of putting backward masked messages on their albums that subliminally influence poor unsuspecting teenagers into doing things like blowing their heads off with shotguns. A psychologist was interested in whether backward masked messages could have an

effect. He created a version of Taylor Swift's 'Shake it off' that contained the masked message 'deliver your soul to the dark lord' repeated in the chorus. He took this version, and the original, and played one version (randomly) to a group of 32 people. Six months later he played them whatever version they hadn't heard the time before. So each person heard both the original and the version with the masked message, but at different points in time. The psychologist measured the number of goats that were sacrificed in the week after listening to each version. Test the hypothesis that the backward message would lead to more goats being sacrificed using a Wilcoxon signed-rank test (**dark_lord.sav**). Ⓐ

- **Task 3**: A media researcher was interested in the effect of television programmes on domestic life. She hypothesized that through 'learning by watching', certain programmes encourage people to behave like the characters within them. She exposed 54 couples to three popular TV shows, after which the couple were left alone in the room for an hour. The experimenter measured the number of times the couple argued. Each couple viewed all TV shows but at different points in time (a week apart) and in a counterbalanced order. The TV shows were *EastEnders* (a UK TV show which portrays the lives of extremely miserable, argumentative, London folk who spend their lives assaulting each other, lying and cheating), *Friends* (which portrays unrealistically considerate and nice people who love each other oh so very much – but I love it anyway) and a *National Geographic* programme about whales (this was a control). Test the hypothesis with Friedman's ANOVA (**eastenders.sav**). Ⓐ
- **Task 4**: A researcher was interested in preventing coulrophobia (fear of clowns) in children. She did an experiment in which different groups of children (15 in each) were exposed to positive information about clowns. The first group watched adverts in which Ronald McDonald is seen dancing with children and singing about how they should love their mums. A second group was told a story about a clown who helped some children when they got lost in a forest (what a clown was doing in a forest remains a mystery). A third group was entertained by a real clown, who made balloon animals for the children. A final, control, group had nothing done to them at all. Children rated how much they liked clowns from 0 (not scared of clowns at all) to 5 (very scared of clowns). Use a Kruskal–Wallis test to see whether the interventions were successful (**coulrophobia.sav**). Ⓐ
- **Task 5**: Thinking back to Labcoat Leni's Real Research 4.1, test whether the number of offers was significantly different in people listening to Bon Scott compared to those listening to Brian Johnson (**acdc.sav**). Compare your results to those reported by Oxoby (2008). Ⓐ
- **Task 6**: Repeat the analysis above, but using the minimum acceptable offer – see Chapter 4, Task 3. Ⓐ
- **Task 7**: Using the data in **shopping.sav** (Chapter 4, Task 4), test whether men and women spent significantly different amounts of time shopping? Ⓐ
- **Task 8**: Using the same data, test whether men and women walked significantly different distances while shopping. Ⓐ
- **Task 9**: Using the data in **pets.sav** (Chapter 4, Task 5), test whether people with fish or cats as pets differed significantly in their life satisfaction. Ⓐ
- **Task 10**: Use the **spss_exam.sav** (Chapter 6, Task 2) data to test whether students at the Universities of Sussex and Duncetown differed significantly in their SPSS exam scores, their numeracy, their computer literacy, and the number of lectures attended. Ⓐ
- **Task 11**: Use the **download.sav** data from Chapter 6 to test whether hygiene levels changed significantly over the three days of the festival. Ⓐ

View the solutions at www.discoverspss.com





CORRELATION

8.1 What will this chapter tell me? 374
8.2 Modelling relationships 375
8.3 Data entry for correlation analysis 384
8.4 Bivariate correlation 384
8.5 Partial and semi-partial correlation 397
8.6 Comparing correlations 404
8.7 Calculating the effect size 405
8.8 How to report correlation coefficients 405
8.9 Jane and Brian's story 406
8.10 What next? 409
8.11 Key terms that I've discovered 409
Smart Alex's tasks 409

Letters A to D indicate increasing difficulty

8.1 What will this chapter tell me? Ⓐ

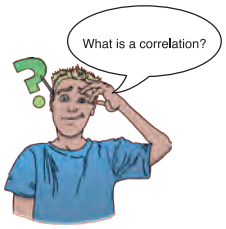
When I was 8 years old, my parents bought me a guitar for Christmas. Even then, I'd desperately wanted to play the guitar for years. I could not contain my excitement at getting this gift (had it been an *electric* guitar I think I would have exploded with excitement). The guitar came with a 'learn to play' book, and after some time trying to play what was on page 1 of this book, I readied myself to unleash a riff of universe-crushing power onto the world (well, 'Skip to my Lou'). But, I couldn't do it. I burst into tears and ran upstairs to hide.[1] My dad sat with me and said something like 'Don't worry, Andy, everything is hard to begin with, but the more you practise the easier it gets.' With his comforting words, my dad was inadvertently teaching me about the relationship, or correlation, between two variables. These two variables could be related in three ways: (1) *positively related*, meaning that the more I practised my guitar, the better a guitar player I would become (i.e., my dad was telling me the truth); (2) *not related* at all, meaning that as I practised the guitar my playing ability would remain completely constant (i.e., my dad had fathered a fool); or (3) *negatively related*, which would mean that the more I practised the guitar the worse a guitar player I would become (i.e., my dad had fathered an indescribably strange child). This chapter looks first at how we can express the relationships between variables statistically by looking at two measures: *covariance* and the *correlation coefficient*. We discover how to carry out and interpret correlations in IBM SPSS Statistics software. The chapter ends by looking at more complex measures of relationships; in doing so, it acts as a precursor to the chapter on the linear model.



Figure 8.1 I don't have a photo from Christmas 1981, but this was taken about that time at my grandparents' house. I'm trying to play an 'E' by the looks of it, no doubt because it's in 'Take on the world'.

1 This is not a dissimilar reaction to the one I have when publishers ask me for new editions of statistics textbooks.

8.2 Modelling relationships Ⓐ



In Chapter 5 I stressed the importance of looking at your data graphically before fitting models to them. Our starting point with a correlation analysis is, therefore, to look at scatterplots of the variables we have measured. I am not going to repeat how to produce these graphs, but I urge you (if you haven't done so already) to read the relevant section before embarking on the rest of this chapter.

Way back in Chapter 1 we started talking about fitting models to your data that represent the hypothesis you're trying to test. We touched on a particular model (the general linear model or GLM), and in Chapter 6 we started to look at that model in detail. We looked at how we fit a model to sample data to estimate a 'population model' that we hope represents some kind of truth about the relationship between variables. We saw that the linear model with one predictor is represented by this equation:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i \tag{8.1}$$

Remember that this model expresses the relationship between an outcome variable (Y_i) and a predictor variable (X_i). The bs in the model are known as parameters. The parameter attached to a predictor variable, b_1, quantifies the relationship it has with the outcome variable, and the parameter b_0 tells us the value of the outcome when the predictor is zero. It's also important to have in mind that we can't know the true value of the bs for certain, but we can estimate them using a sample of data in which we measure the variables in the model. We can express the fact that we're estimating values by putting hats on them (little-known fact, but whenever you wear a hat you turn into an estimate of yourself):

$$Y_i = \hat{b}_0 + \hat{b}_1 X_i + e_i \tag{8.2}$$

The form of the model is unchanged, but the hats remind us that the values of b are estimates (and there will be uncertainty around those values). If we work with the raw scores then to make predictions about the outcome variable we need to know where the outcome is centred: in other words, what the value of the outcome is when the predictor is absent from the model (i.e., the predictor is zero). This gives us a starting point for our prediction (i.e., if there were no predictor variable, what value of the outcome would we expect?). We add this value into the model as a constant, b_0, known as the intercept (we discussed this in detail in the previous chapter).

If we were to standardize our scores before fitting the model by converting them to z-scores then both the predictor and outcome will have a mean of 0 (because that's what standardizing does). In this situation, we know that the average value of the outcome when the predictor isn't in the model is 0. In other words, b_0 is zero and drops out of the model, leaving us with b_1:

$$z(Y)_i = \hat{b}_1 z(X_i) + e_i \tag{8.3}$$

This equation means that the outcome expressed as z-scores can be predicted from the predictor variable (also expressed as z-scores) multiplied by b_1. When working with standardized scores like this, b_1 is called the *Pearson product-moment correlation coefficient*, and when we're not formally expressing the model as in Eq. 8.3, it is denoted by the letter r. Remember that we use the sample to estimate a value for b_1 (i.e., r) in the population, and its value quantifies the strength and direction of relationship between the predictor and outcome. How do we estimate this parameter? Like a quest for fire, we could search across the land … or we could use maths. Within the context of the GLM we typically use ordinary least squares estimation (Section 2.7); we'll now see what that looks like.

8.2.1 A detour into the murky world of covariance Ⓐ

The simplest way to look at whether two variables are associated is to look at whether they *covary*. To understand what covariance is, we first need to think back to the concept of variance that we met in Chapter 1. Remember that the variance of a single variable represents the average amount that the data vary from the mean. Numerically, it is described by:

$$\text{variance}\left(s^2\right) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{N-1} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)}{N-1} \tag{8.4}$$

The mean of the sample is represented by $\bar{x}$, x_i is the data point in question and N is the number of observations. If two variables are related, then changes in one variable should be met with similar changes in the other variable. Therefore, when one variable deviates from its mean we would expect the other variable to deviate from its mean in a similar way.

To illustrate what I mean, imagine we took five people and subjected them to a certain number of advertisements promoting toffee sweets, and then measured how many packets of those sweets each person bought during the next week. The data are in Table 8.1 as well as the mean and standard deviation (s) of each variable.

Table 8.1 Some data about toffees and adverts

Participant:	1	2	3	4	5	Mean	s
Adverts Watched	5	4	4	6	8	5.4	1.673
Packets Bought	8	9	10	13	15	11.0	2.915

If there were a relationship between these two variables, then as one variable deviates from its mean, the other variable should deviate from its mean in the same or the directly opposite way. Figure 8.2 shows the data for each participant (green circles represent the number of packets bought and blue circles represent the number of adverts watched); the green line is the average number of packets bought and the blue line is the average number of adverts watched. The vertical lines represent the differences (remember that these differences are called *deviations* or *residuals*) between the observed values and the mean of the relevant variable. The first thing to notice about Figure 8.2 is that there is a similar pattern of deviations for both variables. For the first three participants the observed values are below the mean for both variables, for the last two people the observed values are above the mean for both variables. This pattern is indicative of a potential relationship between the two variables (because it seems that if a person's score is below the mean for one variable then their score for the other will also be below the mean).

So, how do we calculate the exact similarity between the patterns of differences of the two variables displayed in Figure 8.2? One possibility is to calculate the total amount of deviation, but we would have the same problem as in the single-variable case: the positive and negative deviations would cancel out (see Section 1.7.5). Also, by adding the deviations, we would gain little insight into the *relationship* between the variables. In the single-variable case, we squared the deviations to eliminate the problem of positive and negative deviations cancelling each other out. When there are two variables, rather than squaring each deviation, we can multiply the deviation for one variable by the corresponding deviation for the second variable. If both deviations are positive or negative then this will give us a positive value (indicative of the deviations being in the same direction), but if one deviation is positive and one negative then the resulting product will be negative (indicative of the deviations being opposite in direction). When we multiply the deviations of one variable by the corresponding deviations of a second variable, we get the **cross-product deviations**. As with the variance, if we

want an average value of the combined deviations for the two variables, we divide by the number of observations (we actually divide by $N-1$ for reasons explained in Jane Superbrain Box 2.5). This averaged sum of combined deviations is known as the **covariance**. We can write the covariance in equation form:

$$\text{covariance}\left(\text{cov}_{xy}\right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1} \tag{8.5}$$

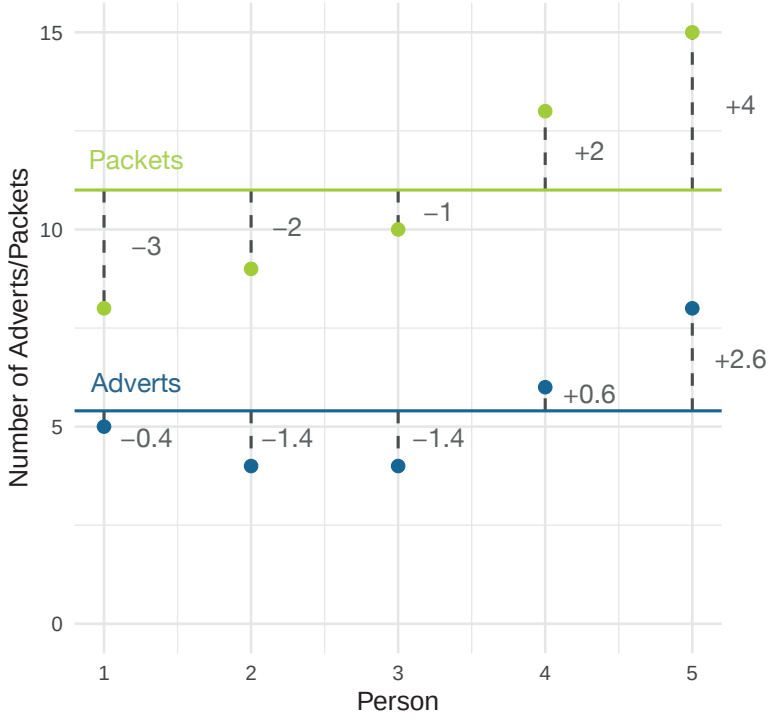


Figure 8.2 Graphical display of the differences between the observed data and the means of two variables

Notice that the equation is the same as the equation for variance (Eq. 1.7), except that instead of squaring the deviances, we multiply them by the corresponding deviance of the second variable. For the data in Table 8.1 and Figure 8.2 we get a value of 4.25:

$$\begin{aligned}\text{cov}_{xy} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1} \\ &= \frac{(-0.4)(-3)+(-1.4)(-2)+(-1.4)(-1)+(0.6)(2)+(2.6)(4)}{N-1} \\ &= \frac{1.2+2.8+1.4+1.2+10.4}{4} = \frac{17}{4} = 4.25\end{aligned} \tag{8.6}$$

A positive covariance indicates that as one variable deviates from the mean, the other variable deviates in the same direction. On the other hand, a negative covariance indicates that as one variable deviates from the mean (e.g., increases), the other deviates from the mean in the opposite direction (e.g., decreases). However, the covariance depends upon the scales of measurement used: it is not a

standardized measure. For example, if we use the data above and assume that they represented two variables measured in miles then the covariance is 4.25 square miles (as calculated above). If we convert these data into kilometres (by multiplying all values by 1.609) and calculate the covariance then we find that it increases to 11 square kilometres. This dependence on the scale of measurement is a problem because we cannot compare covariances in an objective way – we cannot say whether a covariance is particularly large or small relative to another data set unless both data sets were measured in the same units.

8.2.2 Standardization and the correlation coefficient Ⓐ

To overcome the problem of dependence on the measurement scale, we need to convert the covariance into a standard set of units. This process is known as **standardization**. We need a unit of measurement into which any variable can be converted, and typically we use the *standard deviation*. We came across this measure in Section 1.7.5 and saw that, like the variance, it is a measure of the average deviation from the mean. If we divide any distance from the mean by the standard deviation, it gives us that distance in standard deviation units. For example, for the data in Table 8.1, the standard deviation for the number of packets bought is approximately 3.0 (the exact value is 2.92). In Figure 8.2 we can see that the observed value for participant 1 was 3 packets less than the mean (so there was an error of –3 packets of sweets). If we divide this deviation, –3, by the standard deviation, which is approximately 3, then we get a value of –1. This tells us that the difference between participant 1's score and the mean was –1 standard deviation. In this way we can express the deviation from the mean for a participant in standard units by dividing the observed deviation by the standard deviation.

It follows from this logic that if we want to express the covariance in a standard unit of measurement we can divide by the standard deviation. However, there are two variables and, hence, two standard deviations. When we calculate the covariance we calculate two deviations (one for each variable) and multiply them. We do the same for the standard deviations: we multiply them and divide the covariance by the product of this multiplication. The standardized covariance is known as a *correlation coefficient* and is defined as follows:

$$r = \frac{\text{cov}_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y} \tag{8.7}$$

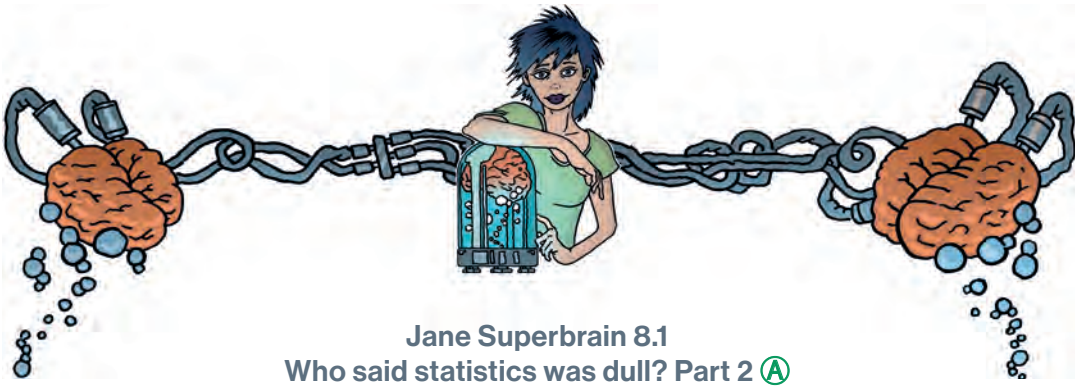
in which s_x is the standard deviation of the first variable and s_y is the standard deviation of the second variable (all other letters are the same as in the equation defining covariance). This coefficient, the *Pearson product-moment correlation coefficient* or **Pearson correlation coefficient**, r, was invented by Karl Pearson, with Florence Nightingale David[2] doing a lot of the hard maths to derive distributions for it (see Figure 8.3 and Jane Superbrain Box 8.1).[3] If we look back at Table 8.1 we see that the standard deviation for the number of adverts watched (s_x) was 1.673, and for the number of packets of crisps bought (s_y) was 2.915. If we multiply these standard deviations we get 1.673 × 2.915 = 4.877. Now all we need to do is take the covariance, which we calculated a few pages ago as being 4.250, and divide by these multiplied standard deviations. This gives us r = 4.250/4.877 = 0.871.

2 Not to be confused with the Florence Nightingale in Chapter 5 whom she was named after.

3 Pearson's product-moment correlation coefficient is denoted by r but, just to confuse us, when we square r (as in Section 8.4.2) an upper-case R is typically used.



Figure 8.3 Karl Pearson and Florence Nightingale David



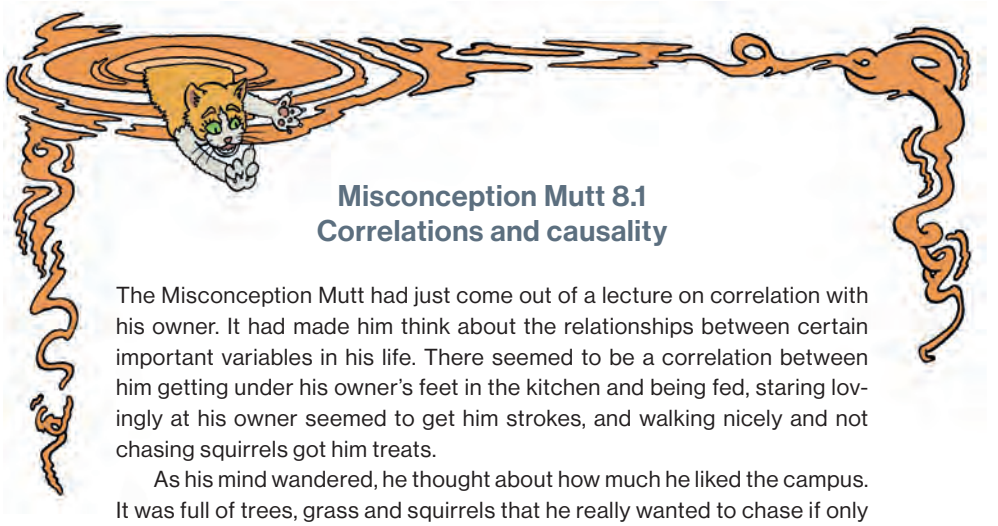
Jane Superbrain 8.1
Who said statistics was dull? Part 2 Ⓐ

We saw in Jane Superbrain Box 2.6 that Fisher and Neyman had a feud over their different views of hypothesis testing. Fisher seemed to believe that if you're going to feud with one of your prominent peers, you may as well feud with them all, and he didn't care much for Karl Pearson either. This wasn't a great career move on Fisher's part because the more senior Pearson wielded huge influence through his journal *Biometrika*. The feud began when Pearson published a paper of Fisher's in his journal but belittled it in his editorial. Two years later, Pearson's group published work following on from Fisher's paper without consulting him. Fisher turned down a job in Pearson's group and published 'improvements' on Pearson's ideas that were greeted by Pearson like a fish would greet a cat as she entered his tank. For his part, Pearson wrote in his own journal about apparent errors made by Fisher (Barnard, 1963; Field, 2005; Savage, 1976). Statistics was never dull back then.



By standardizing the covariance we end up with a value that has to lie between –1 and +1 (if you find a correlation coefficient less than –1 or more than +1 you can be sure that something has gone hideously wrong). We saw in Section 3.7.4 that a coefficient of +1 indicates that the two variables are perfectly positively correlated: as one variable increases, the other increases by a proportionate amount. This does not mean that the change in one variable *causes* the other to change, only that their changes coincide (Misconception Mutt 8.1). Conversely, a coefficient of –1 indicates a perfect negative relationship: if one variable increases, the other decreases by a proportionate amount. A coefficient of zero indicates no linear relationship at all. We also saw that because the correlation coefficient is a standardized measure of an observed effect, it is a commonly used effect size measure and that values of ±0.1 represent a small effect, ±0.3 is a medium effect and ±0.5 is a large effect (although we should interpret the effect size within the context of the research literature and not use these canned effect sizes).

We have just described a **bivariate correlation**, which is a correlation between two variables. Later in the chapter we'll look at variations on the correlation that adjust for one or more additional variables.



Misconception Mutt 8.1
Correlations and causality

The Misconception Mutt had just come out of a lecture on correlation with his owner. It had made him think about the relationships between certain important variables in his life. There seemed to be a correlation between him getting under his owner's feet in the kitchen and being fed, staring lovingly at his owner seemed to get him strokes, and walking nicely and not chasing squirrels got him treats.

As his mind wandered, he thought about how much he liked the campus. It was full of trees, grass and squirrels that he really wanted to chase if only it didn't stop the supply of treats. He was contemplating his conundrum. 'I really like treats,' he thought, 'and I really like chasing squirrels. But chasing squirrels is negatively correlated with treats, so if I chase the squirrel it will cause the number of treats to go down.'

He was hungry, so he started walking extra nicely. A squirrel darted up to him and jumped on his head as if trying to goad him into a chase. It started to grow, becoming heavier and more ginger until the grinning Correcting Cat stared down into the mutt's eyes from on top of his head.

'Correlation coefficients give no indication of the direction of causality,' he said to the dog. 'You might think that chasing squirrels causes fewer treats, but statistically speaking there's no reason why fewer treats isn't causing you to chase squirrels.'

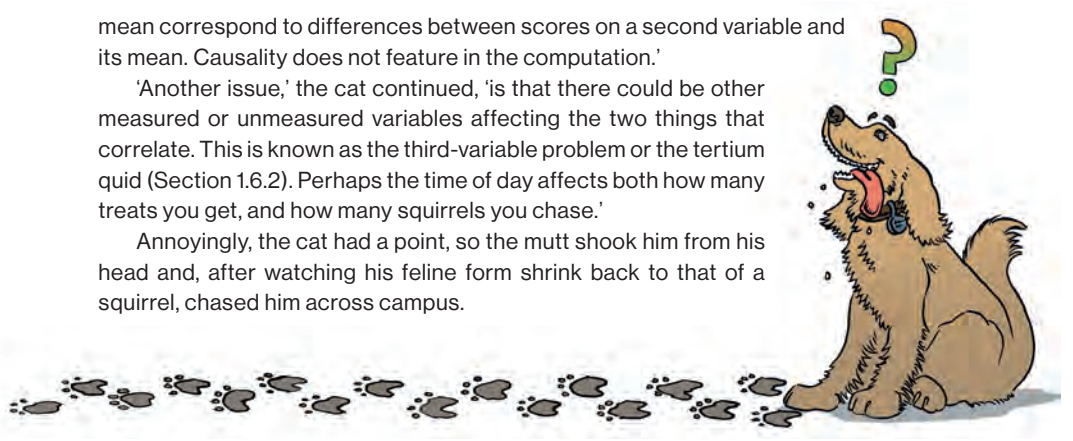
'That's ridiculous,' replied the mutt.

'It may be less intuitive to think of fewer treats causing you to chase squirrels more, but statistically speaking the correlation between those variables provides no information about cause: it is purely a measure of the degree to which variables covary. Think back to the lecture: the correlation measures whether differences between scores on one variable and its

mean correspond to differences between scores on a second variable and its mean. Causality does not feature in the computation.'

'Another issue,' the cat continued, 'is that there could be other measured or unmeasured variables affecting the two things that correlate. This is known as the third-variable problem or the tertium quid (Section 1.6.2). Perhaps the time of day affects both how many treats you get, and how many squirrels you chase.'

Annoyingly, the cat had a point, so the mutt shook him from his head and, after watching his feline form shrink back to that of a squirrel, chased him across campus.



8.2.3 The significance of the correlation coefficient ©

Although we can interpret the size of a correlation coefficient directly (Section 3.7.4), we have seen in Chapter 1 that scientists like to test hypotheses using probabilities. In the case of a correlation coefficient we can test the hypothesis that the correlation is different from zero (i.e., different from 'no relationship'). If we find that our observed coefficient was very unlikely to be (at least) as big as it would be if there were no effect in the population then we might gain confidence that the relationship that we have observed is statistically meaningful.

There are two ways that we can go about testing this hypothesis. The first is to use the trusty *z*-scores that keep cropping up in this book. As we have seen, *z*-scores are useful because we know the probability of a given value of *z* occurring if the distribution from which it comes is normal. There is one problem with Pearson's *r*, which is that it is known to have a sampling distribution that is not normally distributed. This would be a bit of a nuisance, except that thanks to our friend Fisher (1921) *r* can be adjusted so that its sampling distribution *is* normal:

$$z_r = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \tag{8.8}$$

The resulting z_r has a standard error given by:

$$SE_{z_r} = \frac{1}{\sqrt{N-3}} \tag{8.9}$$

For our advert example, our $r = 0.871$ becomes 1.337 with a standard error of 0.707.

We can transform this adjusted *r* into a *z*-score in the usual way. If we want a *z*-score that represents the size of the correlation relative to a particular value, then we compute a *z*-score by subtracting the value that we want to test against and then dividing by the standard error. Normally we want to see whether the correlation is different from 0, in which case we subtract 0 from the observed value of z_r and divide by the standard error, which is the same as dividing z_r by its standard error:

$$z = \frac{z_r}{SE_{z_r}} \tag{8.10}$$

For our advert data this gives us 1.337/0.707 = 1.891. We can look up this value of *z*(1.89) in the table for the normal distribution in the Appendix and get the one-tailed probability from the column labelled 'Smaller Portion' (think back to Section 1.7.6). In this case the value is 0.02938. To get the two-tailed probability we multiply this value by 2, which gives us 0.05876. As such the correlation is not significant, because $p > 0.05$.

In fact, the hypothesis that the correlation coefficient is different from 0 is usually (SPSS, for example, does this) tested not using a *z*-score, but using a different test statistic called a *t*-statistic with $N-2$ degrees of freedom. This statistic can be obtained directly from *r*:

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{8.11}$$

So you might wonder why I told you about *z*-scores. Partly it was to keep the discussion framed in concepts with which you are already familiar (we don't encounter the *t*-statistic properly for a few chapters), but also it is useful background information for the next section.

8.2.4 Confidence intervals for *r* Ⓒ

We saw in Chapter 1 that 95% confidence intervals tell us about the likely value (in this case of the correlation) in the population (assuming that your sample is one of the 95% for which the confidence interval contains the true value). To compute confidence intervals for *r*, we take advantage of what we learnt in the previous section about converting *r* to z_r (to make the sampling distribution normal), and use the associated standard errors. By using z_r we can construct a confidence interval in the usual way. For example, a 95% confidence interval is calculated (see Eq. 2.23) as:

$$\begin{aligned}\text{lower boundary of confidence interval} &= \bar{X} - (1.96 \times SE)\\ \text{upper boundary of confidence interval} &= \bar{X} + (1.96 \times SE)\end{aligned} \tag{8.12}$$

In the case of our transformed correlation coefficients, these equations become:

$$\begin{aligned}\text{lower boundary of confidence interval} &= z_r - \left(1.96 \times \text{SE}_{z_r}\right)\\ \text{upper boundary of confidence interval} &= z_r + \left(1.96 \times \text{SE}_{z_r}\right)\end{aligned} \tag{8.13}$$

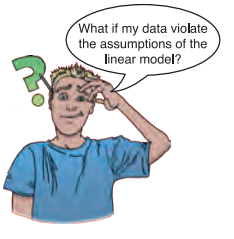
For our advert data we get 1.337 – (1.96 × 0.707) = –0.049, and 1.337 + (1.96 × 0.707) = 2.723. Remember that these values are in the z_r metric, but we can convert back to a correlation coefficient using:

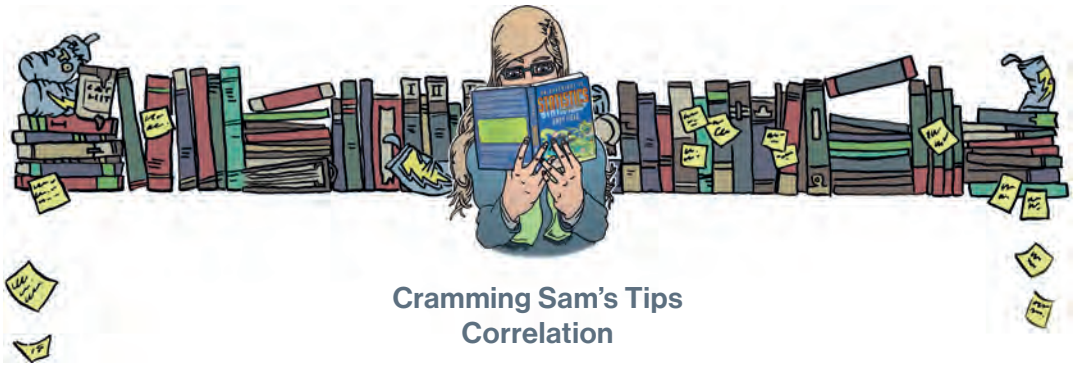
$$r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1} \tag{8.14}$$

This gives us an upper bound of *r* = 0.991 and a lower bound of –0.049 (because this value is so close to zero the transformation to *z* has no impact).

8.2.5 Nonparametric correlation coefficients Ⓐ

The Pearson correlation coefficient requires the same assumptions as the GLM (Chapter 9). In particular, the variables should be continuous and measured at an interval or ratio level. In addition, residuals (tested by looking at the distributions of the variables themselves) should follow a normal distribution. Pearson's *r* can also be affected by extreme scores, influential cases and outliers. If you have concerns about these assumptions, then fear not because **Spearman's correlation coefficient**, denoted by r_s or ρ, is a non-parametric statistic that is useful to minimize the effects of extreme scores or the effects of violations of the assumptions discussed in Chapter 6. Charles Spearman is pictured in Figure 8.4. Because the statistic is sometimes denoted by the Greek letter rho (ρ) this correlation coefficient is also referred to as Spearman's rho (pronounced 'row', as in 'row your boat gently down the stream'). Spearman's statistic is computed by first ranking the data (see Section 7.4.1), and then applying Pearson's equation (Eq. 8.7) to those ranks (Spearman, 1910).





Cramming Sam's Tips Correlation

- A crude measure of the relationship between variables is the covariance.
- If we standardize this value we get Pearson's correlation coefficient, *r*.
- The correlation coefficient has to lie between −1 and +1.
- A coefficient of +1 indicates a perfect positive relationship, a coefficient of −1 indicates a perfect negative relationship, and a coefficient of 0 indicates no linear relationship.
- The correlation coefficient is a commonly used measure of the size of an effect: values of ±0.1 represent a small effect, ±0.3 is a medium effect and ±0.5 is a large effect. However, interpret the size of correlation within the context of the research you've done rather than blindly following these benchmarks.



Kendall's tau, denoted by τ, is another non-parametric correlation and it should be used rather than Spearman's coefficient when you have a small data set with a large number of tied ranks. Although Spearman's statistic is the more popular of the two coefficients, there is much to suggest that Kendall's statistic is a better estimate of the correlation in the population (see Howell, 2012). As such, we can draw more accurate generalizations from Kendall's statistic than from Spearman's.

One issue that arises with Spearman's r_s is around confidence intervals. Because Pearson's *r* can be transformed such that it has a normal sampling distribution (see earlier) we can use normal theory to construct confidence intervals around *r*. This is not the case for Spearman's r_s. The variance of r_s changes as a function of the population correlation and so, therefore, does its sampling distribution (which we need to construct confidence intervals). Various solutions have been proposed to estimate the variance of r_s. Fielleret al. (1957) proposed a variance estimate derived from simulation and based on a *z*-transformed r_s. It is sometimes referred to as the FCI method. However, they concluded that the FCI method would be

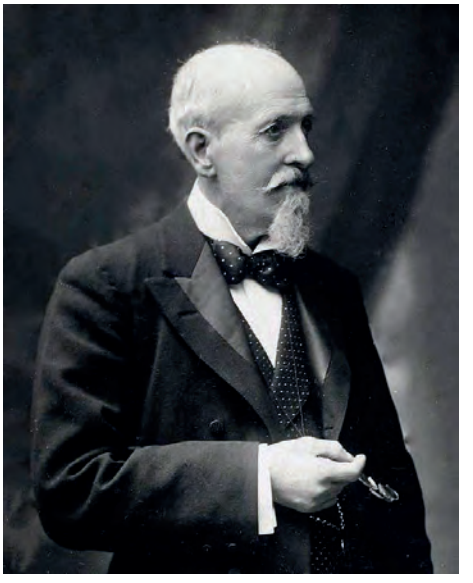


Figure 8.4 Charles Spearman, ranking furiously

inaccurate if the sample size was not between 10 and 50 observations and if the population correlation was above about 0.8 (Caruso & Cliff, 1997). The Caruso and Cliff (1997) adjusted FCI (AFCI) method (also derived through simulation) tweaks the FCI and adds an adjustment for sample size bias. Bonett and Wright (2000) suggest a third variance estimate (again determined empirically) and there are others too. Based on a simulation study, Bonett and Wright (2000) concluded that confidence intervals based on the three estimators contain the value of population parameter the correct amount of the time when the population correlation is (1) below 0.7 when using the FCI and AFCI estimators (above this threshold confidence intervals were too liberal), (2) below 0.95 using their own estimator. The take-home message is that unless you have very high correlations it won't make much difference which estimator you choose, but if the population correlation is high (above 0.7) use the Bonett and Wright (2000) method.

8.3 Data entry for correlation analysis Ⓐ

When looking at relationships between variables each variable is entered in a separate column in the data editor. So, for each variable you have measured, create a variable with an appropriate name, and enter a participant's scores across one row. If you have categorical variables (such as the participant's sex) these can also be entered in a column (remember to define appropriate value labels). For example, if we wanted to calculate the correlation between the two variables in Table 8.1 we would enter these data as in Figure 8.5; each variable is entered in a separate column, and each row represents a single individual's data (so the first consumer saw 5 adverts and bought 8 packets).

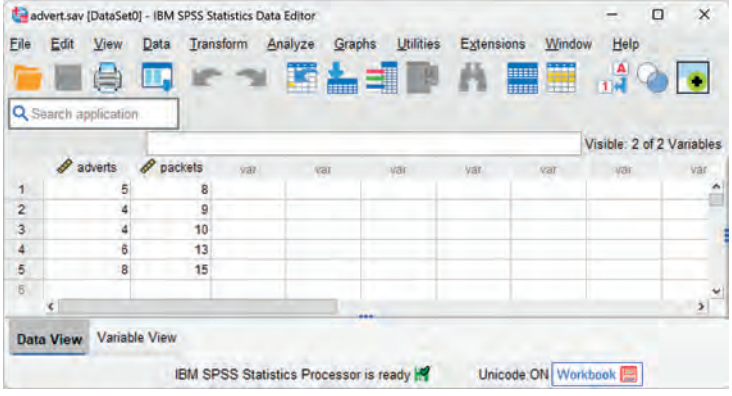


Figure 8.5 Data entry for correlation



Enter the advert data and use the chart editor to produce a scatterplot (number of packets bought on the *y*-axis, and adverts watched on the *x*-axis) of the data.

8.4 Bivariate correlation Ⓐ

Figure 8.6 shows a general procedure to follow when computing a bivariate correlation coefficient. First, check for sources of bias as outlined in Chapter 6. The two most important ones in this context

are linearity and normality. Remember that we're fitting linear model to the data, so if the relationship between variables is not linear then this model is invalid (a transformation might help to make the relationship linear). To meet this requirement, the outcome variable needs to be measured at the interval or ratio level (see Section 1.5.2), as does the predictor variable (one exception is that a predictor variable can be a categorical variable with only two categories – we'll get onto this in Section 8.4.5). As far as normality is concerned, we care about this only if we want confidence intervals or significance tests and if the sample size is small (Section 6.8.2).

If the data have outliers, are not normal (and the sample is small) or your variables are measured at the ordinal level then you can use Spearman's rho (Section 8.4.3) or Kendall's tau (Section 8.4.4), which are versions of the correlation coefficient applied to ranked data (just like the tests in the previous chapter). Ranking the data reduces the impact of outliers. Furthermore, given that normality matters only for inferring significance and computing confidence intervals, we could use a bootstrap to compute the confidence interval, and then we don't need to worry about the distribution.

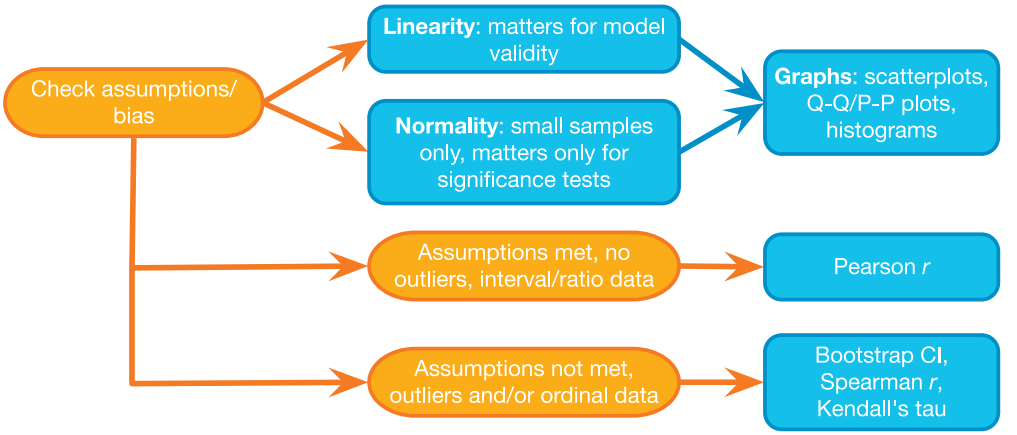
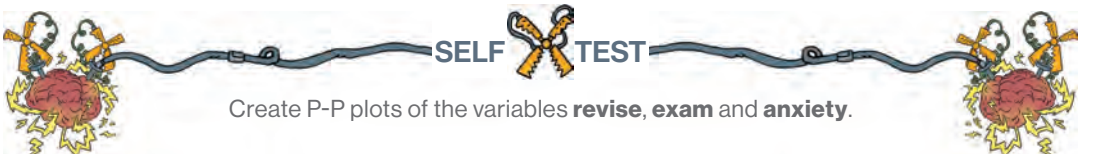


Figure 8.6 The general process for conducting correlation analysis

In Chapter 1 we looked at an example relating to exam anxiety: a psychologist was interested in the effects of exam stress and revision on exam performance. She had devised and validated a questionnaire to assess state anxiety relating to exams (called the Exam Anxiety Questionnaire, or EAQ). This scale produced a measure of anxiety scored out of 100. Anxiety was measured before an exam, and the percentage mark of each student on the exam was used to assess the exam performance. She also measured the number of hours spent revising. These data are in **exam_anxiety.sav**. For this study the researcher set an alpha level of 0.005 in line with Benjamin et al. (2018). We have already created scatterplots for these data (Section 5.8) so we don't need to do that again; however, we will look at the distributions of the three main variables.



It's clear from the P-P plots in Figure 8.7 that exam performance is the most normally distributed of the variables (the dots hover close to the line), but for exam revision and exam anxiety there is evidence of skew (the dots snake around the diagonal line). This skew is a problem if we want to do significance tests or look at confidence intervals. The sample contains 103 observations, which is reasonably large, and possibly large enough for the central limit theorem to relieve of us of concerns

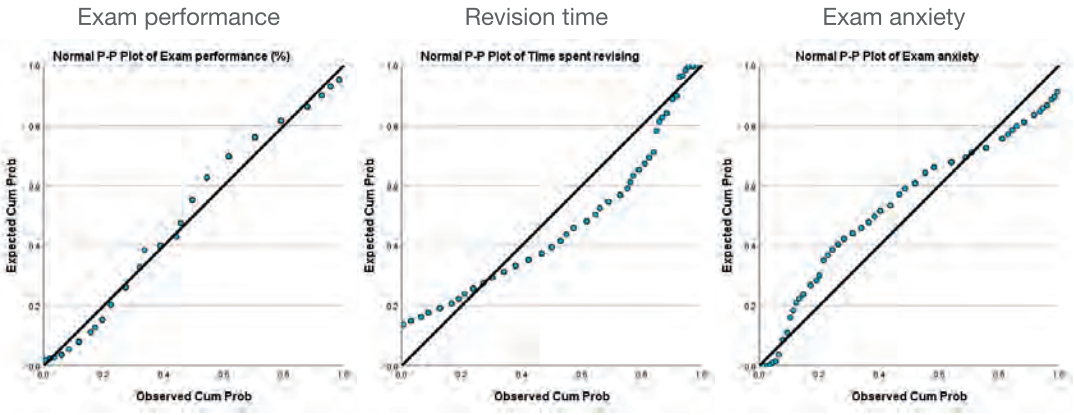
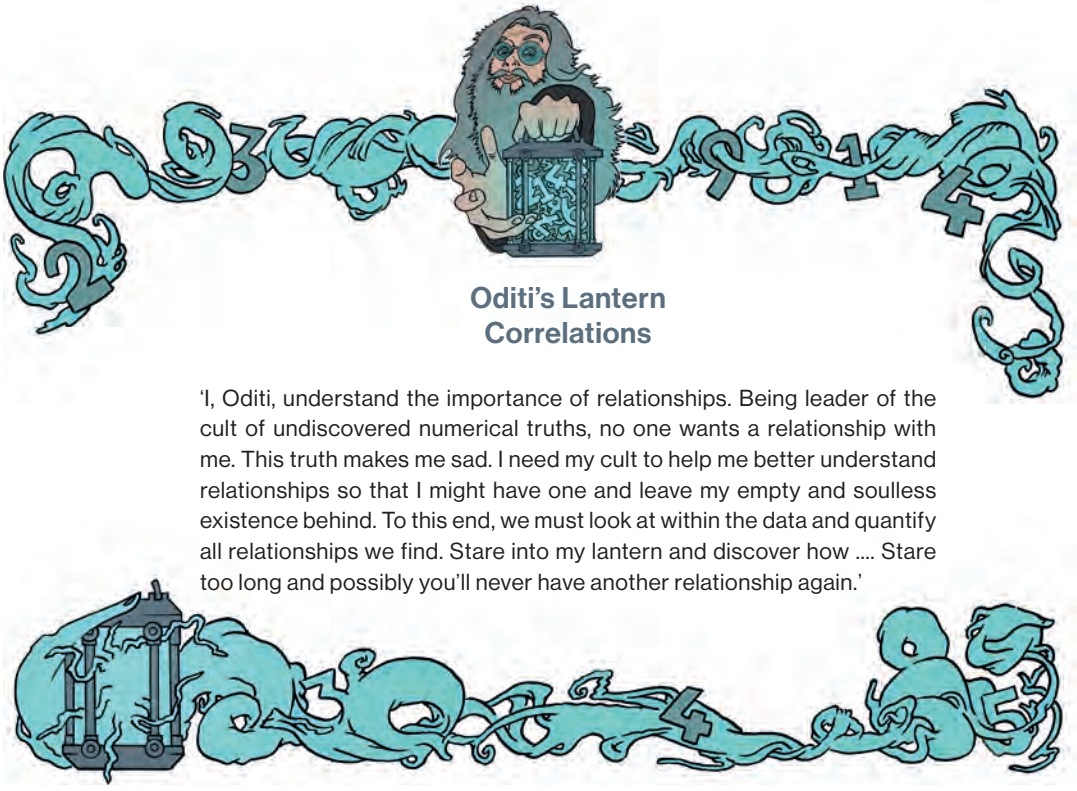


Figure 8.7 P-P plots for the exam anxiety variables

about normality. However, it would be advisable to use a bootstrap to get robust confidence intervals. We might also consider using a rank-based method to compute the correlation coefficient itself.



Oditi's Lantern
Correlations

'I, Oditi, understand the importance of relationships. Being leader of the cult of undiscovered numerical truths, no one wants a relationship with me. This truth makes me sad. I need my cult to help me better understand relationships so that I might have one and leave my empty and soulless existence behind. To this end, we must look at within the data and quantify all relationships we find. Stare into my lantern and discover how .... Stare too long and possibly you'll never have another relationship again.'

8.4.1 General procedure for correlations using SPSS Ⓐ

You conduct bivariate correlation using the dialog box accessed through *Analyze* > *Correlate* > Bivariate... . In this dialog box (Figure 8.8) the variables in the data editor are listed on the left-hand side and there is an empty box labelled *Variables* on the right-hand side. You can select any variables from the list using the mouse and transfer them to the *Variables* box by dragging them there or clicking

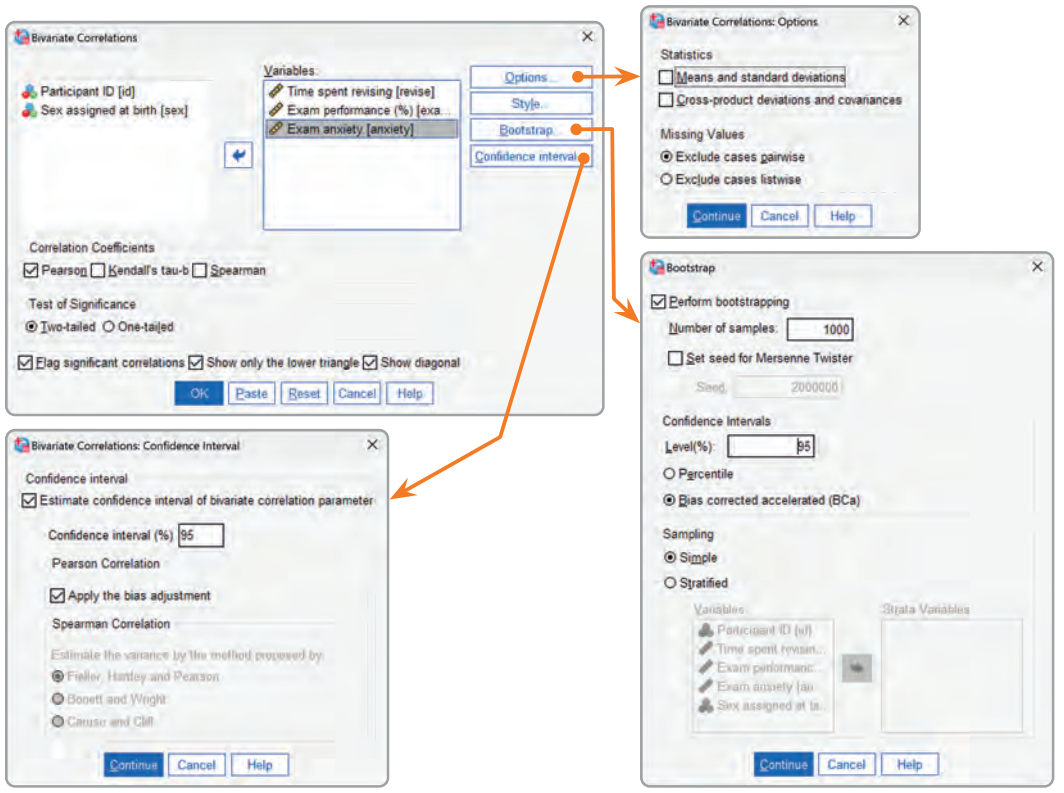


Figure 8.8 Dialog box for conducting a bivariate correlation

on ▸. SPSS creates a table (called a *correlation matrix*) of correlation coefficients for all combinations of specified variables. For our current example, select the variables **Exam performance**, **Exam anxiety** and **Time spent revising** and transfer them to the *Variables* box. Having selected the variables of interest, you can choose between three correlation coefficients: the default is Pearson's product-moment correlation coefficient (☑ Pearson), but you can also select Spearman's r_s (☑ Spearman) and Kendall's tau (☑ Kendall's tau-b) – we'll explore the differences in due course. You can choose more than one of these correlation coefficients if you like. You can also specify whether the test is one- or two-tailed. In Section 2.10.6 I advised against one-tailed tests, so I would leave the default of ◉ Two-tailed, but if you don't like my advice then you might select ◉ One-tailed if your hypothesis is directional (e.g., 'the more anxious someone is about an exam, the worse their mark will be') and ◉ Two-tailed if it is non-directional (i.e., 'I'm not sure whether as exam anxiety increases exam marks will go up or down').

By default SPSS will produce a table of information that follows the format of a correlation matrix. Figure 8.9 shows the layout of a correlation matrix – basically it is a grid of values with i rows and j columns, and each row and column represents a variable. The values in the grid represent the correlation between the variables that intersect at that point. In general r_{ij} in the grid is the correlation between the variables represented by row i and column j, where i and j can be replaced by any number. For example, for r_{12} $i = 1$ and $j = 2$ so it is the correlation between the variables represented by row 1 and column 2. The diagonal values in this grid (r_{11}, r_{22},…, r_{ij}) are the correlations between each variable and itself (which will be a value of 1 because a variable is perfectly correlated with itself). To make things more concrete, I've included the layout for the exam anxiety data where there are three variables and have omitted the diagonal values. Row 1 and column 1 both represent exam performance, row 2 and column 2 both represent exam anxiety and row 3 and column 3 both represent exam revision. Therefore, the value in the second row and first column (r_{21}) is the correlation between exam

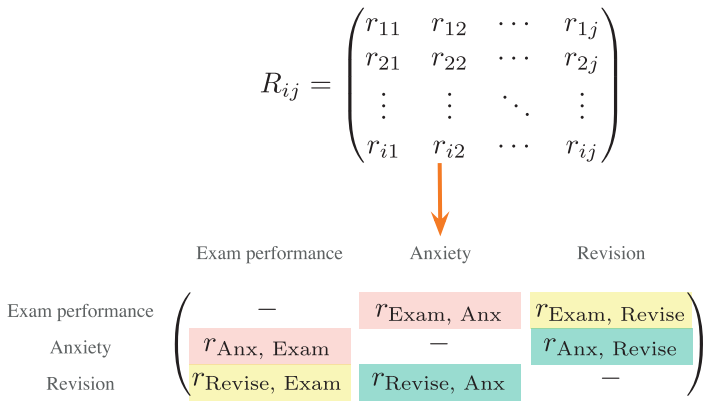


Figure 8.9 Example of a correlation matrix

performance and anxiety ($r_{\text{Anx, Exam}}$) and the value in the third row and second column (r_{32}) is the correlation between exam revision and anxiety ($r_{\text{Revise, Anx}}$). Values with the same background colour are the same, for example, $r_{32} = r_{23}$ because $r_{\text{Revise, Anx}} = r_{\text{Anx, Revise}}$ (the correlation between revision and anxiety is the same regardless of which way around we denote the variables). This colour coding shows that a correlation matrix contains redundant information. The values below the diagonal are the same as those above. Why am I telling you this? It's because the SPSS output will, by default, contain this redundant information and we can make our life a bit easier by selecting ☑ Show only the lower triangle in the main dialog box (Figure 8.8), which will hide the duplicate information. It will basically show us the values below the diagonal but not the values above it.

Clicking Style... opens the dialog box in Figure 8.10, which allows you to apply formatting to the table of correlations in the output. Let's jump down this wormhole for a bit. The drop-down menu in the column called *Value* is for selecting types of cells in the output: you could select the cells containing the correlation coefficients (as I have in the figure), the cells containing the sample sizes (*N*), the means, the significance values, or all the cells in the table. Clicking in the column labelled *Condition* opens a dialog box for setting a condition for formatting. In the figure I have set a condition of formatting cells that contain an absolute value greater than or equal to 0.5. Doing so will apply the formatting only to cells containing correlation coefficients greater than 0.5 or smaller than –0.5. Clicking in the column labelled *Format* opens a dialog box for defining what formatting you wish to apply. In the figure I have chosen to change the background colour of the cells to yellow, but you can do other things such as change the text colour, or make it bold or italic and so on. The settings at the top of Figure 8.10 set up a rule in which I have (1) chosen to format only the cells containing the correlation coefficients; and (2) asked that cells containing a value greater than 0.5 or smaller than –0.5 have a yellow background. The effect will be that these cells are highlighted in the output: I will very quickly be able to see which variables were strongly correlated because their cells will be highlighted. For more on styling see SPSS Tip 8.1.

Going back to the main dialog box (Figure 8.8), clicking Options... opens a dialog box with two *Statistics* options and two options for missing values. The *Statistics* options are enabled only when Pearson's correlation is selected (otherwise they appear 'greyed out'). These two options are meaningful only for interval data, which the Pearson correlation requires, so it makes sense that these options are deactivated if the Pearson correlation has not been selected. Selecting the tick-box labelled *Means and standard deviations* produces the mean and standard deviation of the variables selected for analysis. Selecting the tick-box labelled *Cross-product deviations and covariances* produces the values of these statistics for the variables in the analysis. The cross-product deviations are the values of the numerator (top half) of Eq. 8.5. The covariances between variables, are what you would get from applying Eq. 8.5 to your variables. In other words, the covariance values are the cross-product deviations divided by $N-1$ and represent the unstandardized correlation coefficient. In most instances you won't need these statistics but they occasionally come in handy (see Oliver Twisted). At this point we need to decide how to deal with missing values (look back to SPSS Tip 6.1).

To get confidence intervals for Pearson or Spearman correlations click Confidence interval... and select ☑ Estimate confidence interval of bivariate correlation parameter. If you selected ☑ Pearson in the main dialog box then the area labelled *Pearson Correlation* will be active and you should select ☑ Apply the bias adjustment. If you selected ☑ Spearman in the main dialog box then the area labelled *Spearman Correlation* will be active and you can select between the three variance estimation methods described in Section 8.2.5. On balance I'd select ◉ Bonett and Wright. To get bootstrapped confidence intervals for the correlation coefficient click Bootstrap... in the main dialog box. We discussed this dialog box in Section 6.10.8; to recap, select ☑ Perform bootstrapping to activate bootstrapping for the correlation coefficient, and to get a 95% confidence interval click ◉ Percentile or ◉ Bias corrected accelerated (BCa). In this chapter we'll generally ask for a bias corrected (BCa) confidence interval to get some practice at interpreting them.

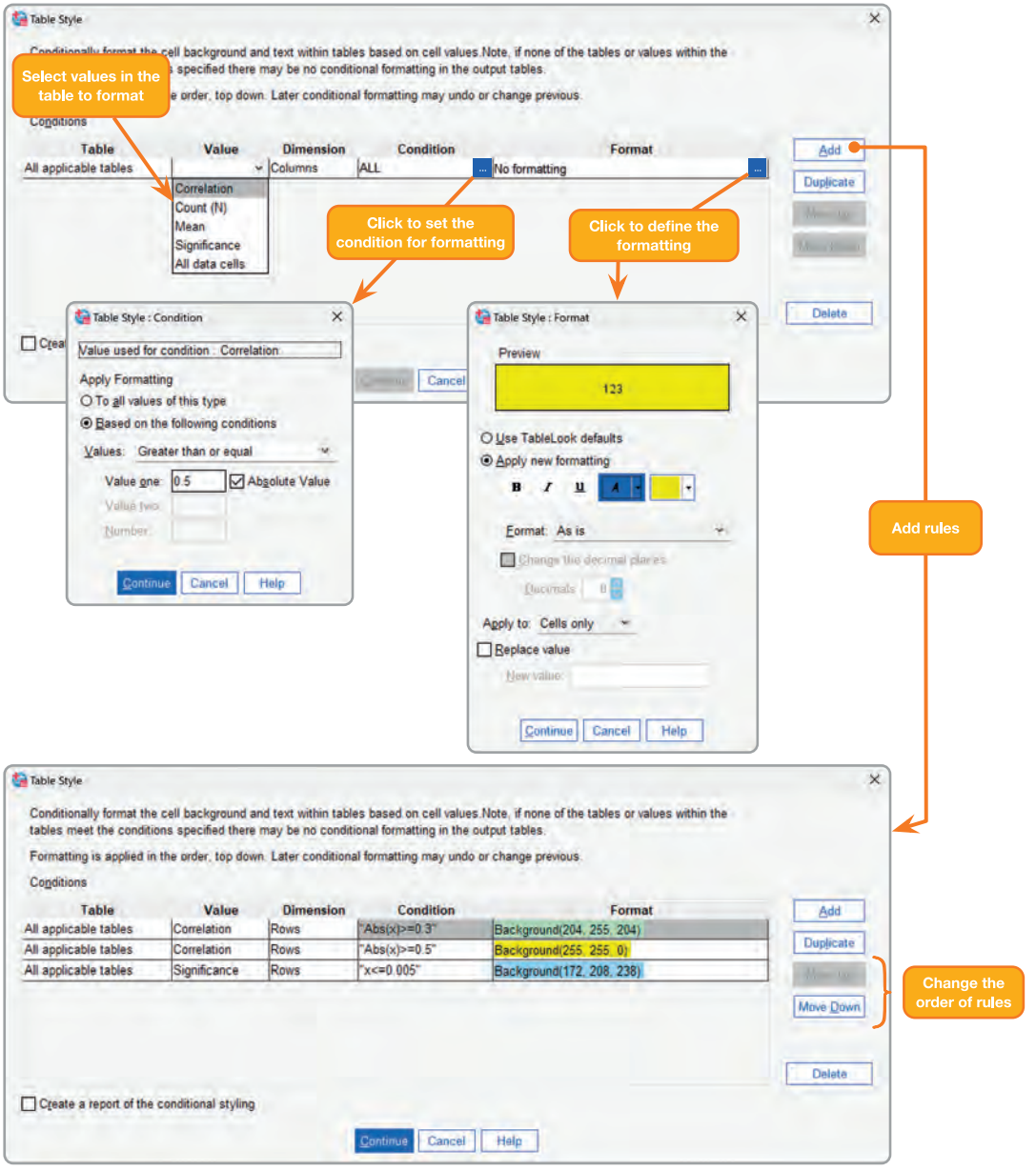
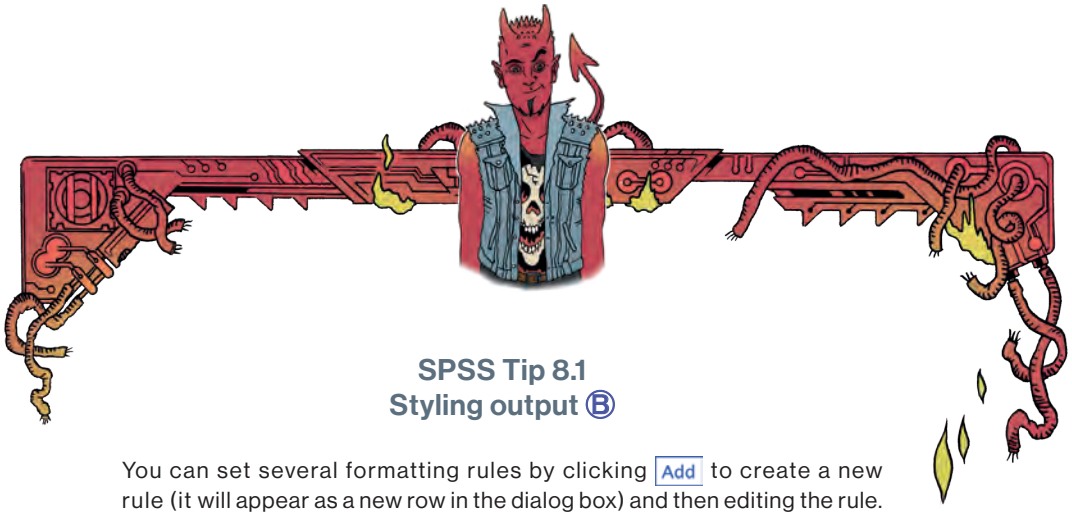


Figure 8.10 The *Table Style* dialog box



Oliver Twisted
Please, Sir, can I have some more ... options?

Oliver is so excited to get onto analysing his data that he doesn't want me to spend pages waffling on about options that you will probably never use. 'Stop writing, you waffling fool,' he says. 'I want to analyse my data.' Well, he's got a point. If you want to find out more about what the Options... do in correlation, then look at www.discoverspss.com.



SPSS Tip 8.1
Styling output Ⓑ

You can set several formatting rules by clicking Add to create a new rule (it will appear as a new row in the dialog box) and then editing the rule. For example, the bottom of Figure 8.10 shows a completed dialog in which I have set up three rules.

- Rule 1: Format cells containing the correlation coefficients so that the background is green when the value is greater than 0.3 or less than −0.3.
- Rule 2: Format cells containing the correlation coefficients so that the background is yellow when the value is greater than 0.5 or less than −0.5.
- Rule 3: Format cells containing significance values so that the background is blue when the value is less than 0.005 (Benjamin et al., 2018).

These rules are processed in order from the top row to the bottom, so we need to think about the order of rules that gives us the desired result. In Figure 8.10 I have set the rules so that, using rule 1, SPSS colours cells green if they contain correlation coefficients greater than 0.3 or smaller than −0.3. This rule will include any cells with correlation coefficients greater than 0.5 or smaller than −0.5. Next, rule 2 colours cells yellow if they contain correlation coefficients greater than 0.5 or smaller than −0.5. This rule means that some of the cells coloured green by rule 1 will now be coloured yellow (the ones containing correlation coefficients greater than 0.5 or smaller than −0.5). These rules result in a table where green cells indicate correlation coefficients between 0.3 and 0.5 (and −0.3 and −0.5), and yellow cells indicate correlation coefficients greater than 0.5 or smaller than −0.5, which is what I was aiming for.

If we swap the order of rules 1 and 2 we'd get a different outcome. Rule 1 will colour cells yellow if they contain correlation coefficients greater than 0.5 or smaller than −0.5. Rule 2 will then colour cells green if they contain correlation coefficients greater than 0.3 or smaller than −0.3, which will include all the cells coloured yellow by rule 1 (because these cells contain correlation coefficients greater than 0.3 or smaller than −0.3). The resulting table will have no yellow cells because rule 2 has overridden what was done by rule 1. Once you have entered your rules you can change their order by selecting a rule and clicking one of the buttons to move it up or down the list.

Incidentally, my third rule flags significant correlations in the table; I've set the criterion at $p \leq 0.005$ (Benjamin et al., 2018). I don't particularly recommend doing this because it's the sort of thing that will encourage you to blindly apply a significance criterion, but it can be useful if you're the sort of person who has internalized my advice to always interpret significance within the wider context of the sample size and effect size. Finally, you can select a row and click Duplicate to create a new rule based on one you selected and you can delete a rule by selecting the row representing the rule and clicking Delete.



8.4.2 Pearson's correlation coefficient using SPSS Ⓐ

To obtain Pearson's correlation coefficient run through the general procedure that we've just looked at (Section 8.4.1), selecting ☑ Pearson (the default) and then clicking OK (Figure 8.8). Assuming you selected ☑ Show only the lower triangle in the main dialog box (Figure 8.8) and applied the same styling rules as I did, you'll see Output 8.1, which is not as bewildering as it looks. The first main row (which is subdivided into two) tells us about time spent revising, but is empty (apart from telling us the sample size) because we asked only to display the lower part of the table. The next main row contains information about exam performance and is subdivided into seven other rows the first column giving us information about the relationship between time revising and exam performance. The first row

(the green cell) tells us the correlation coefficient between exam performance and the time spent revising ($r = 0.397$), underneath (the blue cell) is the significance value of this correlation ($p < .001$), and underneath that is the sample size of this correlation ($N = 103$). The final two rows give us the lower (0.245) and upper (0.517) boundaries of the 95% bootstrap confidence interval. We'll interpret these in due course. The third main row contains information about exam anxiety and is subdivided into seven other rows. The values in the first column relate to the relationship between exam anxiety and time revising and we can see that there was a strong negative association, $r = –0.709$ (yellow cell), that is highly significant (blue cell). The values in the second column relate to the relationship between exam anxiety and exam performance and also show a strong negative association, $r = –0.441$ (green cell), that is highly significant (blue cell). Both correlations are based on 103 observations. The significance values of the correlation coefficient tell us that the probability of getting correlation coefficients at least as big as those observed in a sample of 103 people if the null hypothesis were true (there was no relationship between these variables) is close to zero. All the significance values are below the criterion of 0.005 that we set for this study, indicating a 'statistically significant' relationship.

The second table in Output 8.1 shows the 95% confidence intervals for the correlation coefficients. Remember that these intervals will contain the population value of the association between variables in 95% of samples. Under the assumption (*which could be wrong*) that our sample is one of the 95% that produces confidence intervals that contain the true effect, we could conclude that (1) the true relationship between revision time and exam performance falls somewhere between $r = 0.218$ and $r = 0.547$ (i.e. fairly large and not zero); (2) that the true relationship between revision time and exam anxiety falls somewhere between –0.793 and –0.596 (i.e. extremely large and not zero); and (3) that

Correlations

				Time spent revising	Exam performance (%)	Exam anxiety
Time spent revising	Pearson Correlation			--		
	N			103		
Exam performance (%)	Pearson Correlation			.397**	--	
	Sig. (2-tailed)			<.001		
	N			103	103	
	Bootstrap[c]	Bias		-.004		
		Std. Error		.068		
		BCa 95% Confidence Interval	Lower	.245		
			Upper	.517		
Exam anxiety	Pearson Correlation			-.709**	-.441**	--
	Sig. (2-tailed)			<.001	<.001	
	N			103	103	103
	Bootstrap[c]	Bias		.003	.004	
		Std. Error		.114	.063	
		BCa 95% Confidence Interval	Lower	-.865	-.564	
			Upper	-.462	-.294	

**. Correlation is significant at the 0.01 level (2-tailed).

c. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Confidence Intervals

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a] Lower	Upper
Time spent revising - Exam performance (%)	.397	<.001	.218	.547
Time spent revising - Exam anxiety	-.709	<.001	-.793	-.596
Exam performance (%) - Exam anxiety	-.441	<.001	-.583	-.269

a. Estimation is based on Fisher's r-to-z transformation with bias adjustment.

Output 8.1

the true relationship between exam anxiety and exam performance falls somewhere between –0.583 and –0.269 (i.e. fairly large and not zero). Think about what a correlation of zero represents: it is no effect. None of confidence intervals contained the value of 0, so this suggests that the true value of these correlations is not zero (again, under the possibly incorrect assumption that our sample is one of the 95% that produces confidence intervals that contain the population value). In general, then, if we assume that our sample is one of the 95% of samples that produces a confidence interval containing the true value of the correlation (I can't stress enough that this assumption might be wrong), then if the interval crosses zero it means that (1) the population value could be zero (i.e., no effect at all), and (2) we can't be sure if the true relationship is positive or negative because the population value could plausibly be a negative or positive value.

Given the lack of normality in some of the variables, we should focus on the bootstrapped confidence intervals (labelled *BCa 95% Confidence Interval* in the main table) over the ones based on normal theory. For each correlation we're given two values: the upper boundary and the lower boundary. For the relationship between revision time and exam performance the interval is 0.245 to 0.517, for revision time and exam anxiety it is –0.865 to –0.462, and for exam anxiety and exam performance it is –0.564 to –0.294. These values differ from the non-robust confidence intervals but the interpretation doesn't change. An important point about the bootstrap confidence intervals is that because they are derived empirically using a random sampling procedure (i.e., bootstrapping) the results will be slightly different each time you run the analysis. The confidence intervals you get won't be the same as in Output 8.1, and that's normal and nothing to worry about.

To sum up, there appears to be a complex interrelationship between the three variables: (1) as anxiety about an exam increases, the percentage mark obtained in that exam significantly decreases; (2) as the amount of time revising increases, the percentage obtained in the exam significantly increases; and (3) as revision time increases, the student's anxiety about the exam significantly decreases.



Although we cannot make direct conclusions about causality from a correlation coefficient (Misconception Mutt 8.1), we can take it a step further by squaring it. The correlation coefficient squared (known as the **coefficient of determination**, R^2) is a measure of the amount of variability in one variable that is shared by the other. Exam performance in our data varied, reflecting the fact that people's scores were not identical; they will have varied for any number of reasons (different ability, different levels of preparation and so on). If we add up this individual variability we get an estimate of how much variability in exam performances there was in total (this is computed with Eq. 1.7 from Section 1.7.5). It turns out that the variance is about 672 for the exam performance scores. Imagine that this value is the surface area of a biscuit. The size of the biscuit represents 100% of the variance in exam scores. Imagine that exam anxiety comes along and takes a bite out of the biscuit. R^2 tells us how much of this exam performance 'biscuit' has been eaten by exam anxiety. In other words, it tells us the proportion of variance (what proportion of the 672 units of variation) overlaps with exam anxiety. These two variables had a correlation of –0.4410 and so the value of R^2 is $-0.4410^2 = 0.194$, which means that 0.194 of the variability in exam performance is shared by exam anxiety. It's a bit easier to think of this value as a percentage rather than a proportion, which we can obtain by multiplying by 100. In this example, then, exam anxiety shares 19.4% of the variability in exam performance. To put this value into perspective, this leaves 80.6% of the variability unexplained.

You'll often see people write things about R^2 that imply causality: they might write 'the variance in *y accounted for* by *x*', or 'the variation in one variable *explained* by the other'. Although R^2 is a useful measure of the substantive importance of an effect, it cannot be used to infer causal relationships. Exam anxiety might well share 19.4% of the variation in exam scores, but it does not necessarily cause this variation.

8.4.3 Spearman's correlation coefficient Ⓐ

We discovered in Section 8.2.5 that Spearman's r_s is a non-parametric correlation that can be used if you have concerns that the assumptions of Pearson's *r* have not been met, or there are extreme cases in the data. Let's look at an example. I was born in England, which has some bizarre traditions. One such oddity is the World's Biggest Liar competition held annually at the Santon Bridge Inn in Wasdale (in the Lake District). The contest honours a local publican, 'Auld Will Ritson', who in the nineteenth century was famous in the area for his far-fetched stories (one such tale being that Wasdale turnips were big enough to be hollowed out and used as garden sheds). Each year locals are encouraged to attempt to tell the biggest lie in the world (lawyers and politicians are apparently banned from the competition). Over the years there have been tales of mermaid farms, giant moles and farting sheep blowing holes in the ozone layer. (I am thinking of entering next year and reading out some sections of this book.)

Imagine I wanted to test a theory that more creative people will be able to create taller tales. I gathered 68 past contestants from this competition and noted where they were placed in the competition (first, second, third, etc.); I also gave them a creativity questionnaire (maximum score 60). The position in the competition is an ordinal variable (see Section 1.5.2) because the places are categories but have a meaningful order (first place is better than second place and so on). Therefore, Spearman's correlation coefficient should be used (Pearson's *r* requires interval or ratio data). The data for this study are in the file **biggest_liar.sav**. The data are in two columns: one labelled **creativity** and one labelled **position** (there's a third variable in there but we will ignore it for the time being). For the **position** variable, each of the categories described above has been coded with a numerical value. First place has been coded with the value 1, with other positions being labelled 2, 3 and so on. Note that for each numeric code I have provided a value label (just like we did for coding variables). I have also set the *Measure* property of this variable to Ordinal.

The procedure for doing a Spearman correlation is the same as for a Pearson correlation except that in the *Bivariate Correlations* dialog box (Figure 8.8), we need to select ☑ Spearman and deselect the option for a Pearson correlation. To get confidence intervals click Confidence Interval... and select ☑ Estimate confidence interval of bivariate correlation parameter and ◉ Bonett and Wright. We can also click on Bootstrap... to get robust confidence intervals.

The output for a Spearman correlation (Output 8.2) is like that of the Pearson correlation, giving the correlation coefficient between the two variables (–0.377), the significance value of this coefficient (0.002) and finally the sample size (68).[4] We also have the BCa 95% confidence interval, ranging from –0.622 to –0.126.[5] The fact that the confidence interval does not cross zero (and the significance is less than 0.05) tells us that there is a non-zero negative relationship between creativity scores and how well someone did in the World's Biggest Liar competition: as creativity increased, position decreased.[6] This might seem contrary to what we predicted until you remember that a low number means that you did well in the competition (a low number such as 1 means you came first, and a high number like 4 means you came fourth). Therefore, our hypothesis is supported: as creativity increased, so did success in the competition.

4 It is good to check that the value of *N* corresponds to the number of observations that were made. If it doesn't then data may have been excluded for some reason.

5 Remember that these confidence intervals are based on a random sampling procedure so the values you get will differ slightly from mine, and will change if you rerun the analysis.

6 Again, this conclusion is valid under the (possibly incorrect) assumption that our sample is one of the 95% that yield confidence intervals containing the population parameter.

Correlations

					Creativity	Position in Biggest Liar Competition
Spearman's rho	Creativity	Correlation Coefficient			--	
		Sig. (2-tailed)			.	
		N			68	
	Position in Biggest Liar Competition	Correlation Coefficient			-.377**	--
		Sig. (2-tailed)			.002	.
		N			68	68
		Bootstrap[c]	Bias		.004	
			Std. Error		.120	
			BCa 95% Confidence Interval	Lower	-.622	
				Upper	-.126	

**. Correlation is significant at the 0.01 level (2-tailed).

c. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Confidence Intervals of Spearman's rho

	Spearman's rho	Significance (2-tailed)	95% Confidence Intervals (2-tailed)[a,b] Lower	Upper
Creativity - Position in Biggest Liar Competition	-.377	.002	-.570	-.143

a. Estimation is based on Fisher's r-to-z transformation.

b. Estimation of standard error is based on the formula proposed by Bonett and Wright.

Output 8.2



8.4.4 Kendall's tau (non-parametric) Ⓐ

We discovered in Section 8.2.5 that Kendall's tau is another non-parametric correlation that should be used rather than Spearman's coefficient when you have a small data set with a large number of tied ranks. To carry out Kendall's correlation on the World's Biggest Liar data follow the same steps as for Pearson and Spearman correlations but select ☑ Kendall's tau-b and deselect the Pearson and Spearman options. There's no option for confidence intervals for Kendall's tau other than bootstrap confidence intervals, so we should select those in the usual way.

The output is much the same as for Spearman's correlation (Output 8.3). Notice that the value of the correlation coefficient is closer to zero than the Spearman correlation (it has changed from –0.377 to –0.302). Despite the difference in the correlation coefficients we can interpret this result as being a highly significant relationship because the significance value of 0.001 is less than our criterion of 0.005 and the robust confidence interval does not cross zero (–0.488 to –0.100). However, Kendall's value is likely a more accurate gauge of what the correlation in the population would be. As with any correlation, we cannot assume that creativity caused success in the World's Best Liar competition.



Correlations

					Creativity	Position in Biggest Liar Competition
Kendall's tau_b	Creativity	Correlation Coefficient			--	
		Sig. (2-tailed)			.	
		N			68	
	Position in Biggest Liar Competition	Correlation Coefficient			-.302**	--
		Sig. (2-tailed)			.001	.
		N			68	68
		Bootstrap[c] Bias			.004	
		Std. Error			.101	
		BCa 95% Confidence Interval		Lower	-.488	
				Upper	-.100	

**. Correlation is significant at the 0.01 level (2-tailed).

c. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Confidence Intervals of Kendall's tau_b

	Kendall's tau_b	Significance (2-tailed)	95% Confidence Intervals (2-tailed)[a] Lower	Upper
Creativity – Position in Biggest Liar Competition	-.302	.001	-.441	-.148

a. Estimation is based on Fisher's r-to-z transformation.

Output 8.3

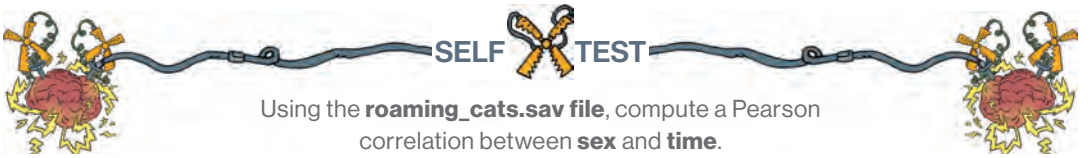
8.4.5 Biserial and point-biserial correlations ©

Often it is necessary to investigate relationships between two variables when one of the variables is dichotomous (i.e., it is categorical with only two categories). The biserial and point-biserial correlation coefficients should be used in these situations. These correlations are distinguished by a conceptual difference and their statistical calculation is quite different. The difference between the use of biserial and point-biserial correlations depends on whether the dichotomous variable is discrete or continuous. This difference is very subtle. A discrete, or true, dichotomy is one for which there is no underlying continuum between the categories. An example of this is whether someone is dead or alive: a person can be only dead or alive, they can't be 'a bit dead'. Although you might describe a person as being 'half-dead' there is a definition of clinically dead and if you don't meet that definition then you are alive. As such, there is no continuum between the two categories. However, it is possible to have a dichotomy for which a continuum does exist. An example is passing or failing a statistics test: some people will only just fail while others will fail by a large margin; likewise, some people will scrape a pass while others will excel. Although participants fall into only two categories (pass or fail) there is an underlying continuum along which people lie.

The **point-biserial correlation** coefficient (r_{pb}) is used when one variable is a discrete dichotomy (e.g., pregnancy), whereas the **biserial correlation** coefficient (r_b) is used when one variable is a continuous dichotomy (e.g., passing or failing an exam). The biserial correlation coefficient cannot be calculated directly in SPSS; first you must calculate the point-biserial correlation coefficient and then adjust the value. Let's look at an example.

Imagine that I was interested in the relationship between the sex of a cat and how much time it spent away from home (I love cats, so these things interest me). I had heard that male cats disappeared for substantial amounts of time on long-distance roams around the neighbourhood (something about hormones driving them to find mates) whereas female cats tended to be more homebound. I used this as a purr-fect (sorry!) excuse to go and visit lots of my friends and their cats. I took a note of the sex of the cat (**sex**) and then asked the owners to note down the number of hours that their cat was absent from home over a week (**time**). The time spent away from home is measured at the ratio level – and let's assume it meets the other relevant assumptions – while the sex of the cat is a discrete dichotomy. The data are in the file **roaming_cats.sav**.

We want to calculate a point-biserial correlation, which is the Pearson correlation when the dichotomous variable is coded with 0 for one category and 1 for the other (in practice, you can use any values because SPSS changes the lower one to 0 and the higher one to 1 when it does the calculations). In the saved data I coded the **sex** variable 1 for male and 0 for female. The **time** variable contains the roaming time over the week in hours.



Congratulations – if you did the self-test task then you have conducted your first point-biserial correlation. Despite the horrible name, it's really easy to do. You should find that you have the same output as Output 8.4, which shows the correlation matrix of **time** and **sex**. The point-biserial correlation coefficient is $r_{pb} = 0.378$, which has a significance value of 0.003. The significance test for this correlation is the same as performing an independent-samples *t*-test on the data (see Chapter 10). The sign of the correlation (i.e., whether the relationship was positive or negative) depends entirely on which way round the dichotomous variable was coded. To prove this point, the data file has an extra variable called **recode** which is the same as the variable **sex** except that the coding is reversed (1 = female, 0 = male). If you repeat the Pearson correlation using **recode** instead of **sex** you will find that the correlation coefficient becomes –0.378. The sign of the coefficient is completely dependent on which category you assign to which code and so we must ignore all information about the direction of the relationship. However, we can still interpret R^2 as before. In this example, $R^2 = 0.378^2 = 0.143$, which equates to the sex of the cat sharing 14.3% of the variability in the time spent away from home.

Correlations

				Time away from home (hours)	Sex of cat
Time away from home (hours)	Pearson Correlation			--	
	N			60	
Sex of cat	Pearson Correlation			.378**	--
	Sig. (2-tailed)			.003	
	N			60	60
	Bootstrap[c]	Bias		.005	
		Std. Error		.114	
		BCa 95% Confidence Interval	Lower	.123	
			Upper	.619	

**. Correlation is significant at the 0.01 level (2-tailed).

c. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 8.4

8.5 Partial and semi-partial correlation Ⓑ

8.5.1 Semi-partial (or part) correlation Ⓒ

I mentioned earlier that there is a type of correlation that can be done that allows you to look at the relationship between two variables while accounting for the effect of a third variable. For example, in



Oliver Twisted
Please, Sir, can I have some more ... biserial correlation?

'Some of the male cats were neutered and so there might be a continuum of maleness that underlies the sex variable, you horns-woggler,' Oliver hurls at me. 'We need to convert the point-biserial correlation into the biserial correlation coefficient (r_b). I think you're Fagin your knowledge of how to do this.' Oliver, if you go to www.discoverspss.com you'll find that I am not artfully dodging how to do the conversion.



Cramming Sam's Tips
Correlations

- Spearman's correlation coefficient, r_s, is a non-parametric statistic and requires only ordinal data for both variables.
- Kendall's correlation coefficient, τ, is like Spearman's r_s but probably better for small samples.
- The point-biserial correlation coefficient, *r*pb, quantifies the relationship between a continuous variable and a variable that is a discrete dichotomy (e.g., there is no continuum underlying the two categories, such as dead or alive).
- The biserial correlation coefficient, r_b, quantifies the relationship between a continuous variable and a variable that is a continuous dichotomy (e.g., there is a continuum underlying the two categories, such as passing or failing an exam).



the exam anxiety data (in the file **exam_anxiety.sav**) exam performance (EP) was negatively related to exam anxiety (EA), but positively related to revision time (RT), and revision time itself was negatively related to exam anxiety. If revision time is related to both exam anxiety and exam performance, then to get a measure of the unique relationship between exam anxiety and exam performance we need to account for revision time.

Let's begin by transforming the correlation coefficients from Output 8.1 into proportions of variance by squaring them:

$$\begin{aligned}&\text{Proportion of variance shared by EP and EA} = r^2_{\text{EP–EA}} = -0.441^2 = 0.194\\&\text{Proportion of variance shared by EP and RT} = r^2_{\text{EP–RT}} = -0.397^2 = 0.158\\&\text{Proportion of variance shared by EA and RT} = r^2_{\text{EA–RT}} = -0.709^2 = 0.503\end{aligned} \tag{8.15}$$

If we multiply the resulting proportions by 100 to turn them into percentages (which people typically find easier to understand) we see that exam performance shares 19.4% of its variance with exam anxiety, and 15.8% with revision time. Revision time shares about half of its variance (50.3%) with exam anxiety.

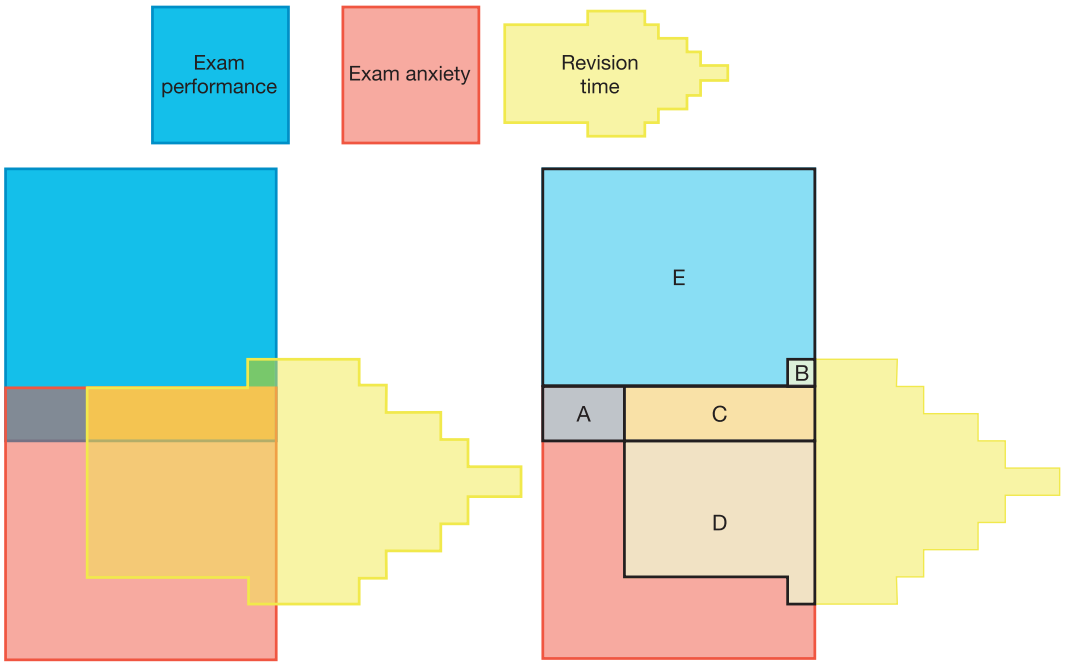
Figure 8.11 depicts the variance of each variable as a shape: exam performance and anxiety are blue and red squares respectively, and revision time is a sort of weird yellow spacecraft or spark plug. The area of each shape is 100% of the variance of the variable it represents. The figure shows the overlap in the variances of these variables. The depicted overlaps are proportional (approximately) to the actual data: if two variables share 10% of variance then about 10% of the surface areas of their shapes overlap in the diagram. The left-hand side shows only the overlap of shapes/variances and the right duplicates the image but with specific areas marked out and labelled.

To see what the overlaps mean, let's look at exam performance and exam anxiety. Notice that the bottom of the square for exam performance overlaps with the top of the square for exam anxiety. For both variables (squares) this overlap is about 19.4% of their surface area, which corresponds to the amount of variance they share (Eq. 8.15). This overlap is (I hope) clear on the left-hand side, but on the right I have marked out these areas as A and C. What distinguishes area A from area C is that area C also overlaps with revision time, whereas area A does not. Together (A + C) these areas are the overlap between exam performance and exam anxiety (19.4%), but this overlap can be decomposed into variance in exam performance that is unique to exam anxiety (A) and variance in exam performance that is not unique to exam anxiety because it is also shared with revision time (C).

Similarly, look at the overlap between exam performance and revision time by focusing on the area where the yellow shape overlaps with the blue square. This overlap is represented by a sort of L-shape that in total represents 15.8% of the surface area of both shapes (i.e., 15.8% of shared variance). I have marked out the overlap between the blue square and yellow shape on the right with areas labelled B and C. As before, these areas combine (B + C) to represent the overlap between exam performance and revision time (15.8%) but this overlap can be decomposed into variance in exam performance that is unique to revision time (B) and variance in exam performance that is not unique to revision time because it is also shared with exam anxiety (C).

We know from Eq. 8.15 that revision time and exam anxiety share 50.3% of their variance. This is represented in Figure 8.11 by the overlap between the yellow shape and red square, which is represented by areas C and D. Like before, these areas combine (C + D) to represent the total overlap between exam anxiety and revision time (50.3%) but this shared variance can be decomposed into variance in exam anxiety that is unique to revision time (D) and variance in exam anxiety that is not unique to revision time because it is also shared with exam performance (C).

Figure 8.11 contains the size of the areas A, B, C and D as well as some of their sums; it also explains what the areas represent in isolation and in combination; for example, area B is the variance in exam performance that is *uniquely shared* with revision time (1.5%), but in combination with area C it is the *total* variance in exam performance that is shared with revision time (15.8%). Think about that. The correlation between exam performance and revision time tells us that they share 15.8% of



A = variance exam performance uniquely shared with exam anxiety (5.1%)

B = variance in exam performance uniquely shared with revision time (1.5%)

C = variance in exam performance shared by both exam anxiety and revision time (14.3%)

D = variance shared by exam anxiety and revision time but not exam performance (36%)

E = variance in exam performance not shared by any measured variable (79.1%)

A + C = variance shared by exam performance and exam anxiety (19.4%)

C + B = variance shared by exam performance and revision time (15.8%)

C + D = variance shared by revision time and exam anxiety (50.3%)

A + B + C = variance in exam performance accounted for by revision time and exam anxiety (20.9%)

Figure 8.11 Diagram showing the principle of partial correlation

the variance, but in reality, only 1.5% is unique to revision time whereas the remaining 14.3% (area C) is also shared with exam anxiety. This issue gets to the heart of **semi-partial correlation** (also referred to as a *part correlation*), which we'll return to in the next chapter. The areas of unique variance in Figure 8.11 represent the semi-partial correlation. For example, area A is the unique variance in exam performance shared with exam anxiety expressed as a proportion of the variance in exam performance. It is 5.1%. This means that 5.1% of the variance in exam performance is shared uniquely with exam anxiety (and no other variable that we're currently using to predict exam performance). As a proportion this value is 0.051, and remember that to get these proportions we square the correlation coefficient, so to get back to r we would take the square root of 0.051, which is 0.226 (because r can be negative the value could also be –0.226). This is the semi-partial correlation between exam performance and exam anxiety. It is the correlation between exam performance and exam anxiety, ignoring the portion of that relationship that is also shared by revision time. Likewise, we can look at area B, which represents the unique variance in exam performance shared with revision time (1.5%). As a proportion this value is 0.015, and to get this proportion back to a correlation coefficient, we take the square root, which gives us 0.122 (or –0.122). This value is the semi-partial correlation between exam performance and revision time. It is the correlation between exam performance and revision time, ignoring the portion of that relationship that is also shared by exam anxiety.

As such, the semi-partial correlation expresses the unique relationship between two variables as a function of their total variance. In general terms, imagine we want to look at the relationship between two variables X and Y, adjusting for the effect of a third variable Z. Again it's easier to think in terms of proportions (r^2) and not r itself; the semi-partial correlation squared is the uniquely shared variance between X and Y, expressed as a proportion of the *total variance in Y*. We can show this in terms of the areas in Figure 8.11:

$$sr^2_{\text{EP–EA}} = \frac{A}{A+B+C+E} = \frac{5.1}{100} = 0.051$$
$$sr_{\text{EP–EA}} = \sqrt{sr^2_{\text{EP–EA}}} = \sqrt{0.051} = \pm 0.226 \quad (8.16)$$

The semi-partial correlation squared (sr^2) between exam performance and exam anxiety is the uniquely overlapping area (A) expressed as a function of the total area of exam performance (A + B + C + E). The values of each area are underneath Figure 8.11 if you want to plug the numbers into Eq. 8.16 as I have. As such a semi-partial correlation is the relationship between X and Y (area A) accounting for the overlap in Y and Z (areas C and B), but not the overlap in X and Z (area D). In our specific example, the semi-partial correlation for exam performance and exam anxiety (area A) quantifies their relationship accounting for the overlap between exam anxiety and revision time (area C) *but not* the overlap between exam performance and revision time (area B).

8.5.2 Partial correlation ©

Another way to express the unique relationship between two variables (i.e., the relationship accounting for other variables) is the **partial correlation**. Remember that the semi-partial correlation expresses the unique relationship between two variables, X and Y (area A), as a function of the *total variance* in Y (area A + B + C + D). We can instead express this unique variance in terms of the *variance* in *Y left over when other variables have been considered*. The following equation shows this in terms of the areas in Figure 8.11:

$$pr^2_{\text{EP–EA}} = \frac{A}{A+E} = \frac{5.1}{5.1+79.1} = 0.061$$
$$pr_{\text{EP–EA}} = \sqrt{sr^2_{\text{EP–EA}}} = \sqrt{0.066} = \pm 0.247 \quad (8.17)$$

Compare this equation with Eq. 8.16 and notice that the denominator has changed from the total area for exam performance (A + B + C + E) to the area for exam performance that is left over once we have considered revision time (A + E). In other words, the areas in exam performance that overlap with revision time (areas B and C) are removed from the denominator. Hopefully this makes clear the point that the partial correlation expresses the unique relationship between X and Y as a function of the *variance* in *Y left over when other variables have been considered*. Returning to our exam example, partial correlation squared (pr^2) between exam performance and exam anxiety is the uniquely overlapping area (A) expressed as a function of the area of exam performance that does *not* overlap with revision time (A + E). Again you can use the values in Figure 8.11 to follow Eq. 8.17.

By ignoring the variance in Y that overlaps with Z, a partial correlation adjusts for both the overlap that X and Z have *and* the overlap in Y and Z, whereas a semi-partial correlation adjusts only for the overlap that X and Z have. In our specific example, the partial correlation for exam performance and exam anxiety (area A) adjusts for both the overlap in exam anxiety and revision time (area C) *and* the overlap in exam performance and revision time (area B)

8.5.3 Partial correlation using SPSS ©

Reload the **exam_anxiety.sav** file so that we can conduct a partial correlation between exam anxiety and exam performance while adjusting for the effect of revision time. Access the *Partial Correlations* dialog box (Figure 8.12) using the *Analyze* > *Correlate* > Partial... menu. Your variables will be listed in the left-hand box, and on the right the box labelled *Variables* is for specifying the variables that you want to correlate and the box labelled *Controlling for* is for declaring the variables for which you want to adjust. If we want to look at the unique effect of exam anxiety on exam performance we'd correlate the variables **exam** and **anxiety** while adjusting for **revise** as shown in Figure 8.12. In this instance we are going to adjust for one variable, which is known as a *first-order partial correlation*. It is possible to control for the effects of two variables (a *second-order partial correlation*), three variables (a *third-order partial correlation*) and so on by dragging more variables to the *Controlling for* box.

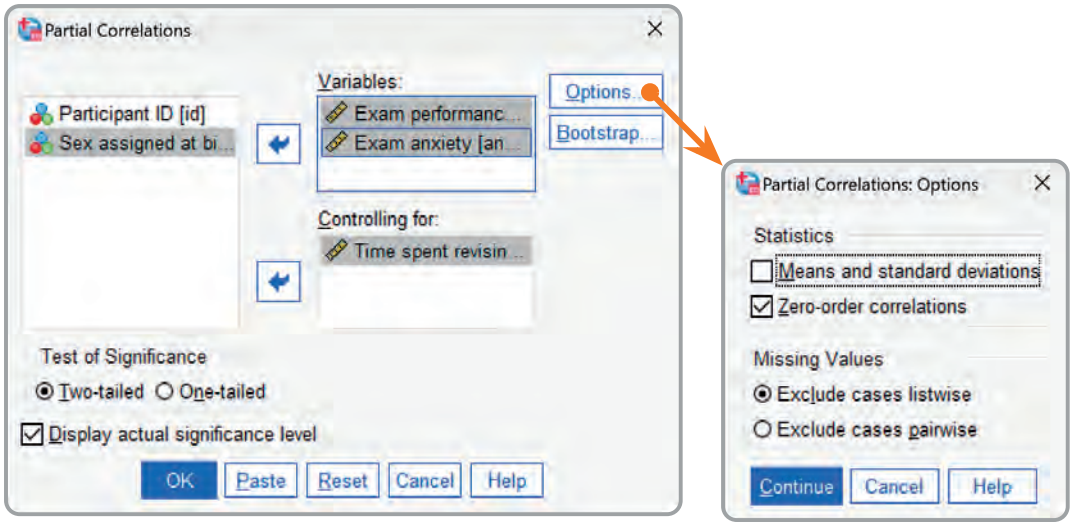


Figure 8.12 Main dialog box for conducting a partial correlation

Clicking Options... accesses options like those in bivariate correlation. Within this dialog box you can select *Zero-order correlations*, which are the Pearson correlation coefficients without adjusting for other variables. If we select this option SPSS will produce a correlation matrix of **anxiety**, **exam** and **revise**, which might be useful if you haven't already looked at the raw (or zero-order) correlations between variables, but we have (Output 8.1) so don't tick this option for this example. As we have done throughout, use the Bootstrap... option to get some robust confidence intervals.

Output 8.5 shows the partial correlation of exam anxiety and exam performance adjusting for revision time. Note that the top and bottom of the table contain identical values, so we can ignore one half of the table. The partial correlation between exam performance and exam anxiety is –0.247, which is the same value as we computed in Eq. 8.17. This value is considerably less than when we didn't adjust for the effect of revision time (r = –0.441). Although this correlation is still statistically significant (its p-value is still below 0.05) and the confidence interval [–0.431, –0.019] still doesn't contain zero, the relationship is diminished. In terms of variance, the value of R^2 for the partial correlation is 0.061, which means that exam anxiety shares only 6.1% of the variance in exam performance that is left over by revision time (compared to 19.4% when revision time was not factored in). Running this analysis has shown us that exam anxiety alone does explain some of the variation in exam scores, but there is a complex relationship between anxiety, revision and exam performance that might otherwise have been ignored. Although causality is still not certain, because relevant variables are being included, the third-variable problem is, at least, being addressed to some degree.

Correlations

Control Variables					Exam performance (%)	Exam anxiety
Time spent revising	Exam performance (%)	Correlation			1.000	-.247
		Significance (2-tailed)			.	.012
		df			0	100
		Bootstrap[a]	Bias		.000	.008
			Std. Error		.000	.103
			BCa 95% Confidence Interval	Lower	.	-.431
				Upper	.	-.019
	Exam anxiety	Correlation			-.247	1.000
		Significance (2-tailed)			.012	.
		df			100	0
		Bootstrap[a]	Bias		.008	.000
			Std. Error		.103	.000
			BCa 95% Confidence Interval	Lower	-.431	.
				Upper	-.019	.

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 8.5 Output from a partial correlation

Partial correlations can be done when variables are dichotomous (including the 'third' variable). For example, we could look at the relationship between bladder relaxation (did the person wet themselves or not?) and the number of large tarantulas crawling up the person's leg, adjusting for fear of spiders (the first variable is dichotomous, but the second variable and 'control' variable are both continuous). Similarly, to use an earlier example, we could examine the relationship between creativity and success in the World's Biggest Liar contest, adjusting for whether someone had previous experience in the competition (and therefore had some idea of the type of tale that would win). In this case the 'control' variable is dichotomous.[7]



Cramming Sam's Tips
Partial and semi-partial correlations

- A partial correlation quantifies the relationship between two variables while adjusting for the effects of a third variable on both variables in the original correlation.
- A semi-partial correlation quantifies the relationship between two variables while adjusting for the effects of a third variable on only one of the variables in the original correlation.

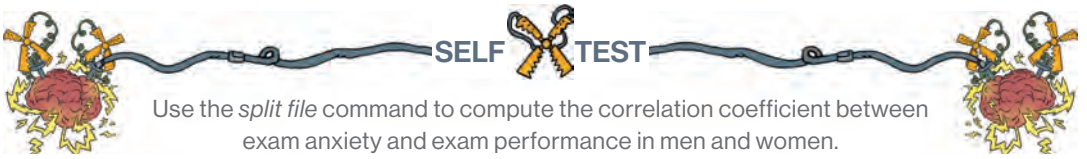


7 Both these examples are, in fact, simple cases of hierarchical regression (see the next chapter) and the first example is also an example of analysis of covariance. This may not mean much yet, but it illustrates what I have repeatedly said about all statistical models being variations of the same linear model.

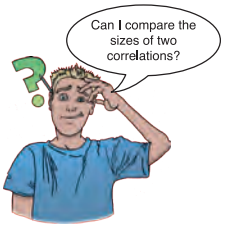
8.6 Comparing correlations ©

8.6.1 Comparing independent *r*s ©

Sometimes we want to know whether one correlation coefficient is bigger than another. For example, when we looked at the effect of exam anxiety on exam performance, we might have been interested to know whether this correlation was different in men and women. We can compute the correlation in these two samples, but how do we assess whether the difference is meaningful?



If you do the self-test you'll find that the correlations are $r_{\text{Male}} = -0.506$ and $r_{\text{Female}} = -0.381$. These two samples are independent (they contain different entities). To compare these correlations we can use what we discovered in Section 8.2.3 to convert these coefficients to z_r (because it makes the sampling distribution normal). Do the conversion and you should obtain z_r (males) = –0.557 and z_r (females) = –0.401. We can calculate a *z*-score of the differences between these correlations using:



$$z_{\text{Difference}} = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} \tag{8.18}$$

We had 52 men and 51 women so the resulting *z* is:

$$z_{\text{Difference}} = \frac{-0.557 - (-0.401)}{\sqrt{\frac{1}{49} + \frac{1}{48}}} = \frac{-0.156}{0.203} = -0.768 \tag{8.19}$$

We can look up this value of *z* (0.768; we can ignore the minus sign) in the table for the normal distribution in the Appendix and get the one-tailed probability from the column labelled 'Smaller Portion'. The value is 0.221, which we need to double to get the two-tailed probability of 0.442. The value 0.768 is less than 1.96 (the 0.05 criterion *z*-score) so the correlation between exam anxiety and exam performance is not significantly different in men and women.

8.6.2 Comparing dependent *r*s ©

If you want to compare correlation coefficients that come from the same entities then things are a little more complicated. You can use a *t*-statistic to test whether a difference between two dependent correlations is significant. For example, in our exam anxiety data we might want to see whether the relationship between exam anxiety (*x*) and exam performance (*y*) is stronger than the relationship between revision (*z*) and exam performance. To address this question, we need the three *r*s that quantify the relationships between these variables: r_{xy}, the relationship between exam anxiety and exam performance (–0.441); r_{zy}, the relationship between revision and exam performance (0.397); and r_{xz}, the relationship between exam anxiety and revision (–0.709). The *t*-statistic is computed as follows (e.g., Chen & Popovich, 2002):

$$t_{\text{Difference}} = \left(r_{xy} - r_{zy}\right)\sqrt{\frac{(n-3)(1+r_{xz})}{2\left(1-r_{xy}^2-r_{xz}^2-r_{zy}^2+2r_{xy}r_{xz}r_{zy}\right)}} \tag{8.20}$$

Admittedly that equation looks hideous, but really it's not too bad when you remember that it just uses the three correlation coefficients and the sample size N.

Place the numbers from the exam anxiety example in (*N* was 103) and you should end up with:

$$t_{\text{Difference}} = (-0.838)\sqrt{\frac{29.1}{2(1-0.194-0.503-0.158+0.248)}} = -5.09 \tag{8.21}$$

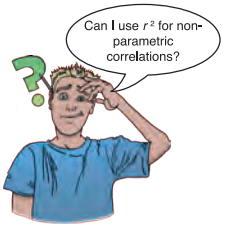
This value can be checked against the appropriate critical value for *t* with *N*–3 degrees of freedom (in this case 100). The critical values in the table (see the Appendix) are 1.98 ($p \leq 0.05$) and 2.63 ($p \leq 0.01$), two-tailed. As such we can say that the correlation between exam anxiety and exam performance was significantly higher than the correlation between revision time and exam performance (this isn't a massive surprise, given that these relationships went in the opposite directions to each other).

8.6.3 Comparing *r*s using SPSS Ⓑ

You can't compare correlations through the dialog boxes in SPSS, but there is a macro available by Weaver and Wuensch (2013).

8.7 Calculating the effect size Ⓐ

Calculating effect sizes for correlation coefficients couldn't be easier because, as we saw earlier in the book, correlation coefficients *are* effect sizes. So, no calculations (other than those you have already done) are necessary. However, although the Spearman and Kendall correlations are comparable to Pearson's *r* in many respects (their power, for example, is similar under parametric conditions), there are important differences (Strahan, 1982).



First, we can square the value of Pearson's *r* to get the proportion of shared variance, R^2. For Spearman's r_s we can do this too because it uses the same equation as Pearson's *r*. However, the resulting R_s^2 is the proportion of variance in the *ranks* that two variables share. Having said this, R_s^2 is usually a good approximation of R^2 (especially in conditions of near-normal distributions). Kendall's τ is not numerically similar to either *r* or r_s and so τ^2 does not tell us about the proportion of variance shared by two variables (or the ranks of those two variables).

The second difference relates to a more general point that when using correlations as effect sizes (both when reporting your own analysis and when interpreting others) you should be mindful that the choice of correlation coefficient can make a substantial difference to the apparent size of the effect. For example, Kendall's τ is 66–75% smaller than both Spearman's r_s and Pearson's *r*, but *r* and r_s are generally similar in size (Strahan, 1982). Therefore, if τ is used as an effect size it is not comparable to *r* and r_s. The point-biserial and biserial correlations differ in size too, so you should think very carefully about whether your dichotomous variable has an underlying continuum, or whether it is a truly discrete variable.

8.8 How to report correlation coefficients Ⓐ

Reporting correlation coefficients is pretty easy: you report how big they are, their confidence intervals and significance value (the significance value is probably the least important because the correlation

coefficient is an effect size). Some general points (see Sections 1.8.3 and 3.8) are as follows: (1) if you use APA style there should be no zero before the decimal point for the correlation coefficient or the probability value (because neither can exceed 1); (2) coefficients are usually reported to 2 or 3 decimal places because this is a reasonable level of precision; (3) generally people report 95% confidence intervals; (4) each correlation coefficient is represented by a different letter (and some of them are Greek); and (5) report exact *p*-values. Let's take a few examples from this chapter:

- ✓ There was not a significant relationship between the number of adverts watched and the number of packets of sweets purchased, $r = 0.87$, $p = 0.054$.
- ✓ Bias corrected and accelerated bootstrap 95% CIs are reported in square brackets. Exam performance was significantly correlated with exam anxiety, $r = -0.44$ [–0.56, –0.29], and time spent revising, $r = 0.40$ [0.25, 0.52]; the time spent revising was also correlated with exam anxiety, $r = -0.71$ [–0.87, –0.46] (all ps < 0.001).
- ✓ Creativity was significantly related to how well people did in the World's Biggest Liar competition, $r_s = -0.38$, 95% BCa CI [–0.62, –0.13], $p = 0.002$.
- ✓ Creativity was significantly related to a person's placing in the World's Biggest Liar competition, $\tau = -0.30$, 95% BCa CI [–0.49, –0.10], $p = 0.001$. (Note that I've quoted Kendall's τ.)
- ✓ The sex of the cat was significantly related to the time the cat spent away from home, $r_{pb} = 0.38$, 95% BCa CI [0.11, 0.62], $p = 0.003$.
- ✓ The sex of the cat was significantly related to the time the cat spent away from home, $r_b = 0.48$, $p = 0.003$.

A table is a good way to report lots of correlations. Our exam anxiety correlations could be reported as in Table 8.2. Note that above the diagonal I have reported the correlation coefficients and used symbols to represent different levels of significance. The confidence intervals are reported underneath. Under the table there is a legend to tell readers what symbols represent. (None of the correlations were non-significant or had *p* bigger than 0.001 so most of the table footer is there to give you a template – you would normally include only symbols that you had actually used in the table.) Finally, in the lower part of the table I have reported the sample sizes. These are all the same (103) but when you have missing data it is useful to report the sample sizes in this way because different values of the correlation will be based on different sample sizes. You could alternatively use the bottom part of the table to report exact *p*-values.

Table 8.2 An example of reporting a table of correlations

	Exam Performance	Exam Anxiety	Revision Time
Exam Performance	1	−0.44*** [−0.56, −0.29]	0.40*** [0.25, 0.52]
Exam Anxiety	103	1	−0.71*** [−0.87, −0.46]
Revision Time	103	103	1

ns = not significant ($p > 0.05$), * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. BCa bootstrap 95% CIs reported in brackets.

8.9 Jane and Brian's story Ⓐ

Brian wandered out of his lecture about correlations. An hour learning about relationships had made him feel flat. He knew it was stupid: the lecturer wasn't talking about those kinds of relationships, but the word kept buzzing around his head, distracting him. Jane was complex. On the one hand she



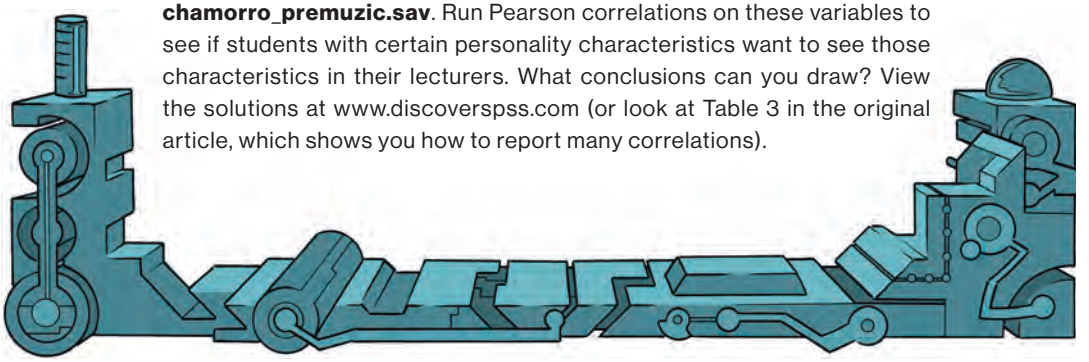
Labcoat Leni's Real Research 8.1
Why do you like your lecturers? Ⓐ

Chamorro-Premuzic, T., et al. (2008). *Personality and Individual Differences*, *44,* 965–976.

As students you probably have to rate your lecturers at the end of the course. There will be some lecturers you like and others you don't. As a lecturer I find this process horribly depressing (although this has a lot to do with the fact that I tend focus on negative feedback and ignore the good stuff). There is some evidence that students tend to pick courses of lecturers they perceive to be enthusiastic and good communicators. In a fascinating study, Tomas Chamorro-Premuzic and his colleagues (Chamorro-Premuzic et al., 2008) tested the hypothesis that students tend to like lecturers who are like themselves. (This hypothesis will have the students on my course who like my lectures screaming in horror.)

The authors measured students' own personalities using a very well-established measure (the NEO-FFI) which measures five fundamental personality traits: neuroticism, extroversion, openness to experience, agreeableness and conscientiousness. Students also completed a questionnaire in which they were given descriptions (e.g., 'warm: friendly, warm: sociable, cheerful, affectionate, outgoing') and asked to rate how much they wanted to see this in a lecturer from −5 (I don't want this characteristic in my lecturer at all) through 0 (the characteristic is not important) to +5 (I really want this characteristic in my lecturer). The characteristics were the same as those measured by the NEO-FFI.

As such, the authors had a measure of how much a student had each of the five core personality characteristics, but also a measure of how much they wanted to see those same characteristics in their lecturer. Tomas and his colleagues could then test whether, for instance, extroverted students want extroverted lecturers. The data from this study are in the file **chamorro_premuzic.sav**. Run Pearson correlations on these variables to see if students with certain personality characteristics want to see those characteristics in their lecturers. What conclusions can you draw? View the solutions at www.discoverspss.com (or look at Table 3 in the original article, which shows you how to report many correlations).



always gave him the brush-off, on the other he sensed a playful humour in her responses to him. Was he kidding himself, or was that her idea of flirting? Should he be trying to get involved with someone who thought that was how to flirt? It was confusing. He wanted to get in her head. Did she think they were developing a positive relationship? Was it negative? Non-existent? He decided to be direct.

Jane's morning had been strange. She'd worked late into the night, but woke early in a mental haze. She was acting on autopilot. She ate breakfast, spent more time on her makeup than she normally would, and put on some of her favourite clothes. She didn't remember doing any of it. It didn't register as odd to her. Neither did it seem unusual to her that she passed by the entrance of the Leviathan lecture theatre at exactly 10.50, the time that Brian's statistics lecture kicked out. Seeing Brian dispersed the fog in her mind. She panicked as she realized that her unconscious mind had bought her to this place at this exact time. She felt overdressed and awkward. He seemed sad, so she asked him if he was OK. He told her about the lecture he had just had, and the familiarity of correlations calmed her.

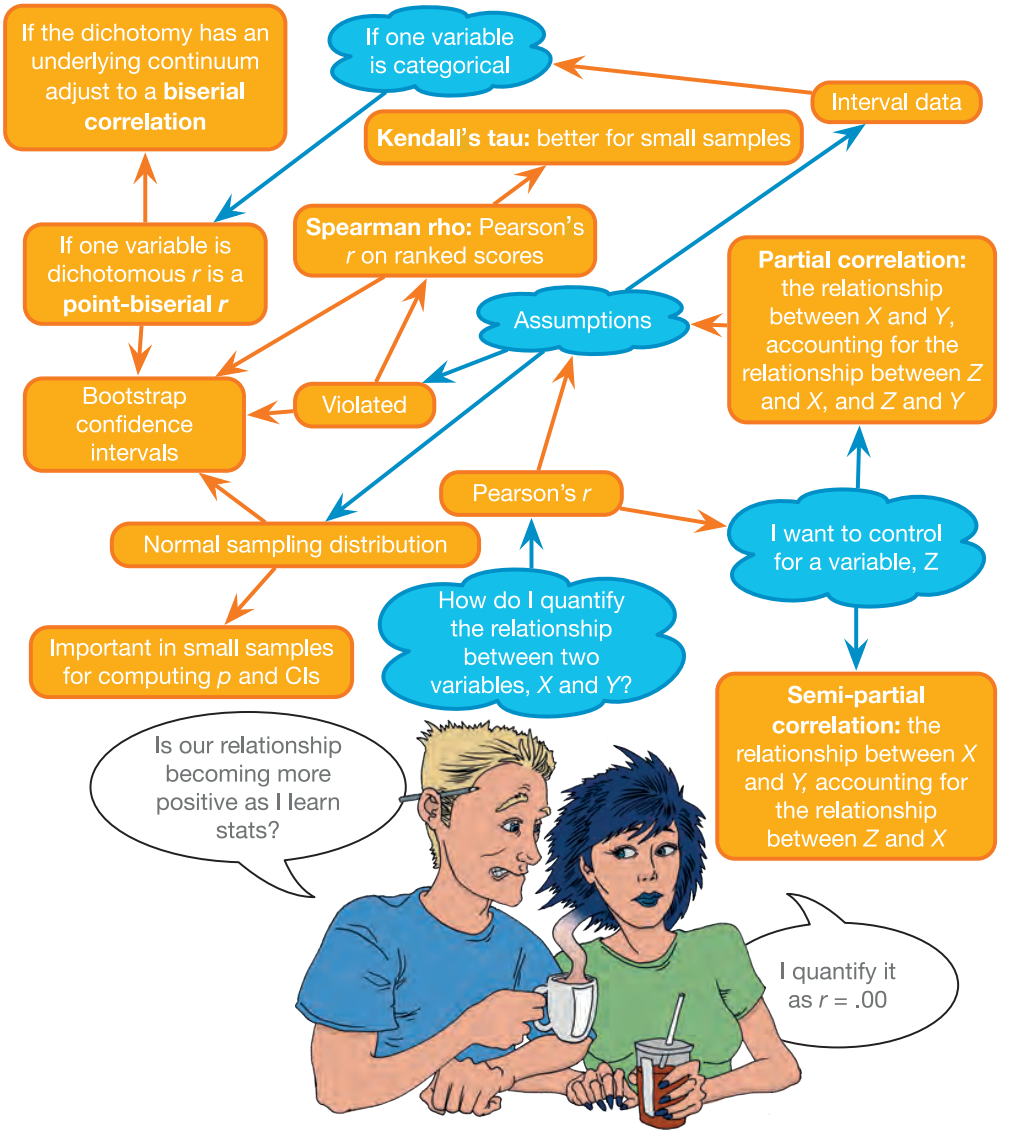


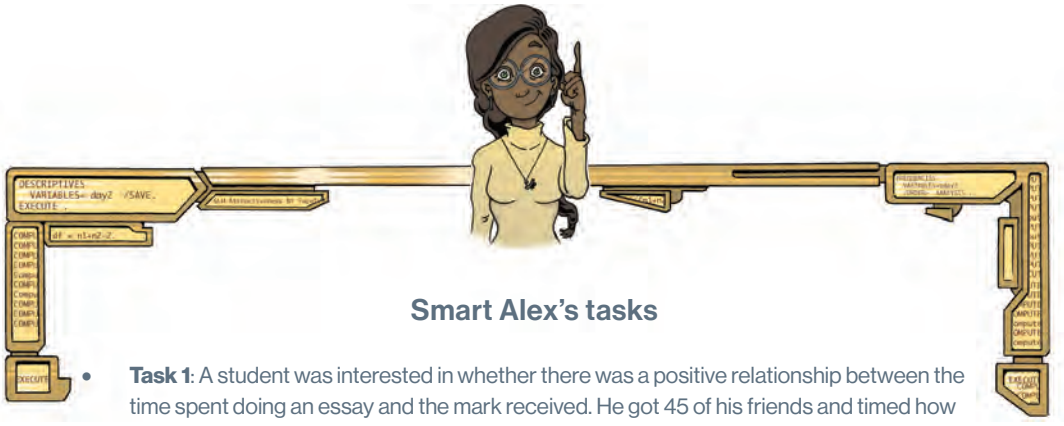
Figure 8.13 What Brian learnt from this chapter

8.10 What next? Ⓐ

At the age of 8 my dad taught me a valuable lesson, which is that if you really want something then you need to work at it, and the harder you work at it the more likely you are to get what you want. I did practise my guitar and before long the tears gave way to a competent version of 'Skip to my Lou'. My dad had also aspired to be a musician when he was young and encouraged my new passion.[8] He found me a guitar teacher and found the money for lessons. These lessons illustrate how being a good student often depends on finding the right teacher. Ken Steers, despite his best efforts, was on a completely different wavelength than me. I wanted to learn some crushing metal riffs, and he wanted me to work through Bert Weedon's 'Play in a day' and learn trad jazz classics. As an adult, I wish I had paid more attention to Ken because I'd have been a better guitar player than I am. I was a terrible student and adopted a strategy of selective practice: I'd practise if I wanted to do something but not if I thought it was 'boring'. Perhaps this is why I am so obsessed with trying not to be a boring teacher. Nevertheless, my dad and Ken did get me going and soon enough, like my favourite record of the time, I was ready to 'Take on the world'. Well, Wales at any rate …

8.11 Key terms that I've discovered

Biserial correlation
Bivariate correlation
Coefficient of determination
Covariance
Cross-product deviations
Kendall's tau
Partial correlation
Pearson correlation coefficient
Point-biserial correlation
Semi-partial correlation
Spearman's correlation coefficient
Standardization



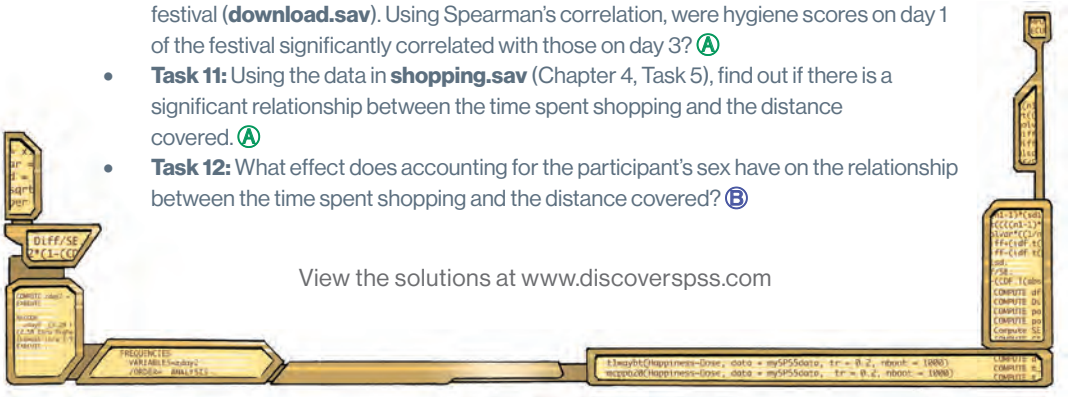
Smart Alex's tasks

- **Task 1**: A student was interested in whether there was a positive relationship between the time spent doing an essay and the mark received. He got 45 of his friends and timed how long they spent writing an essay (**hours**) and the percentage they got in the essay (**essay**). He also translated these grades into their degree classifications (**grade**): in the UK, a student can get a first-class mark (the best), an upper-second-class mark, a lower second, a third, a pass or a fail (the worst). Using the data in the file **essay_marks.sav** find out what the relationship was between the time spent doing an essay and the eventual mark in terms of percentage and degree class (draw a scatterplot too). Ⓐ
- **Task 2**: Using the **notebook.sav** data from Chapter 3, quantify the relationship between the participant's gender and arousal. Ⓐ
- **Task 3**: Using the notebook data again, quantify the relationship between the film watched and arousal. Ⓐ

8 My dad, like me, never made it in his band, but, unlike me, did sing on the UK TV show *Stars in Their Eyes*, which made us all pretty proud.

- **Task 4**: As a statistics lecturer I am interested in the factors that determine whether a student will do well on a statistics course. Imagine I took 25 students and looked at their grades for my statistics course at the end of their first year at university: first, upper second, lower second and third class (see Task 1). I also asked these students what grade they got in their high school maths exams. In the UK GCSEs are school exams taken at age 16 that are graded A, B, C, D, E or F (an A grade is the best). The data for this study are in the file **grades.sav**. To what degree does GCSE maths grade correlate with first-year statistics grade? Ⓐ
- **Task 5:** In Figure 2.3 we saw some data relating to people's ratings of dishonest acts and the likeableness of the perpetrator (for a full description see Jane Superbrain Box 2.1). Compute the Spearman correlation between ratings of dishonesty and likeableness of the perpetrator. The data are in **honesty_lab.sav**.
- **Task 6:** In Chapter 4 (Task 6) we looked at data from people who had fish or cats as pets and measured their life satisfaction and, also, how much they like animals (**pets.sav)**. Is there a significant correlation between life satisfaction and the type of animal the person had as a pet? Ⓑ
- **Task 7:** Repeat the analysis above taking account of animal liking when computing the correlation between life satisfaction and the type of animal the person had as a pet. Ⓑ
- **Task 8**: In Chapter 4 (Task 7) we looked at data based on findings that the number of cups of tea drunk was related to cognitive functioning (Feng et al., 2010). The data are in the file **tea_15.sav**. What is the correlation between tea drinking and cognitive functioning? Is there a significant effect? Ⓐ
- **Task 9**: The research in the previous task was replicated but in a larger sample (N = 716), which is the same as the sample size in Feng et al.'s research (**tea_716.sav**). Conduct a correlation between tea drinking and cognitive functioning. Compare the correlation coefficient and significance in this large sample, with the previous task. What statistical point do the results illustrate? Ⓑ
- **Task 10:** In Chapter 6 we looked at hygiene scores over three days of a rock music festival (**download.sav**). Using Spearman's correlation, were hygiene scores on day 1 of the festival significantly correlated with those on day 3? Ⓐ
- **Task 11:** Using the data in **shopping.sav** (Chapter 4, Task 5), find out if there is a significant relationship between the time spent shopping and the distance covered. Ⓐ
- **Task 12:** What effect does accounting for the participant's sex have on the relationship between the time spent shopping and the distance covered? Ⓑ

View the solutions at www.discoverspss.com





THE LINEAR MODEL (REGRESSION)

9.1 What will this chapter tell me? 412
9.2 The linear model (regression) ... again! 413
9.3 Bias in linear models 422
9.4 Generalizing the model 425
9.5 Sample size and the linear model 429
9.6 Fitting linear models: the general procedure 431
9.7 Using SPSS to fit a linear model with one predictor 432
9.8 Interpreting a linear model with one predictor 433
9.9 The linear model with two or more predictors (multiple regression) 436
9.10 Using SPSS to fit a linear model with several predictors 441
9.11 Interpreting a linear model with several predictors 447
9.12 Robust regression 466
9.13 Bayesian regression 469
9.14 Reporting linear models 471
9.15 Jane and Brian's story 471
9.16 What next? 474
9.17 Key terms that I've discovered 474
Smart Alex's tasks 475

Letters A to D indicate increasing difficulty

9.1 What will this chapter tell me? Ⓐ

Although none of us can know the future, predicting it is so important that organisms are hard-wired to learn about predictable events in their environment. We saw in the previous chapter that I received a guitar for Christmas when I was 8. My first foray into public performance was a weekly talent show at a holiday camp called 'Holimarine' in Wales (it doesn't exist anymore because I am old and this was 1981). I sang a Chuck Berry song called 'My ding-a-ling'[1] and to my absolute amazement I won the competition.[2] Suddenly other 8-year-olds across the land (well, a ballroom in Wales) worshipped me (I made lots of friends after the competition). I had tasted success, it tasted like praline chocolate, and so I wanted to enter the competition in the second week of our holiday. To ensure success, I needed to know why I had won in the first week. One way to do this would have been to collect data and to use these data to predict people's evaluations of children's performances in the contest from certain variables: the age of the performer, what type of performance they gave (singing, telling a joke, magic tricks) and perhaps how cute they looked. Obviously actual talent wouldn't be a factor. A linear model (regression) fitted to these data would enable us to predict the future (success in next week's competition) based on values of the variables we'd measured. If, for example, singing was an important factor in getting a good audience evaluation, I could sing again the following week; but if jokers tended to do better, then I might switch to a comedy routine. When I was 8 I was fortunate enough not to know about linear models (nor did I wish to); however, my dad thought that success was due to the winning combination of a cherub-looking 8-year-old singing songs that can be interpreted in a filthy way. He wrote a song for me to sing about the keyboard player in the Holimarine Band 'messing about with his organ'. He said 'take this song, son, and steal the show' … and that's what I did: I came first again. There's no accounting for taste.



Figure 9.1 Me playing with my ding-a-ling in the Holimarine Talent Show. Note the groupies queuing up at the front

1 It appears that even then I had a passion for lowering the tone.

2 I have a very grainy video of this performance recorded by my dad's friend on a video camera the size of a medium-sized dog that had to be accompanied at all times by a 'battery pack' the size and weight of a tank (see Oditi's Lantern).



Oditi's Lantern
Words that go unspoken, deeds that go undone

'I, Oditi, do not want my followers to get distracted by playing with their ding-a-lings. To warn you all of the dangers of such frivolity, I have uncovered a song, sung by an innocent child, that explains the risks. Stare into my lantern and hear what you can never unhear.'



9.2 The linear model (regression) ... again! Ⓐ

9.2.1 A musical example Ⓑ

Imagine a fictional dream-world where instead of being a not very good academic, I am a statistician employed by a record company. They used to be a thing back in the day. Anyway, in this far-fetched universe, my job depends upon predicting album sales (outcome), which can be made up of physical and digital sales and streaming revenue. If I can predict sales accurately then my boss is happy, and as someone who thrives on external validation, that makes me happy too. Early on in my career, my boss sets me the task of predicting album sales from the amount of money spent advertising the album (predictor). I'm keen to impress.

The data we'll use for this example are in the file **album_sales.csv**. This file has 200 rows, each one representing a different album. There are several columns, one of which contains the sales (in thousands) of each album in the week after release (**sales**), and another contains the amount (in thousands of pounds) spent promoting the album before release (**adverts**). The other columns represent how many times songs from the album were played on a prominent national radio station in the week before release (**airplay**), and how attractive people found the band's image out of 10 (**image**). Ignore these last two variables for now.

9.2.2 The linear model with one predictor Ⓐ

Now, let's look at a model for predicting album sales from advertising. To see this relationship let's create a scatterplot.

SELF TEST

Produce a scatterplot of sales (*y*-axis) against advertising budget (*x*-axis). Include the regression line.

Figure 9.2 shows the resulting plot. A positive relationship exists: the more money spent advertising the album, the more it sells. Of course, there are some albums that sell well regardless of advertising (top left of scatterplot), but there are none that sell badly when advertising levels are high (bottom right of scatterplot).

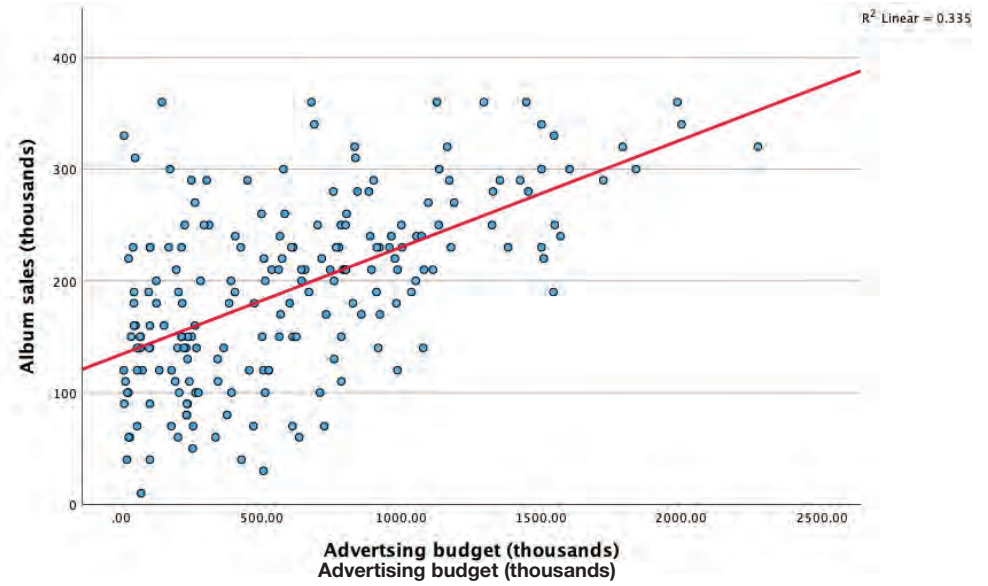


Figure 9.2 Scatterplot showing the relationship between album sales and the amount spent promoting the album

Let's revise some ideas from Chapter 6 (return there for more detail – and by 'detail' I mean 'anguish'). If we want to predict one variable from another we can model this as:

$$\begin{aligned} \text{outcome}_i &= (b_0 + b_1 X_i) + \text{error}_i \\ Y_i &= b_0 + b_1 X_i + \varepsilon_i \end{aligned} \tag{9.1}$$

I mentioned there that this model can be summed up as 'we want to predict the values of an outcome variable (Y_i) from a predictor variable (X_i)'. The bs in the model are known as parameters, and their values represent things in which we're interested. The parameter attached to a predictor variable, b_1, quantifies the relationship it has with the outcome variable, and the parameter b_0 tells us the value of the outcome when the predictor is zero. We also saw that we assume that there are 'true' values of the parameters that we would like to know. In other words, there is a version of this model that quantifies the relationship between the outcome and predictor in the population. However, because we cannot collect data from the entire population we, instead, try to estimate the bs in the model using a sample. When we do this, we specify the model as:

$$Y_i = \hat{b}_0 + \hat{b}_1 X_{1i} + e_i \tag{9.2}$$

The hats remind us that we are estimating values (rather than knowing for certain their true values), and the change in the error term from ε_i to e_i also reminds us that the errors in prediction in our sample model (the residuals) are representative of, but not exactly the same us, the errors in prediction of the population model. These residuals (e_i) are observable (we can know their value) whereas errors in prediction (ε_i) cannot be observed and are not known.

Let's apply this to our example of predicting album sales. We could adapt the (population) linear model (Eq. 9.1) by replacing the predictor and outcome with our variable names:

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{album sales}_i &= b_0 + b_1\ \text{advertising budget}_i + \varepsilon_i \end{aligned} \tag{9.3}$$

We could estimate the b in this model using a sample:

$$\text{album sales}_i = \hat{b}_0 + \hat{b}_1\ \text{advertising budget}_i + e_i \tag{9.4}$$

This equation describes the straight line shown in Figure 9.2. Once we have estimated the values of the bs we would be able to make a prediction about album sales by replacing 'advertising' with a number representing how much we wanted to spend advertising an album. We place a hat over the outcome variable to remind us that it's an estimate, and the error term disappears.

$$\widehat{\text{album sales}}_i = \hat{b}_0 + \hat{b}_1 \text{advertising budget}_i \tag{9.5}$$

For example, imagine that b_0 turned out to be 50 and b_1 turned out to be 100. Our model would be:

$$\widehat{\text{album sales}}_i = 50 + (100 \times \text{advertising budget}_i) \tag{9.6}$$

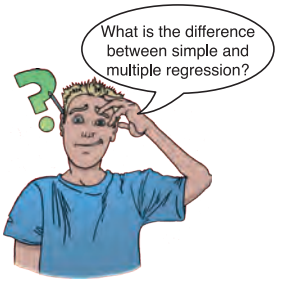
Note that I have replaced the bs with their numeric values. Now, we can make a prediction. Imagine we wanted to spend £5 on advertising. We can replace the variable 'advertising budget' with this value and solve the equation to discover how many album sales we will get:

$$\widehat{\text{album sales}}_i = 50 + (100 \times 5) = 550 \tag{9.7}$$

So, based on our model we can predict that if we spend £5 on advertising, we'll sell 550 albums. This value of 550 album sales is known as a **predicted value**. The error term reminds us that this prediction will probably not be perfectly accurate, and because this error is unobservable, we don't know what value it has.

9.2.3 The linear model with several predictors Ⓑ

Life is usually complicated, and there will be numerous variables that might be related to the outcome that you want to predict. To take our album sales example, variables other than advertising are likely to affect sales. For example, how much someone hears songs from the album on the radio, or the 'look' of the band. One of the beautiful things about the linear model is that it expands to include as many predictors as you like. An additional predictor can be placed in the model and given a b to estimate its relationship to the outcome:



$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i \tag{9.8}$$

All that has changed is the addition of a second predictor (X_2) and an associated parameter (b_2). Again, we can estimate the values of the b by using sample data and fitting this model to it:

$$Y_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} + e_i \tag{9.9}$$

As before, bs get a hat to remind us that they are estimates and the error turns into a residual (because we can observe it in the sample data). To make things more concrete, if we add the number of plays of the band on the radio per week (airplay) to the model in Eq. 9.3, we get:

$$\text{album sales}_i = b_0 + b_1 \text{ advertising budget}_i + b_2 \text{ airplay}_i + \varepsilon_i \tag{9.10}$$

The new model includes a b-value for both predictors (and, of course, the constant, b_0). By estimating the b-values, we can make predictions about album sales based not only on the amount spent on advertising but also on airplay.

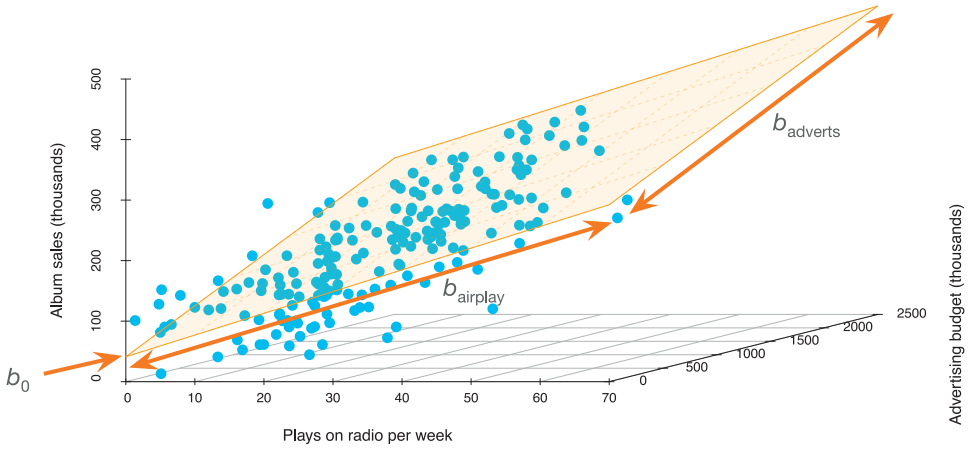


Figure 9.3 Scatterplot of the relationship between album sales, advertising budget and radio play

The resulting model is visualized in Figure 9.3. The tinted trapezium (the regression *plane*) is described by Eq. 9.10 and the dots represent the observed data points. Like a regression line, a regression plane aims to give the best prediction for the observed data. However, there are invariably differences between the model and the real-life data (this fact is evident because most of the dots do not lie exactly on the plane). The vertical distances between the plane and each data point are the errors or *residuals* in the model. The b-value for advertising describes the slope of the left and right sides of the plane, whereas the b-value for airplay describes the slope of the top and bottom of the plane. Just like with one predictor, these two slopes describe the shape of the model (what it looks like) and the intercept locates the model in space.

It is easy enough to visualize a linear model with two predictors, because it is possible to plot the plane using a 3-D scatterplot. However, with three, four or even more predictors you can't immediately visualize what the models looks like, or what the b-values represent, but you can apply the principles of these simpler models to more complex scenarios. For example, in general, we can add as many predictors as we like, provided we give them a b, and the linear model expands accordingly:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni} + \varepsilon_i \tag{9.11}$$

Y is the outcome variable, b_1 is the coefficient of the first predictor (X_{1i}), b_2 is the coefficient of the second predictor (X_{2i}), b_n is the coefficient of the nth predictor (X_{ni}) and ε_i is the error for the ith entity. This equation illustrates that we can add predictors to the model until we reach the final one (X_{ni}), and each time we add one, we assign it a regression coefficient (b).

To sum up, we can fit a linear model to predict values of an **outcome variable** (a.k.a. dependent variable) from one or more **predictor variables** (a.k.a. independent variables). With one predictor variable, the technique is sometimes referred to as **simple regression** but with several predictors it is called **multiple regression**. Both are terms for the linear model.

9.2.4 Assessing the goodness of fit, sums of squares, R and R^2 Ⓐ

We have seen that the linear model is a versatile model for summarizing the relationship between one or more predictor variables and an outcome variable. No matter how many predictors we have, the model can be described entirely by a constant (b_0) and by parameters associated with each predictor (bs). In Section 6.2.4 we met the idea of a residual, which is basically the error in prediction for a given case. Remember that it is the difference between the value of the outcome variable observed for a given case and the value predicted by the model for that case:

$$\begin{aligned} \text{residual}_i &= \text{observed}_i - \text{model}_i \\ e_i &= Y_i - \hat{Y}_i \end{aligned} \tag{9.12}$$

Then, in Section 6.2.5 we discovered that we can measure the total error:

$$\text{Total error} = \sum_{i=1}^{n} (\text{observed}_i - \text{model}_i)^2 \tag{9.13}$$

in which 'observed' are the observed values of the outcome (Y_i). When quantifying the error in prediction (the residual sum of squares, RSS or SS_R) the 'model' in the equation becomes the predicted values ($\hat{Y}_i$), and we get:

$$SS_R = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{9.14}$$

Model parameters are typically estimated using the method of least squares, which finds the b-values that minimizes this sum of squared residuals. With one predictor a little bearded wizard called Nephwick the Line Finder magically finds lines of best fit. Yes, he lives inside your computer. For more complex models, Nephwick invites his brother Clungglewad the Beta Seeker for tea and cake and together they stare into the tea leaves in their cup until the optimal beta values are revealed to them. Then they compare beard growth since their last meeting. I'm pretty sure that's how the method of least squares works.

Once Nephwick and Clungglewad have found the values of b that define the model of best fit we assess how well this model fits the observed data (i.e., the **goodness of fit**). We do this because even though the model is the best one available, it can still be a lousy fit (the best of a bad bunch). We saw above that the residual sum of squares measures how much error there is in the model: it quantifies the error in prediction, but it doesn't tell us whether using the model is better than nothing. We need to compare the model against a baseline to see whether it 'improves' how well we can predict the outcome. So, we fit a baseline model and calculate the error, SS_R, within it (using Eq. 9.14) and then fit the 'best' model and do the same. If the best model is any good, it should have significantly less error within it than the baseline model.

What would be a good baseline model? Let's go back to our example of predicting album sales (Y) from the amount of money spent advertising that album (X). One day my boss bursts into my office.



He says, 'Andy, I know you wanted to be a rock star but have ended up working as my stats-monkey, but how many albums will we sell if we spend £100,000 on advertising?' If I didn't have an accurate model of the relationship between album sales and advertising, what would my best guess be? Probably the best answer I could give would be the mean number of album sales (say, 200,000) because – on average – that's how many albums we expect to sell. This response might satisfy a clueless record company executive (so clueless that they passed up the opportunity to offer my band a record contract). The next day he bursts in again and demands to know how many albums we will sell if we spend £1 on advertising. In the absence of any better information, I'm again going to have to say the average number of sales (200,000). This is getting embarrassing for me: whatever amount of money is spent on advertising I predict the same levels of sales. My boss will think I'm being ridiculous. He might have a point, but statistically speaking he's assuming that there is a relationship between advertising budget and sales, whereas the model I've fitted assumes that there isn't.

The mean of the outcome is a model of 'no relationship' between the variables: as one variable changes the prediction for the other remains constant. I hope this illustrates that the mean of the outcome is a good baseline of 'no relationship'. We again use Eq. 9.13 but replace 'model' with the mean of the outcome variable:

$$SS_T = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \tag{9.15}$$

Basically, we calculate the difference between the observed values and the mean outcome, we square these differences and then add them. This sum of squared differences is known as the **total sum of squares** (denoted by SS_T) and it represents how good the mean is as a model of the observed outcome scores (Figure 9.4, top left).

We then fit a more sophisticated model to the data, such as a linear model, and again work out the differences between what this new model predicts and the observed data. This value is the residual sum of squares (SS_R) that we have already discussed (here and in Section 6.2.5 where we labelled it RSS). It represents the degree of inaccuracy when the best model is fitted to the data (Figure 9.4, top right).

We can use the values of SS_T and SS_R to calculate how much better the linear model is than the baseline model of 'no relationship'. The improvement in prediction resulting from using the linear model rather than the mean is calculated as the difference between SS_T and SS_R (Figure 9.4, bottom). We can again use Eq. 9.13 but the 'observed' values are now the values predicted from the best fit model ($\hat{Y}_i$) and the 'model' is the mean of the outcome variable:

$$SS_M = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = SS_T - SS_R \tag{9.16}$$

This difference shows us the reduction in the inaccuracy of the model resulting from fitting the regression model to the data. This improvement is the **model sum of squares** (SS_M). Figure 9.4 shows each sum of squares graphically where the model is a line (i.e., one predictor) but the same principles apply with more than one predictor.

If the value of SS_M is large, the linear model is very different from using the mean to predict the outcome variable. This implies that the linear model has made a big improvement to predicting the outcome variable. If SS_M is small then using the linear model is little better than using the mean (i.e., the best model is no better than predicting from 'no relationship'). A useful measure arising from these sums of squares is the proportion of improvement due to the model. This is calculated by dividing the sum of squares for the model by the total sum of squares to give a quantity called R^2:

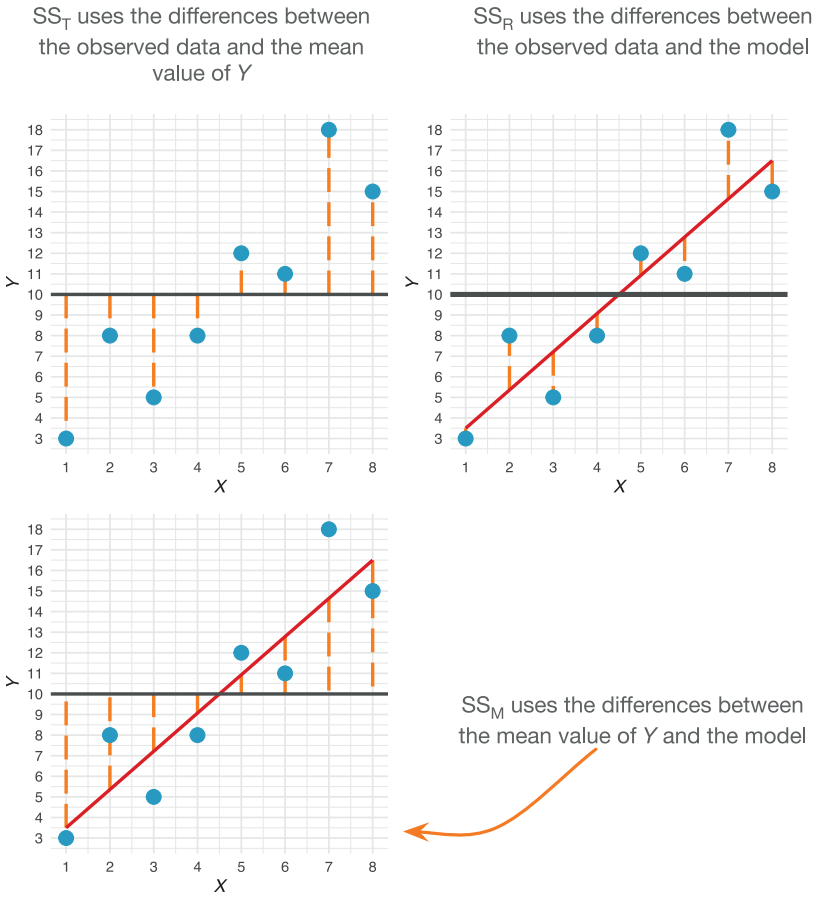


Figure 9.4 Diagram showing from where the sums of squares derive

$$R^2 = \frac{SS_M}{SS_T} \tag{9.17}$$

To express this value as a percentage multiply it by 100. R^2 represents the amount of variance in the outcome explained by the model (SS_M) relative to how much variation there was to explain in the first place (SS_T); it is the same as the R^2 we met in Section 8.4.2 and it is interpreted in the same way: as a percentage, it represents the percentage of the variation in the outcome that can be predicted from the model. We can take the square root of this value to obtain Pearson's correlation coefficient for the relationship between the values of the outcome predicted by the model and the observed values of the outcome.[3] As such, the correlation coefficient for the predicted and observed values provides a good estimate of the overall fit of the linear model, and R^2 is a gauge of the substantive size of the model fit.[4]

3 With only one predictor in the model this value will be the same as the Pearson correlation coefficient between the predictor and outcome variable.

4 When the model contains more than one predictor people sometimes refer to R^2 as multiple R^2. This is another example of how people attempt to make statistics more confusing than it needs to be by referring to the same thing in different ways. The meaning and interpretation of R^2 is the same regardless of how many predictors you have in the model or whether you choose to call it multiple R^2: it is the squared correlation between values of the outcome predicted by the model and the values observed in the data.

One issue with R^2 is that it increases as a function of the number of predictors because unless a predictor has no relationship whatsoever with the outcome its inclusion will increase the variance in the outcome explained by the model (SS_M) – even if only a little – which in turn will increase R^2. The solution is **adjusted R^2**, given by:

$$\text{adjusted } R^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \tag{9.18}$$

which adjusts R^2 for the sample size, n, and the number of predictors, k. Although sometimes discussed in a cross-validation context, adjusted R^2 tells us nothing about how well the model would predict scores of a different sample of data from the same population; that isn't really what it sets out to do. However, by adjusting for the number of predictors, it gives us a better estimate (than the unadjusted value) of the R^2 in the population. For this reason it is said to indicate the loss of predictive power or **shrinkage**.

A second use of the sums of squares in assessing the model is the F-test. I mentioned way back in Chapter 2 that test statistics (like F) are usually the amount of systematic variance divided by the amount of unsystematic variance, or, put another way, the model compared to the error in the model. This is true here: F is based upon the ratio of the improvement due to the model (SS_M) and the error in the model (SS_R). I say 'based upon' because the sums of squares depend on the number of differences that were added up, and so the average sums of squares (referred to as the **mean squares** or MS) are used to compute F. The mean sum of squares is the sum of squares divided by the associated degrees of freedom (this is comparable to calculating the variance from the sums of squares – see Section 2.6.2). For SS_M the degrees of freedom are the number of predictors in the model (k), and for SS_R, as we saw in Section 6.2.5 they are $n - p$ (where n is the sample size and p is the number of parameters (bs) estimated). Another way of expressing these degrees of freedom is in terms of the number of predictors because we typically estimate a b for each predictor and the intercept (b_0), so $p = k + 1$, giving us degrees of freedom of $n - (k + 1)$ or, equivalently, $n - k - 1$ (Eq. 9.19). So we have:

$$\begin{aligned} MS_M &= \frac{SS_M}{k} \\ MS_R &= \frac{SS_R}{n-k-1} \end{aligned} \tag{9.19}$$

There is more on mean squares in Chapter 12. The ***F*-statistic** computed from these mean squares,

$$F = \frac{MS_M}{MS_R} \tag{9.20}$$

is a measure of how much the model has improved the prediction of the outcome compared to the level of inaccuracy of the model. If a model is good, then the improvement in prediction from using the model should be large (MS_M will be large) and the difference between the model and the observed data should be small (MS_R will be small). In short, for a good model the top of Eq. 9.20 will be bigger than the bottom, resulting in a large F-statistic (greater than 1 at least).

This F has an associated probability distribution from which a p-value can be derived to tell us the probability of getting an F at least as big as the one we have if the null hypothesis were true. The null hypothesis in this case is a flat model (predicted values of the outcome are the same regardless of the value of the predictors). If you want to go old school, you can compare the F-statistic against critical values for the corresponding degrees of freedom (as in the Appendix).

The F-statistic is also used to calculate the significance of R^2 using the following equation:

$$F = \frac{(n-k-1)R^2}{k(1-R^2)} \tag{9.21}$$

in which n is the number of cases or participants, and k is the number of predictors in the model. This F tests the null hypothesis that R^2 is zero (i.e., there is no improvement in the sum of squared error due to fitting the model).

9.2.5 Assessing individual predictors Ⓐ

We discovered in Section 6.2.6 that we can use model parameters to conduct significance tests. That was a fairly technical section of the book, so let's have a recap of the key ideas. When we do a hypothesis test we compare a null hypothesis (there is no effect) with an alternative (there is an effect). Any predictor in a linear model has a coefficient (b_1). The value of b represents the change in the outcome resulting from a unit change in a predictor, so it is a measure of effect.



If a predictor had 'no effect', how would the outcome change as the predictor changes? What would be the corresponding value of b? What, therefore, would be a reasonable null hypothesis?

Under the null hypothesis that there is 'no relationship' or 'no effect' between a predictor and an outcome, as the predictor changes we would expect the predicted value of the outcome *not* to change (it is a constant value). In other words, the regression line would be flat like we saw in the panel representing SS_T in Figure 9.4. Therefore, 'no effect' equates to 'flat line' and 'flat line' equates to $b = 0$. If we want to do a hypothesis test then we can compare the 'alternative hypothesis' that there is an effect against this null. When there is an effect, the model will not be flat and b will not equal zero. So, we get the following hypotheses:

- $H_0 : b = 0$. If there is no association between a predictor and outcome variable we'd expect the parameter for that predictor to be zero.
- $H_1 : b \neq 0$. If there is a non-zero association between a predictor and outcome variable we'd expect the parameter for that predictor to be non-zero.

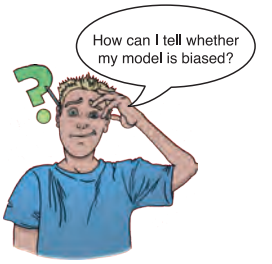
Specifically, from Eq. 6.21 (reproduced in Eq. 9.22) we test this hypothesis using a ***t*-statistic**:

$$t_{n-p} = \frac{\hat{b} - \hat{b}_{\text{Expected}}}{\widehat{SE}\left(\hat{b}\right)} = \frac{\hat{b}}{\widehat{SE}\left(\hat{b}\right)} \tag{9.22}$$

Remember that the top half of Eq. 9.22 asks 'is the parameter estimate ($\hat{b}$) different from the expected value of the parameter if the null hypothesis were true ($\hat{b}_{\text{Expected}}$)?' If the null is true, we expect the parameter value to be 0 ($\hat{b}_{\text{Expected}} = 0$) so the top half of the ratio reduces to the value of the observed $\hat{b}$. The bottom half represents the error in measuring this parameter estimate, which is the degree to which estimates vary across samples, which is quantified by its standard error. Therefore, like F, the t-statistic is a signal-to-noise ratio, it is the size of effect relative to how much variability there is in this effect across samples. When the standard error is small, even a small deviation from zero can reflect a significant difference because b is representative of the range of possible samples.

The statistic t has a probability distribution that differs according to the degrees of freedom for the test. In this context, the degrees of freedom are $n - p$ (remember that n is the sample size and p is the number of bs estimated), or equivalently $n - k - 1$ (where k is the number of predictors). With only one predictor, this expression reduces to $n - 2$. Using the appropriate t-distribution, it's possible to calculate a p-value that indicates the probability of getting a t at least as large as the one we observed if the null hypothesis were true (i.e., if b was in fact 0 in the population). If this observed p-value is less than or equal to 0.05, then scientists choose to assume that b is significantly different from 0; put another way, the predictor makes a significant contribution to predicting the outcome. However, remember the potential pitfalls of blindly applying this 0.05 rule. If you want to pretend it's 1935 then instead of computing an exact p, you can compare your observed t against critical values in a table (in the Appendix).

9.3 Bias in linear models Ⓑ



We spent all of Chapter 6 discovering how statistical models can be biased by unusual cases or by failing to meet certain assumptions. The next questions to ask are whether the model: (1) is influenced by a small number of cases; and (2) generalizes to other samples. These questions are, in some sense, hierarchical because we wouldn't want to generalize a bad model. However, it is a mistake to think that because a model is a good fit of the observed data we can draw conclusions beyond our sample. **Generalization** (Section 9.4) is a critical additional step and if we find that our model is not generalizable, then we must restrict any conclusions to the sample used. First, let's look at bias. To answer the question of whether the model is influenced by a small number of cases, we can look for outliers and influential cases (the difference is explained in Jane Superbrain Box 9.1).

9.3.1 Outliers Ⓑ

We saw in Chapter 6 that outliers, cases that differ substantially from the main trend in the data, can bias a model. We also saw that residuals represent the error present in the model. If a model fits the sample data well then all residuals will be small (if the model was a perfect fit of the sample data – all data points fall on the regression line – then all residuals would be zero). If a model is a poor fit to the sample data then the residuals will be large. Up to now we have discussed *normal* or **unstandardized residuals**. These are the raw differences between predicted and observed values of the outcome variable. They are measured in the same units as the outcome variable, which makes it difficult to apply general rules (because what constitutes 'large' depends on the outcome variable). All we can do is to look for residuals that stand out as being particularly large.

To overcome this problem, we can use **standardized residuals**, which are the residuals converted to z-scores and so are expressed in standard deviation units. Regardless of the variables in your model, standardized residuals (like any z-scores) are distributed around a mean of 0 with a standard deviation of 1. We saw in Jane Superbrain Box 6.6 how to use what we know about z-scores to apply universal guidelines to standardized residuals. To recap: (1) standardized residuals with an absolute value greater than 3.29 (we can use 3 as an approximation) are cause for concern because in an average sample a value this high is unlikely to occur; (2) if more than 1% of our sample cases have standardized residuals with an absolute value greater than 2.58 (2.5 will do) there is evidence that the level of error within our model may be unacceptable; and (3) if more than 5% of cases have standardized residuals with an absolute value greater than 1.96 (2 for convenience) then the model may be a poor representation of the data.

A third form of residual is the **Studentized residual**, which is the unstandardized residual divided by an estimate of its standard deviation that varies point by point. These residuals have the same properties as the standardized residuals but usually provide a more precise estimate of the error variance of a specific case.

9.3.2 Influential cases ©

It is also possible to look at whether certain cases exert undue influence over the parameter estimates. In other words, if we were to delete a certain case, how different would the $\hat{b}$s be? This analysis helps to determine whether the model is stable across the sample, or whether it is biased by a few influential cases. This process can also unveil outliers.

There are several statistics used to assess the influence of a case. The **adjusted predicted value** for a case is the predicted value of the outcome for that case from a model in which the case is excluded. In effect, you estimate the model parameters excluding a particular case and use this new model to predict the outcome for the case that was excluded. If a case does not exert a large influence over the $\hat{b}$s then the adjusted predicted value should be similar to the predicted value when the case is included. Put simply, if the model is stable then the predicted value of a case should be the same regardless of whether that case was used to estimate the model.

We can also look at the **deleted residual**, which is the difference between the adjusted predicted value and the original observed value. The deleted residual can be divided by the standard error to give a standardized value known as the **Studentized deleted residual**. This residual can be compared across different regression analyses because it is measured in standard units.

The deleted residuals are very useful to assess the influence of a case on the ability of the model to predict that case. However, they do not provide any information about how a case influences the model as a whole (i.e., the impact that a case has on the model's ability to predict *all* cases). **Cook's distance** is a measure of the overall influence of a case on the model, and Cook and Weisberg (1982) have suggested that values greater than 1 may be cause for concern.

The **leverage** (sometimes called **hat values**) gauges the influence of the observed value of the outcome variable over the predicted values. The average leverage value is defined as $(k+1)/n$, in which k is the number of predictors in the model and n is the number of cases.[5] The leverage can vary from 0 (no influence) to 1 (the case has complete influence over prediction).

- If no cases exert undue influence over the model then all leverage values should be close to the average value of $(k + 1)/n$.
- We should investigate cases with values greater than twice the average, $2(k + 1)/n$ (Hoaglin & Welsch, 1978).

We will see how to use these cut-off points later. However, cases with large leverage values will not necessarily have a large influence on the regression coefficients because they are measured on the outcome variables, not the predictors.

Related to the leverage values are the **Mahalanobis distances** (Mahalanobis, 1936), which measure the distance of cases from the mean(s) of the predictor variable(s). Look for the cases with the highest values. These distances have a chi-square distribution, with degrees of freedom equal to the number of predictors (Rao, 1945, 1948). With the usual caveats about using significance tests to determine bias, one way to establish a cut-off point is to find the critical value of chi-square for the desired alpha level (values for $p = 0.05$ and 0.01 are in the Appendix). For example, with three predictors, a distance greater than 7.81 ($p = 0.05$) or 11.34 ($p = 0.01$) would be cause for concern. As general context,

5 You may come across the average leverage denoted as p/n in which p is the number of parameters being estimated. I have mentioned a few times that we estimate parameters for each predictor and the intercept, so $p = k + 1$.

based on Barnett and Lewis (1978), with large samples (N = 500) and five predictors, values above 25 are cause for concern. In smaller samples (N = 100) and fewer predictors (namely, three), values greater than 15 are problematic. In very small samples (N = 30) with only two predictors, values greater than 11 should be examined.

Another approach is to look at how the estimates of b in a model change as a result of excluding a case (i.e., compare the values of b estimated from the full data to those estimated from the data excluding the particular case). The change in bs tells us how much influence a case has on the parameters of the model. To take a hypothetical example, imagine two variables that have a perfect negative relationship except for a single case (case 30). These data are in the file **df_beta.csv.**



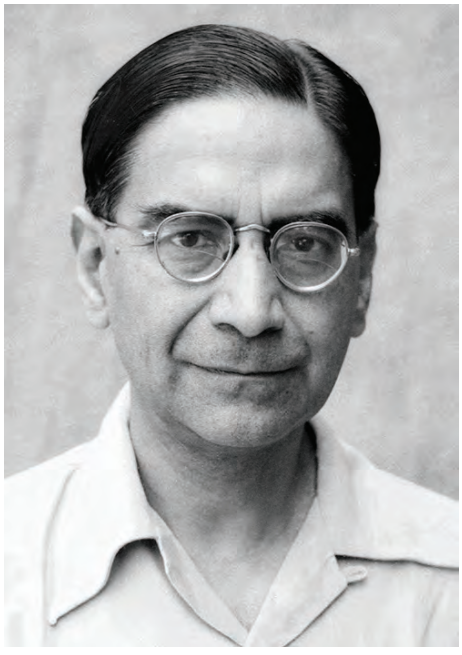


Figure 9.5 Prasanta Chandra Mahalanobis[6]

The results of these two models are summarized in Table 9.1, which shows: (1) the parameters for the linear model when the extreme case is included or excluded; (2) the resulting model equations; and (3) the value of Y predicted from participant 30's score on the X variable (which is obtained by replacing the X in the model equation with participant 30's score for X, which was 1). When case 30 is excluded, these variables have a perfect negative relationship; hence the coefficient for the predictor (b_1) is –1, and the coefficient for the constant (the intercept, b_0) is 31. However, when case 30 is included, both $\hat{b}$s are reduced[7] and this change in their value is in the last column. The difference between a parameter estimated using all cases and estimated when one case is excluded is known as the **DFBeta**. DFBeta is calculated for every case and for each of the parameters in the model. So, in our hypothetical example, the DFBeta for the constant is –2, and the DFBeta for the predictor variable is 0.1. The values of DFBeta help us to identify cases that have a large influence on the parameters of the model. The units of measurement used will affect these values and so you can use **standardized DFBeta** to apply universal cut-off points. Standardized DFBetas with absolute values above 1 indicate cases that substantially influence the model parameters (although Stevens, 2002, suggests looking at cases with absolute values greater than 2).

6 Photo reproduced by kind permission of The Prasanta Chandra Mahalanobis Memorial Museum and Archives (www.isical.ac.in/~museum/).

7 The value of b_1 is reduced because the variables no longer have a perfect linear relationship and so there is now variance that the predictor cannot explain.

Table 9.1 The difference in the parameters of the linear model when one case is excluded

Parameter estimate ($\hat{b}$)	Case 30 Included	Case 30 Excluded	Difference
Constant ($\hat{b}_0$)	29.00	31.00	−2.00
Predictor ($\hat{b}_1$)	−0.90	−1.00	0.10
Model	$\hat{Y}_i = -0.9X_i + 29$	$\hat{Y}_i = -1X_i + 31$	
Value of predictor for case 30 (X_{30})	1	1	
Predicted value for case 30 ($\hat{Y}_{30}$)	28.10	30.00	−1.90

A related statistic is the **DFFit**, which is the difference between the predicted value for a case when the model is estimated including or excluding that case: in this example the value is –1.90 (see Table 9.1). If a case has no influence then its DFFit should be zero – hence, we expect non-influential cases to have small DFFit values. As with DFBeta, this statistic depends on the units of measurement of the outcome, and so a DFFit of 0.5 will be very small if the outcome ranges from 1 to 100, but very large if the outcome varies from 0 to 1. The solution is to use standardized versions of the DFFit values (**standardized DFFit**), which are expressed in standard deviation units.

A final measure is the **covariance ratio (CVR)**, which quantifies the degree to which a case influences the variance of the regression parameters. A description of the computation of this statistic leaves me dazed and confused, so suffice to say that when this ratio is close to 1 the case is having very little influence on the variances of the model parameters. Belsey et al. (1980) recommend the following:

- If $CVR_i > 1 + [3(k + 1)/n]$ then deleting the ith case will damage the precision of some of the model's parameters.
- If $CVR_i < 1 - [3(k + 1)/n]$ then deleting the ith case will improve the precision of some of the model's parameters.

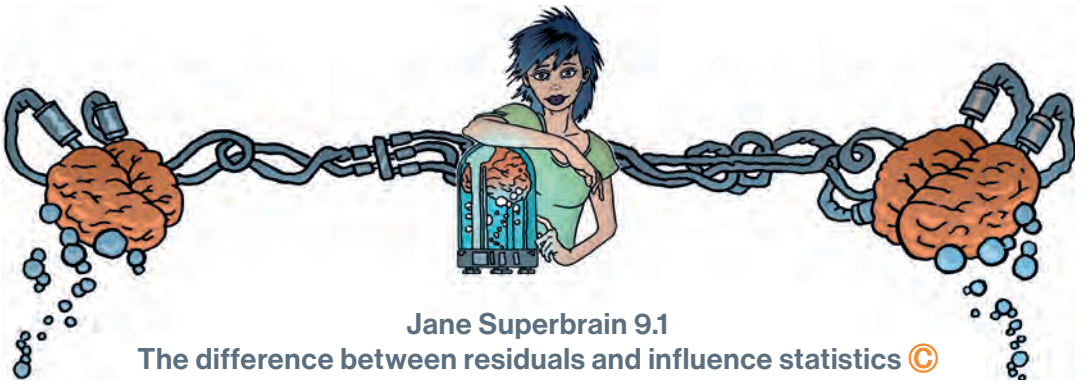
In both inequalities, k is the number of predictors, CVR_i is the covariance ratio for the ith participant and n is the sample size.

9.3.3 A final comment on diagnostic statistics Ⓑ

I'll conclude this section with a point made by Belsey et al. (1980): diagnostics are tools to see how well your model fits the sampled data and *not* a way of justifying the removal of data points to effect some desirable change in the regression parameters (e.g., deleting a case that changes a non-significant b-value into a significant one). Stevens (2002) similarly notes that if a case is a significant outlier but is not having an influence (e.g., Cook's distance is less than 1, DFBetas and DFFit are small) there is no real need to worry about that point because it's not having a large impact on the model parameters. Nevertheless, you should still be interested in *why* the case didn't fit the model.

9.4 Generalizing the model Ⓑ

The linear model produces an equation that is correct for the sample of observed values. However, we are usually interested in generalizing our findings beyond the sample. For a linear model to generalize the underlying assumptions must be met, and to test whether the model does generalize we can cross-validate it.



Jane Superbrain 9.1
The difference between residuals and influence statistics ©

To illustrate how residuals and influence statistics differ, imagine that the Mayor of London in 1900 was interested in how drinking alcohol affected mortality. London is divided up into different regions called boroughs, and so he measured the number of pubs and the number of deaths over a period of time in eight of his boroughs. The data are in a file called **pubs.csv**.

The scatterplot of these data (Figure 9.6) reveals that without the last case there is a perfect linear relationship (the orange line). However, the presence of the last case (case 8) changes the line of best fit dramatically (although this line is still a significant fit to the data – fit the model and see for yourself).

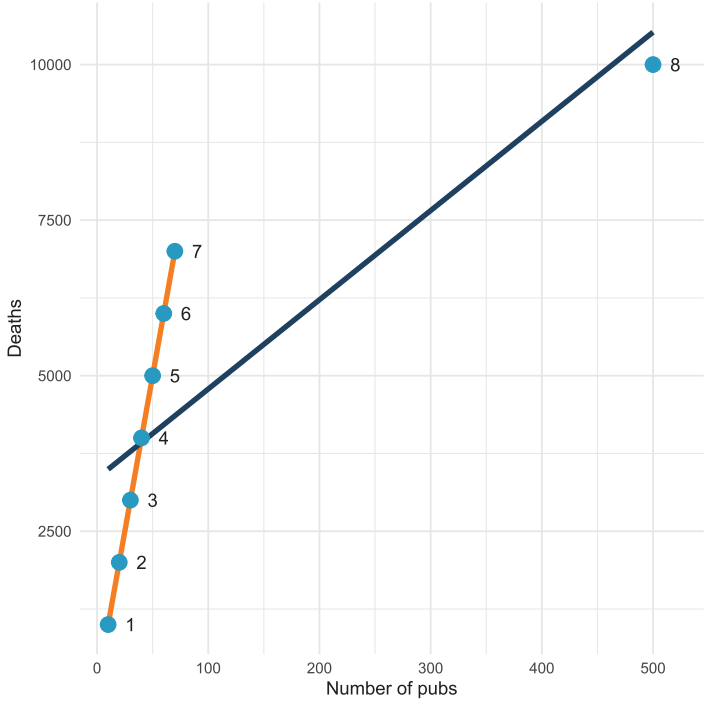


Figure 9.6 Relationship between the number of pubs and the number of deaths in eight London districts

The residuals and influence statistics are interesting (Output 9.1). The residual for case 8 is the second *smallest*: it produces a relatively small residual (all but one of the non-outliers have larger residuals) because it sits very close to the line that has been fitted to the data. According to the residual it is not an outlier, but how is that possible when it is so different from the rest of the data? The answer lies in the influence statistics, which are all massive for case 8: it exerts a huge influence over the model (the leverage is almost at its maximum value of 1), so huge that the model predicts that case quite well.

Case Summaries[a]

	Standardized Residual	Mahalanobis Distance	Cook's Distance	Centered Leverage Value	DFFIT	DFBETA Intercept	DFBETA pubs
1	−1.33839	.28515	.21328	.04074	−495.72692	−509.65184	1.39249
2	−.87895	.22370	.08530	.03196	−305.09716	−321.12768	.80153
3	−.41950	.16969	.01814	.02424	−137.20167	−147.10661	.33016
4	.03995	.12314	.00015	.01759	12.38769	13.45081	−.02658
5	.49940	.08403	.02294	.01200	147.81622	161.44976	−.27267
6	.95885	.05237	.08092	.00748	273.00807	297.67748	−.41116
7	1.41830	.02817	.17107	.00402	391.72124	422.81664	−.44422
8	−.27966	6.03375	227.14286	.86196	−39478.5847	3351.95531	−85.66108
Total N	8	8	8	8	8	8	8

a. Limited to first 100 cases.

Output 9.1

When you see a statistical oddity like this, ask what's happening in the real world. District 8 is the City of London, a tiny area of only 1 square mile in the centre of London where very few people lived but where thousands of commuters (even then) came to work and wanted pubs. Therefore, there was a massive number of pubs. (I'm very grateful to David Hitchin for this example, and he in turn got it from Dr Richard Roberts.)



9.4.1 Assumptions of the linear model Ⓑ

We have already looked at the main assumptions of the linear model and how to assess them in Chapter 6. To recap, the main ones in order of importance (Field & Wilcox, 2017; Gelman & Hill, 2007) are:

- *Additivity and linearity* (Section 6.6): The outcome variable should, in reality, be linearly related to any predictors, and if there are several predictors, their combined effect is best described by adding their effects together. In other words, the process we're trying to model can be described by the linear model. If this assumption isn't met then the model is invalid. You can sometimes transform variables to make their relationships linear.
- *Independent errors* (Section 6.7.3): For any two observations the errors should be uncorrelated (i.e., independent). This eventuality is sometimes described as a lack of **autocorrelation**. If we violate this assumption then the model standard errors will be invalid, as will the confidence intervals and significance tests based upon them. In terms of the model parameters themselves, the estimates from the method of least squares will be valid but not optimal (see Section 6.7.3). This assumption can be tested with the **Durbin–Watson test**, which tests for serial correlations between errors. Specifically, it tests whether adjacent residuals are correlated. As such it is affected by the order of cases and only makes sense when your cases have a meaningful order (which they don't in the album sales example). The test statistic varies between 0 and 4 with a value of 2, meaning that the residuals are uncorrelated. A value greater than 2 indicates a negative correlation between adjacent residuals, whereas a value below 2

indicates a positive correlation. The size of the Durbin–Watson statistic depends upon the number of predictors in the model and the number of observations. If this test is relevant to you look up the critical values in Durbin and Watson (1951). As a very conservative rule of thumb, values less than 1 or greater than 3 are cause for concern.

- *Homoscedasticity* (see Section 6.7): At each level of the predictor variable(s), the variance of the model errors should be constant. This assumption means that the errors at each level of the predictor(s) should have the same variance (**homoscedasticity**); when the variances are very unequal there is said to be **heteroscedasticity**. Violating this assumption invalidates confidence intervals and significance tests; estimates of the model parameters (*b*) using the method of least squares are valid but not optimal. This problem is overcome using weighted least squares regression in which each case is weighted by a function of its variance, or using robust regression.
- *Normally distributed errors* (see Section 6.7): It can be helpful if the errors in the model are random, normally distributed variables with a mean of 0. This assumption means that the differences between the predicted and observed data are most frequently zero or very close to zero, and that differences much greater than zero happen only occasionally. Some people confuse this assumption with the idea that predictors have to be normally distributed, which they don't. In small samples a lack of normality invalidates confidence intervals and significance tests, whereas in large samples it will not because of the central limit theorem. If you are concerned only with estimating the model parameters (and not significance tests and confidence intervals) then this assumption barely matters. If you bootstrap confidence intervals then you can ignore this assumption.

There are other considerations that we haven't touched on (see Berry, 1993):

- *Predictors are uncorrelated with 'external variables'*: *External variables* are variables that haven't been included in the model and that influence the outcome variable.[8] These variables are like the 'third variable' that we discussed in the Chapter 8. This assumption means that there should be no external variables that correlate with any of the variables included in the regression model. Obviously, if external variables do correlate with the predictors, then the conclusions we draw from the model become unreliable (because other variables exist that can predict the outcome just as well).
- *Variable types*: All predictor variables must be quantitative or categorical (with two categories), and the outcome variable must be quantitative, continuous and unbounded. By 'quantitative' I mean that they should be measured at the interval level and by 'unbounded' I mean that there should be no constraints on the variability of the outcome. If the outcome is a measure ranging from 1 to 10 yet the data collected vary between 3 and 7, then these data are constrained.
- *No perfect* **multicollinearity**: If your model has more than one predictor then there should be no perfect linear relationship between two or more of the predictors. So, the predictor variables should not correlate too highly (see Section 9.9.3).
- *Non-zero variance*: The predictors should have some variation in value (i.e., they should not have variances of 0). This is self-evident really.

As we saw in Chapter 6 violating these assumptions has implications mainly for significance tests and confidence intervals; the estimates of *b*s are not dependent on these assumptions (although least squares methods will be optimal when the assumptions are met). However, the 95% confidence

8 Some authors refer to these external variables as part of an error term that includes any random factor in the way in which the outcome varies. However, to avoid confusion with the residual terms in the linear model equations I have chosen the label 'external variables'. Although this term implicitly washes over any random factors, I acknowledge their presence.

interval for a *b* tells us the boundaries within which the population values of that *b* are likely to fall.[9] Therefore, if confidence intervals are inaccurate (as they are when these assumptions are broken) we cannot accurately estimate the likely population value. In other words, we can't generalize our model to the population. When the assumptions are met then *on average* the regression model from the sample is the same as the population model. However, you should be clear that even when the assumptions are met, it is possible that a model obtained from a sample is not the same as the population model – but the likelihood of their being the same is increased.

9.4.2 Cross-validation of the model ©

Even if we can't be confident that the model derived from our sample accurately represents the population, we can assess how well our model might predict the outcome in a different sample. Assessing the accuracy of a model across different samples is known as **cross-validation**. If a model can be generalized, then it must be capable of accurately predicting the same outcome variable from the same set of predictors in a different group of people. If the model is applied to a different sample and there is a severe drop in its predictive power, then the model does *not* generalize. First, we should collect enough data to obtain a reliable model (see the next section). Once we have estimated the model we can cross-validate using data splitting. This approach involves randomly splitting your sample data, estimating the model in both halves of the data and comparing the resulting models. When using stepwise methods (see Section 9.9.1), cross-validation is particularly important; you should run the stepwise regression on a random selection of about 80% of your cases. Then force this model on the remaining 20% of the data. By comparing values of the R^2 and *b* in the two samples you can tell how well the original model generalizes. If you're interested in indexes of cross-validation then Browne (2000) reviews some options.

9.5 Sample size and the linear model ©

In the previous section I said that it's important to collect enough data to obtain a reliable model. Also, larger samples enable us to assume that our *b*s have normal sampling distributions because of the central limit theorem (Section 6.2.7). Well, how much is enough?



You'll find a lot of rules of thumb floating about, the two most common being that you should have 10 cases of data for each predictor in the model, or 15 cases of data per predictor. These rules are very pervasive but they oversimplify the issue to the point of being useless. The sample size required depends on the size of effect that we're trying to detect (i.e., how strong the relationship is that we're trying to measure) and how much power we want to detect these effects. The simplest rule of thumb is that the bigger the sample size, the better: the estimate of *R* that we get from regression is dependent on the number of predictors, *k*, and the sample size, *N*. In fact, the expected *R* for random data is $k/(N-1)$ and so with small sample sizes random data can appear to show a strong effect: for example, with six predictors and 21 cases of data, $R = 6/(21-1) = 0.3$ (a medium effect size by Cohen's criteria described in Section 3.7.4). Obviously for random data we'd want the expected *R* to be 0 (no effect), and for this to be true we need large samples (to take the previous example, if we had 100 cases rather than 21, then the expected *R* would be a more acceptable 0.06).

9 Assuming your sample is one of the 95% that generates a confidence interval containing the population value. Yes, I do have to keep making this point – it's important.

Figure 9.7 shows the sample size required[10] to achieve a high level of power (I've taken Cohen's, 1988, benchmark of 0.8) to detect that R^2 is not equal to zero (i.e. the model fits significantly overall) when that is true in the population. I've varied the number of predictors and the size of expected effect: I used $R^2 = 0.02$ (small), 0.13 (medium) and 0.26 (large), which correspond to benchmarks in Cohen (1988). Broadly speaking, if your aim is to test the overall fit of the model: (1) if you expect to find a large effect then a sample size of 77 will always suffice (with up to 20 predictors) and if there are fewer predictors then you can afford to have a smaller sample; (2) if you're expecting a medium effect, then a sample size of 160 will always suffice (with up to 20 predictors), you should always have a sample size above 55, and with six or fewer predictors you'll be fine with a sample of 100; and (3) if you're expecting a small effect size then just don't bother unless you have the time and resources to collect hundreds of cases of data. Miles and Shevlin (2001) produce more detailed graphs that are worth a look, but the take-home message is that if you're looking for medium to large effects sample sizes don't need to be massive, regardless of how many predictors you have.

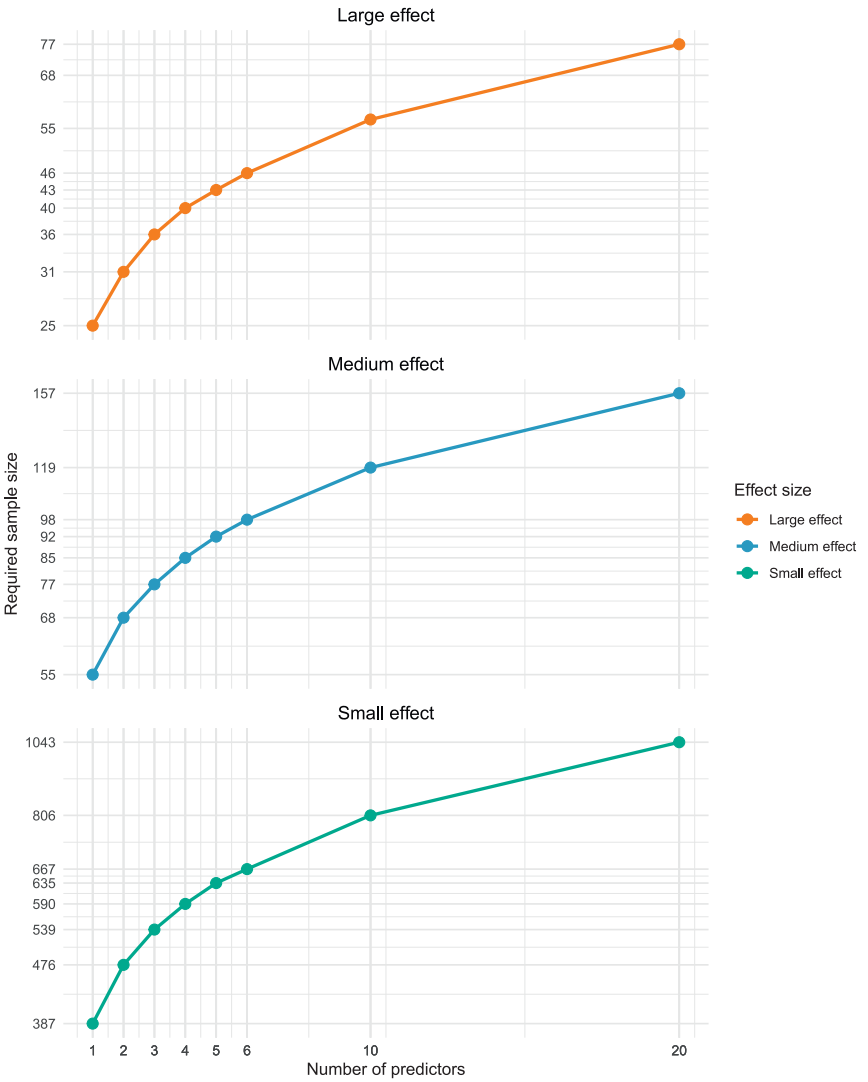


Figure 9.7 The sample size required to test the overall regression model depending on the number of predictors and the size of expected effect, $R^2 = 0.02$ (small), 0.13 (medium) and 0.26 (large)

10 I used the program G*power to compute these values.

9.6 Fitting linear models: the general procedure Ⓐ

Figure 9.8 shows the general process of fitting linear model. First, we should produce scatterplots to get some idea of whether the assumption of linearity is met, and to look for outliers or obvious unusual cases. Having done this initial screen for problems, we fit a model and save the various diagnostic statistics that we discussed in Section 9.3. If we want to generalize our model beyond the sample, or we are interested in interpreting significance tests and confidence intervals, then we examine these residuals to check for homoscedasticity, normality, independence and linearity. If we find problems then we take corrective action, which involves fitting a different model. If the problem is lack of linearity then we fit a non-linear model, for lack of independent errors we'd use a multilevel model, and in all other situations we fit a robust version of the model using either bootstrapping (small samples) or robust standard errors. This process might seem complex, but it's not as bad as it seems.

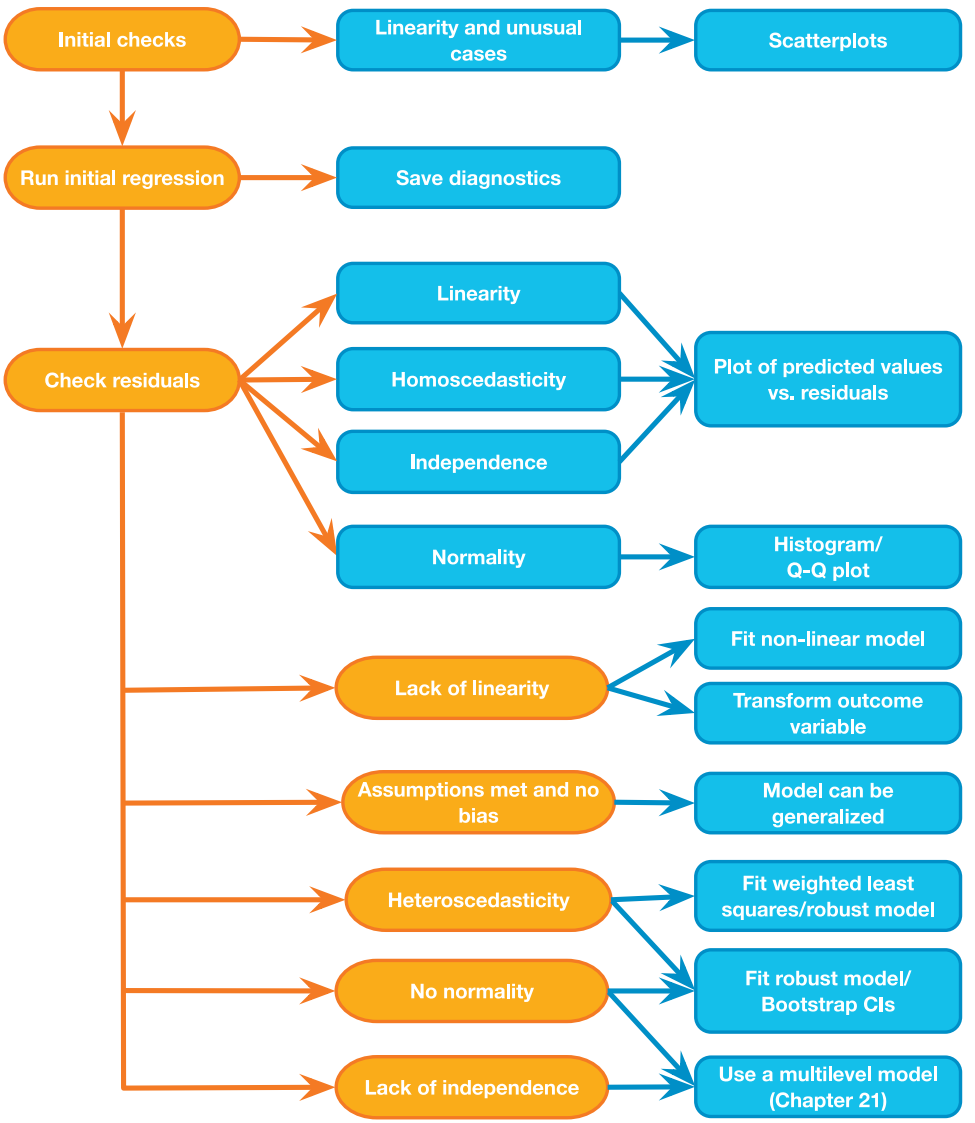


Figure 9.8 The process of fitting a regression model

9.7 Using SPSS to fit a linear model with one predictor Ⓐ

Earlier on I asked you to imagine that I worked for a record company and that my boss was interested in predicting album sales from advertising. There are data for this example in the file **album_sales.sav** and described in Section 9.2.1. Note how the data are laid out (Figure 9.9): each variable is in a column and each row represents a different album (200 in total). So, the first album had £10,260 spent advertising it, sold 330,000 copies, received 43 plays on radio the week before release and was made by a band with a pretty sick image.

Earlier we plotted the relationship between advertising budget and sales. That plot (Figure 9.2) shows a positive relationship: the more money spent advertising the album, the more it sells. Of course there are some albums that sell well regardless of advertising (top left of scatterplot), but there are none that sell badly when advertising levels are high (bottom right of scatterplot). The scatterplot shows the line of best fit for these data: bearing in mind that the mean would be represented by a flat line at around the 200,000 sales mark, the regression line is noticeably different.

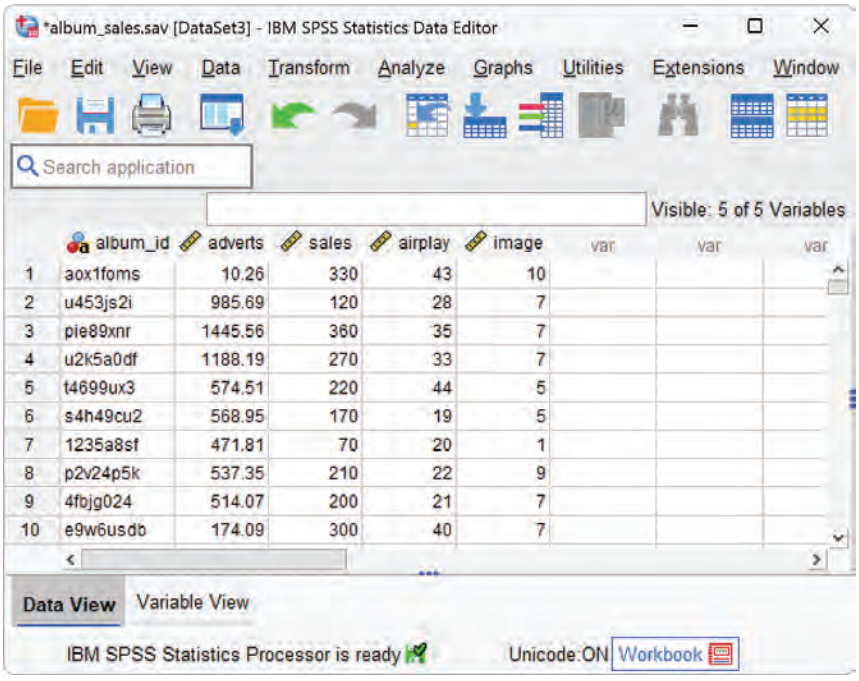


Figure 9.9 The data editor for fitting a linear model

To fit the model access the main dialog box by selecting *Analyze* > *Regression* > Linear... (Figure 9.10). First, we define the outcome variable (in this example **sales**). Select **sales** from the list on the left-hand side, and drag it to the space labelled *Dependent* (or click). In this model we're going to enter only one predictor (**adverts**) so select it from the list and drag it (or click) to the box labelled *Independent(s)*. There are a tonne of options available, but we'll explore these when we build up the model in due course. For now, request bootstrapped confidence intervals for the regression coefficients by clicking Bootstrap... (see Section 6.10.8). Select ☑ Perform bootstrapping to activate bootstrapping, and to get a 95% confidence interval select ◉ Bias corrected accelerated (BCa). Click OK in the main dialog box to fit the model.

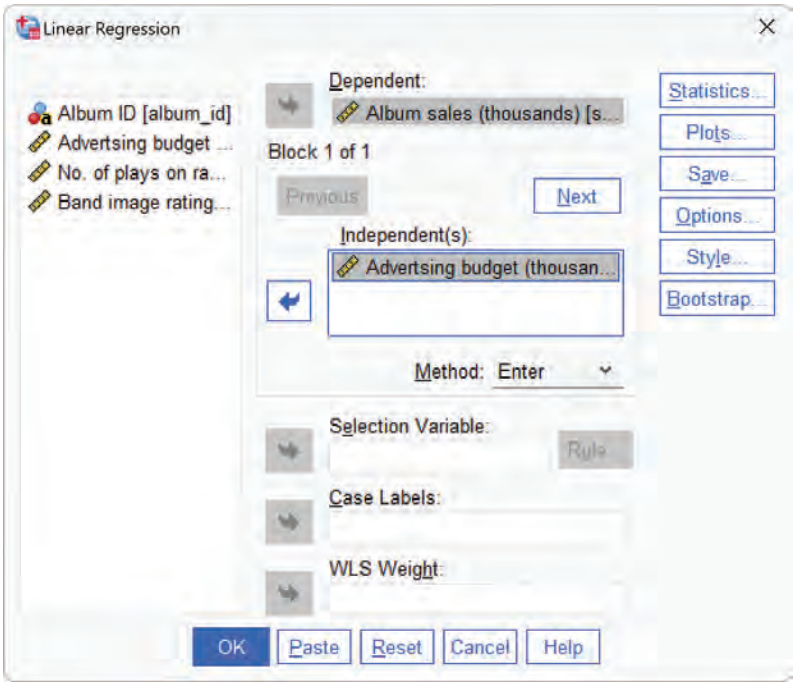


Figure 9.10 Main dialog box for regression

9.8 Interpreting a linear model with one predictor Ⓐ

9.8.1 Overall fit of the model Ⓐ

The first table is a summary of the model (Output 9.2). This summary table provides the value of R and R^2 for the model. For these data, R has a value of 0.578 and because there is only one predictor, this value is the correlation between advertising and album sales (you can confirm this by running a correlation using what you learnt in Chapter 8). The value of R^2 is 0.335, which tells us that advertising expenditure can account for 33.5% of the variation in album sales. This means that 66.5% of the variation in album sales remains unaccounted for: there might be other variables that have an influence also.

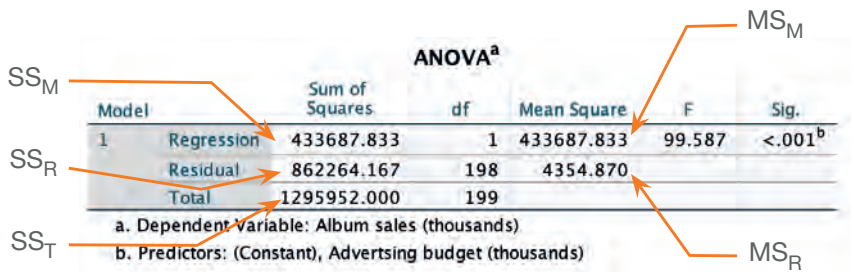
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.991

a. Predictors: (Constant), Advertsing budget (thousands)

Output 9.2

The next part of the output (Output 9.3) reports the various sums of squares described in Figure 9.4, the degrees of freedom associated with each and the resulting mean squares (Eq. 9.19). The most important part of the table is the F-statistic (Eq. 9.20) of 99.59 and its associated significance value of $p < 0.001$ (expressed this way because the value in the column labelled *Sig.* is less than 0.001). This p-value tells us that there is less than a 0.1% chance that an F-statistic at least this large would happen if the null hypothesis were true. Therefore, we could conclude that our model results in significantly better prediction of album sales than if we used the mean value of album sales. In short, the linear model overall predicts album sales significantly.



ANOVA[a]

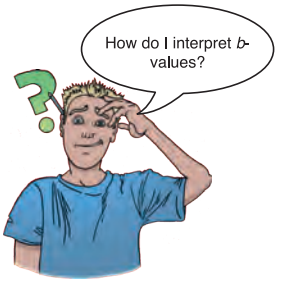
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	<.001[b]
	Residual	862264.167	198	4354.870		
	Total	1295952.000	199			

a. Dependent Variable: Album sales (thousands)

b. Predictors: (Constant), Advertsing budget (thousands)

Output 9.3

9.8.2 Model parameters Ⓐ



The table in Output 9.4 provides estimates of the model parameters (the beta values) and the significance of these values. We saw in Eq. 9.1 that b_0 was the Y intercept and this value is 134.14 (B for the constant in Output 9.4). This value can be interpreted as meaning that when no money is spent on advertising (when $X = 0$), the model predicts that 134,140 albums will be sold (remember that our unit of measurement is thousands of albums). We can also read off the value of b_1 from the table, which is 0.096. Although this value is the slope of the line for the model, it is more useful to think of this value as representing *the change in the outcome associated with a unit change in the predictor*. In other words, if our predictor variable is increased by one unit (if the advertising budget is increased by 1), then our model predicts that 0.096 extra albums will be sold. Our units of measurement were thousands of pounds and thousands of albums sold, so we can say that for an increase in advertising of £1000 the model predicts 96 (0.096 × 1000 = 96) extra album sales. This investment is pretty useless for the record company: it invests £1000 and gets only 96 extra sales! Fortunately, as we already know, advertising accounts for only one-third of album sales.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	134.140	7.537		17.799	<.001
	Advertsing budget (thousands)	.096	.010	.578	9.979	<.001

a. Dependent Variable: Album sales (thousands)

Bootstrap for Coefficients

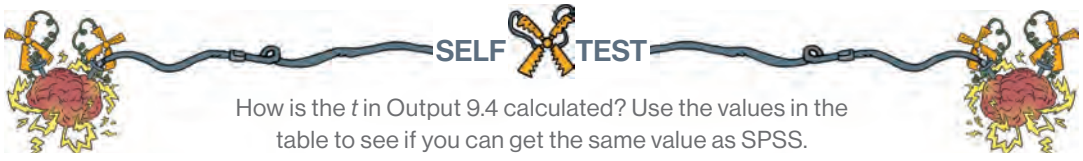
Model		Bootstrap[a] B	Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval Lower	Upper
1	(Constant)	134.140	-.049	8.087	<.001	118.699	150.746
	Advertsing budget (thousands)	.096	.000	.008	<.001	.079	.113

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 9.4

We saw earlier that if a predictor is having a significant impact on our ability to predict the outcome then its b should be different from 0 (and large relative to its standard error). We also saw that the t-test and associated p-value tell us whether the b-value is significantly different from 0. The column *Sig.* contains the exact probability that a value of t at least as big as the one in the table would occur if the value of b in the population were zero. If this probability is less than or equal to 0.05, then people interpret that as the predictor being a 'significant' predictor of the outcome (see Chapter 2). For both ts, the probabilities are given as < 0.001 (less than 0.001), and so we can say that the probability of these t values (or larger) occurring if the values of b in the population were zero is less than 0.001. In other words, the bs are significantly different from 0. In the case of the b for advertising budget this result means that the advertising budget makes a significant contribution ($p < 0.001$) to predicting album sales.

If our sample is one of the 95% producing confidence intervals that contain the population value then the bootstrap confidence interval tells us that the population value of b for advertising budget is likely to fall between 0.079 and 0.113 and because this interval doesn't include zero we might conclude that there is a genuine positive relationship between advertising budget and album sales in the population. Also, the significance associated with this confidence interval is $p < 0.001$, which is highly significant. Note that the bootstrap process involves re-estimating the standard error (it changes from 0.01 in the original table to a bootstrap estimate of 0.008). This is a very small change. The bootstrap confidence intervals and significance values are useful to report and interpret because they do not rely on assumptions of normality or homoscedasticity.



How is the t in Output 9.4 calculated? Use the values in the table to see if you can get the same value as SPSS.

9.8.3 Using the model Ⓐ

We have discovered that we have a model that significantly improves our ability to predict album sales. The next stage is often to use that model to make predictions about the future. First we specify the model by replacing the b-values in Eq. 9.1 with the values from Output 9.4. We can also replace the X and Y with the variable names:

$$\begin{aligned}\widehat{\text{album sales}}_i &= \hat{b}_0 + \hat{b}_1\text{advertising budget}_i \\ &= 134.14 + (0.096 \times \text{advertising budget}_i)\end{aligned} \tag{9.23}$$

We can make a prediction about album sales by replacing the advertising budget with a value of interest. For example, if we spend £100,000 on advertising a new album, remembering that our units are already in thousands of pounds, we replace the advertising budget with 100. We discover that album sales should be around 144,000 for the first week of sales:

$$\begin{aligned}\widehat{\text{album sales}}_i &= 134.14 + (0.096 \times \text{advertising budget}_i) \\ &= 134.14 + (0.096 \times 100) \\ &= 143.74\end{aligned} \tag{9.24}$$



Cramming Sam's Tips
Linear models

- A linear model (regression) is a way of predicting values of one variable from another based on a model that describes a straight line.
- This line is the line that best summarizes the pattern of the data.
- To assess how well the model fits the data use:
    - R^2, which tells us how much variance is explained by the model compared to how much variance there is to explain in the first place. It is the proportion of variance in the outcome variable that is shared by the predictor variable.
    - *F*, which tells us how much variability the model can explain relative to how much it can't explain (i.e., it's the ratio of how good the model is compared to how bad it is).
    - The *b*-value, which tells us the gradient of the regression line and the strength of the relationship between a predictor and the outcome variable. If it is significant (*Sig.* ≤ 0.05 in the SPSS output) then the predictor variable significantly predicts the outcome variable.





How many albums would be sold if we spent £666,000 on advertising the latest album by Deafheaven?

9.9 The linear model with two or more predictors (multiple regression) Ⓑ

Imagine that the record company executive wanted to extend the model of albums sales to incorporate other predictors. Before an album is released, the executive notes the amount spent on advertising, the number of times songs from the album are played on a prominent radio station the week before release (**airplay**), and ratings of the band's image (**image**). He does this for 200 albums (each by a different band). The credibility of the band's image was rated by a random sample of the target audience on a scale from 0 (dad dancing at a disco) to 10 (sicker than a dog that's eaten a bag of onions). The mode rating was used because the executive was interested in what most people thought, not the average opinion.

When we build a model with several predictors everything we have discussed so far applies. However, there are some additional things to think about. The first is what variables to enter into the model. A great deal of care should be taken in selecting predictors for a model because the estimates of the regression coefficients depend upon the variables in the model (and the order in which they are entered). *Do not enter hundreds of predictors, just because you've measured them, and expect the resulting model to make sense*. IBM SPSS Statistics software will happily generate output based on any garbage you decide to feed it, it will not judge you, but others will. Select predictors based on a sound theoretical rationale or well-conducted past research that has demonstrated their importance.[11] In our example, it seems logical that the band's image and radio play ought to affect sales, so these are sensible predictors. It would not be sensible to measure how much the album cost to make because this won't affect sales directly: you would just add noise to the model. If predictors are being added that have never been looked at before (in your research context) then select these variables based on their substantive *theoretical* importance. The key point is that the most important thing when building a model is to use your brain – which is slightly worrying if your brain is as small as mine.

9.9.1 Methods of entering predictors into the model Ⓑ

Having chosen predictors, you must decide the order to enter them into the model. When predictors are completely uncorrelated the order of variable entry has very little effect on the parameters estimated; however, we rarely have uncorrelated predictors and so the method of variable entry has consequences and is, therefore, important.

Other things being equal, use **hierarchical regression** in which you select predictors based on past work and you decide in which order to enter them into the model. Generally speaking, you should enter known predictors (from other research) into the model first in order of their importance in predicting the outcome. After having entered known predictors, you can add new predictors into the model simultaneously, in a stepwise manner, or hierarchically (entering the new predictor suspected to be the most important first).

An alternative is forced entry (or *Enter* as it is known in SPSS), in which you force all predictors into the model simultaneously. Like hierarchical, this method relies on good theoretical reasons for including the chosen predictors, but unlike hierarchical, you don't determine the order in which variables are entered. Some researchers believe that this method is the only appropriate method for theory testing (Studenmund & Cassidy, 1987) because stepwise techniques are influenced by random variation in the data and so seldom give replicable results if the model is retested.

The final option, **stepwise regression**, is generally frowned upon by statisticians. Nevertheless, SPSS makes it easy to do and actively encourages it in the *Automatic Linear Modeling* process (probably because this function is aimed at people who don't know better) – see Oditi's Lantern. I'm assuming that you wouldn't wade through 1100 pages of my drivel unless you wanted to know better, so we'll give stepwise a wide berth. However, you probably ought to know what it does so you can understand why to avoid it. The stepwise method bases decisions about the order in which predictors enter the model on a purely mathematical criterion. In the *forward* method, an initial model is defined that contains only the constant (b_0). The computer then searches for the predictor (out of the ones available) that best predicts the outcome variable – it does this by selecting the predictor that has the highest simple correlation with the outcome. If this predictor significantly improves the model's ability to predict the outcome then it is retained and the computer looks to add a second predictor from the available pool of variables. The next predictor the computer tries will be the one that has the largest semi-partial correlation with the outcome. Remember that the semi-partial correlation quantifies the unique overlap between two variables X and Y: it 'partials out' or accounts for the relationship that X

11 Preferably past research that is methodologically and statistically rigorous and yielded reliable, generalizable models.

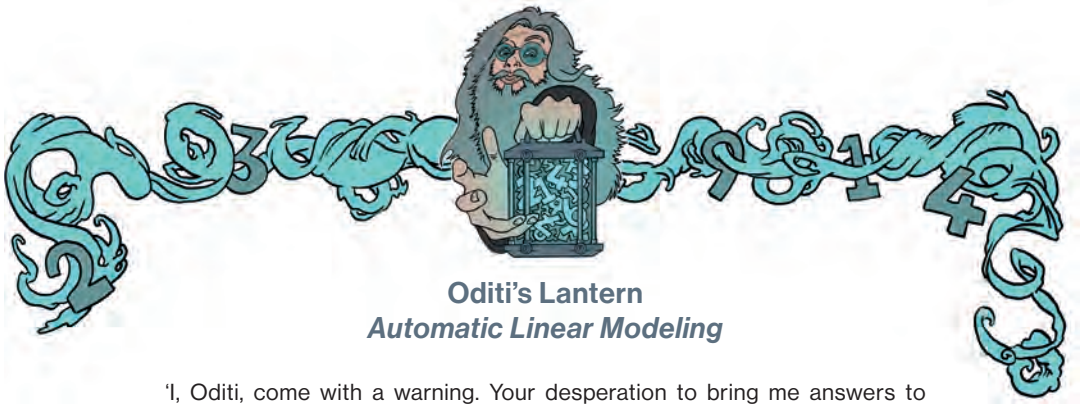
has with other predictors. Therefore, the computer looks for the variable that has the largest *unique* overlap with the outcome. This variable is retained if it significantly improves the fit of the model, otherwise it is rejected and the process stops. If it is retained and there still potential predictors left out of the model then these are reviewed and the one with the largest semi-partial correlation with the outcome is entered, evaluated and retained if it significantly improves the fit and so on until there are no more potential predictors or none of the potential predictors significantly improves the model if it is entered.

Let's make this process a bit more concrete. In Section 8.5, we used an example of the relationships between exam performance, exam anxiety and revision time. Imagine our goal is to predict exam performance from the other two variables. Think back to Figure 8.11. If we build the model stepwise, the first step is to see which of exam anxiety and revision time overlaps more with exam performance. The area of overlap between exam performance and exam anxiety is area A + C (19.4%), whereas for revision time it is area B + C (15.8%). Therefore, exam anxiety will enter the model first and is retained only if it significantly improves the model's predictive power. If not, no predictor variables are entered.

In our exam performance example, there is only one other potential predictor (revision time), so this will be entered next. Remember that its unique overlap with exam performance is area B in Figure 8.11; we ignore the part of its overlap with exam performance that is shared with exam anxiety (area C), because that variable is already in the model. If area B is big enough to improve the fit of the model significantly then revision time will be retained. If not, the final model will contain only exam anxiety.

In the case where we had another potential predictor (let's say we measured the difficulty of the exam) and exam anxiety had been entered first, then the unique overlap for revision time and exam performance (area B) would be compared to the equivalent area for exam difficulty. The variable with the bigger area would be entered next, evaluated and retained only if its inclusion improved the fit of the model.

The *stepwise* method in SPSS is the same as the forward method, except that each time a predictor is added to the equation, a removal test is made of the least useful predictor. As such, the regression



Oditi's Lantern
Automatic Linear Modeling

'I, Oditi, come with a warning. Your desperation to bring me answers to numerical truths so as to gain a privileged place within my heart may lead you into the temptation that is SPSS's *Automatic Linear Modeling*. This feature promises answers without thought, and like a cat that is promised a fresh salmon, you will drool and purr in anticipation. If you want to find out more then stare into my lantern, but be warned, sometimes what looks like a juicy salmon is a rotting pilchard in disguise.'



equation is constantly reassessed to see whether redundant predictors can be removed. The *backward* method is the opposite of the forward method in that the model initially contains all predictors and the contribution of each is evaluated with the *p*-value of its *t*-test. This significance value is compared against a removal criterion (which can be either an absolute value of the test statistic or a *p*-value). If a predictor meets the removal criterion (i.e., it is not making a statistically significant contribution to the model) it is removed and the model is re-estimated for the remaining predictors. The contribution of the remaining predictors is then reassessed.



Which of these methods should you use? The short answer is 'not step wise', because variables are selected based upon mathematical criteria. The issue is that these criteria (e.g., the semi-partial correlation) are at the mercy of sampling variation. That is, a particularly variable might have a large semi-partial correlation in your sample but a small one in a different sample. Therefore, models built using stepwise methods are less likely to generalize across samples because the selection of variables in the model is affected by the sampling process. Also, because the criterion for retaining variables is based on statistical significance, your sample size affects the model you get: in large samples significance tests are highly powered, resulting in predictors being retained that make trivial contributions to predicting the outcome, and in small samples where power is low, predictors that make a large contribution may get overlooked. Consequently, there is the danger of overfitting (having too many variables in the model that essentially make little contribution to predicting the outcome) and underfitting (leaving out important predictors) the model. Stepwise methods also take important methodological decisions out of the hands of the researcher.

The main problem with stepwise methods is that they assess the fit of a variable based on the other variables in the model. Jeremy Miles (who has worked with me on other books) illustrates this problem by imagining getting dressed using a stepwise method. You wake up one morning and on your dressing table (or floor, if you're me) you have underwear, some jeans, a T-shirt and jacket. Imagine these items are predictor variables. It's a cold day and you're trying to keep warm. A stepwise method will put your trousers on first because they fit your goal best. It then looks around and tries the other clothes (variables). It tries to put your underwear on you but it won't fit over your jeans. It decides they are 'a poor fit' and discards them. It tries a jacket – that fits, but your T-shirt doesn't go over the top and is discarded. You end up leaving the house in jeans and a jacket with nothing underneath. You are very cold. Later in the day during a university seminar you stand up and your trousers fall down (because your body has shrunk from the cold) exposing you to your year group. It's a mess. The problem is that the underwear was a poor fit only because when you tried to put it on you were already wearing jeans. In stepwise methods, variables might be considered bad predictors only because of what has already been put in the model.

For these reasons, stepwise methods are best avoided except for exploratory model building. If you do decide to use a stepwise method then let the statistical blood be on your hands, not mine. Use the backward method rather than the forward method to minimize **suppressor effects**, which occur when a predictor has a significant effect only when another variable is held constant. Forward selection is more likely than backward elimination to exclude predictors involved in suppressor effects. As such, the forward method runs a higher risk of making a Type II error (i.e., missing a predictor that does in fact predict the outcome). It is also advisable to cross-validate your model by splitting the data (see Section 9.4.2).

9.9.2 Comparing models Ⓑ

Hierarchical and (although obviously you'd never use them) stepwise methods involve adding predictors to the model in stages, and it is useful to assess the improvement to the model at each stage. Given that larger values of R^2 indicate better fit, a simple way to quantify the improvement when predictors

are added is to compare the R^2 for the new model to that for the old model. We can assess the significance of the change in R^2 using Eq. 9.21, but because we're looking at the change in models we use the change in R^2 (R^2_{change}) and the change in the number of predictors (k_{change}) as well as the R^2 (R^2_{new}) and number of predictors (k_{new}) in the new model:

$$F_{change} = \frac{(N - k_{new} - 1)R^2_{change}}{k_{change}(1 - R^2_{new})} \tag{9.25}$$

We can compare models using this F-statistic. The problem with R^2 is that when you add more variables to the model, it always goes up. So, if you are deciding which of two models fits the data better, the model with more predictor variables in will always fit better. The **Akaike information criterion (AIC)**[12] is a measure of fit that penalizes the model for having more variables. If the AIC is bigger, the fit is worse; if the AIC is smaller, fit is better. If you use the *Automated Linear Modeling* function in SPSS then you can use the AIC to select models rather than the change in R^2. The AIC doesn't mean anything on its own: you cannot say that an AIC of 10 is small, or that a value of 1000 is large. The only thing you do with the AIC is compare it to other models with the same outcome variable: if it's getting smaller, the fit of your model is improving.

9.9.3 Multicollinearity Ⓑ

A final consideration for models with more than one predictor is multicollinearity, which exists when there is a strong correlation between two or more predictors. **Perfect collinearity** exists when at least one predictor is a perfect linear combination of the others (the simplest example being two predictors that are perfectly correlated – they have a correlation coefficient of 1). If there is perfect collinearity between predictors it becomes impossible to obtain unique estimates of the regression coefficients because there are an infinite number of combinations of coefficients that would work equally well. Put simply, if we have two predictors that are perfectly correlated, then the values of b for each variable are interchangeable. The good news is that perfect collinearity is rare in real-life data. The bad news is that less than perfect collinearity is virtually unavoidable. Low levels of collinearity pose little threat to the model estimates, but as collinearity increases there are three problems that arise:

- **Untrustworthy estimates** ($\hat{b}$): As collinearity increases so do the standard errors of the $\hat{b}$s. Big standard errors for $\hat{b}$s mean more variability in these estimates across samples, and a greater chance of (1) predictor equations that are unstable across samples too; and (2) $\hat{b}$s in the sample that are unrepresentative of those in the population. Crudely put, multicollinearity leads to untrustworthy *parameter estimates*. Don't lend them money and don't let them go for dinner with your boy- or girlfriend.
- **It limits the size of *R***: Remember that R is a measure of the correlation between the predicted values of the outcome and the observed values and that R^2 indicates the variance in the outcome for which the model accounts. Imagine a situation in which a single variable predicts the outcome variable with $R = 0.80$ and a second predictor variable is added to the model. This second variable might account for a lot of the variance in the outcome (which is why it is included in the model), but the variance it accounts for is the same variance accounted for by the first variable (the second variable accounts for very little *unique* variance). Hence, the overall variance in the outcome accounted for by the two predictors is little more than when only one predictor is used (R might increase from 0.80 to 0.82). If, however, the two predictors are completely uncorrelated, then the second predictor is likely to account for

12 Hirotsugu Akaike (pronounced A-ka-ee-kay) was a Japanese statistician who gave his name to the AIC, which is used in a huge range of different places.

different variance in the outcome than that accounted for by the first predictor. The second predictor might account for only a little of the variance in the outcome, but the variance it does account for is different from that of the other predictor (and so when both predictors are included, R is substantially larger, say 0.95).

- **Importance of predictors**: Multicollinearity between predictors makes it difficult to assess the individual importance of a predictor. If the predictors are highly correlated, and each accounts for similar variance in the outcome, then how can we know which of the two variables is important? We can't – the model could include either one, interchangeably.

A 'ballpark' method of identifying multicollinearity (that will miss subtler forms) is to scan the correlation matrix for predictor variables that correlate very highly (values of r above 0.80 or 0.90). SPSS can compute the **variance inflation factor (VIF)**, which indicates whether a predictor has a strong linear relationship with the other predictor(s), and the **tolerance** statistic, which is its reciprocal (1/VIF). Some general guidelines have been suggested for interpreting the VIF:

- If the largest VIF is greater than 10 (or the tolerance is below 0.1) then this indicates a serious problem (Bowerman & O'Connell, 1990; Myers, 1990).
- If the average VIF is substantially greater than 1 then the regression may be biased (Bowerman & O'Connell, 1990).
- Tolerance below 0.2 indicates a potential problem (Menard, 1995).

Other measures that are useful in discovering whether predictors are dependent are the *eigenvalues of the scaled, uncentred cross-products matrix*, the *condition indexes* and the *variance proportions*. These statistics will be covered as part of the interpretation of SPSS output (see Section 9.11.5).

9.10 Using SPSS to fit a linear model with several predictors Ⓑ

Remember the general procedure in Figure 9.8. First, we could look at scatterplots of the relationships between the outcome variable and the predictors. Figure 9.11 shows a matrix of scatterplots for our album sales data, but I have shaded all the scatterplots except the three related to the outcome, album sales. Although the data are messy, the three predictors have reasonably linear relationships with the album sales and there are no obvious outliers (except maybe in the bottom left of the scatterplot with band image).



9.10.1 Main options Ⓑ

Past research shows that advertising budget is a significant predictor of album sales, and so we should include this variable in the model first, entering the new variables (**airplay** and **image**) afterwards. This method is hierarchical (we decide the order that variables are entered based on past research). To do a hierarchical regression we enter the predictors in blocks, with each block representing one step in the hierarchy. Access the main *Linear Regression* dialog box by selecting *Analyze* > *Regression* > Linear... We encountered this dialog box when we looked at a model with only one

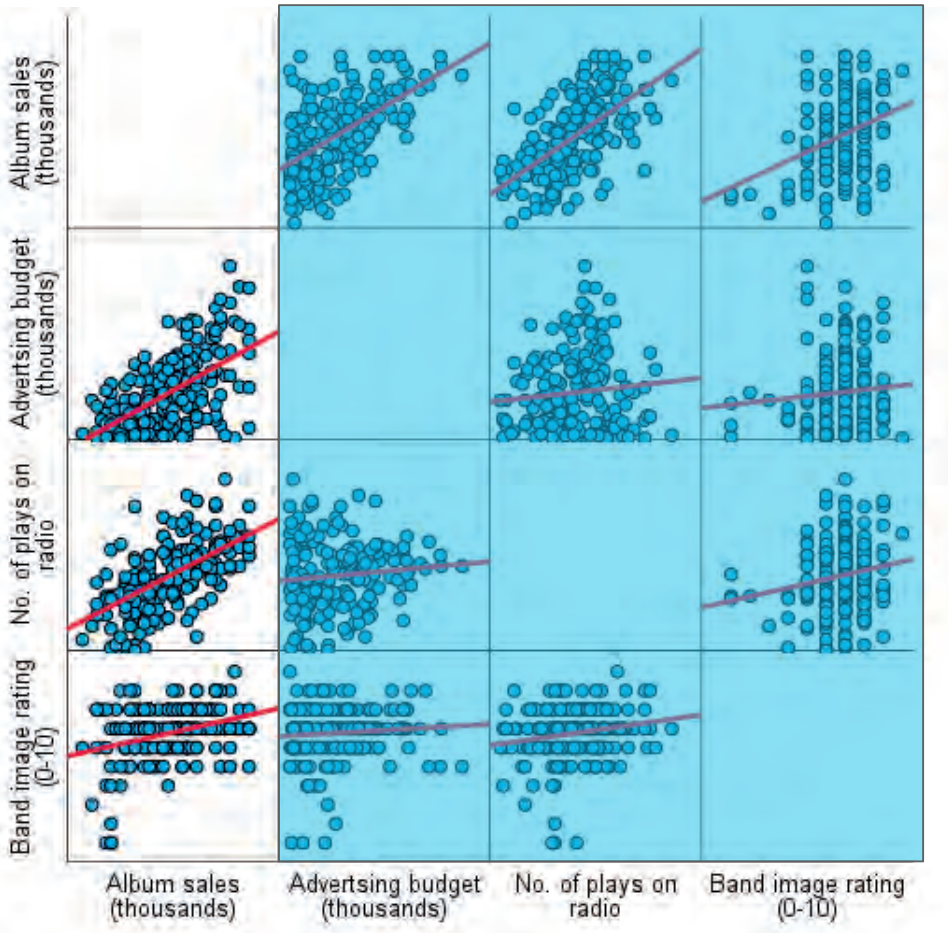


Figure 9.11 Matrix scatterplot of the relationships between advertising budget, airplay, image rating and album sales

predictor (Figure 9.10). To set up the first block we do what we did before: drag **sales** to the box labelled *Dependent* (or click ▸). We also need to specify the predictor variable for the first block, which we decided should be advertising budget. Drag this variable from the left-hand list to the box labelled *Independent(s)* (or click ▸). Underneath the *Independent(s)* box is a drop-down menu for specifying the *Method* of variable entry (see Section 9.9.1). You can select a different method for each block by clicking on Enter. The default option is forced entry, and this is the option we want, but if you were carrying out more exploratory work, you might use a different method.

Having specified the first block, we can specify a second by clicking Next. This process clears the *Independent(s)* box so that you can enter the new predictors (note that it now reads *Block 2 of 2* above this box to indicate that you are in the second block of the two that you have so far specified). We decided that the second block would contain the remaining predictors, so select **airplay** and **image** and drag them to the *Independent(s)* box (or click ▸). The dialog box should look like Figure 9.12. To move between blocks use the Previous and Next buttons (e.g., to move back to block 1, click Previous). It is possible to select different methods of variable entry for different blocks. For example, having specified forced entry for the first block, we could now specify a stepwise method for the second. Given that we have no previous research regarding the effects of image and airplay on album sales, we might be justified in doing this. However, because of the problems with stepwise methods, I am going to stick with forced entry for both blocks.

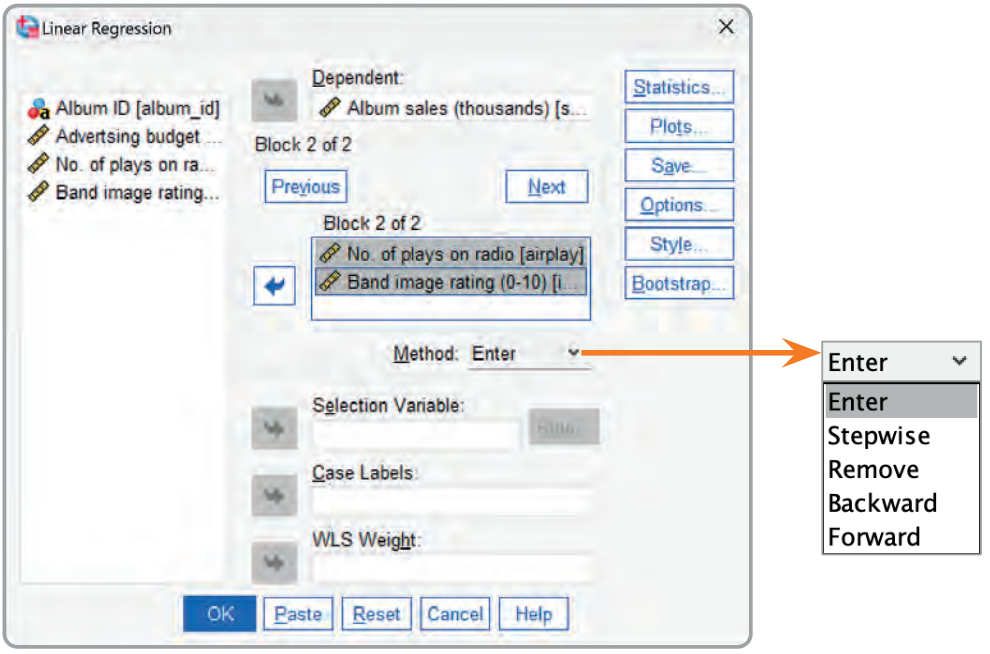


Figure 9.12 Main dialog box for block 2 of the multiple regression

9.10.2 Statistics Ⓑ

In the main *Linear Regression* dialog box click Statistics... to open the dialog box in Figure 9.13. Below is a run-down of the options available. Select the options you require and click Continue to return to the main dialog box.

- *Estimates*: This option is selected by default because it gives us the estimated *b*-values for the model as well as the associated *t*-test and *p*-value.
- *Confidence intervals*: This option produces confidence intervals around each $\hat{b}$ in the model. Remember that if the model assumptions are not met these confidence intervals will be inaccurate and bootstrap confidence intervals should be used instead.
- *Covariance matrix*: This option produces a matrix of the covariances, correlation coefficients and variances between the $\hat{b}$s for each variable in the model. A variance–covariance matrix displays variances along the diagonal and covariances as off-diagonal elements. Correlations are produced in a separate matrix.
- *Model fit*: This option produces the *F*-test, *R*, R^2 and the adjusted R^2 (described in Section 9.2.4).
- *R squared change*: This option displays the change in R^2 resulting from the inclusion of a new predictor (or block of predictors) – see Section 9.9.2.
- *Descriptives*: This option displays a table of the mean, standard deviation and number of observations of the variables included in the model. A correlation matrix is produced too, which can be helpful for spotting multicollinearity.
- *Part and partial correlations*: This option produces the zero-order correlation (the Pearson correlation) between each predictor and the outcome variable. It also produces the semi-partial (part) and partial correlation between each predictor and the outcome (see Section 8.5 and 9.9.1).
- *Collinearity diagnostics*: This option produces collinearity statistics such as the VIF, tolerance, eigenvalues of the scaled, uncentred cross-products matrix, condition indexes and variance proportions (see Section 9.9.3).

- *Durbin-Watson*: This option produces the Durbin–Watson test statistic, which tests the assumption of independent errors when cases have some meaningful sequence. In this case it isn't useful because our cases do not have a meaningful order.
- *Casewise diagnostics*: This option produces a table that lists the observed value of the outcome, the predicted value of the outcome, the difference between these values (the residual) and this difference standardized. You can choose to have this information for all cases, but that will result in a big table in large samples. The alternative option is to list only cases for which the standardized residual is greater than 3 (when the ± sign is ignored). I usually change this to 2 (so that I don't miss cases with standardized residuals not quite reaching the threshold of 3). A summary table of residual statistics indicating the minimum, maximum, mean and standard deviation of both the values predicted by the model and the residuals is also produced (see Section 9.10.4).

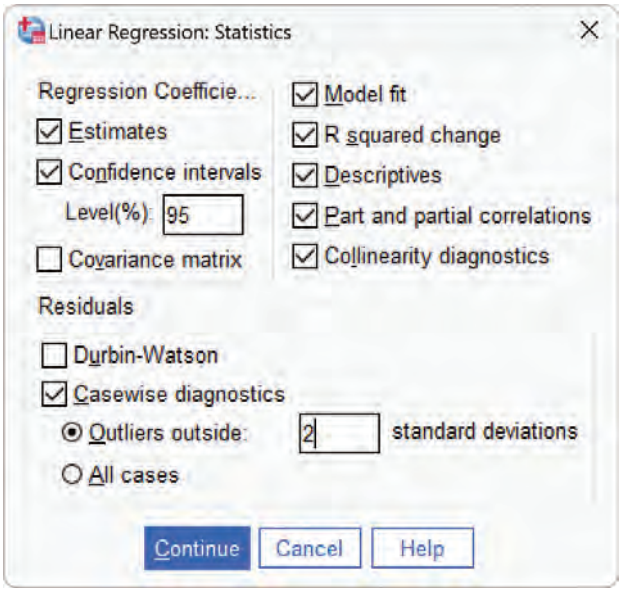


Figure 9.13 *Statistics* dialog box for regression analysis

9.10.3 Regression plots Ⓑ

Once you are back in the main dialog box, click Plots... to activate the dialog box in Figure 9.14, which we can use to test some assumptions of the model. Most of these plots involve various *residual* values, which were described in Section 9.3. The left-hand side lists several variables:

- **DEPENDNT**: the outcome variable.
- ***ZPRED**: the standardized predicted values of the outcome based on the model. These values are standardized forms of the values predicted by the model.
- ***ZRESID**: the standardized residuals, or errors. These values are the standardized differences between the observed values of the outcome and those predicted by the model.
- ***DRESID**: the deleted residuals described in Section 9.3.2.
- ***ADJPRED**: the adjusted predicted values described in Section 9.3.2.
- ***SRESID**: the Studentized residual described in Section 9.3.1.
- ***SDRESID**: the Studentized deleted residual described in Section 9.3.2.

In Section 6.9.6 we saw that a plot of ***ZRESID** (y-axis) against ***ZPRED** (x-axis) is useful for testing the assumptions of independent errors, homoscedasticity, and linearity. A plot of ***SRESID** (y-axis) against ***ZPRED** (x-axis) will show up heteroscedasticity also. Although often these two plots are virtually identical, the latter is more sensitive on a case-by-case basis. To create these plots drag a variable from the list to the space labelled either X or Y (which refer to the axes) or select the variable and click . When you have selected two variables for the first plot (as in Figure 9.14) you can specify a new plot (up to nine different plots) by clicking on Next. This process clears the dialog box and you can specify a second plot. Click Next or Previous to move between plots you have specified.

Ticking the box labelled *Produce all partial plots* will produce scatterplots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors. Regardless of whether the previous sentence made any sense to you, these plots have important characteristics that make them worth inspecting. First, the gradient of the regression line between the two residual variables is equivalent to the coefficient of the predictor in the regression equation. As such, any obvious outliers on a partial plot represent cases that might have undue influence on a predictor's parameter estimate, $\hat{b}$. Second, non-linear relationships between a predictor and the outcome variable are much more evident on these plots. Finally, they are useful for spotting collinearity.

There are a two other tick-boxes labelled *Standardized Residual Plots*. One produces a histogram of the standardized residuals and the other produces a normal probability plot, both of which are useful for checking for normality of errors). Click Continue to return to the main dialog box.

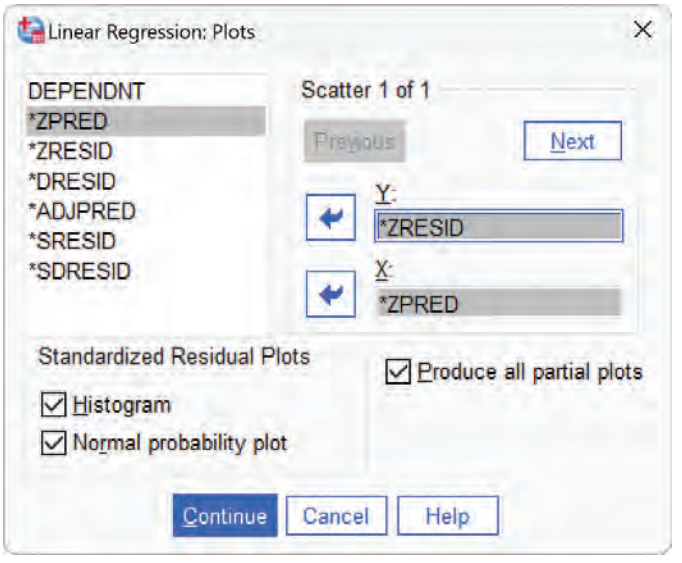


Figure 9.14 The *Plots* dialog box

9.10.4 Saving regression diagnostics Ⓑ

Section 9.3 described numerous variables that we can use to diagnose outliers and influential cases. We can save these diagnostic variables for our model in the data editor (SPSS calculates them and places the values in new columns in the data editor) by clicking Save... to access the dialog box in Figure 9.15. Most of the available options were explained in Section 9.3, and Figure 9.15 shows what I consider to be a reasonable set of diagnostic statistics. Standardized (and Studentized) versions of these diagnostics are generally easier to interpret and so I tend to select them in preference to the unstandardized versions. Once the model has been estimated, SPSS creates a column in your data editor for

each statistic requested; it uses a standard set of variable names to describe each one. After the name, there will be a number that refers to the model from which they were generated. For example, for the first model fitted to a data set the variable names will be followed by a 1, if you estimate a second model it will create a new set of variables with names followed by a 2, and so on. For reference, the names used by SPSS are listed below. Selected the diagnostics you require and click Continue to return to the main dialog box.

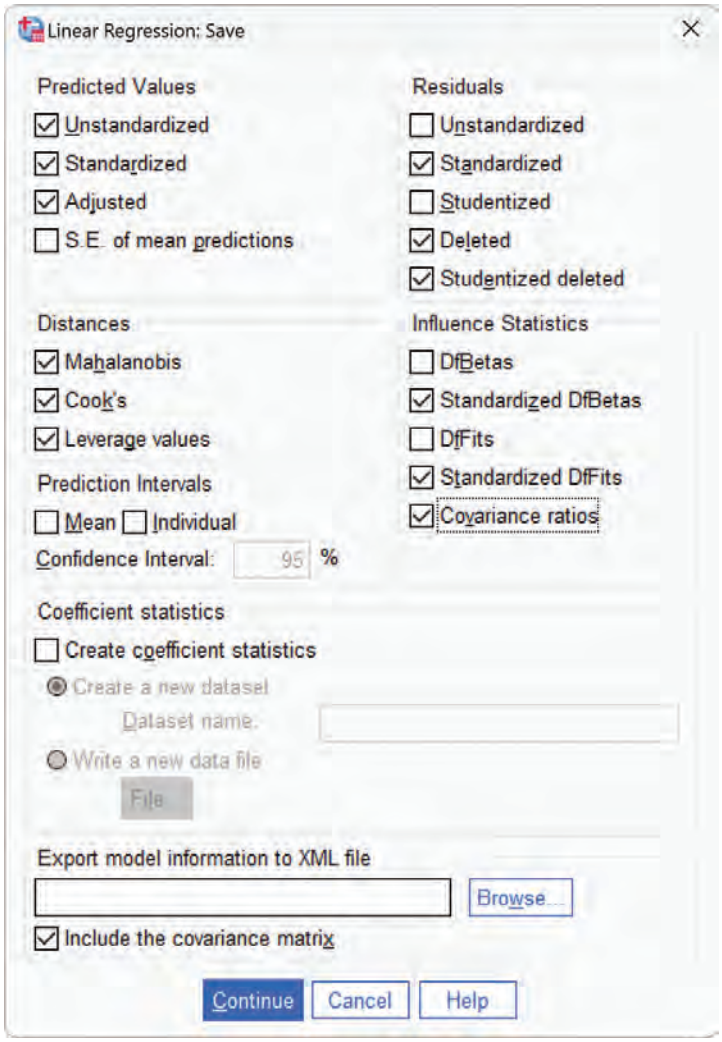


Figure 9.15 Dialog box for regression diagnostics

- **pre_1**: unstandardized predicted value;
- **zpr_1**: standardized predicted value;
- **adj_1**: adjusted predicted value;
- **sep_1**: standard error of predicted value;
- **res_1**: unstandardized residual;
- **zre_1**: standardized residual;
- **sre_1**: Studentized residual;
- **dre_1**: deleted residual;
- **sdr_1**: Studentized deleted residual;

- **mah_1**: Mahalanobis distance;
- **coo_1**: Cook's distance;
- **lev_1**: centred leverage value;
- **sdb0_1**: standardized DFBeta (intercept);
- **sdb1_1**: standardized DFBeta (predictor 1);
- **sdb2_1**: standardized DFBeta (predictor 2);
- **sdf_1**: standardized DFFit;
- **cov_1**: covariance ratio.

9.10.5 Further options Ⓑ

Clicking Options... activates the dialog box in Figure 9.16. The first set of options allows you to change the criteria used for entering variables in a stepwise regression. If you insist on doing stepwise regression, then it's probably best that you leave the default criterion of 0.05 probability for entry alone. However, you can make this criterion more stringent (0.01). There is also the option to build a model that doesn't include a constant (i.e., has no Y intercept). This option should also be left alone. Finally, you can select a method for dealing with missing data points (see SPSS Tip 6.1 for a description). Just a hint: leave the default of listwise alone because using pairwise can lead to absurdities such as R^2 that is negative or greater than 1.0.

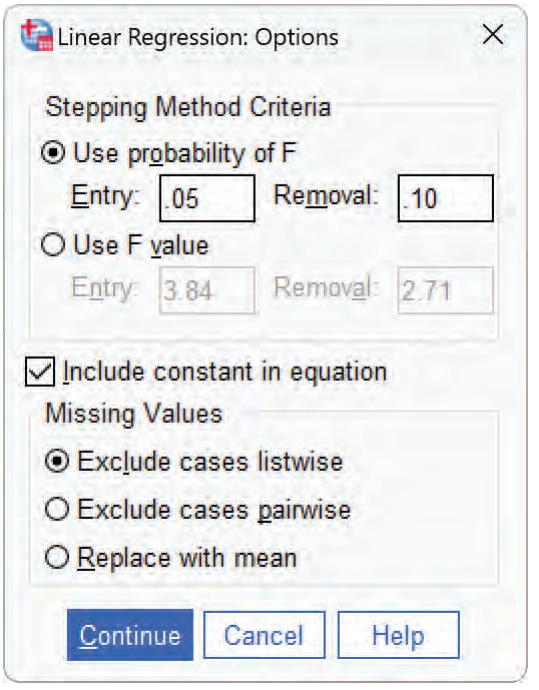
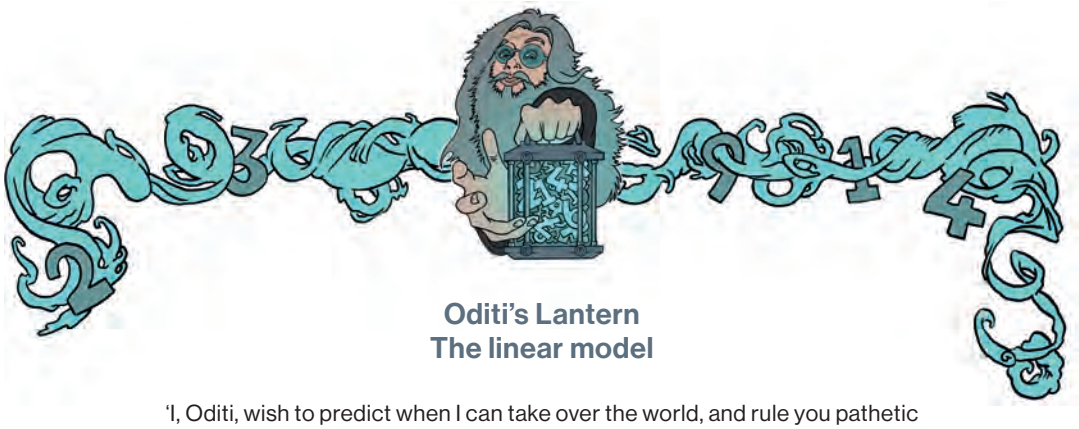


Figure 9.16 Options for linear regression

9.11 Interpreting a linear model with several predictors Ⓑ

Having selected the relevant options and returned to the main dialog box, click OK and watch in awe as SPSS spews forth quite terrifying amounts of output in the viewer window.



Oditi's Lantern
The linear model

'I, Oditi, wish to predict when I can take over the world, and rule you pathetic mortals with a will of pure iron ... erm ... ahem, I mean, I wish to predict how to save cute kittens from the jaws of rabid dogs, because I'm nice like that, and have no aspirations to take over the world. This chapter is so long that some of you will die before you reach the end, so ignore the author's bumbling drivel and stare instead into my lantern of wonderment.'



9.11.1 Descriptives Ⓑ

The output described in this section is produced using the options in Figure 9.13. If you selected the *Descriptives* option, you'll get Output 9.5, which tells us the mean and standard deviation of each variable in our model. This table is a useful summary of the variables. You'll also see a correlation matrix containing the Pearson correlation coefficient between every pair of variables, the one-tailed significance of each correlation, and the number of cases contributing to each correlation. Along the diagonal of the matrix the values for the correlation coefficients are all 1.00 (a perfect positive correlation) because they are the correlation of each variable with itself.

We can use the correlation matrix to get a sense of the relationships between predictors and the outcome, and for a preliminary look for multicollinearity. If there is no multicollinearity in the data then there should be no substantial correlations ($r > 0.9$) between predictors. If we look only at the predictors (ignore album sales) then the highest correlation is between the ratings of the band's image and the amount of airplay which is significant at the 0.01 level ($r = 0.182$, $p = 0.005$). Despite the significance, the coefficient itself is small and so there is no collinearity to worry about. If we look at the outcome variable, then it's apparent that of the predictors airplay correlates best with the outcome ($r = 0.599$, $p < 0.001$).



Cramming Sam's Tip
descriptive statistics

- Use the descriptive statistics to check the correlation matrix for multicollinearity; that is, predictors that correlate too highly with each other, $r > 0.9$.



Descriptive Statistics

	Mean	Std. Deviation	N
Album sales (thousands)	193.20	80.699	200
Advertsing budget (thousands)	614.4123	485.65521	200
No. of plays on radio	27.50	12.270	200
Band image rating (0–10)	6.77	1.395	200

Correlations

		Album sales (thousands)	Advertsing budget (thousands)	No. of plays on radio	Band image rating (0–10)
Pearson Correlation	Album sales (thousands)	1.000	.578	.599	.326
	Advertsing budget (thousands)	.578	1.000	.102	.081
	No. of plays on radio	.599	.102	1.000	.182
	Band image rating (0–10)	.326	.081	.182	1.000
Sig. (1-tailed)	Album sales (thousands)	.	<.001	<.001	<.001
	Advertsing budget (thousands)	.000	.	.076	.128
	No. of plays on radio	.000	.076	.	.005
	Band image rating (0–10)	.000	.128	.005	.
N	Album sales (thousands)	200	200	200	200
	Advertsing budget (thousands)	200	200	200	200
	No. of plays on radio	200	200	200	200
	Band image rating (0–10)	200	200	200	200

Output 9.5

9.11.2 Summary of the model Ⓑ

Output 9.6 describes the overall fit of the model. There are two models in the table because we chose a hierarchical method with two blocks and the summary statistics are repeated for each model/block.

Model 1 refers to the first stage in the hierarchy when only advertising budget is used as a predictor. Model 2 refers to when all three predictors are used. We can tell this from the footnotes under the table. If you selected the *R squared change* and *Durbin-Watson* options, then these values are included also (we didn't select Durbin–Watson so it is missing from Output 9.6).

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.578[a]	.335	.331	65.991	.335	99.587	1	198	<.001
2	.815[b]	.665	.660	47.087	.330	96.447	2	196	<.001

a. Predictors: (Constant), Advertsing budget (thousands)

b. Predictors: (Constant), Advertsing budget (thousands), Band image rating (0–10), No. of plays on radio

c. Dependent Variable: Album sales (thousands)

Output 9.6

The column labelled *R* contains the multiple correlation coefficient between the predictors and the outcome. When only advertising budget is used as a predictor, this is the simple correlation between advertising and album sales (0.578). In fact, the statistics for model 1 are the same as the simple regression model earlier (see Section 9.8). The next column gives us a value of R^2, which we already know is a measure of how much of the variability in the outcome is accounted for by the predictors. For the first model its value is 0.335, which means that advertising budget accounts for 33.5% of the variation in album sales. However, when the other two predictors are included as well (model 2), this value increases to 0.665 or 66.5% of the variance in album sales. If advertising accounts for 33.5%, then image and airplay must account for an additional 33%.[13]

The value of R^2 increasing is good, but it will always increase if we add new predictors, so the adjusted R^2 is helpful in telling us the value of R^2 adjusted for the number of predictors. In a sense it is a better estimate of the population value because it is unaffected by the number of predictors. Ideally, we'd like the adjusted value to be the same as, or very close to, the unadjusted value. The difference between the adjusted and unadjusted R^2 for our final model is small (it is 0.665 – 0.660 = 0.005 or about 0.5%). This shrinkage means that the model fit has been relatively unaffected by the number of predictors. In any case, we can use this value as the better estimate of the population value and say that the predictors account for approximately 66% of the variance in the outcome.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	<.001[b]
	Residual	862264.167	198	4354.870		
	Total	1295952.000	199			
2	Regression	861377.418	3	287125.806	129.498	<.001[c]
	Residual	434574.582	196	2217.217		
	Total	1295952.000	199			

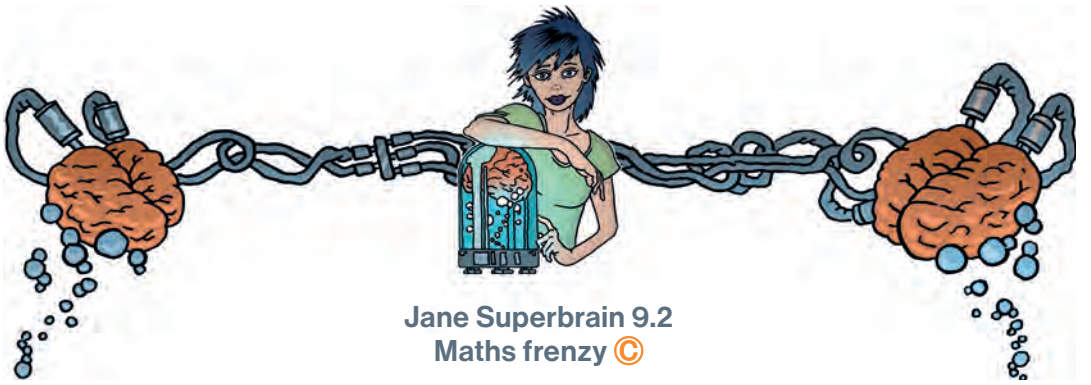
a. Dependent Variable: Album sales (thousands)

b. Predictors: (Constant), Advertsing budget (thousands)

c. Predictors: (Constant), Advertsing budget (thousands), Band image rating (0–10), No. of plays on radio

Output 9.7

13 That is, 33% = 66.5% – 33.5% (this value is the R Square Change in the table).



Jane Superbrain 9.2
Maths frenzy ©

We can have a look at how some of the values in the output are computed by thinking back to the theory part of the chapter. For example, looking at the change in R^2 for the first model, we have only one predictor (so $k = 1$) and 200 cases ($N = 200$), so the F comes from Eq. 9.21:[14]

$$F_{\text{Model1}} = \frac{(200-1-1)0.334648}{1(1-0.334648)} = 99.59$$

In model 2 in Output 9.6 two predictors have been added (image and airplay), so the new model has 3 predictors (k_{new}) and the previous model had only 1, which is a change of 2 (k_{change}). The addition of these two predictors increases R^2 by 0.330 (R^2_{change}), making the R^2 of the new model 0.665 (R^2_{new}).[15] The F-statistic for this change comes from Eq. 9.25:

$$F_{\text{change}} = \frac{(N-3-1)0.33}{2(1-0.664668)} = 96.44$$



The change statistics are provided only if requested and they tell us whether the change in R^2 is significant (i.e. how much does the model fit improve as predictors are added?). The change is reported for each block of the hierarchy: for model 1, R^2 changes from 0 to 0.335, and gives rise to an F-statistic of 99.59, which is significant with a probability less than 0.001. In model 2, in which image and airplay have been added as predictors, R^2 increases by 0.330, making the R^2 of the new model 0.665 with a significant ($p < 0.001$) F-statistic of 96.44 (Jane Superbrain Box 9.2).

Output 9.7 shows the F-test of whether the model is significantly better at predicting the outcome than using the mean outcome (i.e., no predictors). The F-statistic represents the ratio of the improvement

14 To get the same values as SPSS we have to use the exact value of R^2, which is 0.3346480676231 (if you don't believe me, double-click in the table in the SPSS output that reports this value, then double-click the cell of the table containing the value of R^2 and you'll see that what was 0.335 now has a lot more decimal places).

15 The more precise value is 0.664668.

in prediction that results from fitting the model, relative to the inaccuracy that still exists in the model (see Section 9.2.4). This table again reports the information for each model separately. The table contains the sum of squares for the model (the value of SS_M from Section 9.2.4), the residual sum of squares (the value of SS_R from Section 9.2.4 and their respective degrees of freedom. For SS_M the *df* are the number of predictors (1 for the first model and 3 for the second). For SS_R the *df* are the number of observations (200) minus the number of coefficients in the regression model. The first model has two coefficients (one each for the predictor and constant) whereas the second has four (the constant plus one for each of the three predictors). Therefore, model 1 has 198 residual degrees of freedom whereas model 2 has 196. Remember, the mean sum of squares (MS) is the SS divided by the *df* and the *F*-statistic is average improvement in prediction by the model (MS_M) divided by the average error in prediction (MS_R). The *p*-value tells us the probability of getting an *F* at least as large as the one we have if the null hypothesis were true (if we used the outcome mean to predict album sales). The *F*-statistic is 99.59, $p < 0.001$, for the initial model and 129.498, $p < 0.001$, for the second. We can interpret these results as meaning that both models significantly improve our ability to predict the outcome variable compared to not fitting the model.



Cramming Sam's Tip
The model summary

- The fit of the linear model can be assessed using the *Model Summary* and *ANOVA* tables from SPSS.
- R^2 tells you the proportion of variance explained by the model.
- If you have done a hierarchical regression, assess the improvement of the model at each stage by looking at the change in R^2 and whether it is significant (values less than or equal to 0.05 in the column labelled *Sig F Change*).
- The *F*-test tells us whether the model is a significant fit to the data overall (look for values less than or equal to 0.05 in the column labelled *Sig.*).



9.11.3 Model parameters Ⓑ

Output 9.8 shows the model parameters for both steps in the hierarchy. The first step in our hierarchy was to include advertising budget, and so the parameters for this first model are identical to those we obtained earlier in this chapter in Output 9.4. Therefore, we will focus on the parameters for the final model (in which all predictors were included). The format of the table of coefficients depends on the

options selected in Figure 9.13; for example, the confidence intervals for the *b*s, collinearity diagnostics and the part and partial correlations will be present only if you checked those options.

Earlier in the chapter we saw that a linear model with several predictors takes the form of Eq. 9.11, which contains several unknown parameters (the *b*-values). The first column in Output 9.8 contains the estimates ($\hat{b}$) of these *b*-values, which indicate the individual contribution of each predictor to the model. By replacing the *X*s in Eq. 9.11 with variables names and taking the $\hat{b}$s from Output 9.8 we can define our specific model as:

$$\begin{aligned}\widehat{\text{sales}}_i &= \hat{b}_0 + \hat{b}_1\text{advertising}_i + \hat{b}_2\text{airplay}_i + \hat{b}_3\text{image}_i \\ &= -26.61 + (0.085\ \text{advertising}_i) + (3.367\ \text{airplay}_i) + (11.086\ \text{image}_i)\end{aligned} \tag{9.26}$$

The $\hat{b}$s quantify the relationship between album sales and each predictor. The *direction* of the coefficient – positive or negative – corresponds to whether the relationship with the outcome is positive or negative. All three predictors have positive *b*-values, indicating positive relationships. So, as advertising budget, plays on the radio, and image rating increase so do album sales. The *size* of the $\hat{b}$s indicate the degree to which each predictor affects the outcome *if the effects of all other predictors are held constant*:

- **advertising budget**: $\hat{b}$ = 0.085 indicates that as advertising budget increases by one unit, album sales increase by 0.085 units. Both variables were measured in thousands; therefore, for every £1000 more spent on advertising, an extra 0.085 thousand albums (85 albums) are sold. This interpretation is true only if the effects of band image and airplay are held constant.
- **airplay**: $\hat{b}$ = 3.367 indicates that as the number of plays on radio in the week before release increases by one, album sales increase by 3.367 units. Every additional play of a song on radio (in the week before release) is associated with an extra 3.367 thousand albums (3367 albums) being sold. This interpretation is true only if the effects of the band's image and advertising budget are held constant.
- **image**: $\hat{b}$ = 11.086 indicates that if a band can increase their image rating by 1 unit they can expect additional album sales of 11.086 units. Every unit increase in the band's image rating is associated with an extra 11.086 thousand albums (11,086 albums) being sold. This interpretation is true only if the effects of airplay and advertising are held constant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	134.140	7.537		17.799	<.001	119.278	149.002
	Advertsing budget (thousands)	.096	.010	.578	9.979	<.001	.077	.115
2	(Constant)	-26.613	17.350		-1.534	.127	-60.830	7.604
	Advertsing budget (thousands)	.085	.007	.511	12.261	<.001	.071	.099
	No. of plays on radio	3.367	.278	.512	12.123	<.001	2.820	3.915
	Band image rating (0–10)	11.086	2.438	.192	4.548	<.001	6.279	15.894

a. Dependent Variable: Album sales (thousands)

Coefficients[a]

Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	Advertsing budget (thousands)	.578	.578	.578	1.000	1.000
2	Advertsing budget (thousands)	.578	.659	.507	.986	1.015
	No. of plays on radio	.599	.655	.501	.959	1.043
	Band image rating (0–10)	.326	.309	.188	.963	1.038

a. Dependent Variable: Album sales (thousands)

Output 9.8[16]

16 To spare your eyesight I have split this part of the output into two tables; however, it should appear as one long table.

Each $\hat{b}$ has an associated standard error indicating to what extent the estimate varies across different samples. The standard errors are also used to compute a *t*-statistic that tests whether the $\hat{b}$-value is significantly different from 0 (Section 9.2.5). Remember that if a predictor's $\hat{b}$ is zero then its relationship to the outcome is zero also. By testing whether an observed $\hat{b}$ is significantly different from zero, we're testing whether the relationship between the predictor and outcome is different from zero. The *p*-value associated with a $\hat{b}$'s *t*-statistic (in the column *Sig.*) is the probability of getting a *t* at least as big as the one we have if the population value of *b* was zero (i.e. if there was no relationship between that predictor and the outcome).

For this model, the advertising budget, $t(196) = 12.26$, $p < 0.001$, the amount of radio play prior to release, $t(196) = 12.12$, $p < 0.001$, and band image, $t(196) = 4.55$, $p < 0.001$, are all significant predictors of album sales.[17] Remember that these significance tests are accurate only if the assumptions discussed in Chapter 6 are met. From the magnitude of the *t*-statistics we can see that the advertising budget and radio play had a similar impact, whereas the band's image had less impact.

The standardized versions of the $\hat{b}$-values are useful if you want to compare predictors (because they are not dependent on the units of measurement of the variables). The standardized beta values (in the column labelled *Beta*, β_i) tell us the number of standard deviations that the outcome changes when the predictor changes by one standard deviation. Because the standardized beta values are measured in standard deviation units they are directly comparable: the values for airplay and advertising budget are virtually identical (0.512 and 0.511, respectively), suggesting that both variables have a comparably large effect, whereas image (standardized beta of 0.192) has a relatively smaller effect (this concurs with what the magnitude of the *t*-statistics told us). To interpret these values literally, we need to know the standard deviations of the variables, and these values can be found in Output 9.5.

- **advertising budget**: The standardized $\hat{\beta} = 0.511$ indicates that as advertising budget increases by one standard deviation (£485,655), album sales increase by 0.511 standard deviations. The standard deviation for album sales is 80,699, so this constitutes a change of 41,240 sales (0.511 × 80,699). Therefore, for every £485,655 more spent on advertising, an extra 41,240 albums are sold. This interpretation is true only if the effects of the band's image and airplay are held constant.
- **airplay**: The standardized $\hat{\beta} = 0.512$ indicates that as the number of plays on radio in the week before release increases by 1 standard deviation (12.27), album sales increase by 0.512 standard deviations. The standard deviation for album sales is 80,699, so this is a change of 41,320 sales (0.512 × 80,699). Basically, if the station plays the song an extra 12.27 times in the week before release, 41,320 extra album sales can be expected. This interpretation is true only if the effects of the band's image and advertising are held constant.
- **image**: The standardized $\hat{\beta} = 0.192$ indicates that a band rated one standard deviation (1.40 units) higher on the image scale can expect additional album sales of 0.192 standard deviations units. This is a change of 15,490 sales (0.192 × 80,699). A band with an image rating 1.40 higher than another band can expect 15,490 additional sales. This interpretation is true only if the effects of airplay and advertising are held constant.



17 For all of these predictors I wrote *t*(196). The number in brackets is the degrees of freedom. We saw in Section 9.2.5 that the degrees of freedom are $N-k-1$, where N is the total sample size (in this case 200) and k is the number of predictors (in this case 3). For these data we get $200-3-1 = 196$.

Output 9.8 also contains the confidence intervals for the $\hat{b}$s (again these are accurate only if the assumptions discussed in Chapter 6 are met). A bit of revision. Imagine that we collected 100 samples of data measuring the same variables as our current model. For each sample we estimate the same model that we have in this chapter, including confidence intervals for the unstandardized beta values. These boundaries are constructed such that in 95% of samples they contain the population value of b (see Section 2.9). Therefore, 95 of our 100 samples will yield confidence intervals for $\hat{b}$ that contain the population value. The trouble is that we don't know if our sample is one of the 95% with confidence intervals containing the population values or one of the 5% that miss.

The typical pragmatic solution to this problem is to assume that your sample is one of the 95% that hit the population value. If you assume this, then you can reasonably interpret the confidence interval as providing information about the population value of b. A narrow confidence interval suggests that all samples would yield estimates of b that are fairly close to the population value, whereas wide intervals suggest a lot of uncertainty about what the population value of b might be. If the interval contains zero then it suggests that the population value of b might be zero – in other words, no relationship between that predictor and the outcome – and could be positive but might be negative. All of these statements are reasonable *if you're prepared to believe that your sample is one of the 95% for which the intervals contain the population value. Your belief will be wrong 5% of the time, though.*

In our model of album sales, the two best predictors (advertising and airplay) have very tight confidence intervals, indicating that the estimates for the current model are likely to be representative of the true population values. The interval for the band's image is wider (but still does not cross zero) indicating that the parameter for this variable is less representative, but nevertheless significant.

If you asked for part and partial correlations, then they appear in separate columns of the table. The zero-order correlations are the Pearson's correlation coefficients and correspond to the values in Output 9.5. Semi-partial (part) and partial correlations were described in Section 8.5; in effect, the part correlations quantify the unique relationship that each predictor has with the



Cramming Sam's Tip
Coefficients

- The individual contribution of variables to the regression model can be found in the *Coefficients* table. If you have done a hierarchical regression then look at the values for the final model.
- You can see whether each predictor variable has made a significant contribution to predicting the outcome by looking at the column labelled *Sig.* (values less than or equal to 0.05 are significant).
- The standardized beta values tell you the importance of each predictor (bigger absolute value = more important).
- The tolerance and VIF values will also come in handy later, so make a note of them.



outcome. If you opted to do a stepwise regression, you would find that variable entry is based initially on the variable with the largest zero-order correlation and then on the part correlations of the remaining variables. Therefore, airplay would be entered first (because it has the largest zero-order correlation), then advertising budget (because its part correlation is bigger than that of image rating) and then finally the band's image rating – try running a forward stepwise regression on these data to see if I'm right. Finally, Output 9.8 contains collinearity statistics, but we'll discuss these in Section 9.11.5.

9.11.4 Excluded variables Ⓑ

At each stage of fitting a linear model a summary is provided of predictors that are not yet in the model.

We had a two-block hierarchy with one predictor entered (and two excluded) in block 1, and three predictors entered (and none excluded) in block 2. Output 9.9 details excluded variables only for the first block of our hierarchy, because in the second block no predictors were excluded. The table includes an estimate of the *b*-value and associated *t*-statistic for each predictor *if* it entered the model at this point. Using a stepwise method, the predictor with the highest *t*-statistic will enter the model next, and predictors will continue to be entered until there are none left with *t*-statistics that have significance values less than or equal to 0.05. The partial correlation also indicates what contribution (if any) an excluded predictor would make if it entered the model.

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance	Collinearity Statistics: VIF	Collinearity Statistics: Minimum Tolerance
1	No. of plays on radio	.546[b]	12.513	<.001	.665	.990	1.010	.990
	Band image rating (0–10)	.281[b]	5.136	<.001	.344	.993	1.007	.993

a. Dependent Variable: Album sales (thousands)

b. Predictors in the Model: (Constant), Advertsing budget (thousands)

Output 9.9

9.11.5 Assessing multicollinearity Ⓑ

I promised to come back to the measures of collinearity in Output 9.8 so here we go. The output contains the VIF and tolerance statistics (with tolerance being 1 divided by the VIF) and we need to apply the guidelines from Section 9.9.3. The VIF values are well below 10 and the tolerance statistics are well above 0.2. The average VIF, obtained by adding the VIF values for each predictor and dividing by the number of predictors (k), is also very close to 1:

$$\overline{\text{VIF}} = \frac{\sum_{i=1}^{k} \text{VIF}_i}{k} = \frac{1.015 + 1.043 + 1.038}{3} = 1.032 \tag{9.27}$$

It seems unlikely, therefore, that we need to worry about collinearity among predictors.

The other information we get about collinearity is a table of eigenvalues of the scaled, uncentred cross-products matrix, condition indexes and variance proportions. I discuss collinearity and variance proportions at length in Section 20.7.4, so here I'll just give you the headline: look for large variance proportions on the same *small* eigenvalues (Jane Superbrain Box 9.3). Therefore, in Output 9.10 inspect the bottom few rows of the table (these are the small eigenvalues) and look for variables that *both* have high variance proportions for that eigenvalue. The variance proportions vary between

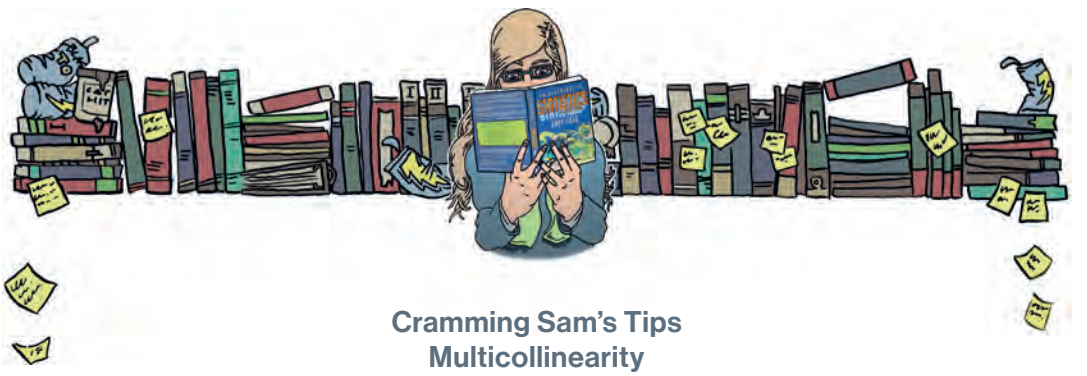
0 and 1, and you'd like to see each predictor having a high proportion on a different eigenvalue to other predictors (in other words, the large proportions are distributed across different eigenvalues). For our model, each predictor has most of its variance loading onto a different dimension than other predictors (advertising has 96% of variance on dimension 2, airplay has 93% of variance on dimension 3 and image rating has 92% of variance on dimension 4). These data represent no multicollinearity. For an example of when collinearity exists in the data and some suggestions about what can be done, see Chapters 20 (Section 20.7.4) and 18 (Section 18.3.3).

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Advertsing budget (thousands)	No. of plays on radio	Band image rating (0–10)
1	1	1.785	1.000	.11	.11		
	2	.215	2.883	.89	.89		
2	1	3.562	1.000	.00	.02	.01	.00
	2	.308	3.401	.01	.96	.05	.01
	3	.109	5.704	.05	.02	.93	.07
	4	.020	13.219	.94	.00	.00	.92

a. Dependent Variable: Album sales (thousands)

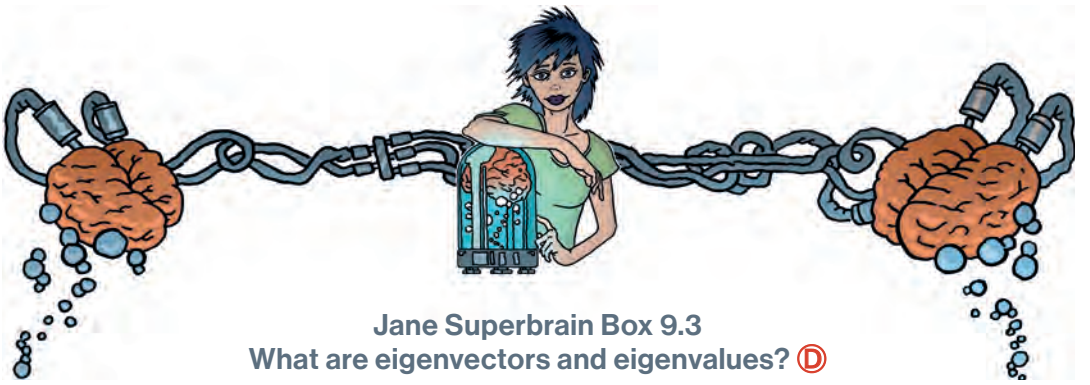
Output 9.10



Cramming Sam's Tips
Multicollinearity

- To check for multicollinearity, use the VIF values from the table labelled *Coefficients*.
- If these values are less than 10 then that indicates there probably isn't cause for concern.
- If you take the average of VIF values, and it is not substantially greater than 1, then there's also no cause for concern.





Jane Superbrain Box 9.3
What are eigenvectors and eigenvalues? Ⓓ

The definitions and mathematics of eigenvalues and eigenvectors are complicated and most of us need not worry about them (although they do crop up again in Chapters 17 and 18). Although the mathematics is hard, we can get a sense of what they represent visually. Imagine we have two variables: the age of a zombie (how long it has been a zombie), and how many blows to the head it takes to kill it.[18] These two variables are normally distributed and can be considered together as a bivariate normal distribution. If these variables are correlated their scatterplot forms an ellipse: if we draw a dashed line around the outer values of the scatterplot we get an oval shape (Figure 9.17). Imagine two lines to measure the length and height of this ellipse: these represent the *eigenvectors* of the correlation matrix for these two variables (a vector is a set of numbers that tells us the location of a line in geometric space). Note that the two straight lines in Figure 9.17 are at 90 degrees to each other, which means that they are independent of one another. So, with two variables, think of eigenvectors as lines measuring the length and height of the ellipse that surrounds the scatterplot of data for those variables. If we add a third variable (e.g., force of blow) our scatterplot gets a third dimension (depth), the ellipse turns into something shaped like a rugby ball (or American football), and we get an extra eigenvector to measure the extra dimension. If we add a fourth variable, a similar logic applies (although it's harder to visualize).

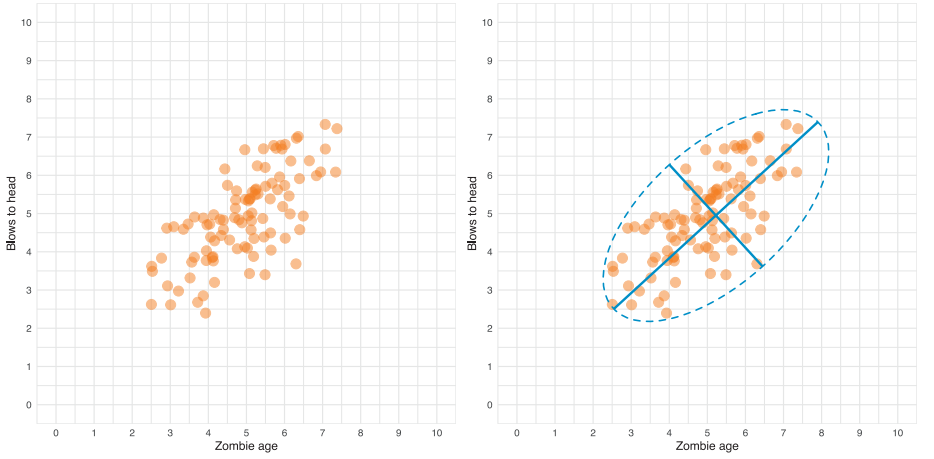


Figure 9.17 A scatterplot of two correlated variables forms an ellipse

Each eigenvector has an *eigenvalue* that tells us its length (i.e., the distance from one end of the eigenvector to the other). By looking at the eigenvalues for a data set, we know the dimensions of the ellipse (length and height) or rugby ball (length, height, depth); more generally, we know the dimensions of the data. Therefore, the eigenvalues quantify how evenly (or otherwise) the variances of the matrix are distributed.

18 Assuming you can ever kill a zombie, that is.

In the case of two variables, the *condition* of the data is related to the ratio of the larger eigenvalue to the smaller. Figure 9.18 shows the two extremes: when there is no relationship at all between variables (left), and when there is a perfect relationship (right). When there is no relationship, the data cloud will be contained roughly within a circle (or a sphere if we had three variables). If we draw lines that measure the height and width of this circle they'll be the same length, which means they'll have the same eigenvalues. Consequently, when we divide the largest eigenvalue by the smallest we'll get a value of 1. When the variables are perfectly correlated (i.e., there is perfect collinearity) the data cloud (and ellipse around it) will collapse to a straight line. The height of the ellipse will be very small indeed (it will approach zero). Therefore, the largest eigenvalue divided by the smallest will tend to infinity (because the smallest eigenvalue is close to zero). An infinite condition index is a sign of deep trouble.

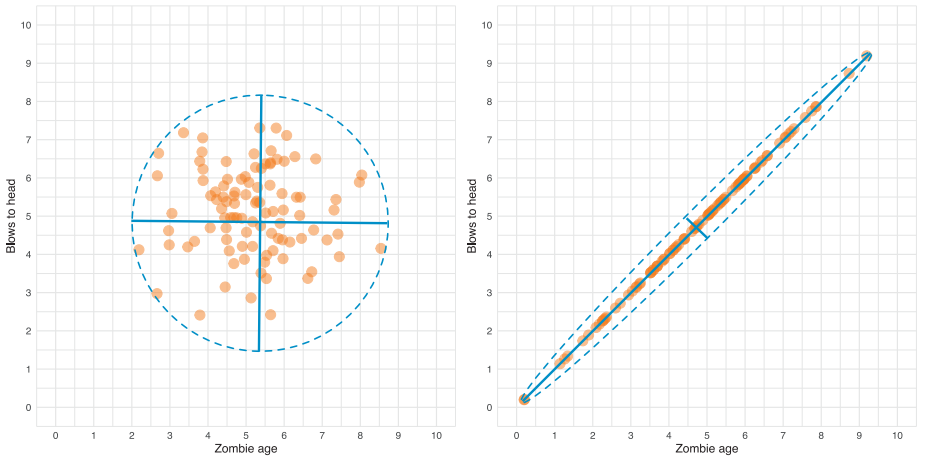


Figure 9.18 Perfectly uncorrelated (left) and correlated (right) variables



9.11.6 Bias in the model: casewise diagnostics Ⓑ

The final stage of the general procedure outlined in Figure 9.8 is to check the residuals for evidence of bias. The first step is to examine the casewise diagnostics. Output 9.11 shows any cases that have a standardized residual less than –2 or greater than 2 (remember that we changed the default criterion from 3 to 2 in Figure 9.13). In an ordinary sample we would expect 95% of cases to have standardized residuals within about ±2 (Jane Superbrain Box 6.6). We have a sample of 200, therefore it is reasonable to expect about 10 cases (5%) to have standardized residuals outside these limits. Output 9.11 shows that we have 12 cases (6%) that are outside of the limits: close to what we would expect. In addition, 99% of cases should lie within ±2.5 and only 1% of cases should lie outside these limits. We have two cases that lie outside of the limits (cases 164 and 169), which is 1% and what we would expect. These diagnostics give us no cause for concern except that case 169 has a standardized residual greater than 3, which is probably large enough for us to investigate this case further.

Casewise Diagnostics[a]

Case Number	Std. Residual	Album sales (thousands)	Predicted Value	Residual
1	2.125	330	229.92	100.080
2	-2.314	120	228.95	-108.949
10	2.114	300	200.47	99.534
47	-2.442	40	154.97	-114.970
52	2.069	190	92.60	97.403
55	-2.424	190	304.12	-114.123
61	2.098	300	201.19	98.810
68	-2.345	70	180.42	-110.416
100	2.066	250	152.71	97.287
164	-2.577	120	241.32	-121.324
169	3.061	360	215.87	144.132
200	-2.064	110	207.21	-97.206

a. Dependent Variable: Album sales (thousands)

Output 9.11

In Section 9.10.4 we opted to save various diagnostic statistics. You should find that the data editor contains columns for these variables. You can check these values in the data editor, or list values in the viewer window. To create a table of values in the viewer, select *Analyze* > *Reports* > Case Summaries... to access the dialog box in Figure 9.19. Select and drag the variables that you want to list into the box labelled *Variables* (or click). By default, the output is limited to the first 100 cases, but if you want to list all cases deselect this option (also see SPSS Tip 9.1). It is also useful to select *Show case numbers* to enable you to identify the case numbers of any problematic cases.

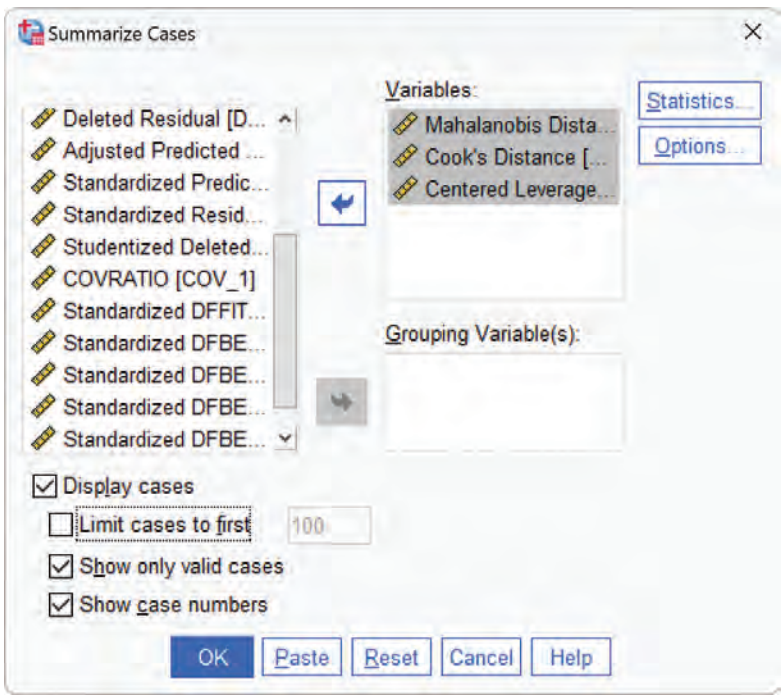
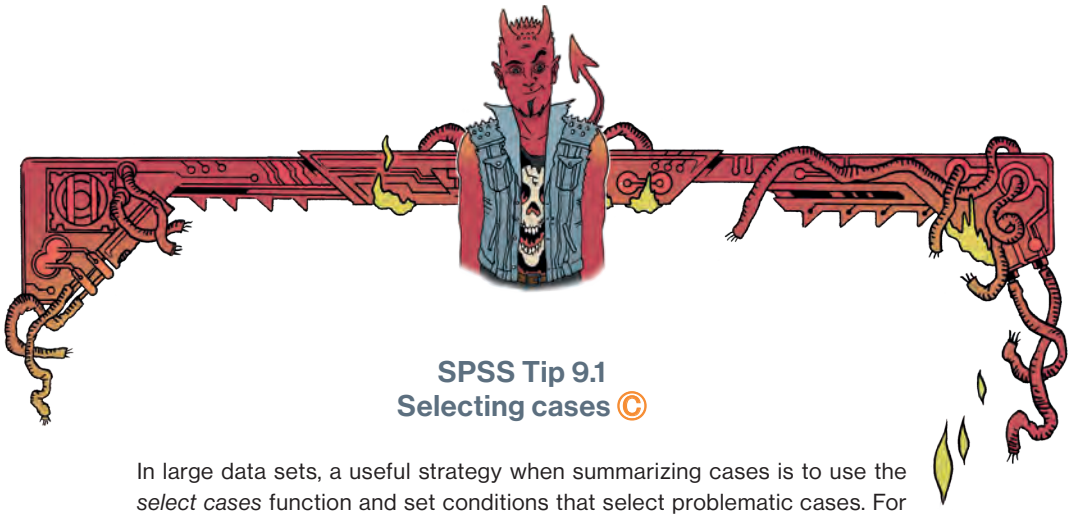


Figure 9.19 The *Summarize Cases* dialog box

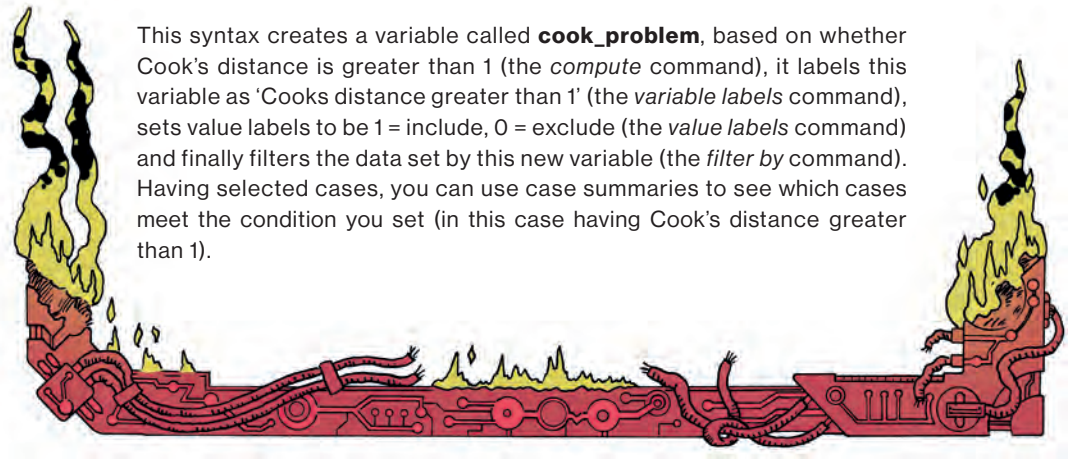


SPSS Tip 9.1
Selecting cases ©

In large data sets, a useful strategy when summarizing cases is to use the *select cases* function and set conditions that select problematic cases. For example, you could create a variable that selects cases with a Cook's distance greater than 1 by running the following syntax:

```
USE ALL.
COMPUTE cook_problem=(COO_1 > 1).
VARIABLE LABELS cook_problem 'Cooks distance greater than 1'.
VALUE LABELS cook_problem 0 'Not Selected' 1 'Selected'.
FILTER BY cook_problem.
EXECUTE.
```

This syntax creates a variable called **cook_problem**, based on whether Cook's distance is greater than 1 (the *compute* command), it labels this variable as 'Cooks distance greater than 1' (the *variable labels* command), sets value labels to be 1 = include, 0 = exclude (the *value labels* command) and finally filters the data set by this new variable (the *filter by* command). Having selected cases, you can use case summaries to see which cases meet the condition you set (in this case having Cook's distance greater than 1).



To save space, Output 9.12 shows the influence statistics for 12 cases that I selected. None of them has a Cook's distance greater than 1 (even case 169 is well below this criterion) and so no case appears to have an undue influence on the model. The average leverage can be calculated as $(k + 1)/n = 4/200 = 0.02$, and we should look for values either twice (0.04) or three times (0.06) this value (see Section 9.3.2). All cases are within the boundary of three times the average, and only case 1 is close to twice the average. For the Mahalanobis distances we saw earlier in the chapter with a sample of 100 and three predictors, values greater than 15 are problematic. Also, with three predictors values greater than 7.81 are significant ($p \leq 0.05$). None of our cases come close to exceeding the criterion of 15, although case 1 would be deemed 'significant'.

Case Summaries

	Case Number	COVRATIO	Standardized DFFIT	Standardized DFBETA Intercept	Standardized DFBETA adverts	Standardized DFBETA airplay	Standardized DFBETA image
1	1	.97127	.48929	-.31554	-.24235	.15774	.35329
2	2	.92018	-.21110	.01259	-.12637	.00942	-.01868
3	10	.94392	.26896	-.01256	-.15612	.16772	.00672
4	47	.91458	-.31469	.06645	.19602	.04829	-.17857
5	52	.95995	.36742	.35291	-.02881	-.13667	-.26965
6	55	.92486	-.40736	.17427	-.32649	-.02307	-.12435
7	61	.93654	.15562	.00082	-.01539	.02793	.02054
8	68	.92370	-.30216	-.00281	.21146	-.14766	-.01760
9	100	.95888	.35732	.06113	.14523	-.29984	.06766
10	164	.92037	-.54029	.17983	.28988	-.40088	-.11706
11	169	.85325	.46132	-.16819	-.25765	.25739	.16968
12	200	.95435	-.31985	.16633	-.04639	.14213	-.25907
Total N		12	12	12	12	12	12

Case Summaries

	Case Number	Mahalanobis Distance	Cook's Distance	Centered Leverage Value
1	1	8.39591	.05870	.04219
2	2	.59830	.01089	.00301
3	10	2.07154	.01776	.01041
4	47	2.12475	.02412	.01068
5	52	4.81841	.03316	.02421
6	55	4.19960	.04042	.02110
7	61	.06880	.00595	.00035
8	68	2.13106	.02229	.01071
9	100	4.53310	.03136	.02278
10	164	6.83538	.07077	.03435
11	169	3.14841	.05087	.01582
12	200	3.49043	.02513	.01754
Total N		12	12	12

Output 9.12

The DFBeta statistics tell us how much influence each case has on the model parameters. An absolute value greater than 1 is a problem and all cases in Output 9.12 have values within ±1, which is good news. For the covariance ratio we need to use the following criteria (Section 9.3.2):

- $CVR_i > 1 + [3(k + 1)/n] = 1 + [3(3 + 1)/200] = 1.06$
- $CVR_i < 1 - [3(k + 1)/n] = 1 - [3(3 + 1)/200] = 0.94$.

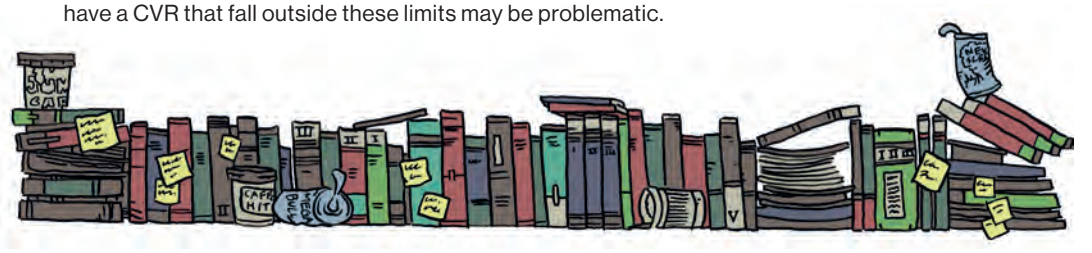
Therefore, we are looking for any cases that deviate substantially from these boundaries. Most of our 12 potential outliers have CVR values within or just outside these boundaries. The only case that causes concern is case 169 (again), whose CVR is some way below the lower limit. However, given the Cook's distance for this case, there is probably little cause for alarm. You will have requested other diagnostic statistics and you can apply what we learnt earlier in the chapter when glancing over them.

From this minimal set of diagnostics there's nothing to suggest that there are influential cases (although we'd need to look at all 200 cases to confirm this conclusion); we appear to have a fairly reliable model that has not been unduly influenced by any subset of cases.



Cramming Sam's Tips
Residuals

- Look for cases that might be influencing the model.
- Look at standardized residuals and check that no more than 5% of cases have absolute values above 2, and that no more than about 1% have absolute values above 2.5. Any case with a value above about 3 could be an outlier.
- Look in the data editor for the values of Cook's distance: any value above 1 indicates a case that might be influencing the model.
- Calculate the average leverage and look for values greater than twice or three times this average value.
- For Mahalanobis distance, a crude check is to look for values above 25 in large samples (500) and values above 15 in smaller samples (100). However, Barnett and Lewis (1978) should be consulted for more refined guidelines.
- Look for absolute values of DFBeta greater than 1.
- Calculate the upper and lower limit of acceptable values for the covariance ratio, CVR. Cases that have a CVR that fall outside these limits may be problematic.



9.11.7 Bias in the model: assumptions Ⓑ

The general procedure outlined in Figure 9.8 suggests that, having fitted a model, we need to look for evidence of bias, and the second stage of this process is to check the assumptions described in Chapter 6. We saw in Section 6.9.6 that we can look for heteroscedasticity and non-linearity using a plot of standardized predicted values against standardized residuals. We asked for this plot in Section 9.10.3. If everything is OK then this graph should look like a random array of dots. Figure 9.20 (top left) shows the plot for our model. Note how the points are randomly and evenly dispersed throughout the plot. This pattern is indicative of a situation in which the assumptions of linearity and homoscedasticity have been met; compare it with the examples in Figure 6.27.

Figure 9.20 also shows the partial plots, which are scatterplots of the residuals of the outcome variable and each of the predictors when both variables are regressed separately on the remaining predictors. Obvious outliers on a partial plot represent cases that might have undue influence on a predictor's $\hat{b}$. Non-linear relationships and heteroscedasticity can be detected using these plots as well. For advertising budget (Figure 9.20, top right) the partial plot shows the strong positive relationship

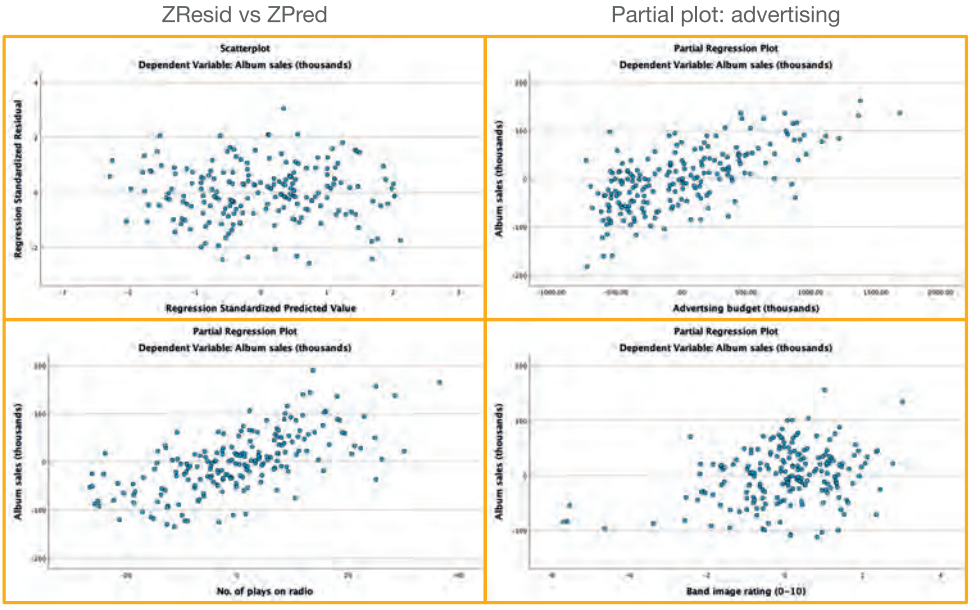


Figure 9.20 Plot of standardized predicted values against standardized residuals (top left), and partial plots of album sales against advertising (top right), airplay (bottom left) and image of the band (bottom right)

to album sales. There are no obvious outliers and the cloud of dots is evenly spaced out around the line, again indicating homoscedasticity. The plot for airplay (Figure 9.20, bottom left) also shows a strong positive relationship to album sales, there are no obvious outliers, and the cloud of dots is evenly spaced around the line, indicating homoscedasticity. For image (Figure 9.20, bottom right) the plot again shows a positive relationship to album sales, but the dots show funnelling, indicating greater spread for bands with a high image rating. There are no obvious outliers on this plot, but the funnel-shaped cloud indicates a violation of the assumption of homoscedasticity.

To test the normality of residuals, we look at the histogram and normal probability plot selected in Figure 9.14 and shown in Figure 9.21. Compare these plots to examples of non-normality in Section 6.9.5. For the album sales data, the distribution is very normal: the histogram is symmetrical and approximately bell-shaped. In the P-P plot the dots lie almost exactly along the diagonal, which we know indicates a normal distribution (see Section 6.9.4); hence this plot also suggests that the residuals are normally distributed.

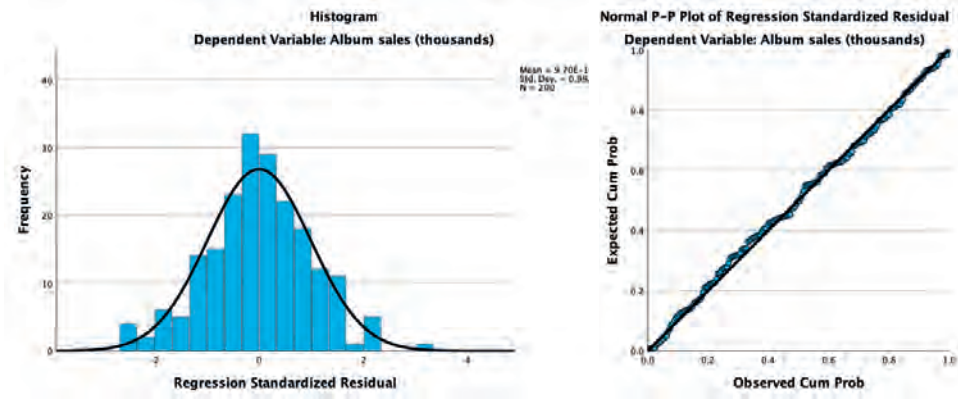
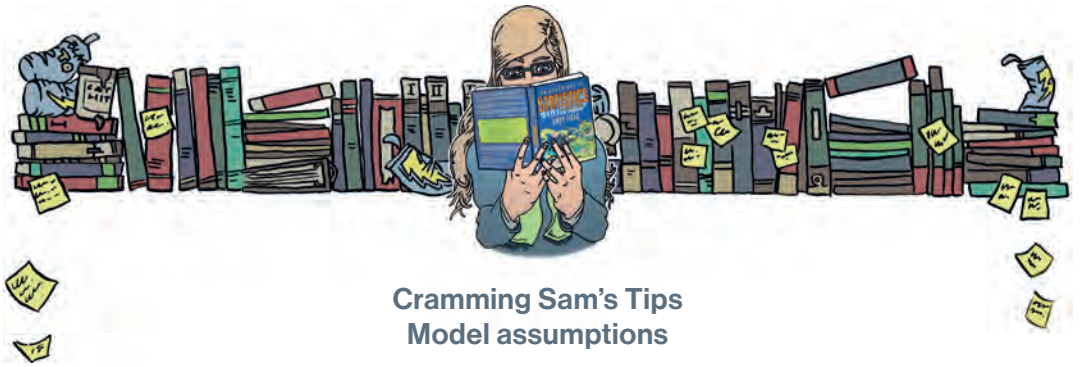


Figure 9.21 Histogram and normal P-P plot for the residuals from our model



Cramming Sam's Tips
Model assumptions

- Look at the plot of **ZRESID*** against **ZPRED***. If it looks like a random array of dots then this is good. If the dots get more or less spread out over the graph (looks like a funnel) then the assumption of homogeneity of variance is probably unrealistic. If the dots have a pattern to them (i.e., a curved shape) then the assumption of linearity is probably not true. If the dots seem to have a pattern and are more spread out at some points on the plot than others then this could reflect violations of both homogeneity of variance *and* linearity. Any of these scenarios puts the validity of your model into question. Repeat the above for all partial plots too.
- Look at the histogram and P-P plot. If the histogram looks like a normal distribution (and the P-P plot looks like a diagonal line), then all is well. If the histogram looks non-normal and the P-P plot looks like a wiggly snake curving around a diagonal line then things are less good. Be warned, though: distributions can look very non-normal in small samples even when they are normal.



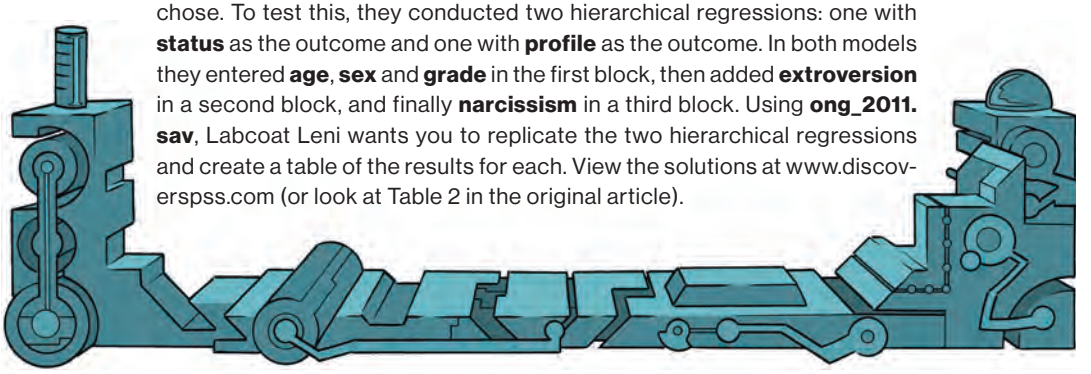


Labcoat Leni's Real Research 9.1
I want to be loved (on Facebook) Ⓐ

Ong et al. (2011). *Personality and Individual Differences*, 50(2), 180–185.

Social media websites such as Facebook offer an unusual opportunity to carefully manage your self-presentation to others (i.e., you can appear rad when in fact you write statistics books and wear 1980s heavy metal band T-shirts). Ong et al. (2011) examined the relationship between narcissism and behaviour on Facebook in 275 adolescents. They measured the **age**, **sex** and **grade** (at school), as well as **extroversion** and **narcissism**. They also measured how often (per week) these people updated

their Facebook status (**status**), and also how they rated their own profile picture on each of four dimensions: coolness, glamour, fashionableness and attractiveness. These ratings were summed as an indicator of how positively they perceived the profile picture they had selected for their page (**profile**). Ong et al. hypothesized that narcissism would predict the frequency of status updates, and how positive a profile picture the person chose. To test this, they conducted two hierarchical regressions: one with **status** as the outcome and one with **profile** as the outcome. In both models they entered **age**, **sex** and **grade** in the first block, then added **extroversion** in a second block, and finally **narcissism** in a third block. Using **ong_2011.sav**, Labcoat Leni wants you to replicate the two hierarchical regressions and create a table of the results for each. View the solutions at www.discoverspss.com (or look at Table 2 in the original article).



9.12 Robust regression Ⓑ

Our model appears, in most senses, to be both accurate for the sample and generalizable to the population. The only slight glitch is some concern over whether image ratings violated the assumption of homoscedasticity. Therefore, we could conclude that in our sample advertising budget and airplay are equally important in predicting album sales. The image of the band is a significant predictor of album sales but is less important than the other predictors (and probably needs verification because of possible heteroscedasticity). The assumptions seem to have been met and so we can probably assume that this model would generalize to any album being released. You won't always (ever?) have such nice data: there will be times when you uncover problems that cast a dark shadow of evil over your model. It will invalidate significance tests, confidence intervals and generalization of the model (use Chapter 6 to remind yourself of the implications of violating model assumptions).

Luckily, a lot of the problems can be overcome. If confidence intervals and significance tests of the model parameters are in doubt then use bootstrapping to generate confidence intervals and *p*-values. If homogeneity of variance is the issue then estimate the model with standard errors designed for heteroscedastic residuals (Hayes & Cai, 2007) – you can do this using the PROCESS tool described in Chapter 11. Finally, if the model parameters themselves are in doubt, estimate them using robust regression.

To get robust confidence intervals and significance tests of the model parameters re-estimate your model, selecting the same options as before but clicking Bootstrap... in the main dialog box (Figure 9.10) to access the dialog box explained in Section 6.10.8. To recap, select ☑ Perform bootstrapping to activate bootstrapping, and to get a 95% confidence interval click ◉ Percentile or ◉ Bias corrected accelerated (BCa). For this analysis, let's ask for a bias corrected (BCa) confidence interval. Bootstrapping won't work if you have set options to save diagnostics, so click Save... to open the dialog box in Figure 9.15 and *deselect everything*. Back in the main dialog box click OK to estimate the model.

The output will contain a table of bootstrap confidence intervals for each predictor and their significance value (Output 9.13).[19] These tell us that advertising, $\hat{b} = 0.09$ [0.07, 0.10], $p < 0.001$, airplay,

19 Remember that because of how bootstrapping works the values in your output will be different than mine, and different if you rerun the analysis.

Bootstrap for Coefficients

Model		B	Bias	Std. Error	Bootstrap[a] Sig. (2-tailed)	BCa 95% Confidence Interval Lower	Upper
1	(Constant)	134.140	-.116	7.952	<.001	120.108	148.793
	Advertsing budget (thousands)	.096	.000	.008	<.001	.079	.112
2	(Constant)	-26.613	.489	16.295	.097	-55.403	8.595
	Advertsing budget (thousands)	.085	.000	.007	<.001	.072	.098
	No. of plays on Radio	3.367	.010	.321	<.001	2.735	4.022
	Band Image Rating (0-10)	11.086	-.119	2.221	<.001	6.458	15.013

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 9.13

$\hat{b} = 3.37$ [2.74, 4.02], $p < 0.001$, and the band's image, $\hat{b} = 11.09$ [6.46, 15.01], $p < 0.001$, all significantly predict album sales. These bootstrap confidence intervals and significance values do not rely on assumptions of normality or homoscedasticity, so they give us an accurate estimate of the population value of b for each predictor (assuming our sample is one of the 95% with confidence intervals that contain the population value).

We can estimate the bs using a robust method using the R integration (Section 4.13.2) to access a function that uses an M-estimator instead of OLS. First, we have to activate the extension that gives us access to this function by selecting *Extensions* > *Extension Hub* ... to access the dialog box in Figure 9.22. Enter a search term (the word 'Robust' will find this extension) and restrict the search to extinctions that use R. The result should be an extension called *Robust Regression*. Check the box to get the extension and click on OK. Another dialog box will appear with a license, and you'll need to select ◉ I accept the terms in the license agreement. and then click Finish to install the extension.

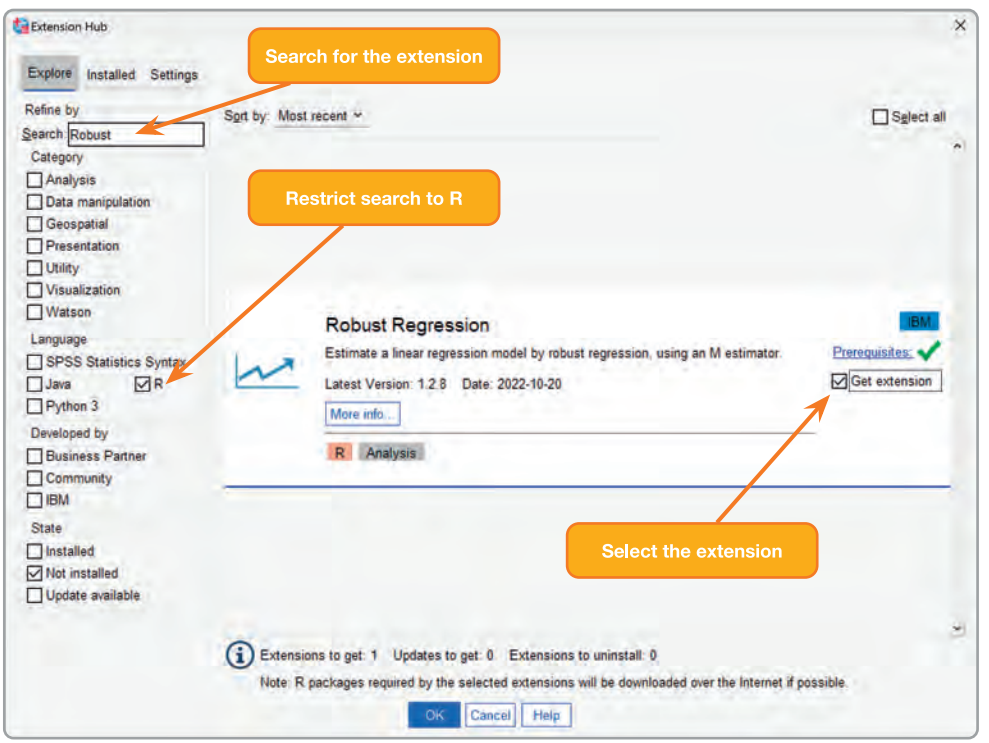


Figure 9.22 Dialog box for the extension hub

Having installed the extension, you can fit a robust regression by selecting *Analyze* > *Regression* > Robust Regression... If you haven't installed the extension this menu option won't be there! Drag the outcome (album sales) to the box labelled *Dependent* and any predictors in the final model (in this case **adverts**, **airplay** and **image**) to the box labelled *Independent Variables*. Click OK to estimate the model.

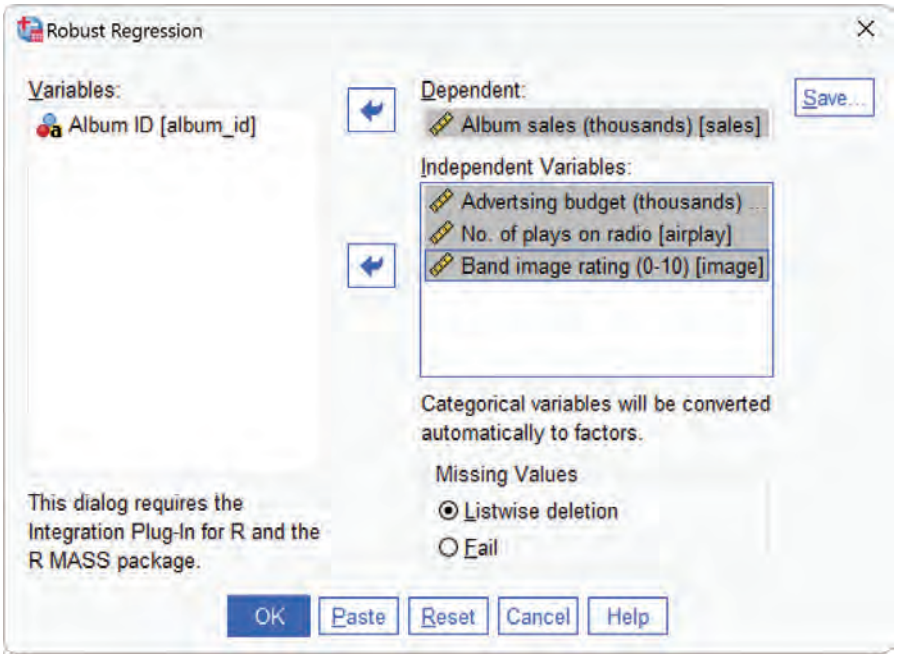


Figure 9.23 Dialog box for robust regression

Output 9.14 shows the resulting robust $\hat{b}$-values, their robust standard errors and t-statistics. Compare these with the non-robust versions in Output 9.8. The values are not much different (mainly because our original model didn't seem to violate its assumptions); for example, the $\hat{b}$ for image rating has changed from 11.09 (Output 9.8) to 11.39 (Output 9.14), the associated standard error was 2.44 and the robust version is 2.47, and the associated t-statistic has changed from 4.55 to 4.62. Essentially our interpretation of the model won't have changed, but this is still a useful sensitivity analysis in that if robust estimates are giving us basically the same results as non-robust estimates then we know that the non-robust estimates have not been unduly biased by properties of the data. So, this analysis is always a useful double check (Field & Wilcox, 2017), and if the robust estimates are hugely different from the original estimates then you can use and report the robust versions.

Coefficients

	Value	Std. Error	t value
(Intercept)	-28.858	17.569	-1.642
adverts	.086	.007	12.219
airplay	3.371	.281	11.984
image	11.394	2.469	4.615

rlm(formula = sales ~ adverts+airplay+image, data = dta, na.action = na.exclude, method = "MM", model = FALSE)
Residual standard error: 45.37370
Degrees of freedom: 196

Output 9.14

9.13 Bayesian regression ©

In Section 3.8 we looked at Bayesian approaches. To access a dialog box (Figure 9.24) to fit a Bayesian linear model select *Analyze* > *Bayesian Statistics* > Linear Regression. You can fit the model either using default priors (so-called Reference priors), which set distributions that represent very diffuse prior beliefs, or conjugate priors, which allow you to specify more specific priors. One of the key strengths of Bayesian statistics (in my opinion) is that you can set evidence-based priors that you update with the data that you collect. However, this is not a trivial undertaking, and it requires a deeper understanding of the models being fit than we have covered. So, to help you to dip your toe in the water of Bayesian statistics we will stick to using the reference priors built into SPSS. The benefit of reference priors is that they enable you to get going with Bayesian models without drowning in a lot of quite technical material, but the cost is that you are building only very uninformative prior information into your models.[20]

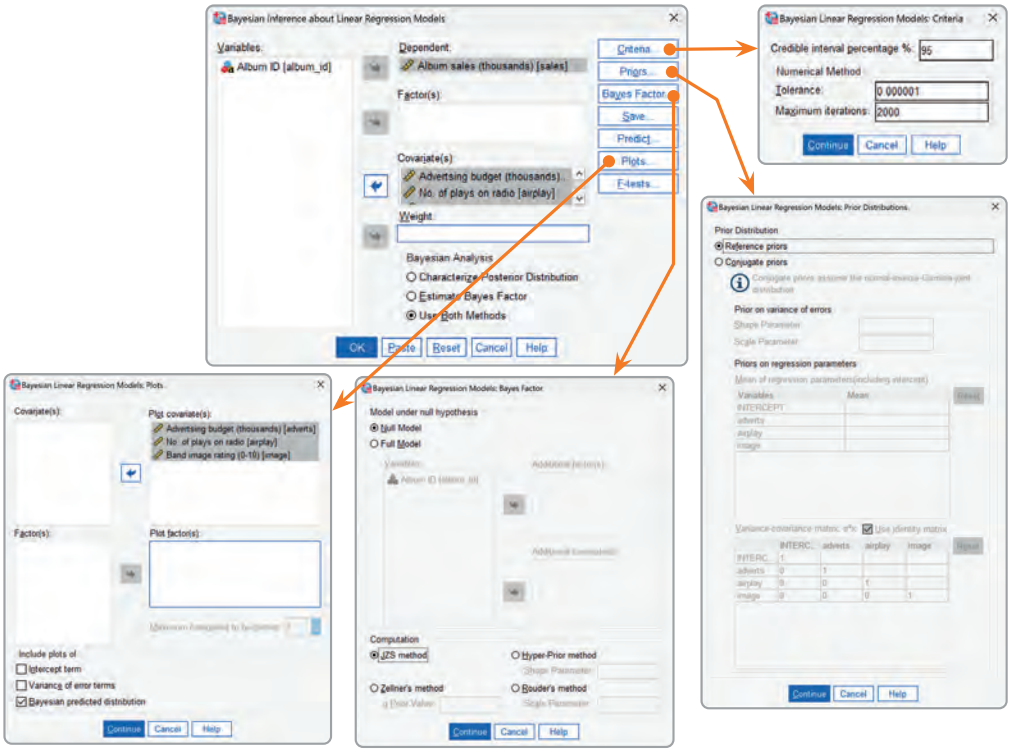


Figure 9.24 Dialog box for Bayesian regression

In the main dialog box (Figure 9.24) drag **sales** to the box labelled *Dependent* (or click ▸) and drag **adverts**, **airplay** and **image** to the *Covariate(s)* box (or click ▸). If your model has categorical predictors (which we'll look at in the following chapter) drag them to the *Factor(s)* box. If you want to both compute Bayes factors and estimate the model parameters then select Use Both Methods.

If you want a credible interval other than 95% then click Criteria... and change the 95 to the value you want. Click Priors... to set your priors, although we'll stick with Reference priors. Click Bayes Factor... if

20 Another downside of this convenience is that I find it hard to know what these priors actually represent (especially in the case of regression).

you want to get a Bayes factor for your model. By default, the full model will be compared the null model and there are four methods to compute them. I have selected ◉ JZS method (Jeffreys, 1961; Zellner & Siow, 1980). Click Plots... to inspect the prior and posterior distributions for each predictor. Drag all predictors to the box labelled *Plot covariate(s)* (or click ▸) and select ☑ Bayesian predicted distribution. In the main dialog box click OK to fit the model.

Output 9.15 (top) shows the Bayes factor for the full model compared to the null model, which is the model including only the intercept. The Bayes factor is 1.066×10^{43} (that's what the E+43 means). In other words, it is massive. The probability of the data given the model including all three predictors is 1.07×10^{43} greater than the probability of the data given the model with only the intercept. We should shift our belief in the model (relative to the null model) by a factor of 1.07×10^{43}! This is very strong evidence for the model.

Bayes Factor Model Summary[a,b]

Bayes Factor[c]	R	R Square	Adjusted R Square	Std. Error of the Estimate
1.066E+43	.815	.665	.660	47.09

a. Method: JZS

b. Model: (Intercept), Advertsing budget (thousands), No. of plays on radio, Band image rating (0–10)

c. Bayes factor: Testing model versus null model (Intercept).

Bayesian Estimates of Coefficients[a,b,c]

Parameter	Posterior Mode	Posterior Mean	Posterior Variance	95% Credible Interval Lower Bound	95% Credible Interval Upper Bound
(Intercept)	-26.613	-26.613	304.126	-60.830	7.604
Advertsing budget (thousands)	.085	.085	.000	.071	.099
No. of plays on radio	3.367	3.367	.078	2.820	3.915
Band image rating (0–10)	11.086	11.086	6.004	6.279	15.894

a. Dependent Variable: Album sales (thousands)

b. Model: (Intercept), Advertsing budget (thousands), No. of plays on radio, Band image rating (0–10)

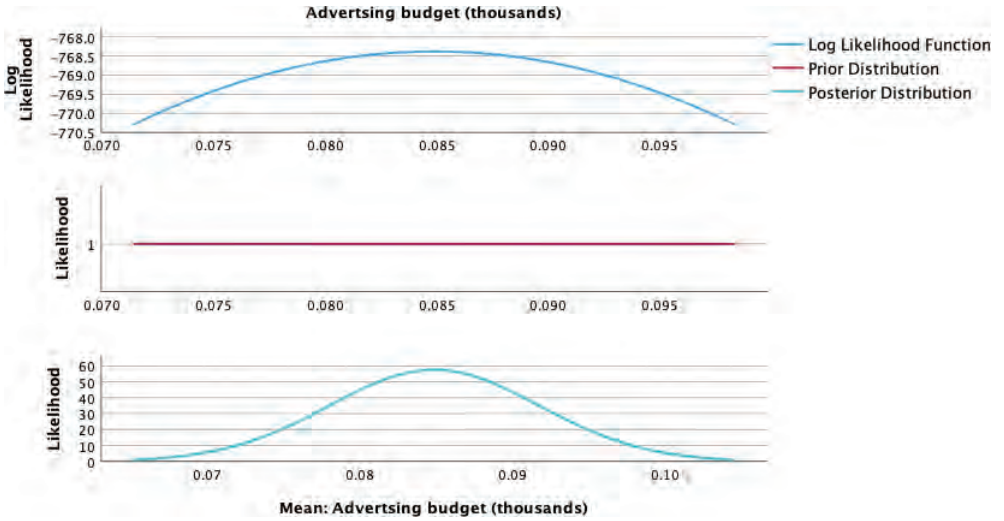
c. Assume standard reference priors.

Output 9.15

Output 9.15 (bottom) shows the $\hat{b}$s based on Bayesian estimation in the columns labelled *Posterior Mode* and *Posterior Mean*. The columns contain identical values, but they won't always. The reason for the two columns is that we use the peak of the posterior distribution as the estimate and that peak can be defined by either the mode of the posterior or its mean. The values are 0.085 for advertising budget, 3.367 for airplay and 11.086 for image, compared to the values of 0.085, 3.37 and 11.09 (Output 9.8) from the non-Bayesian model. They are basically the same, which is not all surprising because we started off with very diffuse priors and with diffuse priors the estimates will be driven almost entirely by the data rather than the priors (think back to Section 3.8). We can see this fact in Output 9.16, which shows the prior distribution for the $\hat{b}$ for advertising budget as a red line (you will see similar plots for the other two predictors in your output): the line is completely flat, representing a completely open and diffuse belief about the potential values of $\hat{b}$. The green line is the posterior distribution which is quantified in Output 9.15 (right).

Perhaps the most useful parts of Output 9.15 are the 95% credible intervals for the model parameters. Unlike confidence intervals, credible intervals contain the population value with a probability of 0.95 (95%). For advertising budget, therefore, there is a 95% probability that the population value of b lies between 0.071 and 0.099, for airplay the population value is plausibly between 2.820 and

3.915 and for image it plausibly lies between 6.279 and 15.894. These intervals are constructed assuming that an effect exists, so you cannot use them to test hypotheses, only to establish plausible population values of the *b*s in the model.



Output 9.16

9.14 Reporting linear models Ⓑ

If your model has several predictors then you can't really beat a summary table as a concise way to report the model. As a bare minimum report the betas along with their standard errors and confidence interval (or credible interval if you've gone Bayesian). If you haven't gone Bayesian, report the significance value and perhaps the standardized beta. Include some general fit statistics about the model such as the R^2 or the Bayes factor. Personally, I like to see the constant as well because then readers of your work can construct the full model if they need to. For hierarchical regression you should report these values at each stage of the hierarchy. For the example in this chapter we might produce a table like that in Table 9.2.

Things to note are as follows: (1) I've rounded off to 2 decimal places throughout because this is a reasonable level of precision given the variables measured; (2) if you are following APA format (which I'm not), do not put zeros before the decimal point for the standardized betas, R^2 and *p*-values (because these values shouldn't exceed 1); (3) I've reported exact *p*-values, which is good practice; (4) the R^2 for the initial model and the change in R^2 (denoted by ΔR^2) for each subsequent step of the model are reported below the table; and (5) in the title I have mentioned that confidence intervals and standard errors in the table are based on bootstrapping, which is important for readers to know.

9.15 Jane and Brian's story Ⓐ

Jane put the fork down next to the jar and suppressed her gag reflex. When she'd started at this university she'd had utter conviction in testing her flatworm theory. She would be her own single-case design. She knew that experimenting on herself would confound everything, but she wanted some evidence to firm up her beliefs. If it didn't work on her then she could move on, but if she found

Table 9.2 Linear model of predictors of album sales. 95% bias corrected and accelerated confidence intervals reported in parentheses. Confidence intervals and standard errors based on 1000 bootstrap samples

		b	SE B	β	p
Step 1					
	Constant	134.14 (118.70, 150.75)	8.09		$p < 0.001$
	Advertising Budget	0.10 (0.08, 0.11)	0.01	0.58	$p < 0.001$
Step 2					
	Constant	−26.61 (−55.40, 8.60)	16.30		$p = 0.097$
	Advertising Budget	0.09 (0.07, 0.10)	0.01	0.51	$p < 0.001$
	Plays on BBC radio 1	3.37 (2.74, 4.02)	0.32	0.51	$p < 0.001$
	Image	11.09 (6.46, 15.01)	2.22	0.19	$p < 0.001$

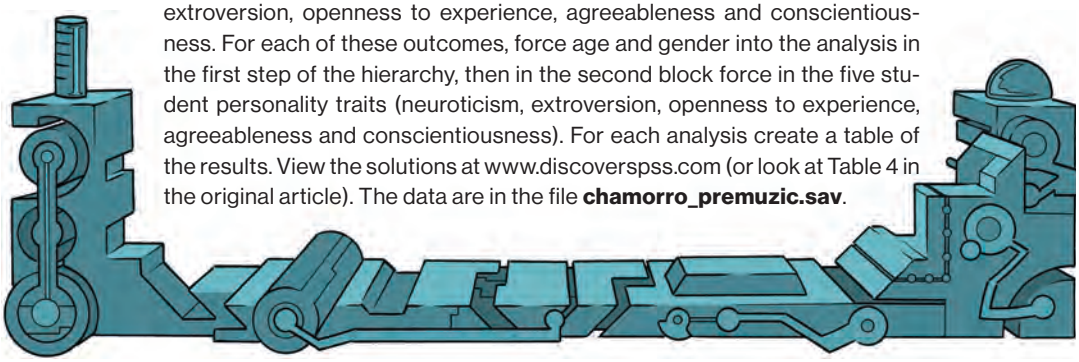
Note. $R^2 = 0.34$ for Step 1; $\Delta R^2 = 0.33$ for Step 2 (all ps < 0.001).



Labcoat Leni's Real Research 9.2
Why do you like your lecturers? Ⓐ

Chamorro-Premuzic, T., et al. (2008). *Personality and Individual Differences*, *44,* 965–976.

In the previous chapter we encountered a study by Chamorro-Premuzic et al. that linked students' personality traits with those they want to see in lecturers (see Labcoat Leni's Real Research 8.1 for a full description). In that chapter we correlated these scores, but now Labcoat Leni wants you to carry out five multiple regression analyses: the outcome variables across the five models are the ratings of how much students want to see neuroticism, extroversion, openness to experience, agreeableness and conscientiousness. For each of these outcomes, force age and gender into the analysis in the first step of the hierarchy, then in the second block force in the five student personality traits (neuroticism, extroversion, openness to experience, agreeableness and conscientiousness). For each analysis create a table of the results. View the solutions at www.discoverspss.com (or look at Table 4 in the original article). The data are in the file **chamorro_premuzic.sav**.



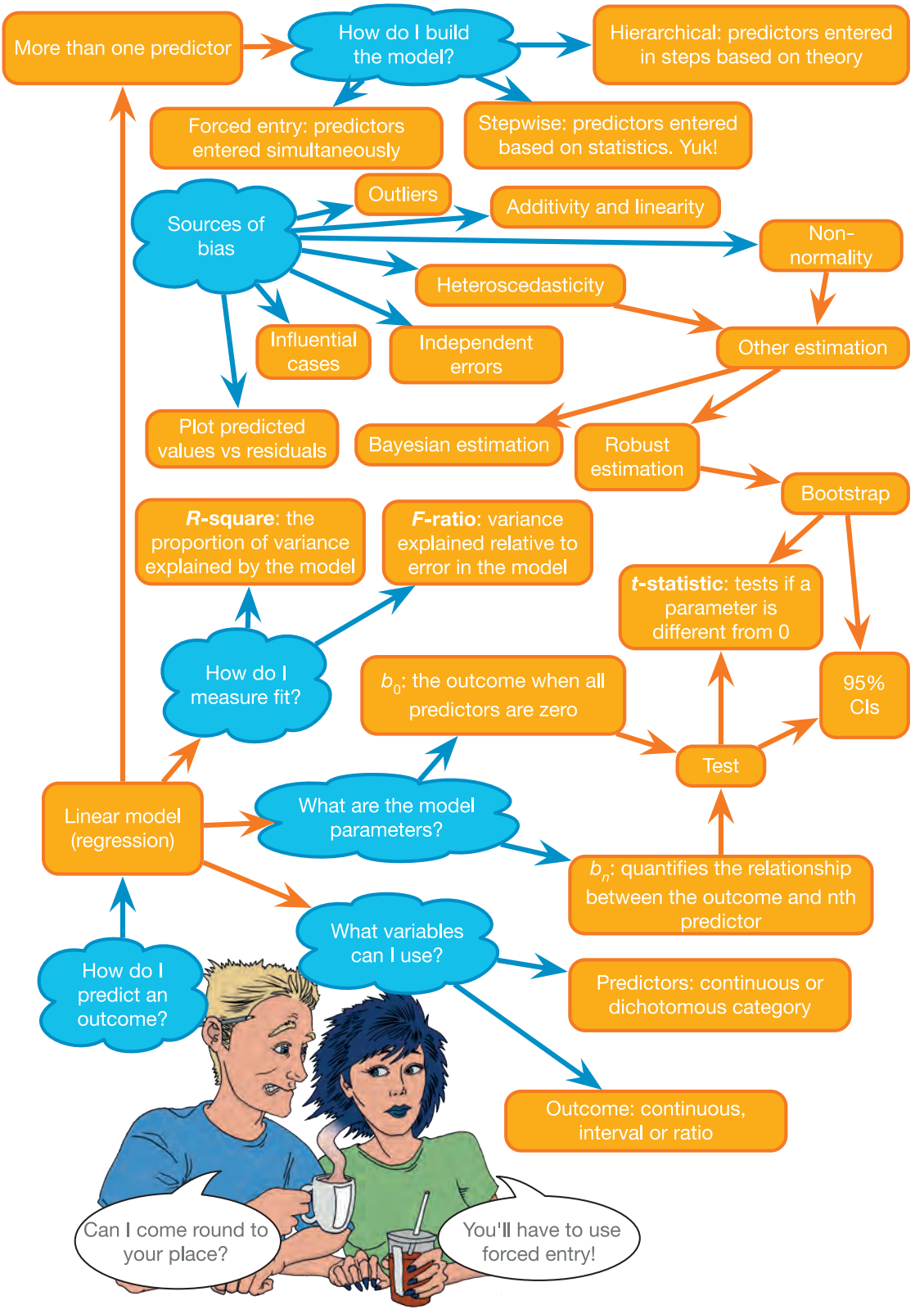


Figure 9.25 What Brian learnt from this chapter

evidence for some effect then that was starting point for better research. She felt conflicted, though. Was it the experiments making her mind so unfocused, or was it the interest from campus guy? She hadn't come here looking for a relationship, she hadn't expected it, and it wasn't in the plan. Usually she was so good at ignoring other people, but his kindness was slowly corroding her shell. As she got up and replaced the jar on the shelf she told herself that the nonsense with campus guy had to stop. She needed to draw a line.

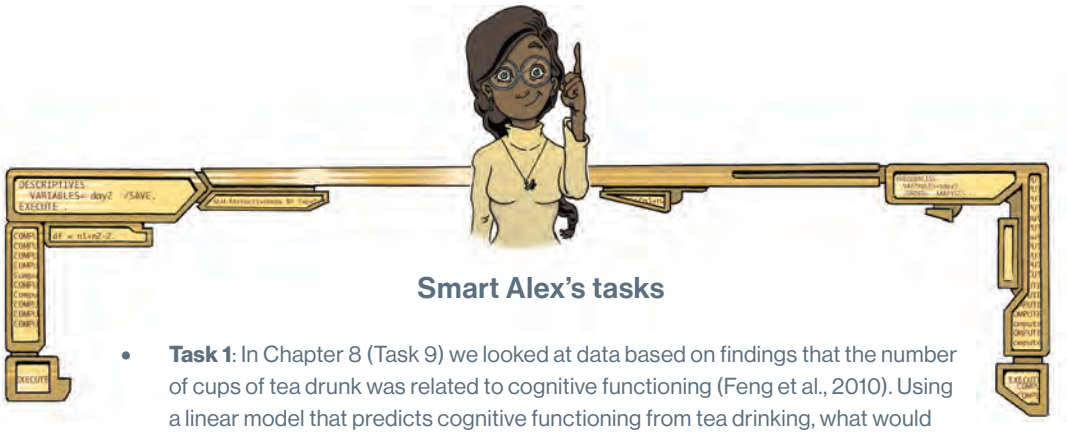
9.16 What next? Ⓐ

This chapter is possibly the longest book chapter ever written since Chapter 6, and if you feel like you aged several years while reading it then, well, you probably have (look around, there are cobwebs in the room, you have a long beard, and when you go outside you'll discover a second ice age has been and gone, leaving only you and a few woolly mammoths to populate the planet). However, on the plus side, you now know more or less everything you'll ever need to know about statistics. Seriously; you'll discover in the coming chapters that everything else we discuss is a variation of this chapter. So, although you may be near death having spent your life reading this chapter (and I'm certainly near death having written it) you are officially a stats genius – well done!

We started the chapter by discovering that at 8 years old I could have really done with a linear model to tell me which variables are important in predicting talent competition success. Unfortunately I didn't have one, but I did have my dad (and he's better than a linear model). He correctly predicted the recipe for superstardom but in doing so he made me hungry for more. I was starting to get a taste for the rock-idol lifestyle: I had friends, a fortune (well, two fake-gold-plated winner's medals), fast cars (a bike) and dodgy-looking 8-year-olds were giving me suitcases full of lemon sherbet to lick off mirrors. The only thing needed to complete the job was a platinum-selling album. However, before I could get this my parents and teachers were about to impress reality upon my young mind …

9.17 Key terms that I've discovered

Adjusted predicted value
Adjusted R^2
Akaike information criterion (AIC)
Autocorrelation
β_i
Cook's distance
Covariance ratio (CVR)
Cross-validation
Deleted residual
DFBeta
DFFit
Durbin–Watson test
F-statistic
Generalization
Goodness of fit
Hat values
Heteroscedasticity
Hierarchical regression
Homoscedasticity
Leverage
Mahalanobis distance
Mean squares
Model sum of squares
Multicollinearity
Multiple regression
Outcome variable
Perfect collinearity
Predicted value
Predictor variable
Shrinkage
Simple regression
Standardized dfbeta
Standardized DFfit
Standardized residuals
Stepwise regression
Studentized deleted residuals
Studentized residuals
Suppressor effects
t-statistic
Tolerance
Total sum of squares
Unstandardized residuals
Variance inflation factor (VIF)

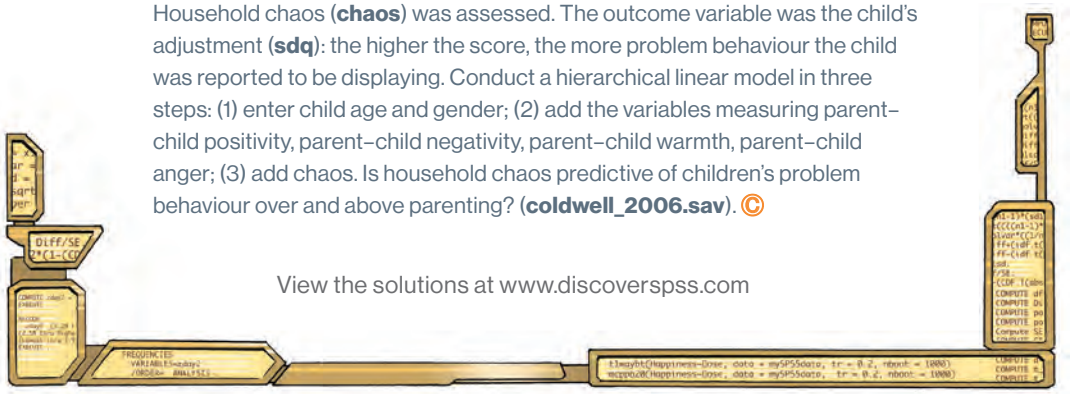


Smart Alex's tasks

- **Task 1**: In Chapter 8 (Task 9) we looked at data based on findings that the number of cups of tea drunk was related to cognitive functioning (Feng et al., 2010). Using a linear model that predicts cognitive functioning from tea drinking, what would cognitive functioning be if someone drank 10 cups of tea? Is there a significant effect? (**tea_716.sav**) Ⓐ
- **Task 2**: Estimate a linear model for the **pubs.sav** data in Jane Superbrain Box 9.1 predicting **mortality** from the number of **pubs**. Try repeating the analysis but bootstrapping the confidence intervals. Ⓑ
- **Task 3:** In Jane Superbrain Box 2.1 we encountered data (**honesty_lab.sav**) relating to people's ratings of dishonest acts and the likeableness of the perpetrator. Run a linear model with bootstrapping to predict ratings of dishonesty from the likeableness of the perpetrator. Ⓑ
- **Task 4**: A fashion student was interested in factors that predicted the salaries of catwalk models. She collected data from 231 models (**supermodel.sav**). For each model she asked them their salary per day (**salary**), their age (**age**), their length of experience as models (**years**) and their industry status as a model as reflected in their percentile position rated by a panel of experts (**beauty**). Use a linear model to see which variables predict a model's salary. How valid is the model? Ⓑ
- **Task 5**: A study was carried out to explore the relationship between **aggression** and several potential predicting factors in 666 children who had an older sibling. Variables measured were **parenting_style** (high score = bad parenting practices), **computer_games** (high score = more time spent playing computer games), **television** (high score = more time spent watching television), **diet** (high score = the child has a good diet low in harmful additives), and **sibling_aggression** (high score = more aggression seen in their older sibling). Past research indicated that parenting style and sibling aggression were good predictors of the level of aggression in the younger child. All other variables were treated in an exploratory fashion. Analyse them with a linear model (**child_aggression.sav**). Ⓑ
- **Task 6:** Repeat the analysis in Labcoat Leni's Real Research 9.1 using bootstrapping for the confidence intervals. What are the confidence intervals for the regression parameters? Ⓒ
- **Task 7:** Coldwell et al. (2006) investigated whether household chaos predicted children's problem behaviour over and above parenting. From 118 families they recorded the age and gender of the youngest child (**child_age** and **child_gender**). They measured dimensions of the child's perceived relationship with their mum: (1) warmth/enjoyment (**child_warmth**) and (2) anger/hostility (**child_anger**). Higher scores indicate more warmth/enjoyment and anger/hostility respectively. They measured the mum's perceived relationship with her child, resulting in dimensions of positivity (**mum_pos)** and negativity (**mum_neg**).

Household chaos (**chaos**) was assessed. The outcome variable was the child's adjustment (**sdq**): the higher the score, the more problem behaviour the child was reported to be displaying. Conduct a hierarchical linear model in three steps: (1) enter child age and gender; (2) add the variables measuring parent–child positivity, parent–child negativity, parent–child warmth, parent–child anger; (3) add chaos. Is household chaos predictive of children's problem behaviour over and above parenting? (**coldwell_2006.sav**). ©

View the solutions at www.discoverspss.com





CATEGORICAL PREDICTORS: COMPARING TWO MEANS

10.1 What will this chapter tell me? 478
10.2 Looking at differences 479
10.3 A mischievous example 479
10.4 Categorical predictors in the linear model 482
10.5 The *t*-test 484
10.6 Assumptions of the *t*-test 492
10.7 Comparing two means: general procedure 492
10.8 Comparing two independent means using SPSS 493
10.9 Comparing two related means using SPSS 502
10.10 Reporting comparisons between two means 515
10.11 Between groups or repeated measures? 516
10.12 Jane and Brian's story 516
10.13 What next? 517
10.14 Key terms that I've discovered 518
Smart Alex's tasks 518

Letters A to D indicate increasing difficulty

10.1 What will this chapter tell me? Ⓐ

We saw in the previous chapter that I had successfully conquered the holiday camps of Wales with my singing and guitar playing (and the Welsh know a thing or two about good singing). I had jumped on a snowboard called oblivion and thrown myself down the black run known as world domination. About 10 metres after starting this slippery descent I hit the lumpy patch of ice called 'adults'. I was 9, life was fun, but every adult I encountered seemed obsessed with my future. 'What do you want to be when you grow up?' they would ask. Would I be a surgeon, a lawyer, a teacher? I was 9 and 'grown-up' was a lifetime away. All I knew was that I was going to marry Clair Sparks (more on her in the next chapter) and be a rock legend who didn't need to worry about such adult matters as having a job. It was a difficult question, but adults require answers and I wasn't going to let them know that I didn't care about 'grown-up' matters. Like all good scientists I drew upon past data: I hadn't tried conducting brain surgery, nor had I taught anyone. I had, however, had a go at singing and playing guitar; therefore, I predicted I would be a rock star. Even at this early age I realized that not all adults would appreciate the raw talent that would see me parading across the lighted stage in front of tens of thousands of people. Some of them might not think that rock stardom was good career prospect. I needed to convince them. Adults tend to think money is important, so I decided I should demonstrate that rock stars earn more money than, say, a 'respectable' profession such as being a teacher. I could gather some teachers and rock stars, find out what their salaries were and compare them. Effectively I'd be 'predicting' salary from two categories: rock star or teacher. I could have done all that but I didn't know about statistics when I was 9. Happy days.



Figure 10.1 Practising for my career as a rock star by slaying the baying throng of Grove Primary School at the age of 10. (Note the girl with her hands covering her ears.)

10.2 Looking at differences Ⓐ

The last two chapters have focused on predicting an outcome from continuous variables, but sometimes researchers want to predict outcomes from categories or group membership. In particular, we often want to look at differences in an outcome variable between groups of entities or entities in different treatment conditions. Experimental research, for example, takes advantage of the fact that if we systematically manipulate what happens to people we can make causal inferences about the effects of those manipulations. The simplest form of experiment is one in which we split the sample into an experimental group and a control group that is identical to the experimental group in all respects except the one expected to have an impact on the outcome (see Field & Hole, 2003). For example, we might want to compare statistics knowledge before and after a lecture. One group, the experimental group, has to sit through the lecture, while the other group, the control group, gets to miss it and stay in bed. Here are some other examples of scenarios where we'd compare two conditions:

- Is the movie *Scream 2* scarier than the original *Scream*? We could measure heart rates (which indicate anxiety) during both films and compare them.
- Is your work better when you do it listening to Andy's favourite music? You could get some people to answer an essay question listening to my favourite music (as listed in the Acknowledgements) and get some other people to write answer the same essay question in silence (this is a control) and compare the grades.
- Do diet pills work? We take two groups of people and randomly assign one group to a programme of dieting pills and the other group to a programme of sugar pills (which they believe will help them lose weight). If the people who take the diet pills lose more weight than those on the sugar pills we can infer that the diet pills caused the weight loss.

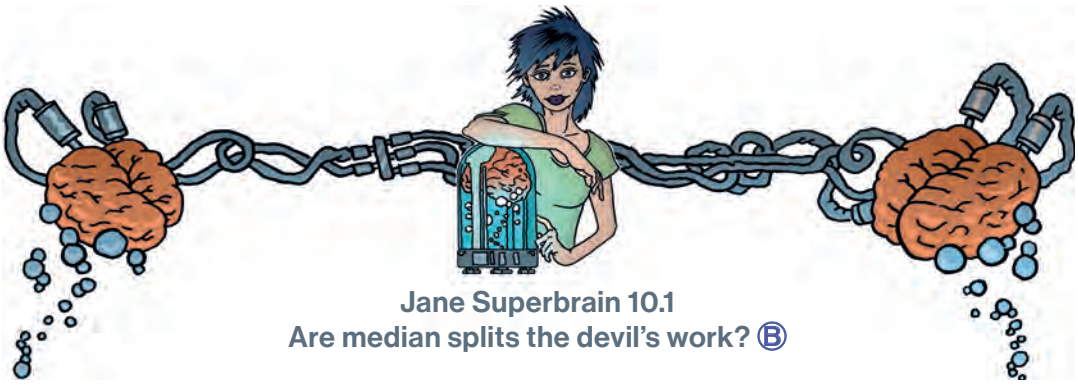
Systematic manipulation of the independent (predictor) variable is a powerful tool because it goes one step beyond merely observing variables.[1] This chapter is the first of many that look at these research scenarios. We start with the simplest scenario: when we have two groups, or, to be more specific, when we want to compare two means. We discovered in Chapter 1 that we can expose different entities to different experimental manipulations (a *between-group* or *independent* design), or take a single group of entities and expose them to different experimental manipulations at different points in time (a *repeated-measures* or *within-subject* design). Researchers sometimes get tempted to compare artificially created groups by, for example, dividing people into groups based on a median score; avoid this temptation (Jane Superbrain Box 10.1).

10.3 A mischievous example Ⓐ

Two news stories caught my eye related to physics (Di Falco et al., 2010). In the first headline (November 2010) the *Daily Mirror* (a UK newspaper) reported 'Scientists make Harry Potter's invisible cloak'. I'm not really a Harry Potter aficionado,[2] so it wasn't his mention that caught my attention, but the idea of being able to don a cloak that would render me invisible and able to get up to mischief. That idea was very exciting indeed; where could I buy one? By February 2011 the same newspaper was reporting on a different piece of research (Chen et al., 2011) with a similarly exaggerated headline: 'Harry Potter-style "invisibility cloak" built by scientists'.

1 People sometimes get confused and think that certain statistical procedures allow causal inferences and others don't (see Jane Superbrain Box 2.4).

2 Perhaps I should be, given that a UK newspaper once tagged me 'the Harry Potter of the social sciences' (http://www.discoveringstatistics.com/docs/thes_170909.pdf). I wasn't sure whether this made me a heroic wizard battling against the evil forces of statistics, or an adult with a mental age of 11.



Jane Superbrain 10.1
Are median splits the devil's work? Ⓑ

Sometimes scientists use a 'median split' of a predictor variable. For example, there is a stereotype that science fiction fans are recluses with no social skills. If you wanted to test this stereotype you might measure social skills and knowledge of the film *Star Wars*. You might then take the median score on *Star Wars* knowledge and classify anyone with a score above the median as a '*Star Wars* fan', and anyone below as a 'non-fan'. In doing this you 'dichotomize' a continuous variable. This practice is quite common, but there are several problems with median splits (MacCallum et al., 2002):

1 Imagine there are four people: Peter, Birgit, Jip and Kiki. We measure how much they know about *Star Wars* as a percentage and get Jip (100%), Kiki (60%), Peter (40%) and Birgit (0%). If we split these four people at the median (50%) then we're saying that Jip and Kiki are the same (they get a score of 1 = fanatic) and Peter and Birgit are the same (they both get a score of 0 = not a fanatic). Median splits change the original information quite dramatically: Peter and Kiki are originally very similar but become opposite after the split, whereas Jip and Kiki are relatively dissimilar originally but become identical after the split.
2 Effect sizes get smaller. If you correlate two continuous variables then the effect size will be larger than if you correlate the same variables after one of them has been dichotomized. Effect sizes also get smaller in linear models.
3 There is an increased chance of finding spurious effects.

So, if your supervisor has just told you to do a median split, have a good think about whether it is the right thing to do and read up on the topic (I recommend DeCoster et al., 2009, 2011; MacCallum et al., 2002). According to MacCallum et al., one of the rare situations in which dichotomizing a continuous variable is justified is when there is a clear theoretical rationale for distinct categories of people based on a meaningful break point (i.e., not the median); for example, phobic versus not phobic based on diagnosis by a trained clinician might be a legitimate dichotomization of anxiety.



Needless to say, scientists hadn't actually made Harry Potter's cloak of invisibility or anything close to it, but never let that get in the way of a headline. What Chen et al. had made wasn't so much a 'cloak' of invisibility as a 'calcite lump' of invisibility. This lump could hide small objects (centimetres and millimetres in scale): you could conceal my brain but little else. Nevertheless, with a suitably large piece of calcite in tow, I could theoretically hide my whole body (although people might get suspicious of the apparently autonomous block of calcite manoeuvring itself around the room on a trolley). Di Falco et al. had created a flexible material (Metaflex) with optical properties that meant that if you

layered it up you might be able to create something around which light would bend. Not exactly a cloak in the clothing sense of the word, but easier to wear than, say, a slab of calcite.

Although the newspapers overstated the case a little, these are exciting pieces of research that bring the possibility of a cloak of invisibility closer to reality. In fact, in the intervening years scientists have developed a skin cloak that 'consists of subwavelength-scale nanoantennas, which provide distinct phase shifts locally to the reflected electromagnetic waves' (Ni et al., 2015), whatever that means. Nevertheless, I imagine a future in which we have some cloaks of invisibility to test out. Given my slightly mischievous streak, the future me is interested in the effect that wearing a cloak of invisibility has on the tendency for mischief. I take 24 participants and placed them in an enclosed community. The community is riddled with hidden cameras so that we can record mischievous acts. Half of the participants are given cloaks of invisibility; they are told not to tell anyone else about their cloak and that they can wear it whenever they liked. I measure how many mischievous acts they performed in a week. These data are in Table 10.1.

Table 10.1 Data from **invisibility.csv**

id	cloak	mischief
Kiara	No cloak	3
Anupama	No cloak	1
Collin	No cloak	5
Ian	No cloak	4
Vanessa	No cloak	6
Darrell	No cloak	4
Tyler	No cloak	6
Steven	No cloak	2
Katheryn	No cloak	0
Kirkpatrick	No cloak	5
Melissa	No cloak	4
Kinaana	No cloak	5
Shajee'a	Cloak	4
Sage	Cloak	3
Jorge	Cloak	6
Conan	Cloak	6
Tamara	Cloak	8
Man	Cloak	5
Joseph	Cloak	5
Israa	Cloak	4
Kathryn	Cloak	2
Luis	Cloak	5
Devante	Cloak	7
Jerry	Cloak	5



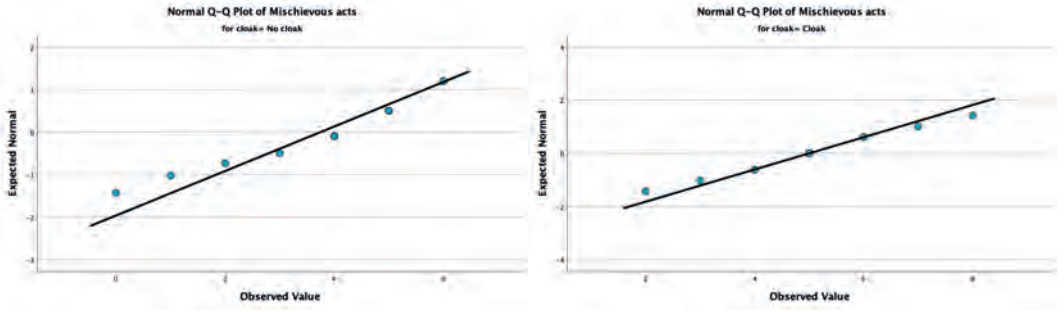
Enter these data into SPSS.

The file **invisibility_cloak.sav** shows how you should have entered the data: the variable **cloak** records whether a person was given a cloak (**cloak** = 1) or not (**cloak** = 0), and **mischief** is how many mischievous acts were performed.



Descriptives

			Cloak of invisibility			
			No cloak		Cloak	
			Statistic	Std. Error	Statistic	Std. Error
Mischievous acts	Mean		3.75	.552	5.00	.477
	95% Confidence Interval for Mean	Lower Bound	2.53		3.95	
		Upper Bound	4.97		6.05	
	5% Trimmed Mean		3.83		5.00	
	Median		4.00		5.00	
	Variance		3.659		2.727	
	Std. Deviation		1.913		1.651	
	Minimum		0		2	
	Maximum		6		8	
	Range		6		6	
	Interquartile Range		3		2	
	Skewness		-.789	.637	.000	.637
	Kurtosis		-.229	1.232	.161	1.232



Output 10.1

Output 10.1 (your table will be arranged differently – I edited mine as in SPSS Tip 6.2) shows that people with invisibility cloaks performed more mischievous acts, $M = 5$, 95% CI [3.95, 6.05], $SD = 1.65$, than those with no cloak, $M = 3.75$, 95% CI [2.53, 4.97], $SD = 1.91$. Both groups' scores look normally distributed because on the Q-Q plots the quantiles from the data approximate the diagonal line that indicates a normal distribution.

10.4 Categorical predictors in the linear model Ⓐ

If we want to compare differences between the means of two groups, all we are doing is predicting an outcome based on membership of two groups. For our invisibility example, we're predicting the number of mischievous acts from whether someone had a cloak of invisibility. This is a linear model with one dichotomous predictor. The b for the model reflects the differences between the mean levels of mischief in the two groups, and the resulting t-test will, therefore, tell us whether the difference between means is different from zero (because, remember, the t-test tests whether $b = 0$).

You might be thinking '*b*s show relationships, not differences between means, what is this fool going on about?' You might be starting to mistrust me, or stuffing the book into a box to post it back for a refund. I wouldn't blame you, because I used to think this too. To tame a land like the complex, thorny, weed-infested and Andy-eating-tarantula-inhabited world of statistics you need an epiphany. Mine was a paper by Cohen (1968) that showed me that when we compare means we are using a special case of the linear model. This revelation turned my statistical world into a beautiful meadow filled with bleating little lambs all jumping for joy at the wonder of life.

Recall from Chapter 6 that a linear model is defined by parameters: b_0 tells us the value of the outcome when the predictor is zero, and b_1 quantifies the relationship between the predictor (X_i) and outcome (Y_i). We've seen this equation lots of times, but let's tailor it to our example. The equation predicting the variable **mischief** from the group to which a person belongs (the variable **cloak**) is:

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{mischief}_i &= b_0 + b_1 \text{cloak}_i + \varepsilon_i \end{aligned} \tag{10.1}$$

Having fitted our model we can predict scores using:

$$\begin{aligned} \hat{Y}_i &= \hat{b}_0 + \hat{b}_1 X_i \\ \widehat{\text{mischief}}_i &= \hat{b}_0 + \hat{b}_1 \text{cloak}_i \end{aligned} \tag{10.2}$$

The variable **cloak** is a nominal variable: people had a 'cloak' or 'no cloak'. We can't put words into a statistical model because it will burn a hole in the ozone layer. Instead, we convert this variable into numbers. The numbers we choose to represent our categories in a mathematical model are important: they change the meaning of the resulting *b*-values. There are different 'standard' ways to code variables (which we won't get into here), one of which is to use **dummy variables**. We'll look at these in Section 12.3, but the summary is that we code a baseline category with a 0, and other categories with a 1. In this example there are two categories. Our baseline category is no cloak (the control condition) and we assign these participants a 0 for the variable **cloak**. The 'experimental' group contains those who were given a cloak and we assign these participants a 1. Let's plug these numbers into the model and see what happens.

First, imagine someone is in the no cloak condition. Knowing that they are in that group, the best prediction we could make of the number of mischievous acts would be the group mean because this value is the summary statistic with the least squared error. So, the predicted value of *the outcome* ($\widehat{\text{mischief}}_i$) in the equation will be the group mean $\bar{X}_{\text{No cloak}}$ (which is 3.75 in Output 10.1) and the value of the **cloak** variable will be 0. Eq. 10.2 becomes:

$$\begin{aligned} \widehat{\text{mischief}}_i &= \hat{b}_0 + \hat{b}_1 \text{cloak}_i \\ \bar{X}_{\text{No cloak}} &= \hat{b}_0 + \left(\hat{b}_1 \times 0\right) \\ \hat{b}_0 &= \bar{X}_{\text{No cloak}} \\ &= 3.75 \end{aligned} \tag{10.3}$$

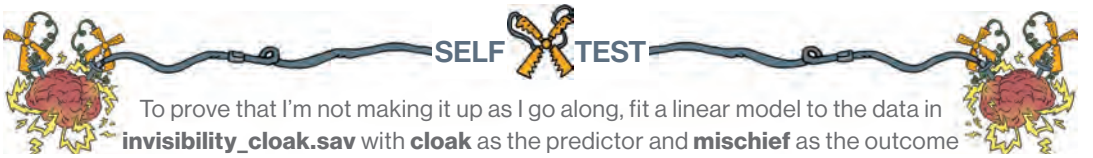
Note that b_0 (the intercept) is equal to the mean of the group coded as 0 (i.e., the no cloak group).

Now let's use the model to predict mischief in people who had an invisibility cloak. As before, the predicted value of the outcome would be the mean of the group to which the person belonged, because this is the statistic with the least squared error. The predicted value for someone in the cloak group is, therefore, the mean of the cloak group $\bar{X}_{\text{cloak}}$, which is 5 in Output 10.1. The value of the **cloak** variable is 1 (because this is the value we used to code group membership). Remember that $\hat{b}_0$ is equal to the mean of the no cloak group ($\bar{X}_{\text{No cloak}}$). If we place all of the values into Eq. 10.2 and rearrange it a bit we get:

$$\begin{aligned}\widehat{\text{mischief}} &= \hat{b}_0 + \hat{b}_1 \text{cloak}_i \\ \bar{X}_{\text{Cloak}} &= \hat{b}_0 + \left(\hat{b}_1 \times 1\right) \\ &= \hat{b}_0 + \hat{b}_1 \\ \hat{b}_1 &= \bar{X}_{\text{Cloak}} - \hat{b}_0 \\ &= \bar{X}_{\text{Cloak}} - \bar{X}_{\text{No cloak}}\end{aligned} \tag{10.4}$$

which shows that $\hat{b}_1$ represents the difference between the group means (in this case 5 – 3.75 = 1.25).

The take-home message is that we use the same linear model that we have used throughout the book to compare group means. In a model with a categorical predictor with two predictors, $\hat{b}_1$ represents the difference between group means, and $\hat{b}_0$ is equal to the mean of the group coded as 0. We have seen in the previous chapter that a *t*-statistic is used to ascertain whether a model parameter ($\hat{b}_1$) is equal to 0; in this context, therefore, it would test whether the difference between group means is equal to 0.



To prove that I'm not making it up as I go along, fit a linear model to the data in **invisibility_cloak.sav** with **cloak** as the predictor and **mischief** as the outcome using what you learnt in the previous chapter. **cloak** is coded using zeros and ones as described above.

If you do the self-test you should obtain the table in Output 10.2. First, notice that the value of the constant ($\hat{b}_0$) is 3.75, the same as the mean of the base category (the no cloak group). Second, notice that the value of the regression coefficient ($\hat{b}_1$) is 1.25, which is the difference between the two group means (5 – 3.75 = 1.25). Finally, the *t*-statistic, which tests whether b_1 is significantly different from zero, is not significant because the significance value is greater than 0.05, which means that the difference between means (1.25) is not significantly different from 0. This section demonstrates that differences between means can be represented by linear models, which is a theme to which we'll return many times over the coming chapters.

Coefficients[a]

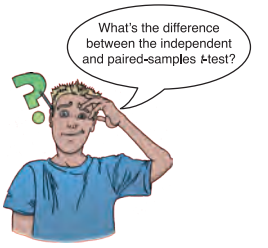
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3.750	.516		7.270	<.001
	Cloak of invisibility	1.250	.730	.343	1.713	.101

a. Dependent Variable: Mischievous acts

Output 10.2

10.5 The *t*-test Ⓐ

In Chapter 9 we saw that a *t*-statistic can be used to test whether a number (in the context of a linear model, the value of a *b*) is different from another number (specifically zero, although it doesn't have to be). Therefore, we're BFFs with the notion that we can use a *t*-statistic to compare two values. We've



also just discovered that when we include a two-category predictor in a linear model the b represents the difference between the means of those two categories. Put those ideas together and you can see that the t-statistic can be used to test the difference between two means. Specifically, it can test whether the difference between two means is different from zero (i.e., no difference at all).

As you read scientific papers, though, you will discover that people comparing two means rarely talk about linear models and instead mention this thing called a t-test (Student, 1908). The t-test is a convenient little package that wraps the t-statistic for the purpose of comparing two means. When wrapped up in this way, the t-statistic comes in two different gift boxes:

- **Independent *t*-test**: This is the t-statistic applied in the way we have looked at it up to now and assumes independent errors. That is, observations are assumed to come from different or 'independent' entities. Therefore, in this version of the test, the t-statistic is used to compare two means that come from conditions consisting of different entities (this test is sometimes called the *independent-measures* or *independent-means* t-test).
- **Paired-samples *t*-test**: This test, also known as the **dependent *t*-test**, is used when you want to compare two means that come from conditions consisting of the same or related entities (Figure 10.2). This test is often preferred over a linear model framework, because this situation violates the assumption of independent errors, which complicates the linear model, as we shall see in a few chapters time.



Figure 10.2 Thanks to the Confusion machine there are lots of terms for the paired-samples t-test

10.5.1 Rationale for the *t*-test Ⓐ

Both t-tests have a similar rationale, which is based on what we learnt in Chapter 1 about hypothesis testing:

- Two samples of data are collected and the sample means calculated. These means might differ by either a little or a lot.
- If the samples come from the same population, then we expect their means to be roughly equal (see Section 2.8). Although it is possible for their means to differ because of sampling variation,

we would expect large differences between sample means to occur very infrequently. Under the null hypothesis we assume that the experimental manipulation has no effect on the participant's behaviour: therefore, we expect means from two random samples to be very similar.

- We compare the difference between the sample means that we collected to the difference between the sample means that we would expect to obtain (in the long run) if there were no effect (i.e., if the null hypothesis were true). We use the standard error (see Section 2.8) as a gauge of the variability between sample means. If the standard error is small, then we expect most samples to have very similar means. When the standard error is large, large differences in sample means are more likely. If the difference between the samples we have collected is larger than we would expect based on the standard error then one of two things has happened:
  - There is no effect but sample means from our population fluctuate a lot and we happen to have collected two samples that produce very different means.
  - The two samples come from different populations, which is why they have different means and this difference is, therefore, indicative of a genuine difference between the samples. In other words, the null hypothesis is unlikely.
- The larger the observed difference between the sample means (relative to the standard error), the more likely it is that the second explanation is correct: that is, that the two sample means differ because of the different testing conditions imposed on each sample.

Most test statistics are signal-to-noise ratios: the 'variance explained by the model' divided by the 'variance that the model can't explain' (re-read Section 2.10.4). In other words, effect divided by error. When comparing two means the 'model' that we fit (the effect) is the difference between the two group means. Means vary from sample to sample (sampling variation), and we can use the standard error as a measure of how much means fluctuate (in other words, the error in the estimate of the mean) – see Chapter 2. Therefore, we can use the standard error of the differences between the two means as an estimate of the error in our model (or the error in the difference between means). Therefore, the *t*-statistic can be expressed as:

$$t = \frac{\text{observed difference between sample means} - \text{expected difference between population means (if null hypothesis is true)}}{\text{estimate of the standard error of the difference between two sample means}} \tag{10.5}$$

The top half of the equation is the 'model', which is that the difference between means is bigger than the expected difference under the null hypothesis, which in most cases will be 0. The bottom half is the 'error'. So, we're basically getting the test statistic by dividing the model (or effect) by the error in the model. The exact form that this equation takes depends on whether scores are independent (e.g., come from different entities) or related to each other (come from the same or related entities).

10.5.2 The paired-samples *t*-test equation explained Ⓐ

We'll start with the simpler scenario of when scores in the two conditions that you want to compare are related; for example, the same entities have been tested in the different conditions of your experiment, or perhaps you have data on a task from twins (you'd expect each person's score to be more similar to that of their twin than that of a stranger). If you choose not to think in terms of a linear model, then you can calculate the *t*-statistic using a numerical version of Eq. 10.5:

$$t = \frac{\bar{D} - \mu_D}{\sigma_{\bar{D}}} = \frac{\bar{D}}{\sigma_{\bar{D}}} \quad (10.6)$$

This equation compares the mean difference between our samples ($\bar{D}$) to the difference that we would expect to find between population means (μ_D) relative to the standard error of the differences ($\sigma_{\bar{D}}$). If the null hypothesis is true, then we expect no difference between the population means and $\mu_D = 0$ and it drops out of the equation.



Let's explore the logic of Eq. 10.6. Imagine you take a pair of samples from a population, calculate their means and then take the difference between them. We know from sampling theory (Section 2.8) that, on average, sample means will be very similar to the population mean; therefore, on average, most samples should have very similar means. Our pair of random samples should, therefore, have similar means, meaning that the difference between their means is zero, or close to zero. Imagine we repeated this process lots of times. We should find that most pairs of samples have differences between means that are close to zero, but sometimes one or both of the samples will have a mean very deviant from the population mean and we'd obtain a large difference between sample means. In short, sampling variation means that it's possible to get a difference between two sample means that is quite large, but it will happen relatively infrequently. If we plotted the frequency distribution of differences between means of pairs of samples we'd get the sampling distribution of differences between means. We might expect this distribution to be normal around zero, indicating that most pairs of samples have differences close to zero, and only very infrequently do we get large differences between sample means. The standard deviation of this sampling distribution is called the **standard error of differences**. Like any standard error (refresh your memory of Section 2.8 if you need to), a small standard error suggests that the difference between means of most pairs of samples will be very close to the population mean (in this case 0 if the null is true) and that substantial differences are very rare. A large standard error tells us that the difference between means of most pairs of samples can be quite variable: although the difference between means of most pairs of samples will still be centred around zero, substantial differences from zero are more common (than when the standard error is small). As such, the standard error is a good indicator of the size of the difference between sample means that we can expect from sampling variation. In other words, it's a good baseline for what could reasonably happen if the conditions under which scores are collected are stable.

The conditions under which scores are collected are not stable though. In experiments, we systematically manipulate the conditions under which scores are collected. For example, to test whether looking like a human affects trust of robots, participants might have two interactions with a robot: in one the robot is concealed under clothes and realistic flesh, whereas in the other their natural titanium exoskeleton is visible in the other. Each person's trust score in the first interaction could be different from their score in the second; the question is whether this difference is the product of how the robot looked, or just what you'd get if you test the same person twice. The standard error helps us to gauge this by giving us a scale of likely variability between samples. If the standard error is small then we know that even a modest difference between scores in the two conditions would be unlikely from two random samples. If the standard error is large then a modest difference between scores is plausible from two random samples. As such, the standard error of differences provides a scale of measurement for how plausible it is that an observed difference between sample means could be the product of taking two random samples from the same population. That's what the bottom of Eq. 10.6 represents: it places the observed difference between sample means in the context of what's plausible for random samples.

The top half of Eq. 10.6 represents the size of the observed effect. $\bar{D}$ is the average difference between people's scores in two conditions. For each person, if we took their score in one condition and subtracted it from their score in the other, this would give us a difference score; $\bar{D}$ is the mean of

these difference scores. Going back to our robot example, if the appearance of the robot had no effect on people's trust, scores would be similar in the two conditions and we'd get an average difference of 0 (or close to it). If the robot's appearance matters, we'd expect scores to differ in the two conditions, and the resulting average difference would be different from 0.

So, $\bar{D}$ represents the effect size, and as I said earlier, we place this effect size within the context of what's plausible for random samples by dividing by the standard error of differences. We know that the standard error can be estimated from the standard deviation divided by the square root of the sample size (Section 2.8). The standard error of differences ($\sigma_{\bar{D}}$) is likewise estimated from the standard deviation of differences within the sample (s_D) divided by the square root of the sample size (N). Replacing this term in Eq. 10.6 gives Eq. 10.7:

$$t = \frac{\bar{D}}{\sigma_{\bar{D}}} = \frac{\bar{D}}{s_D / \sqrt{N}} \tag{10.7}$$

Therefore, t is a signal-to-noise ratio, or the systematic variance compared to the unsystematic variance. The top half of Eq. 10.7 is the signal or effect, whereas the bottom places that effect within the context of the natural variation between samples (the noise, or unsystematic variation). If the experimental manipulation creates difference between conditions, then we would expect the effect (the signal) to be greater than the unsystematic variation (the noise) and, at the very least, t will be greater than 1. We can compare the obtained value of t against the maximum value we would expect to get if the null hypothesis were true in a t-distribution with the same degrees of freedom (these values can be found in the Appendix). If the observed t exceeds the critical value for the predetermined alpha (usually 0.05), scientists choose to assume that this reflects an effect of their independent variable. We *can* compare the observed t to tabulated critical values, but we *can* also urinate in a bucket and throw it from our window into the street. It doesn't mean that we should, and seeing as it's not 1908 we won't do either of these things; instead we'll urinate in a computer and let a toilet compute an exact p-value for t. I think it's that way around. If the exact p-value for t is less than or equal to the predetermined alpha value (usually 0.05), scientists take this to support the conclusion that the differences between scores are not due to sampling variation and that their manipulation (e.g., dressing up a robot as a human) has had a significant effect.

10.5.3 The independent t-test equation explained Ⓐ

When we want to compare scores that are independent (e.g., different entities have been tested in the different conditions of your experiment) we are in the same logical territory as when scores are related. The main difference is how we arrive at the values of interest. The equation for t based on independent scores is still a numerical version of Eq. 10.5. The main difference is that we're not dealing with difference scores because there's no connection between scores in the two conditions that we want to compare.

When scores in two groups come from different participants, pairs of scores will differ not only because of the experimental manipulation reflected by those conditions, but also because of other sources of variance such as individual differences between participants' motivation, IQ, etc.). These individual differences are eliminated when we use the same participants across conditions. Because the scores in the two conditions have no logical connection, we compare means on a *per condition* basis. We compute differences between the two sample means ($\bar{X}_1 - \bar{X}_2$) and not between individual pairs of scores. The difference between sample means is compared to the differences we would expect to get between the means of the two populations from which the samples come ($\mu_1 - \mu_2$):

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\text{estimate of the standard error}} \tag{10.8}$$

If the null hypothesis is true then the samples have been drawn from populations that have the same mean. Therefore, under the null hypothesis $\mu_1 = \mu_2$, which means that $\mu_1 - \mu_2 = 0$, and so $\mu_1 - \mu_2$ drops out of the equation, leaving us with:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{estimate of the standard error}} \tag{10.9}$$

Now, imagine we took several pairs of samples – each pair containing one sample from the two different populations – and compared the means of these samples. From what we have learnt about sampling distributions, we know that many samples from a population will have similar means. If the populations from which we're drawing have the same mean (which under the null hypothesis they do) then pairs of samples should also have the same means, and the difference between sample means should be zero, or close to it. We're now in very familiar territory to the paired-samples *t*-test because the sampling distribution of the differences between pairs of sample means would be normal with a mean equal to the difference between population means ($\mu_1 - \mu_2$), which under the null is zero. The sampling distribution would tell us by how much we can expect the means of two (or more) samples to differ (if the null were true). The standard deviation of this sampling distribution (the standard error) tells us how plausible differences between sample means are (under the null hypothesis). If the standard error is large then large differences between sample means can be expected; if it is small then only small differences between sample means are typical. As with related scores, it makes sense then to use the standard error to place the difference between sample means into the context of what's plausible given the null hypothesis. As such, Eq. 10.9 is conceptually the same as Eq. 10.6; all that differs is how we arrive at the effect (top half) and the standard error (the bottom half). The standard error, in particular, is derived quite differently for independent samples.

I've already reminded you that the standard error can be estimated from the standard deviation and the sample size. It is, therefore, straightforward to estimate the standard error for the sampling distribution of each population by using the standard deviation (s) and size (N) for each sample:

$$\begin{aligned}\text{SE of sampling distribution of population 1} &= \frac{s_1}{\sqrt{n_1}} \\ \text{SE of sampling distribution of population 2} &= \frac{s_2}{\sqrt{n_2}}\end{aligned} \tag{10.10}$$

These values don't tell us about the standard error for the sampling distribution of *differences* between means though. To estimate that, we need to first convert these standard errors to variances by squaring them:[3]

$$\begin{aligned}\text{variance of sampling distribution of population 1} &= \left(\frac{s_1}{\sqrt{n_1}}\right)^2 = \frac{s_1^2}{n_1} \\ \text{variance of sampling distribution of population 2} &= \left(\frac{s_2}{\sqrt{n_2}}\right)^2 = \frac{s_2^2}{n_2}\end{aligned} \tag{10.11}$$

Having converted to variances, we can take advantage of the **variance sum law**, which states that the variance of a difference between two independent variables is equal to the sum of their variances. Put simply, the variance of the sampling distribution of differences between two sample

3 Remember that a standard error is a standard deviation (it's just called a standard error because we're dealing with a sampling distribution), and the standard deviation is the square root of the variance.

means will be equal to the sum of the variances of the two populations from which the samples were taken. This law means that we can estimate the variance of the sampling distribution of differences by adding together the variances of the sampling distributions of the two populations:

$$\text{variance of sampling distribution of differences} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \tag{10.12}$$

We convert this variance back to a standard error by taking the square root:

$$\text{SE of sampling distribution of differences} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.13}$$

If we pop this equation for the standard error of differences into Eq. 10.9 we get:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10.14}$$

Eq. 10.14 is true only when the sample sizes are equal, which in naturalistic studies may not be possible.[4] To compare two groups that contain different numbers of participants we use a pooled variance estimate instead, which takes account of the difference in sample size by *weighting* the variance of each sample by a function of the size of sample on which it's based:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{10.15}$$

This weighting makes sense because (as we saw in Chapter 1) large samples more closely approximate the population than small ones; therefore, they should carry more weight. In fact, rather than weighting by the sample size, we weight by the sample size minus 1 (the degrees of freedom).

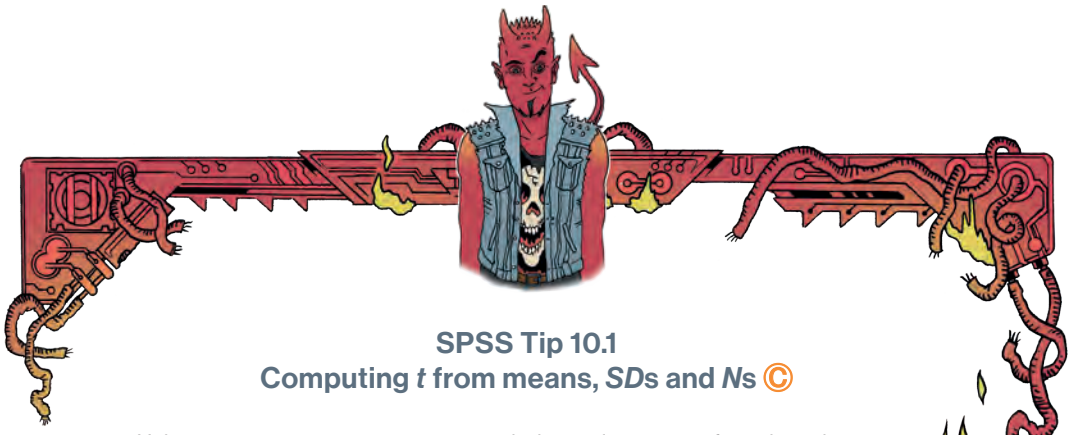
The pooled variance estimate in Eq. 10.15 is a weighted average: each variance is multiplied (weighted) by its degrees of freedom, and then we divide by the sum of weights (or sum of the two degrees of freedom) to get an average. The resulting weighted average variance is plopped into the equation for *t*:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \tag{10.16}$$

One thing that might be apparent from Eq. 10.16 is that you don't actually need any raw data to compute *t*, you just need the group means, standard deviations and sample sizes (see SPSS Tip 10.1).

As with the paired-samples *t*, the Edwardians among you can compare the observed *t* to critical values in a table, but the rest of us flush a toilet and an exact *p*-value comes out that we use as information to decide whether the observed effect is indicative of something more theoretically interesting than sampling variation.

4 You'll find this equation expressed more simply elsewhere because n_1 and n_2 are the same value when sample sizes are equal, but I want to keep continuity with later equations.



SPSS Tip 10.1
Computing *t* from means, *SD*s and *N*s ©

Using syntax, you can compute an independent *t*-test from just the group means, standard deviations and sample sizes. Set up six variables in the data editor: **x1** (mean of group 1), **x2** (mean of group 2), **sd1** (standard deviation of group 1), **sd2** (standard deviation of group 2), **n1** (sample size of group 1) and **n2** (sample size of group 2). Type the values of each of these in the first row of the data editor. Open a syntax window and type the following:

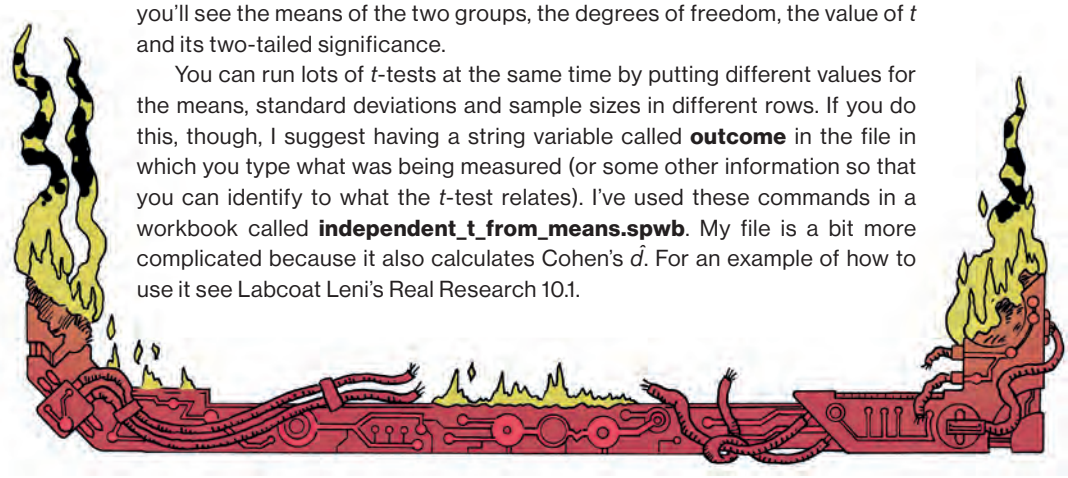
```
COMPUTE df = n1+n2-2.
COMPUTE poolvar = (((n1-1)*(sd1**2))+((n2-1)*(sd2**2)))/df.
COMPUTE t = (x1-x2)/sqrt(poolvar*((1/n1)+(1/n2))).
COMPUTE sig = 2*(1-(CDF.T(abs(t),df))).
Variable labels sig 'Significance (2-tailed)'.
EXECUTE.
```

The first line computes the degrees of freedom, the second the pooled variance, S_p^2, the third *t* and the fourth its two-tailed significance. All these values will be created in new columns in the data editor. The line beginning *Variable labels* labels the significance variable so that we know that it is two-tailed. To display the results in the viewer window we could use the following syntax:

```
SUMMARIZE
  /TABLES= x1 x2 df t sig
  /FORMAT=VALIDLIST NOCASENUM TOTAL LIMIT=100
  /TITLE='T-test'
  /MISSING=VARIABLE
  /CELLS=NONE.
```

These commands produce a table of the variables **x1**, **x2**, **df**, **t** and **sig**, so you'll see the means of the two groups, the degrees of freedom, the value of *t* and its two-tailed significance.

You can run lots of *t*-tests at the same time by putting different values for the means, standard deviations and sample sizes in different rows. If you do this, though, I suggest having a string variable called **outcome** in the file in which you type what was being measured (or some other information so that you can identify to what the *t*-test relates). I've used these commands in a workbook called **independent_t_from_means.spwb**. My file is a bit more complicated because it also calculates Cohen's $\hat{d}$. For an example of how to use it see Labcoat Leni's Real Research 10.1.



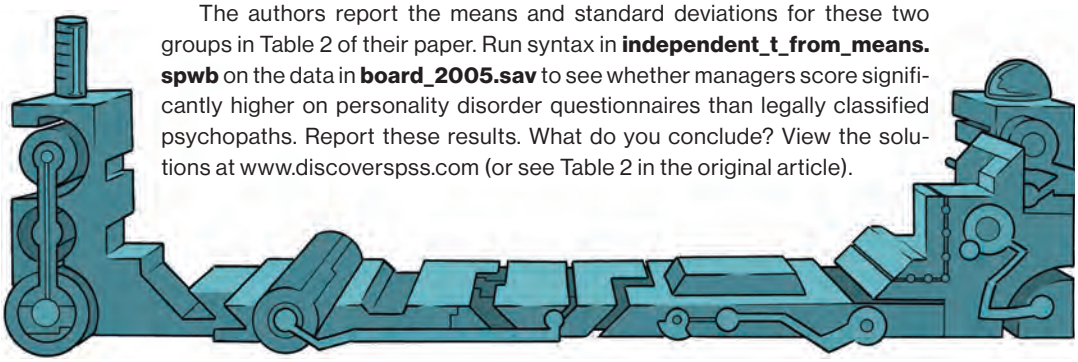


Labcoat Leni's Real Research 10.1 You don't have to be 'mad' to work here, but it helps ©

Board, B. J., & Fritzon, K. (2005). *Psychology, Crime & Law*, *11*, 17–32.

In the UK you often see the 'humorous' slogan 'You don't have to be mad to work here, but it helps' stuck up in workplaces. Board and Fritzon (2005) took this slogan to heart by measuring whether 39 senior business managers and chief executives from leading UK companies had personality disorders (PDs) as indicated by the Minnesota Multiphasic Personality Inventory Scales for DSM III Personality Disorders (MMPI-PD). This questionnaire quantifies people along 11 personality disorder dimensions: histrionic, narcissistic, antisocial, borderline, dependent, compulsive, passive-aggressive, paranoid, schizotypal, schizoid and avoidant. As a comparison group they chose 317 legally classified psychopaths from a high-security psychiatric hospital.

The authors report the means and standard deviations for these two groups in Table 2 of their paper. Run syntax in **independent_t_from_means.spwb** on the data in **board_2005.sav** to see whether managers score significantly higher on personality disorder questionnaires than legally classified psychopaths. Report these results. What do you conclude? View the solutions at www.discoverspss.com (or see Table 2 in the original article).



10.6 Assumptions of the *t*-test Ⓐ

Both the independent *t*-test and the paired-samples *t*-test are *special cases of the linear model* and are prone to the sources of bias discussed in Chapter 6. For the paired-samples *t*-test the assumption of normality relates to the sampling distribution of the *differences* between scores, not the scores themselves. There are variants on these tests that overcome all of the potential problems though.

10.7 Comparing two means: general procedure Ⓐ

I have probably bored most of you to the point of wanting to eat your own legs by now. Equations are tedious, and that is why computers were invented to help us minimize our contact with them. It's time

to move on and do stuff. Figure 10.3 shows the general process for performing a *t*-test. As with fitting any model, we start by looking for the sources of bias identified in Chapter 6. Having satisfied ourselves that assumptions are met and outliers dealt with, we run the test. We can also consider using bootstrapping if any of the test assumptions were not met. Finally, we compute an effect size and Bayes factor (Section 3.8.4).

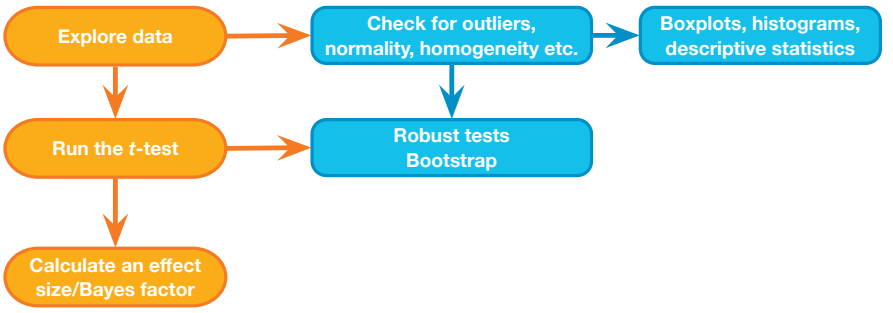


Figure 10.3 The general process for performing a *t*-test

10.8 Comparing two independent means using SPSS Ⓐ

Thinking back to our invisibility cloak data (**invisibility_cloak.sav**), we have 12 people who were given an invisibility cloak and 12 who were not (the groups are coded using the variable **cloak**). Remember that the number of mischievous acts they performed was measured (**mischief**). I have already described how the data are arranged (see Section 10.3), so we can move on to doing the test itself.

10.8.1 Exploring data and testing assumptions Ⓐ

We obtained descriptive statistics and looked at distributional assumptions in Section 10.3. We found evidence of normality in each group, and the mean number of mischievous acts was higher for those with a cloak (M = 5) than for those without (M = 3.75). To look at homogeneity of variance SPSS will produce Levene's test when we run the *t*-test, but as noted in Section 6.9.6 we should ignore this test.



10.8.2 The independent *t*-test using SPSS Ⓐ

To run an independent *t*-test, we access the main dialog box by selecting *Analyze* > *Compare Means* > Independent-Samples T Test...(see Figure 10.4). Once the dialog box is activated, select the outcome variable (**mischief**) and drag it to the box labelled *Test Variable(s)* (or click). If you want to carry out *t*-tests on several outcome variables then select these and transfer them to the *Test Variable(s)* box. However, remember that by doing lots of test you inflate the Type I error rate (see Section 2.10.8).

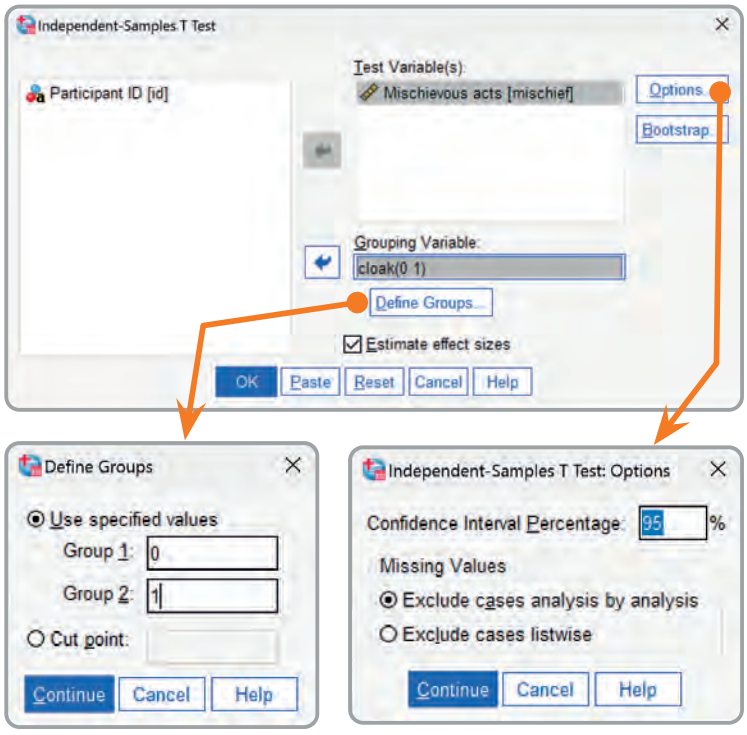


Figure 10.4 Dialog boxes for the independent-samples *t*-test

Make sure that ☑ Estimate effect sizes is selected, which will compute Cohen's $\hat{d}$ (Section 3.7.3) and related effect sizes.

Next, we need to specify a predictor variable (the grouping variable). In this case, transfer **cloak** to the box labelled *Grouping Variable*. The Define Groups... button will become active, so click it to activate the *Define Groups* dialog box. IBM SPSS Statistics software needs to know what numeric codes you assigned to your two groups, and there is a space for you to type the codes. In this example, we coded the no cloak group as 0 and our cloak group as 1, and so these are the values that we type into the boxes (as in Figure 10.4). Alternatively, you can specify a *Cut point* value whereby cases greater than or equal to that value are assigned to one group and values below the cut point to a second group. You can use this option to, for example, compare groups of participants based on something like a median split (see Jane Superbrain Box 10.1) – you would type the median value in the box labelled *Cut point*. When you have defined the groups, click Continue to return to the main dialog box.

Clicking Options... activates a dialog box in which you can change the width of the confidence interval in the output. The default setting of a 95% confidence interval is fine, but there may be times where you want to be stricter (and specify something like a 99% confidence interval), even though you run a higher risk of failing to detect a genuine effect (a Type II error), or more lenient (e.g., a 90% confidence interval), which of course increases the chance of falsely accepting a hypothesis (a Type I error). You can also select how to deal with missing values (see SPSS Tip 6.1).

If you are worried about the assumption of normality, or want confidence intervals that don't rely on this assumption, then use bootstrapping by clicking Bootstrap... to access the *Bootstrap* dialog box. We discussed this dialog box in Section 6.10.8; to recap, select ☑ Perform bootstrapping to activate bootstrapping, and to get a 95% confidence interval click ◉ Percentile or ◉ Bias corrected accelerated (BCa). For this analysis, chose a bias corrected (BCa) confidence interval. Back in the main dialog box click OK to run the analysis.

10.8.3 Output from the independent *t*-test Ⓐ

The first table (Output 10.3) provides summary statistics for the two experimental conditions (if you don't ask for bootstrapping this table will be a bit more straightforward). From this table, we can see that both groups had 12 participants (row labelled *N*). The group that had no cloak performed, on average, 3.75 mischievous acts with a standard deviation of 1.913. What's more, the standard error of that group is 0.552. The bootstrap estimate of this standard error is a little smaller at 0.53, and the bootstrapped confidence interval for the mean ranges from 2.67 to 4.78. Those who were given an invisibility cloak performed, on average, 5 acts, with a standard deviation of 1.651, and a standard error of 0.477. The bootstrap estimate of this standard error is 0.48, and the confidence interval for the mean ranges from 4.18 to 5.90. Note that the confidence intervals for the two groups overlap, implying that they might be from the same population.

Group Statistics

	Cloak of invisibility		Statistic	Bootstrap[a] Bias	Bootstrap[a] Std. Error	Bootstrap[a] BCa 95% CI Lower	Bootstrap[a] BCa 95% CI Upper
Mischievous acts	No cloak	N	12				
		Mean	3.75	.01	.53	2.67	4.78
		Std. Deviation	1.913	-.117	.374	1.166	2.312
		Std. Error Mean	.552				
	Cloak	N	12				
		Mean	5.00	-.02	.48	4.18	5.90
		Std. Deviation	1.651	-.096	.317	1.166	1.954
		Std. Error Mean	.477				

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 10.3

The second table of output (Output 10.4) contains the main test statistics. There are two rows containing values for the test statistics: one is labelled *Equal variances assumed*, while the other is labelled *Equal variances not assumed*. In Chapter 6, we discovered that when we have categorical predictors (as is the case here), the homogeneity of variance assumption relates to groups having equal variances. We also saw that there are adjustments that can be made in situations in which the variances are not equal. The rows of the table relate to whether these adjustments have been applied. You were advised against using Levene's test (see Section 6.9.6) so despite the fact that SPSS runs this test ignore it and always read results from the row labelled *Equal variances not assumed*.

Independent Samples Test

		Levene's Test for Equality of Variances: F	Sig.	t-test for Equality of Means: t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	Std. Error Difference	95% CI of the Difference: Lower	95% CI of the Difference: Upper
Mischievous acts	Equal variances assumed	.545	.468	-1.713	22	.050	.101	-1.250	.730	-2.763	.263
	Equal variances not assumed			-1.713	21.541	.050	.101	-1.250	.730	-2.765	.265

Output 10.4

We are told the mean difference ($\bar{X}_{\text{No cloak}} - \bar{X}_{\text{cloak}} = 3.75 - 5 = -1.25$) and the standard error of the sampling distribution of differences, which is calculated using the lower half of Eq. 10.14. The *t*-statistic is the mean difference divided by this standard error ($t = -1.25/0.730 = -1.71$). A *p*-value associated with this *t* is calculated based on a *t*-distribution with a particular degrees of freedom. For the independent *t*-test, the degrees of freedom are calculated by adding the two sample sizes and then

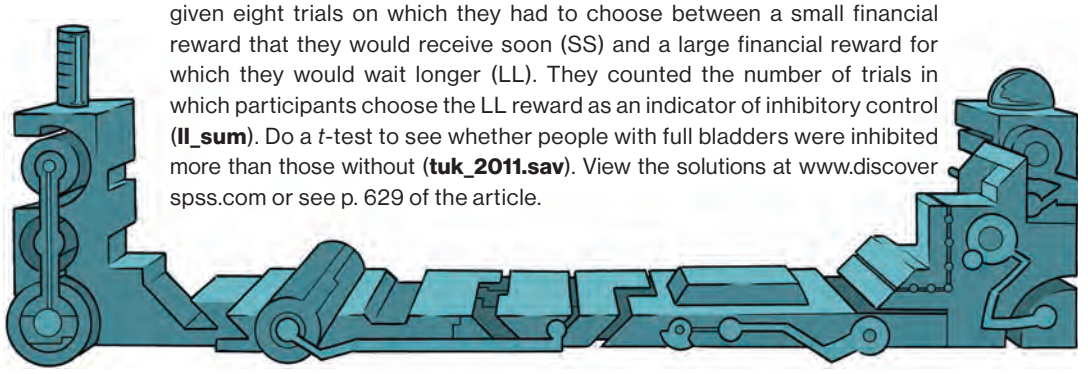
subtracting the number of samples ($df = N_1 + N_2 - 2 = 12 + 12 - 2 = 22$). This value is then reduced to compensate for any imbalance in group variances (in this case it becomes 21.54). The resulting (two-tailed) p-value is 0.101, which represents the probability of getting a t of –1.71 or smaller if the null hypothesis were true. Assuming our alpha is 0.05, we'd conclude that there was no significant difference between the means of these two samples because the observed p of 0.101 is greater than the criterion of 0.05. In terms of the experiment, we can infer that having a cloak of invisibility did not significantly affect the amount of mischief a person got up to. Note that the value of t and the significance value are the same as when we ran the same test as a regression (see Output 10.2).[5] Note that the output also contains a one-tailed significance but you should, in most situations, ignore this value (see 2.10.6 for the nuance).



Labcoat Leni's Real Research 10.2
Bladder control Ⓐ

Tuk et al. (2011). *Psychological Science, 22*(5), 627–633.

Visceral factors that require us to engage in self-control (such as a filling bladder) can affect our inhibitory abilities in unrelated domains. In a fascinating study by Tuk et al. (2011) participants were given five cups of water: one group was asked to drink them all, whereas another was asked to take a sip from each. This manipulation led one group to have full bladders and the other group relatively empty (**urgency**). Later on, these participants were given eight trials on which they had to choose between a small financial reward that they would receive soon (SS) and a large financial reward for which they would wait longer (LL). They counted the number of trials in which participants choose the LL reward as an indicator of inhibitory control (**ll_sum**). Do a t-test to see whether people with full bladders were inhibited more than those without (**tuk_2011.sav**). View the solutions at www.discoverspss.com or see p. 629 of the article.



5 The value of the t-statistic is the same but has a positive sign rather than negative. You'll remember from the discussion of the point-biserial correlation in Section 8.4.5 that when you correlate a dichotomous variable the direction of the correlation coefficient depends entirely upon which cases are assigned to which groups. Therefore, the direction of the t-statistic here is similarly influenced by which group we select to be the base category (the category coded as 0).

10.8.4 Robust tests of two independent means

Assuming you had reason to doubt the assumption of normality, you could interpret the bootstrap confidence intervals (Output 10.5). It's also a legitimate approach to always use bootstrapped confidence intervals and not think about normality one way or another. The table shows a robust re-estimate of the standard error of the mean difference (which is estimated as 0.72 rather than 0.73, the value in Output 10.4).[6] The difference between means was –1.25, and the bootstrap confidence interval ranges from –2.70 to 0.07, which implies that the difference between means in the population could be negative, positive or even zero (because the interval ranges from a negative value to a positive one). In other words, if we assume that this is one of the 95% of intervals that will capture the population value, it's possible that the true difference between means is zero – no difference at all. The bootstrap confidence interval confirms our conclusion that having a cloak of invisibility seems not to affect acts of mischief.

Bootstrap for Independent Samples Test

			Bootstrap[a]			
					BCa 95% Confidence Interval	
		Mean Difference	Bias	Std. Error	Lower	Upper
Mischievous acts	Equal variances assumed	-1.250	.026	.720	-2.702	.070
	Equal variances not assumed	-1.250	.026	.720	-2.702	.070

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 10.5

It's also possible to run a robust version of the *t*-test using the *WRS2* package in R (Section 4.13.2). The workbook file **robust_independent_t_test.spwb** contains syntax that checks whether you have *WRS2* installed, and if not tries to install it, and syntax to run a robust variant of the *t*-test based on Yuen (1974) which uses trimming and a bootstrap. The syntax is explained in SPSS Tip 10.2 and is as follows:

```
BEGIN PROGRAM R.
library(WRS2)
mySPSSdata = spssdata.GetDataFromSPSS(factorMode = "labels")
yuenbt(mischief~cloak, data = mySPSSdata)
END PROGRAM.
```

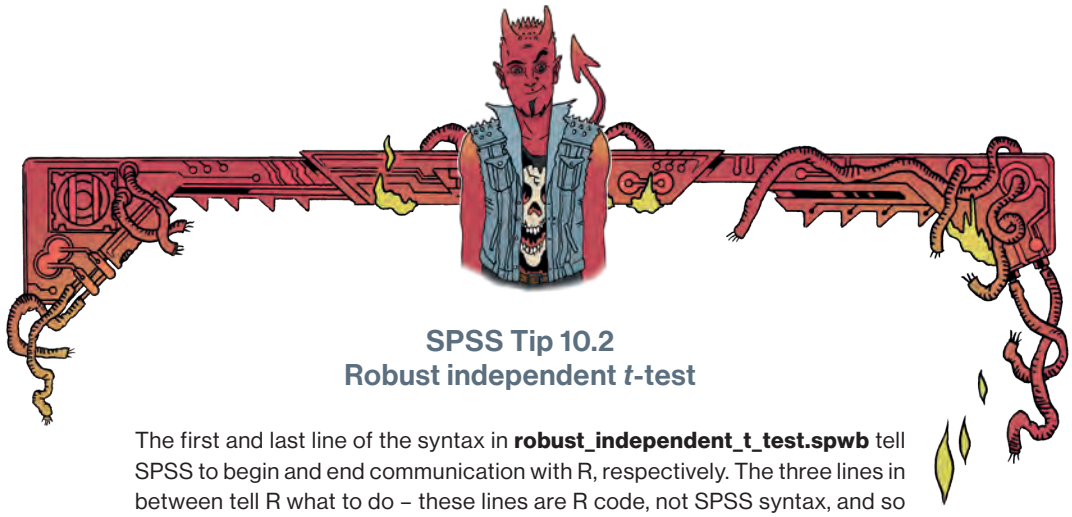
Having run the syntax, you'll find some rather uninspiring text output in the viewer (Output 10.6). This tells us that there is not a significant difference (because the confidence interval crosses zero and the *p*-value is greater than 0.05) in mischief scores across the two cloak groups, Y_t = –1.36 (–2.52, 0.52), p = 0.167. The confidence interval is based on a bootstrap so your values might not match those given here.

```
Test statistic: -1.3607 (df = NA), p-value = 0.16694

Trimmed mean difference: -1
95 percent confidence interval:
-2.5161     0.5161
```

Output 10.6

6 Remember that the values for the standard error and confidence interval you get could differ from mine because of the way bootstrapping works.



SPSS Tip 10.2
Robust independent *t*-test

The first and last line of the syntax in **robust_independent_t_test.spwb** tell SPSS to begin and end communication with R, respectively. The three lines in between tell R what to do – these lines are R code, not SPSS syntax, and so do not end in full stops. The first loads the WRS2 package that contains the function we want to use. The second grabs your data from the SPSS data editor and places it in a data frame called *mySPSSdata* (you can change this name if you like, but you don't need to). The final line runs the test using the function *yuenbt()*. Within this function we specify the data as being the data that we've just grabbed from SPSS, and we specify a formula for the test itself. This formula has the general form of *outcome~predictor*, in which *outcome* is the name of the outcome variable (from the data editor), and *predictor* is the name of the grouping variable. For these data the formula is *mischief~cloak* (which you could read as *predicting mischief from cloak*). If you want to use this syntax on other data sets the only part of it that you need to edit is the formula by replacing **mischief** and **cloak** with the names of your variables.

10.8.5 Effect sizes for two independent means Ⓑ

The fact that the *t*-statistic is not statistically significant doesn't mean that our effect is unimportant, so it's worth quantifying it with an effect size. Output 10.7 shows us the value of Cohen's $\hat{d}$ calculated using the pooled variance estimate (Section 3.7.3). The value suggest that there is a difference of –0.70 standard deviations between the two groups in terms of their mischief making, which is a very substantial effect. The minus sign is because the mean for the cloak group is subtracted from that of the no cloak group, resulting in a negative value (because the mean is larger in the cloak group). The 95% confidence interval for this effect ranges from –1.52 to 0.13, suggesting (if we assume our sample is one of the 95% that produce a confidence interval containing the true value) that the true effect could be anything from a massive effect of the cloak (i.e. large negative $\hat{d}$) to no effect (zero) to a small effect of not having a cloak (a positive $\hat{d}$).

You might reasonably argue that it would be better to use the standard deviation of the no cloak group because it represents a no-treatment control, but it makes little difference to the resulting estimate:

$$\hat{d} = \frac{\bar{X}_{\text{No cloak}} - \bar{X}_{\text{Cloak}}}{s_{\text{No cloak}}} = \frac{3.75 - 5}{1.91} = -0.65 \qquad (10.17)$$

Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	Upper
Mischievous acts	Cohen's d	1.787	-.700	-1.518	.134
	Hedges' correction	1.851	-.675	-1.466	.129
	Glass's delta	1.651	-.757	-1.601	.117

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control group.

Output 10.7



Oditi's Lantern
***t*-tests**

'I, Oditi, leader of the cult of undiscovered numerical truths, do not like differences. Everyone must conform to my cultish ideas, my view is the only view, and we must wage war on those who dare to have different views. Only by locating differences can we eliminate them and turn the world into mindless clones. Stare into my lantern to discover how to detect these differences. The more you stare, the more you will find yourself agreeing with everything I say ...'



10.8.6 Bayesian test of two independent means ©

To access the dialog box to compute a Bayes factor and credible interval for independent means select *Analyze* > *Bayesian Statistics* > Independent Samples Normal (Figure 10.5). Select the outcome variable (**mischief**) and drag it into the box labelled *Test Variable(s)* (or click ▸), then select the group variable (**cloak**) and drag it to the box labelled *Grouping Variable* (or click ▸). As with the regular *t*-test use the Define Groups... button to tell SPSS that the first group (no cloak) was coded with 0 and group 2 (cloak) was coded with 1. To compute Bayes factors as well as estimate the model parameters select ◉ Use Both Methods.

Imagine that prior to collecting data we believed that people without an invisibility cloak might be likely to perform anywhere between 0 and 6 mischievous acts. Obviously, we can't believe in a value less than 0 (you can't have a negative number of mischievous acts) but we're not prepared to believe in a value over 6. With a cloak of invisibility, we think the number of mischievous acts might fall anywhere

between 0 and about 8. We could model these prior beliefs with a normal distribution. The plots labelled *Prior Distribution* in Output 10.8 show these priors. The left plot is the prior for the no cloak group. As you can see it is a normal distribution centred at 3 and ranging from about 0 to 6. This equates to our believing that, without an invisibility cloak, 3 acts of mischief are most probable but we're prepared to entertain values as low as 0 or as high as 6 (but we think these values are unlikely). The right plot is the prior for the cloak group, which is also a normal distribution but centred at 4 and ranging from about 0 to 8. This equates to our believing that, with an invisibility cloak, 4 acts of mischief are most probable but we'll entertain (with low probability) values as low as 0 or as high as 8.

To set up these priors click Priors... and first specify the variances of the two groups (remember that group 1 is the no cloak and group 2 is the cloak group). I've used the variances from the data, which I obtained by squaring the standard deviations in Output 10.3. Using these variances simplifies things (we don't need to set a prior for the variance) but as you get more experience you might want to estimate the variance too. At the bottom select ◉ Normal. For group 1 I have specified a distribution with mean (location) of 3 and standard deviation (scale) of 1. For group 2 I have specified a distribution with a mean of 4 and standard deviation of 1.75. To get an appropriate scale parameter, start by taking the range of plausible scores and divide by 6, then tweak it until the distribution looks correct. For example, for group 2 the range of plausible scores was 8, dividing by 6 gave me 1.33 but the resulting distribution was too narrow so I increased it slightly. If you want a credible interval other than 95% then click Criteria... and change the 95 to the value you want.

Click Bayes Factor... if you want to get a Bayes factor for the model of no difference between means (the null) relative to the model of a difference between means (the alternative). I have selected ◉ Gonen's method (Gönen et al., 2005), because it allows us to incorporate some prior information rather than relying on defaults. Specifically, I have defined a prior belief that the difference between means will be 1 (having an invisibility cloak would lead to one additional mischievous act compared to not having the cloak), and I've set this effect to have a variance of 0.25. Back in the main dialog box click OK to fit the model.

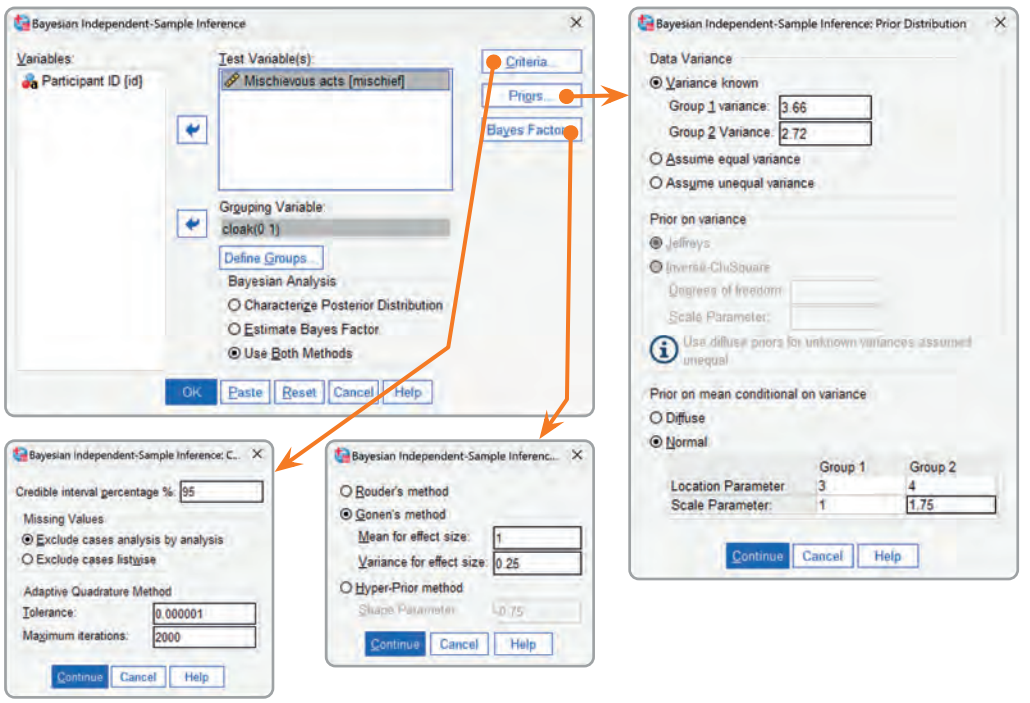


Figure 10.5 Dialog box for a Bayesian test of two independent means

Output 10.8 shows that the Bayes factor is 0.419 SPSS reports the ratio of the null to the alternative hypothesis, so this value means that the data are 0.419 times as probable under the null hypothesis as under the alternative. If we divide 1 by the value (1/0.419 = 2.39) we flip the interpretation to say that the data are 2.39 times as probable under the alternative hypothesis as under the null. In other words, we should shift our belief towards the alternative hypothesis by a factor of 2.39. Although this is in the direction of what we hypothesized, it is not particularly strong evidence for the hypothesis that invisibility cloaks lead to mischief.

Output 10.8 also shows our prior distributions for the group means (the plots labelled *Prior Distribution*), which I have already described. The Bayesian estimates of the difference between means (i.e., the $\hat{b}$-value for group as a predictor of mischief) are in the columns labelled *Posterior Mode* and *Posterior Mean*. The 95% credible interval for this estimate ranged from 0.02 to 2.60. In other words, *assuming that the effect exists*, the population value of the effect will be between 0.02 and 2.60 with 95% probability. This tells us nothing about the null hypothesis (because it assumes the effect exists) but helps us to ascertain the likely population value if we're prepared to accept that the effect exists. So, we can say with 95% probability that having a cloak of invisibility will increase mischievous acts by anything as low as 0.02 (i.e., not at all) to 2.60 (although, of course, you can't have 0.6 of a mischievous act!).

If you want to see what happens when you use reference priors, then repeat this process but select *Assume unequal variance* when you click on Priors... and Rouder's method when you click Bayes Factor... . Doing so will fit a model with diffuse, uninformative priors.

Bayes Factor Independent Sample Test (Method = Gonen)[a]

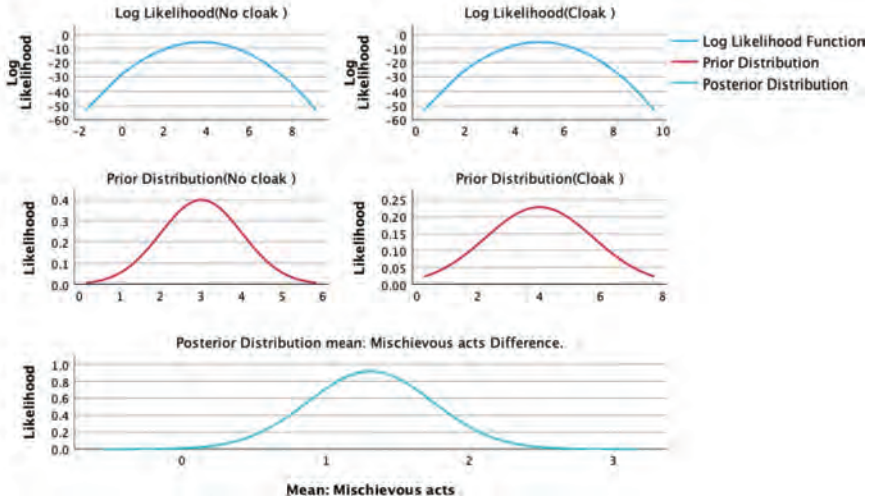
	Mean Difference	Pooled Std. Error Difference	Bayes Factor[b]	t	df	Sig.(2-tailed)
Mischievous acts	1.25	.730	.419	1.713	22	.101

a. Assumes unequal variance between groups.
b. Bayes factor: Null versus alternative hypothesis.

Posterior Distribution Characterization for Independent Sample Mean[a]

	Mode	Posterior Mean	Variance	95% Credible Interval Lower Bound	95% Credible Interval Upper Bound
Mischievous acts	1.31	1.31	.434	.02	2.60

a. Prior for Variance: Diffuse. Prior for Mean: Normal.



Output 10.8



Cramming Sam's Tips
The independent *t*-test

- The independent *t*-test compares two means, when those means have come from different groups of entities.
- Ignore the column labelled *Levene's Test for Equality of Variance* and look at the row in the table labelled *Equal variances not assumed*.
- Look at the column labelled *Two-Sided p* If the value is less than or equal to 0.05 then the means of the two groups are significantly different.
- Look at the table labelled *Bootstrap for Independent Samples Test* to get a robust confidence interval for the difference between means.
- Look at the table labelled *Independent Samples Effect Sizes* to get Cohen's $\hat{d}$.
- Look at the values of the means to tell you how the groups differ.
- A robust version of the test can be computed using syntax.
- A Bayes factor can be computed that quantifies the ratio of how probable the data are under the alternative hypothesis compared to the null.



10.9 Comparing two related means using SPSS Ⓐ

10.9.1 Entering data Ⓐ

Let's imagine that we had collected the cloak of invisibility data using a repeated-measures design: we might have recorded everyone's natural level of mischievous acts in a week, then given them an invisibility cloak and counted the number of mischievous acts in the following week.[7] The data will be identical to the previous example, partly because I am too lazy to generate different scores, but also because it allows me to illustrate various things.

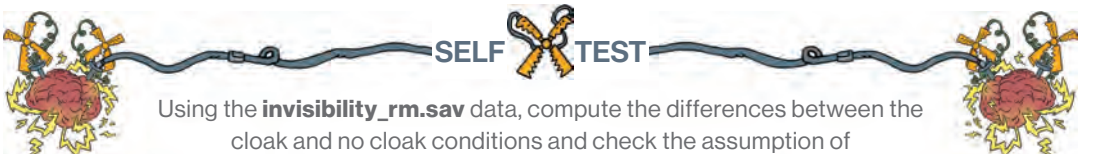
7 In theory we'd counterbalance the weeks so that some people had the cloak and then it was taken away, whereas for others they had no cloak but were then given one. However, given that the research scenario relied on participants not knowing about the cloaks of invisibility it might be best just to have a baseline phase and then give everyone their cloak at the same time (unaware that others were getting cloaks too).



The data would now be arranged differently in the data editor. Instead of having a coding variable, and a single column with mischief scores in, we would arrange the data in two columns (one representing the **cloak** condition and one representing the **no_cloak** condition). The data are in **invisibility_rm.sav** if you had difficulty entering them yourself.

10.9.2 Exploring repeated-measures data Ⓐ

We talked about the assumption of normality in Chapter 6 With the paired-samples *t*-test we're interested in the sampling distribution of the difference scores (not the raw scores). Therefore, if you want to test for normality before a paired-samples *t*-test then you should compute the differences between scores, and then check if this new variable is normally distributed as a proxy for the sampling distribution (or use a big sample or robust test and not worry about normality☺). It is possible to have two measures that are highly non-normal and produce beautifully distributed differences.



We saw in Chapter 5 that you can visualize group differences using error bars. However, there's a problem when we plot error bars from repeated-measures designs.



In one of the earlier self-tests I asked you to produce an error bar graph for the data when we treated them as an independent design. Compare that plot to the one you have just produced. Figure 10.6 shows both plots; can you spot the difference? Hopefully you can't because the plots are identical (apart from axis labels). This is odd, isn't it? We discovered in Chapter 1 that repeated-measures designs eliminate some extraneous variables (such as age and IQ) so why doesn't the plot of the repeated-measures design reflect the increased sensitivity of the design? It's because SPSS treats the data as though scores are independent, and consequently the error bars do not reflect the 'true' error around the means for repeated-measures designs. Let's correct these error bars.

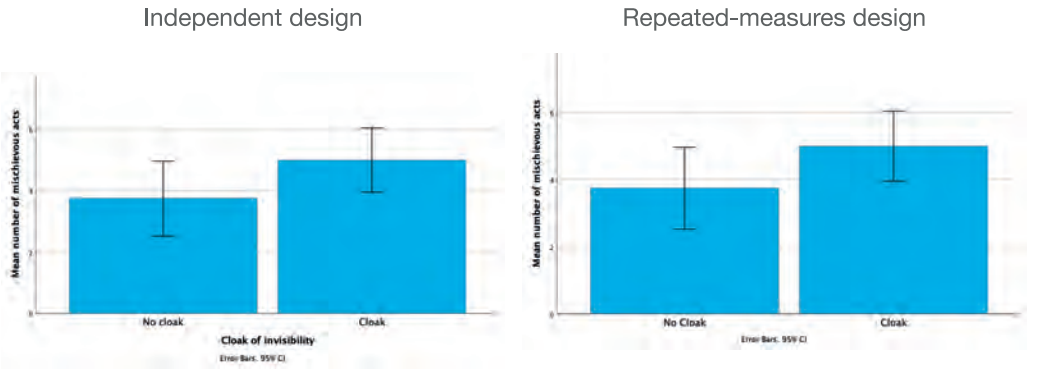


Figure 10.6 Two error bar graphs of the invisibility data

First, we need to calculate the average mischief for each participant. Select *Transform* > Compute Variable... to access the *Compute Variable* dialog box (see Section 6.9.6). Enter the name **mean** into the box labelled *Target Variable*. In the list labelled *Function group* select *Statistical*, and then in the list labelled *Functions and Special Variables* select *Mean*. Click ↑ to transfer this command to the command area. The command will appear as *MEAN(?,?)* and we need to replace the question marks with variable names by typing them or transferring them from the variables list. Replace the first question mark with the variable **no_cloak** and the second one with the variable **cloak**. The completed dialog box should look like Figure 10.7. Clicking OK will create this new variable in the data editor.

The **grand mean** is the mean of all scores and for the current data this value will be the mean of all 24 scores. The means we have just calculated are the average score for each participant; if we take the average of those mean scores, we will have the grand mean – there were a lot of means in that sentence. We will use the *descriptives* command (you could also use the *explore* command, but let's live dangerously) by selecting *Analyze* > *Descriptive Statistics* > Descriptives... (Figure 10.8). Select the variable **mean** from the list and drag it to the box labelled *Variable(s)* (or click →). Clicking Options... activates a second dialog box in which we want to uncheck everything except the mean (because that's all we are interested in). The resulting output gives us the mean of the variable that we called **mean**, which, remember, contains the average score for each participant across the cloak and no cloak conditions. This value (4.375) is the grand mean.

If you look at the variable labelled **mean**, you'll notice that the values for each participant are different, which tells us that some people were, on average, more mischievous than others across the conditions. The fact that participants' mean mischief scores differ represents individual differences (it shows that some participants are generally more mischievous than others). According to Loftus and Masson (1994), these individual differences contaminate the error bars and need to be removed. The method they suggest has some issues (Baguley, 2012; Cousineau, 2005; Morey, 2008), but it is useful for visualizing confidence intervals (although these intervals should not be used to infer anything about hypothesis tests). In effect we want to equalize the means between participants (i.e., adjust the scores in each condition such that when we take the mean score across conditions, it is the same for all participants). To do this, we use the *compute* function to calculate an adjustment factor by subtracting each participant's mean score from the grand mean. Activate the *Compute Variable* dialog box, give the target variable a name (I suggest **adjustment**) and use the command '4.375–*mean*'. This command will take the grand mean (4.375) and subtract from it each participant's average mischief level (see Figure 10.9).

This process creates a new variable in the data editor called **adjustment** which contains the difference between each participant's mean mischief levels and the mean mischief level across all participants. Some of the values are positive (participants who were less mischievous than average) and others negative (participants who were more mischievous than average). We can use these adjustment values to eliminate the between-subject differences in mischief.

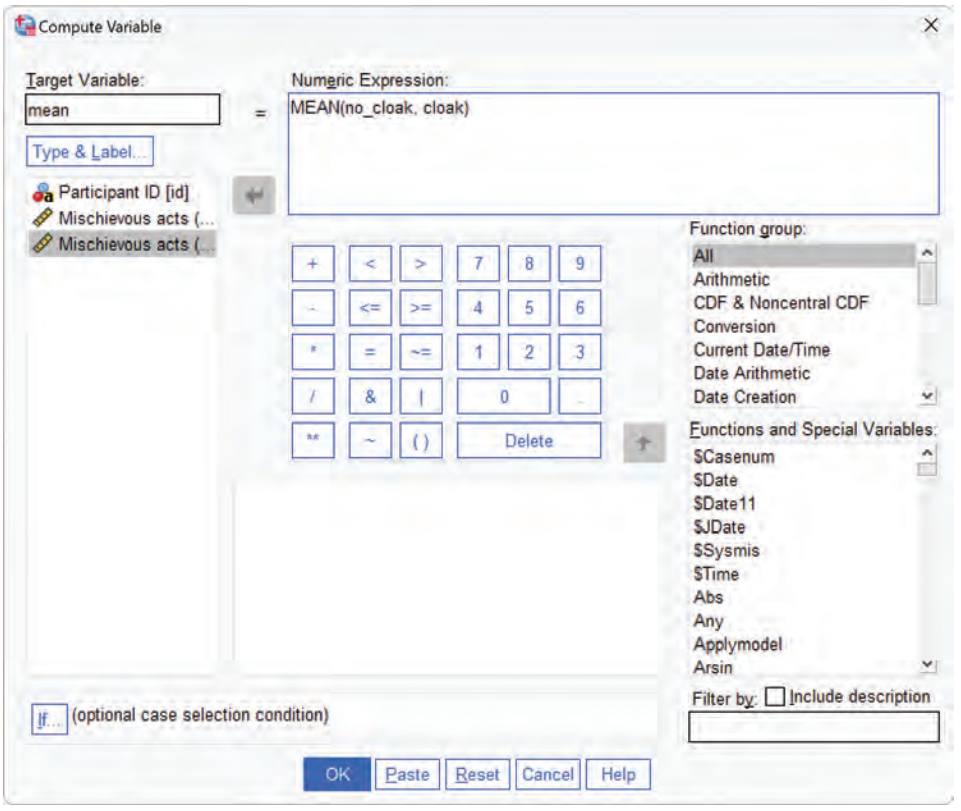


Figure 10.7 Using the *compute* function to calculate the mean of two columns

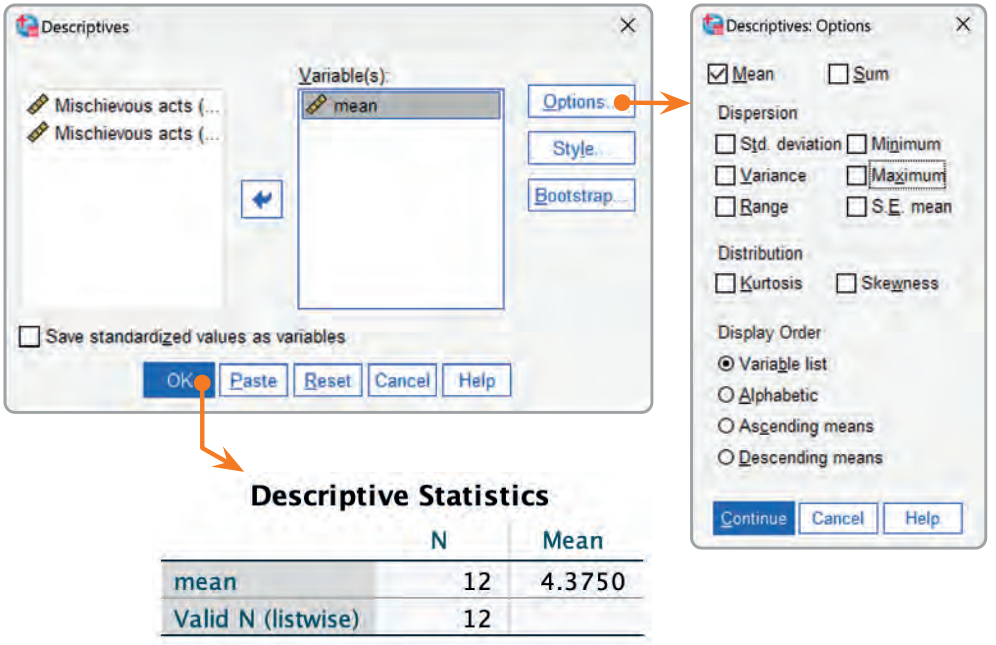


Figure 10.8 Dialog boxes and output for descriptive statistics

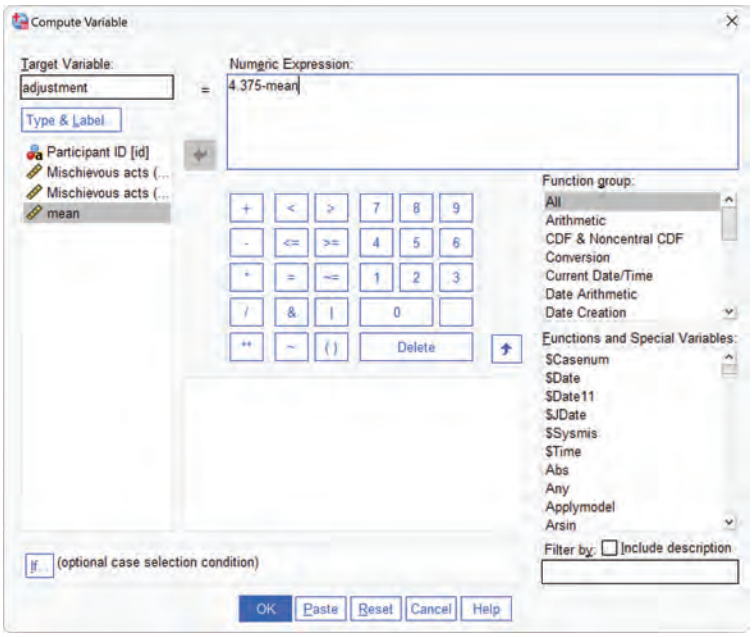


Figure 10.9 Calculating the adjustment factor

First we'll adjust the scores in the **no_cloak** condition again using the *compute* command. Activate the *Compute Variables* dialog box, and title the new variable **no_cloak_adjusted** (you can click Type & Label... to give this variable a label such as 'No cloak Condition: Adjusted Values'). We need to add each participant's score in the **no_cloak** condition to their adjustment value by dragging **no_cloak** to the command area (or select it and click ↪), then click + and drag **adjustment** to the command area. The completed dialog box is in Figure 10.10. We can create a variable called **cloak_adjusted** that contains the values of **cloak** added to the value in the **adjustment** column in the same way.

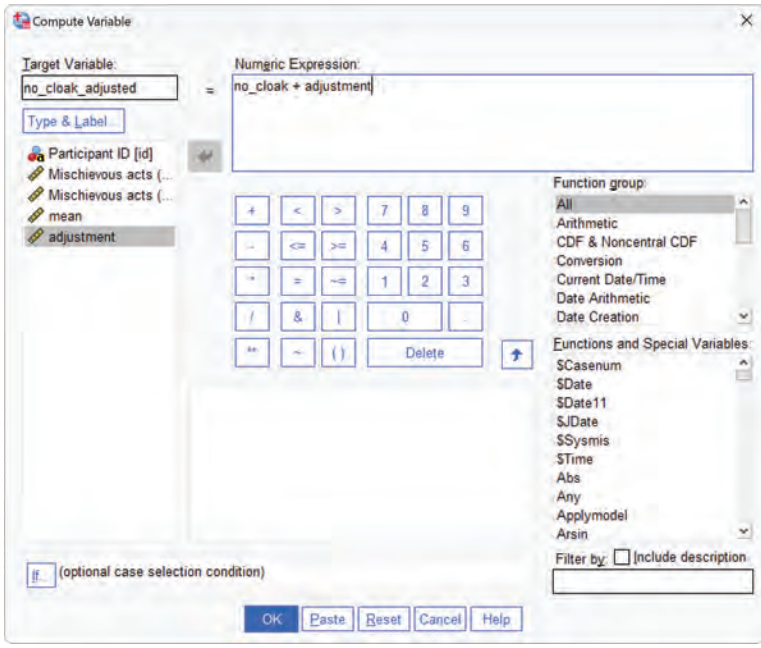


Figure 10.10 Adjusting the values of **no_cloak**

The variables **cloak_adjusted** and **no_cloak_adjusted** represent the mischief experienced in each condition, adjusted to eliminate between-subject differences. If you need proof, use the *compute* command to create a variable **mean2** that is the average of **cloak_adjusted** and **no_cloak_adjusted**. You should find that the value in this column is the same for every participant (it will be 4.375, the grand mean), demonstrating that the between-subject variability in means is gone.



Compare the resulting error bar plot (Figure 10.11) to the plots in Figure 10.6 – what difference do you notice? The means in the two conditions have not changed but the error bars have got narrower. Also, whereas in Figure 10.6 the error bars overlap, in this new plot they do not, which illustrates that the sampling error is smaller in the repeated-measures design – I expand upon this point in Section 10.10.

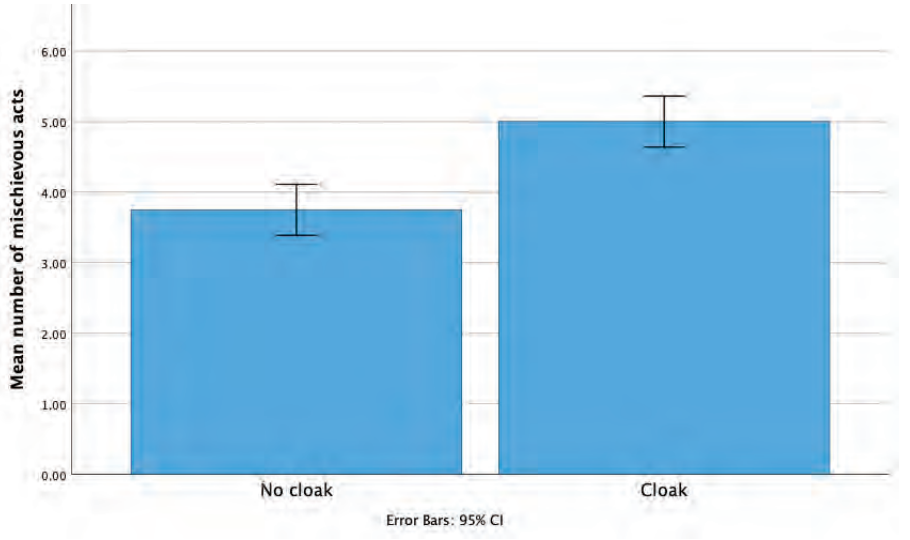


Figure 10.11 Error bar graph of the adjusted values of the data in **invisibility_rm.sav**

10.9.3 The paired-samples *t*-test using SPSS Ⓐ

To conduct a paired-samples *t*-test, access the dialog box in Figure 10.12 by selecting *Analyze* > *Compare Means* > Paired-Samples T Test.... We need to select pairs of variables to be analysed. In this case we have only one pair of interest (**cloak** vs. **no_cloak**). To select a pair click the first variable that you want to select (in this case **no_cloak**), then hold down the *Ctrl* key (*Cmd* on a Mac) and select the second (in this case **cloak**). Transfer these variables to the box labelled *Paired Variables* by clicking ▸. (You can select and transfer each variable individually, but selecting both variables as just described is quicker.) If you want to carry out several *t*-tests then you can select another pair of variables in the same way. Make sure ☑ Estimate effect sizes is selected, which will compute Cohen's $\hat{d}$. For repeated-measures designs the effect size should be corrected for the correlation between conditions, and this is done by selecting ◉ Corrected standard deviation of the difference. Clicking Options... activates another dialog box that gives you the same options as for the independent *t*-test. Similarly, click Bootstrap... to access the

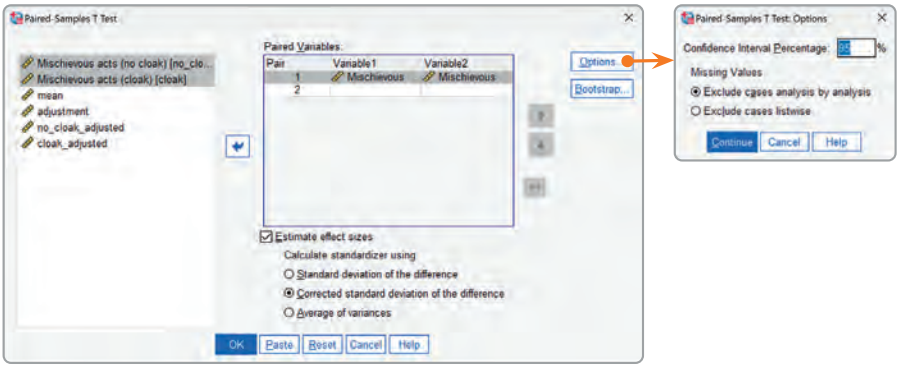


Figure 10.12 Main dialog box for paired-samples *t*-test

bootstrap function (Section 6.10.8). As with the independent *t*-test, select ☑ Perform bootstrapping and ◉ Bias corrected accelerated (BCa). Back in the main dialog box click OK to run the analysis.

10.9.4 Output from the paired-samples *t*-test Ⓐ

Output 10.9 shows a table of summary statistics for the two experimental conditions (if you don't ask for bootstrapping, this table will be a bit more straightforward). For each condition we are told the mean, the number of participants (*N*), the standard deviation and standard error. These values are the same as when we treated the data as an independent design and were described in Section 10.8.3.

Output 10.9 also shows the Pearson correlation between the two conditions. When repeated measures are used scores in the experimental conditions will correlate to some degree (because the

Paired Samples Statistics

				Bootstrap[a]			
						BCa 95% Confidence Interval	
			Statistic	Bias	Std. Error	Lower	Upper
Pair 1	Mischievous acts (no cloak)	Mean	3.75	.00	.54	2.50	4.83
		N	12				
		Std. Deviation	1.913	-.111	.362	1.165	2.300
		Std. Error Mean	.552				
	Mischievous acts (cloak)	Mean	5.00	-.01	.46	4.25	5.75
		N	12				
		Std. Deviation	1.651	-.105	.329	1.087	1.961
		Std. Error Mean	.477				

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Paired Samples Correlations

				Significance		Bootstrap for Correlation[a]			
								BCa 95% Confidence Interval	
		N	Correlation	One-Sided p	Two-Sided p	Bias	Std. Error	Lower	Upper
Pair 1	Mischievous acts (no cloak) & Mischievous acts (cloak)	12	.806	<.001	.002	-.028	.175	.267	.958

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 10.9

Paired Samples Test

		Paired Differences							Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	One-Sided p	Two-Sided p
Pair 1	Mischievous acts (no cloak) - Mischievous acts (cloak)	-1.250	1.138	.329	-1.973	-.527	-3.804	11	.001	.003

Output 10.10

data in each condition come from the same entities you'd expect some constancy in their responses). The output contains the value of Pearson's r and the significance value (see Chapter 8). For these data the experimental conditions yield a very large, highly significant, correlation coefficient, $r = 0.806$, $p = 0.002$, with a bootstrap confidence interval that doesn't include zero, BCa 95% CI [0.27, 0.96].

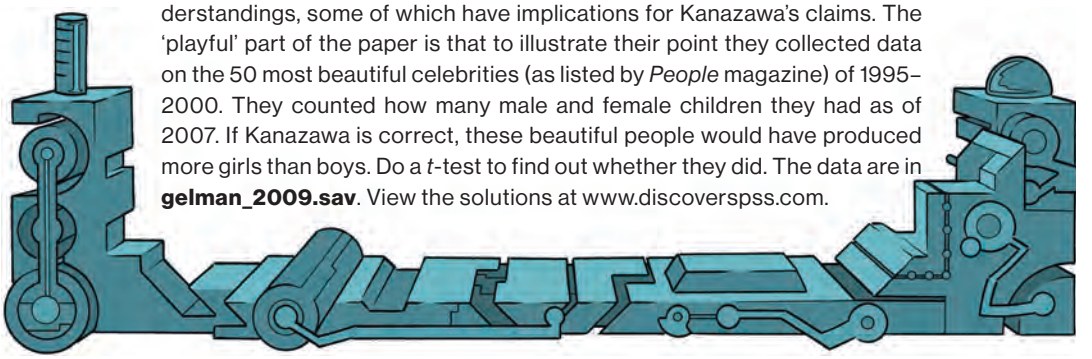
Output 10.10 tabulates the main t-test results. The table contains the average difference score ($\bar{D}$ in Eq. 10.6), which is 3.75 – 5 = –1.25, the standard deviation of the difference scores (1.138), and the standard error of difference scores (0.329, the lower half of Eq. 10.6). The test statistic, t, is calculated using Eq. 10.6 by dividing the mean of differences by the standard error of differences ($t = -1.25/0.329 = -3.804$). For a repeated-measures design the degrees of freedom are the sample size minus 1 ($df = N - 1 = 11$). The p-value in the column labelled *Two-Sided p* is the long-run probability that a value of t at least the size of the one obtained could occur if the population mean of difference scores was 0. This two-tailed probability is very low ($p = 0.003$); it tells us that there is only a 0.3% chance that a value of t of –3.804 or smaller could occur if the null hypothesis were true. Assuming we'd decided on an alpha of 0.05 before the experiment, we'd conclude that there was a significant difference between the means of these two samples, because 0.003 is less than 0.05. In terms of the experiment, we might conclude that having a cloak of invisibility significantly affected the amount of mischief a person got up to, $t(11) = -3.80$, $p = 0.003$.



Labcoat Leni's Real Research 10.3
The beautiful people Ⓐ

Gelman, A., & Weakliem, D. (2009). *American Scientist, 97*, 310–316.

Apparently there are more beautiful women in the world than there are handsome men. Satoshi Kanazawa explains this finding in terms of good-looking parents being more likely to have a baby daughter as their first child than a baby son. Perhaps more controversially, he suggests that, from an evolutionarily perspective, beauty is a more valuable trait for women than for men (Kanazawa, 2007). In a playful and very informative paper, Andrew Gelman and David Weakliem discuss various statistical errors and misunderstandings, some of which have implications for Kanazawa's claims. The 'playful' part of the paper is that to illustrate their point they collected data on the 50 most beautiful celebrities (as listed by *People* magazine) of 1995–2000. They counted how many male and female children they had as of 2007. If Kanazawa is correct, these beautiful people would have produced more girls than boys. Do a t-test to find out whether they did. The data are in **gelman_2009.sav**. View the solutions at www.discoverspss.com.



10.9.5 Robust tests of two dependent means ©

As you've probably noticed, I'm keen on looking at the robust bootstrap confidence intervals (especially in a small sample like we have here). Output 10.11 shows these confidence intervals. Remember that confidence intervals are constructed such that in 95% of samples the intervals contain the true value of the mean difference. So, assuming that this sample's confidence interval is one of the 95 out of 100 that contain the population value, we can say that the true mean difference lies between –1.75 and –0.75. The importance of this interval is that it does not contain zero (both limits are negative), which tells us that the true value of the mean difference is unlikely to be zero. In other words, there is an effect in the population reflecting more mischievous acts performed when someone is given an invisibility cloak.

Bootstrap for Paired Samples Test

					Bootstrap[a]		
						BCa 95% Confidence Interval	
		Mean	Bias	Std. Error	Sig. (2-tailed)	Lower	Upper
Pair 1	Mischievous acts (no cloak) - Mischievous acts (cloak)	-1.250	.003	.310	.010	-1.750	-.750

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 10.11

It's also possible to run a robust version of the paired-samples *t*-test using R. As was the case for the robust independent *t*-test, we need the *Essentials for R* plugin and the *WRS2* package (Section 4.12). The companion website contains a workbook file (**robust_paired_samples_t_test.spwb**) for running a robust variant of the paired-samples *t*-test based on Yuen (1974) and described by Wilcox (2017). The syntax is as follows:

```
BEGIN PROGRAM R.
library(WRS2)
mySPSSdata = spssdata.GetDataFromSPSS()
yuend(mySPSSdata$no_cloak, mySPSSdata$cloak, tr = 0.2)
END PROGRAM.
```

Run these lines of syntax (see SPSS Tip 10.3) and you'll find text output in the viewer (Output 10.12) which says that there was a significant difference (because the confidence interval does not cross zero and the *p*-value is less than 0.05) in mischief scores across the two cloak groups, $Y_t(7) = -2.70$ [–1.87, –0.13], $p = 0.031$.

```
Test statistic: -2.7027 (df = 7), p-value = 0.03052

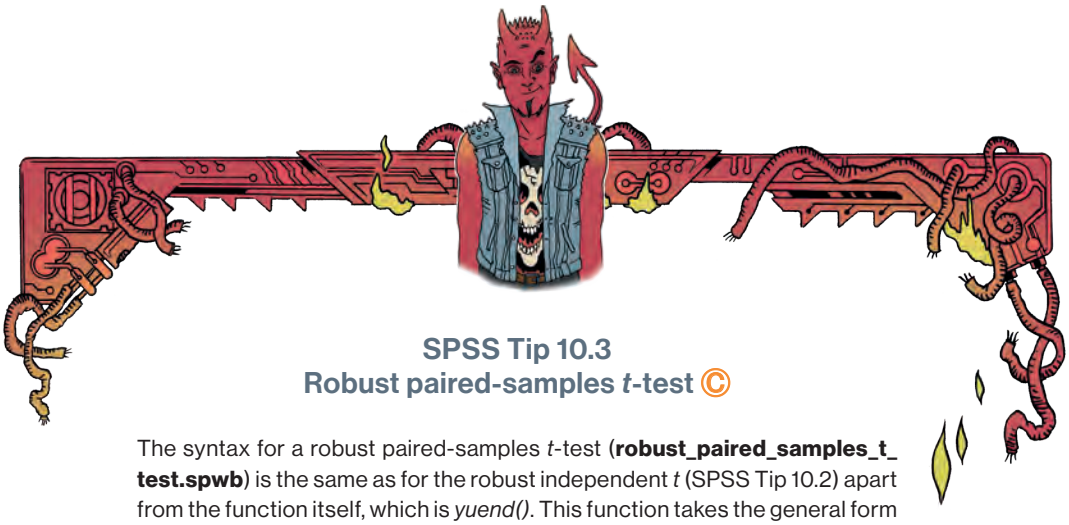
Trimmed mean difference: -1
95 percent confidence interval:
-1.8749 -0.1251

Explanatory measure of effect size: 0.4
```

Output 10.12

10.9.6 Effect sizes for two related means Ⓑ

Output 10.13 shows us the value of Cohen's $\hat{d}$ calculated using the pooled variance estimate and adjusted for the relationship between scores in the two conditions. The value suggest that there is a difference of –0.68 standard deviations between the two groups in terms of their mischief making, which is a very substantial effect. The effect size is very similar to when we used independent groups



SPSS Tip 10.3
Robust paired-samples *t*-test ©

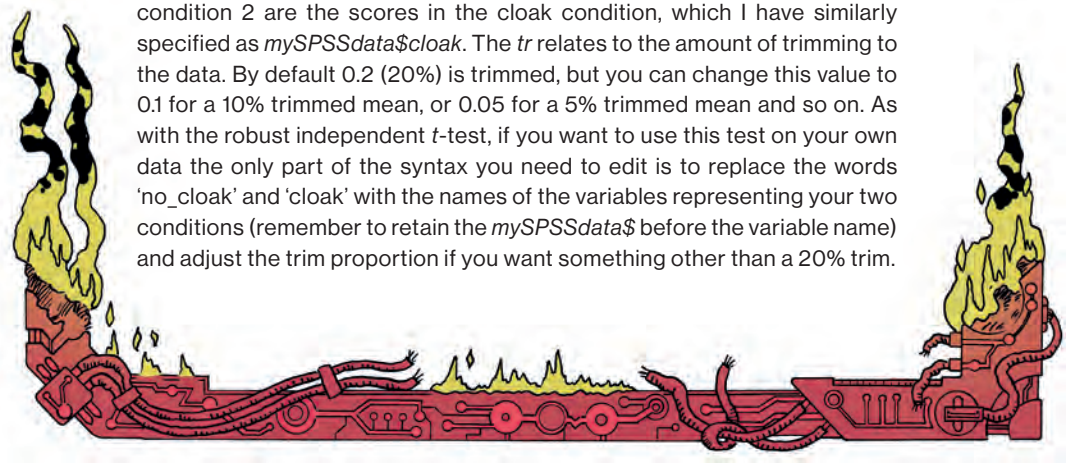
The syntax for a robust paired-samples *t*-test (**robust_paired_samples_t_test.spwb**) is the same as for the robust independent *t* (SPSS Tip 10.2) apart from the function itself, which is *yuend()*. This function takes the general form

```
yuend(scores in condition 1, scores in condition 2,
tr = 0.2)
```

In our example we end up with:

```
yuend(mySPSSdata$no_cloak, mySPSSdata$cloak, tr = 0.2)
```

The scores in condition 1 are the scores in the no cloak condition. In the data editor for the paired-samples example this is the variable **no_cloak**. However, I have written mySPSSdata$no_cloak because R recognizes variables as belonging to data frames and mySPSSdata$no_cloak means *the variable no_cloak in the dataframe mySPSSdata*. Remember that *mySPSSdata* contains the data pulled in from the SPSS data editor, so another way to translate this would be as *the variable no_cloak from the data editor.* The scores in condition 2 are the scores in the cloak condition, which I have similarly specified as *mySPSSdata$cloak*. The *tr* relates to the amount of trimming to the data. By default 0.2 (20%) is trimmed, but you can change this value to 0.1 for a 10% trimmed mean, or 0.05 for a 5% trimmed mean and so on. As with the robust independent *t*-test, if you want to use this test on your own data the only part of the syntax you need to edit is to replace the words 'no_cloak' and 'cloak' with the names of the variables representing your two conditions (remember to retain the *mySPSSdata$* before the variable name) and adjust the trim proportion if you want something other than a 20% trim.



(because the data are the same, and we wouldn't want the study design to affect the effect size). Again, the minus sign just reflects which way around the group means were subtracted. The 95% confidence interval for this effect ranges from –1.19 to –0.18, suggesting (if we assume our sample is one of the 95% that produces a confidence interval containing the true value) that the true effect is not zero. Notice that the confidence interval is narrower than when we used the independent design, reflecting the increased sensitivity of the design.

Paired Samples Effect Sizes

			Standardizer[a]	Point Estimate	95% Confidence Interval Lower	Upper
Pair 1	Mischievous acts (no cloak) – Mischievous acts (cloak)	Cohen's d	1.826	-.684	-1.194	-.175
		Hedges' correction	1.964	-.637	-1.111	-.162

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation of the mean difference adjusted by the correlation between measures.
Hedges' correction uses the sample standard deviation of the mean difference adjusted by the correlation between measures, plus a correction factor.

Output 10.13

10.9.7 Bayesian test of two means from paired samples ©

To access the dialog box to compute a Bayes factor and credible interval for dependent means select *Analyze* > *Bayesian Statistics* > Related Samples Normal (Figure 10.13). Just like with the regular dependent *t*-test, we start by selecting a pair of variables and transferring them to the *Paired Variables* list. We're expecting more mischievous acts in the cloak group than the no cloak group; therefore, if we specify our pair as **cloak** (*Variable 1*) and **no_cloak** (*Variable 2*) we'd predict a positive mean difference (when we subtract scores for the no cloak condition from those for the cloak condition the results should be, on average, positive if our prediction is correct). Conversely, if we specify our pair as **no_cloak** (*Variable 1*) and **cloak** (*Variable 2*) we'd predict a negative mean difference. In terms of setting up the prior distribution it's probably easier to think about a positive value for the mean difference, so let's specify **cloak** as *Variable 1* and **no_cloak** as *Variable 2*. To do this, select **cloak** from the *Variables* list and click ↳, then select **no_cloak** and click ↳. Select the *Variance Known* option and type 1.3 in the cell labelled *Variance Value*. The value of 1.3 is obtained by squaring the standard deviation of difference scores in Output 10.10 (1.138) to convert it to the variance. Using the variance for the data simplifies things (we don't need to set a prior for the variance) but as you get more experience you might want to estimate the variance too. To compute Bayes factors and also estimate the model parameters then select ◉ Use Both Methods.

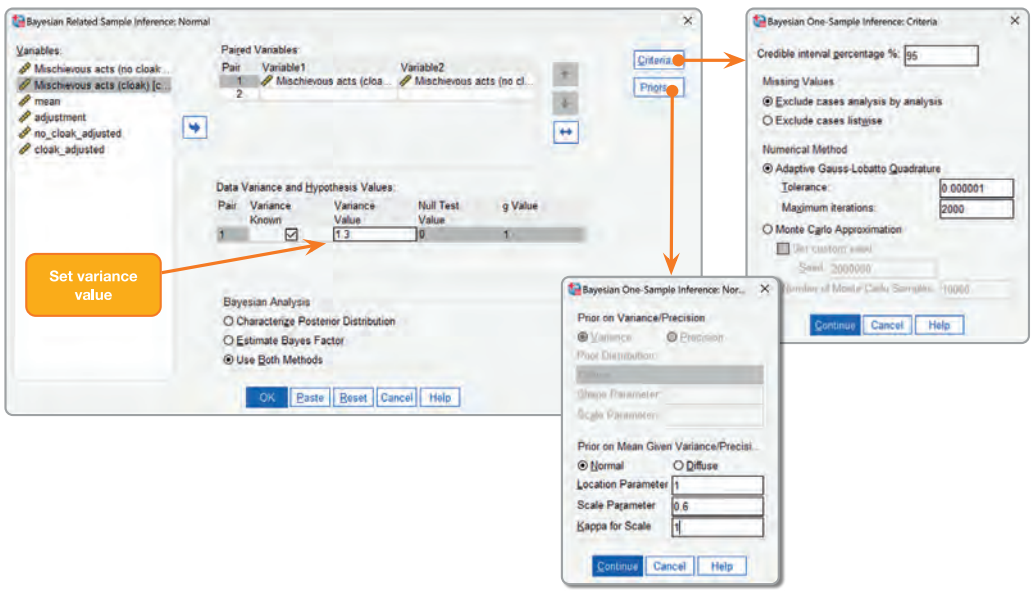


Figure 10.13 Dialog box for a Bayesian test of two related means

Imagine that prior to collecting data we believed that the difference between the number of mischievous acts committed with and without an invisibility cloak would be 1 but you were prepared to believe that it could range from about –1 to 3. In other words, you believe the most probable outcome is that having an invisibility cloak would lead to 1 additional mischievous act compared to not having a cloak. However, you were prepared to accept that, at most, it might lead to 3 additional acts (but you think this possibility has a low probability – it is highly unlikely). Conversely, you're prepared to accept that you might be wrong and that having an invisibility cloak creates a sense of responsibility in the wearer that leads to fewer mischievous acts. At most, you be prepared to believe that a cloak could lead to 1 fewer acts, but again you think this outcome is highly unlikely (you assign it low probability). The plot with labelled *Prior Distribution* in Output 10.14 shows this prior distribution. It is a normal distribution centred at 1 (the most likely outcome in your prior opinion) and ranging from about –1 to 3. This equates to a prior belief that the difference between the number of mischievous acts committed when wearing an invisibility cloak compared to when not is most likely 1, but could be as low as –1 or as high as 3 (although we think these values are unlikely).

To set up these priors click Priors... and select ◉ Normal. I have specified a distribution with mean (location) of 1 and standard deviation (scale) of 0.6. As before, to get this scale parameter I took the range of beliefs, 3–(–1) = 4, and divided by 6, 4/6 = 0.67. I then rounded down (although this won't have had an impact on the results). I set kappa to be the SPSS default of 1. If you'd rather use a non-informative reference prior (i.e., prior to data collection you're prepared to believe any value for the mean difference) then select *Diffuse*. If you want a credible interval other than 95% then click Criteria... and change the 95 to the value you want. In the main dialog box click OK to fit the model.

Bayes Factor for Related-Sample T Test

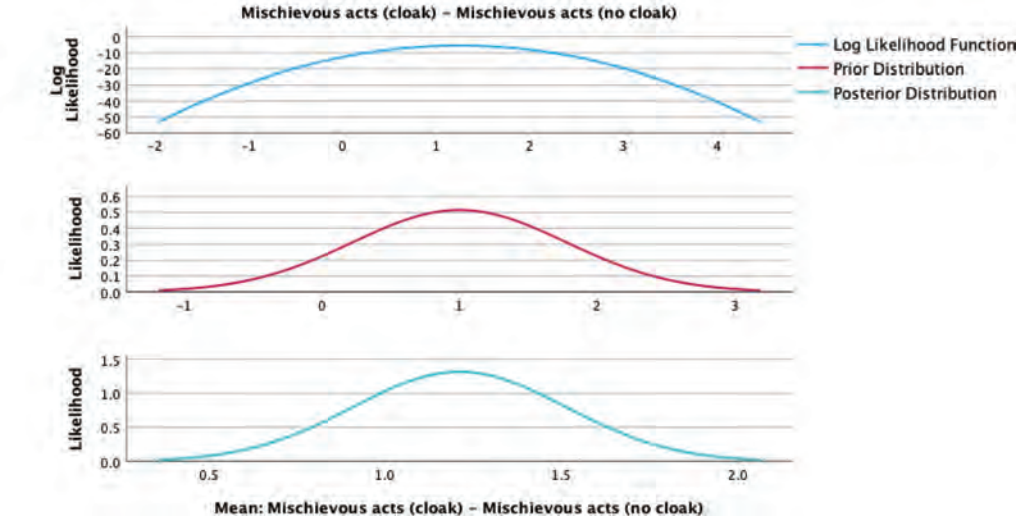
	N	Mean Difference	Std. Deviation	Std. Error Mean	Bayes Factor	t	df	Sig.(2-tailed)
Mischievous acts (cloak) – Mischievous acts (no cloak)	12	1.25	1.138	.329	.005	3.804	11	.003

Bayes factor: Null versus alternative hypothesis.

Posterior Distribution Characterization for Related-Sample Mean Difference

	N	Mode	Posterior Mean	Variance	95% Credible Interval Lower Bound	95% Credible Interval Upper Bound
Mischievous acts (cloak) – Mischievous acts (no cloak)	12	1.21	1.21	.092	.62	1.81

Prior on Variance: Diffuse. Prior on Mean: Normal.



Output 10.14

Output 10.14 shows that the Bayes factor is 0.005. SPSS reports the ratio of the null to the alternative hypothesis, so this value means that the data are 0.005 times as probable under the null hypothesis as under the alternative. We can flip the interpretation by dividing 1 by the value (1/0.005 = 200). This value tells us that the data are 200 times as probable under the alternative hypothesis as under the null. In other words, we should shift our belief towards the alternative hypothesis by a factor of 200. This is strong evidence for the hypothesis that invisibility cloaks lead to more mischief.

Output 10.14 also shows our prior distribution, which I have already described, for the difference between group means (the plot labelled *Prior Distribution*). The Bayesian estimates of the difference between means are in the columns labelled *Posterior Mode* and *Posterior Mean*. The 95% credible interval for this estimate ranged from 0.62 to 1.81. In other words, *assuming that the effect exists*, the population value of the effect will be between 0.62 to 1.81 with 95% probability. This tells us nothing about the null hypothesis (because it assumes the effect exists) but helps us to ascertain the likely population value if we're prepared to accept that the effect exists. So, we can say with 95% probability that having a cloak of invisibility will increase mischievous acts by anything as low as 0.62 to 1.81.



Cramming Sam's Tips
paired-samples *t*-test

- The paired-samples *t*-test compares two means, when those means have come from the same entities.
- Look at the column labelled *Two-Sided p*. If the value is less than or equal to 0.05 (assuming that's your alpha) then the means of the two conditions are significantly different.
- Look at the values of the means to tell you how the conditions differ.
- Look at the table labelled *Bootstrap for Paired Samples Test* to get a robust confidence interval for the difference between means.
- Look at the table labelled *Paired Samples Effect Sizes* to get Cohen's $\hat{d}$.
- A robust version of the test can be computed using syntax.
- A Bayes factor can be computed that quantifies the ratio of how probable the data are under the alternative hypothesis compared to the null.



10.10 Reporting comparisons between two means Ⓐ

As we have seen before, you usually state the finding to which the test relates and then report the test statistic, its degrees of freedom and its probability value. This applies whether we use a *t*-test or a robust test. Ideally report an estimate of the effect size and the Bayes factor too. If you used a robust test you should cite R (R Core Team, 2023) and the *WRS2* package (Mair & Wilcox, 2019), because that's what was used to compute them.

10.10.1 Reporting *t*-tests Ⓐ

For the data based on independent samples we could report this (to see from where the values come look at Outputs 10.4, 10.5 and 10.8:

✓ On average, participants given a cloak of invisibility engaged in more acts of mischief (M = 5, SE = 0.48), than those not given a cloak (M = 3.75, SE = 0.55). This difference, –1.25, BCa 95% CI [–2.70, 0.07,], was not significant, $t(21.54) = -1.71$, $p = 0.101$; however, it represented an effect of $\hat{d} = 0.70$ [–1.52, 0.13].

For the data based on paired samples you might report (see Outputs 10.10, 10.11 and 10.14):

✓ On average, participants given a cloak of invisibility engaged in more acts of mischief (M = 5, SE = 0.48), than those not given a cloak (M = 3.75, SE = 0.55). This difference, –1.25, BCa 95% CI [–1.75, –0.75], was significant, $t(11) = -3.80$, $p = 0.003$, and represented an effect of $\hat{d} = 0.68$ [–1.19, –0.18].

10.10.2 Reporting robust variants for the *t*-test Ⓐ

If you'd done the robust test, you could start with a general statement such as:

✓ R (R Core Team, 2023) was used to compute a robust variant of the *t*-test based on Yuen (1974) using the *WRS2* package (Mair & Wilcox, 2019).

Then for the data based on independent samples report this (see Output 10.6):

✓ On average, participants given a cloak of invisibility engaged in more acts of mischief (M = 5, SE = 0.48), than those not given a cloak (M = 3.75, SE = 0.55). This difference was not significant, $Y_t = -1.36$, 95% CI [–2.52, 0.52], $p = 0.167$.

Or for the data based on paired samples report this (see Output 10.12):

✓ On average, participants given a cloak of invisibility engaged in more acts of mischief (M = 5, SE = 0.48), than those not given a cloak (M = 3.75, SE = 0.55). This difference was significant, $Y_t(7) = -2.70$, 95% CI [–1.87, –0.13], $p = 0.031$.

10.10.3 Reporting Bayesian comparisons of means Ⓐ

For the data based on independent samples, report (see Output 10.8):

- ✓ On average, participants given a cloak of invisibility engaged in more acts of mischief (M = 5, SE = 0.48), than those not given a cloak (M = 3.75, SE = 0.55). The prior distributions for the group means were set as a normal distributions with a mean of 3 and standard deviation of 1 for the no cloak group and a mean of 4 and standard deviation of 1.75 for the cloak group. The Bayes factor was estimated using Gönen's method (Gönen et al., 2005) with a prior difference between means of 1 with a variance of 0.25. The Bayesian estimate of the true difference between means was 1.31, 95% credible interval [0.02, 2.60]. The associated Bayes factor, BF_{10} = 2.39, suggested that the data were weakly more probable under the alternative hypothesis than under the null.

For the data based on paired samples (see Output 10.14):

- ✓ On average, participants given a cloak of invisibility engaged in more acts of mischief (M = 5, SE = 0.48) than those not given a cloak (M = 3.75, SE = 0.55). The prior distribution for the difference between means was set to be a normal distribution with mean of 1 and standard deviation of 0.6. Kappa was set at 1. The Bayesian estimate of the true difference between means was 1.21, 95% credible interval [0.62, 1.81]. The associated Bayes factor, BF_{10} = 200, suggested that the data were 200 times more probable under the alternative hypothesis than under the null.

10.11 Between groups or repeated measures? Ⓐ

The two examples in this chapter illustrate the difference between data collected using the same participants and data collected using different participants. The two examples use the same scores but when analysed as though the data came from the same participants the result was a significant difference between means and large Bayes factor, and when analysed as though the data came from different participants there was no significant difference and a small Bayes factor. This finding may puzzle you – after all, the numbers being analysed were identical, as reflected in the effect size (d) being the same. The explanation is that repeated-measures designs have relatively more power. When the same entities are used across conditions the unsystematic variance (error variance) is reduced dramatically, making it easier to detect the systematic variance. It is often assumed that the way in which you collect data is irrelevant, and in terms of the effect size it sort of is, but if you're interested in significance then it matters. Researchers have carried out studies using the same participants in experimental conditions, then replicated the study using different participants, and have found that the method of data collection interacts significantly with the results found (see Erlebacher, 1977).

10.12 Jane and Brian's story Ⓐ

Brian wished he had a cloak of invisibility. Term was nearly over and Jane was still giving him the brush-off. The nicer he was to her, the more caustic she became. What was it with her? He felt embarrassed by his behaviour – it was a bit desperate. He wanted to be invisible for a while, or at least avoid Jane, and so he'd been hiding out in the library.

A week ago, he'd been struggling through his statistics assignment. He didn't get the whole Bayesian thing. Weeks of learning about p-values and suddenly his lecturer was throwing this whole new approach into the mix. Why don't lecturers see how confusing that is? Anyway, Brian was stuck, frustrated and about to give up when he recognized Alex and Sam from his lectures.

He went over and asked for help. They welcomed him. Alex was smart, that much was obvious, but her generosity and empathy in mentoring him and Sam were quite overwhelming. Alex was nervy around Brian and got a little tetchy whenever he mentioned Jane.



Figure 10.14 What Brian learnt from this chapter

During his week of hiding out in the library he chatted to Alex more and more. She was big-hearted, funny and seemed to enjoy his company.

Today he was alone, though, and reflecting on what a good week it had been. He'd started to feel more human – like a regular guy with normal friends. He felt calm and relaxed for the first time since the term started. That was until a finger tapped his shoulder and he turned to face the knowing smile of Jane.

'Hi, stranger,' she said.

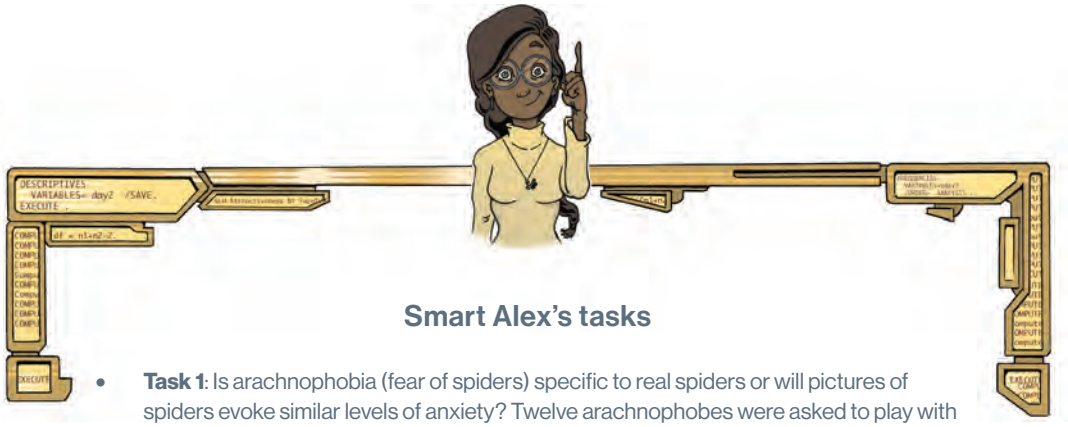
10.13 What next? Ⓐ

I'd announced to my parents that my career of choice was a rock star. Obviously I hadn't presented them with a robust *t*-test showing how much more money I would earn compared to a university professor, but even if I had, I'm not sure it would have mattered. My parents were quite happy for me to live this fantasy if I entertained the possibility that it might not work out and had a plan B. Preferably a plan B that was a little bit more sensible than being a rock star. At the age of 10, I think my plan B was probably to be a soccer star. One way or another I wanted to my career to involve being a star, so if it wasn't rock, then soccer would do. However, we've seen already that I was at a genetic

disadvantage when it came to soccer, but not so much when it came to rock stardom: my dad, after all, was quite musical. All I had to do was make it happen. The first step, I reasoned, was to build a fan base, and the best place to start a fan base is among your friends. With that in mind, I put on my little denim jacket with Iron Maiden patches sewn onto it, threw my guitar over my back and headed off down the rocky road of stardom. The first stop was my school.

10.14 Key terms that I've discovered

Dependent *t*-test
Dummy variables
Grand mean
Independent *t*-test
Paired-samples *t*-test
Standard error of differences
Variance sum law



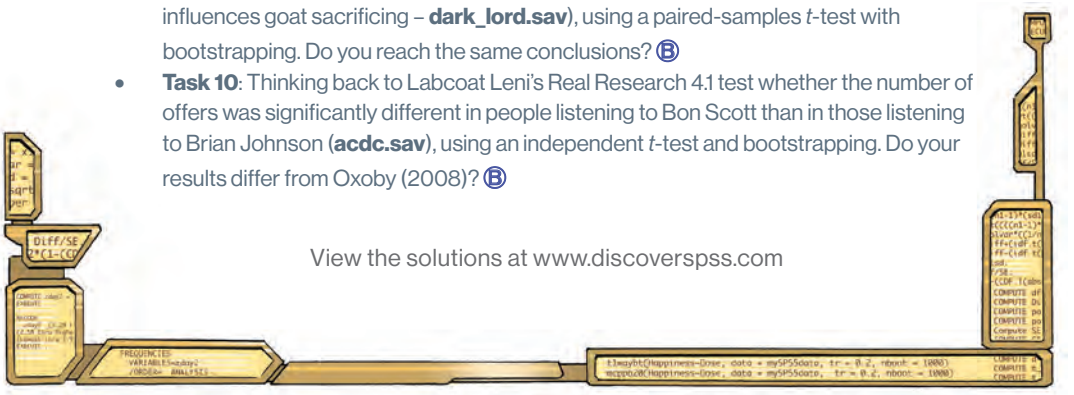
Smart Alex's tasks

- **Task 1**: Is arachnophobia (fear of spiders) specific to real spiders or will pictures of spiders evoke similar levels of anxiety? Twelve arachnophobes were asked to play with a big hairy tarantula with big fangs and an evil look in its eight eyes and at a different point in time were shown only pictures of the same spider. The participants' anxiety was measured in each case. Do a *t*-test to see whether anxiety is higher for real spiders than pictures (**big_hairy_spider.sav**). Ⓐ
- **Task 2**: Plot an error bar graph of the data in Task 1 (remember to adjust for the fact that the data are from a repeated measures design.) Ⓑ
- **Task 3**: 'Pop psychology' books sometimes spout nonsense that is unsubstantiated by science. As part of my plan to rid the world of pop psychology I took 20 people in relationships and randomly assigned them to one of two groups. One group read the famous popular psychology book *Women are from Bras and Men are from Penis*, and the other read *Marie Claire*. The outcome variable was their relationship happiness after their assigned reading. Were people happier with their relationship after reading the pop psychology book (**penis.sav**)? Ⓐ
- **Task 4**: Twaddle and Sons, the publishers of *Women are from Bras and Men are from Penis*, were upset about my claims that their book was as useful as a paper umbrella. They ran their own experiment ($N = 500$) in which relationship happiness was measured after participants had read their book and after reading one of mine (Field & Hole, 2003). (Participants read the books in counterbalanced order with a six-month delay.) Was relationship happiness greater after reading their wonderful contribution to pop psychology than after reading my tedious tome about experiments (**field_hole.sav**)? Ⓐ
- **Task 5**: In Chapter 4 (Task 6) we looked at data from people who had fish or cats as pets and measured their life satisfaction as well as how much they like animals

(**pets.sav**). Conduct a *t*-test to see whether life satisfaction depends upon the type of pet a person has. Ⓐ

- **Task 6:** Fit a linear model to the data in Task 5 to see whether life satisfaction is significantly predicted from the type of animal. What do you notice about the *t*-value and significance in this model compared to Task 5. Ⓐ
- **Task 7:** In Chapter 6 we looked at hygiene scores over threethree days of a rock music festival (**download.sav**). Do a paired-samples *t*-test to see whether hygiene scores on day 1 differed from those on day 3. Ⓐ
- **Task 8:** Analyse the data in Chapter 6, Task 1 (whether men and dogs differ in their dog-like behaviours) using an independent *t*-test with bootstrapping. Do you reach the same conclusions (**men_dogs.sav**)? Ⓑ
- **Task 9:** Analyse the data in Chapter 6, Task 2 (whether the type of music you hear influences goat sacrificing – **dark_lord.sav**), using a paired-samples *t*-test with bootstrapping. Do you reach the same conclusions? Ⓑ
- **Task 10**: Thinking back to Labcoat Leni's Real Research 4.1 test whether the number of offers was significantly different in people listening to Bon Scott than in those listening to Brian Johnson (**acdc.sav**), using an independent *t*-test and bootstrapping. Do your results differ from Oxoby (2008)? Ⓑ

View the solutions at www.discoverspss.com





MODERATION AND MEDIATION

11.1 What will this chapter tell me? 522
11.2 The PROCESS tool 523
11.3 Moderation: interactions in the linear model 523
11.4 Mediation 539
11.5 Jane and Brian's story 554
11.6 What next? 554
11.7 Key terms that I've discovered 555
Smart Alex's tasks 555

Letters A to D indicate increasing difficulty

11.1 What will this chapter tell me? Ⓐ

Having successfully wowed audiences at holiday camps around Britain, my next step towards global domination was my primary school. I had learnt another Chuck Berry song ('Johnny B. Goode'), but broadened my repertoire to include songs by other artists (I have a feeling 'Over the edge' by Status Quo was one of them).[1] When the opportunity came to play at a school assembly I jumped at it. The headmaster tried to ban me,[2] but the show went on. It was a huge success (10-year-olds are very easily impressed). My classmates carried me around the playground on their shoulders. I was a hero. Around this time I had a childhood sweetheart called Clair Sparks. Actually, we had been sweethearts since before my newfound rock legend status. I don't think the guitar playing and singing impressed her much, but she rode a motorbike (really, a little child's one) which impressed *me* quite a lot; I was utterly convinced that we would one day get married and live happily ever after. I was utterly convinced, that is, until she ran off with Simon Hudson. Being 10, she probably literally did run off with him – across the playground. I remember telling my parents and them asking me how I felt about it. I told them I was being philosophical about it. I probably didn't know what 'philosophical' meant at the age of 10, but I knew that it was the sort of thing you said if you were pretending not to be bothered about being dumped.



Figure 11.1 My 10th birthday. From left to right my brother Paul (who still hides behind cakes rather than have his photo taken), Paul Spreckley, Alan Palsey, Clair Sparks and me

If I hadn't been philosophical, I might have wanted to look at what had lowered Clair's relationship satisfaction. We've seen in previous chapters that we could predict things like relationship satisfaction using a linear model. Perhaps it's predicted from your partner's love of rock bands like Status Quo

1 This would have been about 1982, so just before they abandoned hard rock in favour of a long series of increasingly cringeworthy publicity stunts. Nevertheless, up to the age of 10 they were my favourite band.

2 Seriously! I grew up in an age when a headmaster would try to ban a 10-year-old from playing guitar in assembly. The guy used to play hymns on an acoustic guitar, and I can assume only that he lost all perspective on the situation and decided that a 10-year-old blasting out some Quo in a squeaky little voice with an electric guitar was subversive or something.

(I don't recall Clair liking that sort of thing). However, life is usually more complicated than this; for example, your partner's love of rock music probably depends on your own love of rock music. For example, if you both like rock music then your love of the same music might have an additive effect, giving you huge relationship satisfaction (*moderation*), or perhaps the relationship between your partner's love of rock and your own relationship satisfaction can be explained by your own music tastes (*mediation*). This chapter extends what we know about the linear model to these more complicated scenarios.

11.2 The PROCESS tool Ⓑ

The best way to tackle moderation and mediation is with the PROCESS tool, which needs to be installed using the instructions in Section 4.13.1.

11.3 Moderation: interactions in the linear model Ⓒ

11.3.1 The conceptual model Ⓒ

So far we have looked at individual predictors in the linear model. It is possible for a statistical model to include the combined effect of two or more predictor variables on an outcome, which is known conceptually as **moderation**, and in statistical terms as an **interaction effect**. We'll start with the conceptual, using an example of whether violent video games make people antisocial. Video games are among the favourite online activities for young people: two-thirds of 5–16-year-olds have their own video games console, with 88% of boys aged 8–15 owning at least one games console (Ofcom, 2006). Although playing violent video games can enhance visuospatial acuity, visual memory, probabilistic inference, and mental rotation (Feng et al., 2007; Green et al., 2010; Green & Bavelier, 2007, 2012; Mishra et al., 2011) compared to games such as Tetris,[3] these games have also been linked to increased aggression in youths (Anderson & Bushman, 2001). Another predictor of aggression and conduct problems are callous-unemotional traits such as lack of guilt, lack of empathy, and callous use of others for personal gain (Docherty et al., 2019; Rowe et al., 2010). Imagine that a scientist explored the relationship between aggression and playing violent video games such as *Grand Theft Auto*, *MadWorld* and *Manhunt*. She measured aggressive behaviour (**aggress**), callous-unemotional traits (**caunts**) and the number of hours per week they play video games (**vid_game**) in 442 youths (**video_games.sav**).



Let's assume we're interested in the relationship between the hours spent playing these games (predictor) and aggression (outcome). The conceptual model of moderation in Figure 11.2 shows that

3 For completeness, this evidence is contested by others (Boot et al., 2011; van Ravenzwaaij et al., 2014).

a **moderator** variable is one that affects the relationship between two others. If callous-unemotional traits were a moderator then we're saying that the strength or direction of the relationship between game playing and aggression is affected by callous-unemotional traits.

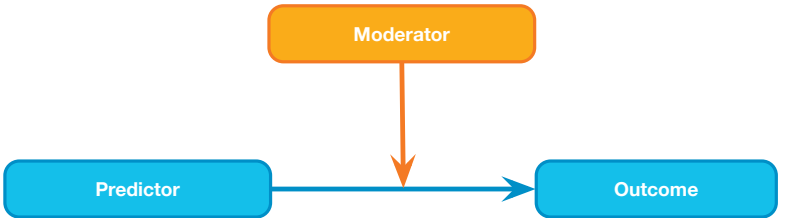


Figure 11.2 Diagram of the *conceptual* moderation model

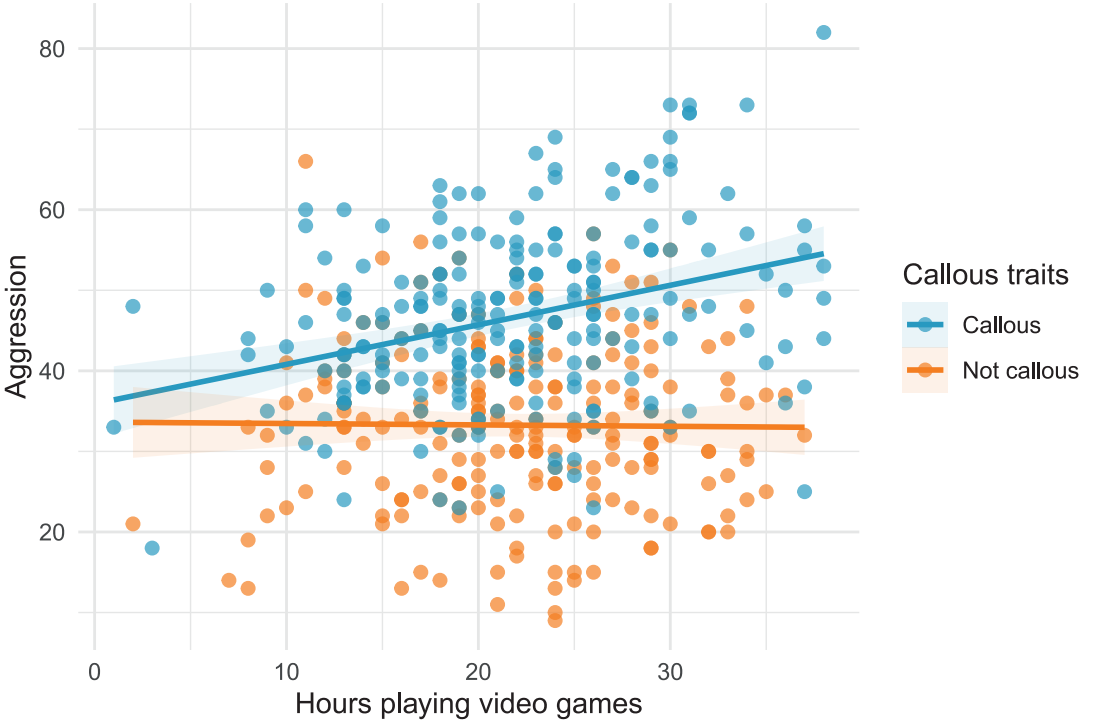


Figure 11.3 A categorical moderator (callous traits)

Suppose that we could classify people as having callous-unemotional traits or not. Our moderator variable would be categorical (callous or not callous). Figure 11.3 shows an example of moderation in this case: for people who are not callous there is no relationship between video games and aggression (the line is completely flat), but for people who are callous there is a positive relationship: as the time spent playing these games increases, so do aggression levels (the line slopes upwards). Therefore, callous-unemotional traits moderate the relationship between video games and aggression: there is a positive relationship for those with callous-unemotional traits but not for those without. This is a simple way to think about moderation but it is not necessary that there is an effect in one group but not in the other, just that there is a difference in the relationship between video games and aggression in the two callousness groups. It could be that the effect is weakened or changes direction.

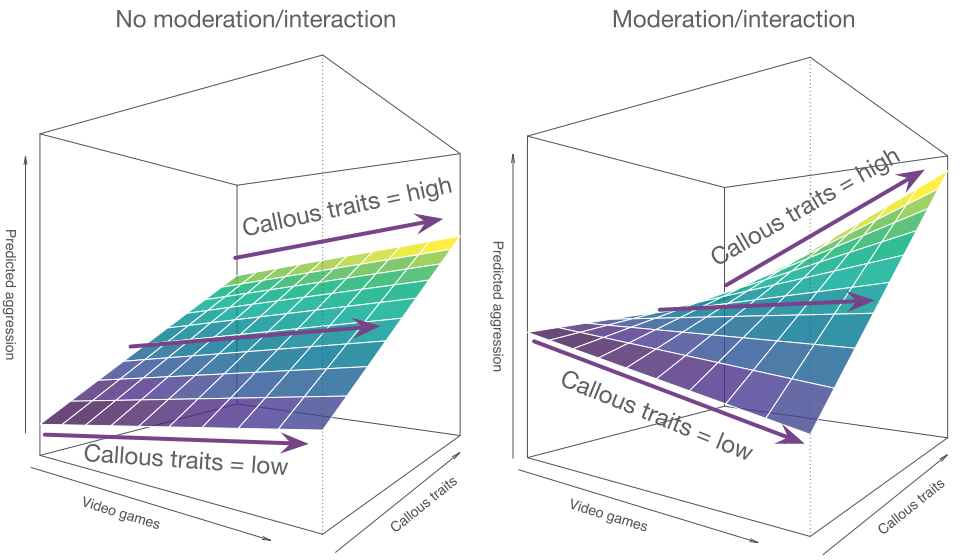


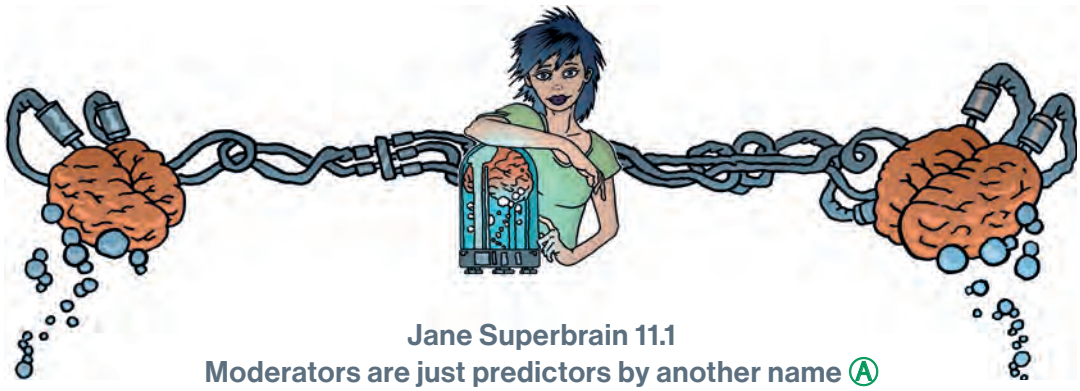
Figure 11.4 A continuous moderator (callous traits)

If we measure the moderator variable along a continuum it becomes trickier to visualize, but the basic interpretation stays the same. Figure 11.4 shows two plots that display the relationships between the time spent playing video games, aggression and callous-unemotional traits (measured along a continuum rather than as two groups). We're still interested in how the relationship between video games and aggression changes as a function of callous-unemotional traits. We can explore this by comparing the slope of the regression plane for time spent gaming as values of callous-unemotional traits increase. To help you I have added arrows that show the relationship between video games and aggression. In Figure 11.4 (left) you can see that at the low end of the callous-unemotional traits scale, there is a slight positive relationship between playing video games and aggression (as time playing games increases so does aggression). At the high end of the callous-unemotional traits scale, we see a very similar relationship between video games and aggression (the top and bottom of the regression plane slope at the same angle). The same is also true at the middle of the callous-unemotional traits scale. This is a case of no interaction or no moderation.

Figure 11.4 (right) shows moderation: at low values of callous-unemotional traits the plane slopes downwards, indicating a slightly negative relationship between playing video games and aggression, but at the high end of callous-unemotional traits the plane slopes upwards, indicating a strong positive relationship between gaming and aggression. At the midpoint of the callous-unemotional traits scale, the relationship between video games and aggression is relatively flat. So, as we move along the callous-unemotional traits variable, the relationship between gaming and aggression changes from slightly negative to neutral to strongly positive. We can say that the relationship between violent video games and aggression is *moderated* by callous-unemotional traits.

11.3.2 The statistical model Ⓑ

Now we know what moderation is conceptually, let's look at how we test for moderation within a statistical model. Figure 11.5 shows the statistical model: we predict the outcome from the predictor variable, the proposed moderator, and the interaction of the two. It is the interaction effect that tells us whether moderation has occurred, but *we must include the predictor and moderator for the interaction term to be valid*. This point is very important. In our example, then, we'd be looking at fitting a linear model predicting aggression (the outcome) from video game playing, callous-unemotional traits and their interaction. See Jane Superbrain Box 11.1 for some important points about moderators and interactions.



Jane Superbrain 11.1
Moderators are just predictors by another name Ⓐ

The term 'moderation' implies that the moderator *alters* the relationship between the other two variables. A significant interaction effect does not justify making this causal assumption, it merely shows that two variables have a combined effect on the outcome. Therefore, although interactions are used to test for moderation, they are not the same thing (Hall & Sammons, 2014). Related to this point, although we label one predictor as a 'moderator', statistically speaking this label changes nothing: it is just a predictor by another name. Therefore, although theoretically it might make sense to think of one variable as a predictor and the other as a moderator, as far as the interaction term is concerned both variables are predictors and either one is moderating the effect of the other on the outcome. In this example, it is theoretically meaningful to think of callous-unemotional traits as moderating the effect of gaming on aggression, but if the interaction term is significant it would be equally valid to interpret the effect as video game use moderating the effect of callous-unemotional traits on aggression.



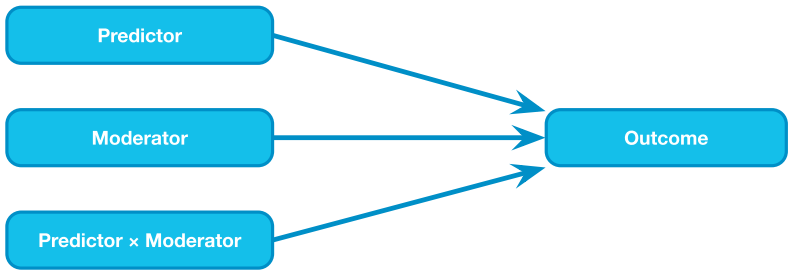


Figure 11.5 Diagram of the *statistical* moderation model

We've encountered the general form of the linear model many times already. The following equation is to refresh your memory (but if it is completely new information to you then read Chapter 9 before continuing):

$$\text{outcome}_i = \text{model} + \text{error}_i$$
$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} + \varepsilon_i \quad (11.1)$$

By replacing the Xs with the names of our predictor variables, and the Y with our outcome variable, the linear model becomes:

$$\text{Aggression}_i = b_0 + b_1\text{Gaming}_i + b_2\text{Callous}_i + \varepsilon_i \quad (11.2)$$

To test for moderation we need to consider the interaction between gaming and callous-unemotional traits. We have seen before that to add variables to a linear model we literally add them in and assign them a parameter (b). Therefore, if we had two predictors labelled A and B, a model that tests for moderation would be expressed as:

$$Y_i = b_0 + b_1A_i + b_2B_i + b_3AB_i + \varepsilon_i \quad (11.3)$$

If we replace the A and B with the names of the variables for this specific example we can express the model as:

$$\text{Aggression}_i = b_0 + b_1\text{Gaming}_i + b_2\text{Callous}_i + b_3\text{Interaction}_i + \varepsilon_i \quad (11.4)$$

11.3.3 Centring variables Ⓑ

When an interaction term is included in the model the b parameters have a specific meaning. For the individual predictors they represent the regression of the outcome on that predictor when the other predictor is zero. So, in Eq. 11.4, b_1 represents the relationship between aggression and gaming when callous traits are zero, and b_2 represents the relationship between aggression and callous traits when someone spends zero hours gaming per week. This interpretation isn't problematic because zero is a meaningful score for both predictors: it's plausible both that a child spends no hours playing video games, and that a child gets a score of 0 on the continuum of callous-unemotional traits. However, there are often situations where it makes no sense for a predictor to have a score of zero. Imagine that rather than measuring how much a child played violent video games we'd measured their heart rate while playing the games as an indicator of their physiological reactivity to them:

$$\text{Aggression}_i = b_0 + b_1\text{HeartRate}_i + b_2\text{Callous}_i + b_3\text{Interaction}_i + \varepsilon_i \quad (11.5)$$

In this model b_2 is the regression of the aggression on callous traits when someone has a heart rate of zero while playing the games. This b makes no sense unless we're interested in knowing something about the relationship between callous traits and aggression in youths who die (and therefore have a heart rate of zero) while playing these games. It's fair to say that if video games killed the players, we'd have more to worry about than their developing aggression. The point is that the presence of the interaction term makes the bs for the main predictors uninterpretable in many situations.

For this reason, it is common to transform the predictors using **grand mean centring**. 'Centring' refers to the process of transforming a variable into deviations around a fixed point. This fixed point can be any value that you choose, but typically it's the grand mean. When we calculated z-scores in Chapter 6 we used grand mean centring because the first step was to take each score and subtract from it the mean of all scores. This is grand mean centring. Like z-scores, the subsequent scores are centred on zero, but unlike z-scores, we don't care about expressing the centred



scores as standard deviations.[4] Therefore, grand mean centring for a given variable is achieved by taking each score and subtracting from it the mean of all scores (for that variable).

Centring the predictors has no effect on the b for highest-order predictor, but will affect the bs for the lower-order predictors. *Order* refers to how many variables are involved: the gaming × callous traits interaction is a higher-order effect than the effect of gaming alone because it involves two variables rather than one. In our model (Eq. 11.4), whether we centre the predictors has no effect on b_3 (the parameter for the interaction) but it changes the values of b_1 and b_2 (the parameters for gaming and callous traits). If we don't centre the gaming and callous variables, the bs represent the effect of the predictor when the other predictor is zero. However, if we centre the gaming and callous variables, the bs represent the effect of the predictor when the other predictor is at its mean value. For example, b_2 would represent the relationship between aggression and callous traits for someone who spends the average number of hours gaming per week.

Centring is important when your model contains an interaction term because it makes the bs for lower-order effects interpretable. There are good reasons for not caring about the lower-order effects when the higher-order interaction involving those effects is significant; for example, if the gaming × callous traits interaction is significant, then it's not clear why we would be interested in the individual effects of gaming and callous traits. However, when the interaction is not significant, centring makes interpreting the main effects easier. With centred variables the bs for individual predictors have two interpretations: (1) they are the effect of that predictor at the mean value of the sample;



Oliver Twisted
Please, Sir, can I have some more ... centring?

'Recentgin,' babbles Oliver as he stumbles drunk out of Mrs Moonshine's alcohol emporium. 'I've had some recent gin.' I think you mean *centring* Oliver, not *recentgin*. If you want to know how to centre your variables using SPSS, then gaze drunkenly at www.discoverspss.com.

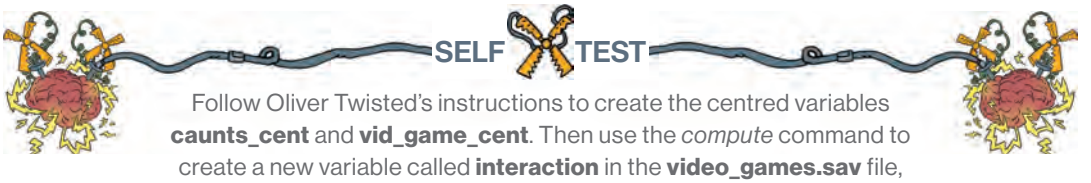
4 Remember that with z-scores we go a step further and divide the centred scores by the standard deviation of the original variable, which changes the units of measurements to standard deviations. You can do this if you want to compare bs across predictors within a model, but remember that b would represent the change in the outcome associated with a 1 standard deviation change in the predictor.

and (2) they are the average effect of the predictor across the range of scores for the other predictors. To explain the second interpretation, imagine we took everyone who spent no hours gaming, estimated the linear model between aggression and callous traits and noted the *b*. We then took everyone who played games for 1 hour and did the same, and then took everyone who gamed for 2 hours per week and did the same. We continued doing this until we had estimated linear models for every different value of the hours spent gaming. We'd have a lot of *b*s, each one representing the relationship between callous traits and aggression for different amounts of gaming. If we took an average of these *b*s then we'd get the same value as the *b* for callous traits (centred) when we use it as a predictor with gaming (centred) and their interaction.

The PROCESS tool does the centring for us so we don't need to worry about how it's done, but because centring is useful in other analyses Oliver Twisted has some additional material that shows you how to do it manually for this example.

11.3.4 Creating interaction variables Ⓑ

Eq. 11.5 contains a variable called 'Interaction', but the data file does not. You might ask how you enter a variable into the model that doesn't exist in the data. You create it, and it's easier than you might think. Mathematically speaking, when we look at the combined effect of two variables (an interaction) we are literally looking at the effect of the two variables multiplied together. The interaction variable in our example would be the scores on the time spent gaming multiplied by the scores for callous-unemotional traits. That's why interactions are denoted by *variable 1* × *variable 2*. The PROCESS tool creates the interaction variable for you, but the self-help task gives you practice at doing it manually (for future reference).



Follow Oliver Twisted's instructions to create the centred variables **caunts_cent** and **vid_game_cent**. Then use the *compute* command to create a new variable called **interaction** in the **video_games.sav** file, which is **caunts_cent** multiplied by **vid_game_cent**.

11.3.5 Following up an interaction effect Ⓑ

If the moderation effect is significant, we need to interpret it. In our example, we're predicting that the moderator (callous traits) influences the relationship between playing violent video games and aggression. If the interaction of callous traits and time spent gaming is a significant predictor of aggression then we know that we have a significant moderation effect, but we don't know the nature of the effect. It could be that the time spent gaming always has a positive relationship with aggression, but that the relationship gets stronger the more a person has callous traits. Alternatively, perhaps the time spent gaming *reduces* aggression in people low on callous traits but *increases* aggression in those high on callous traits (i.e., the relationship reverses). To find out what is going on we need to do a **simple slopes analysis** (Aiken & West, 1991; Rogosa, 1981).

The idea behind simple slopes analysis is no different than what was illustrated in Figure 11.4. When describing that figure I talked about comparing the relationship between the predictor (time spent gaming) and outcome (aggression) at low and high levels of the moderator (callous traits). For example, in Figure 11.4 (right), we saw that time spent gaming and aggression had a slightly negative relationship at low levels of callous traits, but a positive relationship at high levels of callous traits. This is the essence of simple slopes analysis: we work out the model equations for the predictor and

outcome at low, high and average levels of the moderator. The 'high' and 'low' levels can be anything you like, but 1 standard deviation above and below the mean value of the moderator is typically used. Therefore, in our example, we would get the linear model for aggression predicted from hours spent gaming for the average value of callous traits, for 1 standard deviation above the mean value of callous traits and for 1 standard deviation below the mean value of callous traits. We compare these slopes in terms of both their significance and the value and direction of the *b* to see whether the relationship between hours spent gaming and aggression changes at different levels of callous traits.

A related approach is to look at how the relationship between the predictor and outcome changes at many values of the moderator and not at only high, low and mean values (Johnson & Neyman, 1936). Essentially, the method estimates the model including only the predictor and outcome at lots of different values of the moderator. For each model it computes the significance of the *b* for the predictor so you can see for which values of the moderator the relationship between the predictor and outcome is significant. It returns a 'zone of significance',[5] which consists of two values of the moderator. This interval is sometimes referred to as the **Johnson–Neyman interval**. Typically, between these two values of the moderator the predictor does not significantly predict the outcome, whereas below the lower value and above the upper value of the moderator the predictor significantly predicts the outcome.

11.3.6 Moderation analysis using SPSS Ⓑ

Given that moderation is demonstrated through a significant interaction between the predictor and moderator in a linear model, we could follow the general procedure in Chapter 9 (Figure 9.8). We would first centre the predictor and moderator, then create the interaction term as discussed already, then run a forced entry regression with the centred predictor, centred moderator and the interaction of the two centred variables as predictors. The advantage of this approach is that we can inspect sources of bias in the model.



Using the PROCESS tool has several advantages over using the normal regression menu in IBM SPSS Statistics software: (1) it centres predictors for us; (2) it computes the interaction term automatically; and (3) it produces simple slopes analysis. To access the dialog boxes in Figure 11.6 select *Analyze* ❯ *Regression* ❯ *Process v4.2 by Andrew F. Hayes*.[6] Drag (or click ▸) the outcome variable (**aggress**) from the box labelled *Variables* to the box labelled *Y Variable*. Similarly, drag the predictor variable (**vid_game**) to the box labelled *X Variable*. Finally, drag the moderator variable (**caunts**) to the box labelled *Moderator Variable W*. PROCESS can run over 74 different models, and these models are listed in the drop-down box labelled *Model number* (details of the models are in Hayes, 2022). Select 1 from this menu, which is a simple moderation model. To get robust confidence intervals for effects select ☑ Bootstrap inference for model coefficients and if you want something other than a 95% confidence interval then choose a different value from the drop-down list.

5 I must be careful not to confuse this with my wife, who is the Zoë of significance.

6 If the menu isn't there, see Section 4.13.1. The v4.2 reflects the fact I'm using version 4.2, by the time you read this this number might have changed.

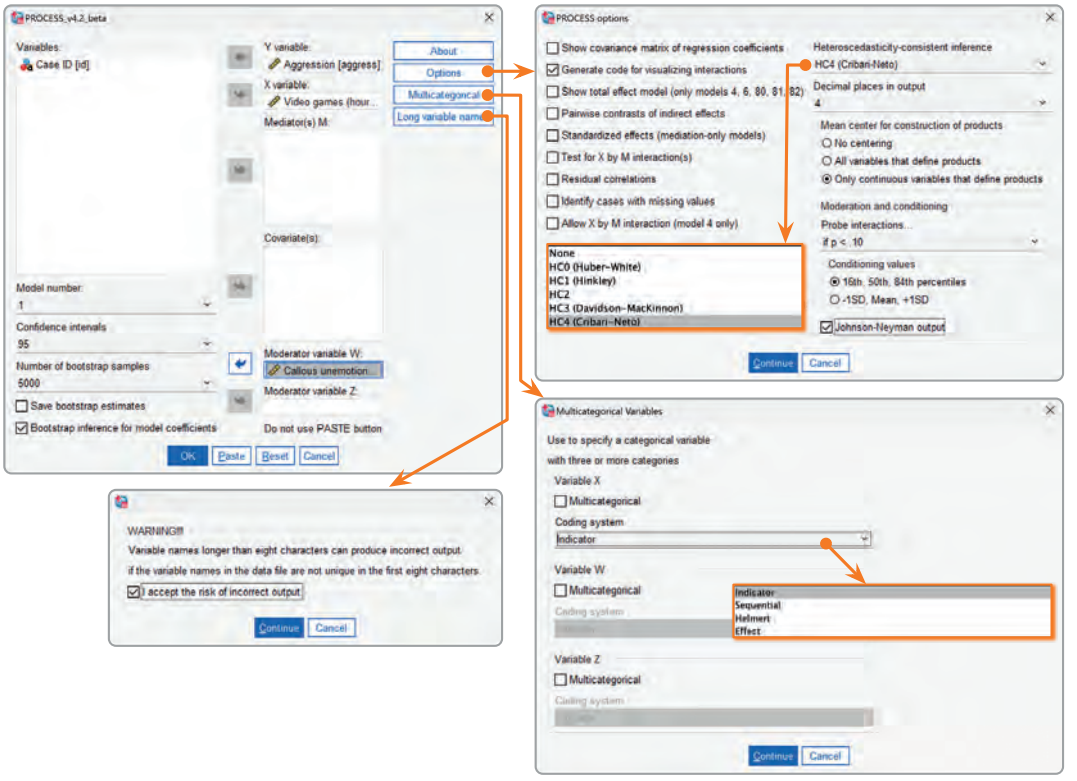


Figure 11.6 The dialog boxes for running moderation analysis

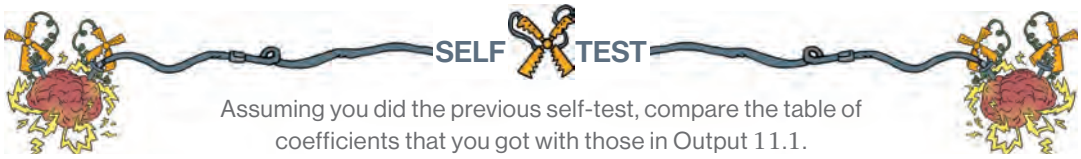
Clicking Options... activates a dialog box containing many useful options for moderation. First, selecting ☑ Generate code for visualizing interactions will generate some code that we can use for plotting the moderation effects. The section labelled *Heteroscedasticity-consistent inference* allows us to select one of the methods for adjusting for heteroscedasticity in Section 6.10.7. I suggest selecting HC4 and basking in the relief of not having to think about the assumption any more. The section labelled *Mean centre for the construction of products* contains options for mean centring. The default is not to centre, and you have to be careful about mean centring variables that are categories, so it makes sense to select ◉ Only continuous variables that define products. In the current example all variables are continuous anyway. The section labelled *Moderation and conditioning* contains options for picking apart the interaction effect. By default PROCESS will produce follow-up tests on any interaction that is significant at $p < 0.1$. Given that researchers very often use an alpha of 0.05 or 0.005, you should think about changing this to a smaller threshold in the drop-down list, but for these data it doesn't make a difference. The options in *Conditioning values* determine the values for the simple slopes analysis. By default PROCESS will look at the effect of the predictor on the outcome at the 16th, 50th and 84th percentile of the moderator. In this example, this would mean looking at the relationship between video games on aggression at low (16th percentile), typical (50th percentile) and high (84th percentile) levels of callous-unemotional traits. This default is fine, but the other option, which is commonly used, takes the mean of the moderator as a typical value, a standard deviation below the mean as the low value, and a standard deviation above the mean as the high value. Select ☑ Johnson-Neyman output to get a zone of significance for the moderator.

Clicking Multicategorical opens a dialog box where you can specify contrasts for the predictor (X) or moderator (W) if that variable is categorical with more than two categories. There is standard indicator or dummy coding (Section 12.3) as well as some others (including Helmert coding, which we look at in Section 12.5.4). Because our predictor and moderator are both continuous we have no use for this dialog box. PROCESS works only with variable names of eight characters or less so if you have

variable names longer than this it won't work. One solution is to select Long variable names and tick the tick-box underneath the foreboding warning. The warning is there because it's possible, if you're not careful, to end up with multiple variables truncated to the same name (SPSS Tip 11.1). To be safe, I'd use variable names of eight characters or less and leave this option unselected. Back in the main dialog box, click OK to run the analysis.

11.3.7 Output from moderation analysis Ⓑ

The output appears as text rather than nicely formatted in tables. Try not to let this formatting disturb you. If your output looks odd or contains warnings, or has a lot of zeros in it, it might be worth checking the variables that you inputted into PROCESS (SPSS Tip 11.1). Assuming everything has gone smoothly, you should see Output 11.1 which is the main moderation analysis. This output is pretty much the same as the table of coefficients that we saw in Chapter 9. We're told the $\hat{b}$-value for each predictor, the associated standard errors (which have been adjusted for heteroscedasticity because we asked for them to be). Each $\hat{b}$ is compared to zero using a t-test, which is computed from the beta divided by its standard error. The confidence interval for the $\hat{b}$ is also produced (because we asked for it). Moderation is shown up by a significant interaction effect, and that's what we've got here, $\hat{b} = 0.027$, 95% CI [0.013, 0.042], $t = 3.66$, $p = 0.0003$, indicating that the relationship between the time spent gaming and aggression is moderated by callous traits. Later in the output you'll find the robust confidence intervals (Output 11.2) based on a bootstrap, $\hat{b} = 0.027$, 95% CI [0.013, 0.042], which match the non-robust intervals and also do not contain zero, confirming that the interaction is likely to be a non-zero effect (if we assume this sample is one of the 95% that produces confidence intervals containing the population value).




```
**************************************************************************
Model  : 1
    Y  : aggress
    X  : vid_game
    W  : caunts

Sample
Size:  442
**************************************************************************
OUTCOME VARIABLE:
 aggress

Model Summary
          R       R-sq        MSE      F(HC4)        df1        df2          p
      .6142      .3773    99.5266    90.1077     3.0000   438.0000      .0000

Model
              coeff     se(HC4)          t          p       LLCI       ULCI
constant    39.9671      .4744    84.2543      .0000    39.0348    40.8994
vid_game      .1696      .0762     2.2255      .0266      .0198      .3194
caunts        .7601      .0467    16.2775      .0000      .6683      .8519
Int_1         .0271      .0074     3.6621      .0003      .0125      .0416

Product terms key:
 Int_1    :        vid_game x        caunts
```

Output 11.1

```
*********** BOOTSTRAP RESULTS FOR REGRESSION MODEL PARAMETERS ************

OUTCOME VARIABLE:
 aggress

              Coeff    BootMean      BootSE   BootLLCI   BootULCI
constant    39.9671    39.9825      .4699    39.0839    40.9030
vid_game      .1696      .1662      .0741      .0197      .3148
caunts        .7601      .7605      .0463      .6710      .8511
Int_1         .0271      .0273      .0072      .0134      .0417
```

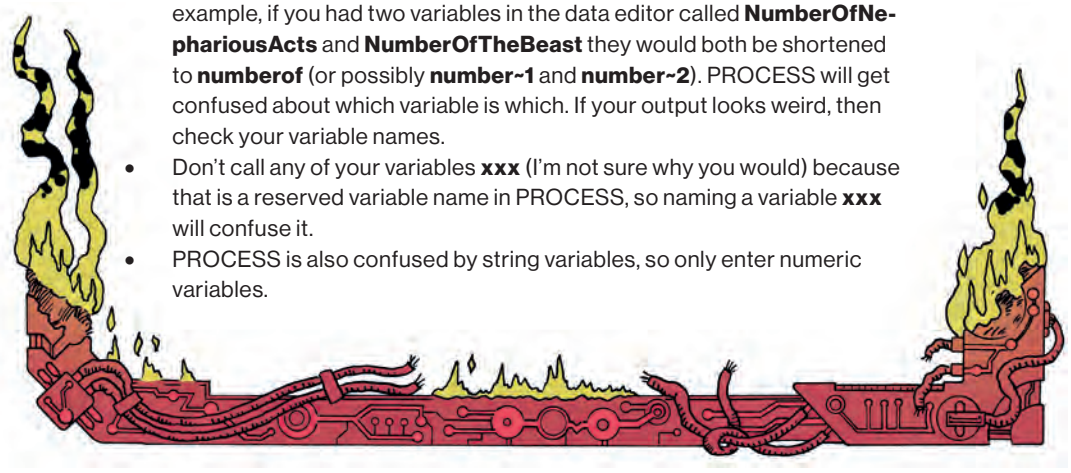
Output 11.2



SPSS Tip 11.1
Troubleshooting PROCESS Ⓑ

Knowing the following about PROCESS might help to prevent weird stuff happening:

- If the variable names entered into PROCESS are longer than eight characters, it blows up. Therefore, make sure your variable names are eight characters or less, or select for PROCESS to truncate them by clicking Long variable names (Figure 11.6). Be careful, though, because if you have variables with similar long names PROCESS will get confused. For example, if you had two variables in the data editor called **NumberOfNephariousActs** and **NumberOfTheBeast** they would both be shortened to **numberof** (or possibly **number~1** and **number~2**). PROCESS will get confused about which variable is which. If your output looks weird, then check your variable names.
- Don't call any of your variables **xxx** (I'm not sure why you would) because that is a reserved variable name in PROCESS, so naming a variable **xxx** will confuse it.
- PROCESS is also confused by string variables, so only enter numeric variables.



To interpret the moderation effect we examine the simple slopes, which are shown in Output 11.3. The output shows us the results of three models: the model for time spent gaming as a predictor of aggression (1) when callous traits are low (to be precise, when the value of callous traits is at the 16th percentile = –10.5950); (2) at the typical value of callous traits (because we centred callous traits the

50th percentile is close to 0, as indicated by the value of –0.5950); and (3) when the value of callous traits is high (i.e., when the value of callous traits is at the 84th percentile = 11.4050). We interpret these models as we would any other linear model by looking at the value of $\hat{b}$ (called *Effect* in the output), and its significance. We can interpret the three models as follows:

1. When callous traits are low, there is a non-significant negative relationship between time spent gaming and aggression, $\hat{b} = -0.117$, 95% CI [–0.336, 0.102], $t = -1.05$, $p = 0.295$.
2. At the typical value of callous traits, there is a significant positive relationship between time spent gaming and aggression, $\hat{b} = 0.154$, 95% CI [0.003, 0.304], $t = 2.01$, $p = 0.046$.
3. When callous traits are high, there is a significant positive relationship between time spent gaming and aggression, $\hat{b} = 0.478$, 95% CI [0.260, 0.697], $t = 4.30$, $p < 0.001$.

These results tell us that the relationship between time spent playing violent video games and aggression only really emerges in people with average or greater levels of callous-unemotional traits.

```
Conditional effects of the focal predictor at values of the moderator(s):

     caunts      Effect     se(HC4)           t           p        LLCI        ULCI
   -10.5950     -.1171      .1116     -1.0496      .2945     -.3364      .1022
     -.5950      .1535      .0765      2.0060      .0455      .0031      .3039
    11.4050      .4783      .1112      4.3006      .0000      .2597      .6968
```

Output 11.3

Output 11.4 shows the results of the Johnson–Neyman method. First we're told that the boundaries of the zone of significance are –17.2582 and –0.7036. These are the values of the centred version of the callous-unemotional traits variable, and define regions within which the relationship between the time spent gaming and aggression is significant. The table underneath gives a detailed breakdown of these regions. Essentially it's a finer-grained simple slopes analysis: it takes different values of callous-unemotional traits and for each one computes the $\hat{b}$ (*Effect*) and its significance for the relationship between the time spent gaming and aggression. I have annotated the output to show the boundaries of the zone of significance. Looking at the first row, we see the relationship between time spent gaming and aggression when the value of callous-unemotional traits is –18.595 (remember this variable has been centred so 0 is the mean value, and negative values are below the mean). At this low value of callous-unemotional traits, there is a significant negative relationship between time spent gaming and aggression, $\hat{b} = -0.334$, 95% CI [–0.648, –0.019], $t = -2.09$, $p = 0.038$. As we move down to the next row, we see the effect at a slightly larger value of callous traits (–17.2582). At this value, the relationship between time spent gaming and aggression is still significant ($p = 0.0500$). In the third row we again see the effect of gaming on aggression at a slightly larger value of callous traits (–16.3319) and at this value the relationship becomes non-significant ($p = 0.062$). Therefore, the threshold for significance ends at –17.2582 (which we were told at the top of the output). As we increase the value of callous-unemotional traits the relationship between time spent gaming and aggression remains non-significant until the value of callous-unemotional traits is –0.7036, at which point it crosses the threshold for significance again. For all subsequent increasing values of callous-unemotional traits the relationship between time spent gaming and aggression is significant. Looking at the $\hat{b}$-values themselves (in the column labelled *Effect*), we can also see that as callous-unemotional traits increase, the strength of relationship between time spent gaming and aggression goes from a small negative effect ($\hat{b} = -0.334$) to a strong positive one ($\hat{b} = 0.830$).

The final way to break down the interaction is to plot it. In Figure 11.6 we asked PROCESS to generate code for plotting and this syntax is in Output 11.5. This code is SPSS syntax. The first part (DATA LIST FREE/ …) creates three variables named after the variables in our original data file. The part

```
Moderator value(s) defining Johnson-Neyman significance region(s):
      Value    % below    % above
   -17.2582     1.3575    98.6425
    -.7036    48.8688    51.1312

Conditional effect of focal predictor at values of the moderator:
      caunts     Effect    se(HC4)          t          p       LLCI       ULCI
   -18.5950     -.3336     .1599    -2.0857     .0376    -.6479    -.0192
   -17.2582     -.2974     .1513    -1.9654     .0500    -.5948     .0000
   -16.3319     -.2724     .1455    -1.8724     .0618    -.5582     .0135
   -14.0687     -.2111     .1315    -1.6054     .1091    -.4695     .0473
   -11.8055     -.1499     .1183    -1.2672     .2058    -.3823     .0826
    -9.5424     -.0886     .1060     -.8359     .4037    -.2970     .1197
    -7.2792     -.0274     .0951     -.2877     .7737    -.2144     .1596
    -5.0161      .0339     .0862      .3932     .6944    -.1355     .2032
    -2.7529      .0951     .0797     1.1933     .2334    -.0616     .2518
    -.7036       .1506     .0766     1.9654     .0500     .0000     .3012
    -.4898       .1564     .0765     2.0452     .0414     .0061     .3066
    1.7734       .2176     .0768     2.8340     .0048     .0667     .3685
    4.0366       .2789     .0807     3.4572     .0006     .1203     .4374
    6.2997       .3401     .0876     3.8821     .0001     .1679     .5123
    8.5629       .4014     .0970     4.1387     .0000     .2108     .5919
   10.8260       .4626     .1081     4.2782     .0000     .2501     .6751
   13.0892       .5238     .1206     4.3444     .0000     .2869     .7608
   15.3523       .5851     .1340     4.3675     .0000     .3218     .8484
   17.6155       .6463     .1480     4.3663     .0000     .3554     .9373
   19.8787       .7076     .1626     4.3517     .0000     .3880    1.0272
   22.1418       .7688     .1775     4.3302     .0000     .4199    1.1178
   24.4050       .8301     .1928     4.3056     .0000     .4512    1.2090
```

Significant

Not significant

Significant

Output 11.4

```
DATA LIST FREE/
   vid_game   caunts     aggress.
BEGIN DATA.
    -6.9639   -10.5950    32.7294
      .1561   -10.5950    31.8956
     7.1561   -10.5950    31.0759
    -6.9639     -.5950    38.4457
      .1561     -.5950    39.5388
     7.1561     -.5950    40.6135
    -6.9639    11.4050    45.3053
      .1561    11.4050    48.7106
     7.1561    11.4050    52.0585
END DATA.
GRAPH/SCATTERPLOT=
 vid_game WITH     aggress  BY       caunts.
```

Output 11.5

between BEGIN DATA and END DATA are values of these variables. The first column contains the values of the 16th, 50th and 84th percentiles[7] of the variable **vid_game**. So, –6.9639, 0.1561 and 7.1561 are the 16th, 50th and 84th percentiles respectively. The second column contains the same percentiles for the variable **caunts**. That is, –10.5950, –0.5950 and 11.4050 are the 16th, 50th and 84th percentiles of **caunts**, respectively. The final column contains the predicted values of the outcome (aggression) for these combinations of the predictors. For example, when **vid_game** and **caunts** are both low (–6.9639 and –10.5950, respectively) the predicted value of aggression is 32.7294, when both variables are at the 50th percentile (0.1561 and –0.5950) the predicted value of aggression is 39.5388, and so on. The final part of the syntax (GRAPH/SCATTERPLOT …) creates a plot of these data.

7 It contains these percentiles because of the option we selected in Figure 11.6. If we'd selected the other option this column would contain the values of **vid_game** that are one standard deviation below the mean, the mean and one standard deviation above the mean.

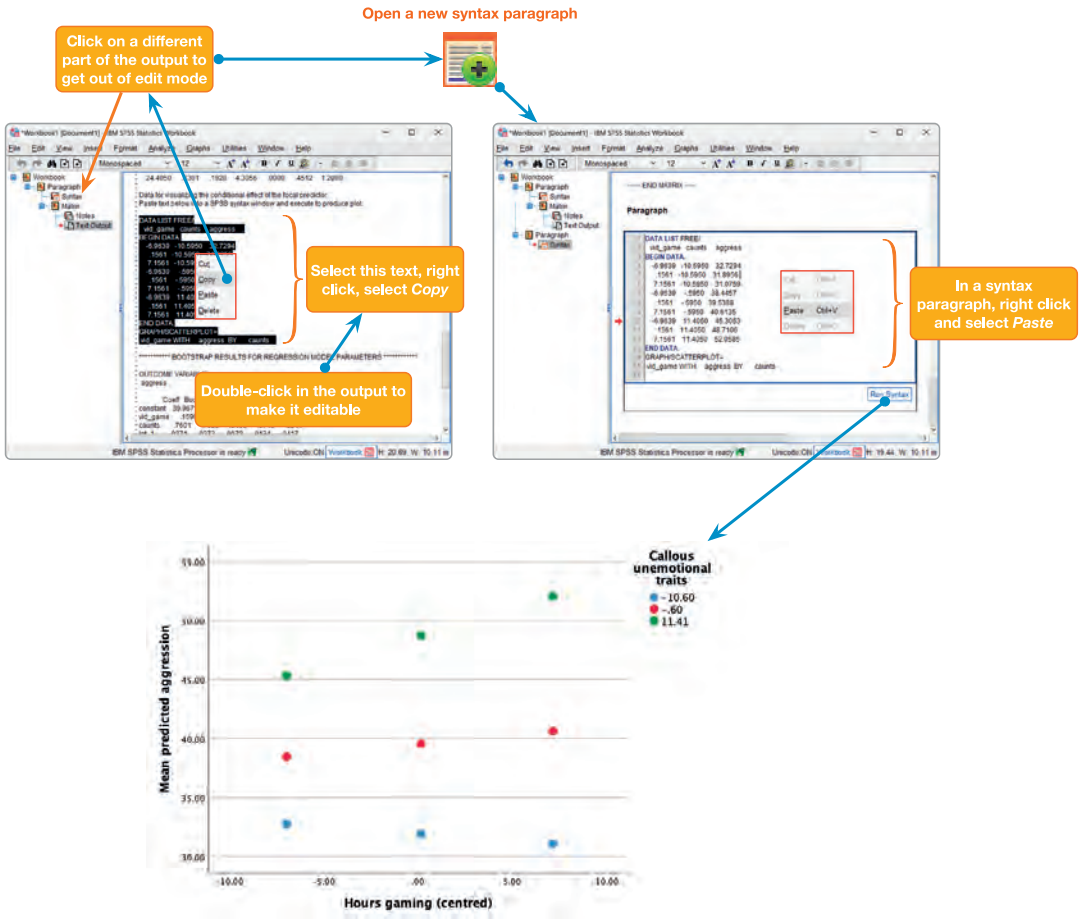


Figure 11.7 Plotting the simple slopes from the PROCESS-generated code

Figure 11.7 shows how to use this syntax to create a plot. First, double-click the output from PROCESS so it is editable. Note that in the navigation tree (the left part of the window) *Text output* will be highlighted. Starting at the top of the syntax, hold the left mouse button down and drag it over the text in Output 11.5 so it becomes highlighted. Then right-click with your mouse to bring up a menu in which you select *Copy*.[8] This action copies the highlighted syntax to memory. Click on a different part of the output in the navigation tree, which will take you out of edit mode and back to a regular workbook/viewer window. Ideally click on an item in the same branch as the PROCESS text output (e.g., *Notes* or *Syntax*). Assuming you're using workbook mode (Section 4.8.2), click  to create a new syntax paragraph.[9] In that new paragraph right-click your mouse and select paste from the menu that appears. This action will paste the syntax that you copied from the process output. Click  to generate the plot. The resulting plot shows what we found from the simple

8 On my computer, keyboard shortcuts for copy and paste were a bit temperamental within workbook files. This method always works!

9 This paragraph will be created directly below the currently active branch in the workbook, which is why in the previous step I suggested selecting something in the same branch as the PROCESS output, because by doing so the plot will be created directly below the main PROCESS output. If you're using classic mode, you need to create a new syntax file and paste the syntax into it and run it from there.

slopes analysis.[10] When callous traits are low (blue dots) there is a non-significant negative relationship between time spent gaming and aggression; at typical values of callous traits (red dots) there is small positive relationship between time spent gaming and aggression; and this relationship gets even stronger at high levels of callous traits (green dots). In short, there is no substantive relationship between video gaming and aggression at low levels of callous-unemotional traits, but a positive relationship emerges as levels of these traits increase.

The fact that we label one variable as a moderator doesn't change the fact that, statistically speaking, both variables are predictors (Jane Superbrain Box 11.1). We can talk of either predictor as moderating the relationship between the other predictor and the outcome. Therefore, it can be useful to plot the simple slopes with the 'predictor' and 'moderator' the other way around. To do this we need to edit this part of the syntax:

```
GRAPH/SCATTERPLOT=
 vid_game WITH aggress BY caunts.
```

To switch the variables **vid_game** and **caunts**. The resulting syntax is

```
GRAPH/SCATTERPLOT=
 caunts WITH aggress BY vid_game.
```

The resulting plot is in Figure 11.8 (I have added lines to highlight the trends). This shows that there is a strong positive relationship between callous-unemotional traits and aggression at low levels of video game use (blue dots) but that this relationship becomes stronger (the lines becomes steeper) at typical levels of video game use (red dots) and stronger still at high levels of video game use (green dots). In short, as video game use increases, the positive relationship between callous-unemotional traits and aggression gets stronger.

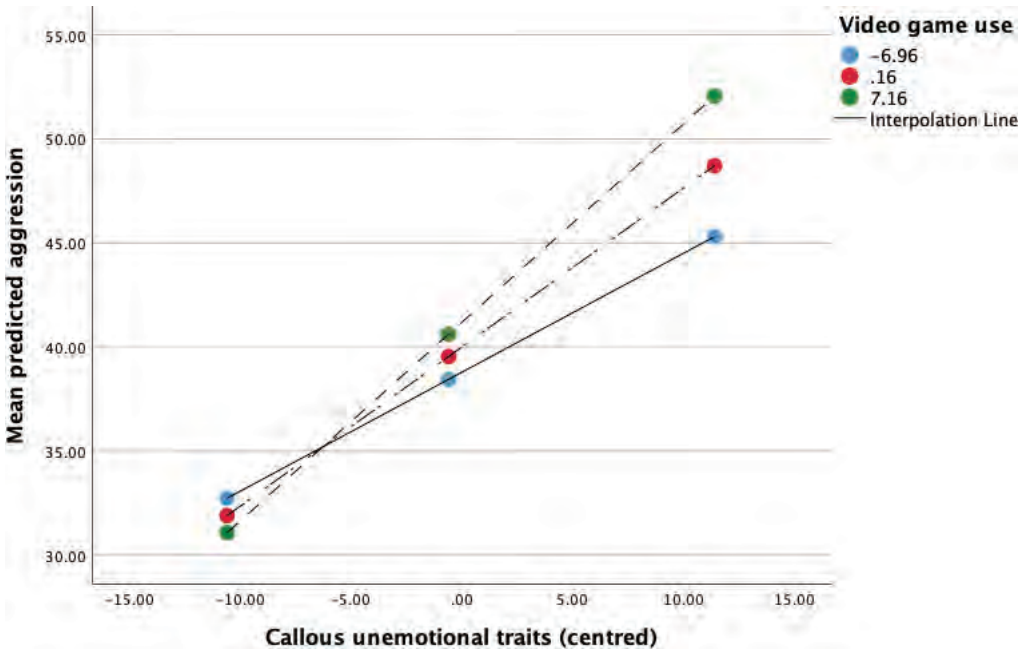


Figure 11.8 Simple slopes of predicted aggression from callous traits at three levels of video game use

10 I've changed the colours from the defaults to make it easier to describe in the text.

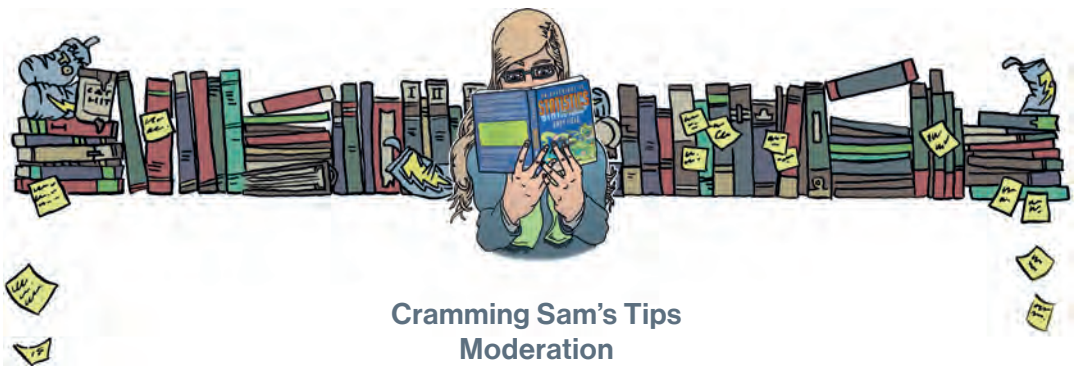
11.3.8 Reporting moderation analysis Ⓑ

Moderation can be reported in the same way as described in Section 9.13. My personal preference would be to produce a table like Table 11.1. You should state that you used the PROCESS tool (with version number) for SPSS and cite the relevant version of Andrew Hayes's book. At the time of writing we might write 'The moderation models were fitted using the PROCESS macro version 4.2 (Hayes, 022) for IBM SPSS Statistics 29 (IBM Corp., 2022)'.

Table 11.1 Linear model of predictors of aggression.

	$\hat{b}$	$SE\,\hat{b}$	t	p
Constant	39.97 [39.03, 40.90]	0.475	84.25	$p < 0.001$
Callous traits (centred)	0.76 [0.67, 0.85]	0.047	16.30	$p < 0.001$
Gaming (centred)	0.17 [0.02, 0.32]	0.076	2.23	$p = 0.027$
Callous traits × Gaming	0.027 [0.01, 0.04]	0.007	3.66	$p = 0.003$

Note. $R^2 = 0.38$



Cramming Sam's Tips
Moderation

- Moderation occurs when the relationship between two variables changes as a function of a third variable. For example, the relationship between watching horror films and feeling scared at bedtime might increase as a function of how vivid an imagination a person has.
- Moderation is tested using a linear model in which the outcome (fear at bedtime) is predicted from a predictor (how many horror films are watched), the moderator (imagination) and the interaction of the predictor variables.
- Predictors should be centred before the analysis.
- The interaction of two variables is their scores multiplied together.
- If the interaction is significant then the moderation effect is also significant.
- If moderation is found, follow up the analysis with simple slopes analysis, which looks at the relationship between the predictor and outcome at low, typical and high levels of the moderator.



11.4 Mediation Ⓑ

11.4.1 The conceptual model Ⓑ



Whereas moderation alludes to the combined effect of two variables on an outcome, **mediation** refers to a situation when the relationship between a predictor variable and an outcome variable can be explained by their relationship to a third variable (the **mediator**). The top of Figure 11.9 shows a relationship between a predictor and an outcome (called the *total effect*). The bottom of the figure shows that this total effect is broken down into a **direct effect** (denoted by c) and an **indirect effect** (which is made up of two paths labelled a and b). In the bottom diagram the predictor and outcome are related to a third variable (the mediator) in specific ways: (1) the predictor predicts the mediator through the path denoted by a; (2) the mediator predicts the outcome through the path denoted by b. The relationship between the predictor and outcome will probably be different when the mediator is also included in the model and so is denoted by c. The letters denoting each path (a, b and c) represent the unstandardized parameter estimate ($\hat{b}$) between the variables connected by the arrow; therefore, they symbolize the strength of relationship between variables. Mediation is said to have occurred if the strength of the relationship between the predictor and outcome (the total effect) is reduced by including the mediator (i.e., the $\hat{b}$-value for c is smaller than for the total effect). Perfect mediation occurs when c is zero: the relationship between the predictor and outcome is completely wiped out by including the mediator in the model.

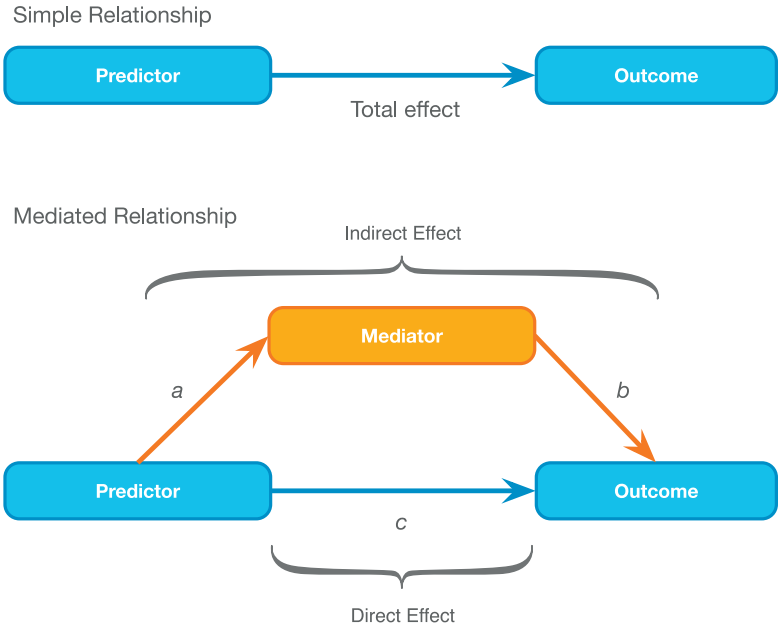


Figure 11.9 Diagram of a mediation model

This description is a bit abstract, so let's use an example. My wife and I often wonder what the important factors are in making a relationship last. For my part, I don't really understand why she'd want to be with a heavy rock fan with an oversized collection of vinyl and musical instruments and an unhealthy love of *Doctor Who* and numbers. It is important I gather as much information as possible

about keeping her happy because the odds are stacked against me. For her part, I have no idea why she wonders: her very existence makes me happy. Perhaps if you are in a relationship you have wondered how to make it last too.

During our cyber-travels, Mrs Field and I have discovered that physical attractiveness (McNulty et al., 2008), conscientiousness and neuroticism (good for us) predict marital satisfaction (Claxton et al., 2012). Pornography use probably doesn't: it is related to infidelity (Lambert et al., 2012). Mediation is really all about the variables that explain relationships like these: it's unlikely that everyone who catches a glimpse of some porn suddenly rushes out of their house to have an affair – presumably it leads to some kind of emotional or cognitive change that undermines the love glue that holds us and our partners together. Lambert et al. tested this hypothesis. Figure 11.10 shows their mediator model: the initial relationship is that between pornography consumption (the predictor) and infidelity (the outcome), and they hypothesized that this relationship is mediated by commitment (the mediator). This model suggests that the relationship between pornography consumption and infidelity isn't a direct effect but operates though a reduction in relationship commitment. For this hypothesis to be true: (1) pornography consumption must predict infidelity in the first place (the total effect is not zero); (2) pornography consumption must predict relationship commitment (path *a*); (3) relationship commitment must predict infidelity (path *b*); and (4) the relationship between pornography consumption and infidelity should be smaller when relationship commitment is included in the model than when it isn't (*c* is smaller than the total effect). We can partition the total effect of pornography consumption on infidelity into the *direct effect*, which is the relationship between them adjusting for relationship commitment, and the *indirect effect*, which is the effect of pornography consumption on infidelity through relationship commitment.

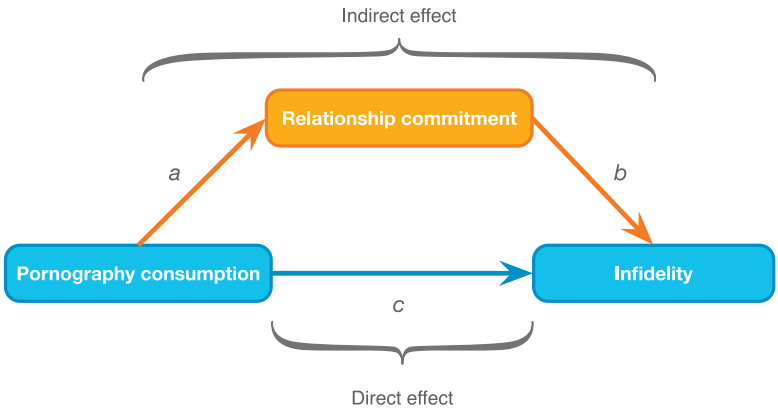


Figure 11.10 Diagram of a mediation model from Lambert et al. (2012)

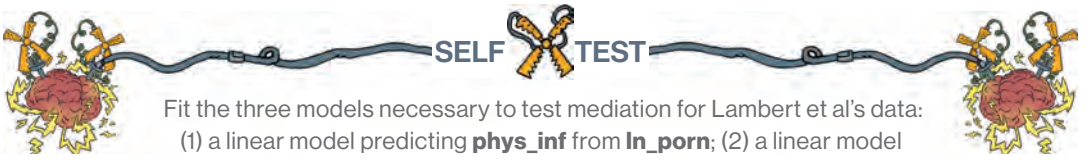
11.4.2 The statistical model Ⓑ

Unlike moderation, the statistical model for mediation is basically the same as the conceptual model: it is characterized in Figure 11.9. Historically, this model was tested through a series of linear models that reflect the four conditions necessary to demonstrate mediation (Baron & Kenny, 1986). I have mentioned already that the letters denoting the paths in Figure 11.9 represent the unstandardized $\hat{b}$-values for the relationships between variables denoted by the path. Therefore, to estimate any one of these paths, we want to know the unstandardized $\hat{b}$ for the two variables involved. For example, Baron and Kenny suggested that mediation is tested through three linear models (see also Judd & Kenny, 1981):

1 A linear model predicting the outcome from the predictor variable. The $\hat{b}$-value coefficient for the predictor gives us the value of the *total effect* in Figure 11.9.
2 A linear model predicting the mediator from the predictor variable. The $\hat{b}$-value for the predictor gives us the value of a in Figure 11.9.
3 A linear model predicting the outcome from both the predictor variable and the mediator. The $\hat{b}$-value for the predictor gives us the value of c in Figure 11.9, and the $\hat{b}$-value for the mediator gives us the value of b.

These models test the four conditions of mediation: (1) the predictor variable must significantly predict the outcome variable in model 1; (2) the predictor variable must significantly predict the mediator in model 2; (3) the mediator must significantly predict the outcome variable in model 3; and (4) the predictor variable must predict the outcome variable less strongly in model 3 than in model 1.

In Lambert et al.'s (2012) study, all participants had been in a relationship for at least a year. The researchers measured pornography consumption on a scale from 0 (low) to 8 (high) but this variable, as you might expect, was skewed (most people had low scores) so they analysed log-transformed values (**ln_porn**). They also measured commitment to their current relationship (**commit**) on a scale from 1 (low) to 5 (high). Infidelity was measured with questions asking whether the person had committed a physical act (**phys_inf**) that they or their partner would consider to be unfaithful (0 = no, 1 = one of them would consider it unfaithful, 2 = both of them would consider it unfaithful),[11] and also using the number of people they had 'hooked up' with in the previous year (**hook_ups**), which would mean during a time period in which they were in their current relationship.[12] The actual data from Lambert et al.'s study are in the file **lambert_2012.csv**.



Fit the three models necessary to test mediation for Lambert et al's data: (1) a linear model predicting **phys_inf** from **ln_porn**; (2) a linear model predicting **commit** from **ln_porn**; and (3) a linear model predicting **phys_inf** from both **ln_porn** and **commit**. Is there mediation?

I haven't included the outputs here because this isn't the way I want you to do mediation, but if you run the model you'll find that:

- Model 1 shows that pornography consumption significantly predicts infidelity, $\hat{b} = 0.59$, 95% CI [0.19, 0.98], $t = 2.93$, $p = 0.004$. As consumption increases, physical infidelity increases also.
- Model 2 shows that pornography consumption significantly predicts relationship commitment, $\hat{b} = -0.47$, 95% CI [–0.89, –0.05], $t = -2.21$, $p = 0.028$. As pornography consumption increases, commitment declines.
- Model 3 shows that relationship commitment significantly predicts infidelity, $\hat{b} = -0.27$, 95% CI [–0.39, –0.16], $t = -4.61$, $p < 0.001$. As relationship commitment increases, physical infidelity declines.
- The relationship between pornography consumption and infidelity is stronger in model 1, $\hat{b} = 0.59$, than in model 3, $\hat{b} = 0.46$.

11 I've coded this variable differently from the original data to make interpretation of it more intuitive, but it doesn't affect the results.

12 A 'hook-up' was defined to participants as 'when two people get together for a physical encounter and don't necessarily expect anything further (e.g., no plan or intention to do it again)'.

As such, the four conditions of mediation have been met. Many people still use this approach to test mediation. I think it is useful for illustrating the principles of mediation and for understanding what mediation means, but it has limitations. The main one is the fourth criterion: *the predictor variable must predict the outcome variable less strongly in model 3 than in model 1*. Although perfect mediation is shown when the relationship between the predictor and outcome is reduced to zero in model 3, in practice it rarely happens. Instead, you see a reduction in the relationship between the predictor and outcome, but it is not wiped out altogether. The question, therefore, arises of how much 'reduction' is sufficient to infer mediation. For example, for the models we have just run, for example, the relationship between the predictor and outcome falls from $\hat{b} = 0.59$ to $\hat{b} = 0.46$. Is that change meaningful?

Although Baron and Kenny advocated looking at the sizes of the $\hat{b}$-values, in practice people tend to look for a change in significance; so, mediation would occur if the relationship between the predictor and outcome was significant ($p \leq 0.05$) when looked at in isolation (model 1) but not significant ($p > 0.05$) when the mediator is included too (model 3). In the models we have just run, this doesn't happen because the relationship between porn consumption and physical infidelity is still significant when commitment is included as a predictor. This significance-based approach leads to all sorts of silliness because of the 'all-or-nothing' thinking that *p*-values encourage. You could have a situation in which the $\hat{b}$-value for the relationship between the predictor and outcome changes very little in models with and without the mediator, but the *p*-values shift from one side of the threshold to the other (e.g., from $p = 0.049$ when the mediator isn't included to $p = 0.051$ when it is). Even though the *p*-values have changed from significant to not significant, the change is very small, and the size of the relationship between the predictor and outcome will not have changed very much at all. Conversely, you could have a situation where the $\hat{b}$ for the relationship between the predictor and the outcome reduces a lot when the mediator is included but remains significant in both cases. For example, perhaps when looked at in isolation the relationship between the predictor and outcome is $b = 0.46$, $p < 0.001$, but when the mediator is included as a predictor as well it reduces to $b = 0.18$, $p = 0.042$. You'd conclude (based on significance) that mediation hadn't occurred despite the relationship between the predictor and outcome reducing to more than half its original value.

An alternative is to estimate the indirect effect and its significance. The indirect effect is illustrated in Figure 11.9 and Figure 11.10; it is the combined effects of paths *a* and *b*, and the significance of this test can be assessed using the **Sobel test** (Sobel, 1982) or (the approach we'll take) by including the combined effect as a predictor and testing whether the estimate is significantly different from zero. If the test of the indirect effect is significant it means that the predictor significantly affects the outcome variable via the mediator. In other words, there is significant mediation. Now that computers make it easy for us to estimate the indirect effect (i.e., the effect of mediation), its confidence interval and associated significance test, this practice is increasingly common and is preferable to Baron and Kenny's regressions and the Sobel test because it's harder to get sucked into the black-and-white thinking of significance testing. People have tended to apply Baron and Kenny's method in a way that is intrinsically bound to looking for 'significant' relationships, whereas estimating the indirect effect and its confidence interval allows us to simply report the degree of mediation observed in the data.

11.4.3 Effect sizes of mediation ©

If we're going to look at the size of the indirect effect to judge the degree of mediation, then it's useful to have effect size measures to help us (see Section 3.5). Many effect size measures have been proposed and are discussed in detail elsewhere (MacKinnon, 2008; Preacher & Kelley, 2011). The simplest is to look at the $\hat{b}$-value for the indirect effect and its confidence interval. Figure 11.9 shows us that the indirect effect is the combined effect of paths *a* and *b*. We have also seen that *a* and *b* are

unstandardized model coefficients for the relationships between variables denoted by the path. To find the combined effect of these paths, we multiply these $\hat{b}$-values:

$$\text{indirect effect} = ab \tag{11.6}$$

The resulting value is an unstandardized regression coefficient like any other, and consequently is expressed in the original units of measurement. As we have seen, it is sometimes useful to look at standardized $\hat{b}$-values, because these can be compared across different studies using different outcome measures (see Chapter 9). MacKinnon (2008) suggested standardizing this measure by dividing by the standard deviation of the outcome variable:

$$\text{indirect effect (partially standardized)} = \frac{ab}{s_{\text{outcome}}} \tag{11.7}$$

This standardizes the indirect effect with respect to the outcome variable, but not the predictor or mediator. As such, it is sometimes referred to as the partially standardized indirect effect. To fully standardize the indirect effect we would need to multiply the partially standardized measures by the standard deviation of the predictor variable (Preacher & Hayes, 2008):

$$\text{indirect effect (standardized)} = \frac{ab}{s_{\text{outcome}}} \times s_{\text{predictor}} \tag{11.8}$$

This measure is sometimes called the **index of mediation**. It is useful in that it can be compared across different mediation models that use different measures of the predictor, outcome and mediator. Reporting this measure would be particularly helpful if anyone decides to include your research in a meta-analysis.

11.4.4 Mediation using SPSS Ⓑ

We can test Lambert's mediation model (Figure 11.10) using the PROCESS tool. Access the dialog boxes in Figure 11.11 by selecting *Analyze* ❯ *Regression* ❯ *Process v4.2 by Andrew F. Hayes*. Drag (or click ▸) the outcome variable (**phys_inf**) from the box labelled *Variables* to the box labelled *Y Variable*, then drag the predictor variable (**ln_porn**) to the box labelled *X Variable*. Finally, drag the mediator variable (**commit**) to the box labelled *Mediator(s) M*, where you specify any mediators (you can have more than one). Simple mediation is represented by model 4, therefore, make sure that 4 is selected in the drop-down list under *Model number*. To test the indirect effects we will use bootstrapping to generate a confidence interval around the indirect effect, so select ☑ Bootstrap inference for model coefficients (we will use a 95% confidence interval but you can choose a different value from the list if you have a different alpha).

Clicking Options... activates the same dialog box that we used for moderation. The section labelled *Heteroscedasticity-consistent inference* allows us to select one of the methods for adjusting for heteroscedasticity in Section 6.10.7. As in our moderation example, I suggest selecting HC4. The other main useful options are to select ☑ Standardized effects to estimate the size of the indirect effect discussed in Section 11.4.3, and to select ☑ Show total effect model to print the direct effect of the predictor on the outcome (in this case the linear model of infidelity predicted from pornography consumption). As discussed in Section 11.3.6, if our predictor variable (X) or mediator (M) was categorical with more than two categories (which it's not), we could click Multicategorical to get PROCESS to automatically dummy code it for us, and we can click Long variable names cautiously if we have long variable names (see SPSS Tip 11.1). Back in the main dialog box, click OK to run the analysis.

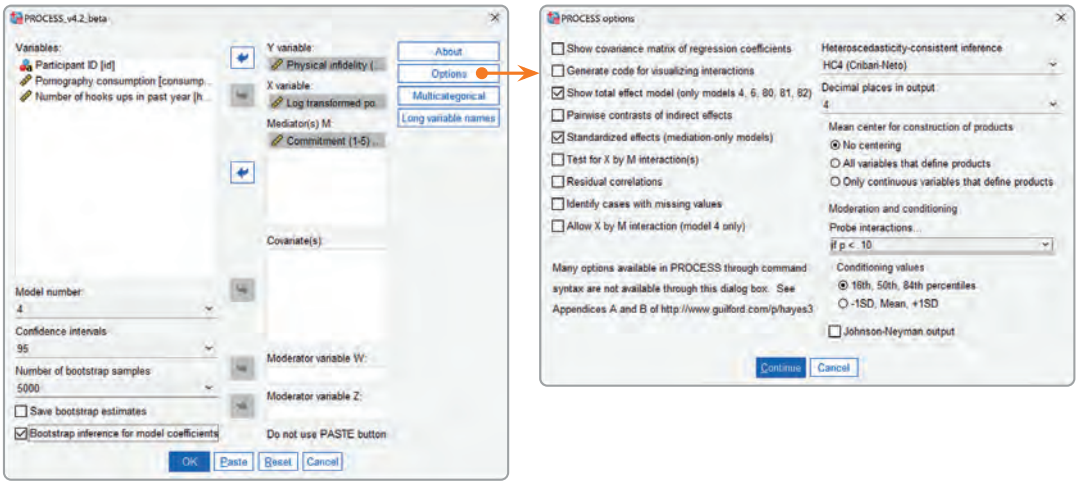
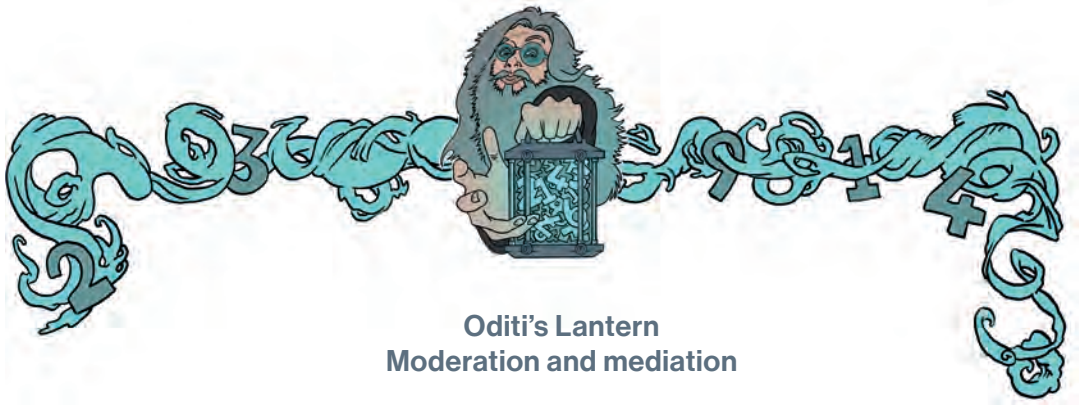


Figure 11.11 The dialog boxes for running mediation analysis



Oditi's Lantern
Moderation and mediation

'I, Oditi, want you to join my cult of undiscovered numerical truths. I also want you to stare into my lantern to gain statistical enlightenment. It's possible that statistical knowledge mediates the relationship between staring into my lantern and joining in my cult ... or it could be mediated by neurological changes to your brain created by the subliminal messages in the videos. Stare into my lantern to find out about mediation and moderation.'



11.4.5 Output from mediation analysis Ⓑ

As with moderation, the output appears as text. Output 11.6 shows the first part of the output, which initially tells us the name of the outcome (Y), the predictor (X) and the mediator (M) variables (SPSS Tip 11.1). This output is useful for double-checking we have entered the variables in the correct place:

the outcome is infidelity, the predictor is (log-transformed) pornography consumption, and the mediator is commitment. The next part of the output shows us the results of the linear model of commitment predicted from pornography consumption (i.e., path *a* in Figure 11.10). This output is interpreted just as we would interpret any linear model: pornography consumption significantly predicts relationship commitment, $\hat{b} = -0.47$, 95% CI [–0.94, –0.00], $t = -1.99$, $p = 0.048$. The R^2 values tells us that pornography consumption explains 2% of the variance in relationship commitment, and the fact that the *b* is negative tells us that as consumption increases, commitment declines (and vice versa). We also get the standardized effect of pornography consumption on predicts commitment, $\hat{\beta} = -0.14$, which isn't a particularly huge effect.

Output 11.7 shows the results of the regression of infidelity predicted from both pornography consumption and commitment. We can see that pornography consumption does not significantly predict infidelity with relationship commitment in the model, $\hat{b} = 0.46$, 95% CI [–0.05, 0.97], $t = 1.77$, $p = 0.08$; relationship commitment significantly predicts infidelity, $\hat{b} = -0.27$, 95% CI [–0.42, –0.12], $t = -3.63$, $p = 0.003$. The R^2 value tells us that the model explains 11.4% of the variance in infidelity. The negative $\hat{b}$ for commitment tells us that as commitment increases, infidelity declines (and vice versa), but the positive $\hat{b}$ for consumptions indicates that as pornography consumption increases, infidelity increases also. These relationships are in the predicted direction. The standardized betas show a small relationship between pornography consumption and infidelity, $\hat{\beta} = 0.15$ (note that even though the significance changes as a result of including commitment in the model, this effect size does not), and a larger relationship between commitment and infidelity of $\hat{\beta} = -0.29$.

```
**************************************************************************
Model  : 4
    Y  : phys_inf
    X  : ln_porn
    M  : commit

Sample
Size:  239

**************************************************************************
OUTCOME VARIABLE:
 commit

Model Summary
          R       R-sq        MSE      F(HC4)        df1        df2          p
      .1418      .0201      .5354      3.9438     1.0000   237.0000      .0482

Model
              coeff    se(HC4)          t          p       LLCI       ULCI
constant     4.2027      .0530    79.2857      .0000     4.0983     4.3071
ln_porn      -.4697      .2365    -1.9859      .0482     -.9356     -.0038

Standardized coefficients
            coeff
ln_porn     -.1418
```

Output 11.6

Output 11.8 shows the total effect of pornography consumption on infidelity (outcome). You will get this bit of the output only if you selected ☑ Show total effect model in Figure 11.11. The total effect is the effect of the predictor on the outcome when the mediator is not present in the model – in other words, path *c* in Figure 11.10. When relationship commitment is not in the model, pornography consumption significantly predicts infidelity, $\hat{b} = 0.58$, 95% CI [0.08, 1.08], $t = 2.30$, $p = 0.022$. The R^2

```
************************************************************************
OUTCOME VARIABLE:
 phys_inf

Model Summary
          R       R-sq        MSE      F(HC4)        df1        df2          p
      .3383      .1144      .4379      9.2618     2.0000   236.0000      .0001

Model
              coeff    se(HC4)          t          p       LLCI       ULCI
constant     1.3704      .3315     4.1337      .0000      .7173     2.0235
ln_porn       .4573      .2588     1.7673      .0785     -.0525      .9672
commit       -.2710      .0746    -3.6310      .0003     -.4180     -.1240

Standardized coefficients
              coeff
ln_porn       .1455
commit       -.2855
```

Output 11.7

value tells us that the model explains 3.46% of the variance in infidelity. As is the case when we include relationship commitment in the model, pornography consumption has a positive relationship with infidelity (as shown by the positive $\hat{b}$-value). The effect size is fairly small, $\hat{\beta} = 0.19$.

```
************************** TOTAL EFFECT MODEL ****************************
OUTCOME VARIABLE:
 phys_inf

Model Summary
          R       R-sq        MSE      F(HC4)        df1        df2          p
      .1859      .0346      .4754      5.3078     1.0000   237.0000      .0221

Model
              coeff    se(HC4)          t          p       LLCI       ULCI
constant      .2315      .0478     4.8434      .0000      .1374      .3257
ln_porn       .5846      .2538     2.3039      .0221      .0847     1.0845

Standardized coefficients
              coeff
ln_porn       .1859
```

Output 11.8

Output 11.9 is the most important because it displays the results for the indirect effect of pornography consumption on infidelity (i.e., the effect via relationship commitment). This is path *ab* in Figure 11.10. First, we're told the effect of pornography consumption on infidelity in isolation (the total effect), and these values replicate the model in Output 11.8. Next, we're told the effect of pornography consumption on infidelity when relationship commitment is included as a predictor as well (the direct effect). These values replicate those in Output 11.7. The first bit of new information is the *Indirect effect of X on Y*, which in this case is the indirect effect of pornography consumption on infidelity. We're given an estimate of this effect ($\hat{b} = 0.127$) as well as a bootstrapped standard error and confidence interval. As we have seen many times before, 95% confidence intervals contain the true value of a parameter in 95% of samples. People tend to assume that their sample isn't one of the 5% that do not contain the true value and use it to infer the population value of an effect. In this case,

assuming our sample is one of the 95% that 'hit' the true value, we know that the true $\hat{b}$-value for the indirect effect falls between 0.0023 and 0.274.[13] This range does not include zero, and remember that $b = 0$ would mean 'no effect whatsoever'; therefore, the fact that the confidence interval does not contain zero supports the idea that relationship commitment really does mediate the relationship between pornography consumption and infidelity. Output 11.9 also contains the effect size measure (the index of mediation from Section 11.4.3). The index of mediation is $\hat{\beta} = 0.041$, 95% BCa CI [0.007, 0.085], which is very small indeed. The final part of the output (Output 11.10) shows the bootstrap confidence intervals for the models we already discussed in Output 11.6 and Output 11.7.

```
************** TOTAL, DIRECT, AND INDIRECT EFFECTS OF X ON Y **************

Total effect of X on Y
      Effect     se(HC4)          t          p       LLCI       ULCI       c_cs
      .5846       .2538     2.3039      .0221      .0847     1.0845      .1859

Direct effect of X on Y
      Effect     se(HC4)          t          p       LLCI       ULCI      c'_cs
      .4573       .2588     1.7673      .0785     -.0525      .9672      .1455

Indirect effect(s) of X on Y:
            Effect     BootSE   BootLLCI   BootULCI
commit      .1273      .0693      .0023      .2741

Completely standardized indirect effect(s) of X on Y:
            Effect     BootSE   BootLLCI   BootULCI
commit      .0405      .0215      .0007      .0850
```

Output 11.9

```
*********** BOOTSTRAP RESULTS FOR REGRESSION MODEL PARAMETERS ************

OUTCOME VARIABLE:
 commit

               Coeff   BootMean     BootSE   BootLLCI   BootULCI
constant      4.2027     4.2017      .0536     4.0955     4.3058
ln_porn       -.4697     -.4611      .2305     -.9266     -.0090

----------

OUTCOME VARIABLE:
 phys_inf

               Coeff   BootMean     BootSE   BootLLCI   BootULCI
constant      1.3704     1.3696      .3150      .7262     1.9629
ln_porn        .4573      .4588      .2448     -.0121      .9282
commit        -.2710     -.2708      .0708     -.4060     -.1260
```

Output 11.10

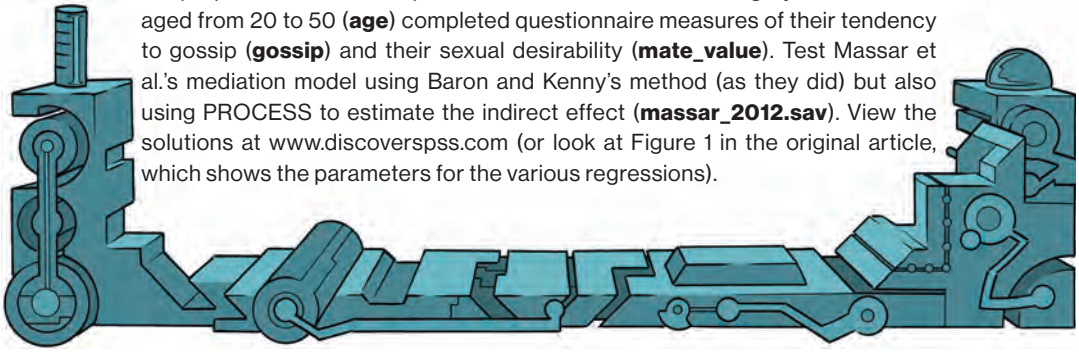
13 Remember that because of the nature of bootstrapping you will get slightly different values in your output.



Labcoat Leni's Real Research 11.1
I heard that Jane has a boil and kissed a tramp Ⓑ

Massar, K., et al. (2012). *Personality and Individual Differences*, *52*, 106–109.

Everyone likes a good gossip from time to time, but apparently it has an evolutionary function. One school of thought is that gossip is used as a way to derogate sexual competitors – especially by questioning their appearance and sexual behaviour. For example, if you've got your eyes on a guy, but he has his eyes on Jane, then a good strategy is to spread gossip that Jane has something unappealing about her or kissed someone unpopular. Apparently men rate gossiped-about women as less attractive, and they are more influenced by the gossip if it came from a woman with a high mate value (i.e., attractive and sexually desirable). Karlijn Massar and her colleagues hypothesized that if this theory is true then (1) younger women will gossip more because there is more mate competition at younger ages; and (2) this relationship will be mediated by the mate value of the person (because for those with high mate value gossiping for the purpose of sexual competition will be more effective). Eighty-three women aged from 20 to 50 (**age**) completed questionnaire measures of their tendency to gossip (**gossip**) and their sexual desirability (**mate_value**). Test Massar et al.'s mediation model using Baron and Kenny's method (as they did) but also using PROCESS to estimate the indirect effect (**massar_2012.sav**). View the solutions at www.discoverspss.com (or look at Figure 1 in the original article, which shows the parameters for the various regressions).



11.4.6 Reporting mediation analysis Ⓑ

Some people report only the indirect effect in mediation analysis. I have repeatedly favoured using bootstrap confidence intervals, so you should report these:

- ✓ The median model was fitted using the PROCESS macro version 4.2 (Hayes, 2022) for IBM SPSS Statistics 29 (IBM Corp., 2022). There was a significant indirect effect of pornography consumption on infidelity though relationship commitment, $\hat{b} = 0.13$, 95% BCa CI [0.02, 0.27]. This effect was tiny, $\hat{\beta} = 0.04$.

This is fine, but it can be quite useful to present a diagram of the mediation model, and indicate on it the regression coefficients, the indirect effect and its bootstrapped confidence intervals. For the current example, we might produce something like Figure 11.12.

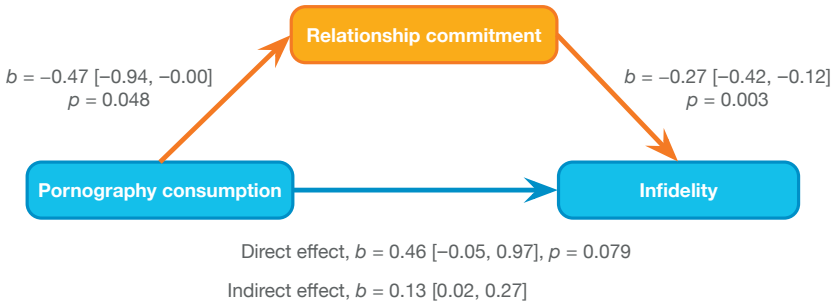


Figure 11.12 Model of pornography consumption as a predictor of infidelity, mediated by relationship commitment. The confidence interval for the indirect effect is a BCa bootstrapped CI based on 5000 samples



Cramming Sam's Tips Mediation

- Mediation is when the strength of the relationship between a predictor variable and outcome variable is reduced by including another variable as a predictor. Essentially, mediation equates to the relationship between two variables being 'explained' by a third. For example, the relationship between watching horror films and feeling scared at bedtime might be explained by scary images appearing in your head.
- Mediation is tested by assessing the size of the *indirect effect* and its confidence interval. If the confidence interval contains zero then we tend to assume that a genuine mediation effect doesn't exist. If the confidence interval doesn't contain zero, then we tend to conclude that mediation has occurred.



11.4.7 Mediation with two mediators ©

The basic mediation model can be extended to include two mediator variables. The rapid increase in 'fake news' and misinformation is a worrying trend in recent years. Perhaps more worrying is how widely some of this news is taken as fact. Researchers have started to look at what characteristics predict susceptibility to fake news. Bronstein et al. (2019) hypothesized that delusion-prone individuals may be more likely to believe fake news because of their tendency to engage in less analytic

and open-minded thinking. They conducted two online studies that got merged into a single analysis to test this hypothesis. We're going to replicate one of the analyses from the paper. The data are in **bronstein_2019.sav**.

The 947 participants completed several measures. We're going to focus on four of them. First, they measured the tendency to believe fake news (**fake_newz**). Participants viewed 12 fake news headlines, each with a brief description and photo, and rated their accuracy (1 = Not at all accurate, 4 = Very accurate). This variable is the average rating converted to a *z*-score. Participants also completed an inventory called the PDI, which uses statements such as 'Do you ever feel as if there is a conspiracy against you?' to gauge a person's propensity for delusion-like thinking. Again, scores were converted to *z*-scores (**delusionz**). Two measures of analytic thinking were used. Open-minded thinking was assessed with the Actively Open-Minded Thinking Scale, on which people endorse statements such as 'A person should always consider new possibilities' using a six-point scale (1 = strongly disagree, 6 = strongly agree). The total score was again converted to *z* (**thinkz_open**). Analytic thinking was assessed using the Cognitive Reflection Test, which uses several problems that have intuitive but incorrect responses. Participants must override their intuition to get the correct answer. Over seven items, higher scores (converted to *z*-scores again) indicate a greater tendency to use an analytic cognitive style (**thinkz_anal**). As a quick reminder, we'd start by plotting the data and looking at correlations between variables and so on.[14]

Figure 11.13 shows one of the mediation models tested in the paper (it corresponds to Figure 3 in the paper). The total effect of delusional thinking on belief in fake news is composed of the direct relationship between those variables, but also an indirect effect via open-minded thinking and analytic thinking. The model is essentially the same as Figure 11.9 except that another mediator has been added along with corresponding paths from the predictor (delusional thinking) and to the outcome (belief in fake news). As such we now have two path *a*s (labelled a_1 and a_2) and two path *b*s (labelled b_1 and b_2). The two mediators are connected by a double-headed arrow to remind us that they are not independent (i.e. they will correlate) and that we need to model the relationship between them.

We can test Bronstein et al. mediation model using the PROCESS tool. Access the dialog boxes in Figure 11.14 by selecting *Analyze* > *Regression* > *Process v4.2 by Andrew F. Hayes*. Drag (or click) the outcome variable (**fake_newz**) from the box labelled *Variables* to the box labelled *Y Variable*, then

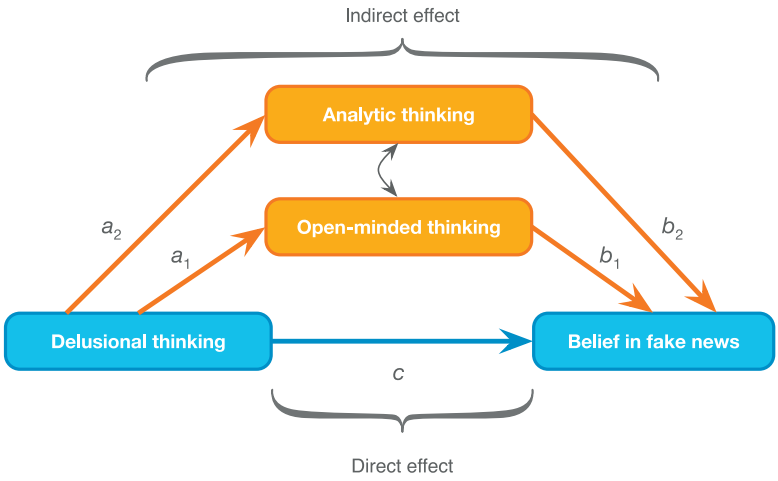


Figure 11.13 A mediation model with two mediators

14 If you do this and compare your results to Table 2 in the paper there are, according to the authors, some transcription errors in that table. In the supplementary materials they produce correlation tables for the two studies separated which will match your output.

drag the predictor variable (**delusionz**) to the box labelled *X Variable*. Finally, drag the two mediator variables (**thinkz_open** and **thinkz_anal**) to the box labelled *Mediator(s) M*. Remember, you can select both variables simultaneously by pressing Ctrl (on Windows) or Cmd (on MacOS) as you click on the variables. As with the previous example, select 4 in the drop-down list under *Model number*. The sample size is massive, so there's no value in requesting bootstrap inference. Click Options... and, as in the previous example, select (1) HC4 in the section labelled *Heteroscedasticity-consistent inference* select and (2) ☑ Show total effect model. All of the variables in the data set are standardized (they are *z*-scores) so the resulting parameter estimates will be the standardized parameter estimates, meaning that there is no point in selecting ☑ Standardized effects (because the standardized and raw effects will be the same). All our variable names have more than eight characters so we need to check that when they are cropped they are unique (see SPSS Tip 11.1). When shortened to eight characters our variables will become **fake_new**, **delusion, thinkz_o** and **thinkz_a**, which are all unique, so we can proceed by clicking Long variable names and accepting the risk. If we'd had two variables that had the same name when shortened we would need to rename one of them in the data editor to give it a unique name before proceeding. Back in the main dialog box, click OK to run the analysis.

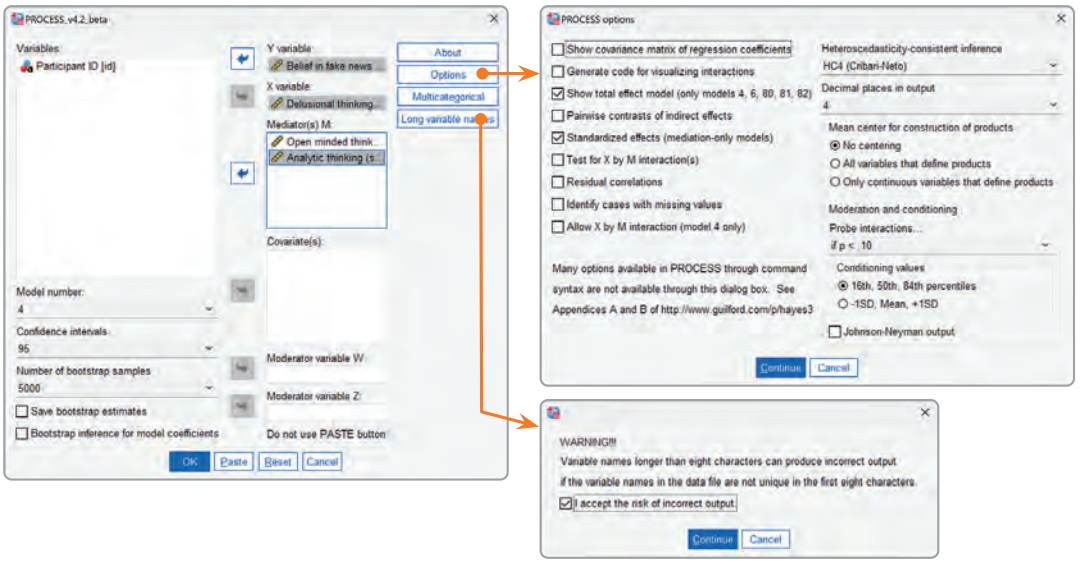


Figure 11.14 The dialog boxes for running mediation analysis with two mediators

We can summarize the output as follows. From Output 11.11 we can conclude that:

- Delusional thinking has a significant negative relationship with open-minded thinking, $\hat{b} = -0.30$, 95% CI [–0.36, –0.24], $t = -9.69$, $p < 0.001$. As delusional thinking increases, open-minded thinking decreases.
- Delusional thinking has a significant negative relationship with analytic thinking, $\hat{b} = -0.27$, 95% CI [–0.33, –0.20], $t = -8.42$, $p < 0.001$. As delusional thinking increases, analytic thinking decreases.

From Output 11.12 we can see that:

- Open-minded thinking has a significant negative relationship with belief in fake news when analytic thinking is fixed, $\hat{b} = -0.17$, 95% CI [–0.24, –0.11], $t = -5.17$, $p < 0.001$. As open-minded thinking increases, belief in fake news decreases.

- Analytic thinking has a significant negative relationship with belief in fake news when open-minded thinking is fixed, $\hat{b} = -0.14$, 95% CI [–0.21, –0.07], $t = -4.09$, $p < 0.001$. As analytic thinking increases, belief in fake news decreases.

From Output 11.13 we can see that:

- From the *Total effect of X on Y*, delusional thinking has a significant positive relationship with belief in fake news, $\hat{b} = 0.24$, 95% CI [0.17, 0.31], $t = 6.77$, $p < 0.001$. As delusional thinking increases, belief in fake news increases.
- From the *Direct effect of X on Y*, with open-minded and analytic thinking in the model delusional thinking still has a significant (but smaller than when the other predictors are excluded) positive relationship with belief in fake news, $\hat{b} = 0.15$, 95% CI [0.08, 0.22], $t = 4.20$, $p < 0.001$. As delusional thinking increases, belief in fake news increases.
- The indirect effect of delusional thinking on belief in fake news via open-minded thinking was $\hat{b} = 0.05$, 95% CI [0.03, 0.07]. The confidence interval does not contain zero suggesting this indirect pathway is significant.
- The indirect effect of delusional thinking on belief in fake news via analytic thinking was $\hat{b} = 0.04$, 95% CI [0.02, 0.06]. The confidence interval does not contain zero suggesting this indirect pathway is significant.

The model is visualized in Figure 11.15 and you can hopefully map the values back to the summary of the output.[15] The key thing is that the initial relationship between delusional thinking and belief in fake news ($\hat{b} = 0.24$) is diminished to $\hat{b} = 0.15$ when the mediators are included, and both indirect pathways/effects are significant ($p \leq 0.05$), albeit small effects, suggesting that part of the relationship

```
************************************************************************
OUTCOME VARIABLE:
 thinkz_o

Model Summary
          R       R-sq        MSE      F(HC4)        df1        df2          p
      .3032      .0920      .9090     93.9215     1.0000   945.0000      .0000

Model
               coeff     se(HC4)          t          p       LLCI       ULCI
constant       .0000      .0310      .0000     1.0000     -.0608      .0608
delusion      -.3032      .0313    -9.6913      .0000     -.3646     -.2418

************************************************************************
OUTCOME VARIABLE:
 thinkz_a

Model Summary
          R       R-sq        MSE      F(HC4)        df1        df2          p
      .2662      .0709      .9301     70.9367     1.0000   945.0000      .0000

Model
               coeff     se(HC4)          t          p       LLCI       ULCI
constant       .0000      .0314      .0000     1.0000     -.0615      .0615
delusion      -.2662      .0316    -8.4224      .0000     -.3283     -.2042
```

Output 11.11

15 This figure matches Figure 3 in the original paper except that I have included confidence intervals, and the statistics for the indirect effects.

```
**************************************************************************
OUTCOME VARIABLE:
 fake_new

Model Summary
          R       R-sq        MSE      F(HC4)        df1        df2          p
      .3397       .1154      .8874     34.5794     3.0000   943.0000      .0000

Model
              coeff    se(HC4)          t          p       LLCI       ULCI
constant      .0000      .0307     .0000     1.0000    -.0602     .0602
delusion      .1529      .0364    4.2031     .0000     .0815      .2243
thinkz_o     -.1709      .0331   -5.1693     .0000    -.2358     -.1060
thinkz_a     -.1410      .0345   -4.0894     .0000    -.2087     -.0733
```

Output 11.12

```
************** TOTAL, DIRECT, AND INDIRECT EFFECTS OF X ON Y **************

Total effect of X on Y
      Effect    se(HC4)          t          p       LLCI       ULCI       c_cs
      .2423      .0358     6.7683     .0000     .1720      .3125      .2423

Direct effect of X on Y
      Effect    se(HC4)          t          p       LLCI       ULCI      c'_cs
      .1529      .0364     4.2031     .0000     .0815      .2243      .1529

Indirect effect(s) of X on Y:
              Effect     BootSE   BootLLCI   BootULCI
TOTAL         .0894      .0142      .0623      .1182
thinkz_o      .0518      .0108      .0310      .0734
thinkz_a      .0375      .0100      .0191      .0579
```

Output 11.13

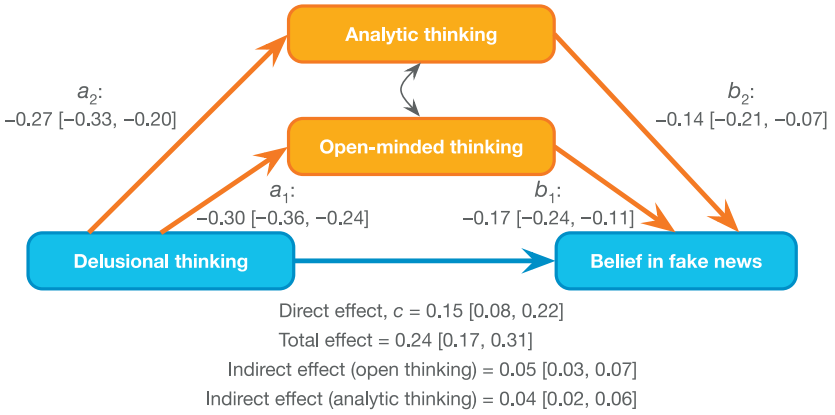


Figure 11.15 Multiple mediation model of belief in fake news. All paths are significant $p < 0.001$.

between delusional thinking and belief in fake news operates through lower levels of analytic and open-minded thinking. We know the effect is through *lower* levels of analytic and open-minded thinking because delusional thinking is *negatively* related to these variables (the higher the delusion score, the lower the scores on these forms of thinking), and these types of thinking are *negatively* related to belief in fake news (the lower your score on analytic and open-minded thinking, the higher your score on believing fake news).

11.5 Jane and Brian's story Ⓐ

Jane didn't like the library: it was full of people, and people made her uneasy. Why had she gone there to find campus guy? It made no sense to her, but as days passed without seeing him, she missed his nervy recitals of his statistics lectures. She'd watched from afar as he left his lectures and scuttled to the library. Was he avoiding her? The thought that he might be made her want to see him. She wanted to make amends for her barbed comments and yet when she found him, and he spat statistical theory at her like one of Pavlov's dogs, she cut him down again. Why did she do that? Why wouldn't she give him a millimetre? Had she spent too much time alone in the basement of the Pleiades building? Jane had realized as she'd looked at his deflated face in the library that he would reach a limit with her quips. She'd wanted to keep him at arm's length, but it was only now that she realized how long her emotional arms could be. If she wanted their chats to continue, she needed to give him some hope.

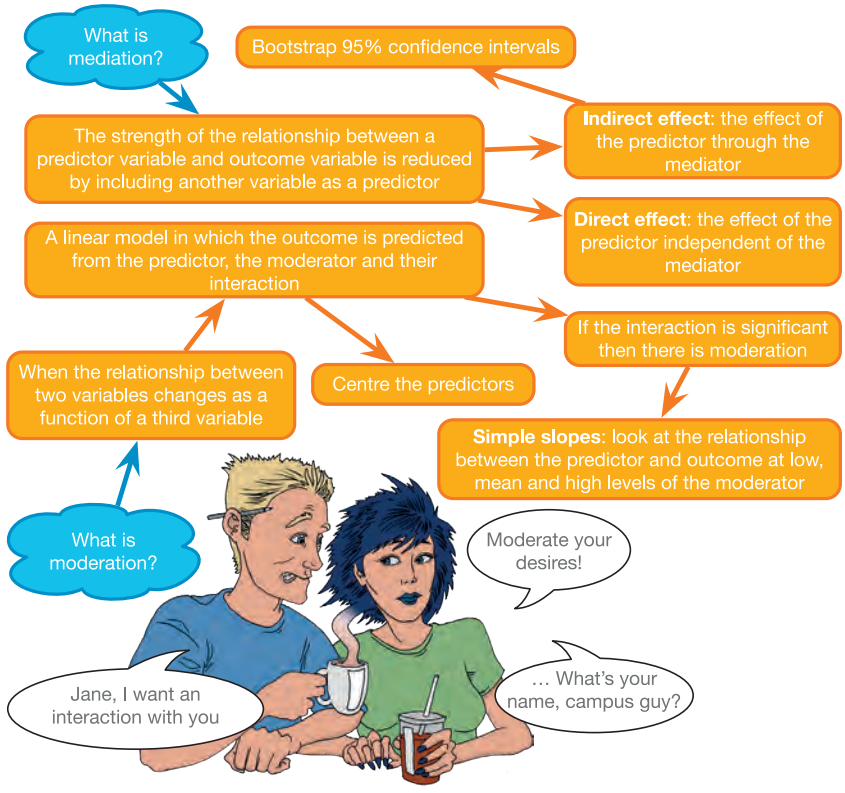


Figure 11.16 What Brian learnt from this chapter

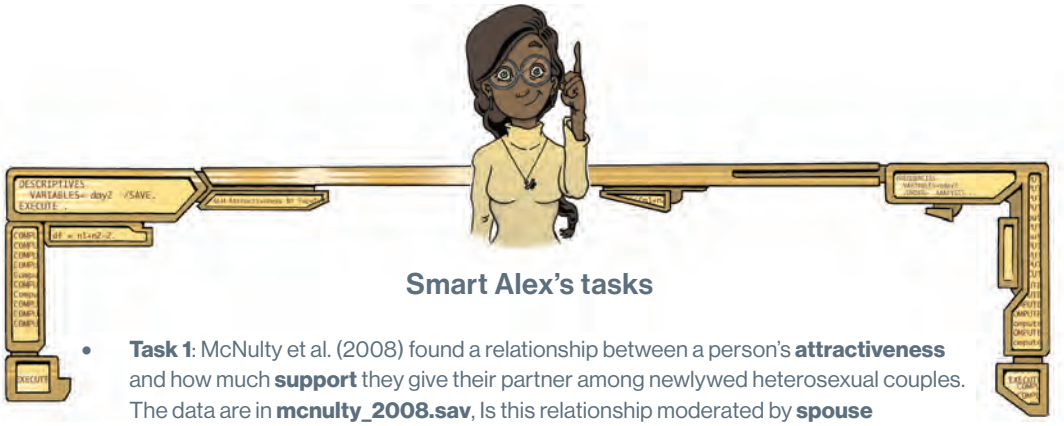
11.6 What next? Ⓐ

We started this chapter by looking at my relative failures as a human being compared to Simon Hudson. I then bleated on excitedly about moderation and mediation, which could explain why Clair Sparks chose Simon Hudson all those years ago. Perhaps she could see the writing on the wall. I was true to my word to my parents though: I *was* philosophical about it. I set my sights on a girl called Zoë during the obligatory lunchtime game of kiss chase (not the same Zoë whom I ended up marrying). I don't think she was all that keen, which was just as well because I was about to be dragged out of her life for ever. Not that I believe in these things, but if I did I would have believed that the scaly,

warty, green, long-nailed hand of fate (I have always assumed that anything named as ominously as the hand of fate would need to look monstrous) had decided that I was far too young to be getting distracted by girls. Waggling its finger at me, it plucked me out of primary school and cast me into the flaming pit of hell, otherwise known as all-boys' grammar school. It's fair to say that my lunchtime primary school kiss chase games were the last I would see of girls for quite some time …

11.7 Key terms that I've discovered

Direct effect
Grand mean centring
Index of mediation
Indirect effect
Interaction effect
Johnson–Neyman interval
Mediation
Mediator
Moderation
Moderator
Simple slopes analysis
Sobel test



Smart Alex's tasks

- **Task 1**: McNulty et al. (2008) found a relationship between a person's **attractiveness** and how much **support** they give their partner among newlywed heterosexual couples. The data are in **mcnulty_2008.sav**, Is this relationship moderated by **spouse** (i.e., whether the data were from the husband or wife)?[16] Ⓑ
- **Task 2**: Produce the simple slopes analysis for Task 1. Ⓑ
- **Task 3**: McNulty et al. (2008) also found a relationship between a person's **attractiveness** and their relationship **satisfaction** among newlyweds. Using the same data as in Tasks 1 and 2, find out if this relationship is moderated by **spouse**. Ⓑ
- **Task 4**: In this chapter we tested a mediation model of infidelity for Lambert et al.'s data. Repeat this analysis but using **hook_ups** as the measure of infidelity. Ⓑ
- **Task 5:** Tablets like the iPad are very popular. A company owner was interested in how to make his brand of tablets more desirable. He collected data on how cool people perceived a product's advertising to be (**advert_cool**), how cool they thought the product was (**product_cool**) and how desirable they found the product (**desirability**). Test his theory that the relationship between cool advertising and product desirability is mediated by how cool people think the product is (**tablets.sav**). Am I showing my age by using the word 'cool'? Ⓒ

View the solutions at www.discoverspss.com

16 These are not the actual data from the study, but are simulated to mimic the findings in Table 1 of the original paper.



GLM 1: COMPARING SEVERAL INDEPENDENT MEANS

12.1 What will this chapter tell me? 558
12.2 A puppy-tastic example 559
12.3 Compare several means with the linear model 560
12.4 Assumptions when comparing means 574
12.5 Planned contrasts (contrast coding) 577
12.6 *Post hoc* procedures 590
12.7 Effect sizes when comparing means 592
12.8 Comparing several means using SPSS 592
12.9 Output from one-way independent ANOVA 599
12.10 Robust comparisons of several means 607
12.11 Bayesian comparison of several means 609
12.12 Reporting results from one-way independent ANOVA 610
12.13 Jane and Brian's story 610
12.14 What next? 611
12.15 Key terms that I've discovered 612
Smart Alex's tasks 612

Letters A to D indicate increasing difficulty

12.1 What will this chapter tell me? Ⓐ

There are pivotal moments in everyone's life, and one of mine was at the age of 11. Where I grew up in England there were three choices when leaving primary school and moving on to secondary school: (1) state school (where most people go); (2) grammar school (where clever people who pass an exam called the Eleven Plus go); and (3) private school (where rich people go). My parents were not rich and I am not clever and consequently I failed my Eleven Plus, so private school and grammar school (where my older brother, 'the clever one', had gone) were out. There was no choice but for me to join my friends at the local state school. I could not have been happier: at age 11 I hadn't yet become a stats hermit and liked having friends. Imagine everyone's shock when my parents received a letter saying that some extra spaces had become available at the grammar school; although the local authority could scarcely believe it and had checked the Eleven Plus papers several million times to confirm their findings, I was next on their list. I could not have been unhappier. So, I waved goodbye to all my friends and trundled off to join my brother at Ilford County High School for Boys (a school at which the head teacher hit students with a cane if they were particularly bad and that, for some considerable time and with good reason, had 'H.M. Prison' painted in huge white letters on its roof). It was goodbye to normality, and hello to seven years of learning how not to function in society. I often wonder how my life would have turned out had I not gone to this school. In the parallel universes where the letter didn't arrive and the parallel Andy went to state school, or where my parents were rich and he went to private school, what became of him? If we wanted to compare these three situations we couldn't use the methods in Chapter 10 because there are more than two conditions.[1] However, this chapter tells us all about the statistical models that we use to analyse situations in which we want to compare more than two independent means. The model is typically called **analysis of variance** (or **ANOVA** to its friends) but, as we shall see, it is just a variant on the linear model. So, in effect, we're going to learn stuff that we already know from previous chapters. Hopefully that's reassuring.



Figure 12.1 My brother Paul (left) and I (right) in our very fetching school uniforms (note the fear in my face)

1 This is the least of our problems – there's also the small issue of reinventing physics to access the parallel universes.

12.2 A puppy-tastic example Ⓐ

Let's start with an example. You're about halfway through the book now, and there are a lot of equations in this chapter, so we probably need some puppy therapy. Puppy therapy is a form of animal-assisted therapy, in which puppy contact is introduced into the therapeutic process. Puppy rooms have been set up to destress students and staff at my own university (Sussex) in the UK along with universities in Bristol, Nottingham, Aberdeen and Lancaster. I've heard of similar things at Dalhousie and Simon Fraser in Canada and Tufts and Caldwell in the USA. My own contribution is to sometimes bring my adorable spaniel, Milton (Figure 12.2), into work to sit in my office and look cute at any students who break down in tears about some stats issue they have.[2] Similarly he may pop up at strategic points in this chapter to help you.

Despite this increase in puppies on campuses (which can only be a good thing) to reduce stress, the evidence base is pretty mixed. A review of animal-assisted therapy in childhood mental health found that of 24 studies, 8 found positive effects of animal-assisted therapy, 10 showed mixed findings and 6 concluded that there was no effect (Hoagwood et al., 2017). Imagine we wanted to contribute to this literature by running a study in which we randomized people into three groups (**dose**):



Figure 12.2 Some puppy therapy for you in the form of my dog, Milton

2 On the off chance you also have the previous edition of this book, I should reluctantly explain why the seemingly identical spaniel in that book is called Ramsey, not Milton. Ramsey was killed in a road traffic accident in 2017 (shortly after the book came out). The following week I had to give two lectures that contained numerous references and photos of him. Thankfully, after 4 days of repeatedly and spontaneously bursting into tears every time I thought about him, I managed to hold my emotions together during the lectures, but it was a close call. Next time you think your lecturer is giving a poor lecture, spare a thought for what might be going on in their life that you don't know about.

(1) a control group, which could be a treatment as usual, a no treatment (no puppies) or ideally some kind of placebo group (we could give people in this group a cat disguised as a dog); (2) 15 minutes of puppy therapy (a low-dose group); and (3) 30 minutes of puppy contact (a high-dose group). The dependent variable was a measure of **happiness** ranging from 0 (as unhappy as I can possibly imagine) to 10 (as happy as I can possibly imagine). The design of this study mimics a very simple randomized controlled trial (as used in pharmacological, medical and psychological intervention trials) because people are randomized into a control group or groups containing the active intervention (in this case puppies, but in other cases a drug or a surgical procedure). We'd predict that any form of puppy therapy should be better than the control (i.e. higher happiness scores) but also formulate a dose-response hypothesis that as exposure time increases (from 0 minutes to 15 and 30) happiness will increase too. The data are in Table 12.1 (first three columns) and the file **puppies.sav**.

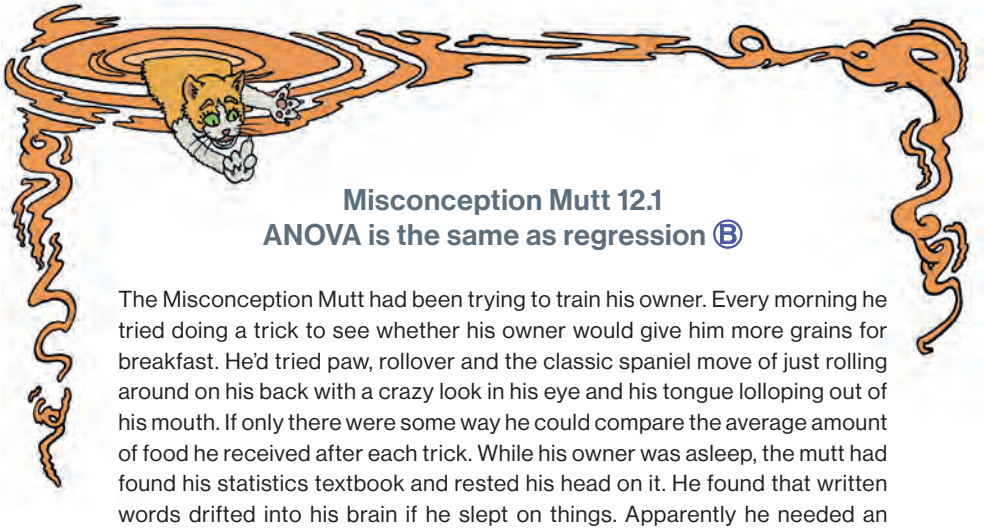
Table 12.1 Data in **puppies.sav** and some summary information

ID	Dose	Happiness	Group mean	Group variance	Group SD
25hto3	No puppies	3	2.20	1.70	1.30
121118	No puppies	2			
t54p42	No puppies	1			
s6u853	No puppies	1			
tcs14p	No puppies	4			
oum4t7	15 mins	5	3.20	1.70	1.30
kfl7lq	15 mins	2			
2gi51b	15 mins	4			
d3j771	15 mins	2			
eu23ns	15 mins	3			
b343ey	30 mins	7	5.00	2.50	1.58
5nvg7h	30 mins	4			
5ta11l	30 mins	5			
82e7va	30 mins	3			
	Grand mean	3.467			
	Grand variance	3.124			
	Grand SD	1.767			

12.3 Comparing several means with the linear model Ⓑ

Having loaded the data, we can think about the model we want to fit. It's not uncommon for researchers to think, and for people to be taught, that you compare means with 'ANOVA' and that this is somehow different from 'regression' (i.e. the linear model), which you apply to look for relationships between variables. This artificial division is unhelpful (in my view) and exists largely for weird historical reasons (Misconception Mutt 12.1). The 'ANOVA' to which some people allude is simply the *F*-statistic that we encountered as a test of the fit of a linear model, it's just that sometimes the linear model consists of group means (as we will see). Having said that, there is a different way of teaching the use of the *F*-statistic to compare means known as the variance-ratio method. This approach is fine for simple designs, but becomes impossibly cumbersome in more complex situations such as analysis

of covariance, or when you have unequal sample sizes.[3] The linear model framework has various advantages. First, we're building on material that you have already learnt (this chapter is a natural progression from the bits of the book that you've hopefully already read). Second, the linear model extends very logically to the more complex situations (e.g., multiple predictors, unequal group sizes) without the need to get bogged down in mathematics.



Misconception Mutt 12.1
ANOVA is the same as regression Ⓑ

The Misconception Mutt had been trying to train his owner. Every morning he tried doing a trick to see whether his owner would give him more grains for breakfast. He'd tried paw, rollover and the classic spaniel move of just rolling around on his back with a crazy look in his eye and his tongue lolloping out of his mouth. If only there were some way he could compare the average amount of food he received after each trick. While his owner was asleep, the mutt had found his statistics textbook and rested his head on it. He found that written words drifted into his brain if he slept on things. Apparently he needed an ANOVA.

The next morning as he munched his grains, he muttered to himself about using an ANOVA to compare means. His dog flap clanked shut, making him jump on his heels to turn. There was nothing there. As he turned back to his bowl he noticed a large ginger cat helping himself to his food. The mutt growled.

'You know,' said the cat, 'ANOVA is the same thing as the linear model or regression.'

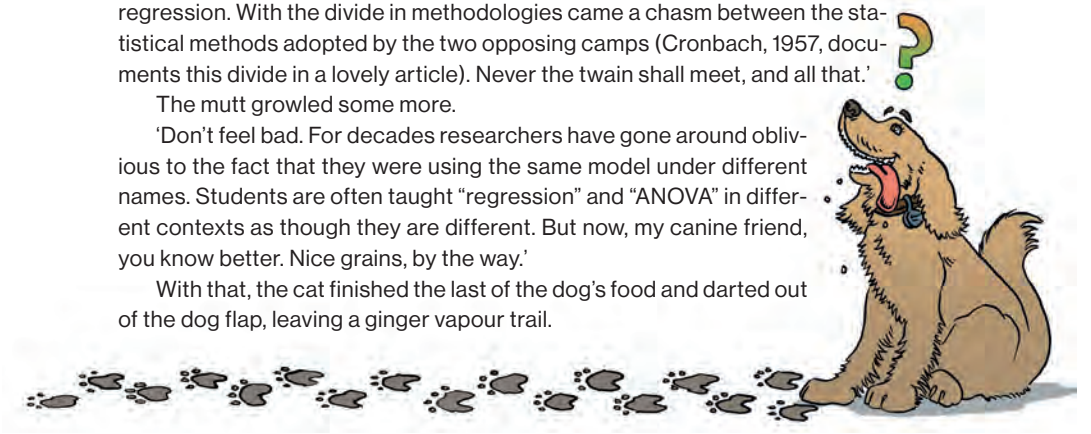
The growling continued.

'The reason why some people think of ANOVA and regression as separate statistical models is historical. Two distinct branches of methodology, correlational and experimental, developed in the social sciences. Researchers interested in controlled experiments adopted ANOVA as their model of choice, whereas those looking for real-world relationships adopted multiple regression. With the divide in methodologies came a chasm between the statistical methods adopted by the two opposing camps (Cronbach, 1957, documents this divide in a lovely article). Never the twain shall meet, and all that.'

The mutt growled some more.

'Don't feel bad. For decades researchers have gone around oblivious to the fact that they were using the same model under different names. Students are often taught "regression" and "ANOVA" in different contexts as though they are different. But now, my canine friend, you know better. Nice grains, by the way.'

With that, the cat finished the last of the dog's food and darted out of the dog flap, leaving a ginger vapour trail.



3 Having said this, it is well worth the effort of trying to obtain equal sample sizes in your different conditions because unbalanced designs do cause statistical complications (see Section 11.5).

So, how do we use the linear model to compare means? Going back to the puppy study, to predict happiness from group membership we can adapt the general equation that keeps popping up:

$$\text{outcome}_i = \text{model} + \text{error}_i \tag{12.1}$$

A more specific version of this equation (Eq. 12.2) shows that we'd estimate (hence the hats) happiness from group membership and two parameters (*b*s) that represent the levels of happiness when **group** is zero ($\hat{b}_0$) and the association between group membership and happiness ($\hat{b}_1$):

$$\text{happiness}_i = \hat{b}_0 + \hat{b}_1 \text{group}_{1i} + e_i \tag{12.2}$$

We saw in Chapter 10 that we can include a predictor variable containing two categories into the linear model and the resulting $\hat{b}$ for that predictor will compare the difference between the mean score for the two categories. Specifically, categorical predictors with two categories can be included in a linear model by coding the categories with 0 and 1.[4] So far so good, except that we have three categories (no puppies, 15 minutes and 30 minutes), not two. You might reasonably think that with three categories we could use 0, 1 and 2, but we can't. The problem is that categories typically have no intrinsic order and are equally important, but by using any sequence of numbers we give them order and magnitude (i.e. numerically some categories become more 'important' than others). We'd also be assuming that a continuum underlies the categories. These points might be easier to think about if you imagine predicting a creator's influence across different social media platforms. Let's say you have some measure of 'influence' for each creator. Your categorical predictor would be what platform they create on: YouTube, TikTok, Instagram, Twitch or whatever else happens to be fashionable to people more hip than I am. Let's say we assign these four platforms the values 1 to 4, respectively. It would be meaningless to assign meaning to the numbers that represent these platforms: there is no continuous variable underlying those categories. The fact that TikTok is assigned 1 and Twitch 2 does not mean that 2 × TikTok = Twitch. Hopefully, you get the point.

Although we can't in general treat multiple categories as a sequence of numbers, we can express them using numbers. As I've already reminded you, categorical predictors with two categories can be included in a linear model by coding the categories with 0 and 1. We can extend this idea and include a categorical predictor with several categories into a linear model by converting it to *several* predictor variables each of which has only two categories. There are different ways to achieve this conversion and we'll cover two of them in this chapter: first we'll explore dummy coding (a.k.a. indicator coding) and later we'll look at contrast coding.[5]

12.3.1 Dummy coding Ⓑ

Dummy coding is the idea of using 0s and 1s to represent categories. We've used it already when we used a single dummy variable to represent two categories in Chapter 10. If we need to represent more than two categories, we increase the number of dummy variables accordingly; the number of dummy variables we need is one fewer than the number of groups we're recoding. There are eight steps:

1. Count the number of groups you want to recode and subtract 1.
2. Create as many new variables as the value you calculated in step 1. These are your dummy variables.
3. Choose one of your groups as a baseline against which all other groups will be compared. Normally you'd pick a group that might be considered a control, or, if you don't have a specific

4 We saw in Section 10.4 why we use 0 and 1.

5 For some more detailed reading see Hardy (1993).

hypothesis, the group that represents the majority of people (because it might be interesting to compare other groups against the majority).
4 Having chosen a baseline group, assign that group values of 0 for all dummy variables.
5 For your first dummy variable, assign the value 1 to the first group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
6 For the second dummy variable assign the value 1 to the second group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
7 Repeat this process until you run out of dummy variables.
8 Place all the dummy variables into the linear model in the same block.

Let's try this out on the puppy data:

- Steps 1–2: We have three groups, so we'll need two dummy variables.
- Step 3: The baseline category should be the condition against which you intend to compare the other groups. In most well-designed experiments there will be a group of participants who act as a control for the other groups and, other things being equal, this will be your baseline category – although the group you choose will depend upon the particular hypotheses you want to test. In designs in which the group sizes are unequal it is important that the baseline category contains a large number of cases to ensure that the estimates of the $\hat{b}$-values are reliable. In the puppy therapy example, we can take the control group (who received no puppy therapy) as the baseline category because we are interested in comparing both the 15- and 30-minute groups to this group.
- Step 4: Having chosen a baseline group, assign that group values of 0 for all dummy variables. Table 12.2 shows this step.
- Step 5: For your first dummy variable, assign the value 1 to the first group that you want to compare against the baseline group. Assign all other groups 0 for this variable. Let's start by comparing the 30-minute group to the no puppy therapy group. We assign 1 to anyone in the 30-minute group and 0 to everyone in the other groups as shown in Table 12.3. This dummy variable compares people in the longer therapy session to those in the no puppy (control) group, so we could call it **long** instead of 'dummy variable 1'.
- Steps 6–7: For the second dummy variable assign the value 1 to the second group that you want to compare against the baseline group. Assign all other groups 0 for this variable. Repeat this process until you run out of dummy variables. The only other group left to assign is the 15-minute group, so we assign 1 to people in the 15-minute group and 0 to all other groups as shown in Table 12.4. This dummy variable compares people in the short therapy session to those in the no puppy (control) group, so let's call it **short** instead of 'dummy variable 2'.

Note from Table 12.4 that every person has their group membership coded by a unique combination of 0s and 1s across the two dummy variables. Someone in the no puppy condition would have a 0 for both dummy variables, someone who had 15 minutes of therapy would have a 0 on the dummy variable labelled **long** and a 1 on the dummy variable labelled **short**, and someone who received 30 minutes of puppy therapy would have a 1 on the dummy variable labelled **long** and a 0 on the dummy variable labelled **short**.

Table 12.2 Dummy coding begins by assigning 0 to the control group for all dummy variables

Group	Dummy variable 1	Dummy variable 2
No puppies (control)	0	0
15 minutes of puppy therapy		
30 minutes of puppy therapy		

Table 12.3 The first dummy variable compares the 15-minuted group to the control group

Group	Dummy variable 1 (long)	Dummy variable 2
No puppies (control)	0	0
15 minutes of puppy therapy	0	
30 minutes of puppy therapy	1	

Table 12.4 Dummy coding for the three-group experimental design

Group	Dummy variable 1 (long)	Dummy variable 2 (short)
No puppies (control)	0	0
15 minutes of puppy therapy	0	1
30 minutes of puppy therapy	1	0

Let's try extending this idea beyond three groups. In Chapter 6 we encountered a biologist who was worried about the potential health effects of music festivals. She originally collected data at a heavy metal festival (Download Festival), but was worried that her findings might not generalize beyond metal fans. Perhaps it's only metal fans who get smellier at festivals (at this point, as a metal fan, I would sacrifice the biologist to Odin for her prejudices). To find out whether the type of music a person likes predicts whether hygiene decreases over the festival the biologist measured hygiene over the three days of the Glastonbury Music Festival, which has an eclectic clientele. Her hygiene measure ranged between 0 (you smell like you've bathed in sewage) and 4 (you smell like you've bathed in freshly baked bread). The data file (**glastonbury.sav**) contains the hygiene scores for each day of the festival and a variable called **change**, which is the change in hygiene from day 1 to day 3 of the festival. Not everyone could be followed up on day 3, so only a subset of the original sample has a change score. The biologist coded the festival-goers' subcultures into the categories 'hipster' (people who mainly like alternative music), 'metalhead' (people who like heavy metal) and 'raver' (people who like dance/electronic/ambient stuff). Anyone not falling into these categories was labelled 'no subculture'. These categories are stored in the variable **music**.

Following the steps for creating dummy variables. With four groups we need three dummy variables (one fewer than the number of groups). The first step is to choose a baseline group. We're interested in comparing those who have different subcultures against those who don't, so our baseline category will be 'no subculture'. We code this group with 0 for all dummy variables. For our first dummy variable, we could compare the 'raver' group to this baseline by assigning anyone who is a raver a code of 1, and everyone else a code of 0. For our second dummy variable, we could compare the 'metalhead' group by assigning a 1 to anyone who is a metalhead, and a 0 to everyone else. Our final dummy variable codes the final category 'hipster' by assigning 1 to anyone who is a hipster, and a 0 to everyone else. Table 12.5 illustrates the resulting coding scheme. Note that each group has a code of 1 on only one of the dummy variables (except the base category, which is always coded as 0).

Table 12.5 Dummy coding for the Glastonbury festival data

Music category	Dummy variable 1 (raver vs. no subculture)	Dummy variable 2 (metalhead vs. no subculture)	Dummy variable 3 (hipster vs. no subculture)
No subculture	0	0	0
Hipster	0	0	1
Metalhead	0	1	0
Raver	1	0	0

12.3.2 Dummy coding and the linear model Ⓑ

Back to puppies. Having converted our categories into dummy variables, we can now add these dummy variables to our model as predictors. Remember that we used the control group (no puppies) as the baseline category and the two dummy variables represent the length of the dose of puppy therapy: one compares the 30-minute dose against no puppies (**long**) and the other compares the 15-minute dose against no puppies (**short**). Putting these dummy variables into the model as predictors gives:

$$\text{happiness}_i = \hat{b}_0 + \hat{b}_1\text{long}_i + \hat{b}_2\text{short}_i + e_i \tag{12.3}$$

in which a person's happiness is predicted from knowing their group code (i.e., the numeric code for the **long** and **short** dummy variables) and the estimates of the parameters for these effects and the intercept ($\hat{b}_0$). Specifically, the predicted value of happiness is calculated Eq. 12.4 (remember from Jane Superbrain Box 2.4 that we put a hat on happiness to remind us that it's an estimate, and these predicted values have no error term):

$$\widehat{\text{happiness}}_i = \hat{b}_0 + \hat{b}_1\text{long}_i + \hat{b}_2\text{short}_i \tag{12.4}$$

When we are predicting an outcome from group membership, predicted values from the model (the value of happiness in Eq. 12.4) are the group means. This is illustrated in Figure 12.3, which splits the data by group membership. If we are trying to predict the happiness of a new person and we know to which group they have been assigned (but we don't have their score yet) then our best guess will be the group mean because, on average, we'll be correct. For example, if we know that someone is going to receive 30 minutes of puppy therapy and we want to predict their happiness, our best guess will be 5 because we know that, on average, people who have 30 minutes with a puppy rate their happiness as 5. If the group means are meaningfully different, then using the group means should be an effective way to predict scores (because we can successfully differentiate people's predicted happiness based on how much puppy time they received). Remembering from Chapter 9, that we test the overall fit of a linear model with an *F*-statistic, we can do the same here: we first use an *F* to test whether we can significantly predict the outcome variable by using group means (which tells us whether, overall, the group means are significantly different) and then use the specific model parameters (the *b*s) to tell us which means differ from which. So let's have a look at what the parameters (*b*s) represent.

Let's first examine the model for the no puppy control group. Both the **long** and **short** dummy variables are coded 0 for people in the control group. Substituting these values into Eq. 12.4 the model becomes:

$$\begin{aligned}\widehat{\text{happiness}}_i &= \hat{b}_0 + \left(\hat{b}_1 \times 0\right) + \left(\hat{b}_2 \times 0\right) \\ &= \hat{b}_0 \\ \bar{X}_{\text{No puppies}} &= \hat{b}_0\end{aligned} \tag{12.5}$$

The 15- and 30-minute groups have dropped out of the model (because they are coded with 0) and we're left with $\hat{b}_0$. As we have just discovered, the predicted value of happiness will be the mean of the control group ($\bar{X}_{\text{No puppies}}$), so we can replace $\widehat{\text{happiness}}_i$ with this value. This leaves us with the epiphany that $\hat{b}_0$ in the model is always the mean of the baseline category.

For someone in the 30-minute group, the value of the dummy variable **long** will be 1 and the value for **short** will be 0. By substituting these values into Eq. 12.4 the model becomes:

$$\begin{aligned}\widehat{\text{happiness}}_i &= \hat{b}_0 + \left(\hat{b}_1 \times 1\right) + \left(\hat{b}_2 \times 0\right) \\ &= \hat{b}_0 + \hat{b}_1\end{aligned} \tag{12.6}$$

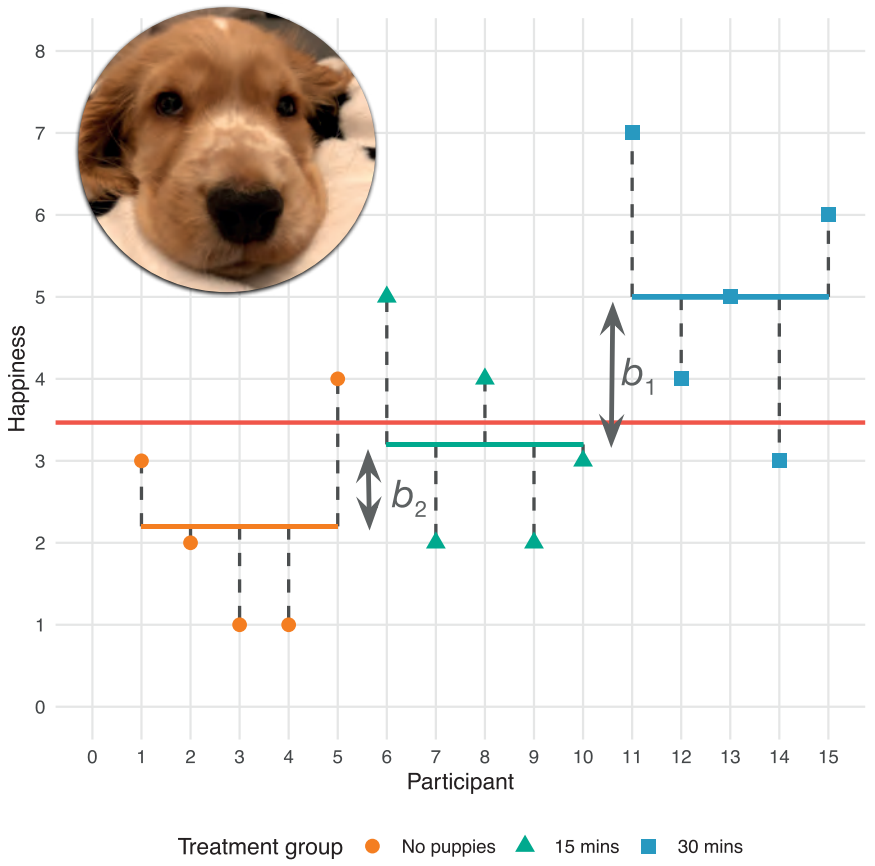


Figure 12.3 The puppy therapy data. The coloured horizontal lines represent the mean happiness of each group. The shapes represent the happiness of individual participants (different shapes indicate different experimental groups). The red horizontal line is the average happiness of all participants. The puppy is my dog Milton, as a pup. You should never needlessly apply puppies to your plots ... unless they're Milton, in which case it's fine.

which tells us that predicted happiness for someone in the 30-minute group is the sum of $\hat{b}_0$ and the b for the dummy variable **long** ($\hat{b}_1$). We know already that $\hat{b}_0$ is the mean of the control group ($\bar{X}_{\text{No puppies}}$) and that the predicted value of **happiness** for someone in the 30-minute group is the mean of that group ($\bar{X}_{30\text{ mins}}$). Therefore, we can replace $\hat{b}_0$ with $\bar{X}_{\text{No puppies}}$ and $\widehat{\text{happiness}}_i$ with $\bar{X}_{30\text{ mins}}$. The result is:

$$\begin{aligned}\widehat{\text{happiness}}_i &= \hat{b}_0 + \hat{b}_1 \\ \bar{X}_{30\text{ mins}} &= \bar{X}_{\text{No puppies}} + \hat{b}_1 \\ \hat{b}_1 &= \bar{X}_{30\text{ mins}} - \bar{X}_{\text{No puppies}}\end{aligned} \tag{12.7}$$

which shows that the $\hat{b}$-value for the dummy variable representing the 30-minute group is the difference between the means of that group and the control.

Finally, for someone assigned to the 15-minute group the dummy variable **short** has a value of 1 and the dummy variable **long** is 0. By substituting these values into Eq. 12.4 the model becomes:

$$\begin{aligned}\widehat{\text{happiness}}_i &= \hat{b}_0 + \left(\hat{b}_1 \times 0\right) + \left(\hat{b}_2 \times 1\right) \\ &= \hat{b}_0 + \hat{b}_2.\end{aligned} \tag{12.8}$$

which tells us that predicted happiness for someone in the 15-minute group is the sum of $\hat{b}_0$ and the $\hat{b}$ for the dummy variable **short** ($\hat{b}_2$). Again, we can replace $\hat{b}_0$ with $\bar{X}_{\text{No puppies}}$. The predicted value of **happiness** for someone in the 15-minute group is the mean of that group, so we can replace $\widehat{\text{happiness}}_i$ with $\bar{X}_{15\text{ mins}}$. The result is:

$$\begin{aligned} \widehat{\text{happiness}}_i &= \hat{b}_0 + \hat{b}_2 \\ \bar{X}_{15\text{ mins}} &= \bar{X}_{\text{No puppies}} + \hat{b}_2 \\ \hat{b}_2 &= \bar{X}_{15\text{ mins}} - \bar{X}_{\text{No puppies}} \end{aligned} \tag{12.9}$$

which shows that the $\hat{b}$-value for the dummy variable representing the 15-minute (**short**) group is the difference between means for the 15-minute group and the control.

Using dummy coding is only one of many ways to code dummy variables. We shall see later in this chapter (Section 12.5) that an alternative is contrast coding, in which you code the dummy variables in such a way that the $\hat{b}$-values represent differences between groups that you specifically hypothesized before collecting data.



To illustrate what is going on I have created a file called **puppies_dummy.sav** that contains the puppy therapy data along with the two dummy variables (**short** and **long**) we've just discussed (Table 12.4). Fit a linear model predicting **happiness** from **short** and **long**.

Output 12.1 shows the model from the self-test. The overall fit of the model has been tested with an *F*-statistic (i.e., ANOVA), which is significant, $F(2, 12) = 5.12$, $p = 0.025$. Given that our model represents the group means, this *F* tells us that using group means to predict happiness scores is significantly better than using the mean of all scores: in other words, the group means are significantly different.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20.133	2	10.067	5.119	.025[b]
	Residual	23.600	12	1.967		
	Total	43.733	14			

a. Dependent Variable: Happiness (0–10)

b. Predictors: (Constant), Short vs none (dummy variable), Long vs. none (dummy variable)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.200	.627		3.508	.004
	Long vs. none (dummy variable)	2.800	.887	.773	3.157	.008
	Short vs none (dummy variable)	1.000	.887	.276	1.127	.282

a. Dependent Variable: Happiness (0–10)

Output 12.1

The *F*-test is an overall test that doesn't identify differences between specific means. However, the model parameters (the $\hat{b}$-values) do. Quickly remind yourself of the group means from Table 12.1. As we just discovered, the constant ($\hat{b}_0$) is equal to the mean of the baseline category (the control group), 2.2. The $\hat{b}$-value for the first dummy variable ($\hat{b}_1$) is equal to the difference between the means of the 30-minute group and the control group (5.0 – 2.2 = 2.8). Finally, the $\hat{b}$-value for the second dummy variable ($\hat{b}_2$) is equal to the difference between the means of the 15-minute group and the control group (3.2 – 2.2 = 1). This demonstrates what we saw in Eq. 12.5, Eq. 12.7 and Eq. 12.9. We can see from the significance values of the associated *t*-tests that the difference between the 30-minute group and the control group ($\hat{b}_1$) is significant because $p = 0.008$, which is less than 0.05; however, the difference between the 15-minute and the control group is not ($p = 0.282$). This gives you a flavour of what this chapter is about, but we'll return to it all in more detail.

12.3.3 An overview of the *F*-statistic Ⓑ

We learnt in Chapter 9 that the *F*-statistic (or *F*-ratio as it's also known) tests the overall fit of a linear model to a set of observed data. *F* is the ratio of how good the model is compared to how bad it is (its error). When the model is based on group means, our predictions from the model are those means. If the group means are the same then our ability to predict the observed data will be poor (*F* will be small) but if the means differ we will be able to better discriminate between cases from different groups (*F* will be large). So, in this context *F* basically tells us whether the group means are significantly different. Let me elaborate.

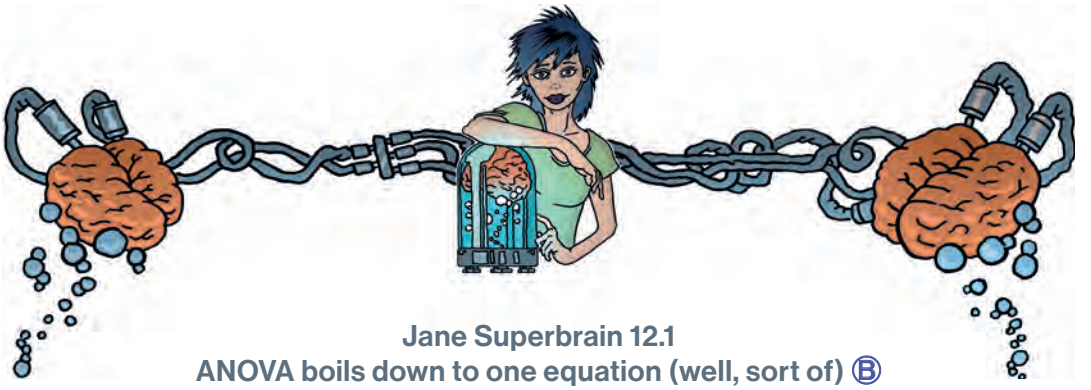
Figure 12.3 shows the puppy therapy data including the group means, the overall mean and the difference between each case and the group mean. We want to test the hypothesis that the means of three groups are different (so the null hypothesis is that the group means are the same). If the group means were all the same, then we would not expect the control group to differ from the 15- or 30-minute groups, and we would not expect the 15-minute group to differ from the 30-minute group. In this situation, the three coloured horizontal lines representing the group means in Figure 12.3 would be in the same vertical position (the exact position would be the grand mean – the red horizontal line in the figure). This is not the case in the figure: the coloured lines are in different vertical positions, showing that the group means are different. We have just found out that in the model, $\hat{b}_1$ represents the difference between the control and 30-minute group means, and $\hat{b}_2$ represents the difference between the 15-minute and control group means. These two distances are represented in Figure 12.3 by the vertical arrows. If the null hypothesis is true and all the groups have the same mean, then these *b* coefficients should be zero (because if the group means are equal then the difference between them will be zero).

We can apply the same logic as for any linear model:

- The model that represents 'no effect' or 'no relationship between the predictor variable and the outcome' is one where the predicted value of the outcome is always the grand mean (the mean of the outcome variable). We'll call this the null model.
- We can fit a different model to the data that represents our alternative hypotheses. We compare the fit of this model to the fit of the null model.
- The intercept and one or more parameters (*b*) describe the model, and we can estimate their values from the data ($\hat{b}$).
- The parameters determine the shape of the model that we have fitted; therefore, the bigger the coefficients, the greater the deviation between the model and the null model (grand mean).
- The parameters (*b*) represent the differences between group means. The bigger the differences between group means, the greater the difference between the model and the null model.

- If the differences between group means are large enough, then the resulting model will be a better fit to the data than the null model.
- If this is the case, we can infer that our model (i.e., predicting scores from the group means) is better than not using a model (i.e., predicting scores from the grand mean). Put another way, our group means are significantly different from the null (that all means are the same).

We use the *F*-statistic to compare the improvement in fit due to using the model (rather than the null, or grand mean, model) to the error that still remains. In other words, the *F*-statistic is the ratio of the explained to the unexplained variation. We calculate this variation using sums of squares, which might sound complicated, but isn't as bad as you think (see Jane Superbrain Box 12.1).



Jane Superbrain 12.1
ANOVA boils down to one equation (well, sort of) Ⓑ

At every stage of the ANOVA we're assessing variation (or deviance) from a particular model (be that the null model, or the model representing our hypothesis). Back in Section 2.6.1 we saw that the extent to which a model deviates from the observed data can be expressed as a sum of squared error in the general form of Eq. 2.17, which I repeat here:

$$\text{Total error} = \sum_{i=1}^{n}(\text{observed}_i - \text{model}_i)^2 \qquad (12.10)$$

So, when comparing means, as in the linear model generally, we use Eq. 12.10 to calculate the fit of the null model, and then the fit of the alternative model that represents our hypothesis. If the alternative model is any good then it should fit the data significantly better than the null model.

All of the sums of squares we look at in this chapter are variations on Eq. 12.10: all that changes is what we use as the model and observed data. As you read about the sums of squares, think back to Eq. 12.10 to remind yourself that the equations are just variants of looking at the difference between observed values and values predicted by the model.



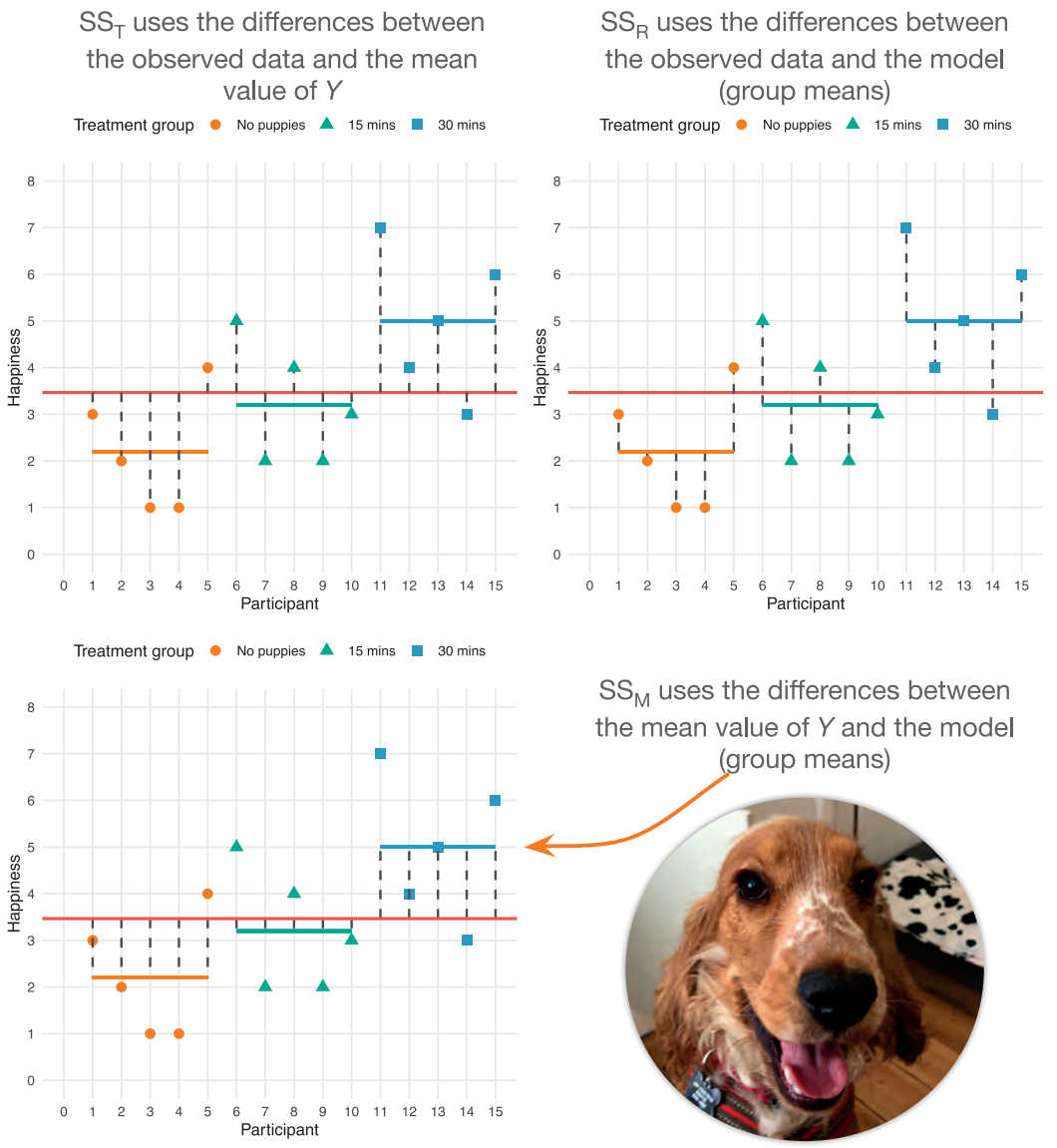


Figure 12.4 Graphical representation of the different sums of squares when comparing several means using a linear model. Also a picture of Milton as a puppy. Tufte would call him chartjunk, but I call him my loveable, crazy spaniel

12.3.4 Total sum of squares (SS_T) Ⓑ

To find the total amount of variation within our outcome we calculate the difference between each observed data point and the grand mean. We square these differences and add them to give us the total sum of squares (SS_T):

$$SS_T = \sum_{i=1}^{N} \left(x_i - \bar{x}_{grand}\right)^2 \tag{12.11}$$

The variance and the sum of squares are related such that variance, $s^2 = SS/(N - 1)$, where N is the number of observations (Section 2.6.1). Therefore, we can calculate the total sum of squares from the variance of all observations (the **grand variance**) by rearranging the relationship ($SS = s^2(N - 1)$).

The grand variance is the variation between all scores, regardless of the group from which the scores come. Figure 12.4 shows the different sums of squares graphically (note the similarity to Figure 9.4 which we looked at when we learnt about the linear model). The top left panel shows the total sum of squares: it is the sum of the squared distances between each point and the solid horizontal line (which represents the mean of all scores).

The grand variance for the puppy therapy data is given in Table 12.1, and there were 15 scores in all. Therefore, SS_T is 43.74:

$$\begin{aligned} SS_T &= s^2_{grand}(N-1) \\ &= 3.124(15-1) \\ &= 3.124 \times 14 \\ &= 43.736 \end{aligned} \tag{12.12}$$

Before we move on, have a look back at Jane Superbrain Box 2.5 to refresh your memory on degrees of freedom. When we estimate population values, the degrees of freedom are typically one less than the number of scores used to calculate the estimate. This is because to get the estimates we hold something constant in the population (e.g., to get the variance we hold the mean constant), which leaves all but one of the scores free to vary. For SS_T, we used the entire sample (i.e., 15 scores) to calculate the sums of squares and so the total degrees of freedom (df_T) are one less than the total sample size ($N - 1$). For the puppy therapy data, this value is 14.

12.3.5 Model sum of squares (SS_M) Ⓑ

So far, we know that the total amount of variation within the outcome is 43.74 units. We now need to know how much of this variation the model can explain. Because our model predicts the outcome from the means of our treatment (puppy therapy) groups, the model sum of squares tells us how much of the total variation in the outcome can be explained by the fact that different scores come from entities in different treatment conditions.

The model sum of squares is calculated by taking the difference between the values predicted by the model and the grand mean (see Section 9.2.4, Figure 9.4). When making predictions from group membership, the values predicted by the model are the group means (the coloured horizontal lines in Figure 12.4). The bottom panel in Figure 12.4 shows the model sum of squared error: it is the sum of the squared distances between what the model predicts for each data point – the mean of the group to which the score belongs represented by the orange (control), green (15 mins) and blue (30 mins) horizontal lines – and the overall mean of the outcome (the red horizontal line). For example, the predicted value for the five participants in the control group (circles) is 2.2, for the participants (triangles) in the 15-minute condition is 3.2, and for the participants (squares) in the 30-minute condition is 5. The model sum of squares requires us to calculate the differences between each participant's predicted value and the grand mean. These differences are squared and added together (for reasons I've explained before). Given that the predicted value for participants in a group is the same value (the group mean), the easiest way to calculate SS_M is by using:

$$SS_M = \sum_{g=1}^{k} n_g\left(\bar{x}_g - \bar{x}_{grand}\right)^2 \tag{12.13}$$

This equation basically says:

- Calculate the difference between the mean of each group ($\bar{x}_g$) and the grand mean ($\bar{x}_{grand}$).
- Square each of these differences.
- Multiply each result by the number of participants within that group (n_g).
- Add the values for each group together.

If we do this for the puppy therapy data, we get:

$$\begin{aligned} SS_M &= 5(2.200 - 3.467)^2 + 5(3.200 - 3.467)^2 + 5(5.000 - 3.467)^2 \\ &= 5(-1.267)^2 + 5(-0.267)^2 + 5(1.533)^2 \\ &= 8.025 + 0.355 + 11.755 \\ &= 20.135 \end{aligned} \tag{12.14}$$

For SS_M, the degrees of freedom (df_M) are one less than the number of 'things' used to calculate the SS. We used the three group means, so df_M is the number of groups minus one (which you'll see denoted by $k - 1$). So, in this example, we have three groups and the degrees of freedom are 2 (because the calculation of the sums of squares is based on the group means, two of which will be free to vary in the population if the third is held constant).

12.3.6 Residual sum of squares (SS_R) Ⓑ

We now know that there are 43.74 units of variation to be explained in our outcome variable, and that our model explains 20.14 of them (nearly half). The residual sum of squares (SS_R) tells us how much of the variation *cannot* be explained by the model. This value is the amount of variation created by things that we haven't measured such as measurement error and individual differences in things that might affect happiness. The simplest way to calculate SS_R is to subtract SS_M from SS_T ($SS_R = SS_T - SS_M$), but this provides little insight into what SS_R represents and, of course, if you've messed up the calculations of either SS_M or SS_T (or both!) then SS_R will be incorrect also.

We saw in Section 9.2.4 that the residual sum of squares is the difference between what the model predicts and what was observed. When using group membership to predict an outcome the values predicted by the model are the group means (the coloured horizontal lines in Figure 12.4). The top right panel of Figure 12.4 shows the residual sum of squared error: it is the sum of the squared distances between each point and the horizontal line for the group to which the score belongs.

We already know that for a given participant, the model predicts the mean of the group to which that person belongs. Therefore, SS_R is calculated by looking at the difference between the score obtained by a person and the mean of the group to which the person belongs. In graphical terms the dashed vertical lines in Figure 12.4 represent this sum of squares. These distances between each data point and the group mean are squared and added together to give the residual sum of squares, SS_R:

$$SS_R = \sum_{g=1}^{k}\sum_{i=1}^{n}\left(x_{ig} - \bar{x}_g\right)^2 \tag{12.15}$$

Eq. 12.15 says that the sum of squares for each group is the squared difference between each participant's score in a group (x_{ig}) and the group mean ($\bar{x}_g$), and the two sigma signs mean that we repeat this calculation for the first participant ($i = 1$) through to the last (n), in the first group ($g = 1$) through to the last (k). As such, we can also express SS_R as $SS_R = SS_{\text{group 1}} + SS_{\text{group 2}} + SS_{\text{group 3}} + \ldots + SS_{\text{group } k}$. We know that the variance is the sums of squares divided by $n - 1$, we can express the residual sum of squares in terms of the variance like we did for the total sum of squares. The result is:

$$SS_R = \sum_{g=1}^{k} s_g^2\left(n_g - 1\right) \tag{12.16}$$

which translates as 'multiply the variance for each group (s_g^2) by one less than the number of people in that group ($n_g - 1$), then add the results for each group together'. For the puppy therapy data, we get:

$$\begin{aligned} SS_R &= s^2_{\text{group }1}(n_1 - 1) + s^2_{\text{group }2}(n_2 - 1) + s^2_{\text{group }3}(n_3 - 1) \\ &= 1.70(5-1) + 1.70(5-1) + 2.50(5-1) \\ &= (1.70 \times 4) + (1.70 \times 4) + (2.50 \times 4) \\ &= 6.8 + 6.8 + 10 \\ &= 23.60 \end{aligned} \tag{12.17}$$

The degrees of freedom for SS_R (df_R) are the total degrees of freedom minus the degrees of freedom for the model ($df_R = df_T - df_M = 14 - 2 = 12$). Put another way, this is $N - k$: the total sample size, N, minus the number of groups, k.

12.3.7 Mean squares Ⓑ

SS_M tells us the *total* variation that the model (in this case the experimental manipulation) explains and SS_R tells us the *total* variation that is due to unmeasured factors. Because both values are sums their size depends on the number of scores; for example, SS_M used the sum of three values (the group means), whereas SS_R and SS_T used the sum of 15 values. To eliminate this inequality we calculate an average sum of squares (known as the *mean squares*, MS). Rather than dividing by the number of scores used for each SS, we divide by the degrees of freedom because we are trying to extrapolate to a population and so some parameters within that populations will be held constant (remember we did this when calculating the variance, see Jane Superbrain Box 2.5). For the puppy therapy data we get a model mean squares (MS_M) of 10.067 and a residual mean squares (MS_R) of 1.96:

$$\begin{aligned} MS_M &= \frac{SS_M}{df_M} = \frac{20.135}{2} = 10.067 \\ MS_R &= \frac{SS_R}{df_R} = \frac{23.60}{12} = 1.97 \end{aligned} \tag{12.18}$$

MS_M represents the average amount of variation explained by the model (e.g., the systematic variation), whereas MS_R is a gauge of the average amount of variation explained by unmeasured variables (the unsystematic variation).

12.3.8 The *F*-statistic Ⓑ

The *F*-statistic (a.k.a. the *F*-ratio) is a measure of the ratio of the variation explained by the model and the variation attributable to unsystematic factors. In other words, it is the ratio of how good the model is to how bad it is (how much error there is). It is calculated by dividing the model mean squares by the residual mean squares:

$$F = \frac{MS_M}{MS_R} \tag{12.19}$$

As with other test statistics that we have looked at (e.g., *t*), the *F*-statistic is a signal-to-noise ratio. In experimental research, it is the ratio of the experimental effect to the individual differences in performance. Because *F* is the ratio of systematic to unsystematic variance, if it is less than 1 it means that MS_R is greater than MS_M and that there is more unsystematic than systematic variance. In experimental research this means that the effect of natural variation is greater than differences bought about by the experiment. In this scenario, we can, therefore, be sure that our experimental manipulation has been unsuccessful (because it has brought about less change than if we left our participants alone) and *F* will be non-significant. For the puppy therapy data, the *F*-statistic is:

$$F = \frac{MS_M}{MS_R} = \frac{10.067}{1.967} = 5.12 \tag{12.20}$$

meaning that the systematic variation is 5 times larger than the unsystematic variation; basically the experimental manipulation (puppy therapy groups) had some effect above and beyond the effect of individual differences in performance. Typically, researchers are interested in whether this ratio is significantly different from zero; put another way, what the probability would be of getting an F at least this big if the experimental manipulation, in reality, had no effect at all on happiness (i.e. the null hypothesis). When the time comes that society collapses and we revert to a simple bartering system and all technology has been jettisoned into space, we will compare the obtained value of F against the critical value for an F-distribution with the same degrees of freedom (see the Appendix). If the observed value exceeds the critical value we will probably conclude that our independent variable has had a genuine effect (because an F at least as big as the one we have observed would be very unlikely if there were no effect in the population). With 2 and 12 degrees of freedom the critical values are 3.89 (p = 0.05) and 6.93 (p = 0.01). The observed value, 5.12, is, therefore, significant at the 0.05 level of significance but not significant at the 0.01 level. Until the dawn of this anti-technological future, we can get the exact p from a computer: if it is less than the alpha level that we set before the experiment (e.g., 0.05 or 0.005) then scientists typically conclude that the variable that they manipulated had a genuine effect (more generally, predicting the outcome from group membership improves prediction).

12.3.9 Interpreting F Ⓑ



I've already mentioned that F assesses the overall fit of the model to the data. When the model is one that predicts an outcome from group means, F evaluates whether 'overall' there are differences between means; it does not provide specific information about which groups were affected. It is an *omnibus* test. In our puppy therapy example in which there are three groups, a significant F tells us that the means of these three samples are not equal (i.e., that $\bar{X}_1 = \bar{X}_2 = \bar{X}_3$ is *not* true). There are several ways in which the means can differ: (1) all three sample means could be significantly different ($\bar{X}_1 \neq \bar{X}_2 \neq \bar{X}_3$); (2) the means of groups 1 and 2 could be similar to each other but different from that of group 3 ($\bar{X}_1 = \bar{X}_2 \neq \bar{X}_3$); (3) the means of groups 2 and 3 could be similar to each other but different from that of group 1 ($\bar{X}_1 \neq \bar{X}_2 = \bar{X}_3$); or (4) the means of groups 1 and 3 could be similar to each other but different from that of group 2 ($\bar{X}_1 = \bar{X}_3 \neq \bar{X}_2$).

You might feel that F is a bit pointless because, given that you've gone to the trouble of running an experiment, you probably had more specific predictions than 'there's a difference, somewhere or other'. You might wonder why you don't fit lot of models each of which compares only two means at a time; after all, this would tell you specifically whether pairs of group means differ. The reason why we don't do this was explained in Section 2.10.8: every time you run a test on the same data you inflate the Type I error rate. We'll return to this point in Section 12.6 when we discover how to establish where the group differences lie. For now, though, the reason why the F-test is useful is that as a single test (albeit of a non-specific hypothesis) it controls the Type I error rate. Having established that overall group means differ (i.e. the outcome can be significantly predicted using the group means) we can use the parameters of the model (the $\hat{b}$-values) to tell us where the differences lie.

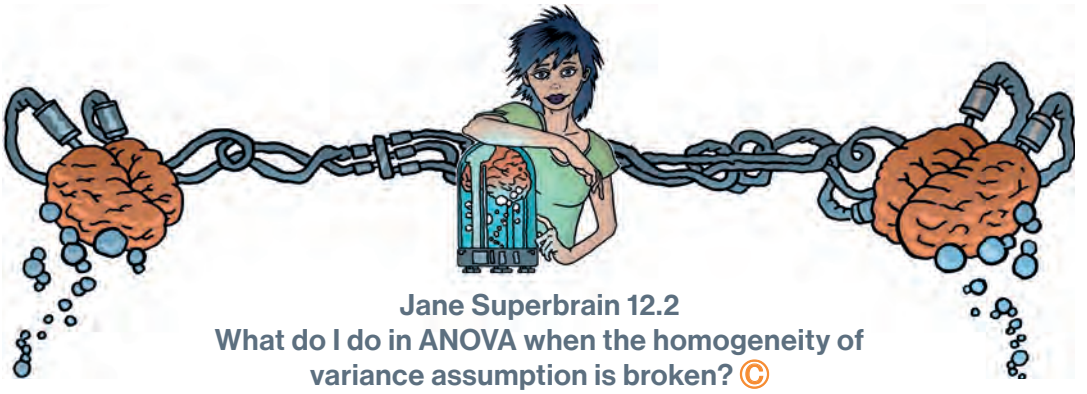
12.4 Assumptions when comparing means Ⓒ

To compare means we use a linear model, so all of the potential sources of bias discussed in Chapter 6 apply.

12.4.1 Homoscedasticity Ⓑ

As with any linear model, we assume that the variance of the model errors is steady as the predictor changes (in this context it means that variances of model errors are equal in each group or category). The assumption is the same as for continuous predictors, but the residual plot will look quite different (more on this later).

It's common in research designs that aim to test differences between means to use Levene's test to test this assumption, but I don't recommend it (see Section 6.9.6) because we can routinely adjust the model for the extent of heteroscedasticity. When using a linear model to compare means, two such corrections are the **Brown–Forsythe *F*** (Brown & Forsythe, 1974) and **Welch's *F*** (Welch, 1951). If you're really bored, these two statistics are discussed in Jane Superbrain Box 12.2. You can also use a robust version of *F* that does not assume homogeneity (Section 12.10).



Jane Superbrain 12.2
What do I do in ANOVA when the homogeneity of variance assumption is broken? Ⓒ

The Brown–Forsythe *F* is fairly easy to explain. When group sizes are unequal and the large groups have the biggest variance, *F* is conservative. If you think back to Eq. 12.16, this makes sense because to calculate SS_R variances are multiplied by their sample size (minus 1), so you get a large sample size cross-multiplied with a large variance, which will inflate the value of SS_R. *F* is proportionate to SS_M/SS_R, so if SS_R is big, then the *F*-statistic gets smaller (which is why it becomes conservative: its value is overly reduced). Brown and Forsythe get around this problem by weighting the group variances not by their sample size, but by $1 - n/N$. The fraction n/N is the sample size as a proportion of the total sample size, and (because proportions sum to 1) $1 - n/N$ reverses this proportion such that large sample sizes (large proportions) become small and vice versa. Therefore, this adjustment reduces the impact of large sample sizes with large variances:

$$F_{BF} = \frac{SS_M}{SS_{R_{BF}}} = \frac{SS_M}{\sum s_k^2\left(1 - \frac{n_k}{N}\right)}$$

For the puppy therapy data, SS_M is the same as before (20.135), so *F* becomes:

$$\begin{aligned} F_{BF} &= \frac{20.135}{s^2_{\text{group 1}}\left(1 - \frac{n_{\text{group 1}}}{N}\right) + s^2_{\text{group 2}}\left(1 - \frac{n_{\text{group 2}}}{N}\right) + s^2_{\text{group 3}}\left(1 - \frac{n_{\text{group 3}}}{N}\right)} \\ &= \frac{20.135}{1.7\left(1 - \frac{5}{15}\right) + 1.7\left(1 - \frac{5}{15}\right) + 2.5\left(1 - \frac{5}{15}\right)} \\ &= \frac{20.135}{3.933} \\ &= 5.119 \end{aligned}$$

This statistic is evaluated using degrees of freedom for the model and error terms. For the model, df_M is the same as before (i.e., $k - 1 = 2$), but an adjustment is made to the residual degrees of freedom, df_R. Welch's (1951) F is an alternative adjustment that is more involved to explain – if you're interested, see Oliver Twisted. Both adjustments control the Type I error rate well (i.e., when there's no effect in the population you do get a non-significant F), but Welch's F has more power (i.e., is better at detecting an effect that exists) except when there is an extreme mean that has a large variance (Tomarken & Serlin, 1986).





Oliver Twisted
Please, Sir, can I have some more ... Welch's *F*?

'You're only telling us about the Brown–Forsythe *F* because you don't understand Welch's *F*,' taunts Oliver. 'Andy, Andy, brains all sandy ....' Whatever, Oliver. Like the Brown–Forsythe *F*, Welch's *F* adjusts *F* and the residual degrees of freedom to combat problems arising from violations of the homogeneity of variance assumption. There is a lengthy explanation about Welch's *F* in the additional material available at www.discoverspss.com. Oh, and Oliver, microchips are made of sand.

12.4.2 Is ANOVA robust? ©

People often say 'ANOVA is a robust test', which means that if we break the assumptions, *F* will still be accurate. Remember from Chapter 6 that we mainly care about normality of the sampling distribution if we want to assess significance or construct confidence intervals. There are two issues to consider around the significance of *F*. First, does the *F* control the Type I error rate or is it significant even when there are no differences between means? Second, does *F* have enough power (i.e., is it able to detect differences when they are there)?



The myth that 'ANOVA is robust' is quite pervasive. It probably stems from research done a long time ago that investigated a limited range of situations. Examples of this work suggested that *when group sizes are equal* the *F*-statistic can be quite robust to violations of normality (Lunney, 1970), especially non-normality in the form of skew (Donaldson, 1968), and that when variances are proportional to the means the power of *F* is unaffected by heterogeneity of variance (Budescu, 1982; Budescu & Appelbaum, 1981).

The trouble is that this research is over 35 years old, so let's update things (for a more detailed review see Field & Wilcox, 2017). Recent simulations show that differences in skewness, non-normality and heteroscedasticity interact in complicated ways that impact power (Wilcox, 2017). For example, in the absence of normality, violations of homoscedasticity will affect *F* even when group sizes are equal (Wilcox, 2010, 2016, 2017), and when means are equal the error rate (which should be 5%) can be as high as 18%. Wilcox (2016) suggests that *F* can be considered robust only if the group distributions are identical; for example, if groups are skewed to an identical degree, which in practice is probably unlikely. Heavy-tailed distributions are particularly problematic: if you set up a situation with power of 0.9 to detect an effect in a normal distribution and contaminate that distribution with 10% of scores from a normal distribution with a bigger variance (so you get heavier tails), power drops to 0.28 (despite the fact that only 10% of scores have changed). Similarly, Cohen's *d* drops from 1 when distributions are normal to 0.28 (Wilcox et al., 2013). Also, heavy-tailed samples have implications for the central limit theorem, which says that in samples of 30 or more the sampling distribution should be normal (Section 6.8.2); for heavy-tailed distributions samples need to be much larger, up to 160 in some cases (Wilcox, 2010). To sum up, *F* is not robust, despite what your supervisor might tell you.

Violations of the assumption of independence are very serious indeed. Scariano and Davenport (1987) showed that if scores are made to correlate moderately (say, with a Pearson coefficient of 0.5) to create dependencies in the model errors, then when comparing three groups of 10 observations per group the Type I error rate is 0.74 (remember that we'd expect it to be 0.05). In other words, you think you'll make a Type I error 5% of the time but in fact you'll make one 74% of the time!

12.4.3 What to do when assumptions are violated Ⓑ

Violations of assumptions are nowhere near the headache they used to be. In Chapter 6 we discussed methods for correcting problems (e.g., the bias reduction methods in Section 6.10), but these can mostly be avoided. If you routinely interpret Welch's *F* then you need never even think about homoscedasticity; in small samples you can bootstrap parameter estimates so that their confidence intervals and *p*-values are robust. There are also robust tests that use 20% trimmed means and a bootstrap, which we can implement. In fact, there is a school of thought that you should apply robust tests in all situations and not bother thinking about assumptions. Finally, you can use the Kruskal–Wallis test (Chapter 7) although I prefer the aforementioned robust tests. If you do apply the usual *F*-statistic then, at the very least, conduct a sensitivity analysis (i.e., apply a robust test to check that your conclusion doesn't change).

12.5 Planned contrasts (contrast coding) Ⓑ

I've already alluded to the need to follow up a significant *F* by looking at the model parameters, which tell us about specific differences between means. In fact, in Section 12.3.2 we saw that if we dummy-code the groups then the $\hat{b}$ for each dummy variable compares the mean of the baseline group to the group coded with a 1 in the particular dummy variable. In the puppy therapy example we ended up with two *b*s, one comparing the mean of the 15-minute group to the control group and the other comparing the 30-minute group to the control. Each of these $\hat{b}$s is tested with a *t*-statistic that tells us whether the $\hat{b}$ is significantly different from 0 (i.e. the means differ). The trouble is that

with two dummy variables we end up with two *t*-tests, which inflates the familywise error rate (see Section 2.10.8). The other problem is that the dummy variables might not make all the comparisons that we want to make (e.g., the 15- and 30-minute groups are never compared).

There are a couple of solutions to these problems. The first is to use contrast coding rather than dummy coding. Contrast coding is a way of assigning weights to groups in dummy variables to carry out **planned contrasts** (also known as planned comparisons). Weights are assigned in such a way that the contrasts are independent, which means that the overall Type I error rate is controlled. A second option is to compare every group mean to all others (i.e., to conduct several overlapping tests using a *t*-statistic each time) but using a stricter acceptance criterion that keeps the familywise error rate at 0.05. These are known as using ***post hoc* tests** (see Section 12.6). Typically planned contrasts are done to test specific hypotheses, whereas *post hoc* tests are used when there were no specific hypotheses. Let's first look at contrast coding and planned contrasts.

12.5.1 Choosing which contrasts to do Ⓑ

In the puppy therapy example the primary hypothesis would be that any dose of puppy therapy should change happiness compared to the control group. A second hypothesis might be that a 30-minute session should increase happiness more than a 15-minute one. To do planned contrasts, these hypotheses must be derived *before* the data are collected. The *F*-statistic is based upon splitting the total variation into two component parts: the variation due to the model or experimental manipulation (SS_M) and the variation due to unsystematic factors (SS_R) (see Figure 12.5).

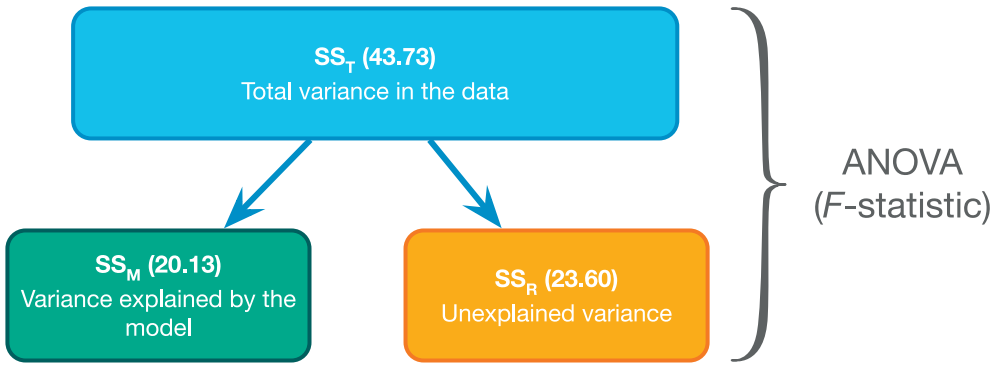


Figure 12.5 Partitioning variance for ANOVA

Planned contrasts extend this logic by breaking down the variation due to the model/experiment into component parts (see Figure 12.6). The exact contrasts will depend upon the hypotheses you want to test. Figure 12.6 shows a situation in which the model variance is broken down based on the two hypotheses that we ready discussed. Contrast 1 looks at how much variation in happiness is created by the two puppy conditions compared to the no puppy control condition. Next, the variation explained by puppy therapy (in general) is broken down to see what proportion is accounted for by a 30-minute session relative to a 15-minute one (contrast 2).

Typically, students struggle with the notion of designing planned contrasts, but there are three rules that can help you to work out what to do.

1. If you have a control group this is usually because you want to compare it against any other groups.
2. Each contrast must compare only two 'chunks' of variation.
3. Once a group has been singled out in a contrast it can't be used in another contrast.

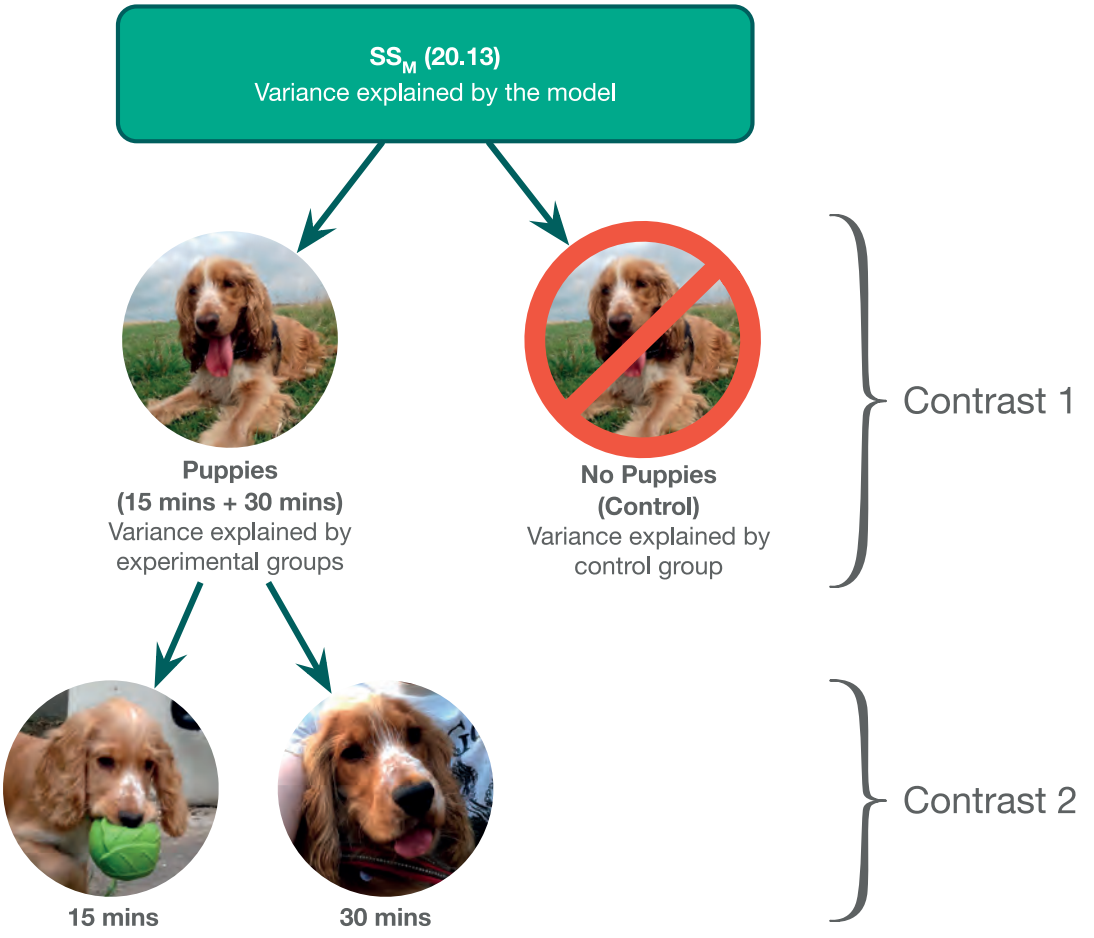


Figure 12.6 Partitioning of the model/experimental variance into component contrasts

Let's look at these rules in reverse order. First, if a group is singled out in one contrast, then it should not reappear in another contrast. The important thing is that we are breaking down one chunk of variation into smaller independent chunks. This independence matters for controlling the Type I error rate. It's like slicing up a cake: you begin with a cake (the total sum of squares) and then cut it into two pieces (SS_M and SS_R), then you take the piece of cake that represents SS_M and slice it again into smaller pieces. Once you have cut off a piece of cake you cannot stick that piece back onto the original slice, and you cannot stick it onto other pieces of cake, but you can divide it into smaller pieces of cake. Likewise, once a slice of variance has been split from a larger chunk, it cannot be attached to any other pieces of variance, it can only be subdivided into smaller chunks of variance. All this talk of cake is making me hungry, but hopefully it illustrates a point. So, in Figure 12.6 contrast 1 compares the control group to the experimental groups and because the control group is singled out, it is not incorporated into contrast 2.

Second, each contrast must compare only two chunks of variance. This rule is so that we can interpret the contrast. The original F tells us that some of our means differ, but not which ones, and if we were to perform a contrast on more than two chunks of variance we would be no better off. By comparing only two chunks of variance we know that the result represents a significant difference (or not) between these two portions of variation. If you follow the independence of contrasts rule (the cake slicing), and always compare only two pieces of variance, then you should end up with $k - 1$ contrasts (where k is the number of conditions you're comparing); in other words, one contrast fewer than the number of conditions you have in your design.

The first rule in the list reminds us that we often use at least one control condition, and it (or they) usually exists because we predict that the experimental conditions will differ from it (or them). As such, when planning contrasts the chances are that your first contrast will be one that compares all the experimental groups with the control group (or groups). Once you have done this first comparison, any remaining contrasts will depend upon which groups you predict will differ based on the theory you're testing.

To further illustrate these principles Figure 12.7 and Figure 12.8 show potential sets of contrasts for two different four-group experiments. In both examples there are three possible contrasts (one fewer than the number of groups) and every contrast compares only two chunks of variance. The first contrast is the same in both cases: the experimental groups are compared against the control group or groups. In Figure 12.7 there is only one control condition and this portion of variance is used only in the first contrast (because it cannot be broken down any further). In Figure 12.8 there were two control groups, and so the portion of variance due to the control conditions (contrast 1) can be broken down further to see whether the scores in the control groups differ from each other (contrast 3).

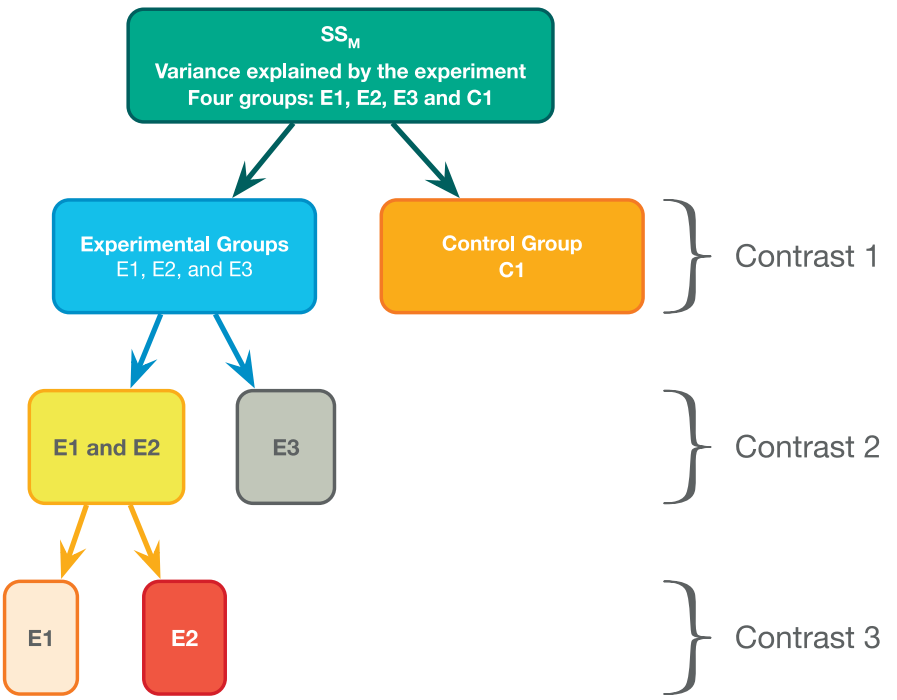


Figure 12.7 Partitioning variance for planned contrasts in a four-group experiment using one control group

In Figure 12.7, the first contrast contains a chunk of variance that is due to the three experimental groups, and this chunk of variance is broken down by first looking at whether groups E1 and E2 differ from E3 (contrast 2). It is equally valid to use contrast 2 to compare groups E1 and E3 to E2, or to compare groups E2 and E3 to E1. The exact contrast that you choose depends upon your hypotheses. For contrast 2 in Figure 12.7 to be valid we need to have a good theoretical reason to expect group E3 to be different from the other two groups. The third contrast in Figure 12.7 depends on the comparison chosen for contrast 2. Contrast 2 necessarily had to involve comparing two experimental groups against a third, and the experimental groups chosen to be combined must be separated in the final contrast. As a final point about Figure 12.7 and Figure 12.8, notice that once a group has been singled out in a contrast, it is never used in any subsequent one.

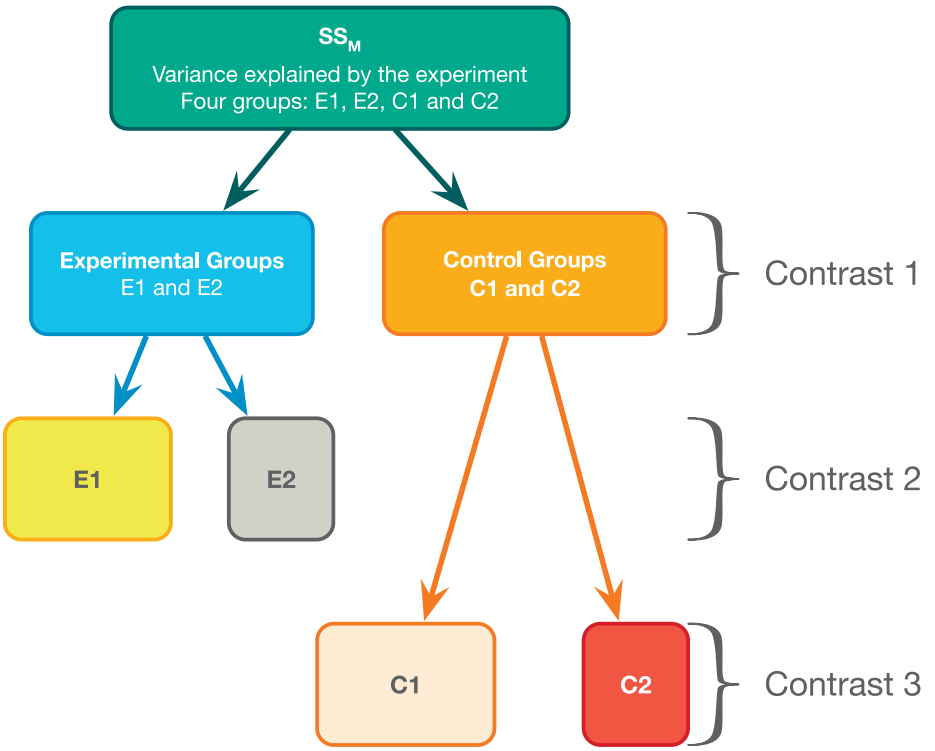


Figure 12.8 Partitioning variance for planned contrasts in a four-group experiment using two control groups



When we carry out a planned contrast, we compare 'chunks' of variance and these chunks often consist of several groups. When you design a contrast that compares several groups to one other group, you are comparing the means of the groups in one chunk with the mean of the group in the other chunk. As an example, for the puppy therapy data I suggested that an appropriate first contrast would be to compare the two dose groups with the control group. The means of the groups are 2.20 (control), 3.20 (15 minutes) and 5.00 (30 minutes), and so the first comparison, which compared the two experimental groups to the control, is comparing 2.20 (the mean of the control group) to the average of the other two groups ((3.20 + 5.00)/2 = 4.10). If this first contrast turns out to be significant, then we can conclude that 4.10 is significantly greater than 2.20, which in terms of the experiment tells us that the average of the experimental groups is significantly different from the average of the controls. You can probably see that logically this means that, if the standard errors are the same, the experimental group with the highest mean (the 30-minute group) will be significantly different from the mean of the control group. However, the experimental group with the lower mean (the 15-minute group) might not necessarily differ from the control group; we have to use the final contrast to make sense of the experimental conditions. For the puppy data the final contrast looked at whether the two experimental groups differ (i.e., is the mean of the 30-minute group significantly different from the mean of the 15-minute group?). If this contrast turns out to be significant then we can conclude that having 30 minutes of puppy therapy significantly affected happiness compared to having 15 minutes. If the contrast is non-significant then we conclude that the dosage of puppy therapy made no significant difference to happiness. In this latter scenario it is likely that both doses affect happiness more than control, whereas the former case implies that having 15 minutes may be no different than having a control. However, the

word *implies* is important here: it is possible that the 15-minute group might not differ from the control. To be completely sure you would need *post hoc* tests.

12.5.2 Defining contrasts using weights Ⓑ

Hopefully by now you have got some idea of how to plan which contrasts to do (i.e., if your brain hasn't exploded yet). The next issue is how to do them. To carry out contrasts we need to code our dummy variables such that the resulting $\hat{b}$s compare the 'chunks' that we set out in our contrasts. Remember that when we used dummy coding, we used values of 0 and 1 for the dummy variables and ended up with $\hat{b}$-values that compared each group mean to a baseline group mean. We need to work out what values (instead of 0 and 1) to assign to each group to give us (in this example) two $\hat{b}$-values, the first of which compares all puppy therapy scores to the control group and the second of which compares 15 minutes of puppy therapy with 30 minutes (and ignores the control group). The values assigned to the dummy variables are known as **weights**.

This procedure is horribly confusing, but there are a few basic rules for assigning values to the dummy variables to obtain the contrasts you want. I will explain these simple rules before showing how the process works. Remember the previous section when you read through these rules, and remind yourself of what I mean by a 'chunk' of variation.

- **Rule 1**: Choose sensible contrasts. Remember that you want to compare only two chunks of variation and that if a group is singled out in one contrast, that group should be excluded from any subsequent contrasts.
- **Rule 2**: Groups coded with positive weights will be compared against groups coded with negative weights. So, assign one chunk of variation positive weights and the opposite chunk negative weights.
- **Rule 3**: If you add up the weights for a given contrast the result should be zero.
- **Rule 4**: If a group is not involved in a contrast, automatically assign it a weight of zero, which will eliminate it from the contrast.
- **Rule 5**: For a given contrast, the weight assigned to the group(s) in one chunk of variation should be equal to the number of groups in the opposite chunk of variation.

OK, let's follow these rules to derive the weights for the puppy therapy data. The first contrast we chose was to compare the two experimental groups against the control (Figure 12.9). The first chunk of variation contains the two experimental groups, and the second chunk contains only the control group. Rule 2 states that we should assign one chunk positive weights, and the other negative. It doesn't matter which way round we do this, but for convenience let's assign chunk 1 positive weights, and chunk 2 negative weights as in Figure 12.9. Using rule 5, the weight we assign to the groups in chunk 1 should be equivalent to the number of groups in chunk 2. There is only one group in chunk 2 and so we assign each group in chunk 1 a weight of 1. Likewise, we assign a weight to the group in chunk 2 that is equal to the number of groups in chunk 1. There are two groups in chunk 1 so we give the control group a weight of 2. Then we combine the sign of the weights with the magnitude to give us weights of –2 (control), 1 (15 minutes) and 1 (30 minutes) as in Figure 12.9. Rule 3 states that for a given contrast, the weights should add up to zero, and by following rules 2 and 5 this should be true (if you haven't followed the rules properly then it will become clear when you add the weights). Let's check by adding the weights: sum of weights = 1 + 1 – 2 = 0. Happy days.

The second contrast was to compare the two experimental groups and so we want to ignore the control group. Rule 4 tells us that we should automatically assign this group a weight of 0 (to eliminate it). We are left with two chunks of variation: chunk 1 contains the 15-minute group and chunk 2 contains the 30-minute group. By following rules 2 and 5 one group is assigned a weight of +1 while the other is assigned –1, as in Figure 12.10. If we add the weights for contrast 2 we find that they again add up to zero: sum of weights = 1 – 1 + 0 = 0. Even happier days.

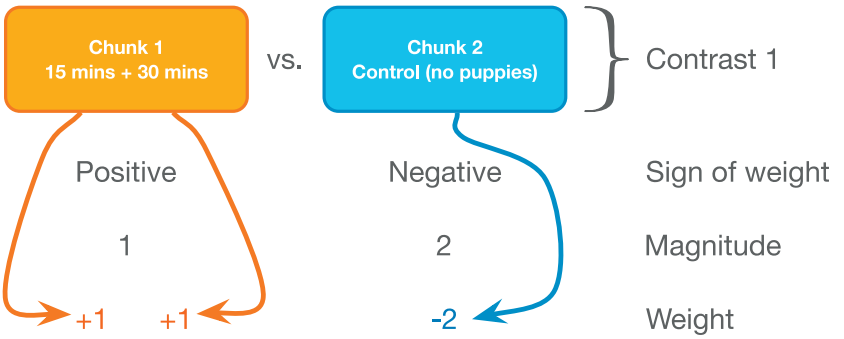


Figure 12.9 Assigning weights for the first contrast

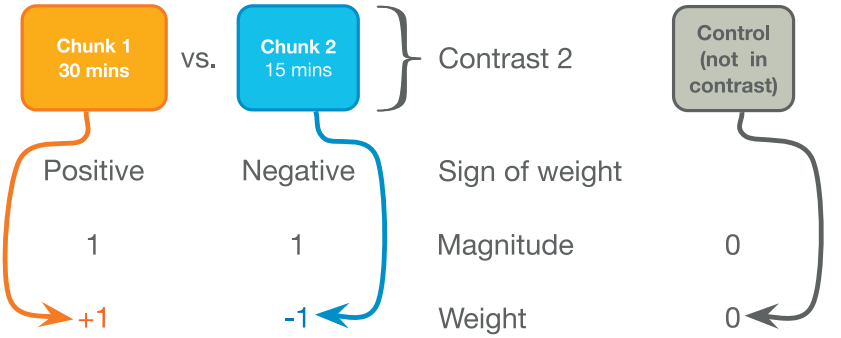


Figure 12.10 Assigning weights for contrast 2

The weights for each contrast are placed in the two dummy variables in the following equation:

$$\widehat{\text{happiness}}_i = \hat{b}_0 + \hat{b}_1 \text{contrast 1}_i + \hat{b}_2 \text{contrast 2}_i \qquad (12.21)$$

Hence, the weights are used in a linear model in which $\hat{b}_1$ represents contrast 1 (comparing the experimental groups to the control), $\hat{b}_2$ represents contrast 2 (comparing the 30-minute group to the 15-minute group), and $\hat{b}_0$ is the grand mean. Each group is specified now not by the 0 and 1 coding scheme that we initially used, but by the coding scheme for the two contrasts. Participants in the control group are identified by a code of –2 for contrast 1 and a code of 0 for contrast 2. Likewise, the 30-minute group is identified by a code of 1 for both contrasts, and the 15-minute group has a code of 1 for contrast 1 and a code of –1 for contrast 2 (Table 12.6).

Table 12.6 Orthogonal contrasts for the puppy therapy data

Group	Dummy variable 1 (Contrast 1)	Dummy variable 2 (Contrast 2)	Product Contrast 1 × Contrast 2
No puppies (control)	–2	0	0
15 minutes	1	–1	–1
30 minutes	1	1	1
Total	0	0	0



It is important that the weights for a contrast sum to zero because it ensures that you are comparing two unique chunks of variation. Therefore, a *t*-statistic can be used (remember it assumes independence). A more important consideration is that when you multiply the weights for a particular group, these products should also add up to zero (see the final column of Table 12.6). If the products sum to zero then the contrasts are *independent* or **orthogonal**. When we used dummy coding and fitted a linear model to the puppy therapy data, I commented that the familywise error rate for the *t*-statistics for the $\hat{b}$-values would be inflated (see Section 2.10.8). This is because these contrasts are not independent (both $\hat{b}$-values involve a comparison with the control group). However, if the contrasts *are* independent then the *t*-statistics for the $\hat{b}$-values are also independent and the resulting *p*-values are uncorrelated. You might think that it is very difficult to ensure that the weights you choose for your contrasts conform to the requirements for independence but, provided you follow the rules I have laid out, you should always derive a set of *orthogonal* contrasts. You should double-check by looking at the sum of the multiplied weights, and if this total is not zero then go back to the rules and see where you have gone wrong.

Earlier on, I mentioned that when you contrast-code dummy variables in a linear model the $\hat{b}$-values represent the differences between the means that the contrasts were designed to test. Let's have a look at how this works (this next part is not for the faint-hearted). When we do planned contrasts, the intercept $\hat{b}_0$ is equal to the grand mean (i.e., the value predicted by the model when group membership is not known), which in this example (and because group sizes are equal) is:

$$\hat{b}_0 = \text{grand mean} = \frac{\bar{X}_{\text{30 mins}} + \bar{X}_{\text{15 mins}} + \bar{X}_{\text{No puppies}}}{3} \tag{12.22}$$

For a participant in the control group, we define their group membership using the values for the control group in Table 12.6 and the predicted value of happiness is the mean of the control group. The model can be expressed as:

$$\begin{aligned} \widehat{\text{happiness}}_i &= \hat{b}_0 + \hat{b}_1\text{contrast } 1_i + \hat{b}_2\text{contrast } 2_i \\ \bar{X}_{\text{No puppies}} &= \left(\frac{\bar{X}_{\text{30 mins}} + \bar{X}_{\text{15 mins}} + \bar{X}_{\text{No puppies}}}{3}\right) + \left(-2\hat{b}_1\right) + \left(\hat{b}_2 \times 0\right) \end{aligned} \tag{12.23}$$

If we rearrange this equation and then multiply everything by 3 (to get rid of the fraction) we get:

$$\begin{aligned} 2\hat{b}_1 &= \left(\frac{\bar{X}_{\text{30 mins}} + \bar{X}_{\text{15 mins}} + \bar{X}_{\text{No puppies}}}{3}\right) - \bar{X}_{\text{No puppies}} \\ 6\hat{b}_1 &= \bar{X}_{\text{30 mins}} + \bar{X}_{\text{15 mins}} + \bar{X}_{\text{No puppies}} - 3\bar{X}_{\text{No puppies}} \\ &= \bar{X}_{\text{30 mins}} + \bar{X}_{\text{15 mins}} - 2\bar{X}_{\text{No puppies}} \end{aligned} \tag{12.24}$$

We then divide everything by 2 to reduce the equation to its simplest form:

$$\begin{aligned} 3\hat{b}_1 &= \left(\frac{\bar{X}_{\text{30 mins}} + \bar{X}_{\text{15 mins}}}{2}\right) - \bar{X}_{\text{No puppies}} \\ \hat{b}_1 &= \frac{1}{3}\left[\left(\frac{\bar{X}_{\text{30 mins}} + \bar{X}_{\text{15 mins}}}{2}\right) - \bar{X}_{\text{No Puppies}}\right] \end{aligned} \tag{12.25}$$

We planned contrast 1 to look at the difference between the average of the two puppy groups and the control, and a final rearrangement of the equation shows how $\hat{b}_1$ represents this difference:

$$\begin{aligned} 3\hat{b}_1 &= \left(\frac{\bar{X}_{30\text{ mins}} + \bar{X}_{15\text{ mins}}}{2}\right) - \bar{X}_{\text{No puppies}} \\ &= \frac{5 + 3.2}{2} - 2.2 \\ &= 1.9 \end{aligned} \tag{12.26}$$

Rather than being the true value of the difference between experimental and control groups, $\hat{b}_1$ is a third of this difference ($\hat{b}_1$ = 1.9/3 = 0.633) – it is divided by the number of groups in the contrast. Nevertheless, it is proportionate to the difference it set out to test.

Let's move on to look at someone in the 30-minute group. The predicted value of happiness for this person is the mean for the 30-minute group, and their group membership is coded using the values for the 30-minute group in Table 12.6. The resulting model is:

$$\begin{aligned} \widehat{\text{happiness}}_i &= \hat{b}_0 + \hat{b}_1\text{contrast } 1_i + \hat{b}_2\text{contrast } 2_i \\ \bar{X}_{30\text{ mins}} &= \hat{b}_0 + \left(\hat{b}_1 \times 1\right) + \left(\hat{b}_2 \times 1\right) \\ \hat{b}_2 &= \bar{X}_{30\text{ mins}} - \hat{b}_1 - \hat{b}_0 \end{aligned} \tag{12.27}$$

We know already what $\hat{b}_1$ and $\hat{b}_0$ represent, so we substitute these values into the equation:

$$\begin{aligned} \hat{b}_2 &= \bar{X}_{30\text{ mins}} - \hat{b}_1 - \hat{b}_0 \\ &= \bar{X}_{30\text{ mins}} - \frac{1}{3}\left[\left(\frac{\bar{X}_{30\text{ mins}} + \bar{X}_{15\text{ mins}}}{2}\right) - \bar{X}_{\text{No Puppies}}\right] \\ &\quad - \frac{\bar{X}_{30\text{ mins}} + \bar{X}_{15\text{ mins}} + \bar{X}_{\text{No puppies}}}{3} \\ 3\hat{b}_2 &= 3\bar{X}_{30\text{ mins}} - \left[\left(\frac{\bar{X}_{30\text{ mins}} + \bar{X}_{15\text{ mins}}}{2}\right) - \bar{X}_{\text{No puppies}}\right] \\ &\quad - \bar{X}_{30\text{ mins}} + \bar{X}_{15\text{ mins}} + \bar{X}_{\text{No puppies}} \end{aligned} \tag{12.28}$$

Next, multiply everything by 2 to get rid of the fractions. (This step is long-winded, but it makes the working easier to follow if your maths is a bit rusty.)

$$\begin{aligned} 6\hat{b}_2 &= 6\bar{X}_{30\text{ mins}} - \left(\bar{X}_{30\text{ mins}} + \bar{X}_{15\text{ mins}} - 2\bar{X}_{\text{No puppies}}\right) \\ &\quad - 2\left(\bar{X}_{30\text{ mins}} + \bar{X}_{15\text{ mins}} + \bar{X}_{\text{No puppies}}\right) \\ &= 6\bar{X}_{30\text{ mins}} - \bar{X}_{30\text{ mins}} - \bar{X}_{15\text{ mins}} + 2\bar{X}_{\text{No puppies}} - 2\bar{X}_{30\text{ mins}} \\ &\quad - 2\bar{X}_{15\text{ mins}} - 2\bar{X}_{\text{No puppies}} \\ &= 3\bar{X}_{30\text{ mins}} - 3\bar{X}_{15\text{ mins}} \end{aligned} \tag{12.29}$$

Finally, let's divide the equation by 6 to find out what $\hat{b}_2$ represents (remember that 3/6 = 1/2):

$$\hat{b}_2 = \frac{\bar{X}_{30\text{ mins}} - \bar{X}_{15\text{ mins}}}{2} \tag{12.30}$$

We planned contrast 2 to look at the difference between the experimental groups and $\hat{b}_2$ represents this difference.

$$\bar{X}_{30\text{mins}} - \bar{X}_{15\text{mins}} = 5 - 3.2 = 1.8 \tag{12.31}$$

Again, rather than being the value of the difference between the experimental groups, $\hat{b}_2$ is this difference divided by the number of groups in the contrast (1.8/2 = 0.9), but the key thing is what it represents, and it is proportional to the difference between experimental group means.

SELF TEST

To illustrate these principles, I have created a file called **puppies_contrast.sav** in which the puppy therapy data are coded using the contrast coding scheme used in this section. Fit a linear model using happiness as the outcome and **puppies_vs_none and short_vs_long** as the predictor variables (leave all default options).

Output 12.2 shows the result of the model from the self-test. The main ANOVA for the model is the same as when dummy coding was used (compare it to Output 12.1), showing that the model fit is the same (it should be because the model represents the group means and these have not changed); however, the $\hat{b}$-values have changed because the values of our dummy variables have changed. The first thing to notice is that the intercept is the grand mean, 3.467 (see, I wasn't telling lies). Second, the $\hat{b}$ for contrast 1 is one-third of the difference between the average of the experimental conditions and the control condition. Finally, the $\hat{b}$ for contrast 2 is half of the difference between the experimental groups (see above). The significance values of the *t*-statistics tell us that our puppy groups were significantly different from the control ($p = 0.029$) but that the 15- and 30-minutes of puppy therapy were not significantly different ($p = 0.065$).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.467	.362		9.574	<.001
	Puppy therapy vs. none (contrast 1)	1.900	.768	.525	2.474	.029
	30 mins vs 15 mins (contrast 2)	1.800	.887	.430	2.029	.065

a. Dependent Variable: Happiness (0-10)

Output 12.2



Cramming Sam's Tips
Planned contrasts

- If the *F* for the overall model is significant you need to find out which groups differ.
- When you have generated specific hypotheses before the experiment, use *planned contrasts*.
- Each contrast compares two 'chunks' of variance. (A chunk can contain one or more groups.)
- The first contrast will usually be experimental groups against control groups.
- The next contrast will be to take one of the chunks that contained more than one group (if there were any) and divide it into two chunks.

- You repeat this process: if there are any chunks in previous contrasts that contained more than one group that haven't already been broken down into smaller chunks, then create a new contrast that breaks them down into smaller chunks.
- Carry on creating contrasts until each group has appeared in a chunk on its own in one of your contrasts.
- The number of contrasts you end up with should be one less than the number of experimental conditions. If not, you've done it wrong.
- In each contrast assign a 'weight' to each group that is the value of the number of groups in the opposite chunk in that contrast.
- For a given contrast, randomly select one chunk, and for the groups in that chunk change their weights to be negative numbers.
- Breathe a sigh of relief.



12.5.3 Non-orthogonal contrasts Ⓑ



Contrasts don't have to be orthogonal: non-orthogonal contrasts are contrasts that are related. The best way to get them is to disobey rule 1 in the previous section. Using my cake analogy again, non-orthogonal contrasts are where you slice up your cake and then try to stick slices of cake together again. Standard dummy coding (Section 12.3.1) is an example of non-orthogonal contrasts because the baseline group is used in each contrast. For the puppy therapy data another set of non-orthogonal contrasts might be to have the same initial contrast (comparing experimental groups against the control), but then to compare the 30-minute group to the control. This disobeys rule 1 because the control group is singled out in the first contrast but used again in the second contrast. The coding for this set of contrasts is shown in Table 12.7, and the last column makes clear that when you multiply and add the codings from the two contrasts the sum is not zero. This tells us that the contrasts are not orthogonal.

There is nothing intrinsically wrong with non-orthogonal contrasts, but you must be careful about how you interpret them because the contrasts are related and so the resulting test statistics and *p*-values will be correlated to some degree. Basically, the Type I error rate isn't controlled so you should use a more conservative probability level to accept a given contrast as statistically significant (see Section 12.6).

Table 12.7 Non-orthogonal contrasts for the puppy therapy data

Group	Dummy Variable 1 (Contrast$_1$)	Dummy Variable 2 (Contrast$_2$)	Product Contrast$_1$× Contrast$_2$
Control	−2	−1	2
15 minutes	1	0	0
30 minutes	1	1	1
Total	0	0	3

12.5.4 Built-in contrasts Ⓑ

Although under most circumstances you will design your own contrasts, there are also 'off-the-shelf' contrasts that you can choose. Table 12.8 shows the built-in contrasts in IBM SPSS Statistics software for procedures such as logistic regression (see Chapter 20), factorial ANOVA and repeated-measures ANOVA (see Chapters 14 and 15). The exact codings are not provided in Table 12.8, but I give examples of the contrasts done in a three- and four-group situation (where the groups are labelled 1, 2, 3 and 1, 2, 3, 4, respectively). When you code categorical variables in the data editor, SPSS treats the lowest-value code as group 1, the next highest code as group 2, and so on. Therefore, depending on which contrasts you want, you should code your grouping variable appropriately (and then use Table 12.8 as a guide to which contrasts you'll get). Some contrasts in Table 12.8 are orthogonal (i.e., Helmert and difference contrasts) while others are non-orthogonal (deviation, simple and repeated). You might also notice that simple contrasts look the same as those given by using the dummy variable coding described in Table 12.4 but there is a subtle difference. For standard dummy coding the intercept is the mean of the reference group (i.e. the no puppy control) whereas for simple contrasts the intercept represents the grand mean for the outcome.

Table 12.8 Standard contrasts available in SPSS

Name	Definition	Contrast	Three Groups	Four Groups
Deviation (first)	Compares the effect of each category (except the first) to the overall experimental effect	1	2 vs. (1,2,3)	2 vs. (1,2,3,4)
		2	3 vs. (1,2,3)	3 vs. (1,2,3,4)
		3		4 vs. (1,2,3,4)
Deviation (last)	Compares the effect of each category (except the last) to the overall experimental effect	1	1 vs. (1,2,3)	1 vs. (1,2,3,4)
		2	2 vs. (1,2,3)	2 vs. (1,2,3,4)
		3		3 vs. (1,2,3,4)
Simple (first)	Each category is compared to the first category	1	1 vs. 2	1 vs. 2
		2	1 vs. 3	1 vs. 3
		3		1 vs. 4
Simple (last)	Each category is compared to the last category	1	1 vs. 3	1 vs. 4
		2	2 vs. 3	2 vs. 4
		3		3 vs. 4
Repeated	Each category (except the first) is compared to the previous category	1	1 vs. 2	1 vs. 2
		2	2 vs. 3	2 vs. 3
		3		3 vs. 4
Helmert	Each category (except the last) is compared to the mean effect of all subsequent categories	1	1 vs. (2, 3)	1 vs. (2, 3, 4)
		2	2 vs. 3	2 vs. (3, 4)
		3		3 vs. 4
Difference (reverse Helmert)	Each category (except the first) is compared to the mean effect of all previous categories	1	3 vs. (2, 1)	4 vs. (3, 2, 1)
		2	2 vs. 1	3 vs. (2, 1)
		3		2 vs. 1

12.5.5 Polynomial contrasts: trend analysis Ⓑ

The **polynomial contrast** is deliberately omitted from Table 12.8. This contrast tests for trends in the data, and in its most basic form it looks for a linear trend (i.e., that the group means increase proportionately). However, there are other trends such as quadratic, cubic and quartic trends that

can be examined. Figure 12.11 shows examples of the types of trend that can exist in data sets. The *linear* trend should be familiar to you and represents a simple proportionate change in the value of the dependent variable across ordered categories (the diagram shows a positive linear trend but of course it could be negative). A **quadratic trend** is where there is a curve in the line (the curve can be subtler than in the figure and in the opposite direction). An example of this is a situation in which a drug enhances performance on a task at first but then as the dose increases the performance tails off or drops. To find a quadratic trend you need at least three groups because with two groups the means of the dependent variable can't be connected by anything other than a straight line. A **cubic trend** is where there are two changes in the direction of the trend. So, for example, the mean of the dependent variable at first goes up across categories of the independent variable, then across the succeeding categories the means go down, but then across the last few categories the means rise again. To have two changes in the direction of the mean you must have at least four categories of the independent variable. The final trend that you are likely to come across is the **quartic trend**, and this trend has three changes of direction (so you need at least five categories of the independent variable).

Polynomial trends should be examined in data sets in which it makes sense to order the categories of the independent variable (so, for example, if you have administered five doses of a drug it makes sense to examine the five doses in order of magnitude). For the puppy therapy data there are three groups, and so we can expect to find only a linear or quadratic trend (it would be pointless to test for any higher-order trends).

Each of these trends has a set of codes for the dummy variables in the model, so we are doing the same thing that we did for planned contrasts except that the codings have already been devised to represent the type of trend of interest. In fact, the plots in Figure 12.11 have been constructed by plotting the code values for the five groups. If you add the codes for a given trend the sum will equal zero, and if you multiply the codes you will find that the sum of the products also equals zero. Hence, these contrasts are orthogonal.

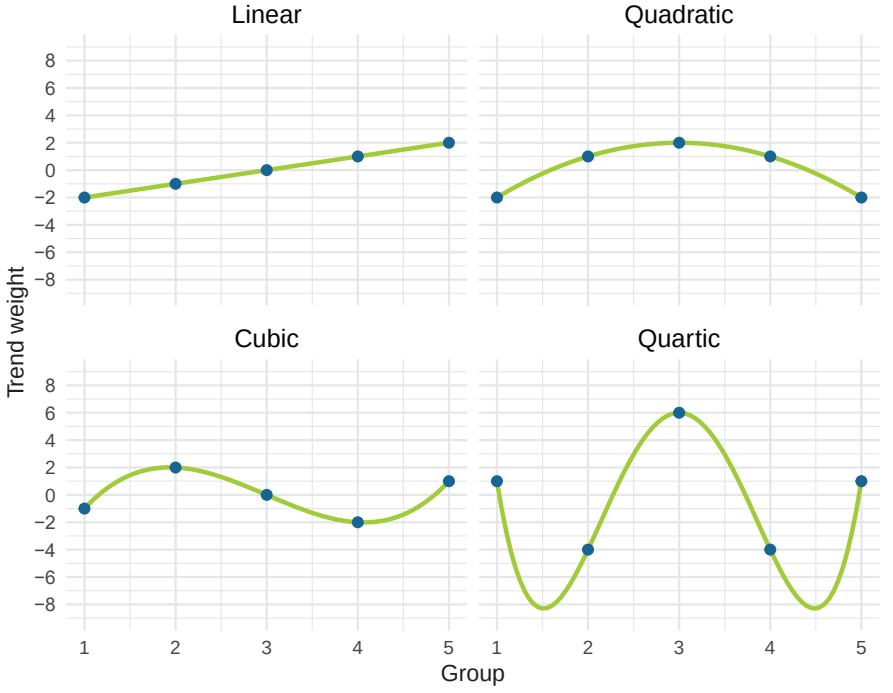


Figure 12.11 Examples of linear, quadratic, cubic and quartic trends across five groups

12.6 *Post hoc* procedures Ⓑ

Often people have no specific *a priori* predictions about the data they have collected and instead they rummage around the data looking for any differences between means that they can find. It's a bit like *p*-hacking (Section 3.3.3), except that you adjust the *p* to make it harder to find significant differences. *Post hoc* tests consist of **pairwise comparisons** that are designed to compare all different combinations of the treatment groups. So, it is taking every pair of groups and performing a separate test on each. Now, this might seem like a particularly stupid thing to do in light of what I have already told you about the problems of inflated familywise error rates (Section 2.10.8). However, pairwise comparisons control the familywise error by correcting the level of significance for each test such that the overall Type I error rate (α) across all comparisons remains at 0.05. There are several ways in which the familywise error rate can be controlled, and we have already discussed one of the most popular ones: the Bonferroni correction (Section 2.10.8).

There are other methods too (SPSS does about 18 different ones). Although I would love to go into the tedious detail of all 18 procedures, there really is little point. One reason is that there are excellent texts already available for those who wish to know (Klockars & Sax, 1986; Toothaker, 1993), but the main one is that to explain them I would have to learn about them first, and I may be a nerd, but even I draw the line at reading up on 18 different *post hoc* tests. However, it *is* important that you know which *post hoc* tests perform best according to three important criteria. First, does the test control the Type I error rate? Second, does the test control the Type II error rate (i.e., does the test have good statistical power)? Third, is the test robust?

12.6.1 Type I (α) and Type II error rates for *post hoc* tests Ⓑ

The Type I error rate and the statistical power of a test are linked. Therefore, there is always a trade-off: if a test is conservative (the probability of a Type I error is small) then it is likely to lack statistical power (the probability of a Type II error will be high). Therefore, it is important that multiple comparison procedures control the Type I error rate but without a substantial loss in power. If a test is too conservative then we are likely to reject differences between means that are, in reality, meaningful.

The *least-significant difference* (LSD) pairwise comparison makes no attempt to control the Type I error and is equivalent to performing multiple tests on the data. The only difference is that the LSD requires the overall ANOVA to be significant. The *Studentized Newman–Keuls* procedure is also a very liberal test and lacks control over the familywise error rate. *Bonferroni's* and *Tukey's t*ests both control the Type I error rate very well but are conservative (they lack statistical power). Of the two, Bonferroni has more power when the number of comparisons is small, whereas Tukey is more powerful when testing large numbers of means. Tukey generally has greater power than *Dunn* and *Scheffé*. The *Ryan, Einot, Gabriel and Welsch Q* procedure (REGWQ) has good power and tight control of the Type I error rate. In fact, when you want to test all pairs of means this procedure is probably the best. However, when group sizes are different this procedure should not be used.

12.6.2 Are *post hoc* procedures robust? Ⓑ

Most research on *post hoc* tests has looked at whether the test performs well when the group sizes are different (an unbalanced design), when the population variances are very different, and when data are not normally distributed. The good news is that most multiple comparison procedures perform relatively well under small deviations from normality. The bad news is that they perform badly when group sizes are unequal and when population variances are different.

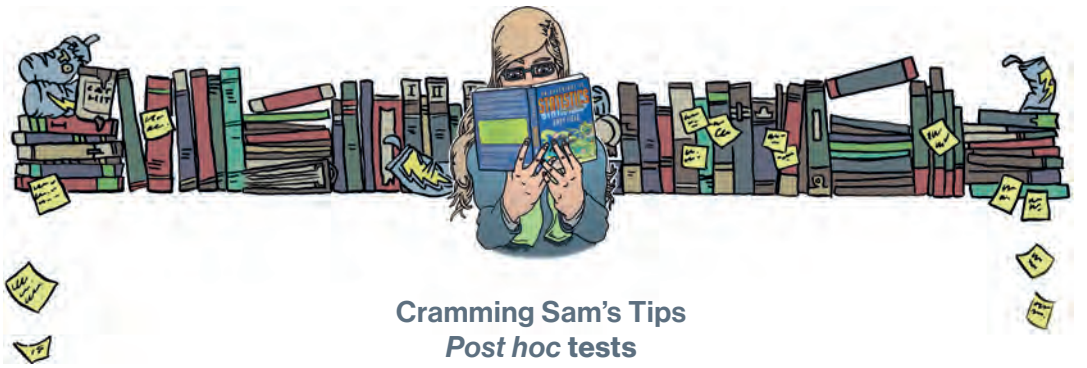
Hochberg's GT2 and *Gabriel's* pairwise test procedure were designed to cope with situations in which sample sizes are different. Gabriel's procedure is generally more powerful but can become too liberal when the sample sizes are very different. Also, Hochberg's GT2 is very unreliable when the

population variances are different, and so should be used only when you are sure that this is not the case. There are several multiple comparison procedures that have been specially designed for situations in which population variances differ. SPSS provides four options for this situation: *Tamhane's T2*, *Dunnett's T3*, *Games–Howell* and *Dunnett's C*. Tamhane's T2 is conservative and Dunnett's T3 and C keep very tight Type I error control. The Games–Howell procedure is the most powerful but can be liberal when sample sizes are small. However, Games–Howell is also accurate when sample sizes are unequal.

12.6.3 Summary of *post hoc* procedures Ⓑ

The choice of comparison procedure will depend on the exact situation you have and whether it is more important for you to keep strict control over the familywise error rate or to have greater statistical power. However, some general guidelines can be drawn (Toothaker, 1993). When you have equal sample sizes and you are confident that your population variances are similar then use REGWQ or Tukey as both have good power and tight control over the Type I error rate. Bonferroni is generally conservative, but if you want guaranteed control over the Type I error rate then this is the test to use. If sample sizes are slightly different then use Gabriel's procedure because it has greater power, but if sample sizes are very different use Hochberg's GT2. If there is any doubt that the population variances are equal then use the Games–Howell procedure because this generally seems to offer the best performance. I recommend running the Games–Howell procedure in addition to any other tests you might select because of the uncertainty of knowing whether the population variances are equivalent.

Although these general guidelines provide a convention to follow, be aware of the other procedures available and when they might be useful to use (e.g., Dunnett's test is the only multiple comparison that allows you to test means against a control mean).



Cramming Sam's Tips
Post hoc tests

- When you have no specific hypotheses before the experiment, follow up the model with *post hoc tests*.
- When you have equal sample sizes and group variances are similar, use REGWQ or Tukey.
- If you want guaranteed control over the Type I error rate then use Bonferroni.
- If sample sizes are slightly different then use Gabriel's, but if sample sizes are very different use Hochberg's GT2.
- If there is any doubt that group variances are equal then use the Games–Howell procedure.



12.7 Effect sizes when comparing means Ⓑ

We saw in Eq. 9.17 that we can compute R^2 for a linear model:

$$R^2 = \frac{SS_M}{SS_T} \tag{12.32}$$

We have just computed these values, so we can calculate R^2 using SS_M (Eq. 12.14) and SS_T (Eq. 12.12) but in the context of comparing means we refer to it as **eta squared**, $\hat{\eta}^2$

$$\hat{\eta}^2 = R^2 = \frac{SS_M}{SS_T} = \frac{20.135}{43.736} = 0.46 \tag{12.33}$$

Other things being equal, the effect of puppy therapy on happiness is a substantive finding. However, eta squared is slightly biased because it is based purely on sums of squares from the sample and no adjustment is made for the fact that we're trying to estimate the effect size in the population. Therefore, we often use a slightly more complex measure called **omega squared** ($\hat{\omega}^2$). This effect size estimate is still based on the sums of squares that we've met in this chapter – it uses the variance explained by the model, and the average error variance:

$$\hat{\omega}^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R} \tag{12.34}$$

The df_M and MS_R in the equation are values we encountered in Eq. 12.18. In this example we'd get:

$$\begin{aligned}\hat{\omega}^2 &= \frac{20.135 - (2)1.97}{43.736 + 1.97} \\ &= \frac{16.195}{45.706} \\ &= 0.354\end{aligned} \tag{12.35}$$

This adjustment has led to a slightly lower estimate, and in general $\hat{\omega}^2$ is a more accurate measure. It has been suggested that values of 0.01, 0.06 and 0.14 represent small, medium and large effects, respectively (Kirk, 1996). Remember, though, that these are rough guidelines and that effect sizes need to be interpreted within the context of the research literature. It is possible to construct confidence intervals for these effect size measures (see Steiger, 2004) and SPSS will do this for us.

Most of the time it isn't that interesting to have effect sizes for the overall effect because it's testing a general hypothesis (means are different). Instead, we really want effect sizes for the contrasts (because these compare only two things, so the effect size is easier to interpret) and we have a lovely raw effect size in the form of the $\hat{b}$ associated with the contrast. This is directly interpretable. If you have used *post hoc* tests you can also, of course, compute Cohen's *d* for each pair of means.

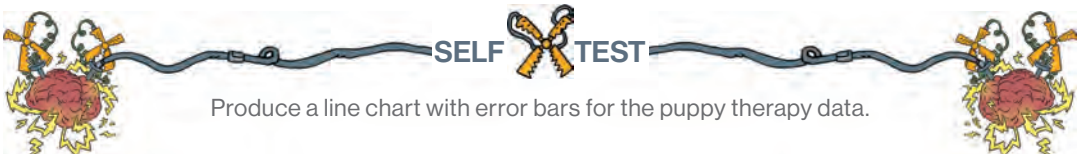
12.8 Comparing several means using SPSS Ⓑ

Because to compare means we fit a variant of the linear model, we could set up contrast dummy variables and run the analysis using the linear regression menus from Chapter 9. However, SPSS has a designated menu for situations where you're predicting an outcome from several group means (historically known as a one-way independent analysis of variance, ANOVA). Using this designated menu has the advantage that you don't have to create dummy variables manually, as we shall see.

12.8.1 General procedure of one-way ANOVA Ⓑ

The data are in the file **puppies.sav** (although it's good practice to enter them yourself). They are entered in three columns. The first contains the participant **id**, the second (**dose**) tells us how much puppy therapy the participant was given by coding the treatment group to which they were assigned (1 = control, 2 = 15 minutes and 3 = 30 minutes), and the third (**happiness**) contains the person's happiness score.

Because one-way **independent ANOVA** is a linear model with a different label attached, look back at the general procedure for linear models in Chapter 9 Figure 12.12 highlights the steps that are specific to this version of the linear model. As with any analysis, begin by plotting the data and looking for and correcting sources of bias.



Noting my earlier comments about homogeneity of variance and normality (Section 6.9), there is a case for always proceeding with a robust test and I recommend using Welch's *F*.

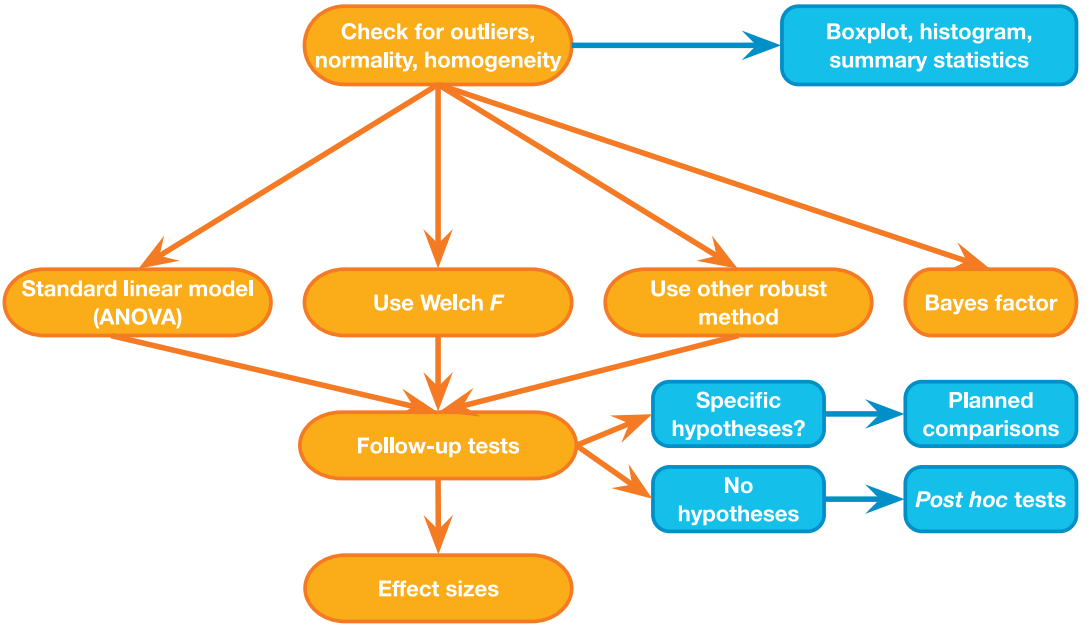


Figure 12.12 Overview of the general procedure for one-way ANOVA

To fit the model select *Analyze* > *Compare Means* > One-Way ANOVA... to access the main dialog in Figure 12.13. There is an empty box in which to place one or more dependent variables and another to specify a categorical predictor or *factor*. 'Factor' is another term for independent variable or categorical predictor and not to be confused with a very different type of factor that we learn about in Chapter 18. For the puppy therapy data drag **happiness** from the variables list into the box labelled *Dependent List* (or click ▸). Then drag the grouping variable **dose** to the box labelled *Factor* (*or* click ▸). Select ☑ Estimate effect size for overall tests to get an estimate of the overall effect size for the model.

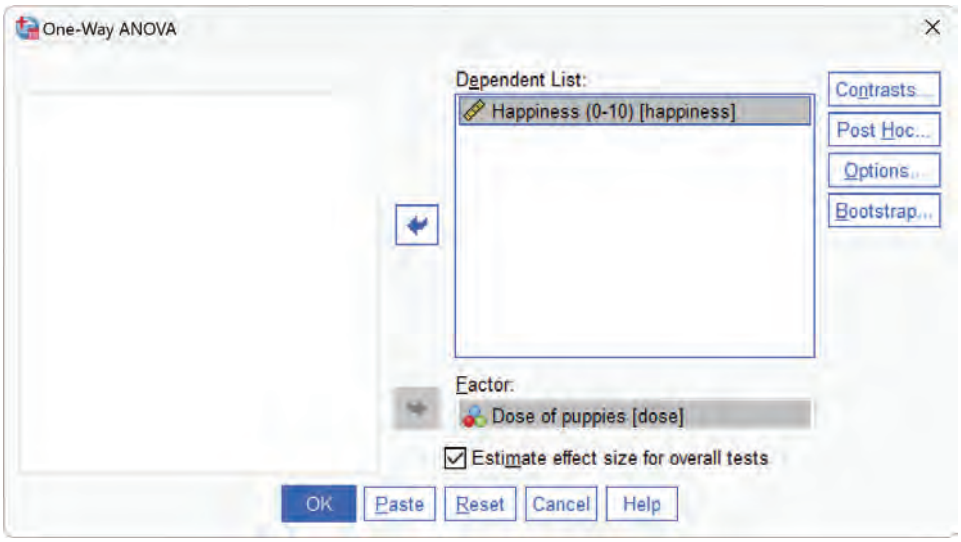


Figure 12.13 Main dialog box for one-way ANOVA

One thing that I dislike about SPSS is that in various procedures, such as this one, you are encouraged to carry out multiple tests (in this case by allowing you to specify several outcome variables in the *Dependent List*). This is not a good thing because you lose control of the Type I error rate. If you had measured several dependent variables (e.g., you had measured not just happiness but other indicators of stress such as cortisol levels, non-verbal behaviour and heart rate) it would be preferable to analyse these data using MANOVA (Chapter 17) or think about ways to combine the measures into a single index of whatever it is you're trying to measure rather than treating each outcome measure separately.

12.8.2 Planned contrasts and trend analysis Ⓑ

Click Contrasts... to access a dialog box for specifying the contrasts described in Section 12.5. The dialog box (Figure 12.14) has three sections. The first section is for specifying trend analyses. If you want to test for trends select ☑ Polynomial. Once this option is active you can select the degree of polynomial you would like from the drop-down list (Linear). The puppy therapy data have only three groups and so the highest degree of trend there can be is quadratic (see Section 12.5.5). It is important from the point of view of trend analysis that we have coded the grouping variable in a meaningful order. We expect happiness to be smallest in the control group, to increase in the 15-minute group and then to increase again in the 30-minute group, so to detect a meaningful trend, these groups must be coded in ascending order. We have done this by coding the control group with the lowest value 1, the 15-minute group with the middle value 2 and the 30-minute group with the highest value 3. If we coded the groups differently, this would influence both whether a trend is detected and, if one is detected, whether it has a meaningful interpretation. For the puppy therapy data select ☑ Polynomial and Quadratic . SPSS will test for the trend requested and all lower-order trends, so with *Quadratic* selected we'll get tests both for a linear and a quadratic trend.

The middle section of the dialog box in Figure 12.14 is for specifying weights for planned contrasts. We went through the process of generating weights for the contrasts we want in Section 12.5.2. The weights for contrast 1 were –2 (control group), +1 (15-minute group) and +1 (30-minute group). We will specify this contrast first. It is important to make sure that you enter the correct weight for each group; the first weight that you enter is the weight for the *first* group (i.e., the group

coded with the lowest value in the data editor). For the puppy therapy data, the group coded with the lowest value was the control group (which had a code of 1) so we enter the weight for this group first. Click in the box labelled *Coefficients*, type '–2' and click Add. Next, we input the weight for the second group, which for the puppy therapy data is the 15-minute group (because this group was coded in the data editor with the second-highest value). Click in the box labelled *Coefficients*, type '1' and click Add. Finally, we input the weight for the last group, which for the puppy therapy data is the 30-minute group (because this group was coded with the highest value in the data editor). Click in the box labelled *Coefficients*, type '1' and click Add. The dialog box should now look like Figure 12.14 (middle left).

Once the weights are assigned you can change or remove any one of them by selecting the weight that you want to change. The weight will appear in the box labelled *Coefficients* where you can type a new weight and then click Change. Alternatively, you can click a weight and remove it completely by selecting Remove. Underneath the weights the sum of weights is displayed, which, as we saw in Section 12.5.2, should equal zero. If the *Coefficient Total* is anything other than zero you should go back and check that the contrasts you have planned make sense and that you have followed the appropriate rules for assigning weights.

Once you have specified the first contrast, click Next. The weights that you have just entered will disappear and the dialog box will now read *Contrast 2 of 2*. We know from Section 12.5.2 that the weights for contrast 2 were: 0 (control group), –1 (15-minute group) and +1 (30-minute group). We specify this contrast as before. Remembering that the first weight we enter will be for the control group, we enter the value 0 as the first weight. Click in the box labelled *Coefficients*, then type '0' and click Add. Next, we input the weight for the 15-minute group by clicking in the box labelled *Coefficients*, typing '–1' and clicking Add. Finally, we input the weight for the 30-minute group by clicking in the box labelled *Coefficients*, typing '+1' and clicking Add. The dialog box should now look like Figure 12.14 (bottom right). Notice that the weights add up to zero as they did for contrast 1. You must remember to input zero weights for any groups that are not in the contrast.

The final section of the dialog box has an option to get effect sizes for our contrasts, which is a good idea. So select ☑ Estimate effect size for contrasts. There are two options for computing these effects sizes. If your groups have equal variances then use the default of standardizing the effect using all group standard deviations, but if your groups have unequal variances select to standardize using only the standard deviations of the groups involved in the contrast. When you have finished, click Continue to return to the main dialog box.

12.8.3 *Post hoc* tests in SPSS Ⓑ

In theory, if we have done planned contrasts we wouldn't do *post hoc* tests (because we have already tested the hypotheses of interest). Likewise, if we choose to conduct *post hoc* tests then we would not do planned contrasts (because we have no specific hypotheses to test). For the puppy example, contrasts make most sense because we have specific hypotheses, but because you might one day want to use *post hoc* tests we will conduct *post hoc* tests but purely as an educational exercise. Click Post Hoc... in the main dialog box to access the dialog box in Figure 12.15.

In Section 12.6.3, I recommended various *post hoc* procedures for particular situations. For the puppy data there are equal sample sizes and so we need not use Gabriel's test. A reasonable course of action would be to use ☑ Tukey and ☑ R-E-G-W Q but check the findings with the ☑ Games-Howell procedure. It's also worth mentioning that because we have a specific hypothesis that both the 30- and 15-minute groups should differ from the control group we could use Dunnett's test to examine these hypotheses. To do this, select ☑ Dunnett and change the control category from Last to First so that the no puppies category (which, remember, we coded with the lowest value and so is the first group) is used as the comparison group. You can choose whether to conduct a two-tailed (◉ 2-sided) or a one-tailed test. If you choose a one-tailed test (which I advised against in Section 2.10.6) then you need to tell SPSS

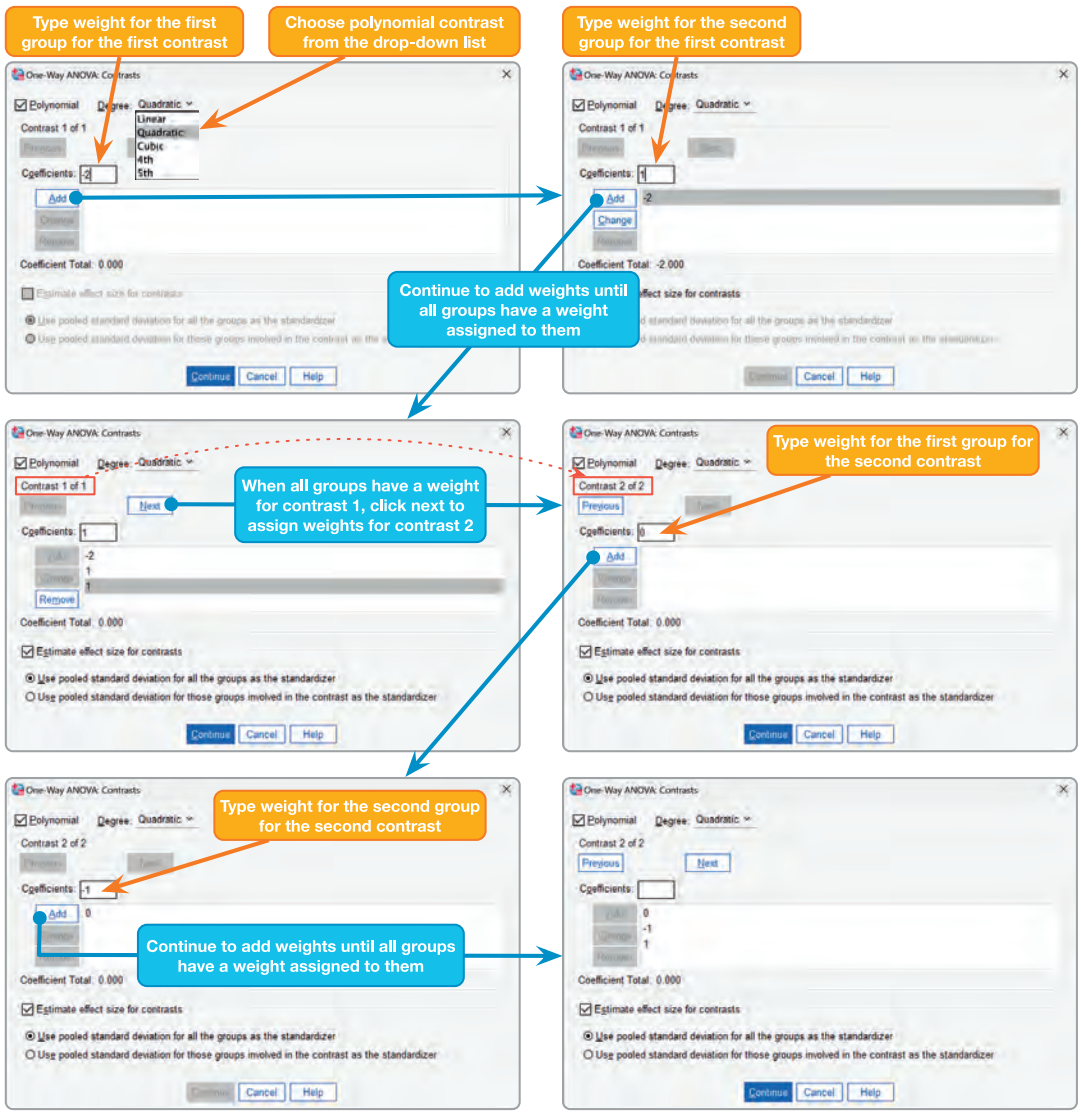


Figure 12.14 Setting up planned contrasts

whether you believe that the mean of each experimental group will be greater than the first category (i.e., no puppies) using ◉ > Control, or less than the first category by selecting ◉ < Control. These are all the *post hoc* tests that we need to specify (Figure 12.15).

By default SPSS will use the same alpha for *post hoc* tests as for all tests in this model (you can specify this significance level in the options described in the next section). However, you can override this behaviour by specifying a different alpha for *post hoc* tests (perhaps a more conservative one). Click Continue to return to the main dialog box.

12.8.4 Options Ⓑ

Click Options... to access the dialog box in Figure 12.16. You can ask for descriptive statistics, which will produce a table of the means, standard deviations, standard errors, ranges and confidence intervals within each group. This information will help us to interpret the results but ideally you'd have done

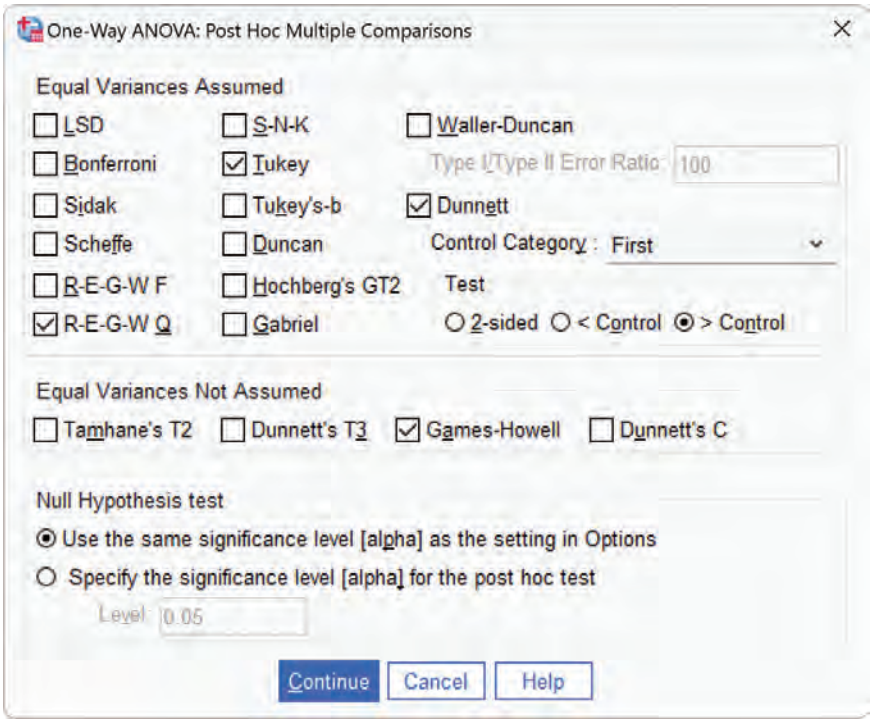


Figure 12.15 Dialog box for specifying *post hoc* tests

a more thorough exploration of your data before you get to this point. Leave ☐ Homogeneity of variance test unchecked because we don't want to use Levene's test (Section 6.9.6). Instead, select ☑ Brown-Forsythe test or ☑ Welch test to get versions of the *F*-statistic that are robust to heterogeneity in group variances. There is an option to produce a *Means plot* (a line plot of the group means), but the resulting plot is a leprous tramp compared to what you can create using the Chart Builder and it's best to plot your data *before* the analysis, not during it. We can also specify whether we want to exclude cases on a listwise basis or on a per analysis basis (see SPSS Tip 6.1 for an explanation). This option is useful only if you are fitting models to several outcome variables simultaneously (which hopefully you're not – ever). Finally, we can specify the alpha level for confidence intervals (by default this is set to 0.95, which equates to 95% confidence intervals and an alpha of 0.05). Ideally you'd set an idiosyncratic alpha level (Maier & Lakens, 2022) or possibly set this as 0.995 in line with the recommendation to use $\alpha = 0.005$ (Benjamin et al., 2018). However, we'll leave the default as it is.

12.8.5 Bootstrapping Ⓑ

Also in the main dialog box is the alluring Bootstrap... button. We know that bootstrapping is a good way to overcome bias, and this button glistens and tempts us with the promise of untold riches, like a diamond in a bull's rectum. However, if you use bootstrapping it'll be as disappointing as if you reached for that diamond only to discover that it's a piece of Lego. You might, not unreasonably, think that if you select bootstrapping you'd get a nice bootstrap of the *F*-statistic. You won't. You will get bootstrap confidence intervals around the means (if you ask for descriptive statistics), contrasts and *post hoc* tests. All of which are useful, but the main test won't be bootstrapped. For this example, we have a very small data set so bootstrapping is going to go haywire, so we won't select it. Click OK in the main dialog box to run the analysis.

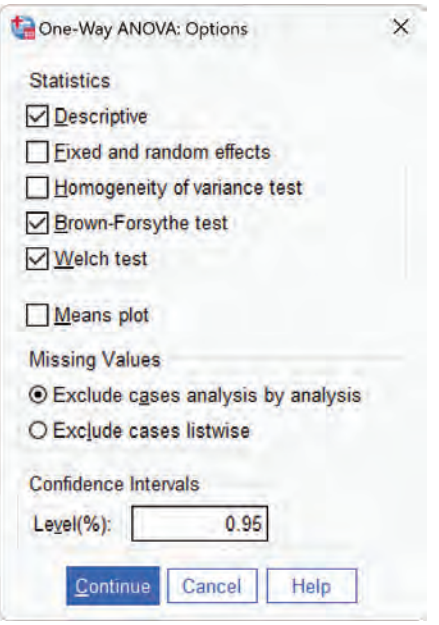


Figure 12.16 Options for one-way ANOVA



Oditi's Lantern
One-Way ANOVA

'I, Oditi, have made great progress in unearthing the hidden truth behind the numbers. This morning, one of my loyal followers reported to me that, based on an ANOVA he'd done, all dogs are controlled by cats who hide small remote controls up their rectums and manipulate them with their tongues. Every time you see a cat 'cleaning' itself, there will be a dog nearby chasing its tail. Listen carefully and you can hear the cat laughing to itself. Be warned, cats are merely piloting the technology, and soon they will control us too, turning us into heated chairs and food vendors. We must find out more. Stare into my lantern so that you too can use ANOVA.'



12.9 Output from one-way independent ANOVA Ⓑ

You should find that the output looks the same as what follows. If not, we should panic because one of us has done it wrong – hopefully not me or a lot of trees have died for nothing. Figure 12.17 shows a line chart with error bars from the self-test earlier in the chapter (I have edited my plot; see if you can make yours look like mine). All the error bars overlap a fair bit, indicating that, at face value, there might not be between-group differences (see Section 2.10.10). The line that joins the means seems to indicate a linear trend in that, as the dose of puppy therapy increases, so does the mean level of happiness.

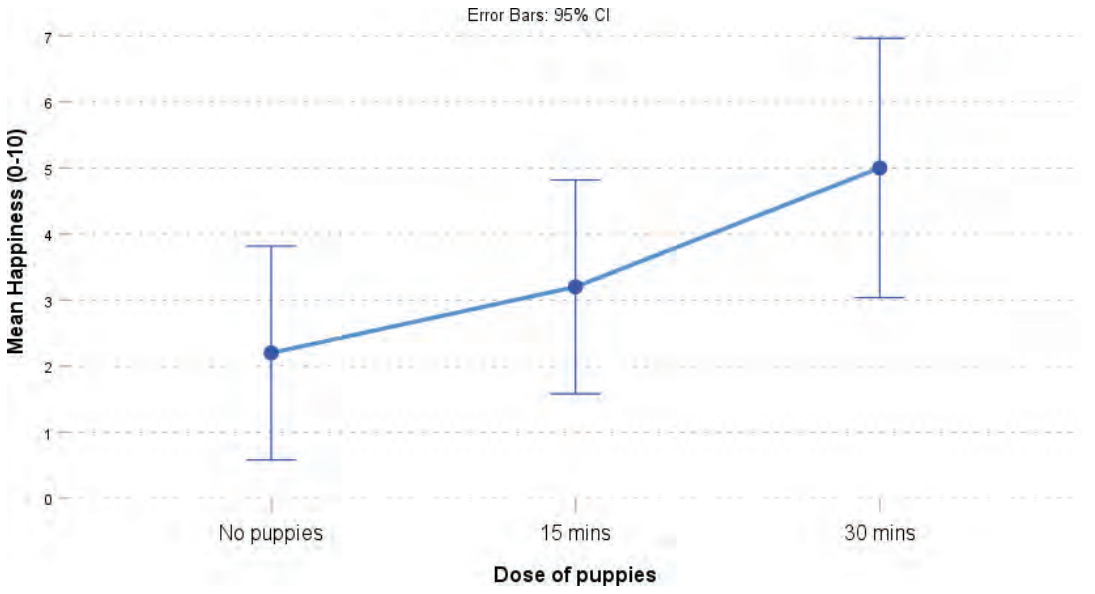


Figure 12.17 Error bar (95% CI) chart of the puppy therapy data

12.9.1 Output for the main analysis Ⓑ

Output 12.3 shows the table of descriptive statistics for the puppy data. The means and standard deviations correspond to those shown in Table 12.1, which is reassuring. We are also given the standard error and confidence intervals for the mean. If this sample is one of the 95% that have confidence intervals containing the true value then the true value of the mean is between 0.58 and 3.82 for the control group. Note that the standard deviations of the control and 15-minute groups are identical, which means the variances are too. The standard deviation of the 30-minute group is also similar to the other groups. We will refer to this table as we wade through the output.

Descriptives

Happiness (0-10)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
No puppies	5	2.20	1.304	.583	.58	3.82	1	4
15 mins	5	3.20	1.304	.583	1.58	4.82	2	5
30 mins	5	5.00	1.581	.707	3.04	6.96	3	7
Total	15	3.47	1.767	.456	2.49	4.45	1	7

Output 12.3

Output 12.4 shows the main ANOVA summary table. The table is divided into between-group effects (effects due to the model – the experimental effect) and within-group effects (this is the unsystematic variation in the data). The between-group effect is further broken down into a linear and quadratic component as requested in Figure 12.14. The between-group effect labelled *Combined* is the overall experimental effect, or, put another way, the improvement in the prediction of happiness scores resulting from using the group means. We are told the model sums of squares (SS_M = 20.13), the degrees of freedom are 2 and the mean squared error 10.067. These values correspond to those we calculated in Eq. 12.14 and Eq. 12.18.

The row labelled *Within groups* gives details of the unsystematic variation within the data (the variation due to natural individual differences in happiness and different reactions to puppy therapy). The table tells us that the residual sum of squares (SS_R) is 23.60, as calculated in Eq. 12.17. The average amount of unsystematic variation, the mean squares (MS_R), is 1.967, as calculated in Eq. 12.18.

The test of whether the group means are the same is represented by the *F*-statistic for the combined between-group effect. This statistic tests whether predicting **happiness** from **group** means significantly improves the fit of the model. The value of *F* is 5.12 (Eq. 12.20). The final column labelled *Sig.* tells us the probability of getting an *F* at least this big if there wasn't a difference between means in the population (see also SPSS Tip 12.1). In this case, there is a probability of 0.025 that an *F*-statistic of at least this size would occur if in reality the effect was zero. Assuming we set a cut-off point of 0.05 as a criterion for statistical significance before collecting data, most scientists would take the fact that 0.025 is less than the criterion of 0.05 as support for a significant effect of puppy therapy. At this stage we still do not know exactly what the effect of puppy therapy was (we don't know which groups differed). One interesting point is that we obtained a significant experimental effect yet the error bar plot suggested that no significant difference would be found. This contradiction illustrates how the error bar chart can act only as a rough guide to the data.

ANOVA

Happiness (0–10)

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		20.133	2	10.067	5.119	.025
	Linear Term	Contrast	19.600	1	19.600	9.966	.008
		Deviation	.533	1	.533	.271	.612
	Quadratic Term	Contrast	.533	1	.533	.271	.612
Within Groups			23.600	12	1.967		
Total			43.733	14			

Output 12.4

Knowing that the overall effect of puppy therapy was significant, we can look at the trend analysis. First, let's look at the linear component. This contrast tests whether the means increase across groups in a linear way. For the linear trend the *F*-statistic is 9.97 and this value is significant at p = 0.008. Therefore, we can say that as the dose of puppy therapy increased from nothing to 15 minutes to 30 minutes, happiness increased proportionately. The quadratic trend tests whether the pattern of means is curvilinear (i.e., is represented by a curve that has one bend). The error bar plot of the data suggests that the means cannot be represented by a curve and the results for the quadratic trend bear this conclusion out. The *F*-statistic for the quadratic trend is non-significant (in fact, the value of *F* is less than 1, which immediately indicates that this contrast will not be significant).

Output 12.5 shows the Welch and the Brown–Forsythe *F*-statistics. If you're interested in how these values are calculated then look at Jane Superbrain Box 12.2, but to be honest you're better off avoiding the mental anguish by looking at the values in Output 12.6 and trusting that they do what they're supposed to do (note that the error degrees of freedom have been adjusted and you should remember

this when you report the values). Based on whether the observed p is less than or equal to 0.05, the Welch F yields a non-significant result ($p = 0.054$) whereas Brown–Forsythe is significant ($p = 0.026$). This is confusing, but only if you like to imbue the 0.05 threshold with magical powers and engage in black-and-white thinking of the sort that people who use p-values so often do.

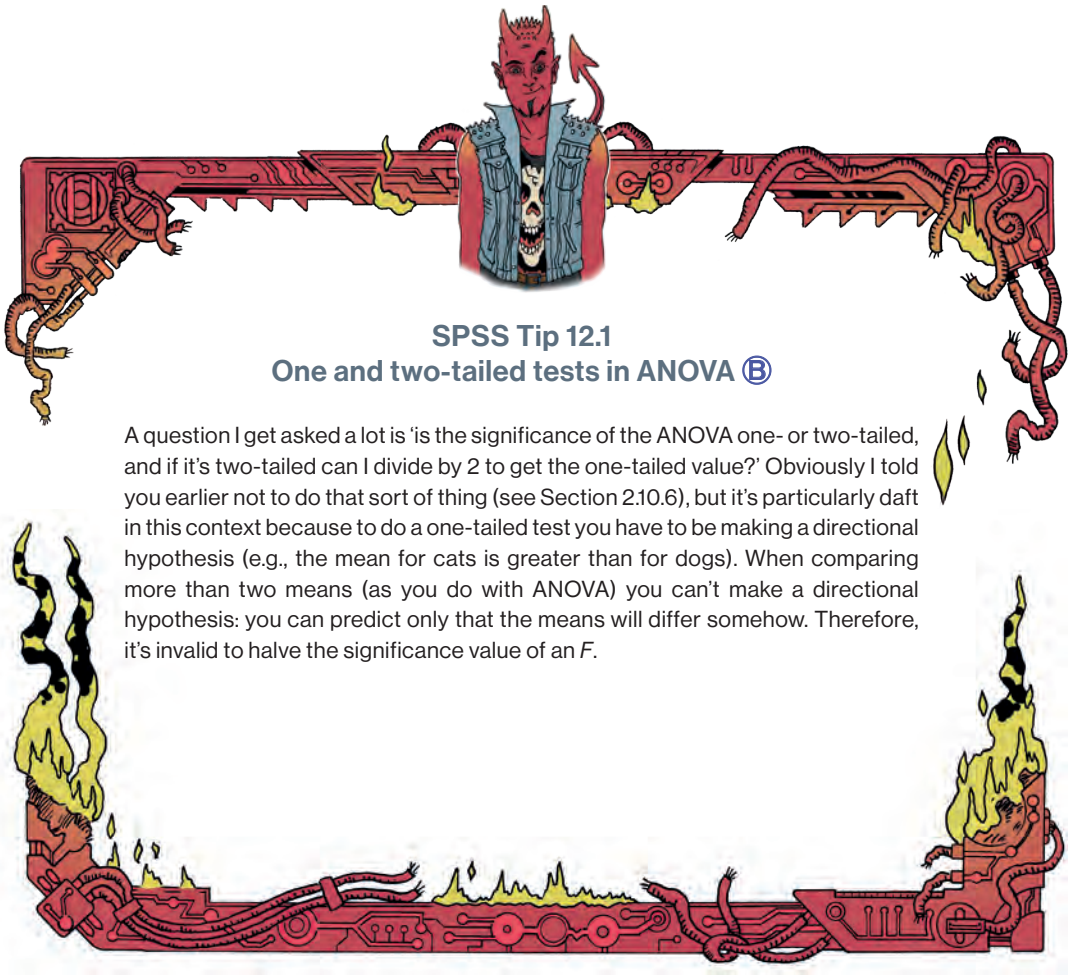
Robust Tests of Equality of Means

Happiness (0–10)

	Statistic[a]	df1	df2	Sig.
Welch	4.320	2	7.943	.054
Brown-Forsythe	5.119	2	11.574	.026

a. Asymptotically F distributed.

Output 12.5



SPSS Tip 12.1
One and two-tailed tests in ANOVA Ⓑ

A question I get asked a lot is 'is the significance of the ANOVA one- or two-tailed, and if it's two-tailed can I divide by 2 to get the one-tailed value?' Obviously I told you earlier not to do that sort of thing (see Section 2.10.6), but it's particularly daft in this context because to do a one-tailed test you have to be making a directional hypothesis (e.g., the mean for cats is greater than for dogs). When comparing more than two means (as you do with ANOVA) you can't make a directional hypothesis: you can predict only that the means will differ somehow. Therefore, it's invalid to halve the significance value of an F.

Output 12.6 shows the overall effect sizes for the model. The *Point Estimate* of eta squared (η^2) matches the calculation in Eq. 12.33 and the one for *omega-squared Fixed-effect* (ω^2) matches the calculation in Eq. 12.35. In both cases we but we also get the corresponding 95% confidence interval. Overall, it looks like puppy therapy has a fairly substantive effect on happiness. One interesting oddity is that

the lower bound of confidence interval for ω^2 is negative. We characterized ω^2 as the proportion of variance, so it should vary between 0 and 1. A negative value is impossible. This fact is acknowledged by the footnote to the table, which says *negative but less biased estimates are retained, not rounded to zero*. In other words, although negative values are impossible they are retained rather than rounding them to the nearest possible value of 0. All of the confidence intervals are very wide. In the case of ω^2, if we assume our sample is one of the 95% that have intervals containing the population values, the population effect could lie anywhere between –0.15 (or, in practice, 0 or no effect) and 0.59 (a massive effect). These wide intervals are because the sample size is so small that the intervals lack precision, but also alert us to the need for a larger study that can produce a more informative model.

ANOVA Effect Sizes[a,b]

			95% Confidence Interval	
		Point Estimate	Lower	Upper
Happiness (0–10)	Eta-squared	.460	.001	.658
	Epsilon-squared	.370	-.166	.601
	Omega-squared Fixed-effect	.354	-.153	.585
	Omega-squared Random-effect	.215	-.071	.413

a. Eta-squared and Epsilon-squared are estimated based on the fixed-effect model.

b. Negative but less biased estimates are retained, not rounded to zero.

Output 12.6

12.9.2 Output for planned contrasts Ⓑ

In Section 12.8.2 we defined two planned contrasts: one to test whether the control group was different from the two groups which received puppy therapy, and one to see whether the two doses of puppy therapy made a difference to happiness. Output 12.7 shows the results of these contrasts. The first table displays the contrast coefficients that we entered in Section 12.8.2, and it is worth checking this table to make sure that the contrasts compare what they are supposed to.

The second table gives the statistics for each contrast in their raw form, but also corrected for unequal variances. I advise routinely looking at the rows labelled *Does not assume equal variances* (see Jane Superbrain Box 6.7). The table tells us the value of the contrast itself, which is the weighted sum of the group means. This value is obtained by taking each group mean, multiplying it by the weight for the contrast of interest, and then adding these values together.[6] The last two columns contain the 95% confidence interval for this estimate, which for the first contrast gives us $V_1 = 3.8$ [0.43, 7.17]. If we assume this sample is one of the 95% for which the confidence interval contains the true value, then the population difference between the no puppy group and the combined dose groups could be as small as 0.43 (virtually no difference) or as large as 7.17 (massive considering happiness is measured on a 10-point scale).

Each contrast estimate has a standard error. In the lower part of the table the standard error is corrected for unequal variances. This standard error is used to compute a test statistic, t, which is the contrast value divided by its standard error ($t = 3.8/1.483 = 2.56$). The significance value of the contrast is given in the final column, and this value is two-tailed. For contrast 1, we can say that taking puppy therapy significantly increased happiness compared to the control group ($p = 0.031$), but contrast 2

6 For the first contrast this value is:

$$v_1 = \sum_{i=1}^{k} \bar{X}_i W_i = (2.2 \times -2) + (3.2 \times 1) + (5 \times 1) = 3.8$$

tells us that 30 minutes of puppy therapy did not significantly affect happiness compared to 15 minutes ($p = 0.086$). Contrast 2 is almost significant, which again demonstrates how the NHST process can lead you to very all-or-nothing thinking (Section 3.2.2).

Contrast Coefficients

Contrast	Dose of puppies: No puppies	15 mins	30 mins
1	−2	1	1
2	0	−1	1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)	95% Confidence Interval Lower	Upper
Happiness (0–10)	Assumes equal variances	1	3.80	1.536	2.474	12	.029	.45	7.15
		2	1.80	.887	2.029	12	.065	−.13	3.73
	Does not assume equal variances	1	3.80	1.483	2.562	8.740	.031	.43	7.17
		2	1.80	.917	1.964	7.720	.086	−.33	3.93

Output 12.7

Output 12.8 shows the effect size (*Point Estimate*) and its 95% confidence interval for each contrast. The *Standardizer* is the estimate of the pooled variance. The table shows us Cohen's $\hat{d}$, but also a version that has a small sample correction applied (Hedge's correction). We should report the corrected value because we have a small sample. With the usual caveat that we will assume (perhaps incorrectly) that our sample is one of the 95% that have intervals containing the population values, we can interpret these as follows. The effect of puppy therapy (in general) compared to no therapy was huge, $\hat{d} = 2.54$ [0.25, 4.74], huge enough that if we found an effect this large in our own data we should be suspicious of it. The confidence interval alerts us to this because it is so wide, suggesting that the population effect could be anything between a fairly modest effect and an effect that is so large as to be completely implausible (it suggests happiness is nearly five standard deviations higher in the puppy therapy group compared to no therapy). The effect of 30 minutes of puppies compared to 15 minutes on happiness is also huge, $\hat{d} = 1.20$ [–0.07, 2.43], but again the confidence interval suggests that we really don't know very much at all about the size of effect because it could be zero, show a slight advantage to having 15 minutes (the minus value) or show a massive advantage to having 30 minutes (the value of 2.4). The large effects and wide confidence intervals are indicative of the tiny sample size.

Contrast Effect Sizes

		Contrast	Standardizer[a]	Point Estimate	95% Confidence Interval Lower	Upper
Happiness (0–10)	Cohen's d	1	1.402	2.710	.265	5.065
		2	1.402	1.284	−.079	2.601
	Hedges' correction	1	1.498	2.536	.248	4.740
		2	1.498	1.201	−.074	2.434

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation for all the groups.
Hedges' uses pooled standard deviation for all the groups, plus a correction factor.

Output 12.8

12.9.3 Output for *post hoc* tests Ⓑ

If we had no specific hypotheses about the effect of puppy therapy on happiness then we would have selected *post hoc* tests to compare all group means to each other. Even though we wouldn't do

contrasts and *post hoc* tests, as an educational exercise we did (Section 12.8.3). Output 12.9 shows tables containing Tukey's test (known as Tukey's HSD),[7] the Games–Howell procedure and Dunnett's test. If we look at Tukey's test first (because we have no reason to doubt that the population variances are unequal) we see that for each pair of groups the difference between group means is displayed, the standard error of that difference, the significance level of that difference and a 95% confidence interval. The first row of Output 12.9 compares the control group to the 15-minute group and reveals a non-significant difference (the *Sig.* of 0.516 is greater than 0.05), and the second row compares the control group to the 30-minute group where there is a significant difference (the *Sig.* of 0.021 is less than 0.05). It might seem odd that the planned contrast showed that any dose of puppy therapy produced a significant increase in happiness, yet the *post hoc* tests indicate that a 15-minute dose does not.



In Section 12.5, I explained that the first planned contrast would compare the experimental groups to the control group. Specifically, it would compare the average of the two group means of the experimental groups ((3.2 + 5.0)/2 = 4.1) to the mean of the control group (2.2). So, it was assessing whether the difference between these values (4.1 – 2.2 = 1.9) was significant. In the *post hoc* tests, when the 15-minute group is compared to the control it is testing whether the difference between the means of these two groups is significant. The difference in this case is only 1, compared to a difference of 1.9

Multiple Comparisons

Dependent Variable: Happiness (0–10)

	(I) Dose of puppies	(J) Dose of puppies	Mean Difference (I–J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	No puppies	15 mins	-1.000	.887	.516	-3.37	1.37
		30 mins	-2.800*	.887	.021	-5.17	-.43
	15 mins	No puppies	1.000	.887	.516	-1.37	3.37
		30 mins	-1.800	.887	.147	-4.17	.57
	30 mins	No puppies	2.800*	.887	.021	.43	5.17
		15 mins	1.800	.887	.147	-.57	4.17
Games-Howell	No puppies	15 mins	-1.000	.825	.479	-3.36	1.36
		30 mins	-2.800*	.917	.039	-5.44	-.16
	15 mins	No puppies	1.000	.825	.479	-1.36	3.36
		30 mins	-1.800	.917	.185	-4.44	.84
	30 mins	No puppies	2.800*	.917	.039	.16	5.44
		15 mins	1.800	.917	.185	-.84	4.44
Dunnett t (>control)[b]	15 mins	No puppies	1.000	.887	.227	-.87	
	30 mins	No puppies	2.800*	.887	.008	.93	

*. The mean difference is significant at the 0.05 level.

b. Dunnett t-tests treat one group as a control, and compare all other groups against it.

Output 12.9

7 The HSD stands for 'honestly significant difference', which has a slightly dodgy ring to it if you ask me.

for the planned contrast. This explanation illustrates how it is possible to have apparently contradictory results from planned contrasts and *post hoc* comparisons. More important, it illustrates the need to think carefully about what our planned contrasts test.

The third and fourth rows of Output 12.9 compared the 15-minute group to both the control group and the 30-minute group. The test involving the 15-minute and 30-minute groups shows that these group means did not differ (because the p of 0.147 is greater than our alpha of 0.05). Rows 5 and 6 repeat comparisons already discussed.

The second block of the table describes the Games–Howell test, and a quick inspection reveals the same pattern of results: the only groups that differed significantly were the 30-minute and control groups. These results give us confidence in our conclusions from Tukey's test because even if the population variances are not equal (which seems unlikely given that the sample variances are very similar), the profile of results holds true. Finally, Dunnett's test is described and you'll hopefully remember that we asked the computer to compare both experimental groups against the control using a one-tailed hypothesis that the mean of the control group would be smaller than both experimental groups. Even as a one-tailed hypothesis, levels of happiness in the 15-minute group are equivalent to the control group. However, the 30-minute group has a significantly higher happiness than the control group.

Output 12.10 shows the results of Tukey's test and the REGWQ test. These tests display subsets of groups that have the same means with associated p-values for each subset. Tukey's test has created two subsets of groups with statistically similar means. The first subset contains the control and 15-minute groups (indicating that these two groups have similar means, $p = 0.516$), whereas the second subset contains the 30- and 15-minute groups (which also have similar means, $p = 0.147$). The REGWQ test agrees with Tukey's test: the first subset suggests that the control and 15-minute groups have similar means ($p = 0.282$) and the second that the 15- and 30-minute groups have similar means ($p = 0.065$).

Tukey's LSD uses the **harmonic mean** sample size, which is a weighted version of the mean that takes account of the relationship between variance and sample size. Although you don't need to know the intricacies of the harmonic mean, it is useful that the harmonic sample size is used because it reduces bias that might be introduced through having unequal sample sizes. However, as we have seen, these tests are still biased when sample sizes are unequal.

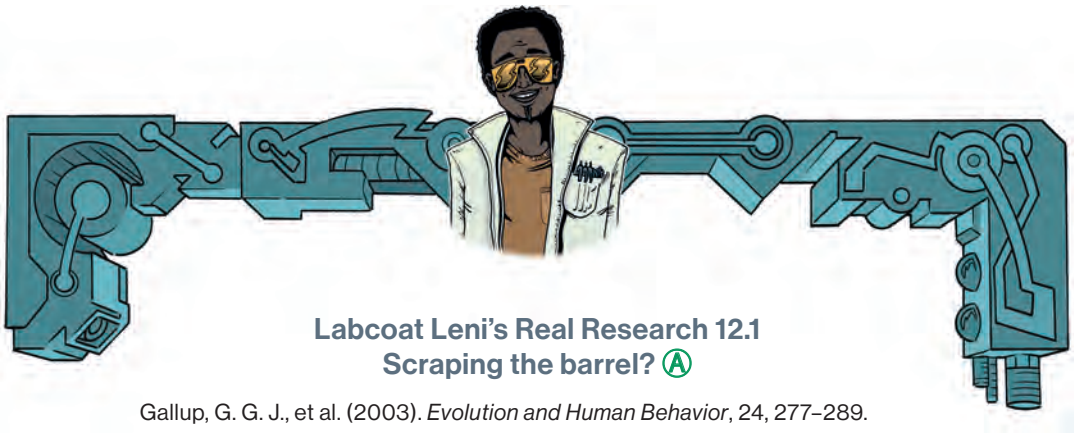
Happiness (0–10)

	Dose of puppies	N	Subset for alpha = 0.05 1	2
Tukey HSD[a]	No puppies	5	2.20	
	15 mins	5	3.20	3.20
	30 mins	5		5.00
	Sig.		.516	.147
Ryan–Einot–Gabriel–Welsch Range	No puppies	5	2.20	
	15 mins	5	3.20	3.20
	30 mins	5		5.00
	Sig.		.282	.065

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 5.000.

Output 12.10



Labcoat Leni's Real Research 12.1
Scraping the barrel? Ⓐ

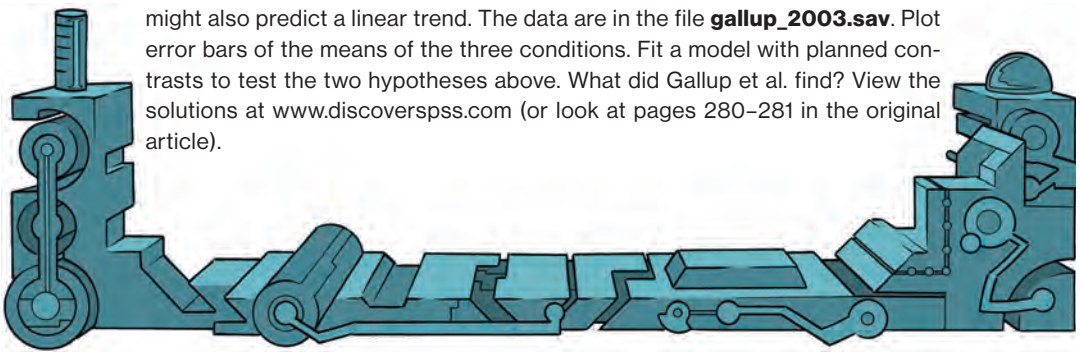
Gallup, G. G. J., et al. (2003). *Evolution and Human Behavior*, 24, 277–289.

Evolution has endowed us with many beautiful things (cats, dolphins, the Great Barrier Reef, etc.), all selected to fit their ecological niche. Given evolution's seemingly limitless capacity to produce beauty, it's something of a wonder how it managed to produce the human penis. One theory is sperm competition: the human penis has an unusually large glans (the 'bell-end') compared to other primates, and this may have evolved so that the penis can displace seminal fluid from other males by 'scooping it out' during intercourse. Armed with various female masturbatory devices from Hollywood Exotic Novelties, an artificial vagina from California Exotic Novelties, and some water and cornstarch to make fake sperm, Gallup and colleagues put this theory to the test. They loaded an artificial vagina with 2.6 ml of fake sperm and inserted one of three female sex toys into it before withdrawing it: a control phallus that had no coronal ridge (i.e., no bell-end), a phallus with a minimal coronal ridge (small bell-end) and a phallus with a coronal ridge.

They measured sperm displacement as a percentage using the following expression (included here because it is more interesting than all of the other equations in this book):

$$\frac{\text{Weight of vagina with semen} - \text{weight of vagina following insertion and removal of phallus}}{\text{Weight of vagina with semen} - \text{weight of empty vagina}} \times 100$$

A value of 100% means that all the sperm was displaced, and 0% means that none of the sperm was displaced. If the human penis evolved as a sperm displacement device then Gallup et al. predicted: (1) that having a bell-end would displace more sperm than not; and (2) that the phallus with the larger coronal ridge would displace more sperm than the phallus with the minimal coronal ridge. The conditions are ordered (no ridge, minimal ridge, normal ridge) so we might also predict a linear trend. The data are in the file **gallup_2003.sav**. Plot error bars of the means of the three conditions. Fit a model with planned contrasts to test the two hypotheses above. What did Gallup et al. find? View the solutions at www.discoverspss.com (or look at pages 280–281 in the original article).



12.10 Robust comparisons of several means ©

It's possible to run a robust test of several means using R and the *WRS2* package (Section 4.12). There is a workbook file (**t1waybt.spwb**) at www.discoverspss.com for running a robust variant of one-way independent ANOVA (*t1waybt*) with *post hoc* tests (*mcppb20*) described by Wilcox (2017). These tests assume neither normality nor homogeneity of variance, so you can ignore these assumptions and plough ahead. The syntax in the file is as follows:

```
BEGIN PROGRAM R.
library(WRS2)
mySPSSdata =  spssdata.GetDataFromSPSS(factorMode = "labels")
t1waybt(happiness~dose, data = mySPSSdata, tr = 0.2, nboot = 1000)
mcppb20(happiness~dose, data = mySPSSdata,  tr = 0.2, nboot = 1000)
END PROGRAM.
```

Run these six lines of syntax and you'll get Output 12.11, which tells us that there was not a significant difference (because the *p*-value is greater than 0.05) in happiness scores across the puppy therapy groups, $F_t = 3$, $p = 0.089$. The *post hoc* tests (which technically we should ignore because the overall test wasn't significant) show no significant difference between the control and 15 minutes ($p = 0.381$), or between the 15- and 30-minute groups ($p = 0.081$), but there is a significant difference (which we should ignore because the overall test wasn't significant) between the control and 30-minute groups ($p = 0.010$).



SPSS Tip 12.2
Robust one-way independent ANOVA ©

To break down the syntax, we're using two functions (*t1waybt* and *mcppb20*) that both have the same form:

```
t1waybt(outcome~predictor, data = mySPSSdata, tr = 0.2, nboot = 1000)
mcppb20(outcome~predictor, data = mySPSSdata,  tr = 0.2, nboot = 1000)
```

In our example we use our variable names to replace *outcome* with **happiness** and *predictor* with **dose**. The *tr* relates to the amount of trimming to the data (by default 0.2, or 20%, but you could change it to a different proportion). The *nboot* refers to the number of bootstrap samples, which I have set to 1000, which should be fine, but feel free to increase it.

```
Effective number of bootstrap samples was 671.

Test statistic: 3
p-value: 0.08942
Variance explained 0.623
Effect size 0.789

mcppb20(formula = Happiness ~ Dose, data = mySPSSdata, tr = 0.2,
    nboot = 1000)

                    psihat ci.lower ci.upper p-value
Control vs. 15 mins     -1 -3.33333  1.33333   0.381
Control vs. 30 mins     -3 -5.00000 -0.33333   0.010
15 mins vs. 30 mins     -2 -4.00000  0.66667   0.081
```

Output 12.11



Cramming Sam's Tips
One-way independent ANOVA

- One-way independent ANOVA compares several means, when those means have come from different groups of people; for example, if you have several experimental conditions and have used different participants in each condition. It is a special case of the linear model.
- When you have generated specific hypotheses before the experiment use *planned contrasts*, but if you don't have specific hypotheses use *post hoc* tests.
- There are lots of different *post hoc* tests. When you have equal sample sizes and homogeneity of variance is met, use REGWQ or Tukey's HSD. If sample sizes are slightly different then use Gabriel's procedure, but if sample sizes are very different use Hochberg's GT2. If there is any doubt about homogeneity of variance use the Games–Howell procedure.
- Use a *robust* test in all situations (the Welch or Brown–Forsythe *F*) or Wilcox's *t1way()* function.
- Locate the *p*-value (usually in a column labelled *Sig.*). If the value is less than or equal to 0.05 then scientists typically interpret this as the group means being significantly different.
- For contrasts and *post hoc* tests, again look to the columns labelled *Sig.* to discover if your comparisons are significant (they will be if the significance value is less than or equal to 0.05).



12.11 Bayesian comparison of several means ©

Because we've been using a linear model to compare means, we can compute a Bayes factor (Section 3.8) just as we did in Section 9.13. You can access the dialog box in Figure 12.18 by selecting *Analyze* > *Bayesian Statistics* > One-way ANOVA. First, specify the model by dragging the outcome (**happiness**) to the box labelled *Dependent* and the predictor (**dose**) to the box labelled *Factor*. If we use estimation on a model with a categorical predictor we will get Bayesian estimates for the group means (*not* the $\hat{b}$s from the model), which are not particularly interesting. Therefore, I have selected to compute only Bayes factors (Estimate Bayes Factor). Click Bayes Factor... and select the default JZS method (Jeffreys, 1961; Zellner & Siow, 1980). In the main dialog box click OK to fit the model.

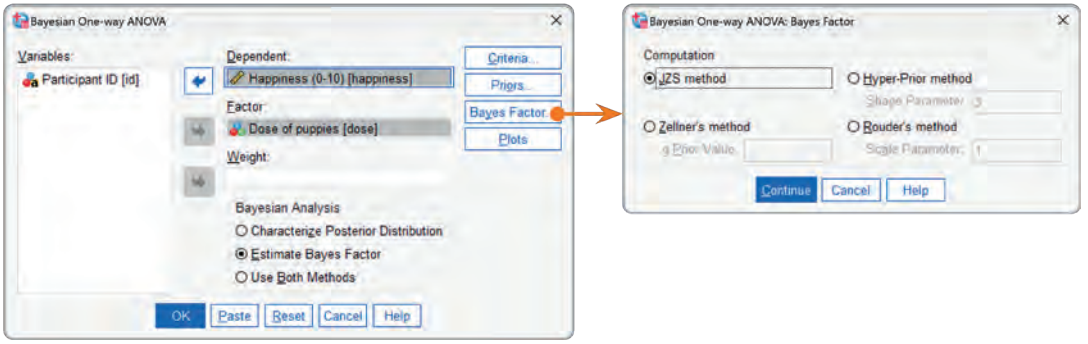


Figure 12.18 Dialog boxes for Bayesian ANOVA

The Bayes factor here is essentially the same as the one we computed for the linear model (Section 9.13): it compares the full model (predicting **happiness** from **dose** and the intercept) to the null model (predicting **happiness** from only the intercept). Therefore, it quantifies the overall effect of **dose**, which is reflected by the fact that the Bayes factor appears as part of the table displaying the overall fit of the model (the *F*-statistic). Output 12.12 shows that the Bayes factor for including **dose** as a predictor (compared to not including it) is 2.158. This means that the data are 2.158 times more likely under the alternative hypothesis (dose of puppy therapy has an effect) than under the null (dose of puppy therapy has no effect). This value is not strong evidence, but nevertheless suggests we should shift our belief about puppy therapy towards it being effective by a factor of about 2.

ANOVA

Happiness (0–10)	Sum of Squares	df	Mean Square	F	Sig.	Bayes Factor[a]
Between Groups	20.133	2	10.067	5.119	.025	2.158
Within Groups	23.600	12	1.967			
Total	43.733	14				

a. Bayes factor: JZS

Output 12.12

12.12 Reporting results from one-way independent ANOVA Ⓑ

We report the *F*-statistic and the degrees of freedom associated with it. Remember that *F* is the model mean squares divided by the residual mean squares for the residual, so the associated degrees of freedom are those for the effect of the model ($df_M = 2$) and those for the residuals of the model ($df_R = 12$). Also include an effect size estimate. Finally, report the exact *p*-value. Based on this advice we could report the overall effect as:

- ✓ There was a significant effect of puppy therapy on levels of happiness, $F(2, 12) = 5.12$, $p = 0.025$, $\hat{\omega}^2 = 0.35$ [–0.15, 0.59].

Notice that the degrees of freedom are in parentheses. I've advised you to always report Brown–Forsythe or Welch's *F*, so let's do that instead (note the adjusted residual degrees of freedom, and change in *F* and *p*):

- ✓ There was not a significant effect of puppy therapy on levels of happiness, $F(2, 7.94) = 4.32$, $p = 0.054$, $\hat{\omega}^2 = 0.35$ [–0.15, 0.59].
- ✓ There was a significant effect of puppy therapy on levels of happiness, $F(2, 11.57) = 5.12$, $p = 0.026$, $\hat{\omega}^2 = 0.35$ [–0.15, 0.59].

The linear contrast can be reported in much the same way:

- ✓ There was a significant linear trend, $F(1, 12) = 9.97$, $p = 0.008$, indicating that as the dose of puppy therapy increased, happiness increased proportionately.

The degrees of freedom have changed to reflect how *F* was calculated. We can do something similar for the planned contrasts:

- ✓ Planned contrasts revealed that having any dose of puppy therapy significantly increased happiness compared to having a control, $\hat{V} = 3.8$ [0.43, 7.17], $t(8.74) = 2.56$, $p = 0.031$, $\hat{d} = 2.54$ [0.25, 4.74], but having 30 minutes did not significantly increase happiness compared to having 15 minutes, $\hat{V} = 1.8$ [–0.33, 3.93], $t(7.72) = 1.96$, $p = 0.086$, $\hat{d} = 1.20$ [–0.07, 2.43].

12.13 Jane and Brian's story Ⓐ

Brian had been astonished to see Jane in the library, and even more so when she dropped her guard enough to ask his name. He'd momentarily cursed his parents for calling him Brian, but Jane didn't seem to care. They'd even continued a bit of an awkward conversation about music before Jane quickly excused herself and ran off. She'd looked pale, like she'd remembered something horrific. Brian wasn't used to seeing her like that: she was always so self-assured. He wanted to know what was up, but despite his attempt to follow her he lost sight of her as she darted across campus. The episode played on his mind: he didn't want her to feel bad. He had no way to contact her, though, other than by going to places where she hung out. He loitered on campus, but it was hopeless. He bumped into Alex. She seemed irked about his encounter with Jane. She said Jane was bad news. He pressed Alex for information, but she gave nothing away except that Jane often went to Blow Your Speakers, the record store in town, at the weekends. Brian knew it well: they had vinyl and coffee mornings on a Saturday. He'd always avoided it, fearing it would be full of pretentious hipsters. This Saturday he made an exception and sat among people in skinny trousers and strange hats, who sipped coffee,

twiddled their moustaches and tried to choose more obscure cuts to play than the last person. He'd had enough and got up to leave as Jane walked in.

Jane was shocked to see Brian, but resisted the urge to flee. Brian smiled, rushed over and placed a reassuring hand on her shoulder. Jane flinched. Brian asked if she was alright and explained that he was worried he'd upset her. Jane was silent, unsure what to say, so Brian filled the silence with his statistics lecture. He needed to work on his social skills.

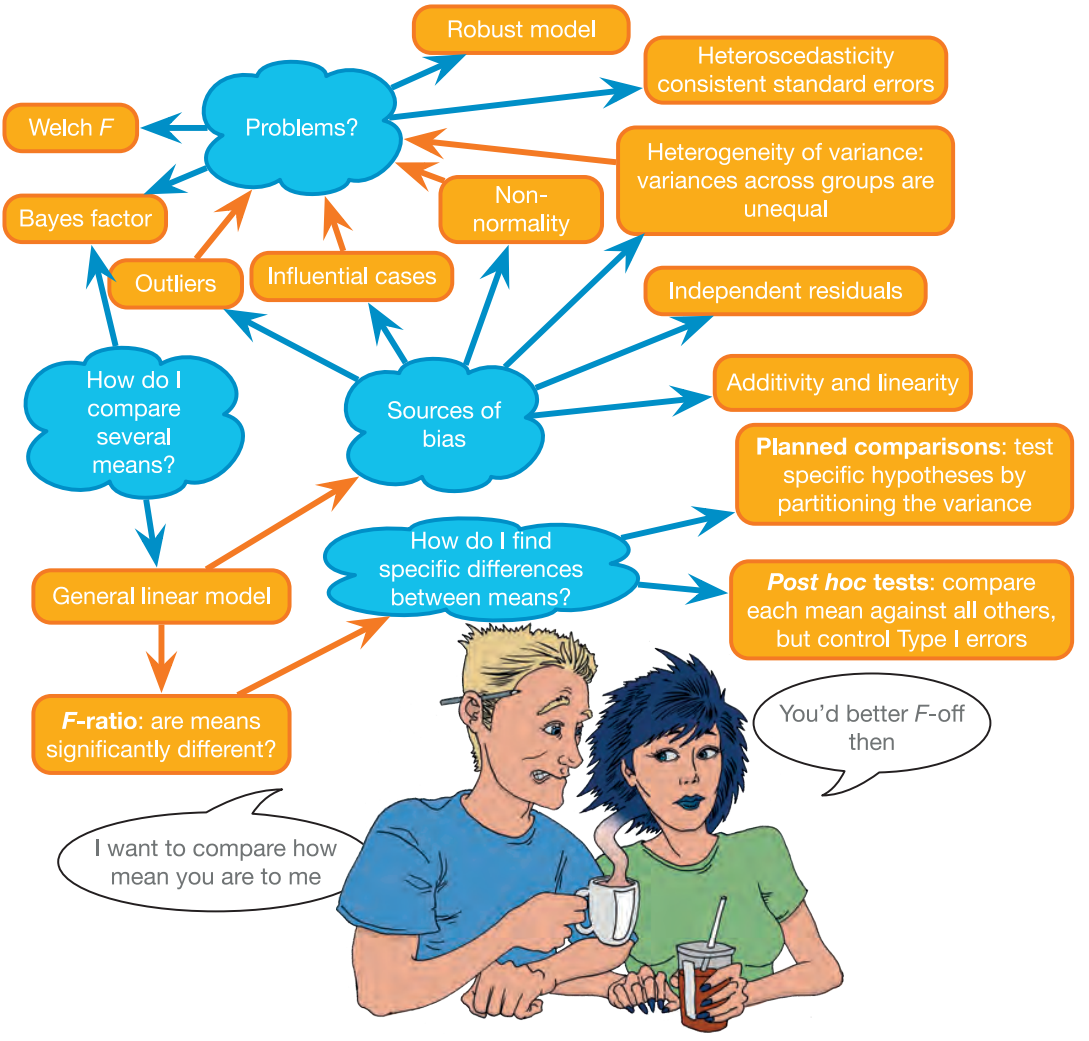


Figure 12.19 What Brian learnt from this chapter

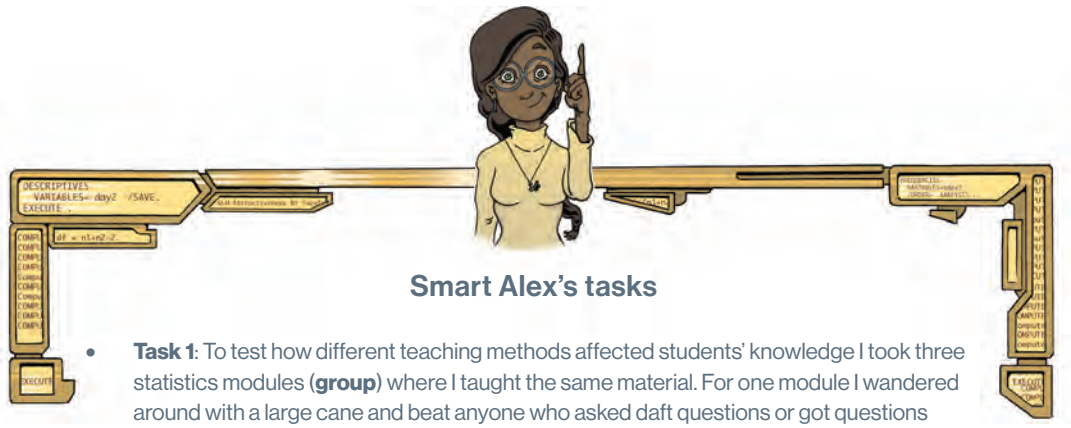
12.14 What next? Ⓐ

My life was changed by a letter popping through the letterbox one day saying that I could go to the local grammar school. When my parents told me, they were not as celebratory as you might expect: they knew how much I wanted to be with my friends. I had come to terms with failing the exam, so my initial reaction was to say that I wanted to go to the local school. I was unwavering in this view.

Unwavering, that is, until my brother convinced me that being at the same school as him would be really cool. It's hard to measure how much I looked up to him, and still do, but the fact that I willingly subjected myself to a lifetime of social dysfunction just to be with him is a measure of sorts. As it turned out, being at school with him was not always cool – he was bullied for being a 'boffin' (that's what they called nerds in the UK in the 1980s), and being the younger brother of a boffin made me a target. Luckily, unlike my brother, I was not clever and played football, which seemed to be good enough reasons for them to leave me alone. Most of the time.

12.15 Key terms that I've discovered

Analysis of variance (ANOVA)
Brown–Forsythe *F*
Cubic trend
Deviation contrast
Difference contrast (reverse Helmert contrast)
Eta squared, $\hat{\eta}^2$
General linear model
Grand variance
Harmonic mean
Helmert contrast
Independent ANOVA
Omega squared, $\hat{\omega}^2$
Orthogonal
Pairwise comparisons
Planned contrasts
Polynomial contrast
Post hoc tests
Quadratic trend
Quartic trend
Repeated contrast
Simple contrast
Weights
Welch's *F*



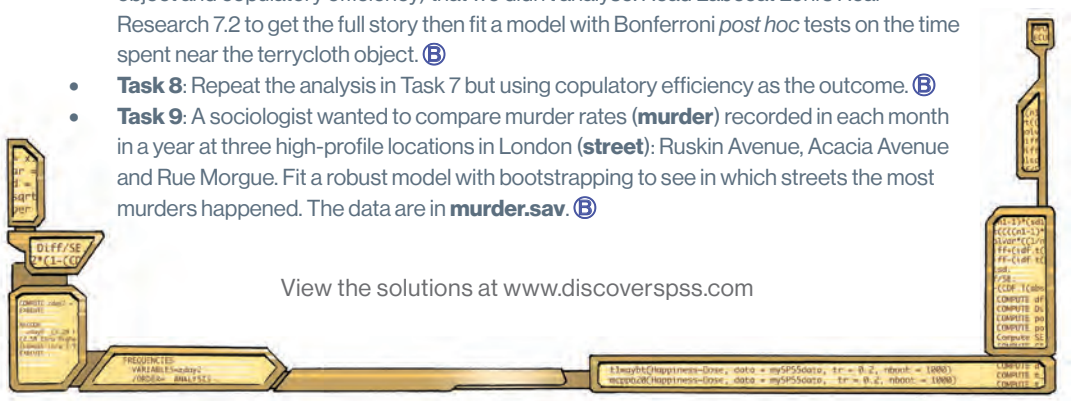
Smart Alex's tasks

- **Task 1**: To test how different teaching methods affected students' knowledge I took three statistics modules (**group**) where I taught the same material. For one module I wandered around with a large cane and beat anyone who asked daft questions or got questions wrong (punish). In the second I encouraged students to discuss things that they found difficult and gave anyone working hard a nice sweet (reward). In the final course I neither punished nor rewarded students' efforts (indifferent). I measured the students' exam marks (**percentage**). The data are in the file **teach_method.sav**. Fit a model with planned contrasts to test the hypotheses that: (1) reward results in better exam results than either punishment or indifference; and (2) indifference will lead to significantly better exam results than punishment. Ⓑ
- **Task 2** Fit a robust model for Task 1 Ⓑ
- **Task 3**: Children wearing superhero costumes are more likely to injure themselves because of the unrealistic impression of invincibility that these costumes could create. For example, children have reported to hospital with severe injuries because of trying 'to initiate flight without having planned for landing strategies' (Davies et al., 2007). I can relate to the imagined power that a costume bestows upon you; indeed, I have been known to dress up as Fisher by donning a beard and glasses and trailing a goat around on a lead in the hope that it might make me more knowledgeable about statistics. Imagine we had data (**superhero.sav**) about the severity of **injury** (on a scale from 0, no injury, to 100, death) for

children reporting to the accident and emergency department at hospitals, and information on which superhero costume they were wearing (**hero**): Spiderman, Superman, the Hulk or a teenage mutant ninja turtle. Fit a model with planned contrasts to test the hypothesis that those wearing costumes of flying superheroes (Superman and Spiderman) have more severe injuries. Ⓑ

- **Task 4**: In Chapter 7 (Section 7.6) there are some data looking at whether eating soya meals reduces your sperm count. Analyse these data with a linear model (ANOVA). What's the difference between what you find and what was found in Section 7.6.5? Why do you think this difference has arisen? The data are in **soya.sav**. Ⓑ
- **Task 5**: Mobile phones emit microwaves, and so holding one next to your brain for large parts of the day is a bit like sticking your brain in a microwave oven and pushing the 'cook until well done' button. If we wanted to test this experimentally, we could get six groups of people and strap a mobile phone on their heads, then by remote control turn the phones on for a certain amount of time each day. After 6 months, we measure the size of any tumour (in mm^3) close to the site of the phone antenna (just behind the ear). The six groups experienced 0, 1, 2, 3, 4 or 5 hours per day of phone microwaves for 6 months. Do tumours significantly increase with greater daily exposure? The data are in **tumour.sav**. Ⓑ
- **Task 6**: Using the data in **glastonbury.sav**, fit a model to see if the change in hygiene (**change**) is significant across people with different musical tastes (**music**). Use a simple contrast to compare each group against the no subculture group. Ⓑ
- **Task 7**: Labcoat Leni's Real Research 7.2 describes an experiment on quails with fetishes for terrycloth objects. There were two outcome variables (time spent near the terrycloth object and copulatory efficiency) that we didn't analyse. Read Labcoat Leni's Real Research 7.2 to get the full story then fit a model with Bonferroni *post hoc* tests on the time spent near the terrycloth object. Ⓑ
- **Task 8**: Repeat the analysis in Task 7 but using copulatory efficiency as the outcome. Ⓑ
- **Task 9**: A sociologist wanted to compare murder rates (**murder**) recorded in each month in a year at three high-profile locations in London (**street**): Ruskin Avenue, Acacia Avenue and Rue Morgue. Fit a robust model with bootstrapping to see in which streets the most murders happened. The data are in **murder.sav**. Ⓑ

View the solutions at www.discoverspss.com





GLM 2: COMPARING MEANS ADJUSTED FOR OTHER PREDICTORS (ANALYSIS OF COVARIANCE)

13.1 What will this chapter tell me? 616
13.2 What is ANCOVA? 617
13.3 The general linear model with covariates 618
13.4 Effect size for ANCOVA 622
13.5 Assumptions and issues in ANCOVA designs 622
13.6 Conducting ANCOVA using SPSS 627
13.7 Interpreting ANCOVA 634
13.8 The non-parallel slopes model and the assumption of homogeneity of regression slopes 641
13.9 Robust ANCOVA 644
13.10 Bayesian analysis with covariates 645
13.11 Reporting results 647
13.12 Jane and Brian's story 647
13.13 What next? 647
13.14 Key terms that I've discovered 648
Smart Alex's tasks 649

Letters A to D indicate increasing difficulty

13.1 What will this chapter tell me? Ⓑ

My road to rock stardom had taken a bit of a knock with my unexpected entry to an all-boys' grammar school (rock bands and grammar schools really didn't go together). I needed to be inspired and I turned to the masters: Iron Maiden. I first heard Iron Maiden at the age of 11 when a friend lent me *Piece of Mind* on a cassette and told me to listen to 'The Trooper'. It was, to put it mildly, an epiphany. I became their smallest (I was 11) biggest fan and obsessed about them in the unhealthiest of ways. I bombarded the man who ran their fan club (a guy called Keith) with letters, and, bless him, he replied to them all. Eventually my stalking paid off and Keith arranged for me to go backstage when they played what was then (and to me always will be), the Hammersmith Odeon in London on 5 November 1986 (*Somewhere on Tour*, in case you're interested). Not only was it the first time I had seen them live, but I got to meet them too. It is difficult to convey how anxiety-evoking that night was. It was all quite overwhelming. I was so utterly awestruck that I managed to say precisely nothing to any of the band (but I do have some good photos where my speechlessness is tangible; see Figure 13.1). Soon to become a theme in my life, a social situation had provoked me to make an utter fool of myself.[1] When it was over I was in no doubt that this was the best day of my life. I thought nothing would ever be as good.[2] This may be true, but I have subsequently had other very nice experiences, so who is to say that they were not better? I could compare experiences to see which one is the best, but there is an important confound: my age. At the age of 13, meeting Iron Maiden was beyond exciting, but adulthood (sadly) dulls your capacity for this kind of unqualified excitement. To see which experience was best, I would have to take account of the variance in enjoyment that is attributable to my age at the time. Doing so will give me a purer measure of how much variance in my enjoyment is attributable to the event itself. This chapter extends the previous one to look at situations in which you want to compare group means, but also adjust those means for another variable



Figure 13.1 Dave Murray (guitarist from Iron Maiden) and me backstage in London in 1986 (my grimace reflects the utter terror I was feeling at meeting my hero)

1 In my teens I met many bands that I liked, and Iron Maiden were by far the nicest.

2 Apart from my wedding day, as it turned out.

(or variables) that you expect to affect the outcome. This process involves fitting a linear model in which an outcome is predicted from dummy variables representing group membership but one or more other predictors (usually continuous variables) are included. These additional predictors are sometimes labelled covariates, and this configuration of the linear model is sometimes known as *analysis of covariance*.

13.2 What is ANCOVA? Ⓑ

In the previous chapter we saw how we can compare multiple group means with the linear model by using dummy variables to code group membership. In addition, in Chapter 9 we saw how the linear model can incorporate several continuous predictor variables. It should, therefore, be no surprise that the linear model to compare means can be extended to include one or more continuous variables that predict the outcome (or dependent variable). When the main focus of the model is to compare means (perhaps from different experimental groups) then these additional predictors in the model are sometimes referred to as **covariates**. Also, this form of the linear model is sometimes referred to as **analysis of covariance** (or **ANCOVA** for short).[3]



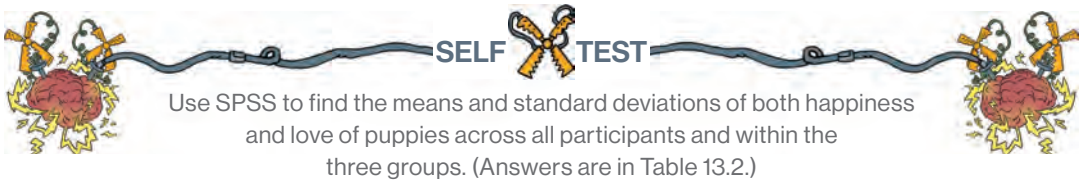
In the previous chapter, we used an example about the effects of puppy therapy on happiness. Let's think about things other than puppy therapy that might influence happiness. The obvious one is how much you like dogs (a dog phobic is going to be about as happy after puppy therapy as I would be after tarantula therapy), but there are other things too such as individual differences in temperament. If these variables (the covariates) are measured, then it is possible to adjust for the influence they have on the outcome variable by including them in the linear model. From what we know of hierarchical regression (see Chapter 9) it should come as no surprise that if we enter the covariate into the model first, and then enter the dummy variables representing the group means (e.g., the experimental manipulation), we can see what effect a predictor variable has, *adjusting for* the effect of the covariate. In essence, rather than predicting the outcome from group means, we predict it from group means that have been adjusted for the effect of covariate(s). There are two main reasons to include covariates in ANOVA:

- **To reduce within-group error variance**: When we predict an outcome from group means (e.g., when these represent the effect of an experiment), we compute an *F*-statistic by comparing the amount of variability in the outcome that the experiment can explain against the variability that it cannot explain. If we can attribute some of this 'unexplained' variance (SS_R) to other measured variables (covariates), then we reduce the error variance, allowing us to assess more sensitively the difference between group means (SS_M).
- **Elimination of confounds**: There may be unmeasured variables that confound the results (i.e., variables other than the experimental manipulation that affect the outcome variable). If variables exist that are known to influence the outcome that we're trying to predict, then including them as covariates can remove these variables as potential explanations for the effect of interest.

3 As we've discussed before, these labels for special cases of the linear model (such as one-way independent ANOVA in the previous chapter, and ANCOVA here) reflect historical divisions in methods (see Misconception Mutt 11.1). They are unhelpful because they create the impression that we're using distinct statistical models when we're not. I want you to focus on the general linear model that underpins these special cases, but I can't really avoid using the ANOVA/ANCOVA labels completely because otherwise you'll be lost when your supervisor tells you to do ANOVA/ANCOVA!

13.3 The general linear model with covariates Ⓑ

The researchers who conducted the puppy therapy study in the previous chapter suddenly realized that a participant's love of dogs could affect whether puppy therapy affected happiness. Therefore, they replicated the study on different participants, but included a self-report measure of love of puppies from 0 (I am a weird person who hates puppies, please be deeply suspicious of me) to 7 (puppies are the best thing ever, one day I might marry one). The data are in Table 13.1 and in the file **puppy_love.sav**, which contains the variables **id** (the participant's id code), **dose** (1 = control, 2 = 15 minutes, 3 = 30 minutes), **happiness** (the person's happiness on a scale from 0 to 10) and **puppy_love** (the love of puppies from 0 to 7).



Use SPSS to find the means and standard deviations of both happiness and love of puppies across all participants and within the three groups. (Answers are in Table 13.2.)

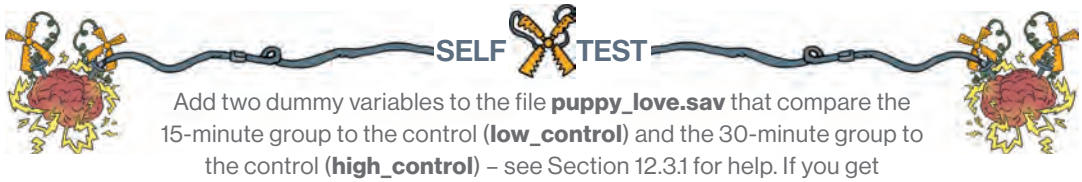
In the previous chapter, we characterized this experimental scenario as Eq. 12.3, and knowing what we know about the linear model (Chapter 9) hopefully you can see that that equation can be extended to include the covariate as follows:

$$\begin{aligned}\text{happiness}_i &= \hat{b}_0 + \hat{b}_1\text{long}_i + \hat{b}_2\text{short}_i + \hat{b}_3\text{Covariate}_i + e_i \\ &= \hat{b}_0 + \hat{b}_1\text{long}_i + \hat{b}_2\text{short}_i + \hat{b}_3\text{puppy love}_i + e_i\end{aligned} \tag{13.1}$$

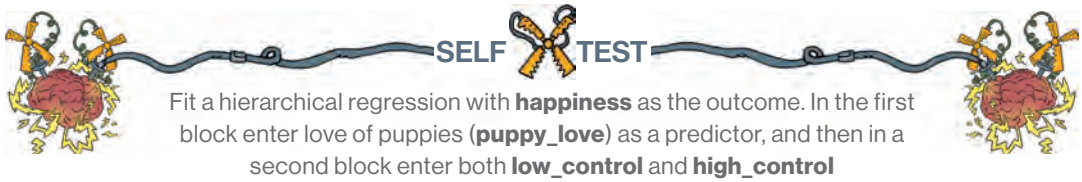
The predicted values of happiness are obtained from (see Jane Superbrain Box 2.4)

$$\widehat{\text{happiness}}_i = \hat{b}_0 + \hat{b}_1\text{long}_i + \hat{b}_2\text{short}_i + \hat{b}_3\text{puppy love}_i \tag{13.2}$$

We can compare the means of different groups using a linear model (see Section 12.3) in which groups are coded as the dummy variables **long** and **short**: **long** takes the value of 1 only for the 30-minute group, **short** takes a value of 1 only for the 15-minute group, and in all other situations they have a value of 0. We can add a covariate as a predictor to the model to test the difference between group means *adjusted for the covariate*. Let's look at a practical example.



Add two dummy variables to the file **puppy_love.sav** that compare the 15-minute group to the control (**low_control**) and the 30-minute group to the control (**high_control**) – see Section 12.3.1 for help. If you get stuck use **puppy_love_dummy.sav.**



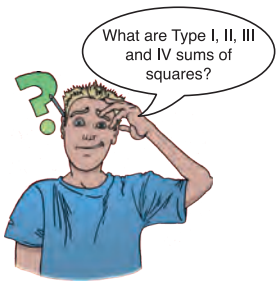
Fit a hierarchical regression with **happiness** as the outcome. In the first block enter love of puppies (**puppy_love**) as a predictor, and then in a second block enter both **low_control** and **high_control** (forced entry) – see Section 9.10 for help.

Table 13.1 More puppy therapy data

Dose	Participant's happiness	Love of puppies
Control	3	4
	2	1
	5	5
	2	1
	2	2
	2	2
	7	7
	2	4
	4	5
15 minutes	7	5
	5	3
	3	1
	4	2
	4	2
	7	6
	5	4
	4	2
30 minutes	9	1
	2	3
	6	5
	3	4
	4	3
	4	3
	4	2
	6	0
	4	1
	6	3
	2	0
	8	1
	5	0

Table 13.2 Means (and standard deviations) from **puppy_love.sav**

Dose	Participant's happiness	Love of puppies
Control	3.22 (1.79)	3.44 (2.07)
15 minutes	4.88 (1.46)	3.12 (1.73)
30 minutes	4.85 (2.12)	2.00 (1.63)
Total	4.37 (1.96)	2.73 (1.86)



The summary of the model resulting from the self-test (Output 13.1) shows us the goodness of fit of the model first when only the covariate is used in the model, and second when both the covariate and the dummy variables are used. The difference between the values of R^2 (0.288 – 0.061 = 0.227) represents the individual contribution of puppy therapy to predicting happiness. Puppy therapy accounted for 22.7% of the variation in happiness, whereas love of puppies accounted for only 6.1%. This additional information provides some insight into the substantive importance of puppy therapy. The next table is the ANOVA table, which is also divided into two sections. The top half represents the effect of the covariate alone, whereas the bottom half represents the whole model (i.e., covariate and puppy therapy included). Notice at the bottom of the ANOVA table (the bit for model 2) that the entire model (love of puppies and the dummy variables) accounts for 31.92 units of variance (SS_M), there are 110.97 units in total (SS_T) and the unexplained variance (SS_R) is 79.05.

The interesting bit is the table of model coefficients (Output 13.2). The top half shows the effect when only the covariate is in the model, and the bottom half, which we'll focus on, contains the whole model. The $\hat{b}$-values for the dummy variables represent the difference between the means of the 15-minute group and the control group and the 30-minute group and the control group – see Section 12.3.2 for an explanation of why. The means of the 15- and 30-minute groups were 4.88 and 4.85 respectively, and the mean of the control group was 3.22 (see Table 13.2). Therefore, the $\hat{b}$-values for the two dummy variables should be roughly the same (4.88 – 3.22 = 1.66 for the control group vs. the 15-minute group and 4.85 – 3.22 = 1.63 for the control group vs. the 30-minute group). The astute among you might notice that the $\hat{b}$-values in Output 13.2 are not only very different from each other

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.246[a]	.061	.027	1.929
2	.536[b]	.288	.205	1.744

a. Predictors: (Constant), Love of puppies (0–7)

b. Predictors: (Constant), Love of puppies (0–7), Control vs. 15 mins (Dummy 1), Control vs. 30 mins (Dummy 2)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.734	1	6.734	1.809	.189[b]
	Residual	104.232	28	3.723		
	Total	110.967	29			
2	Regression	31.920	3	10.640	3.500	.030[c]
	Residual	79.047	26	3.040		
	Total	110.967	29			

a. Dependent Variable: Happiness (0–10)

b. Predictors: (Constant), Love of puppies (0–7)

c. Predictors: (Constant), Love of puppies (0–7), Control vs. 15 mins (Dummy 1), Control vs. 30 mins (Dummy 2)

Output 13.1

(which shouldn't be the case because the 15- and 30-minute groups means are virtually the same), but also different from the values I've just calculated. Does this mean I've been lying to you for the past 50 pages about what the beta values represent? I'm evil, but I'm not *that* evil. The reason for this apparent anomaly is that with a covariate present, the $\hat{b}$-values represent the differences between the means of each group and the control *adjusted for the covariate(s)*. In this case, they represent the difference in the means of the puppy therapy groups adjusted for the love of puppies.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3.657	.634		5.764	<.001
	Love of puppies (0–7)	.260	.193	.246	1.345	.189
2	(Constant)	1.789	.867		2.063	.049
	Love of puppies (0–7)	.416	.187	.395	2.227	.035
	Control vs. 15 mins (Dummy 1)	1.786	.849	.411	2.102	.045
	Control vs. 30 mins (Dummy 2)	2.225	.803	.573	2.771	.010

a. Dependent Variable: Happiness (0–10)

Output 13.2

These **adjusted means** come directly from the model. If we replace the $\hat{b}$-values in Eq. 13.2 with the values in Output 13.2, our model becomes:

$$\widehat{\text{happiness}}_i = 1.789 + 2.225\text{long}_i + 1.786\text{short}_i + 0.416\ \text{puppy love}_i \qquad (13.3)$$

Remember that **long** and **short** are dummy variables such that **long** takes the value of 1 only for the 30-minute group, and **short** takes a value of 1 only for the 15-minute group; in all other situations they have a value of 0. To get the adjusted means, we use this equation, but rather than replacing the covariate with an individual's score, we replace it with the mean value of the covariate from Table 13.2 (i.e., 2.73) because we're interested in the predicted value for each group at the *mean level of the covariate*. For the control group, the dummy variables are both coded as 0, so we replace **long** and **short** in the model with 0. The adjusted mean (which is the predicted value of happiness from the model) will, therefore, be 2.925:

$$\begin{aligned}\widehat{\text{happiness}}_{\text{no puppies}} &= 1.789 + (2.225 \times 0) + (1.786 \times 0) + \left(0.416 \times \bar{X}_{\text{puppy love}}\right) \\ \widehat{\text{happiness}}_{\text{no puppies}} &= 1.789 + (0.416 \times 2.73) \\ &= 2.925\end{aligned} \qquad (13.4)$$

For the 15-minute group, the dummy variable **short** is 1 and **long** is 0, so the adjusted mean is 4.71:

$$\begin{aligned}\widehat{\text{happiness}}_{\text{15 mins}} &= 1.789 + (2.225 \times 0) + (1.786 \times 1) + \left(0.416 \times \bar{X}_{\text{puppy love}}\right) \\ \widehat{\text{happiness}}_{\text{15 mins}} &= 1.789 + 1.786 + (0.416 \times 2.73) \\ &= 4.71\end{aligned} \qquad (13.5)$$

For the 30-minute group, the dummy variable **short** is 0 and **long** is 1, so the adjusted mean is 5.15:

$$\begin{aligned}\widehat{\text{happiness}}_{\text{30 mins}} &= 1.789 + (2.225 \times 1) + (1.786 \times 0) + \left(0.416 \times \bar{X}_{\text{puppy love}}\right) \\ \widehat{\text{happiness}}_{\text{30 mins}} &= 1.789 + 2.225 + (0.416 \times 2.73) \\ &= 5.15\end{aligned} \qquad (13.6)$$

We can now see that the $\hat{b}$-values for the two dummy variables represent the differences between these *adjusted* means (4.71 – 2.93 = 1.78 for the control group vs. the 15-minute group and 5.15 – 2.93 = 2.22 for the control group vs. the 30-minute group). These adjusted means are the average amount of happiness for each group at the mean level of love of puppies. Some people think of this kind of model (i.e., ANCOVA) as 'controlling' for the covariate, because it compares the predicted group means at the average value of the covariate, so the groups are being compared at a level of the covariate that is the same for each group. However, as we shall see, the 'controlling for the covariate' analogy is not a good one.

13.4 Effect size for ANCOVA Ⓑ

In the previous chapter we used eta squared, η^2, as an effect size measure when comparing means (Section 12.7). When we include a covariate in the model we could calculate eta squared for both the categorical predictor and the covariate. However, an effect size measure called **partial eta squared (partial η^2)** is typically a better option. Unlike eta squared, which quantifies the proportion of total variance that a variable explains, partial eta squared quantifies the proportion of variance that a variable explains that *is not explained by other variables in the model*. Let's look at this with our example. Suppose we want to know the effect size of the dose of puppy therapy. Partial eta squared is the proportion of variance in happiness that the dose of puppy therapy shares that is not attributed to love of puppies (the covariate). If you think about the variance that the covariate cannot explain, there are two sources: it cannot explain the variance attributable to the dose of puppy therapy, $SS_{\text{puppy therapy}}$, and it cannot explain the error variability, SS_R. Therefore, we use these two sources of variance instead of the total variability, SS_T, in the calculation. The difference between eta squared and partial eta squared is illustrated by comparing the following two equations:

$$\eta^2 = \frac{SS_{\text{Effect}}}{SS_{\text{Total}}} \tag{13.7}$$

$$\text{Partial}\,\eta^2 = \frac{SS_{\text{Effect}}}{SS_{\text{Effect}} + SS_{\text{Residual}}} \tag{13.8}$$

SPSS will produce partial eta squared for us, and we'll look at its calculation in Section 13.7.2. You can also use omega squared (ω^2) as an effect size measure, but IBM SPSS Statistics software does not produce it.

13.5 Assumptions and issues in ANCOVA designs Ⓒ

Including covariates doesn't change the fact were using the general linear model, so all the sources of potential bias (and counteractive measures) discussed in Chapter 6 apply. There are two additional considerations: (1) independence of the covariate and treatment effect; and (2) homogeneity of regression slopes.

13.5.1 Independence of the covariate and treatment effect Ⓒ

I said in the previous section that covariates can be used to reduce within-group error variance if the covariate explains some of this error variance, which will be the case if the covariate is independent of the experimental effect (group means). Figure 13.2 shows three different scenarios. Part A shows a

basic model that compares group means (it is like Figure 12.5). The variance in the outcome (in our example happiness) can be partitioned into two parts that represent the experimental or treatment effect (in this case the administration of puppy therapy) and the error or unexplained variance (i.e., factors that affect happiness that we haven't measured). Part B shows the ideal scenario when including a covariate, which is that the covariate shares its variance only with the bit of happiness that is currently unexplained. In other words, it is completely independent of the treatment effect (it does not overlap with the effect of puppy therapy at all). Some argue that this scenario is the only one in which ANCOVA is appropriate (Wildt & Ahtola, 1978). Part C shows a situation in which the effect of the covariate overlaps with the experimental effect. In other words, the experimental effect is confounded with the effect of the covariate. In this situation, the covariate will reduce (statistically speaking) the experimental effect because it explains some of the variance that would otherwise be attributable to the experiment. When the covariate and the experimental effect (independent variable) are not independent, the treatment effect is obscured, spurious treatment effects can arise, and at the very least the interpretation of the ANCOVA is seriously compromised (Wildt & Ahtola, 1978).

The problem of the covariate and treatment sharing variance is common and is ignored or misunderstood by many people (Miller & Chapman, 2001). Miller and Chapman are not the only people to point this out; their paper is very readable and they cite many examples of people misapplying ANCOVA. Their main point is that when treatment groups differ on the covariate,

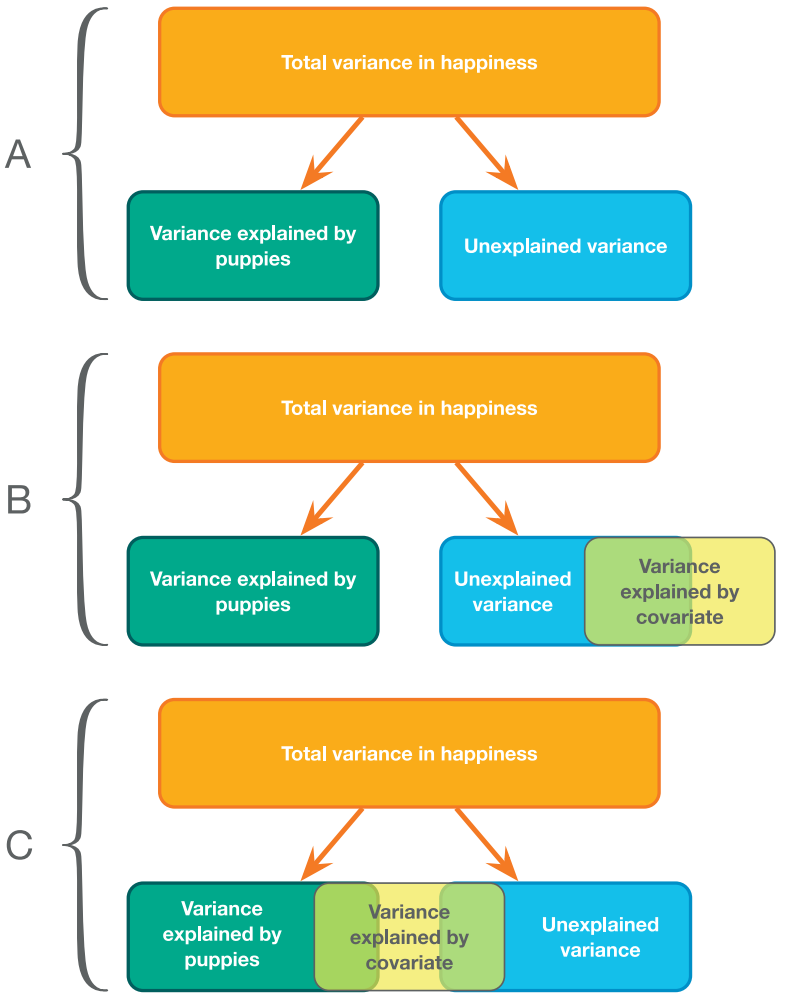
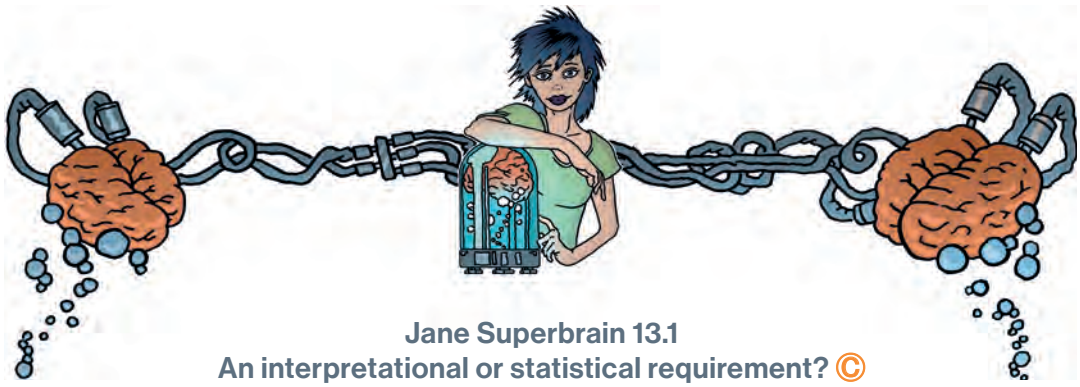


Figure 13.2 The role of the covariate in ANCOVA (see text for details)



Jane Superbrain 13.1
An interpretational or statistical requirement? ©

The treatment effect and covariate are simply predictor variables in a general linear model, yet despite several hundred pages discussing linear models, at no point has it been mentioned that predictors should be completely independent. It has been noted that predictors shouldn't overlap too much (e.g., collinearity) but that's quite different from saying that they shouldn't overlap at all. If, in general, we don't care about predictors being independent in linear models, why should we care now? The short answer is we don't, there is no *statistical* requirement for the treatment variable and covariate to be independent.

However, there are situations in which ANCOVA can be biased when the covariate is not independent of the treatment variable. One situation, common in medical research, has been discussed a lot: an outcome (e.g., hypertension) is measured at baseline, and after a treatment intervention (with participants assigned to a treatment or control group). This design can be analysed using a model in which treatment effects on post-intervention hypertension are modelled while covarying baseline levels of hypertension. In this scenario the independence of treatment and covariate variables means that baseline levels of hypertension are equal in the different treatment groups. According to Senn (2006), the idea that ANCOVA is biased unless treatment groups are equal on the covariate applies only when there is *temporal additivity*. To use our hypertension example, temporal additivity is the assumption that both treatment groups would experience the same change in hypertension over time if the treatment had no effect. In other words, had we left the two groups alone, their hypertension would change by exactly the same amount. Given that the groups have different overall levels of hypertension to begin with, this assumption might not be reasonable, which undermines the argument for requiring group equality in baseline measures.

To sum up, the independence of the covariate and treatment makes interpretation more straightforward but is not a statistical requirement. ANCOVA can be unbiased when groups differ on levels of the covariate, but it creates an interpretational problem that ANCOVA cannot magic away.



putting the covariate into the analysis will not 'control for' or 'balance out' those differences (Lord, 1967, 1969). This situation arises mostly when participants are not randomly assigned to experimental treatment conditions. For example, anxiety and depression are closely correlated (anxious people tend to be depressed), so if you wanted to compare an anxious group of people against a non-anxious group on some task, the chances are that the anxious group would also be more depressed than the non-anxious group. You might think that by adding depression as a covariate

into the analysis you can look at the 'pure' effect of anxiety, but you can't. This situation matches part C of Figure 13.2 because the effect of the covariate (depression) would contain some of the variance from the effect of anxiety. Statistically speaking, all that we know is that anxiety and depression share variance; we cannot separate this shared variance into 'anxiety variance' and 'depression variance', it will always be 'shared'. Another common example is if you happen to find that your experimental groups differ in their ages. Placing age into the analysis as a covariate will not solve this problem – it is still confounded with the experimental manipulation. The use of covariates cannot solve this problem (see Jane Superbrain Box 13.1).

This problem can be avoided by randomizing participants to experimental groups, or by matching experimental groups on the covariate (in our anxiety example, you could try to find participants for the low-anxiety group who score high on depression). We can see the degree to which this problem exists by checking whether experimental groups differ on the covariate before fitting the model. To use our anxiety example again, we could test whether our high- and low-anxiety groups differ on levels of depression. If the groups do not significantly differ then we might consider it reasonable to use depression as a covariate.

13.5.2 Homogeneity of regression slopes ©

When a covariate is used to adjust means we fit a model that includes at least one categorical predictor and one continuous predictor (the covariate) but we do not look at the combined effect of these predictors (the interaction between them). For example, remember that the main example in this chapter looks at whether different doses of puppy therapy (a categorical predictor) affect happiness when including love of puppies as a covariate (continuous predictor). If we simplify Eq. 13.1 to ignore the dummy variables, the model looks like this:

$$\begin{aligned}\text{happiness}_i &= \hat{b}_0 + \hat{b}_1\text{categorical predictor}_i + \hat{b}_2\text{covariate}_i + e_i \\ &= \hat{b}_0 + \hat{b}_1\text{dose}_i + \hat{b}_2\text{puppy love}_i + e_i\end{aligned} \tag{13.9}$$

This model is a parallel slopes model. That is, we estimate a single slope parameter for the covariate (puppy love), and in doing so we model the relationship between the covariate and the outcome as though it is the same in each category of the categorical predictor variable. Figure 13.3 (left) shows the model we are fitting in Eq. 13.9. The predicted values of happiness are shown by the three lines. Notice that the three lines are parallel (hence the name of the parallel slopes model). The slope of each of the three lines is the same and its value will be $\hat{b}_2$ from Eq. 13.9. This makes sense because the model contains a single slope parameter for the covariate ($\hat{b}_2$) so the predicted values form the model must follow this slope irrespective of the treatment the person receives. However, the treatment received affects the vertical height of the line. For example, the line for the no puppy group is lower than the lines for the other two groups. To sum up, with a parallel slopes model the predicted values from the model are constrained such that their relationship to the covariate (their slope) is the same in each group (and equal to $\hat{b}$ for the covariate), but group membership affects the average magnitude the predicted values.

Therefore, when we fit a parallel slopes model we assume that the relationship between the covariate and the outcome holds across different groups. This is known as the assumption of **homogeneity of regression slopes** and it means that the relationship between the outcome and the covariate is the same in each category of the categorical predictor. The assumption is illustrated in Figure 13.3 (left), which shows the parallel slopes model for the puppy data: the relationship between love of puppies and happiness is identical in the no puppy, 15-minute and 30-minute treatment groups. To sum up, when fitting a model such as in Eq. 13.9 (which is typically what people

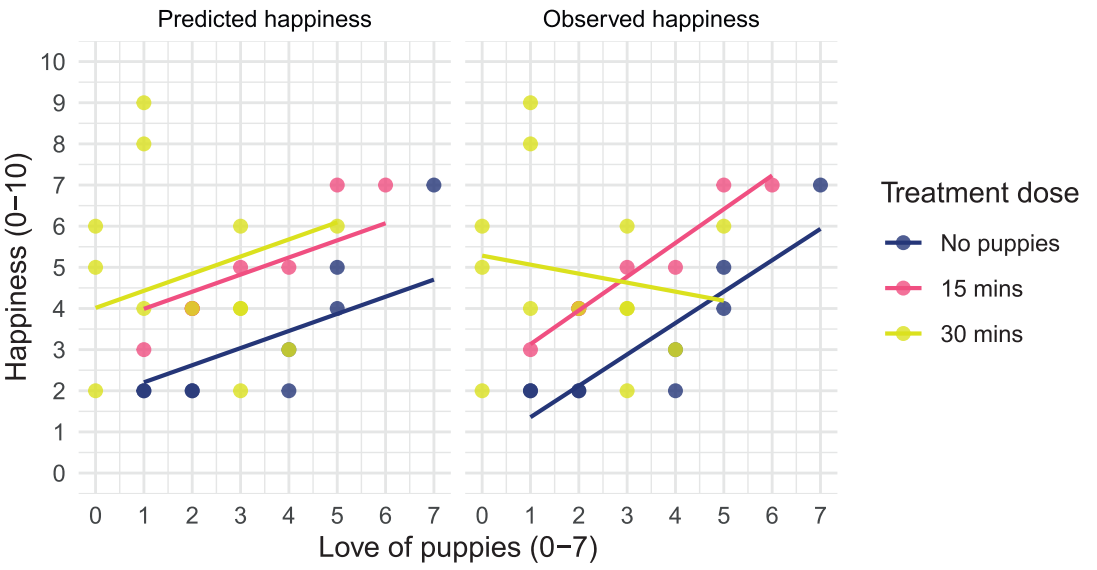


Figure 13.3 Scatterplot and linear models of happiness against love of puppies for each therapy condition

are doing when they talk about using 'ANCOVA') you are assuming homogeneity of regression slopes, so you need to check whether this assumption is reasonable. If it's not reasonable then you're fitting the wrong model.

Figure 13.3 (right) shows the relationship between puppy love and observed happiness in each of the treatment groups. These lines reflect the happiness scores we observed rather than those predicted from the model, which are in the left panel. Put another way, Figure 13.3 (left) shows what the model we're fitting assumes, whereas Figure 13.3 (right) shows the data we observed. Compare the two panels. The right-hand panel shows a positive relationship between love of puppies and participants' happiness in both the control and 15-minute conditions (the lines slope upwards from left to right). In fact, the slopes of the lines for these two groups are very similar, showing that the relationship between happiness and love of puppies is very similar in these two groups. These two groups appear to have *homogeneity* of regression slopes. However, in the 30-minute condition there is a slightly negative relationship between love of puppies and happiness. The slope of this line differs from the slopes in the other two groups, suggesting *heterogeneity of regression slopes* (the relationship between love of puppies and happiness is different in the 30-minute group compared to the other two groups). In the current data, then, the parallel slopes model does not seem reasonable.

In discussions of 'ANCOVA', heterogeneity of regression slopes is seen as a bad thing (Jane Superbrain Box 13.2). However, this isn't an issue with the linear model, it's an issue with what you're trying to get out of the model. If homogeneity of regressions slopes is not a reasonable assumption then you would fit a model that captures how the slope for the covariate varies across groups. This non-parallel slopes model is given by Eq. 13.10:

$$\begin{aligned}\text{happiness}_i &= \hat{b}_0 + \hat{b}_1\text{categorical predictor}_i + \hat{b}_2\text{Covariate}_i + \hat{b}_3\text{interation}_i + e_i \\ &= \hat{b}_0 + \hat{b}_1\text{dose}_i + \hat{b}_2\text{puppy love}_i + \hat{b}_3(\text{dose} \times \text{puppy love})_i + e_i\end{aligned} \tag{13.10}$$

which differs from Eq. 13.9 in that it contains the interaction between the covariate and categorical predictor. This interaction has a parameter estimate ($\hat{b}_3$) that captures the degree to which the slope for the covariate in a particular group differs from the overall slope for the covariate ($\hat{b}_2$).



Jane Superbrain 13.2
What are the consequences of violating the assumption of homogeneity of regression slopes? Ⓑ

When the assumption of homogeneity of regression slopes is met, the resulting *F*-statistic can be assumed to have the corresponding *F*-distribution; however, when the assumption is not met it can't, meaning that the resulting *F*-statistic is being evaluated against a distribution different than the one that it actually has. Consequently, the Type I error rate of the test is inflated and the power to detect effects is not maximized (Hollingsworth, 1980). This is especially true when group sizes are unequal (Hamilton, 1977) and when the standardized regression slopes differ by more than 0.4 (Wu, 1984).



Although the parallel slopes model is used a lot in experimental research, there are numerous situations where you might expect slopes to differ across groups and that variability is interesting. As such, heterogeneity of regression slopes is not a bad thing *per se*, but it does mean that you should fit a model that captures the variability in slopes across groups.

13.5.3 What to do when assumptions are violated Ⓑ

When assumptions are violated you do the same as for any other linear model: fit a robust model. A bootstrap for the model parameters and *post hoc* tests can be used so that these, at least, are robust (see Chapter 6). The bootstrap won't help for the *F*-tests though. There is a robust variant of ANCOVA that can be implemented using R, and we'll discuss this in Section 13.9.

13.6 Conducting ANCOVA using SPSS Ⓑ

13.6.1 General procedure Ⓐ

The general procedure is much the same as for any linear model, so remind yourself of the steps for fitting a linear model (Chapter 9). Figure 13.4 shows a simpler overview of the process that highlights some of the specific issues for ANCOVA-style models. As with any analysis, begin by plotting the data and looking for and correcting sources of bias.

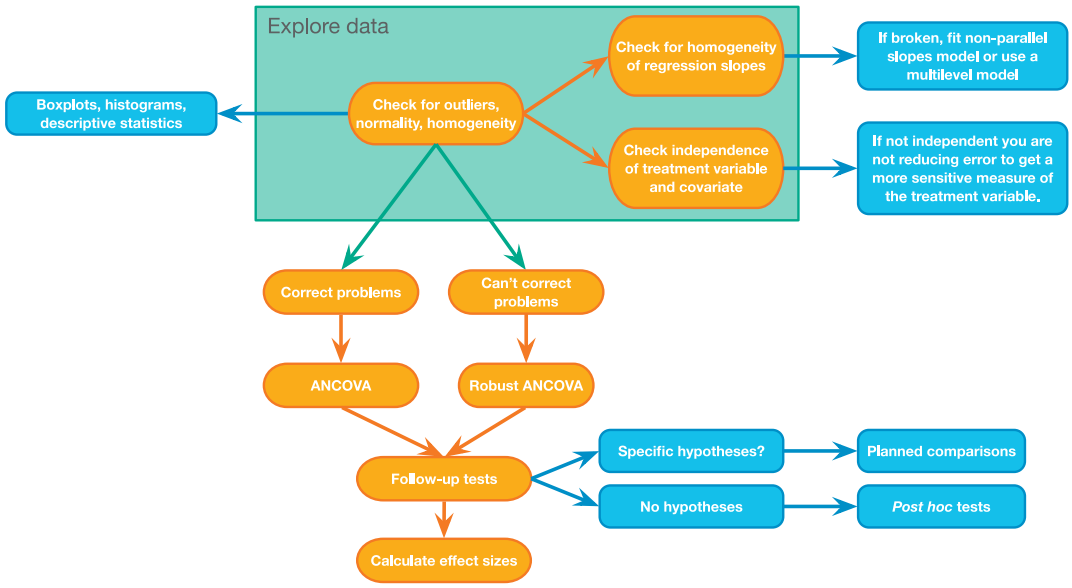


Figure 13.4 General procedure for analysis of covariance

13.6.2 Inputting data Ⓐ

We have already looked at the data (Table 13.1) and the data file (**puppy_love.sav**). To remind you, the data file contains three columns: a coding variable called **dose** (1 = control, 2 = 15 minutes, 3 = 30 minutes), a variable called **happiness** containing the scores for the person's happiness and a variable called **puppy_love** containing the scores for love of puppies from 0 to 7. The 30 rows correspond to each person's scores on these three variables.

13.6.3 Testing the independence of the treatment variable and covariate Ⓑ

In Section 13.5.1, I mentioned that if the covariate and group means (independent variable) are independent then the interpretation of ANCOVA models is more straightforward. In this case, the covariate is love of puppies, so we'd want to check that the mean level of love of puppies is roughly equal across the three puppy therapy groups by fitting a linear model with **puppy_love** as the outcome and **dose** as the predictor.



Output 13.3 shows that the main effect of dose is not significant, $F(2, 27) = 1.98$, $p = 0.16$, which shows that the average level of love of puppies was roughly the same in the three puppy therapy groups. In other words, the means for love of puppies in Table 13.2 are not significantly different across the control, 15- and 30-minute groups. If our aim was to include puppy love (the covariate) to reduce the error variance and gain a more sensitive measure of the effect of puppy therapy then

this result is good news because it suggests that puppy love is independent of the dose of puppy therapy. However, it's worth noting that a non-significant result here does not mean that the therapy group is completely independent of puppy love, just that there is a not a significant link between them in a sample of the size we have. A large enough sample will have yielded a significant difference in the covariate across our treatment groups, and large differences in the covariate across groups will be non-significant in small samples. So, as ever, apply some common sense to this method as a test of independence.

ANOVA

Love of puppies (0–7)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	12.769	2	6.385	1.979	.158
Within Groups	87.097	27	3.226		
Total	99.867	29			

Output 13.3

13.6.4 The main analysis Ⓑ

Most of the *General Linear Model* (GLM) procedures in SPSS contain the facility to include one or more covariates. For designs that don't involve repeated measures it is easiest to include covariates by selecting *Analyze* > *General Linear Model* > Univariate... to activate the dialog box in Figure 13.5. Drag the variable **happiness** into the box labelled *Dependent Variable* (or click), drag **dose** into the box labelled *Fixed Factor(s)* and drag **puppy_love** into the box labelled *Covariate(s)*.

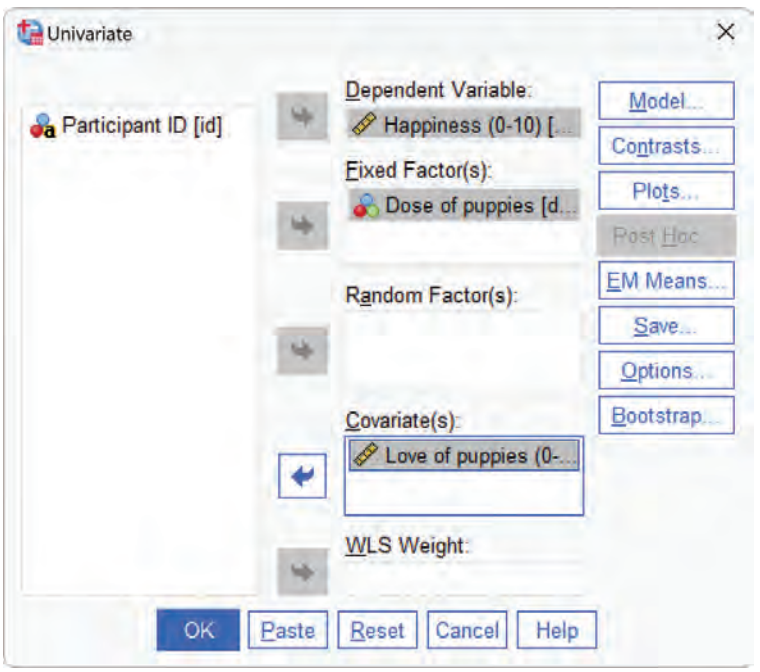


Figure 13.5 Main dialog box for GLM univariate

13.6.5 Contrasts Ⓑ

There are various dialog boxes that can be accessed from the main dialog box. If a covariate is selected, the *post hoc* tests are disabled because the tests that we used in the previous chapter are not designed for models that include covariates. However, comparisons can be done by clicking Contrasts... to access the *Contrasts* dialog box in Figure 13.6. You cannot enter codes to specify user-defined contrasts (but see SPSS Tip 13.1); instead you can select one of the standard contrasts that we met in Table 12.8. In this example, there was a control condition (coded as the first group), so a sensible set of contrasts would be simple contrasts comparing each experimental group to the control (this results in the same contrasts as dummy coding). Click the drop-down list (None) and select a type of contrast (in this case *Simple*) from this list. For simple contrasts you need to specify the reference category (i.e., the category against which all other groups are compared). By default, the first category is used (First), which for our data is the no puppies group. This group makes sense as a control because it is the group that had no treatment. By using the first category as the reference we will get contrasts that compare the 15- and 30-minute groups to the group that had no therapy. We will leave the default, but for your own data, you can alternatively choose to use the last category as the reference category by selecting Last. For the current data this change would compare each group to the 30-minute group. Figure 13.6 shows the completed dialog box. Having selected a contrast, click Continue to return to the main dialog box.

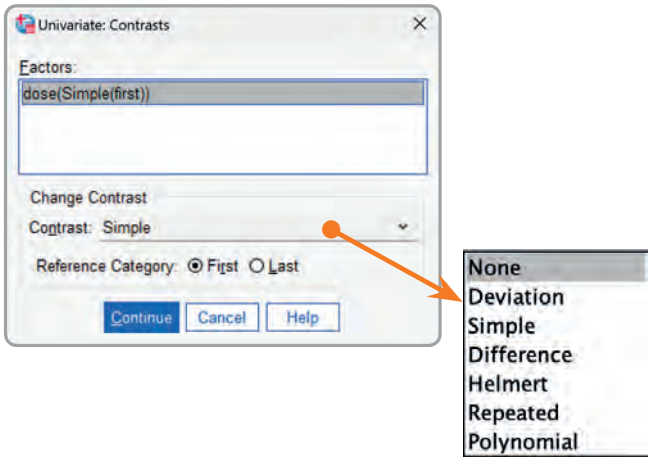


Figure 13.6 Options for standard contrasts in GLM univariate

13.6.6 Other options Ⓑ

You can get a limited range of *post hoc* tests by clicking on EM Means... to access the *Estimated Marginal Means* dialog box (see Figure 13.7). To specify *post hoc* tests, drag the independent variable (in this case **dose**) from the box labelled *Factor(s) and Factor Interactions* and to the box labelled *Display Means for* (or click ▶). Once a variable has been transferred, you'll be able to select ☑ Compare main effects to activate the drop-down list (LSD(none)) of *post hoc* tests. The default is to perform a Tukey LSD *post hoc* test which makes no adjustment for multiple tests (and which I don't recommend). The other options are a Bonferroni *post hoc* test (recommended) and a **Šidák correction**, which is like the Bonferroni correction but is less conservative and so should be selected if you are concerned about the loss of power associated with Bonferroni. For this example we'll use the Šidák correction just for



SPSS Tip 13.1
Planned contrasts for ANCOVA ©

There is no option for specifying planned contrasts like we used in the previous chapter (see Section 12.8.2). However, these contrasts can be done if we fit the model using the regression menu. Imagine you chose some planned contrasts as in Chapter 12, in which the first contrast compared the control group to all doses of puppy therapy, and the second contrast then compared the 30- and 15-minute groups. We saw in Sections 12.5 and 12.8.2 that we specify these contrasts with codes. For the first contrast we discovered that an appropriate set of codes was −2 for the control group and then 1 for both the 30- and 15-minute groups. For the second contrast the codes were 0 for the control group, −1 for the 15-minute group and 1 for the 30-minute group (see Table 12.4). To do these contrasts when a covariate is included in the model, enter these values as two dummy variables. In other words, add a column called **puppies_vs_none** in which every person in the control group has a value of −2 and all other participants have a value of 1. Add a second column called **long_vs_short**, in which everyone in the control group has the value 0, everyone in the 15-minute group has the value −1 and those in the 30-minute group have a value of 1. The file **puppy_love_contrast.sav** includes these dummy variables.

Run the analysis as described in Section 13.3. The model summary and ANOVA table for the model will be identical to Output 13.1 (because we've fitted the same model as before; the only difference is how the model variance is subsequently broken down with the contrasts). The $\hat{b}$-values for the dummy variables will be different than before because we've specified different contrasts. Output 13.4 shows the model parameters. The first dummy variable compares the control group with the 15- and 30-minute groups. As such, it compares the adjusted mean of the control group (2.93) with the average of the adjusted means for the 15- and 30-minute groups ((4.71+5.15)/2 = 4.93). The $\hat{b}$-value for the first dummy variable should reflect the difference between these values: 4.93 − 2.93 = 2. We discovered in a rather complex and boring bit of Section 12.5 that this value gets divided by the number of groups within the contrast (i.e., 3)

Coefficients[a]

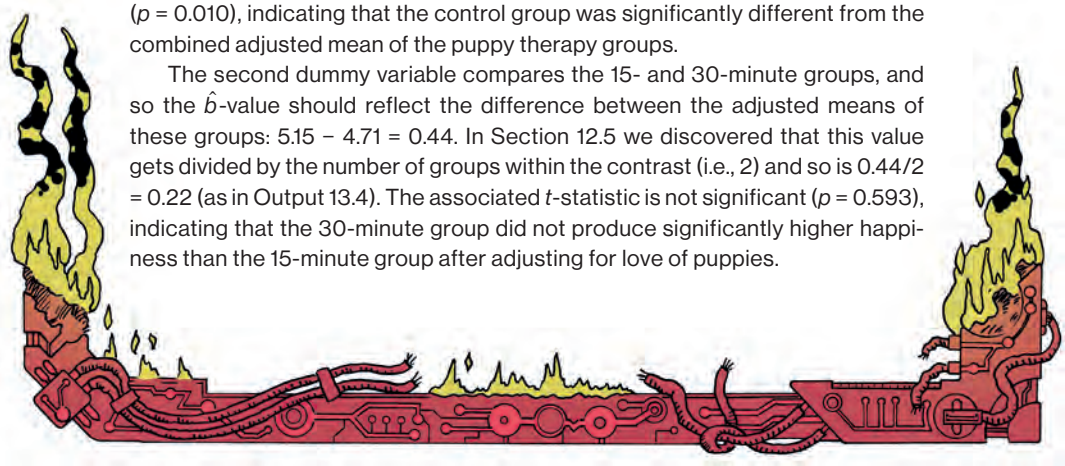
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3.657	.634		5.764	<.001
	Love of puppies (0–7)	.260	.193	.246	1.345	.189
2	(Constant)	3.126	.625		5.002	<.001
	Love of puppies (0–7)	.416	.187	.395	2.227	.035
	Puppy therapy vs. none (contrast 1)	.668	.240	.478	2.785	.010
	30 mins vs 15 mins (contrast 2)	.220	.406	.094	.541	.593

a. Dependent Variable: Happiness (0–10)

Output 13.4

and so is 2/3 = 0.67 (as in Output 13.4).[4] The associated *t*-statistic is significant (p = 0.010), indicating that the control group was significantly different from the combined adjusted mean of the puppy therapy groups.

The second dummy variable compares the 15- and 30-minute groups, and so the $\hat{b}$-value should reflect the difference between the adjusted means of these groups: 5.15 − 4.71 = 0.44. In Section 12.5 we discovered that this value gets divided by the number of groups within the contrast (i.e., 2) and so is 0.44/2 = 0.22 (as in Output 13.4). The associated *t*-statistic is not significant (p = 0.593), indicating that the 30-minute group did not produce significantly higher happiness than the 15-minute group after adjusting for love of puppies.



variety (we have used Bonferroni in previous examples). As well as producing *post hoc* tests for the **dose** variable, the options we've selected will create a table of estimated marginal means for this variable: these are the group means adjusted for the effect of the covariate. Click Continue.

Clicking Options... opens a dialog box to containing the options described in Jane Superbrain Box 13.3. The most useful are (in my opinion) descriptive statistics, estimates of effect size, parameter estimates, residual plot and HC4 robust standard errors (see Figure 13.7).

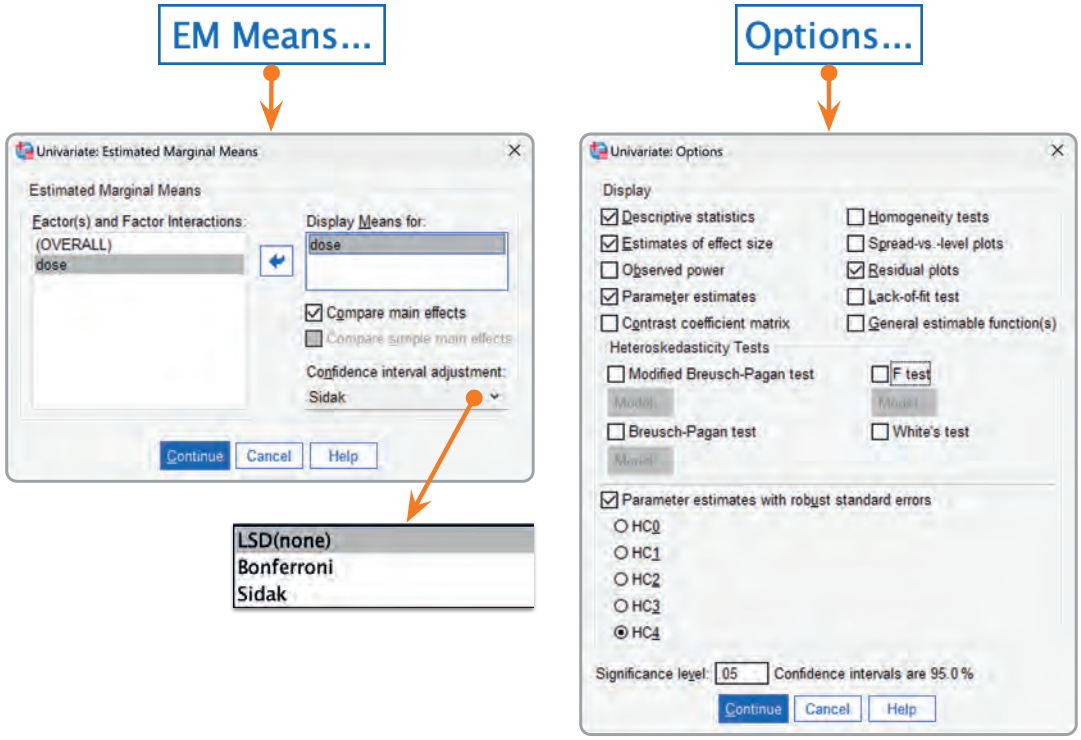


Figure 13.7 *Estimated Marginal Means* and *Options* dialog boxes for GLM univariate

4 The output contains the value 0.668 rather than 0.67. This difference is because we've rounded values to 2 decimal places in our calculations, whereas SPSS retains all decimal places in its calculations.



Jane Superbrain 13.3
Options for ANCOVA Ⓑ

The options in this dialog box are as follows:

- *Descriptive statistics*: This option produces a table of means and standard deviations for each group.
- *Estimates of effect size*: This option produces the value of partial eta squared (partial η^2) – see Section 13.4.
- *Observed power*: This option provides an estimate of the probability that the statistical test could detect the difference between the observed group means. This measure is pointless because if the *F*-test is significant then the probability that the effect was detected will, of course, be high. Likewise, if group differences were small, the observed power would be low. Do power calculations before the experiment is conducted, not after (see Section 2.10.9).
- *Parameter estimates*: This option produces a table of model parameters ($\hat{b}$-values) and their tests of significance for the variables in the model (see Section 13.7.2).
- *Contrast coefficient matrix*: This option produces matrices of the coding values used for any contrasts in the analysis, which is useful for checking which groups are being compared in which contrast.
- *Homogeneity tests*: This option produces Levene's test of the homogeneity of variance assumption (see Section 9.3). You'll have seen by now that I don't recommend this test.
- *Spread-vs.-level plots*: This option produces a chart that plots the mean of each group of a factor (*x*-axis) against the standard deviation of that group (*y*-axis). This plot is useful to check that there is no relationship between the mean and standard deviation. If a relationship exists then the data may need to be stabilized using a logarithmic transformation (see Chapter 6).
- *Residual plots*: This option produces a matrix scatterplot of all combinations of pairs of the following variables: observed values of the outcome, predicted values from the model, and standardized residuals from the model. These plots are used to assess the assumption of homoscedasticity. In particular, the plot of the standardized residuals against the predicted values from the model can be interpreted in a similar way to the *zpred vs. zresid* plot that we have discussed before.
- *Heteroskedasticity Tests*: There are four tests for heteroscedasticity that you can select (two variants of the Breusch–Pagan test, White's test and an *F*-test). For the same reasons that I don't recommend Levene's test, I also don't recommend these – that is, because they are significance tests your decisions based on them will be confounded by your sample size.
- *Parameter estimates with robust standard errors*: This option produces uses one of five methods (HC0 to HC4) to estimate standard errors (and, therefore, confidence intervals) for the model parameters that are robust to heteroscedasticity. These methods are described clearly in Hayes and Cai (2007). In short, HC3 has been shown to outperform HC0 to HC2 (Long & Ervin, 2000) but HC4 outperforms HC3 in some circumstances (Cribari-Neto, 2004). Basically, choose HC3 or HC4.



13.6.7 Bootstrapping and plots Ⓑ

There are other options available from the main dialog box. For example, if you have several independent variables you can plot them against each other (which is useful for interpreting interaction effects – see Section 14.9.3). There's also the Bootstrap... button, which you can use to activate bootstrapping. Selecting this option will bootstrap confidence intervals around the estimated marginal means, parameter estimates ($\hat{b}$-values) and *post hoc* tests, but not the main *F*-statistic. Select the options described in Section 6.10.8 and click OK in the main dialog box to run the analysis.



Oditi's Lantern: ANCOVA

'I, Oditi, have discovered that covariates give us greater control. I like control, especially controlling people's minds and making them worship me, erm, I mean controlling people's minds for the benevolent purpose of helping them to seek truth and personal enlightenment. As long as they are personally enlightened to worship me. In any case, stare into my lantern to discover more about using covariates and ANCOVA.'



13.7 Interpreting ANCOVA Ⓑ

13.7.1 What happens when the covariate is excluded? Ⓑ



Fit the model without the covariate to see whether the three groups differ in their levels of happiness.

Output 13.5 shows (for illustrative purposes) the ANOVA table for these data when the covariate is not included. Let's assume we're using a 0.05 alpha level. The significance value, which is greater than 0.05, shows that that puppy therapy has no significant effect on happiness. Note that the total

Tests of Between-Subjects Effects

Dependent Variable: Happiness (0–10)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16.844[a]	2	8.422	2.416	.108
Intercept	535.184	1	535.184	153.522	<.001
dose	16.844	2	8.422	2.416	.108
Error	94.123	27	3.486		
Total	683.000	30			
Corrected Total	110.967	29			

a. R Squared = .152 (Adjusted R Squared = .089)

Output 13.5

amount of variation in happiness (SS_T) was 110.97 (*Corrected Total*), of which the therapy condition accounted for 16.84 units (SS_M), while 94.12 were unexplained (SS_R).

13.7.2 The main analysis Ⓑ



The format of the ANOVA table in Output 13.6 is largely the same as without the covariate, except that there is an additional row of information about the covariate (**puppy_love**) and an extra column containing the effect size (because we selected this option in Figure 13.7). Looking first at the significance values in the column labelled *Sig.*, the covariate significantly predicts the dependent variable (p = 0.035, which is less than 0.05). Therefore, the person's happiness is significantly influenced by their love of puppies. What's more interesting is that after adjusting for the effect of love of puppies, the effect of puppy therapy is significant (p = 0.027, which is less than 0.05). The amount of variation accounted for by puppy therapy has increased to 25.19 units and the unexplained variance (SS_R) has been reduced to 79.05 units. Notice that SS_T has not changed; all that has changed is how that total variation is partitioned.[5]

The final column of Output 13.6 contains the effect size (partial eta squared) for each variable. These effect sizes are calculated by plugging the sums of squares in the column labelled *Type III Sum of Squares* into Eq. 13.8 (Section 13.4). Using the sums of squares for the effect of dose (25.19), the covariate (15.08) and the error (79.05), we get the following values:

$$\text{Partial } \eta^2_{\text{Dose}} = \frac{SS_{\text{Dose}}}{SS_{\text{Dose}} + SS_{\text{Residual}}} = \frac{25.19}{25.19 + 79.05} = \frac{25.19}{104.24} = 0.24$$

$$\text{Partial } \eta^2_{\text{Puppy love}} = \frac{SS_{\text{Puppy love}}}{SS_{\text{Puppy love}} + SS_{\text{Residual}}} = \frac{15.08}{15.08 + 79.05} = \frac{15.08}{94.13} = 0.16 \quad (13.11)$$

5 The *Corrected Model* in this table is the fit of the model overall (i.e., the model containing the intercept, puppy_love and dose). Note that the SS of 31.92, *df* of 3, *F* of 3.5 and *p* of 0.03 are identical to the values in Output 13.1 (model 2), which tested the overall fit of this model when we ran the analysis as a regression.

Tests of Between-Subjects Effects

Dependent Variable: Happiness (0–10)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	31.920[a]	3	10.640	3.500	.030	.288
Intercept	76.069	1	76.069	25.020	<.001	.490
puppy_love	15.076	1	15.076	4.959	.035	.160
dose	25.185	2	12.593	4.142	.027	.242
Error	79.047	26	3.040			
Total	683.000	30				
Corrected Total	110.967	29				

a. R Squared = .288 (Adjusted R Squared = .205)

Output 13.6

which match those in Output 13.6. These values show that **dose** explains a bigger proportion of the variance (about 24%) not attributable to other variables than **puppy_love** (about 16%).

So far we've seen that when we fit the model without the covariate, puppy therapy does not have a significant effect on happiness, but when the covariate is included it does. This example, therefore, illustrates how covariates can help us to exert stricter experimental control by taking account of confounding variables to give us a 'purer' measure of effect of the experimental manipulation. However, it might also seem odd to you that we're concluding that therapy has an effect on happiness when the 15- and 30-minute groups have very similar mean happiness (4.88 and 4.85) in Table 13.2. Perhaps, you might think, the significant effect reflects a difference between the means in these two therapy groups and the control (which had a lower mean happiness of 3.22). At this point it's worth remembering that we can't use the means in Table 13.2 to interpret the effect of the therapy group: these original means tell us nothing about the group differences reflected by the significant *F*. Instead we need to use the means that have been adjusted for the covariate, which are in Output 13.7 and were obtained by dragging **dose** to *Display Means for* in the *Estimated Marginal Means* dialog box (Figure 13.7, left). The values in Output 13.7 match those calculated in Section 13.3 and we use these values for interpretation. From these adjusted means you can see that happiness increased across the three doses.

Output 13.8 shows the parameter estimates selected in the *Options* dialog box (Figure 13.7, right). and their bootstrapped confidence intervals and *p*-values (bottom table). These estimates result from **dose** being coded using two dummy coding variables. The dummy variables are coded with the last category (the category coded with the highest value in the data editor, in this case the 30-minute group) as the reference category. This reference category (labelled *[dose=3]* in the output) is coded with a 0 for both dummy variables (see Section 12.3.1 for a reminder of how dummy coding works). The

Estimates

Dependent Variable: Happiness (0–10)

Dose of puppies	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound	Bootstrap for Mean[gc] Bias	Bootstrap for Mean[gc] Std. Error	BCa 95% Confidence Interval Lower	BCa 95% Confidence Interval Upper
No puppies	2.926[a]	.596	1.701	4.152	.009	.421	2.190	3.892
15 mins	4.712[a]	.621	3.436	5.988	.030	.370	4.021	5.608
30 mins	5.151[a]	.503	4.118	6.184	.001	.621	4.058	6.455

a. Covariates appearing in the model are evaluated at the following values: Love of puppies (0–7) = 2.73.

gc. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 13.7

Parameter Estimates

Dependent Variable: Happiness (0–10)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	4.014	.611	6.568	<.001	2.758	5.270
puppy_love	.416	.187	2.227	.035	.032	.800
[dose=1]	−2.225	.803	−2.771	.010	−3.875	−.575
[dose=2]	−.439	.811	−.541	.593	−2.107	1.228
[dose=3]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Bootstrap for Parameter Estimates

Dependent Variable: Happiness (0–10)

Parameter	B	Bootstrap[a] Bias	Bootstrap[a] Std. Error	Bootstrap[a] Sig. (2-tailed)	Bootstrap[a] BCa 95% Confidence Interval Lower	Bootstrap[a] BCa 95% Confidence Interval Upper
Intercept	4.014	.038	.815	.002	2.023	5.790
puppy_love	.416	−.024	.192	.046	.015	.686
[dose=1]	−2.225	.007	.719	.016	−3.734	−.765
[dose=2]	−.439	.029	.696	.556	−1.901	.904
[dose=3]	0	0	0		.	.

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 13.8

parameter labelled *[dose=2]* represents the difference between the group coded as 2 (15 minutes) and the reference category (30 minutes), and *[dose=1]* represents the difference between the group coded as 1 (control) and the reference category (30 minutes). The $\hat{b}$-values represent the differences between the adjusted means in Output 13.7 and the significances of the *t*-tests tell us whether these adjusted group means differ significantly. The $\hat{b}$ for *[dose=1]* in Output 13.8 is the difference between the adjusted means for the control group and the 30-minute group, 2.926 – 5.151 = –2.225, and the $\hat{b}$ for *[dose=2]* is the difference between the adjusted means for the 15-minute group and the 30-minute group, 4.712 – 5.151 = –0.439.

The degrees of freedom for the *t*-test of the $\hat{b}$-values are $N - p$ (see Section 6.2.6.), in which N is the total sample size (in this case 30) and p is the number of parameters (in this case 4, the intercept and one each for the two dummy variables and the covariate). For these data, $df = 30 - 4 = 26$. Based on the bootstrapped significance and confidence intervals (remember you'll get different values than me because of how bootstrapping works), we could conclude that the 30-minute group differs significantly from the control group, $p = 0.021$ (*[dose=1]* in the table), but not from the 15-minute group, $p = 0.558$, (*[dose=2]* in the table). The conclusions about **dose** using bootstrapping are consistent with the non-robust model.

The final thing to note is the value of $\hat{b}$ for the covariate (0.416), which is the same as in Output 13.2 (when we ran the analysis through the regression menu). This value tells us that if love of puppies increases by one unit, then the person's happiness should increase by just under half a unit (although there is nothing to suggest a causal link between the two); because the coefficient is positive, as love of puppies increases, so does happiness. A negative coefficient would mean the opposite: as one

Parameter Estimates with Robust Standard Errors

Dependent Variable: Happiness (0–10)

Parameter	B	Robust Std. Error[a]	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	4.014	.805	4.989	<.001	2.360	5.668
puppy_love	.416	.190	2.187	.038	.025	.807
[dose=1]	-2.225	.690	-3.226	.003	-3.642	-.807
[dose=2]	-.439	.695	-.632	.533	-1.868	.990
[dose=3]	0[b]	.	.	.	.	.

a. HC4 method

b. This parameter is set to zero because it is redundant.

Output 13.9

variable increases, the other decreases. This relationship is significant, $p = 0.035$, and has a confidence interval that does not contain zero (0.032, 0.800). The conclusions from the bootstrap procedure are consistent with the non-robust model because it yields a $p = 0.046$ and also produces a confidence interval that does not contain zero (0.015, 0.686).[6]

Output 13.9 repeats the parameter estimates from Output 13.8 but with standard errors, *p*-values and confidence intervals robust to heteroscedasticity (the HC4 estimates that we asked for). We can interpret the effects for **dose** in the same way as for the regular and bootstrap *p*-values and confidence intervals; that is the 30-minute group differs significantly from the control group, $p = 0.003$, but not from the 15-minute group, $p = 0.533$. For the effect of **puppy_love**, The HC4 robust confidence interval and *p*-value supports the conclusion from the non-robust model: the *p*-value is 0.038, which less than 0.05, and the confidence interval does not contain zero (0.025, 0.807).

13.7.3 Contrasts and *post hoc* tests Ⓑ

Output 13.10 shows the result of the contrast analysis specified in Figure 13.6 and compares level 2 (15 minutes) against level 1 (control) as a first comparison, and level 3 (30 minutes) against level 1 (control) as a second comparison. The group differences are quantified using a difference value, its standard error, significance value and 95% confidence interval. These results show that both the 15-minute group (contrast 1, $p = 0.045$) and 30-minute group (contrast 2, $p = 0.010$) had significantly different happiness compared to the control group (note that the parameter of contrast 2 is the same magnitude as that for *[dose=1]* in the previous section).

Output 13.11 shows the results of the Šidák corrected *post hoc* comparisons that were requested in Section 13.6.6. The bottom table shows the bootstrapped significance and confidence intervals for these tests, and because these will be robust we'll interpret this table (again, remember your values will differ because of how bootstrapping works). There is a significant difference between the control group and both the 15- ($p = 0.003$) and 30-minute ($p = 0.016$) groups. The 30- and 15-minute

6 Because of the way bootstrapping works, the *p*-value and confidence interval can change each time you run the procedure. For the effect of **puppy_love**, one time I ran the model I got $p = 0.052$ and a confidence interval that *did* contain zero, which contradicts the conclusions in the main text. Such things remind us of the absurdity of using all-or-nothing thinking when it comes to significance tests: your scientific beliefs are at the mercy of small variations around an arbitrary threshold such as 0.05.

Contrast Results (K Matrix)

Dose of puppies Simple Contrast[a]			Dependent Variable: Happiness (0–10)
Level 2 vs. Level 1	Contrast Estimate		1.786
	Hypothesized Value		0
	Difference (Estimate – Hypothesized)		1.786
	Std. Error		.849
	Sig.		.045
	95% Confidence Interval for Difference	Lower Bound	.040
		Upper Bound	3.532
Level 3 vs. Level 1	Contrast Estimate		2.225
	Hypothesized Value		0
	Difference (Estimate – Hypothesized)		2.225
	Std. Error		.803
	Sig.		.010
	95% Confidence Interval for Difference	Lower Bound	.575
		Upper Bound	3.875

a. Reference category = 1

Output 13.10

Pairwise Comparisons

Dependent Variable: Happiness (0–10)

(I) Dose of puppies	(J) Dose of puppies	Mean Difference (I–J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
No puppies	15 mins	−1.786	.849	.130	−3.953	.381
	30 mins	−2.225*	.803	.030	−4.273	−.177
15 mins	No puppies	1.786	.849	.130	−.381	3.953
	30 mins	−.439	.811	.932	−2.509	1.631
30 mins	No puppies	2.225*	.803	.030	.177	4.273
	15 mins	.439	.811	.932	−1.631	2.509

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Sidak.

Bootstrap for Pairwise Comparisons

Dependent Variable: Happiness (0–10)

(I) Dose of puppies	(J) Dose of puppies	Mean Difference (I–J)	Bootstrap[a] Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval Lower	Upper
No puppies	15 mins	−1.786	−.021	.514	.003	−2.737	−.815
	30 mins	−2.225	.007	.719	.016	−3.734	−.765
15 mins	No puppies	1.786	.021	.514	.003	.643	2.842
	30 mins	−.439	.029	.696	.556	−1.901	.904
30 mins	No puppies	2.225	−.007	.719	.016	.976	3.599
	15 mins	.439	−.029	.696	.556	−.830	1.768

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 13.11

groups did not significantly differ ($p = 0.556$). It is interesting that the significant difference between the 15-minute and control groups when bootstrapped ($p = 0.003$) is not present for the normal *post hoc* tests ($p = 0.130$). This anomaly could reflect properties of the data that have biased the non-robust version of the *post hoc* test.

13.7.4 Interpreting the covariate Ⓑ

I've already mentioned that the parameter estimates (Output 13.8) tell us how to interpret the covariate: the sign of the $\hat{b}$-value tells us the direction of the relationship between the covariate and outcome variable. For these data the $\hat{b}$-value was positive, indicating that as the love of puppies increases, so does the participant's happiness. Another way to discover the same thing is to draw a scatterplot of the covariate against the outcome.



Figure 13.8 confirms that the effect of the covariate is that as love of puppies increases, so does the participant's happiness (as shown by the slope of the line).

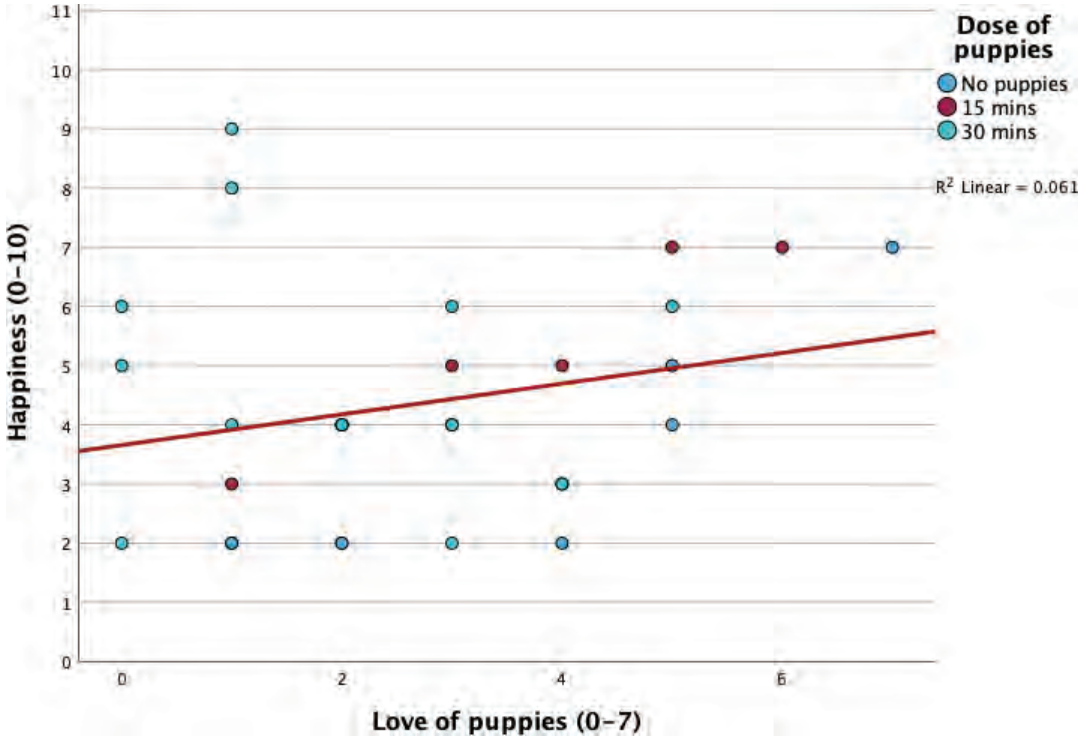
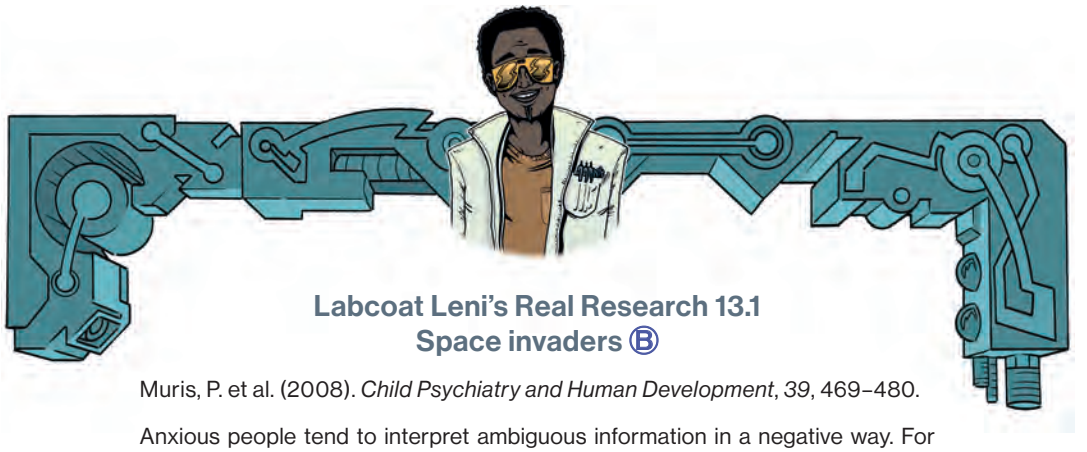


Figure 13.8 Scatterplot of happiness against love of puppies



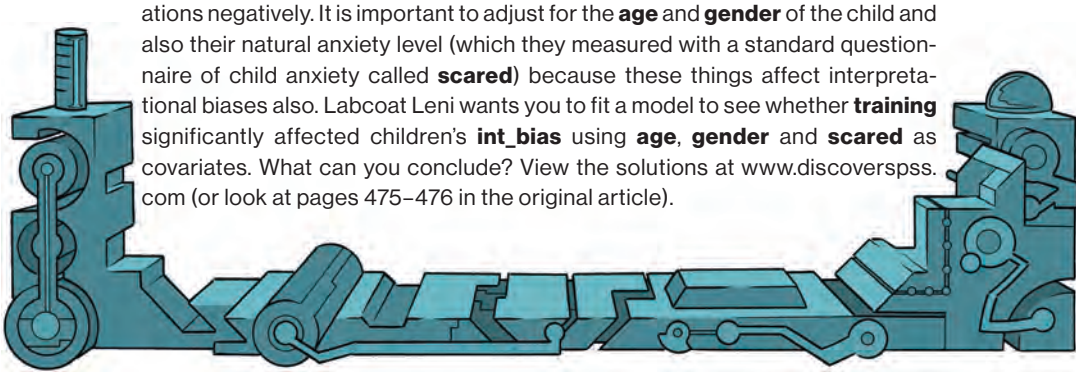
Labcoat Leni's Real Research 13.1
Space invaders Ⓑ

Muris, P. et al. (2008). *Child Psychiatry and Human Development*, *39*, 469–480.

Anxious people tend to interpret ambiguous information in a negative way. For example, being highly anxious myself, if I overheard a student saying 'Andy Field's lectures are really *different*' I would assume that 'different' meant rubbish, but it could also mean 'refreshing' or 'innovative'. Muris et al. (2008) addressed how these interpretational biases develop in children. Children imagined that they were astronauts who had discovered a new planet. They were given scenarios about their time on the planet (e.g., 'On the street, you encounter a spaceman. He has a toy handgun and he fires at you ...') and the child had to decide whether a positive ('You laugh: it is a water pistol and the weather is fine anyway') or negative ('Oops, this hurts! The pistol produces a red beam which burns your skin!') outcome occurred. After each response the child was told whether their choice was correct. Half of the children were *always* told that the negative interpretation was correct, and the remainder were told that the positive interpretation was correct.

Over 30 scenarios children were trained to interpret their experiences on the planet as negative or positive. Muris et al. then measured interpretational biases in everyday life to see whether the training had created a bias to interpret things negatively. In doing so, they could ascertain whether children might learn interpretational biases through feedback (e.g., from parents).

The data from this study are in the file **muris_2008.sav**. The independent variable is **training** (positive or negative) and the outcome is the child's interpretational bias score (**int_bias**) – a high score reflects a tendency to interpret situations negatively. It is important to adjust for the **age** and **gender** of the child and also their natural anxiety level (which they measured with a standard questionnaire of child anxiety called **scared**) because these things affect interpretational biases also. Labcoat Leni wants you to fit a model to see whether **training** significantly affected children's **int_bias** using **age**, **gender** and **scared** as covariates. What can you conclude? View the solutions at www.discoverspss.com (or look at pages 475–476 in the original article).



13.8 The non-parallel slopes model and the assumption of homogeneity of regression slopes Ⓒ

Remember that the assumption of homogeneity of regression slopes means that the relationship between the covariate and outcome variable (in this case **puppy_love** and **happiness**) should be

similar at different levels of the predictor variable (in this case in the three **dose** groups). Figure 13.3 (right) suggested that the relationship between **puppy_love** and **happiness** was comparable in the 15-minute and control groups, but is different in the 30-minute group. In this situation we should fit a non-parallel slopes model. With the usual caveat about using significance tests to make these sorts of decisions, we can also use this model to test whether the assumption of homogeneity of regression slopes is reasonable.

To fit the non-parallel slopes model we need to refit the model but customize it to include the interaction between the covariate and categorical predictor. Access the main dialog box as before and place the variables in the same boxes as before (the finished dialog box should look like Figure 13.5). To customize the model, click Model... to access the dialog box in Figure 13.9 and select Build terms. The variables specified in the main dialog box are listed on the left-hand side. We need a model that includes the interaction between the covariate and grouping variable. It's essential that we also include the main effects, otherwise variance in the outcome (happiness) may be attributed to the interaction term that would otherwise be attributed to the main effects. To begin with, select **dose** and **puppy_love** (you can select both simultaneously by holding down *Ctrl*, or *Cmd* on a Mac), change the drop-down menu to Main effects, and click ➔ to transfer the main effects of **dose** and **puppy_love** to the box labelled *Model*. Next specify the interaction term by selecting **dose** and **puppy_love** simultaneously (as just described), change the drop-down menu to Interaction and click ➔ to transfer the interaction of **dose** and **puppy_love** to the box labelled *Model*. The finished dialog box should look like Figure 13.9. Click Continue to return to the main dialog box and OK to run the analysis.

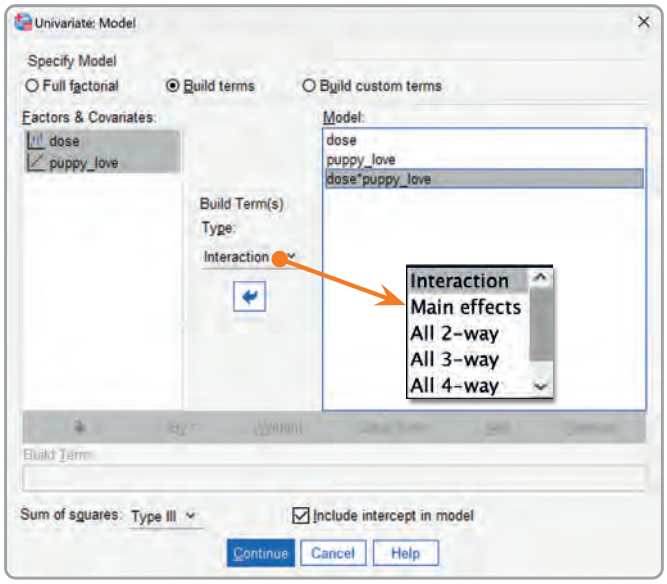


Figure 13.9 *Model* dialog box for GLM univariate

Output 13.12 shows the main summary table for the model including the interaction term. The effects of the dose of puppy therapy and love of puppies are still significant, but so is the covariate by outcome interaction (**dose × puppy_love**), implying that the slope of **puppy_love** (i.e., its relationship to happiness) varies across the three puppy therapy groups. This significant interaction ($p = 0.028$) suggests that the assumption of homogeneity of regression slopes is not realistic and that we should actually interpret this model, because the interaction effect explains a significant amount of variance in happiness (in fact, looking at the values of partial eta squared the interaction effect explains a similar amount of unique variance in happiness to both the therapy group and love of puppies).

Tests of Between-Subjects Effects

Dependent Variable: Happiness (0–10)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	52.346[a]	5	10.469	4.286	.006	.472
Intercept	53.542	1	53.542	21.921	<.001	.477
dose	36.558	2	18.279	7.484	.003	.384
puppy_love	17.182	1	17.182	7.035	.014	.227
dose * puppy_love	20.427	2	10.213	4.181	.028	.258
Error	58.621	24	2.443			
Total	683.000	30				
Corrected Total	110.967	29				

a. R Squared = .472 (Adjusted R Squared = .362)

Output 13.12



Cramming Sam's Tips
Covariates

- When the linear model is used to compare several means adjusted for the effect of one or more other variables (called *covariates*) it can be referred to as analysis of covariance (ANCOVA).
- Before the analysis check that the covariate(s) are independent of any independent variables by seeing whether those independent variables predict the covariate (i.e. the covariate should not differ across groups).
- In the table labelled *Tests of Between-Subjects Effects*, assuming you're using an alpha of 0.05, look to see if the value in the column labelled *Sig.* is less than or equal to 0.05 for both the covariate and the independent variable. If it is for the covariate then this variable has a significant relationship to the outcome variable; if it is for the independent variable then the means (adjusted for the effect of the covariate) are significantly different across categories of this variable.
- If you have generated specific hypotheses before the experiment, use planned contrasts; if not, use *post hoc* tests.
- For parameters and *post hoc* tests, look at the columns labelled *Sig.* to discover if your comparisons are significant (they will be if the significance value is less than or equal to 0.05). Use bootstrapping to get robust versions of these tests.
- In addition to the assumptions in Chapter 6, test for *homogeneity of regression slopes* by customizing the model to look at the independent variable × covariate interaction. If the assumption is not plausible then retain the model with the interaction term.



13.9 Robust ANCOVA

We have already looked at robust confidence intervals and *p*-values for the model parameters that were computed using bootstrapping and heteroscedasticity robust standard errors (Section 13.7.2). In addition, www.discoverspss.com contains a workbook file (**robust_ancova.spwb**) for running a robust variant of ANCOVA (*ancboot*) that works on trimmed means and is described by Wilcox (2017). We need to use the WRS2 package in R (Section 4.12). This test is limited to the situation where the independent variable (the categorical predictor) has two categories, and there is one covariate. But it does enable you to ignore assumptions and get on with your life. Because this syntax only works when you have two groups, I have provided a data file called **puppies_two_group.sav**, which contains the example data for this chapter but excluding the 15-minute condition. This data file compares the control (no puppies) with the 30-minute group (**dose**), and has the scores for the love of puppies covariate too (**puppy_love**). The syntax to run the robust test is as follows:

```
BEGIN PROGRAM R.
library(WRS2)
mySPSSdata = spssdata.GetDataFromSPSS(factorMode = "labels")
ancboot(happiness ~ dose + puppy_love, data = mySPSSdata, tr = 0.2, nboot = 1000)
END PROGRAM.
```

Run this syntax (see SPSS Tip 10.3) to obtain Output 13.13. The test works by identifying values of the covariate for which the relationship between the covariate and outcome are comparable in the

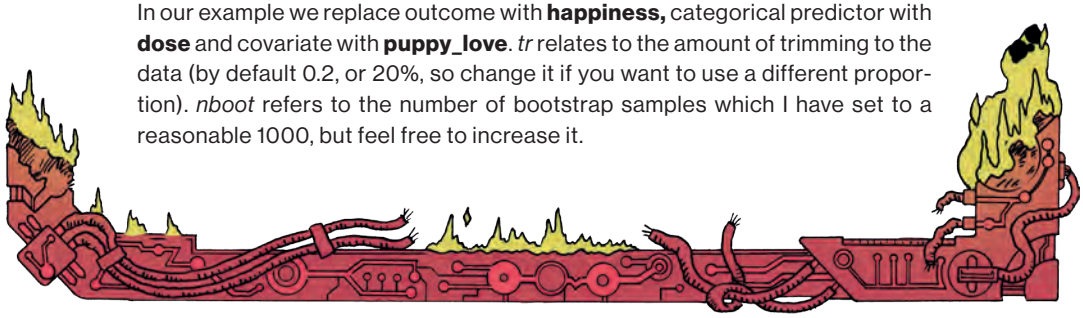


SPSS Tip 13.2
Robust ANCOVA ©

The function *ancboot* takes the form:

```
ancboot(outcome ~ categorical predictor + covariate, data =
mySPSSdata, tr = 0.2, nboot = 1000)
```

In our example we replace outcome with **happiness,** categorical predictor with **dose** and covariate with **puppy_love**. *tr* relates to the amount of trimming to the data (by default 0.2, or 20%, so change it if you want to use a different proportion). *nboot* refers to the number of bootstrap samples which I have set to a reasonable 1000, but feel free to increase it.



```
                n1 n2     diff lower CI upper CI statistic p-value
puppy_love = 2 22 13 -1.0873  -3.0837   0.9091   -1.6952   0.101
puppy_love = 3 27 15 -1.0719  -2.7501   0.6063   -1.9881   0.060
puppy_love = 5 30 22 -0.6508  -2.3612   1.0596   -1.1843   0.264
puppy_love = 6 23 21 -0.9846  -3.2477   1.2785   -1.3542   0.202
puppy_love = 8 12 13 -1.5278  -4.2264   1.1708   -1.7622   0.109
```

Output 13.13

two groups. In this example it identifies five values of **puppy_love** (2, 3, 5, 6, and 8) for which the relationship between love of puppies and happiness is comparable. At each of these design points, we're told the number of cases for the two groups (*n1* and *n2*) that have a value of the covariate (**puppy_love**) close to these design points (not exactly *x*, but close to it). Based on these two samples, trimmed means (20% by default) are computed and the difference between them tested. This difference is stored in the column *Diff* along with the boundaries of the associated 95% bootstrap confidence interval (corrected to control for doing five tests) in the next two columns. The test statistic comparing the difference is in the column *statistic*, with its *p*-value in the final column. Output 13.13 show no significant differences between trimmed means for any of the design points (all *p*-values are greater than 0.05).

13.10 Bayesian analysis with covariates Ⓑ

Because the model we have fitted is a linear model with a categorical predictor and a continuous predictor, you can use what you learned in Section 9.13 to run a Bayesian regression. Select *Analyze* > *Bayesian Statistics* > Linear Regression. You would need to manually create dummy variables (as in the file **puppy_love_dummy.sav**) and drag these to the box labelled *Factor(s)* and drag **puppy_love** to the box labelled *Covariate(s)* as in Figure 13.10. The resulting Output 13.14 shows the Bayes factor for the full model compared to the null model, which is the model including only the intercept. The Bayes factor is 0.651. In other words, the probability of the data given the model including all three predictors is 0.651 that of the probability of the data given the model with only the intercept. This value is less than 1, meaning that we should shift our beliefs *away* from the model and towards the null. Specifically, we should shift our belief in the null (relative to the model) by a factor of 1/0.651 = 1.54. This is weak evidence for the null (and also the model). In short, despite all of the significant results with our frequentist models, the data provide evidence for the model that is too weak to warrant shifting our beliefs away from the idea that puppy therapy doesn't affect happiness.

Output 13.14 (bottom) shows the $\hat{b}$s based on Bayesian estimation in the columns labelled *Posterior Mode* and *Posterior Mean*. The values are 1.786 (for the difference between the 15-minute and control group), 2.225 (for the difference between the 30-minute and control group) and 0.416 for puppy love. Because we used reference priors, the data are doing all of the work in the estimation and the values match those in the non-Bayesian model in Output 13.2. The 95% credible intervals for these parameters contain the population value with a probability of 0.95 (95%) assuming that the effect exists.

The true difference in happiness between the 15-minute and control group budget is between 0.04 (i.e. virtually no difference) and 3.53 (a lot on a 10-point scale) with 95% probability. The true difference in happiness between the 30-minute and control group budget is between 0.56 (i.e. a small difference) and 3.88 (a lot on a 10-point scale) with 95% probability. The true relationship between puppy love and happiness is between 0.03 (i.e. virtually no relationship) and 0.80 (almost a unit shift along the 10-point happiness scale from a unit shift along the 7-point puppy love scale) with 95% probability.

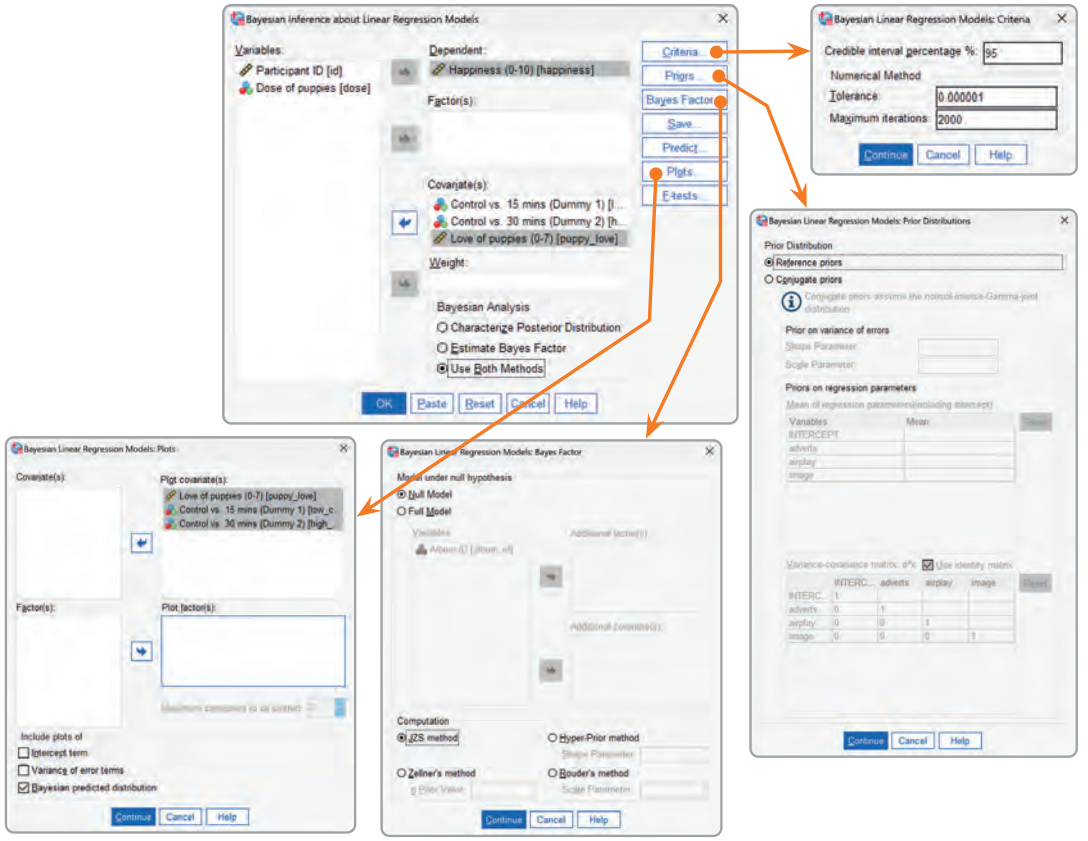


Figure 13.10 Dialog boxes for Bayesian ANCOVA

Bayes Factor Model Summary[a,b]

Bayes Factor[c]	R	R Square	Adjusted R Square	Std. Error of the Estimate
.651	.536	.288	.205	1.74

a. Method: JZS

b. Model: (Intercept), Control vs. 15 mins (Dummy 1), Control vs. 30 mins (Dummy 2), Love of puppies (0–7)

c. Bayes factor: Testing model versus null model (Intercept).

Bayesian Estimates of Coefficients[a,b,c]

Parameter	Posterior Mode	Posterior Mean	Posterior Variance	95% Credible Interval Lower Bound	95% Credible Interval Upper Bound
(Intercept)	1.789	1.789	.815	.007	3.572
Control vs. 15 mins (Dummy 1)	1.786	1.786	.782	.040	3.532
Control vs. 30 mins (Dummy 2)	2.225	2.225	.698	.575	3.875
Love of puppies (0–7)	.416	.416	.038	.032	.800

a. Dependent Variable: Happiness (0–10)

b. Model: (Intercept), Control vs. 15 mins (Dummy 1), Control vs. 30 mins (Dummy 2), Love of puppies (0–7)

c. Assume standard reference priors.

Output 13.14

13.11 Reporting results Ⓑ

When using covariates you can report the model much the same way as any other. For the covariate and the experimental effect give details of the *F*-statistic and the degrees of freedom from which it was calculated. In both cases, the *F*-statistic was derived by dividing the mean squares for the effect by the mean squares for the residual. Therefore, the degrees of freedom used to assess the *F*- statistic are the degrees of freedom for the effect of the model (df_M = 1 for the covariate and 2 for the experimental effect) and the degrees of freedom for the residuals of the model (df_R = 26 for both the covariate and the experimental effect) – see Output 13.6. The correct way to report the main findings would be:

- The covariate, love of puppies, was significantly related to the participant's happiness, $F(1, 26) = 4.96$, $p = 0.035$, $\eta^2 = 0.16$. There was also a significant effect of puppy therapy on levels of happiness after controlling for the effect of love of puppies, $F(2, 26) = 4.14$, $p = 0.027$, partial $\eta^2 = 0.24$.

We can also report some contrasts (see Output 13.8):

- Planned contrasts revealed that having 30 minutes of puppy therapy significantly increased happiness compared to having a control, $t(26) = -2.77$, $p = 0.01$, but not compared to having 15 minutes, $t(26) = -0.54$, $p = 0.59$.

13.12 Jane and Brian's story Ⓐ

The encounter in Blow Your Speakers had been beyond weird. Jane felt terrible. This Brian guy was so nice to her, and she'd just told him where to go – again! It had been easy to dismiss Brian at first, he seemed like a loser, a waste of her time. But there was more to him than that: he'd been working hard to learn statistics, and he'd made impressive progress. She liked how awkward he was around her, and how he always defaulted to talking stats. It was endearing. It could derail her research, though, and he could never know about that. She was a monster, and if he found out the truth it would be another let-down. Best to keep her distance.

The phone rang. It was her brother, Jake. She loved and admired Jake like no one else. Until he left home, he'd been her sanity. Their parents, both highly successful academics, were at home only long enough to pile the pressure on them both to succeed. Jane reacted by spending her youth in books, in a futile pursuit of their attention. Every set of straight As was met with a 'these are just a step towards the exams that really matter, you'll need to up your game'. She was tired of trying to impress them. Jake was her opposite – he'd realized early on that he could never win. He let the pressure roll off him and left home as soon as he could. But he always looked out for Jane.

'Mum is in hospital,' he said as the blood drained from Jane's legs.

'I don't care,' she replied, but she did. She also wanted to see Brian, because he was the closest thing she had to a friend in this town.

13.13 What next? Ⓑ

At the age of 13 I met my heroes, Iron Maiden, and very nice they were too. I've met them a couple of times since (not because they're my best buddies or anything exciting like that, but over the years the fan club has put on various events where you were allowed to stand next to them and gibber like a fool while they humour you politely). You'll notice that the photo at the start of this chapter is signed by Dave Murray. This wasn't possible because I had my own dark room installed backstage at

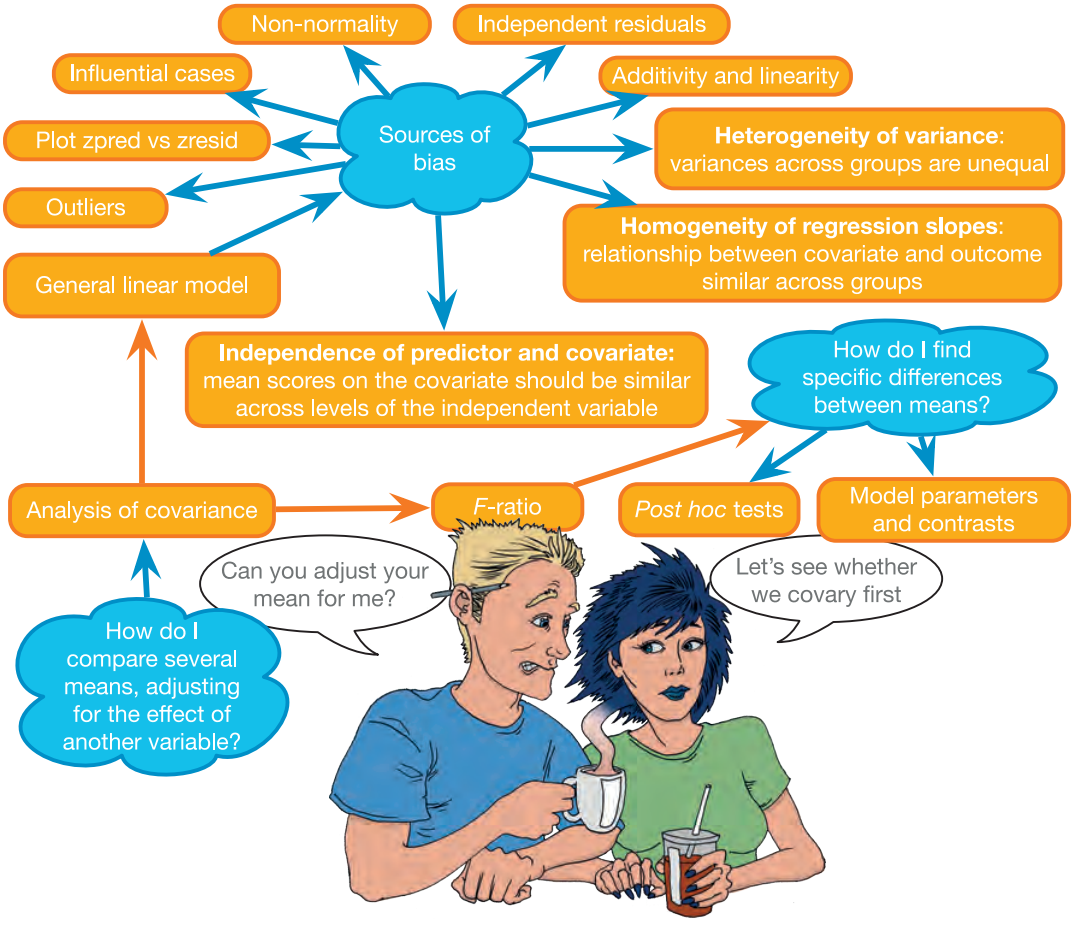


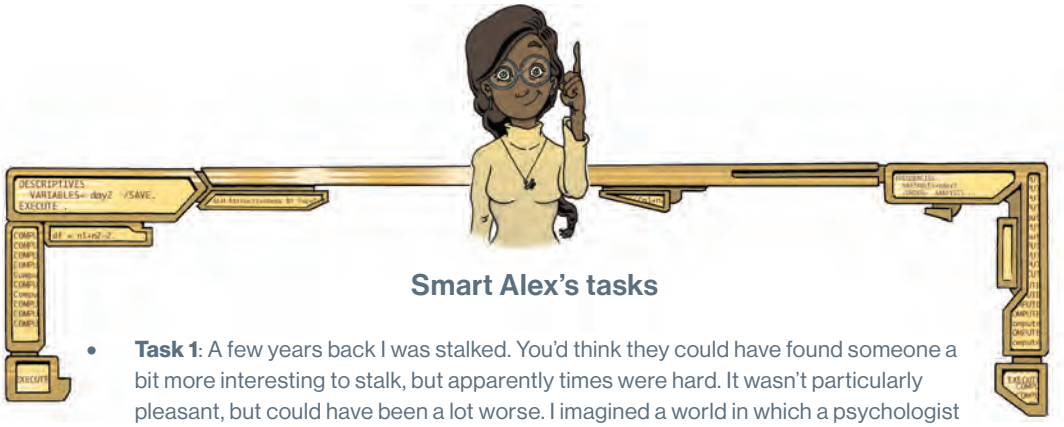
Figure 13.11 What Brian learnt from this chapter

the Hammersmith Odeon in which I could quickly process photographs, or because I had access to time travel (sadly), but because I took the photo with me when I met him again in 2000. I recounted the tale of how terrified I had been about meeting him in 1986. If he thought I was some strange stalker he certainly didn't let on. Uncharacteristically of most people who've sold millions of albums, they're lovely blokes.

Anyway, having seen Iron Maiden in their glory, I was inspired. They still inspire me: I still rate them as the best live band I've ever seen (and I've seen them over 45 times, so I ought to know). Although I had briefly been deflected from my destiny by the shock of grammar school, I was back on track. I *had* to form a band. There was just one issue: no one else played a musical instrument. The solution was easy: through several months of covert subliminal persuasion, I convinced my two best friends (both called Mark, oddly enough) that they wanted nothing more than to start learning the drums and bass guitar. A power trio was in the making.

13.14 Key terms that I've discovered

Adjusted mean
Analysis of covariance (ANCOVA)
Covariate
Homogeneity of regression slopes
Partial eta squared (partial η^2)
Šidák correction

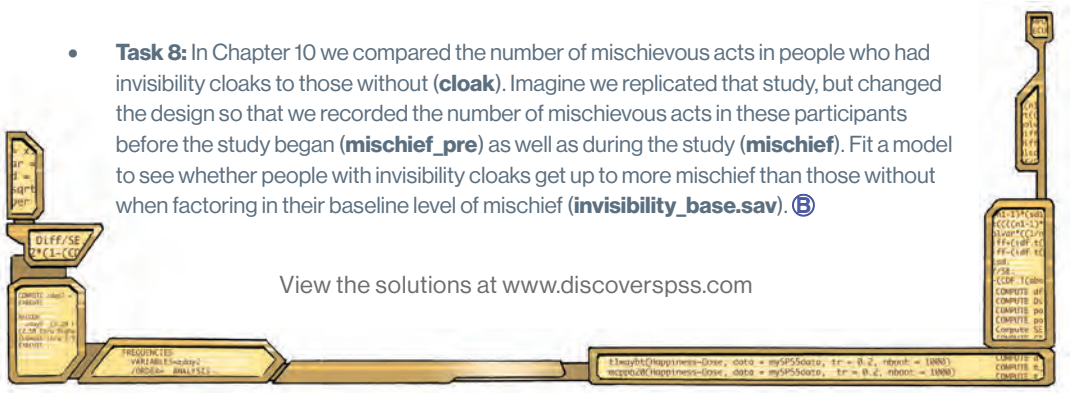


Smart Alex's tasks

- **Task 1**: A few years back I was stalked. You'd think they could have found someone a bit more interesting to stalk, but apparently times were hard. It wasn't particularly pleasant, but could have been a lot worse. I imagined a world in which a psychologist tried two different therapies on different groups of stalkers (25 stalkers in each group – this variable is called **therapy**). To the first group he gave cruel-to-be-kind therapy (every time the stalkers followed him around, or sent him a letter, the psychologist attacked them with a cattle prod). The second therapy was psychodyshamic therapy, in which stalkers were hypnotized and regressed into their childhood to discuss their penis (or lack of penis), their dog's penis, the seventh penis of a seventh penis, and any other penis that sprang to mind. The psychologist measured the number of hours stalking in one week both before (**stalk_pre**) and after (**stalk_post**) treatment (**stalker.sav**). Analyse the effect of therapy on stalking behaviour after therapy, adjusting for the amount of stalking behaviour before therapy. Ⓑ
- **Task 2**: Fit a robust model for Task 1. Ⓑ
- **Task 3**: A marketing manager tested the benefit of soft drinks for curing hangovers. He took 15 people and got them drunk. The next morning as they awoke, dehydrated and feeling as though they'd licked a camel's sandy feet clean with their tongue, he gave five of them water to drink, five of them Lucozade (a very nice glucose-based UK drink) and the remaining five a leading brand of cola (this variable is called **drink**). He measured how well they felt (on a scale from 0 = I feel like death to 10 = I feel really full of beans and healthy) two hours later (this variable is called **well**). He measured how **drunk** the person got the night before on a scale of 0 = straight edge to 10 = flapping about like a haddock out of water (**hangover.sav**). Fit a model to see whether people felt better after different drinks when adjusting for how drunk they were the night before. Ⓑ
- **Task 4:** Compute effect sizes for Task 3 and report the results. Ⓑ
- **Task 5**: The highlight of the elephant calendar is the annual elephant soccer event in Nepal (google it). A heated argument burns between the African and Asian elephants. In 2010, the president of the Asian Elephant Football Association, an elephant named Boji, claimed that Asian elephants were more talented than their African counterparts. The head of the African Elephant Soccer Association, an elephant called Tunc, issued a press statement that read 'I make it a matter of personal pride never to take seriously any remark made by something that looks like an enormous scrotum'. I was called in to settle things. I collected data from the two types of elephants (**elephant**) over a season and recorded how many goals each elephant scored (**goals**) and how many years of experience the elephant had (**experience**). Analyse the effect of the type of elephant on goal scoring, covarying for the amount of football experience the elephant has (**elephooty.sav**). Ⓑ
- **Task 6**: In Chapter 4 (Task 6) we looked at data from people who had fish or cats as pets and measured their life satisfaction and, also, how much they like animals (**pets.sav)**. Fit a model predicting life satisfaction from the type of pet a person had and their animal liking score (covariate). Ⓑ
- **Task 7:** Compare your results from Task 6 to those for the corresponding task in Chapter 11. What differences do you notice and why? Ⓑ

- **Task 8:** In Chapter 10 we compared the number of mischievous acts in people who had invisibility cloaks to those without (**cloak**). Imagine we replicated that study, but changed the design so that we recorded the number of mischievous acts in these participants before the study began (**mischief_pre**) as well as during the study (**mischief**). Fit a model to see whether people with invisibility cloaks get up to more mischief than those without when factoring in their baseline level of mischief (**invisibility_base.sav**). Ⓑ

View the solutions at www.discoverspss.com





GLM 3: FACTORIAL DESIGNS

14.1 What will this chapter tell me? 652
14.2 Factorial designs 653
14.3 A goggly example 653
14.4 Independent factorial designs and the linear model 655
14.5 Interpreting interaction plots 659
14.6 Simple effects analysis 662
14.7 *F*-statistics in factorial designs 663
14.8 Model assumptions in factorial designs 668
14.9 Factorial designs using SPSS 669
14.10 Output from factorial designs 675
14.11 Robust models of factorial designs 683
14.12 Bayesian models of factorial designs 687
14.13 More effect sizes 689
14.14 Reporting the results of factorial designs 691
14.15 Jane and Brian's story 692
14.16 What next? 693
14.17 Key terms that I've discovered 693
Smart Alex's tasks 693

Letters A to D indicate increasing difficulty

14.1 What will this chapter tell me? Ⓑ

After persuading my two friends (Mark and Mark) to learn the bass and drums, I took the rather odd decision to *stop* playing the guitar. I didn't stop, as such, but I focused on singing instead. In retrospect this was a bad decision because I am *not* a good singer. Mind you, I'm not a good guitarist either. The upshot was that a classmate, Malcolm, ended up as our guitarist. I can't recall how or why we ended up in this configuration, but we called ourselves Andromeda, we learnt several Queen and Iron Maiden songs and we were truly awful. I have some recordings somewhere to prove just what a cacophony of tuneless drivel we produced, but the chances of their appearing on the companion website are slim at best. Suffice it to say, you'd be hard pushed to recognize *which* Iron Maiden and Queen songs we were trying to play. The fact that we were only 14 or 15 at the time cannot even begin to mitigate the depths of ineptitude to which we sank. On the plus side, we garnered a reputation for being too loud in school assembly and we did a successful tour of London's East End (well, our friends' houses). It's common for bands to tire of cover versions and to get lofty ambitions to write their own tunes. I wrote one called 'Escape From Inside' about the film *The Fly*. The only thing we did that resembled the activities of a 'proper' band was to split up due to 'musical differences': Malcolm wanted to write 15-part symphonies about a boy's journey to worship electricity pylons and discover a mythical beast called the cuteasaurus, whereas I wanted to write songs about flies. When we could not agree on a musical direction the split became inevitable. Had I had the power of statistics in my hands back then, rather than split up we could have tested empirically the best musical direction for the band. Suppose Malcolm and I had each written a 15-part symphony and a 3-minute song about a fly. We could have played these songs to some very lucky people and measured their screams. From these data, we could have ascertained the best musical direction to gain popularity. We have two variables that predict screams: whether Malcolm or I wrote the song (songwriter), and whether the song was a 15-part symphony or a song about a fly (song type). This design is called a factorial design, and this chapter looks at how the linear model extends to incorporate multiple categorical predictors.



Figure 14.1 Andromeda coming to a living room near you in 1988 (I'm the one wearing the Anthrax T-shirt)

14.2 Factorial designs Ⓑ



In the previous two chapters we have used the linear model to test for differences between group means when those groups have belonged to a single predictor variable (in experimental designs one independent variable has been manipulated). This chapter extends the linear model to situations in which there are two categorical predictors (independent variables).

Predictor variables (independent variables) often get lonely and want to have friends. Scientists are obliging individuals and often put a second (or third) independent variable into their designs to keep the others company. When an experiment has two or more independent variables it is known as a *factorial design* (because, as we have seen, independent variables are sometimes referred to as *factors*). There are several types of factorial design:

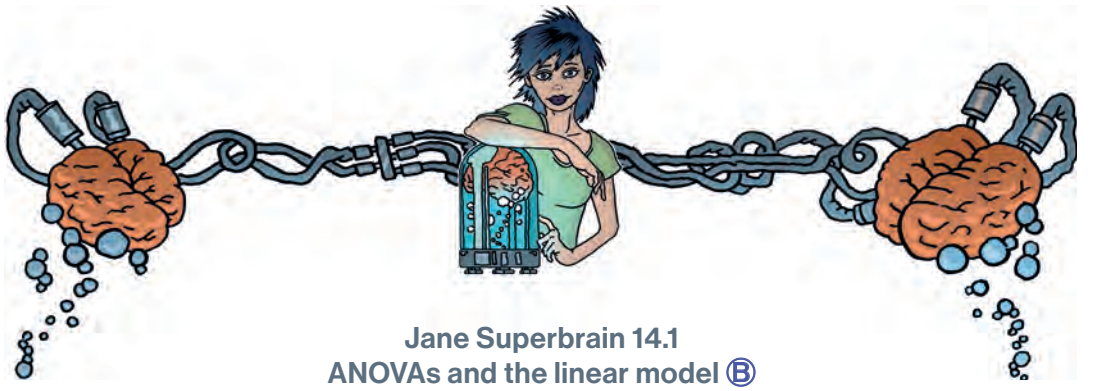
- **Independent factorial design**: There are several independent variables or predictors and each has been measured using different entities (between groups). We discuss this design in this chapter.
- **Repeated-measures (related) factorial design**: Several independent variables or predictors have been measured, but the same entities have been used in all conditions (see Chapter 15).
- **Mixed design**: Several independent variables or predictors have been measured; some have been measured with different entities whereas others used the same entities (see Chapter 16).

As you might imagine, analysing factorial designs can get quite complicated. Fortunately, we can offset this complexity by realizing that we can still fit a linear model to the design, which at least grounds everything in a model with which (I hope) by now you are familiar. Remembering that for weird historical reasons (see Misconception Mutt 12.1) people often label this form of the linear model as 'ANOVA', people often refer to the linear model with two or more categorical predictors that represent experimental independent variables as **factorial ANOVA**. There are also a family of more specific labels that reflect the experimental design that is being analysed (see Jane Superbrain Box 14.1). However, I think our lives would be simpler if we abandoned these labels because they deflect attention from the fact that the underlying model is the same.

14.3 A goggly example Ⓐ

Throughout this chapter we'll use an example of an experimental design with two independent variables (a two-way independent design – Jane Superbrain Box 14.1). The study tested the prediction that subjective perceptions of physical attractiveness become inaccurate after drinking alcohol (the well-known **beer-goggles effect**). The example is based on real research by Chen et al. (2014) who looked at whether the beer-goggles effect was influenced by the attractiveness of the faces being rated. The logic is that alcohol consumption has been shown to reduce accuracy in symmetry judgements, and symmetric faces have been shown to be rated as more attractive. If the beer-goggles effect is driven by alcohol impairing symmetry judgements then you'd expect a stronger effect for unattractive (asymmetric) faces (because alcohol will affect the perception of asymmetry) than attractive (symmetric) ones. The data we'll analyse are fictional, but the results mimic the findings of this research paper.

An anthropologist was interested in the effects of facial attractiveness on the beer-goggles effect. She randomly selected 48 participants. Participants were randomly subdivided into three groups of 16: (1) a placebo group drank 500 ml of alcohol-free beer; (2) a low-dose group drank 500 ml of average strength beer (4% ABV); and (3) a high-dose group drank 500 ml of strong beer (7% ABV).



Jane Superbrain 14.1
ANOVAs and the linear model Ⓑ

Statistical models of experimental designs can appear confusing because there appears to be an endless list of them. It can create the false impression that the models are completely distinct, rather than the truth that they are variations on a common model (the linear model). To add to the confusion, scientists often refer to these models as 'ANOVA' rather than the linear model, because the *F*-statistic that tests the fit of the model (Section 9.2.4) partitions variance, and so is called 'analysis of variance'.

The names people use have two things in common: (1) they involve some quantity of categorical independent (predictor) variables; and (2) they label whether these variables were measured using the same or different entities. If the same entities are tested multiple times the term *repeated measures* is applied, and if different entities take part in the treatment conditions the term *independent* is used. With two or more independent variables, it's possible that some were measured with the same entities and others with different entities; this is referred to as a *mixed* design. In general, people name models that compare means as:

- A [number of independent/predictor variables]-way [how these variables were measured] ANOVA.

Let's put some examples through my magical model translation machine

- One-way independent ANOVA
  - Linear model with *one* categorical predictor in which *different entities* were tested in each category
- Two-way repeated-measures ANOVA
  - Linear model with *two* categorical predictors in which *the same entities* were tested across categories
- Two-way mixed ANOVA
  - Linear model with *two* categorical predictors. For one predictor different entities were tested in the different categories, but for the other *the same entities* were tested across categories
- Three-way independent ANOVA
  - Linear model with *three* categorical predictors. Within each, *different entities* were tested in each category



Within each group, half ($n = 8$) rated the attractiveness of 50 photos of unattractive (asymmetric) faces on a scale from 0 (pass me a paper bag) to 10 (pass me their phone number) and the remaining half rated 50 photos of attractive (symmetric) faces.[1] The outcome for each participant was their median rating across the 50 photos (Table 14.1 and **goggles.sav**).

Table 14.1 Data for the beer-goggles effect

alcohol	placebo		Low dose		high dose	
facetype	attractive	unattractive	attractive	unattractive	attractive	unattractive
	6	2	7	3	5	5
	7	4	6	5	6	6
	6	3	8	7	7	8
	7	3	7	5	5	6
	6	4	6	4	7	7
	5	6	7	4	6	8
	8	5	6	5	5	7
	6	1	5	6	8	6
Total	51	28	52	39	49	53
Mean	6.375	3.500	6.500	4.875	6.125	6.625
Variance	0.839	2.571	0.857	1.554	1.268	1.125
Grand mean = 5.667 Grand variance = 2.525						

14.4 Independent factorial designs and the linear model ©

To simplify things, imagine for now that we have only two levels of the alcohol variable (placebo and high dose). As such, we have two predictor variables, each with two levels. We've seen many times that the general linear model takes the following general form:

$$Y_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i} + \ldots + \hat{b}_n X_{ni} + e_i \quad (14.1)$$

When we first looked at using the linear model to compare means we used an example of the effect of three doses of puppy therapy (none, 15 and 30 minutes) on happiness. We saw that the linear model became:

$$\text{happiness}_i = \hat{b}_0 + \hat{b}_1 \text{long}_i + \hat{b}_2 \text{short}_i + e_i \quad (14.2)$$

in which the **long** and **short** predictor variables were dummy variables that coded to which of the groups a participant belongs using values of 0 and 1. In our current example, we have two variables that represent categories: **facetype** (unattractive or attractive) and **alcohol** (placebo and high dose). Just like we did for puppy therapy, we can code participant's category membership on these variables with zeros and ones; for example, we might code type of face as unattractive = 0, attractive = 1, and

1 These photographs were from a larger pool of 500 that had been pre-rated for facial symmetry by a different sample. The 50 photos with the highest and lowest ratings were used.

alcohol group as 0 = placebo, 1 = high dose. We can copy the puppy therapy model (Eq. 14.2) but replacing the predictors with our two independent variables:

$$\text{attractiveness}_i = \hat{b}_0 + \hat{b}_1\text{facetype}_i + \hat{b}_2\text{alcohol}_i + e_i \tag{14.3}$$

However, this model does not consider the interaction between type of face and alcohol. To include this term, we extend the model to become:

$$\begin{aligned}\text{attractiveness}_i &= \hat{b}_0 + \hat{b}_1 A_i + \hat{b}_2 B_i + \hat{b}_3 AB_i + e_i \\ &= \hat{b}_0 + \hat{b}_1\text{facetype}_i + \hat{b}_2\text{alcohol}_i + \hat{b}_3\text{interaction}_i + e_i\end{aligned} \tag{14.4}$$

You might wonder how we code the interaction term, but we saw how to do this in Section 11.3. The interaction term represents the combined effect of **alcohol** and **facetype** and it is obtained by multiplying the variables involved. This multiplication is the reason why interaction terms are written as **facetype** × **alcohol**. Table 14.2 shows the resulting predictor variables for the model (the group means for the combinations of type of face and alcohol are included because they'll come in useful in due course). Note that the interaction variable is the type of face dummy variable multiplied by the alcohol dummy variable. For example, someone receiving a high dose of alcohol and rating unattractive faces would have a value of 0 for the type of face variable, 1 for the alcohol variable and 0 for the interaction variable.

As with other linear models we have encountered, we can predict values of attractiveness from values of the predictors and the parameter estimates ($\hat{b}$) using Eq. 14.5 (see Section 10.4):

$$\widehat{\text{attractiveness}}_i = \hat{b}_0 + \hat{b}_1\text{facetype}_i + \hat{b}_2\text{alcohol}_i + \hat{b}_3\text{interaction}_i \tag{14.5}$$

Table 14.2 Coding scheme for predictors in a factorial design

Type of faces rated	alcohol	Dummy (facetype)	Dummy (alcohol)	Interaction	Mean
unattractive	placebo	0	0	0	3.500
unattractive	high dose	0	1	0	6.625
attractive	placebo	1	0	0	6.375
attractive	high dose	1	1	1	6.125

To see what the $\hat{b}$-values in Eq. 14.5 represent we can insert values of our predictors and see what happens. Let's start with participants in the placebo group rating unattractive faces. In this case, the values of **facetype**, **alcohol** and the interaction are all 0. The predicted value of the outcome, as we have seen in Chapter 12, will be the group mean (3.500). Our model becomes:

$$\begin{aligned}\widehat{\text{attractiveness}}_i &= \hat{b}_0 + \hat{b}_1\text{facetype}_i + \hat{b}_2\text{alcohol}_i + \hat{b}_3\text{interaction}_i \\ \bar{X}_{\text{unattractive, placebo}} &= \hat{b}_0 + \left(\hat{b}_1 \times 0\right) + \left(\hat{b}_2 \times 0\right) + \left(\hat{b}_3 \times 0\right) \\ \hat{b}_0 &= \bar{X}_{\text{unattractive, placebo}} \\ \hat{b}_0 &= 3.500\end{aligned} \tag{14.6}$$

and you can see that the constant $\hat{b}_0$ represents the mean of the group for which all variables are coded as 0. As such it's the mean value of the base category (in this case participants in the placebo group rating unattractive faces).

Now, let's see what happens when we look at participants in the placebo group rating attractive faces. The outcome is the mean rating of attractive faces after a placebo drink, the type of face variable is 1 and the alcohol and interaction variables are 0. Remember that $\hat{b}_0$ is the mean rating of unattractive faces after a placebo drink. The model becomes:

$$\begin{aligned}\bar{X}_{\text{attractive, placebo}} &= \hat{b}_0 + (\hat{b}_1 \times 1) + (\hat{b}_2 \times 0) + (\hat{b}_3 \times 0)\\ &= \hat{b}_0 + \hat{b}_1\\ &= \bar{X}_{\text{unattractive, placebo}} + \hat{b}_1\\ \hat{b}_1 &= \bar{X}_{\text{attractive, placebo}} - \bar{X}_{\text{unattractive, placebo}}\\ &= 6.375 - 3.500\\ &= 2.875\end{aligned} \tag{14.7}$$

which shows that $\hat{b}_1$ represents the difference between ratings of unattractive and attractive faces in people who drank a placebo. More generally we can say it's the effect of type of face for the base category of alcohol (the category coded with 0, in this case the placebo).

Let's look at people who had a high dose of alcohol and rated unattractive faces. The outcome is the mean rating of unattractive faces after a high dose of alcohol, the type of face variable is 0, the alcohol variable is 1 and the interaction variable is 0. We can replace $\hat{b}_0$ with the mean rating of unattractive faces after a placebo drink. The model becomes:

$$\begin{aligned}\bar{X}_{\text{unattractive, high dose}} &= \hat{b}_0 + \left(\hat{b}_1 \times 0\right) + \left(\hat{b}_2 \times 1\right) + \left(\hat{b}_3 \times 0\right)\\ &= \hat{b}_0 + \hat{b}_2\\ &= \bar{X}_{\text{unattractive, placebo}} + \hat{b}_2\\ \hat{b}_2 &= \bar{X}_{\text{unattractive, high dose}} - \bar{X}_{\text{unattractive, placebo}}\\ &= 6.625 - 3.500\\ &= 3.125\end{aligned} \tag{14.8}$$

which shows that $\hat{b}_2$ in the model represents the difference between ratings of unattractive faces after a high dose of alcohol compared to a placebo drink. Put more generally, it's the effect of alcohol in the base category of **facetype** (i.e., the category coded with a 0, in this case unattractive).

Finally, we can look at ratings of attractive faces after a high dose of alcohol. The predicted outcome is the mean rating of attractive faces by people after a high dose of alcohol, and the type of face, alcohol and interaction variables are all 1. We can replace $\hat{b}_0$, $\hat{b}_1$ and $\hat{b}_2$ with what we now know they represent from Eq. 14.6 to Eq. 14.8. The model becomes:

$$\begin{aligned}\bar{X}_{\text{attractive, high dose}} &= \hat{b}_0 + \left(\hat{b}_1 \times 1\right) + \left(\hat{b}_2 \times 1\right) + \left(\hat{b}_3 \times 1\right)\\ &= \hat{b}_0 + \hat{b}_1 + \hat{b}_2 + \hat{b}_3\\ &= \bar{X}_{\text{unattractive, placebo}} + \left(\bar{X}_{\text{attractive, placebo}} - \bar{X}_{\text{unattractive, placebo}}\right)\\ &\quad + \left(\bar{X}_{\text{unattractive, high dose}} - \bar{X}_{\text{unattractive, placebo}}\right) + \hat{b}_3\\ &= \bar{X}_{\text{attractive, placebo}} + \bar{X}_{\text{unattractive, high dose}} - \bar{X}_{\text{unattractive, placebo}} + \hat{b}_3\\ \hat{b}_3 &= \bar{X}_{\text{unattractive, placebo}} - \bar{X}_{\text{attractive, placebo}} + \bar{X}_{\text{attractive, high dose}}\\ &\quad - \bar{X}_{\text{unattractive, high dose}}\\ &= 3.500 - 6.375 + 6.125 - 6.625\\ &= -3.375\end{aligned} \tag{14.9}$$

which is scary, but we'll break it down. It shows that $\hat{b}_3$ compares the difference between ratings of unattractive and attractive faces in the placebo group to the same difference in the high-dose group.

Put more generally, it compares the effect of type of face after a placebo drink to the effect of type of face after a high dose of alcohol.[2]

This explanation makes more sense if you think in terms of an interaction plot. Figure 14.2 (top left) shows the interaction plot for these data. The difference between ratings of unattractive and attractive faces in the two placebo groups is the distance between the lines on the plot for the placebo group (the difference between group means, which is 2.875). If we look at the same difference for the high-dose group, we find that the difference between ratings of unattractive and attractive faces is –0.500. If we plotted these two 'difference' values as a new plot we'd get a line connecting 2.875 to –0.500 (see Figure 14.2, bottom left). The slope of this line reflects the difference between the effect of type of face after a placebo compared to after a high dose of alcohol. We know that $\hat{b}$-values represent gradients of lines and $\hat{b}_3$ is the gradient of the line that connects the difference scores (this is (–0.500) – 2.875 = –3.375).

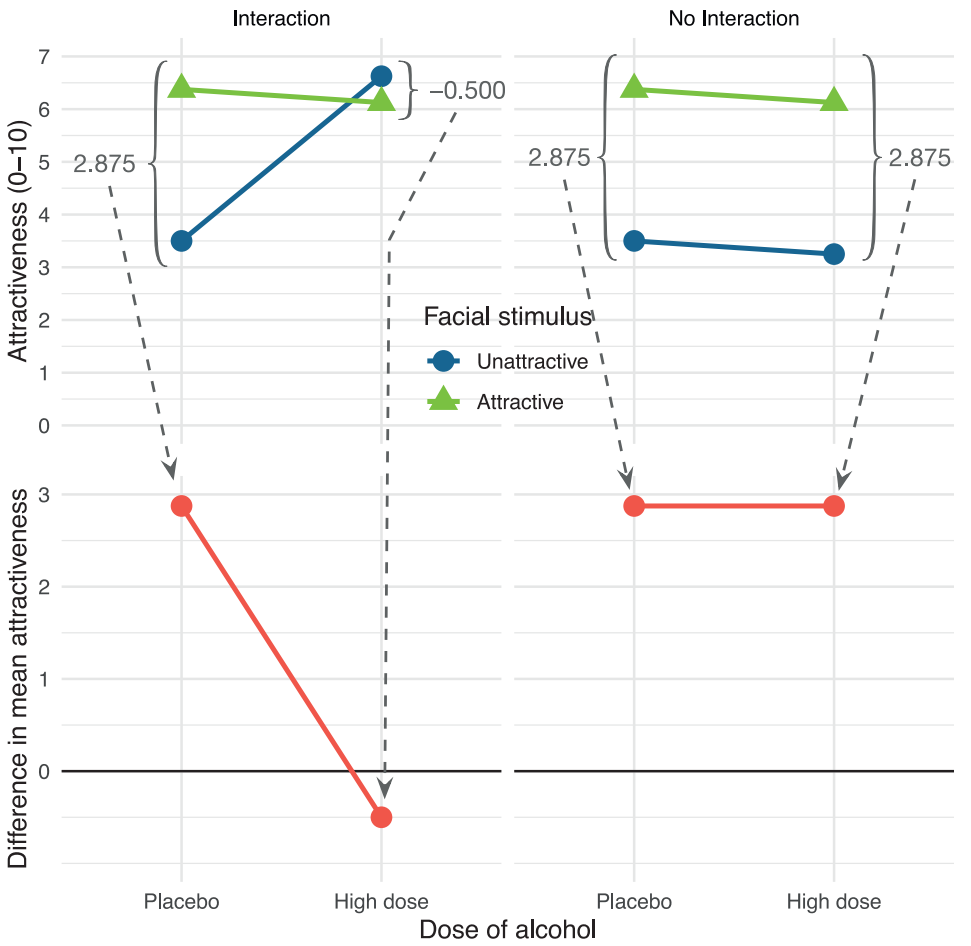
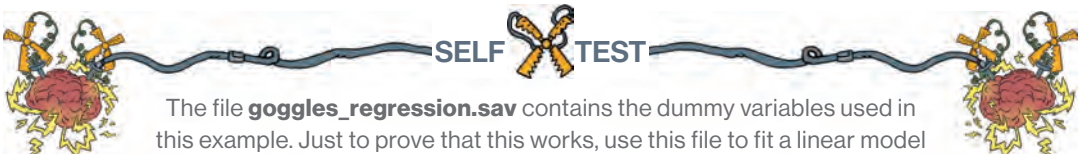


Figure 14.2 Breaking down what an interaction represents

2 If you rearrange the terms in the equation you'll see that you can also phrase the interaction the opposite way around: it represents the effect of alcohol on ratings of attractiveness for attractive faces compared to unattractive ones.

The right-hand side of Figure 14.2 illustrates what happens if there isn't an interaction effect by showing the same data as on the left, except that the mean rating of unattractive pictures after a high dose of alcohol has been changed to 3.25. If we calculate the difference between ratings of unattractive and attractive faces after a placebo drink we get the same as before: 2.875. However, if we calculate the difference between ratings of unattractive and attractive faces after a high dose of alcohol we now get 2.875. If we again plot these differences on a new plot, we find a completely flat line. So, when there's no interaction, the line connecting the effect of type of face after a placebo drink and after a high dose is flat and the resulting $\hat{b}_3$ would be 0 (remember that a zero gradient means a flat line). If we calculate this difference, this is what we get: 2.875 – 2.875 = 0.



The file **goggles_regression.sav** contains the dummy variables used in this example. Just to prove that this works, use this file to fit a linear model predicting attractiveness ratings from **facetype, alcohol** and the **interaction** variable.

Output 14.1 shows the resulting table of coefficients from the self-test. Note that the $\hat{b}$-value for the type of face matches Eq. 14.7, the $\hat{b}$ for the alcohol group matches Eq. 14.8 and the $\hat{b}$ for the interaction matches Eq. 14.9. All of which I hope convinces you that we can use a linear model to analyse designs that incorporate multiple categorical predictors.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.500	.426		8.219	<.001
	Attractivenss of facial stimuli	2.875	.602	.851	4.774	<.001
	Alcohol consumption	3.125	.602	.925	5.189	<.001
	Interaction	-3.375	.852	-.866	-3.963	<.001

a. Dependent Variable: Median attractiveness rating

Output 14.1

14.5 Interpreting interaction plots Ⓑ

Having seen what an interaction term represents mathematically, this section looks at how to visualize them. Figure 14.2 has already shown us that if we plot lines between the means of all of our conditions then an interaction usually results in non-parallel lines (Figure 14.2, left), whereas parallel lines typically suggest no interaction (Figure 14.2, right), or more accurately an interaction effect that is too small to be interesting. Let's have a look at other examples of interaction graphs to practise interpreting them. Imagine we got the profile of results in Figure 14.3 (left). Do you think the interaction effect would still have been significant?

This profile of data probably would also give rise to a significant interaction term because the ratings of attractive and unattractive faces are different in the placebo and high-dose group, but relatively similar in the low-dose group. Visually, the blue line (and circles) shows a different pattern than the green line (and triangles). This situation reflects a world in which the beer-goggles effect kicks in at low doses of alcohol, but after a higher dose reality settles back in again. Theoretically more difficult

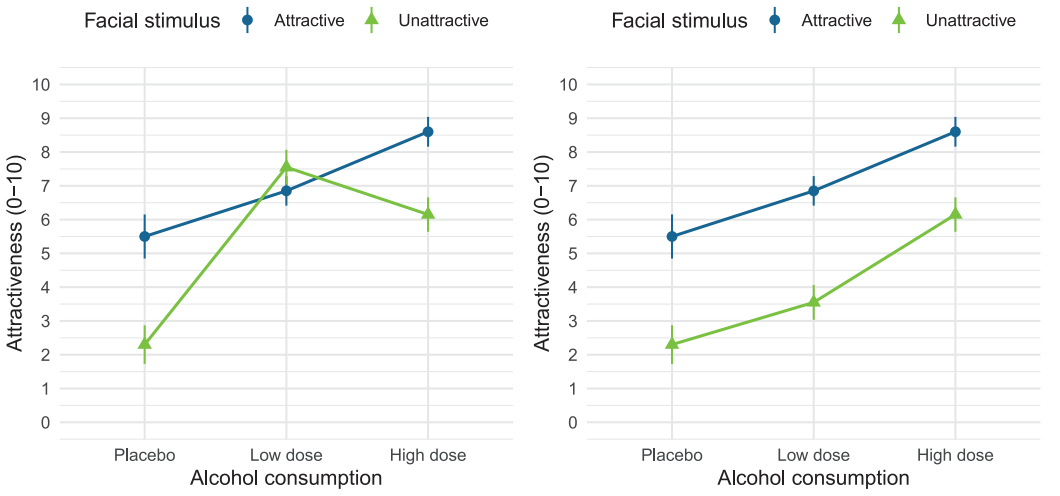


Figure 14.3 Interaction plots

to explain, but nevertheless these results reflect an interaction because the difference between ratings of attractive and unattractive faces varies depending on how much alcohol has been drunk.

Let's try another example. Is there a significant interaction in Figure 14.3 (right)?

The data in Figure 14.3 (right) are unlikely to reflect a significant interaction because the effect of alcohol is the same for unattractive and attractive photos: ratings of attractiveness increase as the dose of alcohol increases, and this is true for both attractive and unattractive faces. Visually, the blue line (and circles) shows the same pattern as the green line (and triangles). Two general points that we can make from these examples are as follows:

1 Non-parallel lines on an interaction plot indicate some degree of interaction, but how strong and whether the interaction is significant depends on how non-parallel the lines are.
2 Lines on an interaction graph that cross are very non-parallel, which hints at a possible significant interaction. However, crossing lines don't *always* reflect a significant interaction.

Sometimes people use bar charts to show interactions, so let's look at those in Figure 14.4. Panels (a) and (b) display the data from the example used in this chapter (in fact, why not have a go at plotting them for yourself?). The data are presented in two different ways: panel (a) shows the data when levels of alcohol are placed along the *x*-axis and different-coloured bars show the means for unattractive and attractive faces; and panel (b) shows the opposite scenario where type of face is plotted on the *x*-axis and different colours distinguish the dose of alcohol. Both plots show a possible interaction effect. You're looking for the differences in heights of coloured bars to vary at different points along the *x*-axis. For example, in panel (a) you'd look at the difference between the blue and green bars for placebo, and then look at, say, the high dose and ask, 'Is the difference in heights between the bars the same as for the placebo?' In this case the difference in height between the blue and green bars for the placebo is bigger than the same difference for the high-dose group, hence a possible interaction. Panel (b) shows the same data, but plotted the other way around. Again look at the pattern of responses. First look at the unattractive faces and see that the pattern is that attractiveness ratings increase as the dose of alcohol increases (the bars increase in height). Now look at the attractive faces. Is the pattern the same? No, the bars are all the same height, showing that ratings don't change as a function of alcohol. The interaction effect is shown up by the fact that for the attractive faces the bars follow a different pattern than the unattractive faces.

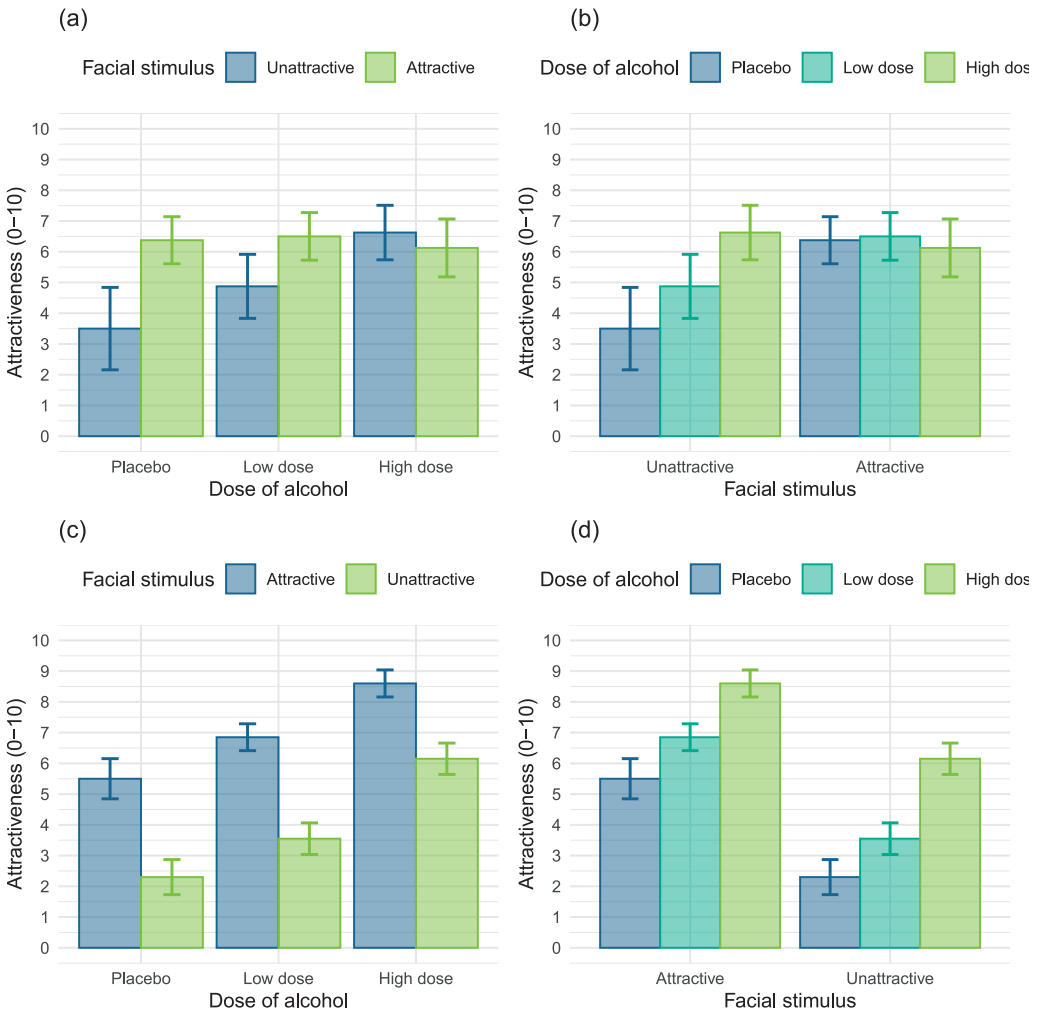


Figure 14.4 Bar charts showing interactions between two variables



Panels (c) and (d) display the different data from a study with the same factorial design. Again the panels show the same data but switch the variable represented by the colours. In the placebo group in panel (c) the blue bar is a lot taller than the green; moving on to the low-dose group, the blue bar is about as much taller than the green bar as was found in the placebo group; and finally, for the high-dose group the blue bar perhaps slightly less taller than the green bar compared to the other groups, but not by a lot. In short, in all conditions the same pattern is shown – the blue bar is a similar amount taller than the green one (i.e., the amount by which attractive faces are rated more attractive than unattractive faces is similar in all three alcohol groups) – therefore, there is unlikely to be a significant interaction. Panel (d) shows the same data and, therefore, pattern. For unattractive faces the pattern is that attractiveness ratings increase as more alcohol is drunk (the bars increase in height), and for

the attractive faces the same pattern emerges, with ratings increasing as more alcohol is consumed. This pattern again suggests no significant interaction: the change in attractiveness ratings due to alcohol is similar for unattractive and attractive faces.

14.6 Simple effects analysis ©

In factorial designs the interaction terms are typically the effect of most interest because they represent the degree to which one predictor moderates the effect of another on the outcome variable. We looked generally at moderation in Chapter 11, and saw that we can break down interactions using something called simple slopes analysis. Factorial designs test a specific case of moderation where all of the variables involved are categorical. In these designs we can't use simple slopes analysis (because none of the predictors are continuous), but we can use a similar tool known as a **simple effects analysis**, which looks at the effect of one independent variable at individual levels of the other independent variable.

For example, we could do a simple effects analysis looking at the effect of type of face at each level of alcohol. This would mean taking the average attractiveness rating of unattractive faces and comparing it to that for attractive faces after a placebo drink, then making the same comparison after a low dose of alcohol, and then finally for a high dose. To make this idea more concrete, imagine we have collected the data in the top left of Figure 14.2, and have a significant interaction between alcohol (in this example we have two levels: placebo and high dose) and type of face stimulus (with two levels: attractive or unattractive). These data are reproduced in Figure 14.5. One way to do a simple effects analysis on this interaction is to look at the effect of facial stimulus at each level of alcohol (Figure 14.5, left). The analysis produces two *F*-statistics. The first tests whether the difference in attractiveness ratings between attractive and unattractive faces is significant for those in the placebo group. In other words, is the distance between the circle and triangle in the placebo group (i.e., 2.875) significantly different from zero? The second tests whether the difference in attractiveness ratings between attractive and unattractive faces is significant for those in the high-dose group. In other words, is the distance between the circle and triangle in the high-dose group (i.e., –0.5) significantly different from zero?

An alternative simple effects analysis is to look at the effect of alcohol at each level of facial stimulus (Figure 14.5, right). Again we'd get two *F*-statistics, but this time the first tests whether those that had alcohol rated *attractive* faces differently than those in the placebo group. In other words, is the vertical distance between the two triangles (i.e., 0.25) different from zero? The second tests whether those that had alcohol rated *unattractive* faces differently to those in the placebo group. In other words, is the vertical distance between the two circles (i.e., –3.125) significantly different from zero? We'll see how to obtain a simple effects analysis in Section 14.9.6.

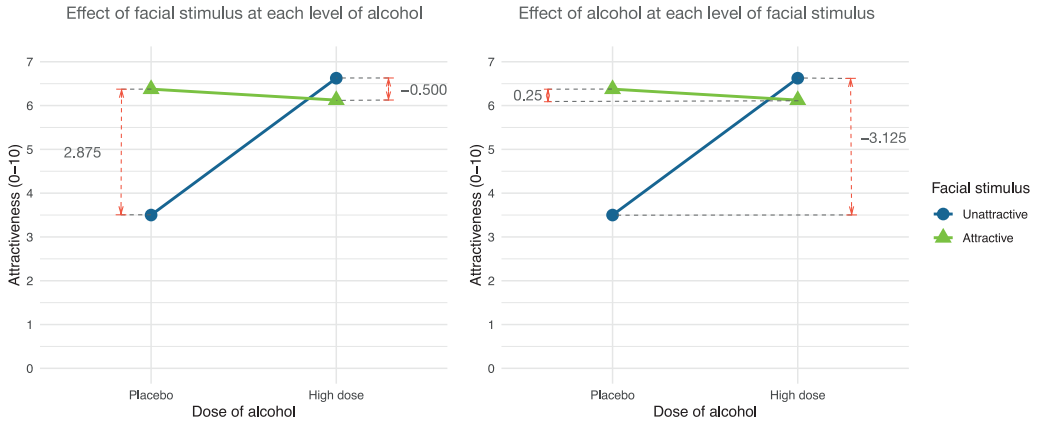


Figure 14.5 Breaking down an interaction using simple effects analysis



Oliver Twisted
Please, Sir, can I have some more ... simple effects?

'I want to impress my friends by doing a simple effects analysis by hand,' boasts Oliver. You don't really need to know how simple effects analyses are calculated to run them, Oliver, but seeing as you asked, it is explained at www.discoverspss.com.

14.7 *F*-statistics in factorial designs Ⓑ

Now that we have a good conceptual understanding of factorial designs as an extension of the linear model, we will turn our attention to the specific calculations that go on behind the scenes. The reason for doing this is that it should help you to understand what the output of the analysis means. Calculating the *F*-statistic with two categorical predictors is very similar to when we had only one: we still find the total sum of squared errors (SS_T) and break this variance down into variance that can be explained by the model/experiment (SS_M) and variance that cannot be explained (SS_R). The main difference is that with factorial designs, the variance explained by the model/experiment is made up of not one predictor (experimental manipulation) but two. Therefore, the model sum of squares gets further subdivided into variance explained by the first predictor/independent variable (SS_A), variance explained by the second predictor/independent variable (SS_B) and variance explained by the interaction of these two predictors ($SS_{A \times B}$) – see Figure 14.6.

14.7.1 Total sum of squares (SS_T) Ⓑ

We start off in the same way as we did for a one-way ANOVA. That is, we calculate how much variability there is between scores when we ignore the experimental condition from which they came. Remember from one-way ANOVA (Eq. 12.11) that SS_T is calculated using the following equation:

$$\begin{aligned} SS_T &= \sum_{i=1}^{N} \left(x_i - \bar{x}_{grand}\right)^2 \\ &= s^2_{grand}\left(N-1\right) \end{aligned} \qquad (14.10)$$

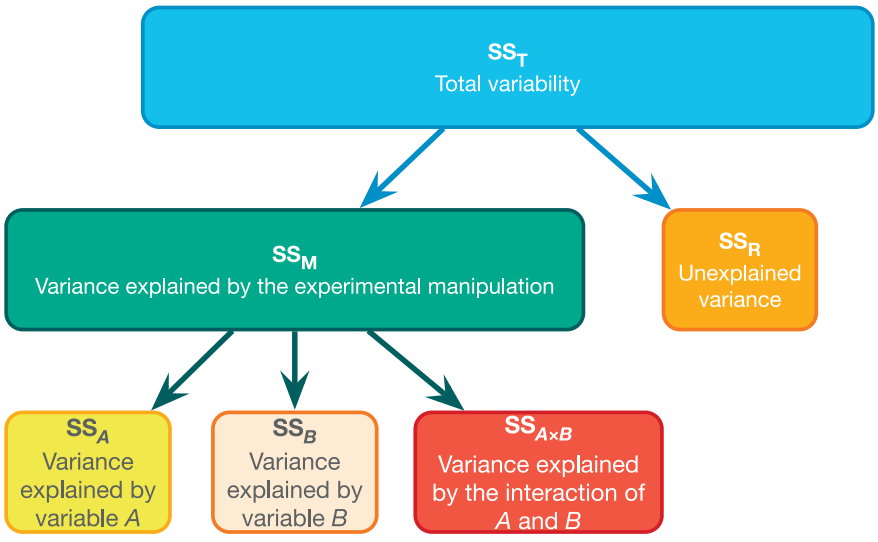


Figure 14.6 Breaking down the variance in a two-way factorial design

The grand variance is the variance of all scores when we ignore the group to which they belong. So we treat the data as one big group, calculate the variance of all scores, and find it's 2.525 as in Table 14.1 (try this on your calculator). We used 48 scores to generate this value, and so N is 48. As such the total sum of squares is 118.675:

$$\begin{aligned} SS_T &= s^2_{grand}(N-1) \\ &= 2.525(48-1) \\ &= 118.675 \end{aligned} \tag{14.11}$$

The degrees of freedom for this SS are N – 1, or 47.

14.7.2 The model sum of squares, SS_M Ⓑ

The next step is to work out the model sum of squares, which is then further broken down into the variance attributable to the first independent variable (SS_A), the variance attributable to the second independent variable (SS_B) and the variance attributable to the interaction of these two variables ($SS_{A\times B}$). The model sum of squares (as we've seen a few times) is the difference between what the model predicts and the overall mean of the outcome variable. We've also seen that with predictors that represent group membership 'what the model predicts' is the group means. Therefore, we work out the model sum of squares by looking at the difference between each group mean and the overall mean (see Section 12.3.5). If we combine all levels of the two independent variables (three alcohol groups that rated unattractive faces and another three that rated attractive faces) we have six groups. We can apply the equation for the model sum of squares that we've used before (Eq. 12.13):

$$SS_M = \sum_{g=1}^{k} n_g\left(\bar{x}_g - \bar{x}_{grand}\right)^2 \tag{14.12}$$

The grand mean is the mean of all scores and is 5.667 (Table 14.1) and n is the number of scores in each group (8 in this case). Therefore, the model sum of squares is 61.17:

$$\begin{aligned} SS_M &= 8(6.375 - 5.667)^2 + 8(3.500 - 5.667)^2 + 8(6.500 - 5.667)^2 \\ &\quad + 8(4.875 - 5.667)^2 + 8(6.125 - 5.667)^2 + 8(6.625 - 5.667)^2 \\ &= 8(0.708)^2 + 8(-2.167)^2 + 8(0.833)^2 + 8(-0.792)^2 + 8(0.458)^2 + 8(0.958)^2 \\ &= 4.010 + 37.567 + 5.551 + 5.018 + 1.678 + 7.342 \\ &= 61.17 \end{aligned} \quad (14.13)$$

The degrees of freedom are the number of groups, k, minus 1; we had six groups and so $df = 5$. At this stage we know that the model (our experimental manipulations) can explain 61.17 units of variance out of the total of 118.675 units. The next stage is to break down this model sum of squares to separate the variance in attractiveness ratings explained by each of our independent variables.

14.7.3 The main effect of type of face, SS_A Ⓑ

To work out the variance accounted for by the first predictor/independent variable (in this case, type of face) we group attractiveness ratings according to which type of face was being rated. So, basically we ignore the dose of alcohol that was drunk, and we place all the ratings of unattractive faces into one group and all of the ratings of attractive ones into another. The data will look like Figure 14.7 (note that the first box contains the three attractive columns from Table 14.1 and the second box contains the unattractive columns).

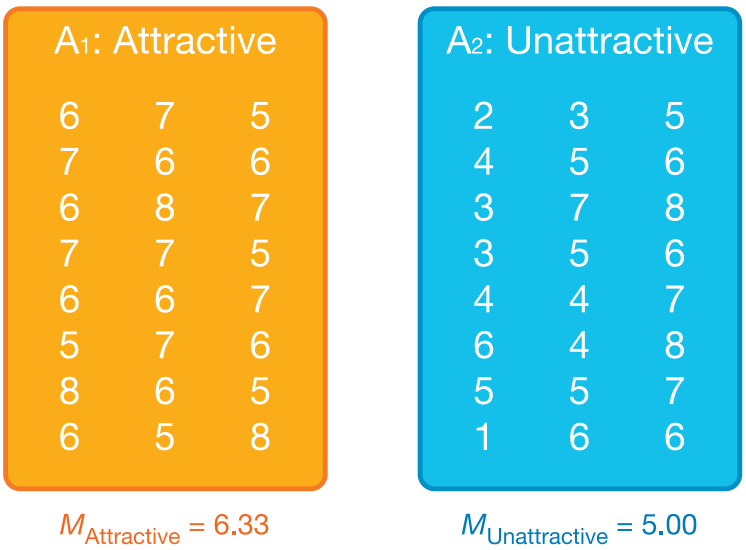


Figure 14.7 The main effect of type of face

We then apply the same equation for the model sum of squares that we just used (compare with Eq. 14.12):

$$SS_A = \sum_{g=1}^{k} n_g \left(\bar{x}_g - \bar{x}_{grand}\right)^2 \quad (14.14)$$

The grand mean is the mean of all scores (5.667, as above) and n is the number of scores in each group (i.e., the number of participants who rated unattractive and attractive faces; 24 in both cases). The means of the two groups have been calculated in Figure 14.7. The resulting model sum of squares for the main effect of the type of face is 21.32:

$$\begin{aligned} SS_{\text{Type of face}} &= 24(6.33 - 5.667)^2 + 24(5.00 - 5.667)^2 \\ &= 24(0.666)^2 + 24(-0.667)^2 \\ &= 10.645 + 10.677 \\ &= 21.32 \end{aligned} \tag{14.15}$$

The degrees of freedom for this SS are the number of groups used, k, minus 1. We used two groups (unattractive and attractive) and so $df = 1$. To sum up, the main effect of type of face compares the mean of all ratings of unattractive faces to the corresponding mean for attractive faces (regardless of how much alcohol was consumed).

14.7.4 The main effect of alcohol, SS_B Ⓑ

To work out the variance accounted for by the second independent variable (dose of alcohol) we group the attractiveness ratings according to the dose of alcohol. In other words, we ignore the type of face the participant was rating, and we place all of the scores after the placebo drink in one group, the ratings after a low dose in another group and the ratings after a high dose in a third group. The data will look like Figure 14.8. We apply the same equation as for the main effect of type of face:

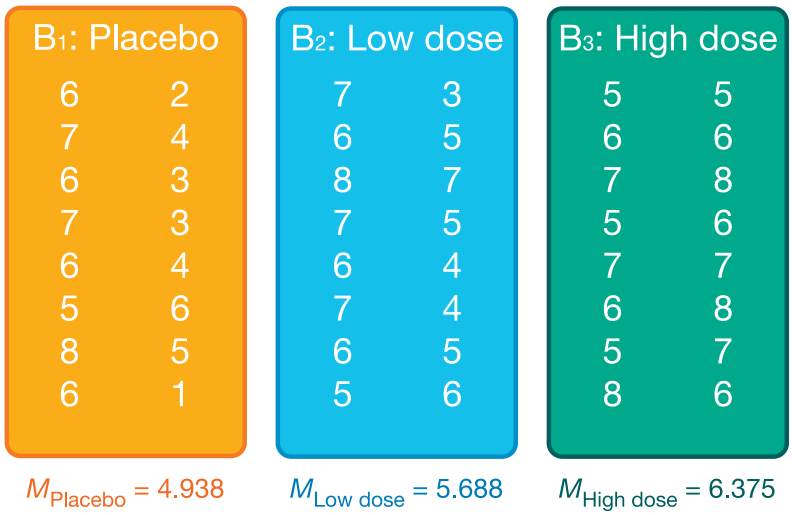


Figure 14.8 The main effect of alcohol

$$SS_B = \sum_{g=1}^{k} n_g \left(\bar{x}_g - \bar{x}_{\text{grand}}\right)^2 \tag{14.16}$$

The grand mean is the mean of all scores (5.667 as before), n is the number of scores in each group (i.e., the number of scores in each alcohol condition, in this case 16) and the group means are given in Figure 14.8. The resulting sum of squares is 16.53:

$$\begin{aligned} SS_{\text{alcohol}} &= 16(4.938 - 5.667)^2 + 16(5.688 - 5.667)^2 + 16(6.375 - 5.667)^2 \\ &= 16(-0.729)^2 + 16(0.021)^2 + 16(0.708)^2 \\ &= 8.503 + 0.007 + 8.020 \\ &= 16.53 \end{aligned} \tag{14.17}$$

The degrees of freedom are the number of groups used minus 1 (see Section 12.3.5). We used three groups and so $df = 2$. To sum up, the main effect of alcohol compares the means of the placebo, low dose and high dose (regardless of whether the ratings were of unattractive or attractive faces).

14.7.5 The interaction effect, $SS_{A\times B}$ Ⓑ

The final stage is to calculate how much variance is explained by the interaction of the two variables. The simplest way to do this is to remember that the SS_M is made up of three components (SS_A, SS_B and $SS_{A\times B}$). Therefore, given that we know SS_A and SS_B, we can calculate the interaction term by subtraction:

$$SS_{A\times B} = SS_M - SS_A - SS_B \tag{14.18}$$

For these data, the value is 23.32:

$$\begin{aligned} SS_{A\times B} &= SS_M - SS_A - SS_B \\ &= 61.17 - 21.32 - 16.53 \\ &= 23.32 \end{aligned} \tag{14.19}$$

The degrees of freedom can be calculated in the same way, but are also the product of the degrees of freedom for the main effects – either method works:

$$\begin{aligned} df_{A\times B} &= df_M - df_A - df_B \\ &= 5 - 1 - 2 \\ &= 2 \\ df_{A\times B} &= df_A \times df_B \\ &= 1 \times 2 \\ &= 2 \end{aligned} \tag{14.20}$$

14.7.6 The residual sum of squares, SS_R Ⓑ

The residual sum of squares is calculated in the same way as in Section 12.3.6. As ever, it represents errors in prediction from the model, but in experimental designs this also reflects individual differences in performance or variance that can't be explained by factors that were systematically manipulated. The value is calculated by taking the squared error between each data point and its corresponding group mean. An alternative way to express this is (see Eq. 12.16):

$$\begin{aligned} SS_R &= \sum_{g=1}^{k} s_g^2 (n_g - 1) \\ &= s^2_{\text{group 1}}(n_1 - 1) + s^2_{\text{group 2}}(n_2 - 1) + \cdots + s^2_{\text{group n}}(n_n - 1) \end{aligned} \tag{14.21}$$

We use the individual variances of each group from Table 14.1 and multiply them by one less than the number of people within the group (n), in this case $n = 8$. The resulting residual sum of squares is 57.50:

$$\begin{aligned} SS_R &= s^2_{\text{group 1}}(n_1 - 1) + s^2_{\text{group 2}}(n_2 - 1) + \cdots + s^2_{\text{group 6}}(n_6 - 1) \\ &= 0.839(8-1) + 2.571(8-1) + 0.857(8-1) + 1.554(8-1) + 1.268(8-1) \\ &\quad + 1.125(8-1) \\ &= (0.839\times 7) + (2.571\times 7) + (0.857\times 7) + (1.554\times 7) + (1.268\times 7) \\ &\quad + (1.125\times 7) \\ &= 5.873 + 17.997 + 5.999 + 10.878 + 8.876 + 7.875 \\ &= 57.50 \end{aligned} \tag{14.22}$$

The degrees of freedom for each group will be one less than the number of scores per group (i.e., 7). We add the degrees of freedom for each group to get a total sum of squares of 6 × 7 = 42.

14.7.7 The *F*-statistics Ⓑ

Each effect in a factorial design has its own *F*-statistic. In a two-way design this means we compute *F* for the two main effects and the interaction. To calculate these *F*s we first calculate the mean squares for each effect by taking the sum of squares and dividing by the respective degrees of freedom (think back to Section 12.3.7). We also need the mean squares for the residual term. So, for this example we'd have four mean squares:

$$\begin{aligned} MS_A &= \frac{SS_A}{df_A} = \frac{21.32}{1} = 21.32 \\ MS_B &= \frac{SS_B}{df_B} = \frac{16.53}{2} = 8.27 \\ MS_{A\times B} &= \frac{SS_{A\times B}}{df_{A\times B}} = \frac{23.32}{2} = 11.66 \\ MS_R &= \frac{SS_R}{df_R} = \frac{57.50}{42} = 1.37 \end{aligned} \tag{14.23}$$

The *F*-statistic for each effect is then calculated, as we've seen before, by dividing its mean squares by the residual mean squares:

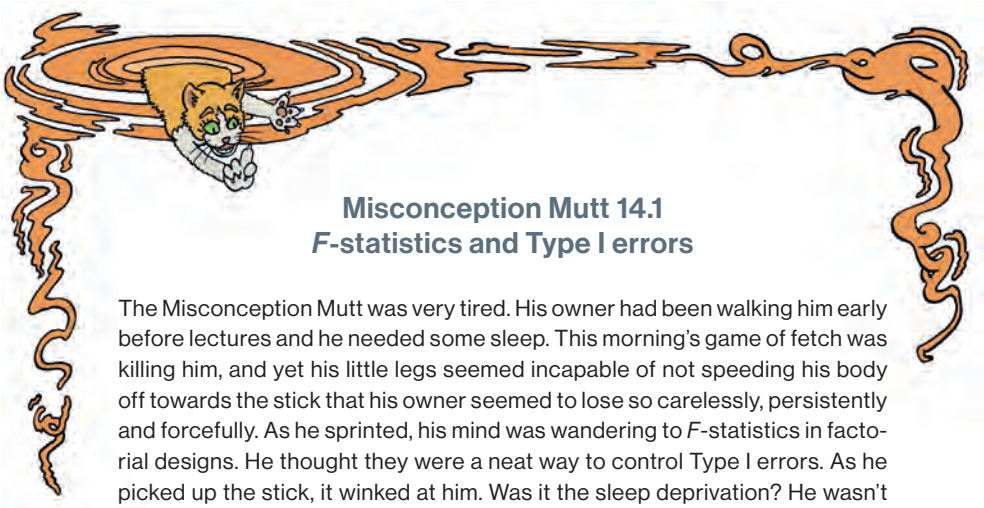
$$\begin{aligned} F_A &= \frac{MS_A}{MS_R} = \frac{21.32}{1.37} = 15.56 \\ F_B &= \frac{MS_B}{MS_R} = \frac{8.27}{1.37} = 6.04 \\ F_{A\times B} &= \frac{MS_{A\times B}}{MS_R} = \frac{11.66}{1.37} = 8.51 \end{aligned} \tag{14.24}$$

IBM SPSS Statistics software will compute an exact *p*-value for each of these *F*-statistics to tell us the probability (in the long run) of getting an *F* at least as big as the one we have if there were no effect in the population. One important issue here is that these *F*s do not control the familywise Type I error (Misconception Mutt 14.1).

To sum up, when you have a factorial design, the computations behind the linear model are basically the same as when you have only one categorical predictor, except that the model sum of squares is partitioned into three parts: the effect of each of the independent variables and the effect of how these variables interact.

14.8 Model assumptions in factorial designs Ⓒ

When using the linear model to analyse a factorial design the sources of potential bias (and counteractive measures) discussed in Chapter 6 apply. If you have violated the assumption of homogeneity of variance there are corrections based on the Welch procedure that was described a couple of chapters back. However, this is quite technical, SPSS doesn't do it, and if you have anything more complicated than a 2 × 2 design then, really, it would be less painful to cover your body in paper cuts and then bathe in chilli sauce (see Algina & Olejnik, 1984). One practical solution is to bootstrap the *post hoc* tests so that these will be robust. You can also ask for confidence intervals and *p*-values for parameter estimates that are robust to heteroscedasticity (refer back to Section 13.6.6). This won't help for the *F*- statistics. There is a robust test based on trimmed means that can be done using R within SPSS (see Section 14.11).



Misconception Mutt 14.1
F-statistics and Type I errors

The Misconception Mutt was very tired. His owner had been walking him early before lectures and he needed some sleep. This morning's game of fetch was killing him, and yet his little legs seemed incapable of not speeding his body off towards the stick that his owner seemed to lose so carelessly, persistently and forcefully. As he sprinted, his mind was wandering to *F*-statistics in factorial designs. He thought they were a neat way to control Type I errors. As he picked up the stick, it winked at him. Was it the sleep deprivation? He wasn't sure but then its bark turned a ginger hue, and he noticed fur shooting out from it. Soon legs began to poke out, and a tail. Before he knew it he was carrying a very heavy ginger cat in his mouth. This was all he needed.

'Did you notice how the *F*-statistics in factorial designs are computed?' the cat said.

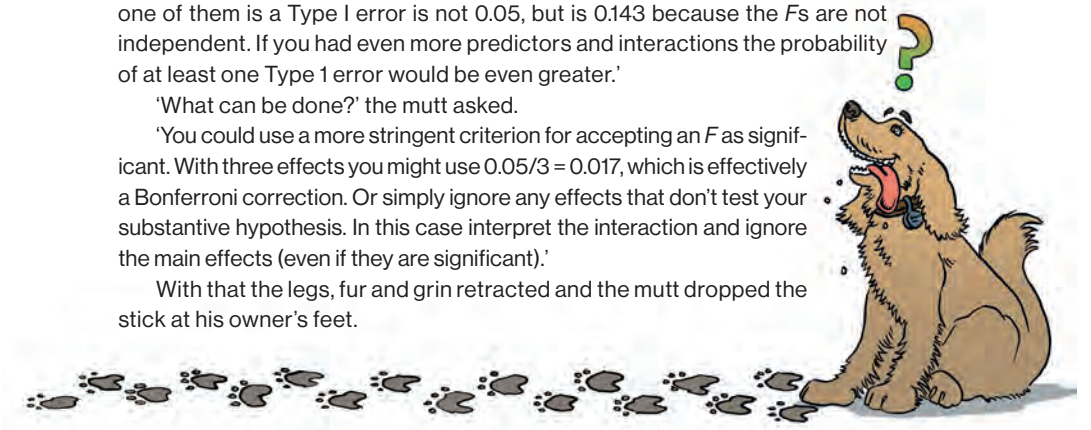
Determined not to be outwitted by a cat, the mutt replied, 'Yes, the mean squares for the effect divided by the residual mean squares'.

'Exactly,' the cat grinned. 'They all use the residual mean squares. This means that they are not independent.' The cat waited as though expecting applause. The mutt was confused, so the cat continued. 'The *F*-statistics are not independent, which means that they do not control the Type I error rate. In the example from your lecture there are three *F*s. The probability that at least one of them is a Type I error is not 0.05, but is 0.143 because the *F*s are not independent. If you had even more predictors and interactions the probability of at least one Type 1 error would be even greater.'

'What can be done?' the mutt asked.

'You could use a more stringent criterion for accepting an *F* as significant. With three effects you might use 0.05/3 = 0.017, which is effectively a Bonferroni correction. Or simply ignore any effects that don't test your substantive hypothesis. In this case interpret the interaction and ignore the main effects (even if they are significant).'

With that the legs, fur and grin retracted and the mutt dropped the stick at his owner's feet.



14.9 Factorial designs using SPSS Ⓑ

14.9.1 General procedure for factorial designs Ⓐ

Using the linear model to test differences between means on several predictors/independent variables requires the same steps as for one predictor/independent variable, so refer back to Figure 12.12 as a guide.

14.9.2 Entering the data and accessing the main dialog box Ⓑ

To enter the data we need two different coding variables to represent type of face and alcohol consumption. Create a variable called **facetype** in the data editor. We've had a lot of experience with coding variables, so hopefully you are able to define value labels to represent the two types of faces: I've used the codes unattractive = 0 and attractive = 1. Having created the variable, enter a code of 0 or 1 in the **facetype** column indicating the group to which the participant was assigned. Create a second variable called **alcohol** and assign value labels of placebo = 0, low dose = 1 and high dose = 2. In the data editor, enter 0, 1 or 2 into the alcohol column to represent the amount of alcohol consumed by the participant. Remember that if you turn on the *value labels* option you will see text in the data editor rather than the numerical codes. The coding scheme I have suggested is summarized in Table 14.3. Once you have created the two coding variables, create a third called **attractiveness** and use the *labels* option to give it the fuller name of *Median attractiveness rating*. Enter the scores in Table 14.1 into this column, taking care to ensure that each score is associated with the correct combination of type of face and alcohol.

Table 14.3 Coding two independent variables

Facial stimulus	Alcohol	The participant ...
0	0	Rated unattractive faces after a placebo drink
0	1	Rated unattractive faces after a low dose of alcohol
0	2	Rated unattractive faces after a high dose of alcohol
1	0	Rated attractive faces after a placebo drink
1	1	Rated attractive faces after a low dose of alcohol
1	2	Rated attractive faces after a high dose of alcohol



Use the Chart Builder to plot an error bar graph of the attractiveness ratings with alcohol consumption on the *x*-axis and different coloured lines to represent whether the faces being rated were unattractive or attractive.

To fit a linear model for independent factorial designs select *Analyze* > *General Linear Model* > Univariate... to access the dialog box in Figure 14.9. Drag the outcome variable **attractiveness** from the variables list on the left-hand side and to the space labelled *Dependent Variable* (or click ▸). Drag **alcohol** and **facetype** (to select them simultaneously hold down *Ctrl*, or *Cmd* on a Mac, while clicking each one) from the variables list to the *Fixed Factor(s)* box (or click ▸). There are various other spaces that are available for conducting more complex analyses such as factorial ANCOVA, which extends the model at the beginning of this chapter to include a covariate (as in the previous chapter).

14.9.3 Plotting effects Ⓑ

Click Plots... to access the dialog box in Figure 14.10, which allows you to specify line graphs of your data. These plots can be useful for interpreting interaction effects but they do tend to be scaled so as to create the impression of massive differences between means, so, seriously, plot your data properly *before* the main analysis. To plot the main effect of **facetype** drag this variable to the space



Oliver Twisted
Please, Sir, can I ... customize my model?

'My friend told me that there are different types of sums of squares,' complains Oliver with an air of impressive authority. 'Why haven't you told us about them? Is it because you have a microbe for a brain?' No, it's not, Oliver, it's because everyone but you will find this very tedious. If you want to find out more about what the Model... button does, and the different types of sums of squares that can be used, then visit www.discoverspss.com.



Figure 14.9 Main dialog box for independent factorial designs

labelled *Horizontal Axis* (or click ▸) and click Add to add the plot to the list at the bottom of the box (Figure 14.10, top). Plot the main effect of **alcohol** in the same way. To create an **interaction graph** to show the combined effect of type of face and alcohol consumption, drag **alcohol** from the variables list to the space labelled *Horizontal Axis* (or click ▸) and drag **facetype** to the space

labelled *Separate Lines*. (You could plot the variables the other way round – it doesn't matter, so decide which way produces a plot that helps you to make sense of the interaction.) When you have moved the two variables to the appropriate boxes, click Add to add the plot to the list at the bottom of the box (Figure 14.10, bottom). You can choose either a bar or line chart (I have chosen a line chart), and you should ask for error bars displaying confidence intervals. Using the default scaling, you can end up with a chart showing apparently massive group differences that turn out to be tiny when you inspect the *y*-axis. To avoid this crushing disappointment, it is generally (although not always) a good idea to select *Y axis starts at 0* to scale the *y*-axis from zero. Click Continue to return to the main dialog box.

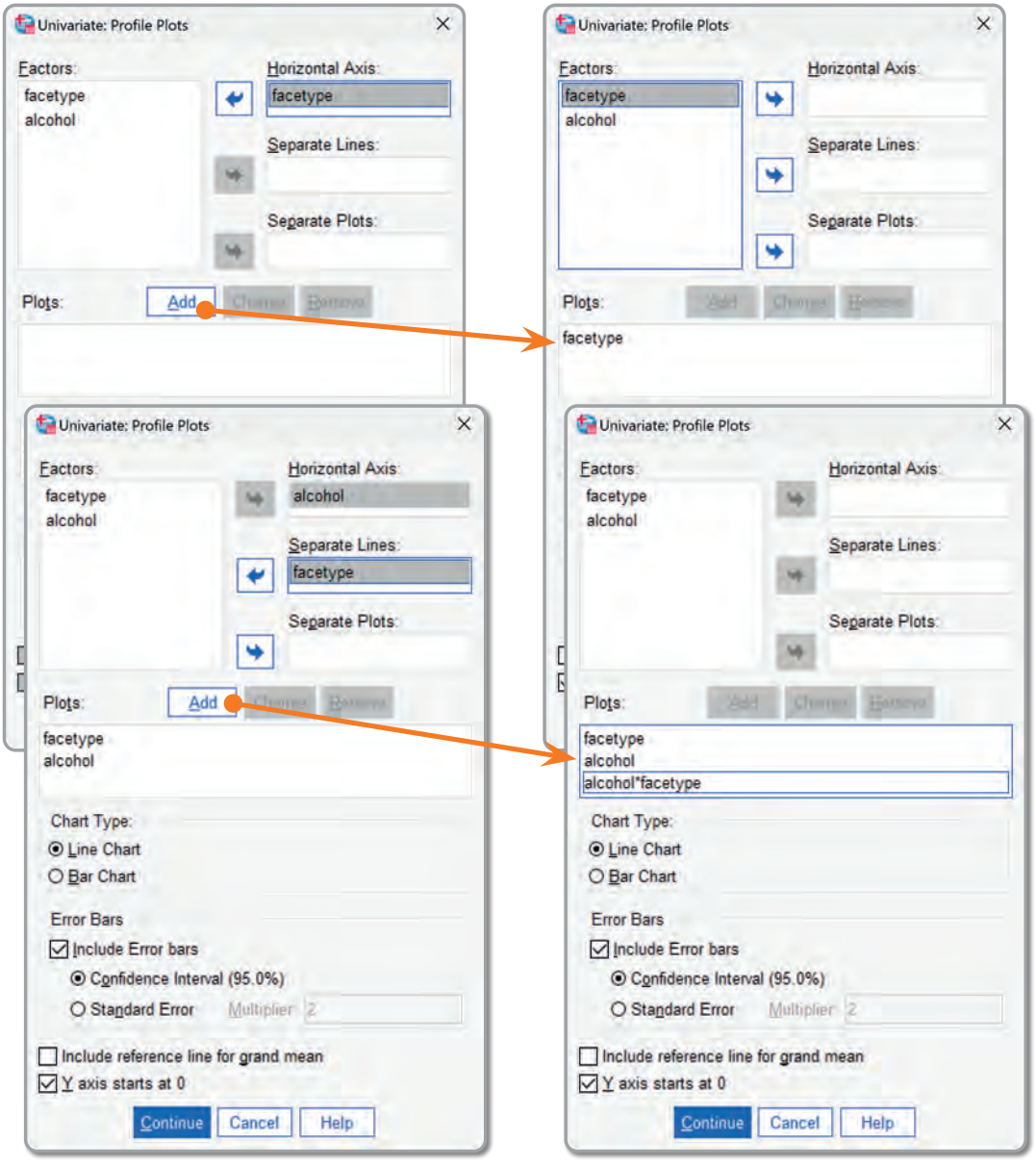


Figure 14.10 Defining plots for independent factorial designs

14.9.4 Contrasts Ⓑ

We saw in Chapter 12 that contrasts help us to break down main effects and tell us where the differences between groups lie. With one independent variable we could enter codes to define the contrasts we want. However, with two independent variables no such facility exists (although it can be done using syntax – see Oliver Twisted). Instead we are restricted to several standard contrasts that were described in Table 12.8.



Oliver Twisted
Please, Sir, can I have some more ... contrasts?

'I don't want to use standard contrasts,' sulks Oliver as he stamps his feet on the floor. 'They smell of rotting cabbage.' Actually, Oliver, I think the stench of rotting cabbage is because you stood your Dickensian self under Mr. Mullycents' window when he emptied his toilet bucket into the street. Nevertheless, I do get asked a fair bit about how to do contrasts with syntax and because I'm a complete masochist I've prepared a fairly detailed guide in the additional material for this chapter. These contrasts are useful to follow up a significant interaction effect.

Standard contrasts will be fine for this example. The type of face had only two levels, so we don't need contrasts for that main effect because it compares only two means. However, the effect of alcohol had three levels: placebo, low dose and high dose. We could select a simple contrast and use the first category as a reference. Doing so would compare the low-dose group to the placebo group, and then compare the high-dose group to the placebo group. As such, the alcohol groups would get compared to the placebo group. We could also select a *repeated* contrast, which would compare the low-dose group to the placebo group, and then the high-dose group to the low-dose group (i.e., it moves through the groups comparing each group to the one before). Again, this might be useful. We could also do a *Helmert* contrast, which compares each category against all subsequent categories. In this case it would compare the placebo group to the remaining categories (i.e., all of the groups that had some alcohol) and then would move on to the low-dose group and compare it to the high. Any of these contrasts would be fine, but they give us contrasts only for the main effects. In reality, most of the time we want contrasts for our interaction term, and they can be obtained only through syntax (it looks like you might have to look at Oliver Twisted after all!).

To get contrasts for the main effect of alcohol click Contrasts... in the main dialog box. We've used the C*ontrasts* dialog box before, so refer back to Section 13.6.5 to help you select a Helmert contrast for the **alcohol** variable (Figure 14.11). Click Continue to return to the main dialog box.

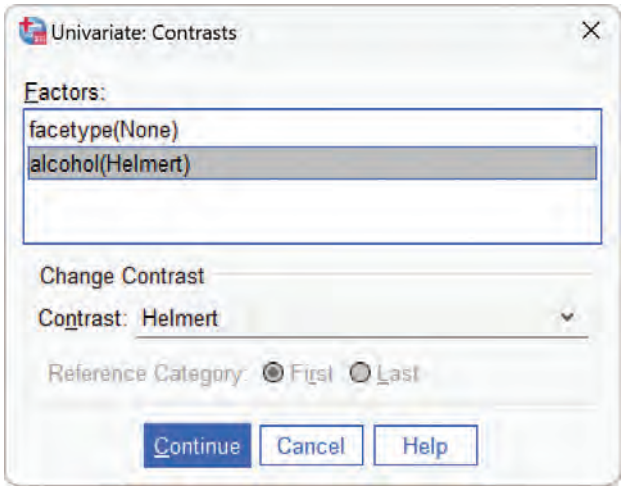


Figure 14.11 Defining contrasts for independent factorial designs

14.9.5 *Post hoc* tests Ⓑ

Click Post Hoc... to access *post hoc* tests (Figure 14.12). The variable **facetype** has only two levels and so we don't need *post hoc* tests (because any significant effects must reflect the difference between unattractive and attractive faces). However, there were three levels of **alcohol** (placebo, low dose and high dose) so if we didn't have prior hypotheses to test we might want *post hoc* tests (remember that normally you would conduct contrasts *or post hoc* tests, not both). Drag the variable **alcohol** from the box labelled *Factors* to the box labelled *Post Hoc Tests for:*. For my recommendations for *post hoc* procedures see Section 12.6 or select the ones in Figure 14.12. Click Continue to return to the main dialog box.

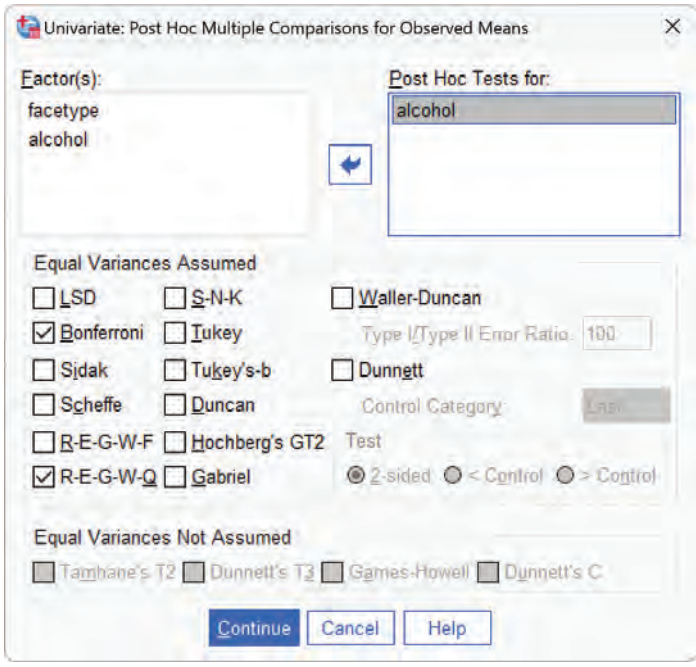


Figure 14.12 Dialog box for *post hoc* tests

14.9.6 Simple effects, bootstrapping and other options Ⓑ

We explored many of the other available options in the previous chapter (see Section 13.6.6 and Jane Superbrain Box 13.3). Click EM Means... to access the dialog box in Figure 14.13 (left). To get estimated marginal means transfer all effects into the box labelled *Display Means for*. By dragging an interaction term into that box the option to ☑ Compare simple main effects becomes available. Select this option to conduct simple effects analysis on the interaction (Section 14.5). Click Options... to activate the dialog box in Figure 14.13 (right). We'll choose the same options as we did in Section 13.6.6 including ☑ Estimate effect sizes to obtain partial eta squared.

The main dialog box contains the Bootstrap... button, which can be used to specify bootstrapped confidence intervals for the estimated marginal means, descriptive statistics and *post hoc* tests, but not the main *F*-statistic. This option is mainly useful if you plan to look at *post hoc* tests, which we will, so select the options described in Section 6.10.8. If you have asked for plots then you'll need to run the model *without* the bootstrap options selected to get the plots, then rerun it with the bootstrap options selected to get the robust confidence intervals in the output. Once these options have been selected click Continue to return to the main dialog box, then click OK.

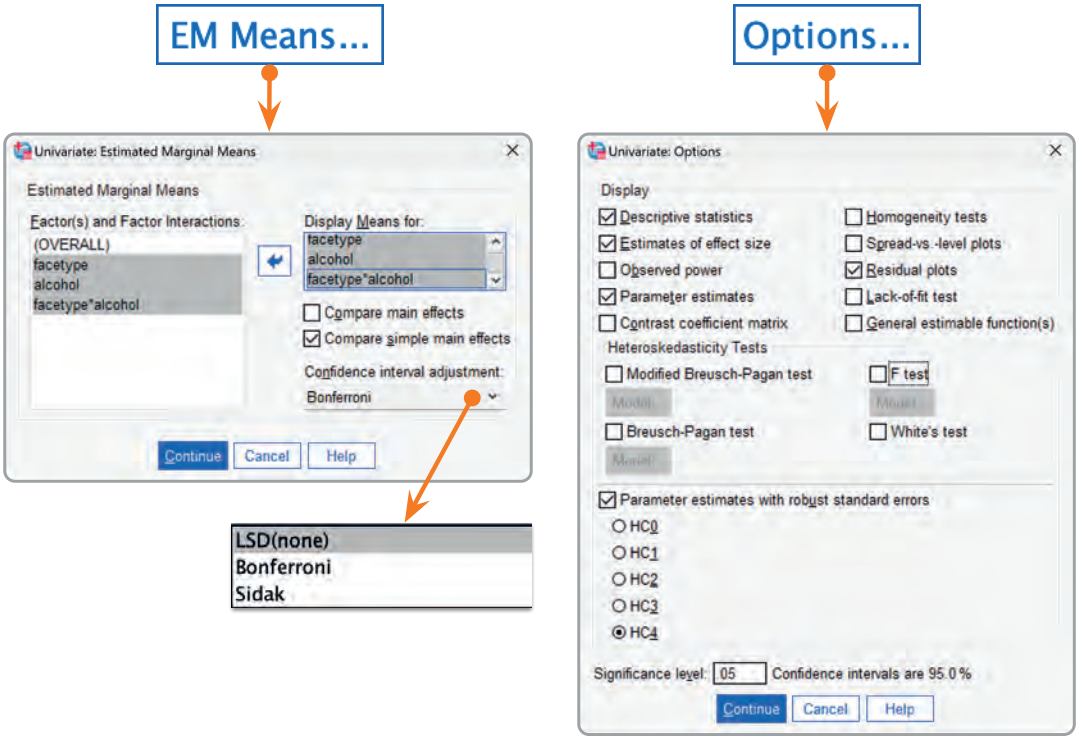


Figure 14.13 *Estimated Marginal Means* and *Options* dialog boxes for GLM univariate

14.10 Output from factorial designs Ⓑ

14.10.1 The main effect of type of face Ⓑ

Output 14.2 tells us whether any of the predictors/independent variables had a significant effect on attractiveness ratings. Note that the sums of squares, mean squares and *F*-statistics match (to within rounding error) the values we calculated in Section 14.7. Let's assume we adopted an alpha of 0.05.



Oditi's Lantern
Factorial ANOVA

'I, Oditi, enjoy interactions immensely. I want to interact with all of my followers, invite them around to my large desert ranch and let them sup on my tasty mint tea. I grow mint in my special mushroom patch, which gives it a unique flavour, and sometimes makes people obey my every command. I have learnt that interactions like these are powerful tools to understand the secrets of global domina... erm, I mean "life" and how to breed cute bunny rabbits of love. Stare into my lantern and discover more about factorial designs.'



The main effect of type of face is significant because the p associated with the F-statistic is given as 0.000, which is less than 0.05. This effect means that overall *when we ignore how much alcohol had been drunk* the type of face being rated significantly affected attractiveness ratings. This effect has $\eta_p^2 = 0.27$, suggesting that the type of face explains 27% of the variance in attractiveness ratings that is not explained by the other predictors.

Tests of Between-Subjects Effects

Dependent Variable: Median attractiveness rating

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	61.167[a]	5	12.233	8.936	<.001	.515
Intercept	1541.333	1	1541.333	1125.843	<.001	.964
facetype	21.333	1	21.333	15.583	<.001	.271
alcohol	16.542	2	8.271	6.041	.005	.223
facetype * alcohol	23.292	2	11.646	8.507	<.001	.288
Error	57.500	42	1.369			
Total	1660.000	48				
Corrected Total	118.667	47				

a. R Squared = .515 (Adjusted R Squared = .458)

Output 14.2

We can visualize this effect by plotting the average attractiveness rating at each level of type of face (ignoring the dose of alcohol completely) as in Figure 14.14. You'll get a less pretty version of this plot if you asked for a profile plot of this effect (Section 14.9.3). The resulting plot shows that the significant main effect reflects the fact that average attractiveness ratings were higher for the photos of attractive faces than unattractive ones. Of course, this result is not at all surprising because the attractive faces were pre-selected to be more attractive than the unattractive faces. This result is a useful manipulation check though: our participants, other things being equal, found the attractive faces more attractive than the unattractive ones.

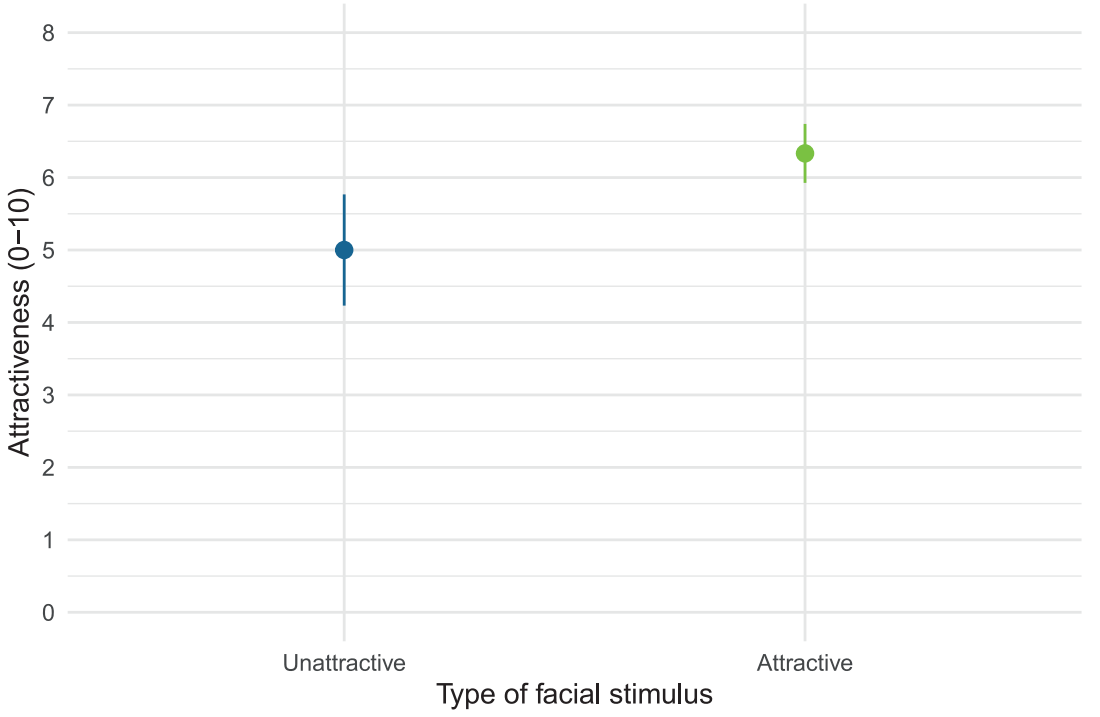


Figure 14.14 Graph to show the main effect of type of face on attractiveness ratings

14.10.2 The main effect of alcohol Ⓑ

Output 14.2 also shows a significant main effect of alcohol on attractiveness ratings (because the significance value of 0.005 is less than 0.05). This result means that *when we ignore whether the participant rated unattractive or attractive faces* the amount of alcohol influenced their attractiveness ratings. This effect has $\eta_p^2 = 0.22$, suggesting that the type of face explains 22% of the variance in attractiveness ratings that is not explained by the other predictors. The best way to understand this effect is to plot the average attractiveness rating at each level of alcohol (ignoring the type of face completely) as in Figure 14.15. Again, if you asked for a profile plot of this effect (Section 14.9.3) you'll find something similar in your output. The plot shows that the mean attractiveness ratings increase quite linearly as more alcohol is drunk. This significant main effect is *likely* to reflect this trend. Looking at the error bars (95% confidence intervals), there is a lot of overlap between the placebo and low-dose groups (implying that these groups have similar average ratings), but the overlap between the placebo and high-dose group is less, and quite possibly within what you'd expect from a significant difference (see Section 2.10.10). The confidence intervals for the low- and high-dose group also overlap a lot, suggesting that these

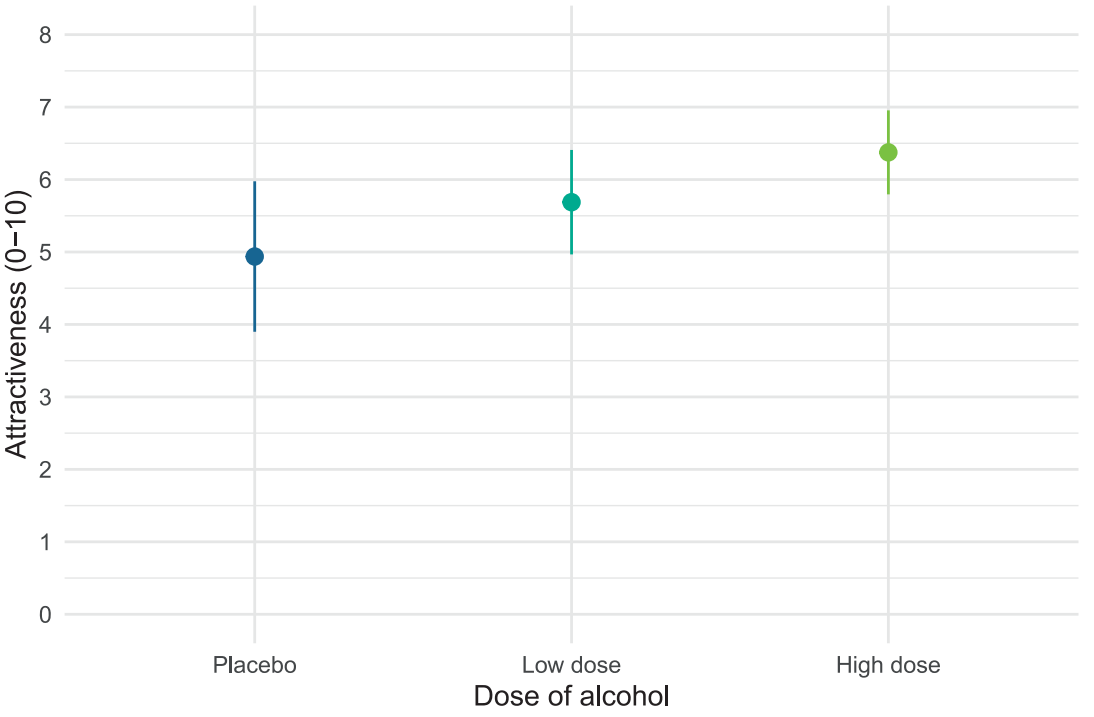


Figure 14.15 Graph showing the main effect of alcohol on attractiveness ratings

groups do not differ. Therefore, we might speculate based on the confidence intervals that this main effect reflects a difference between the placebo and high-dose groups but that no other groups differ. It could also reflect the linear increase in ratings as the dose of alcohol increases.

14.10.3 Contrasts for alcohol Ⓑ

Output 14.3 shows the Helmert contrasts on the effect of alcohol. Ordinarily we wouldn't look at these because the interaction involving alcohol was significant, but for the sake of explanation I will. The top of the table shows the contrast for *Level 1 vs. Later*, which in this case means the placebo group compared to the two alcohol groups. This contrast tests whether the mean of the placebo group (4.938) is different from the mean of the low- and high-dose groups combined ((5.688 + 6.375)/2 = 6.032). This is a difference of –1.094 (4.938 – 6.032), which both the *Contrast Estimate* and the *Difference* in the table tell us. The value of *Sig.* (0.004) tells us that this difference is significant because it is smaller than the criterion of 0.05. The confidence interval for this difference also doesn't cross zero so, assuming this sample is one of the 95 out of 100 that produce a confidence interval containing the true value of the difference, the real difference between the placebo and alcohol groups is not zero (between –1.817 and –0.371, to be precise). We could conclude that the effect of alcohol is that any amount of alcohol increases the attractiveness ratings of pictures compared to when a placebo was drunk. However, we need to look at the remaining contrast to qualify this statement.

The bottom of the table shows the contrast for *Level 2 vs. Level 3*, which in this case means the low-dose group compared to the high. This contrast compares the mean of the low-dose group (5.688) to the mean of the high (6.375). This is a difference of –0.687 (5.688 – 6.375), which is the value of both the *Contrast Estimate* and the *Difference* in the table. This difference is not significantly different from zero (because *Sig.* is 0.104, which is greater than 0.05). The confidence interval for this difference also crosses zero so, assuming this sample is one of the 95 out of 100 that produced

confidence intervals that contain the true value of the difference, the real difference is between −1.522 and 0.147 and could be zero. This contrast tells us that having high dose of alcohol doesn't significantly affect attractiveness ratings compared to having a low dose.

Contrast Results (K Matrix)

Alcohol consumption Helmert Contrast			Dependent Variable: Median attractiveness rating
Level 1 vs. Later	Contrast Estimate		-1.094
	Hypothesized Value		0
	Difference (Estimate – Hypothesized)		-1.094
	Std. Error		.358
	Sig.		.004
	95% Confidence Interval for Difference	Lower Bound	-1.817
		Upper Bound	-.371
Level 2 vs. Level 3	Contrast Estimate		-.687
	Hypothesized Value		0
	Difference (Estimate – Hypothesized)		-.687
	Std. Error		.414
	Sig.		.104
	95% Confidence Interval for Difference	Lower Bound	-1.522
		Upper Bound	.147

Output 14.3

14.10.4 *Post hoc* tests for alcohol Ⓑ

The REGWQ test (Output 14.4) confirms that the means of the placebo and low-dose conditions were equivalent ($p = 0.077$), as were the means of the low- and high-dose groups ($p = 0.104$). I note again that we wouldn't normally interpret *post hoc* tests when there is a significant interaction involving that main effect (as there is here).

Median attractiveness rating

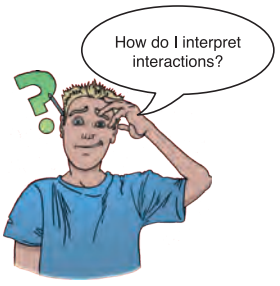
	Alcohol consumption	N	Subset 1	Subset 2
Ryan-Einot-Gabriel-Welsch Range[a]	Placebo	16	4.94	
	Low dose	16	5.69	5.69
	High dose	16		6.37
	Sig.		.077	.104

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 1.369.

a. Alpha = .05.

Output 14.4

14.10.5 The interaction effect Ⓑ



Finally, Output 14.2 tells us about the interaction between the effect of type of face and the effect of alcohol. The *F*-statistic is highly significant (because the observed *p*-value of 0.001 is less than 0.05). This effect means that the effect of alcohol on attractiveness ratings was different when rating unattractive faces compared to when rating attractive faces. This effect has $\eta_p^2 = 0.29$, suggesting that the type of face explains 29% of the variance in attractiveness ratings that is not explained by the other predictors. The output will include the (probably badly scaled) plot that we asked for in Figure 14.10, which plots the means in Output 14.5. Figure 14.16 is a nicer version of this plot, which we can use to get a handle on the interaction effect. Focus first on the green line connecting the triangles, which is flat and shows very little difference in average attractiveness ratings across the alcohol conditions. This line shows that *when rating attractive faces, alcohol has very little effect*. Now look at the blue line connecting the circles, which slopes upwards, showing that when rating unattractive faces ratings increase with the dose of alcohol. This line shows that *when rating unattractive faces, alcohol has an effect*. The significant interaction seems to reflect the differing effect of alcohol when rating attractive and unattractive faces; that is, that alcohol has an effect on ratings of unattractive faces but not for attractive ones. This example illustrates an important point about interaction effects. We concluded earlier that alcohol significantly affected attractiveness ratings (the **alcohol** main effect) but the interaction effect qualifies this conclusion by showing that this is true only when rating unattractive faces (ratings of attractive faces seem unaffected by alcohol). In short, *you should not interpret a main effect in the presence of a significant interaction involving that main effect*.

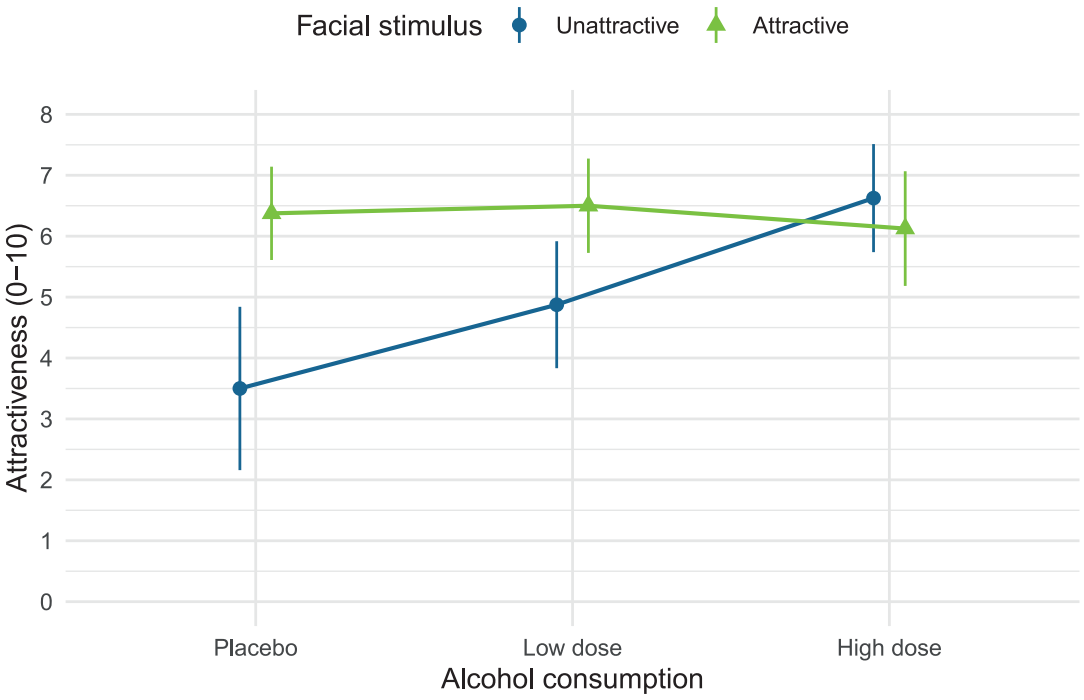


Figure 14.16 Graph of the interaction of type of face and alcohol consumption on attractiveness ratings

3. Attractivenss of facial stimuli * Alcohol consumption

Estimates

Dependent Variable: Median attractiveness rating

Attractivenss of facial stimuli	Alcohol consumption	Mean	Std. Error	95% Confidence Interval		Bootstrap for Mean[a]			
				Lower Bound	Upper Bound	Bias	Std. Error	BCa 95% Confidence Interval Lower	Upper
Unattractive	Placebo	3.500	.414	2.665	4.335	−.018	.570	2.429	4.571
	Low dose	4.875	.414	4.040	5.710	.006	.433	4.000	5.800
	High dose	6.625	.414	5.790	7.460	.002	.363	6.000	7.333
Attractive	Placebo	6.375	.414	5.540	7.210	−.002	.312	5.826	7.000
	Low dose	6.500	.414	5.665	7.335	.015	.316	5.833	7.200
	High dose	6.125	.414	5.290	6.960	.007	.404	5.375	7.000

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 14.5

14.10.6 Simple effects Ⓑ

We can use the simple effects analysis (Section 14.5) to pick apart the interaction and confirm (or otherwise) our speculation about what's driving the effect. Output 14.6 shows the effect of facial stimulus (labelled *Contrast* in the output) at each level of alcohol. From this output it appears that there was a significant difference in ratings of unattractive and attractive faces in the placebo group, $F(1, 42) = 24.15$, $p < 0.001$, $\eta_p^2 = 0.37$, and in the low-dose group, $F(1, 42) = 7.72$, $p = 0.008$, $\eta_p^2 = 0.16$, but not in the high-dose group, $F(1, 42) = 0.73$, $p = 0.398$, $\eta_p^2 = 0.02$. These results confirm our speculation from the previous section.

Univariate Tests

Dependent Variable: Median attractiveness rating

Alcohol consumption		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Placebo	Contrast	33.063	1	33.063	24.150	<.001	.365
	Error	57.500	42	1.369			
Low dose	Contrast	10.563	1	10.563	7.715	.008	.155
	Error	57.500	42	1.369			
High dose	Contrast	1.000	1	1.000	.730	.398	.017
	Error	57.500	42	1.369			

Each F tests the simple effects of Attractivenss of facial stimuli within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

Output 14.6

You will also find the table in Output 14.7 somewhere among the stampede of tables in your workbook. This table shows the effect of alcohol (again labelled *Contrast*) for each type of facial stimulus. It appears that there was a significant difference in ratings across the alcohol groups when rating unattractive, $F(2, 42) = 14.34$, $p < 0.001$, $\eta_p^2 = 0.41$, but not when rating attractive faces, $F(2, 42) = 0.21$, $p = 0.809$, $\eta_p^2 = 0.01$. Again, these results confirm our conclusions from the previous section.

14.10.7 *Post hoc* tests for the interaction Ⓑ

Let's start by saying you wouldn't normally do simple effects analysis *and post hoc* tests (Jane Superbrain Box 14.2). SPSS produces *post hoc* tests both for the effects of facial stimulus within each level of alcohol (Output 14.8) and for the effect of alcohol within each level of facial stimulus

Univariate Tests

Dependent Variable: Median attractiveness rating

Attractivenss of facial stimuli		Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Unattractive	Contrast	39.250	2	19.625	14.335	<.001	.406
	Error	57.500	42	1.369			
Attractive	Contrast	.583	2	.292	.213	.809	.010
	Error	57.500	42	1.369			

Each F tests the simple effects of Alcohol consumption within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

Output 14.7

(Output 14.9). In both cases I have reproduced the non-robust and bootstrapped versions. I'll interpret the robust bootstrapped versions, but the conclusions are the same if you use the non-robust tests. Output 14.8 basically replicates the simple effects in Output 14.6 because one of our variables has only two categories. Note that within each alcohol group the two rows contain the same effect but reversed. For example, the first row is the difference between ratings of unattractive faces compared to attractive ($\bar{X}_{\text{Difference}} = \bar{X}_{\text{Unattractive}} - \bar{X}_{\text{Attractive}}$) whereas the second is the difference between ratings of attractive faces compared to unattractive ($\bar{X}_{\text{Difference}} = \bar{X}_{\text{Attractive}} - \bar{X}_{\text{Unattractive}}$), which flips the sign of the difference from –2.875 to 2.875. Basically, these rows contain the same information, so choose to interpret either the even rows or the odd rows, but not both. I will report the rows that compare attractive faces to unattractive so that positive effects mean that attractive faces were rated higher than unattractive ones (these rows are highlighted in Output 14.8). We can conclude that attractive faces were rated significantly higher than unattractive ones in the placebo group, $\bar{X}_{\text{Difference}} = 2.88$ [4.14, 1.66], $p < 0.001$, and in the low-dose group, $\bar{X}_{\text{Difference}} = 1.63$ [0.64, 2.71], $p = 0.006$, but not in the high-dose group, $\bar{X}_{\text{Difference}} = -0.50$ [–1.68, 0.68], $p = 0.398$.

Output 14.9 shows the difference in attractiveness ratings between the alcohol groups within each of the facial stimulus types. Again there is redundancy because, for example, the table shows

Pairwise Comparisons

Dependent Variable: Median attractiveness rating

Alcohol consumption	(I) Attractivenss of facial stimuli	(J) Attractivenss of facial stimuli	Mean Difference (I–J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Placebo	Unattractive	Attractive	-2.875*	.585	<.001	-4.056	-1.694
	Attractive	Unattractive	2.875*	.585	<.001	1.694	4.056
Low dose	Unattractive	Attractive	-1.625*	.585	.008	-2.806	-.444
	Attractive	Unattractive	1.625*	.585	.008	.444	2.806
High dose	Unattractive	Attractive	.500	.585	.398	-.681	1.681
	Attractive	Unattractive	-.500	.585	.398	-1.681	.681

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Bootstrap for Pairwise Comparisons

Dependent Variable: Median attractiveness rating

Alcohol consumption	(I) Attractivenss of facial stimuli	(J) Attractivenss of facial stimuli	Mean Difference (I–J)	Bootstrap[a] Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval Lower	Upper
Placebo	Unattractive	Attractive	-2.875	-.016	.653	<.001	-4.141	-1.664
	Attractive	Unattractive	2.875	.016	.653	<.001	1.535	4.219
Low dose	Unattractive	Attractive	-1.625	-.009	.540	.006	-2.700	-.654
	Attractive	Unattractive	1.625	.009	.540	.006	.643	2.708
High dose	Unattractive	Attractive	.500	-.006	.539	.355	-.585	1.495
	Attractive	Unattractive	-.500	.006	.539	.355	-1.500	.599

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 14.8

Pairwise Comparisons

Dependent Variable: Median attractiveness rating

Attractivenss of facial stimuli	(I) Alcohol consumption	(J) Alcohol consumption	Mean Difference (I–J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
Unattractive	Placebo	Low dose	−1.375	.585	.071	−2.834	.084
		High dose	−3.125*	.585	<.001	−4.584	−1.666
	Low dose	Placebo	1.375	.585	.071	−.084	2.834
		High dose	−1.750*	.585	.014	−3.209	−.291
	High dose	Placebo	3.125*	.585	<.001	1.666	4.584
		Low dose	1.750*	.585	.014	.291	3.209
Attractive	Placebo	Low dose	−.125	.585	1.000	−1.584	1.334
		High dose	.250	.585	1.000	−1.209	1.709
	Low dose	Placebo	.125	.585	1.000	−1.334	1.584
		High dose	.375	.585	1.000	−1.084	1.834
	High dose	Placebo	−.250	.585	1.000	−1.709	1.209
		Low dose	−.375	.585	1.000	−1.834	1.084

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Bootstrap for Pairwise Comparisons

Dependent Variable: Median attractiveness rating

Attractivenss of facial stimuli	(I) Alcohol consumption	(J) Alcohol consumption	Mean Difference (I–J)	Bias	Std. Error	Bootstrap[a] Sig. (2-tailed)	BCa 95% Confidence Interval Lower	Upper
Unattractive	Placebo	Low dose	−1.375	−.024	.719	.068	−2.833	.084
		High dose	−3.125	−.019	.681	<.001	−4.516	−1.828
	Low dose	Placebo	1.375	.024	.719	.068	−.094	2.878
		High dose	−1.750	.004	.557	.007	−2.850	−.603
	High dose	Placebo	3.125	.019	.681	<.001	1.828	4.517
		Low dose	1.750	−.004	.557	.007	.708	2.783
Attractive	Placebo	Low dose	−.125	−.017	.447	.794	−.961	.667
		High dose	.250	−.009	.507	.614	−.717	1.170
	Low dose	Placebo	.125	.017	.447	.794	−.833	1.167
		High dose	.375	.008	.509	.470	−.684	1.333
	High dose	Placebo	−.250	.009	.507	.614	−1.273	.856
		Low dose	−.375	−.008	.509	.470	−1.333	.667

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

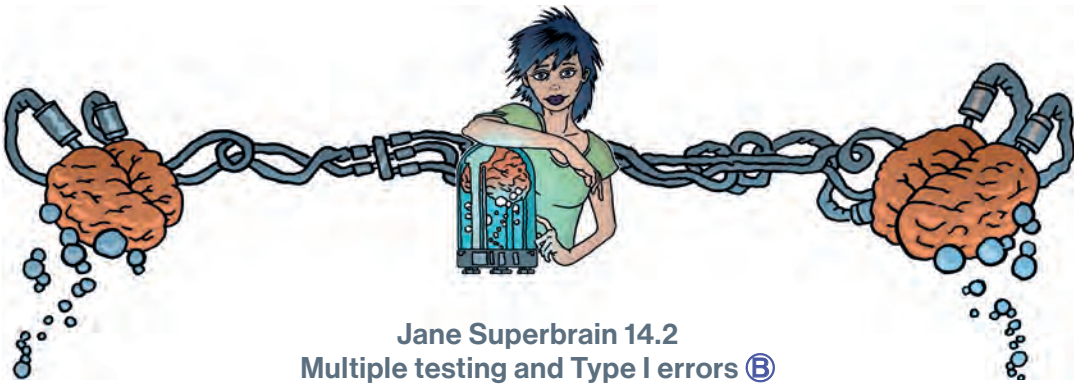
Output 14.9

the difference between the placebo and low-dose group ($\bar{X}_{\text{Difference}} = \bar{X}_{\text{Placebo}} - \bar{X}_{\text{Low dose}}$) and also the difference between the low-dose group and the placebo ($\bar{X}_{\text{Difference}} = \bar{X}_{\text{Low dose}} - \bar{X}_{\text{Placebo}}$). These are the same effects but with the direction of the effect reversed. Again I have highlighted the rows in Output 14.9 that show unique effects in the directions that make the most sense (i.e., we'd expect ratings to increase as a function of alcohol consumption, so it makes sense to look at the low and high doses compared to placebo, and high compared to low (rather than the other way around).

For unattractive facial stimuli we can conclude that faces were not rated significantly differently in the low-dose group compared to placebo, $\bar{X}_{\text{Difference}} = 1.38$ [–0.09, 2.88], $p = 0.068$, but were rated significantly higher in the high-dose group compared to both placebo, $\bar{X}_{\text{Difference}} = 3.13$ [1.83, 4.52], $p < 0.001$, and low-dose groups, $\bar{X}_{\text{Difference}} = 1.75$ [0.71, 2.78], $p = 0.007$. However, for attractive facial stimuli faces were not rated significantly differently in the low-dose group compared to placebo, $\bar{X}_{\text{Difference}} = 0.13$ [–0.83, 1.17], $p = 0.794$, nor were they rated significantly differently in the high-dose group compared to both placebo, $\bar{X}_{\text{Difference}} = -0.25$ [–1.27, 0.86], $p = 0.614$, and low-dose groups, $\bar{X}_{\text{Difference}} = -0.38$ [–1.33, 0.67], $p = 0.470$.

14.11 Robust models of factorial designs ©

If you requested robust standard errors in Figure 14.13 then you'll find a table of parameter estimates for the model in the output with confidence intervals and *p*-values robust to heteroscedasticity (Output 14.10). These are a little hard to unpick. SPSS automatically uses the last category



Jane Superbrain 14.2
Multiple testing and Type I errors Ⓑ

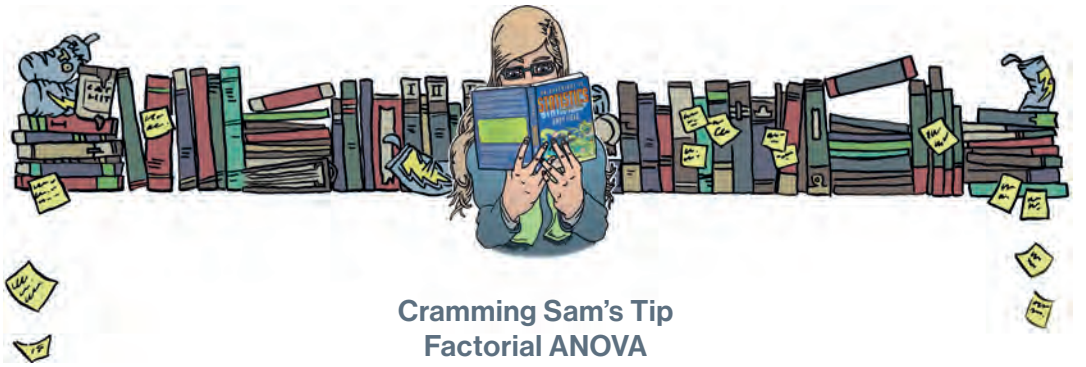
Textbooks need to try to cover most of the eventualities that readers face, so for educational reasons they often are unselective in the options they select. For example, you wouldn't normally do simple effects analysis *and post hoc* tests because the number of significance tests you're conducting quickly escalates and so will your Type I error rate (the probability of falsely discovering at least one significant effect). For example, with the *post hoc* tests in this chapter, although the outputs contain repetition, there are three unique tests in Output 14.8 and six in Output 14.9 and although we chose a Bonferroni adjustment, this is applied *within* each table not *across* tables. If we also factor in the five simple effects tests (Output 14.6 and Output 14.7), we've done a whopping 14 significance tests to break down the interaction. We either need to adopt a more stringent alpha level when interpreting these tests, or we need to do fewer tests. In this example, I would do one of the following (in order of preference):

- Interpret only the simple effects in Output 14.6. To my mind, looking at the effect of facial stimulus within each level of alcohol more elegantly illustrates what the interaction effect shows than looking at the effect of alcohol within each facial stimulus.
- Interpret only the *post hoc* tests in Output 14.9 because these very thoroughly and clearly show the pattern of differences across means.



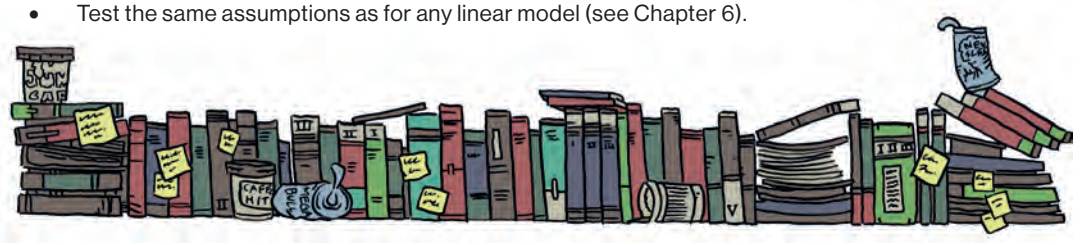
as the reference category. The parameter estimate for face type, because it had only two categories, gives us a robust estimate of the effect on ratings of attractive faces compared to unattractive ones. For the main effect of alcohol the first estimate *[alcohol = 0]* compares the placebo to the high dose and *[alcohol = 1]* compares the low dose to the high dose. Both parameters are not significant. For the interaction term, the first parameter, *[facetype = 0]*[alcohol = 0]* suggests that the difference between ratings of attractive and unattractive faces in the placebo was significantly different from the same difference in the high-dose group ($p < 0.001$). The first parameter, *[facetype = 0]*[alcohol = 1]*, suggests that the difference between ratings of attractive and unattractive faces in the low-dose group was significantly different from the same difference in the high-dose group ($p = 0.009$). These conclusions are based on robust confidence intervals and *p*-values, but confirm our earlier conclusions.

It's possible to run a robust test for two- and three-way factorial designs using R and the *WRS2* package (see Section 4.12). The workbook **robust_2way.spwb** (at www.discoverspss.com) contains syntax



Cramming Sam's Tip
Factorial ANOVA

- Two-way independent designs compare several means when there are two independent variables and different entities have been used in all experimental conditions. For example, if you wanted to know whether different teaching methods worked better for different topics, you could take students from four courses (psychology, geography, management and statistics) and assign them to either lecture-based or book-based teaching. The two variables are topic and method of teaching. The outcome might be the end-of-year mark (as a percentage).
- In the table labelled *Tests of Between-Subjects Effects*, look at the column labelled *Sig.* for all main effects and interactions; assuming you're using a criterion of 0.05, if the value is less than or equal to 0.05 then the effect is significant.
- To interpret a significant interaction, plot the interaction and conduct either simple effects analysis or *post hoc* tests.
- You don't need to interpret main effects if an interaction effect involving that variable is significant.
- If significant main effects are not qualified by an interaction then consult contrasts of *post hoc* tests to see which groups differ: significance is shown by values less than or equal to 0.05 in the columns labelled *Sig.*, and bootstrap confidence intervals that do not contain zero.
- Test the same assumptions as for any linear model (see Chapter 6).



Parameter Estimates with Robust Standard Errors

Dependent Variable: Median attractiveness rating

Parameter	B	Robust Std. Error[a]	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound	Partial Eta Squared
Intercept	6.125	.398	15.386	<.001	5.322	6.928	.849
[facetype=0]	.500	.547	.914	.366	−.604	1.604	.020
[facetype=1]	0[b]	.	.	.	.	.	.
[alcohol=0]	.250	.513	.487	.629	−.786	1.286	.006
[alcohol=1]	.375	.515	.728	.471	−.665	1.415	.012
[alcohol=2]	0[b]	.	.	.	.	.	.
[facetype=0] * [alcohol=0]	−3.375	.852	−3.963	<.001	−5.094	−1.656	.272
[facetype=0] * [alcohol=1]	−2.125	.775	−2.742	.009	−3.689	−.561	.152
[facetype=0] * [alcohol=2]	0[b]	.	.	.	.	.	.
[facetype=1] * [alcohol=0]	0[b]	.	.	.	.	.	.
[facetype=1] * [alcohol=1]	0[b]	.	.	.	.	.	.
[facetype=1] * [alcohol=2]	0[b]	.	.	.	.	.	.

a. HC4 method

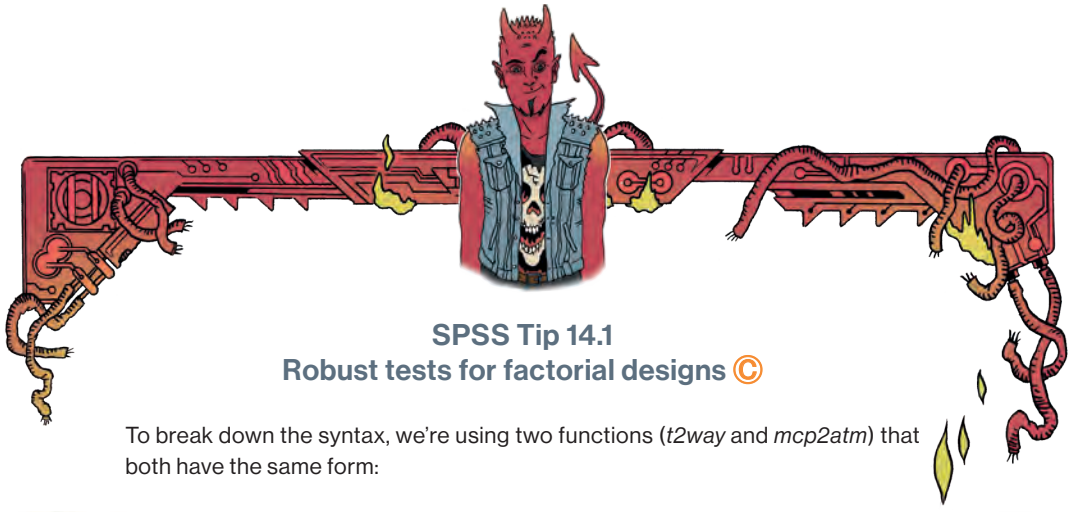
b. This parameter is set to zero because it is redundant.

Output 14.10

to run a robust test of trimmed means for two independent variables/predictors (*t2way*) and associated robust *post hoc* tests (*mcp2atm*) described by Wilcox (2017).[3] The syntax in the file is as follows:

```
BEGIN PROGRAM R.
library(WRS2)
mySPSSdata = spssdata.GetDataFromSPSS(factorMode = "labels")
t2way(attractiveness ~ facetype*alcohol, data = mySPSSdata, tr = 0.2)
mcp2atm(attractiveness ~ facetype*alcohol, data = mySPSSdata, tr = 0.2)
END PROGRAM.
```

Run these six lines of syntax to get the output in Output 14.11. These results confirm what the linear model showed: significant main effects of alcohol, F_t = 10.31, p = 0.019, face type, F_t = 14.57, p = 0.001, and their interaction, F_t = 16.60, p = 0.003. In the *post hoc* tests, *alcohol1* is the difference between placebo and low dose, *alcohol2* is the difference between placebo and high dose, and *alcohol3* is the difference between low and high dose. We can see from the *p*-values and confidence intervals that attractiveness ratings were significantly different between the placebo and high-dose groups,

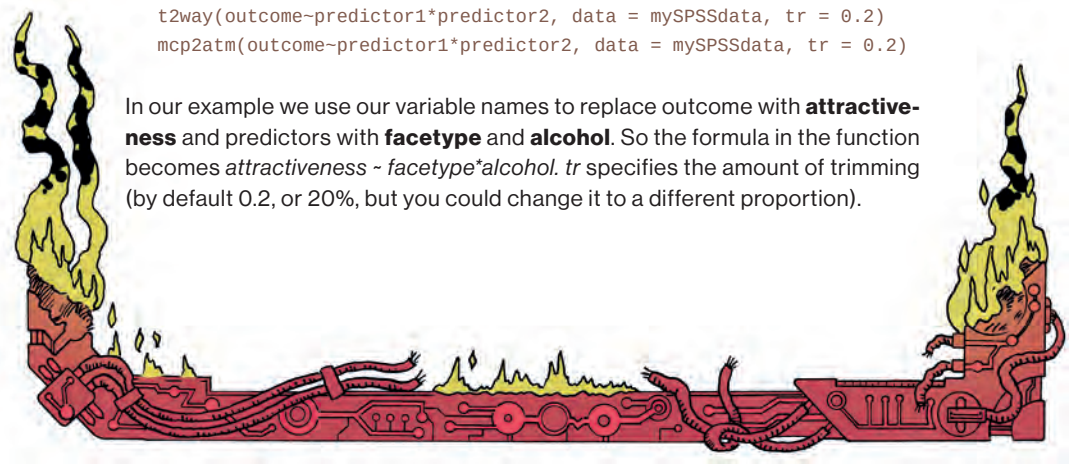


SPSS Tip 14.1
Robust tests for factorial designs ©

To break down the syntax, we're using two functions (*t2way* and *mcp2atm*) that both have the same form:

```
t2way(outcome~predictor1*predictor2, data = mySPSSdata, tr = 0.2)
mcp2atm(outcome~predictor1*predictor2, data = mySPSSdata, tr = 0.2)
```

In our example we use our variable names to replace outcome with **attractiveness** and predictors with **facetype** and **alcohol**. So the formula in the function becomes *attractiveness ~ facetype*alcohol*. *tr* specifies the amount of trimming (by default 0.2, or 20%, but you could change it to a different proportion).



3 You'll also find the file **robust_3way.spwb**, which includes the function *t3way* that performs the same test for designs with three independent variables and predictors, but you'll need to edit this file to include the variable names from your data.

$\hat{\Psi}$ = −2.83, p = 0.005, but not between the placebo and low-dose group, $\hat{\Psi}$ = −1.5, p = 0.080, or between the low- and high-dose groups, $\hat{\Psi}$ = −1.33, p = 0.106. Most interesting are the interaction terms, which look at the difference between ratings of attractive and unattractive faces across the group comparisons we've just examined. These effects tell us that the difference in ratings between attractive and unattractive faces in the high-dose group was significantly different from the corresponding difference in the placebo group, $\hat{\Psi}$ = −3.5, p = 0.001, and the low-dose group, $\hat{\Psi}$ = −2.33, p = 0.008. However, the difference in ratings between attractive and unattractive faces in the low-dose group was not significantly different from the corresponding difference in the placebo group, $\hat{\Psi}$ = −1.17, p = 0.163. In a nutshell, the interaction seems to be driven by effects of a high dose of alcohol compared to the other conditions.

```
t2way(formula = attractiveness ~ facetype * alcohol, data = mySPSSdata,
    tr = 0.2)

                  value p.value
facetype         14.5730   0.001
alcohol          10.3117   0.019
facetype:alcohol 16.6038   0.003

Call:
mcp2atm(formula = attractiveness ~ facetype * alcohol, data = mySPSSdata,
    tr = 0.2)

                     psihat ci.lower ci.upper p-value
facetype1          -3.83333 -5.90979 -1.75688 0.00088
alcohol1           -1.50000 -3.63960  0.63960 0.07956
alcohol2           -2.83333 -5.17113 -0.49554 0.00543
alcohol3           -1.33333 -3.38426  0.71759 0.10569
facetype1:alcohol1 -1.16667 -3.30627  0.97294 0.16314
facetype1:alcohol2 -3.50000 -5.83780 -1.16220 0.00110
facetype1:alcohol3 -2.33333 -4.38426 -0.28241 0.00802
```

Output 14.11

14.12 Bayesian models of factorial designs ©

To analyse a factorial design we have used a linear model, so we can fit a Bayesian model by following the process in Section 9.13. However, although we could specify **facetype** and **alcohol** as *Factor(s)* in this model there is no easy way to specify the interaction term. This is a problem because the interaction term is typically where our interests lie. The way around this is to manually dummy-code **facetype** and **alcohol** (remembering that this variable would become two dummy variables), and then manually create the interaction term by multiplying the **facetype** dummy variable by each of the **alcohol** dummy variables to create two dummy variables for the interaction term. The file **goggles_dummy.sav** contains these dummy variables. Table 14.4 illustrates the coding. The variable **facetype** is coded with 0s for unattractive faces and 1s for attractive faces. The variable **alcohol** is split into two dummy variables (columns 4 and 5), the first of which codes the high dose against placebo (**alc_high_pla**) and the second of which codes the low dose against placebo (**alc_low_pla**). The final two columns code the interaction term into two dummy variables. The first variable (**int_high_pla**) is the codes from columns 3 and 4 multiplied (i.e. **facetype** × **alc_high_pla**). This variable represents the difference between ratings for unattractive faces compared to attractive ones in the placebo group relative to the same difference in the high-dose group. The second variable

(**int_low_pla**) is the codes from columns 3 and 5 multiplied (i.e. **facetype** × **alc_low_pla**). This variable represents the difference between ratings for unattractive faces compared to attractive ones in the placebo group relative to the same difference in the low-dose group.

Table 14.4 Dummy coding scheme for the beer goggles data

Type of faces rated	Alcohol	Face stimulus	alc_high_pla (High vs. Placebo)	alc_low_pla (Low vs. Placebo)	int_high_pla (High vs. Placebo)	int_low_pla (Low vs. Placebo)
Unattractive	Placebo	0	0	0	0	0
Unattractive	Low dose	0	0	1	0	0
Unattractive	High dose	0	1	0	0	0
Attractive	Placebo	1	0	0	0	0
Attractive	Low dose	1	0	1	0	1
Attractive	High dose	1	1	0	1	0

To get Bayes factors and estimates with the default reference priors select *Analyze* ❯ *Bayesian Statistics* ❯ Linear Regression and follow the process in Section 9.13. To specify the model, select and drag **attractiveness** to the box labelled *Dependent* (or click ▸) then select all five dummy variables (**facetype, alc_high_pla, alc_low_pla, int_high_pla, int_low_pla**) and drag them to the box labelled *Covariate(s)* (or click ▸). If you drag the dummy variables to the box labelled *Factor(s)* SPSS will recode them; by using the *Covariate(s)* box the output will match the dummy coding that we used in the data editor. Select ◉ Use Both Methods and leave the default options alone.

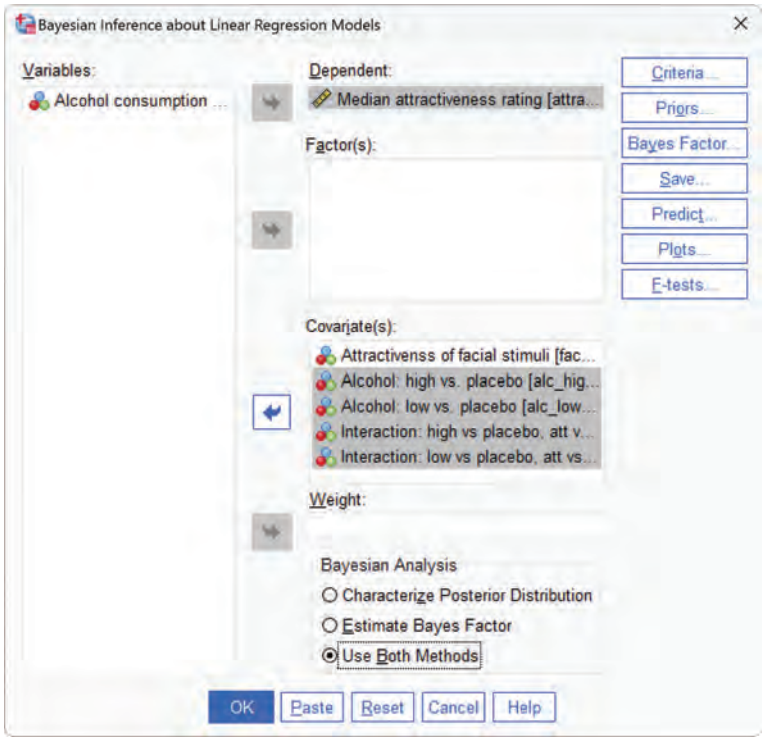


Figure 14.17 Dialog box for Bayesian analysis of factorial designs

Output 14.12 shows the Bayes factor for the model including all five dummy variables against the null model (which includes only the intercept). This Bayes factor is 1195.598. This is very strong evidence for the model. It means that the probability of the data given the model including all five predictors is about 1196 times as great as the probability of the data given the model with only the intercept. We should shift our belief in the model (relative to the null model) by a factor of 1196!

The Bayesian estimate of the *b*s for the dummy variables can be found in the columns labelled *Posterior Mode* and *Posterior Mean*. The 95% credible intervals for the model parameters contain the population value with a probability of 0.95 (95%). For the difference between ratings for unattractive faces compared to attractive ones in the placebo group relative to the same difference in the high-dose group (i.e., the first interaction term), the relative difference has a 95% probability of being between about −5 and −1.7. For the difference between ratings for unattractive faces compared to attractive ones in the placebo group relative to the same difference in the low-dose group (i.e., the second interaction term), the relative difference has a 95% probability of being between about −2.9 and 0.4. These credible intervals assume that the effect exists, but in the latter case it suggests that the effect could plausibly be positive, negative and very small indeed.

Bayes Factor Model Summary[a,b]

Bayes Factor[c]	R	R Square	Adjusted R Square	Std. Error of the Estimate
1195.598	.718	.515	.458	1.17

a. Method: JZS

b. Model: (Intercept), Attractivenss of facial stimuli, Alcohol: high vs. placebo, Alcohol: low vs. placebo, Interaction: high vs placebo, att vs unatt, Interaction: low vs placebo, att vs unatt

c. Bayes factor: Testing model versus null model (Intercept).

Bayesian Estimates of Coefficients[a,b,c]

Parameter	Mode	Posterior Mean	Variance	95% Credible Interval Lower Bound	95% Credible Interval Upper Bound
(Intercept)	3.500	3.500	.180	2.665	4.335
Attractivenss of facial stimuli	2.875	2.875	.359	1.694	4.056
Alcohol: high vs. placebo	3.125	3.125	.359	1.944	4.306
Alcohol: low vs. placebo	1.375	1.375	.359	.194	2.556
Interaction: high vs placebo, att vs unatt	−3.375	−3.375	.719	−5.045	−1.705
Interaction: low vs placebo, att vs unatt	−1.250	−1.250	.719	−2.920	.420

a. Dependent Variable: Median attractiveness rating

b. Model: (Intercept), Attractivenss of facial stimuli, Alcohol: high vs. placebo, Alcohol: low vs. placebo, Interaction: high vs placebo, att vs unatt, Interaction: low vs placebo, att vs unatt

c. Assume standard reference priors.

Output 14.12

14.13 More effect sizes ©

As we have seen, SPSS produces partial eta squared, η^2 (Section 13.9). However, we've seen before that omega squared (ω^2) is less biased. The calculation of omega squared becomes somewhat more cumbersome in factorial designs ('somewhat' being an understatement). Howell (2012), as ever, does a wonderful job of explaining it all (and has a nice table summarizing the various components for a variety of situations). Condensing the information from his book, we first compute a variance component for

each of the effects (the two main effects and the interaction term) and the error, and then use these to calculate effect sizes for each. If we call the first main effect A, the second main effect B and the interaction effect $A \times B$, then the variance components for each of these are calculated from the mean squares of each effect and the sample sizes on which they're based:

$$\begin{aligned}\hat{\sigma}_{\alpha}^{2} &= \frac{(a-1)(\mathrm{MS}_{A}-\mathrm{MS}_{\mathrm{R}})}{nab}\\ \hat{\sigma}_{\beta}^{2} &= \frac{(b-1)(\mathrm{MS}_{B}-\mathrm{MS}_{\mathrm{R}})}{nab}\\ \hat{\sigma}_{\alpha\beta}^{2} &= \frac{(a-1)(b-1)(\mathrm{MS}_{A\times B}-\mathrm{MS}_{\mathrm{R}})}{nab}\end{aligned} \tag{14.25}$$

In these equations, a is the number of levels of the first independent variable, b is the number of levels of the second independent variable and n is the number of people per condition.

Let's calculate these for our data. Output 14.2 contains the mean squares for each effect, and for the error term. The first predictor, alcohol, had three levels (hence $a = 3$) and a mean square of 8.271. Type of face had two levels (hence $b = 2$) and a mean square of 21.333. The interaction had a mean square of 11.646. The number of people in each group was 8 and the residual mean square was 1.369. Therefore, our equations become:

$$\begin{aligned}\hat{\sigma}_{\alpha}^{2} &= \frac{(3-1)(8.271-1.369)}{8\times 3\times 2} = 0.288\\ \hat{\sigma}_{\beta}^{2} &= \frac{(2-1)(21.333-1.369)}{8\times 3\times 2} = 0.416\\ \hat{\sigma}_{\alpha\beta}^{2} &= \frac{(3-1)(2-1)(11.646-1.369)}{8\times 3\times 2} = 0.428\end{aligned} \tag{14.26}$$

We estimate the total variability by adding the estimates in Eq. 14.26 to the residual mean square:

$$\begin{aligned}\hat{\sigma}_{\text{total}}^{2} &= \hat{\sigma}_{\alpha}^{2} + \hat{\sigma}_{\beta}^{2} + \hat{\sigma}_{\alpha\beta}^{2} + \mathrm{MS}_{\mathrm{R}}\\ &= 0.288 + 0.416 + 0.428 + 1.369\\ &= 2.501\end{aligned} \tag{14.27}$$

The effect size is the variance estimate for the effect in which you're interested divided by the total variance estimate from Eq. 14.27:

$$\omega_{\text{effect}}^{2} = \frac{\hat{\sigma}_{\text{effect}}^{2}}{\hat{\sigma}_{\text{total}}^{2}} \tag{14.28}$$

For the main effect of alcohol we get:

$$\omega_{\text{alcohol}}^{2} = \frac{\hat{\sigma}_{\text{alcohol}}^{2}}{\hat{\sigma}_{\text{total}}^{2}} = \frac{0.288}{2.501} = 0.115 \tag{14.29}$$

For the main effect of type of face we get:

$$\omega_{\text{facetype}}^{2} = \frac{\hat{\sigma}_{\text{facetype}}^{2}}{\hat{\sigma}_{\text{total}}^{2}} = \frac{0.416}{2.501} = 0.166 \tag{14.30}$$

For the interaction of type of face and alcohol we get:

$$\omega_{\text{alcohol}\times\text{facetype}}^{2} = \frac{\hat{\sigma}_{\text{alcohol}\times\text{facetype}}^{2}}{\hat{\sigma}_{\text{total}}^{2}} = \frac{0.428}{2.501} = 0.171 \tag{14.31}$$

To make these values comparable to r we can take the square root, which gives us effect sizes of 0.34 for alcohol and 0.41 for both type of face and the interaction term.

14.14 Reporting the results of factorial designs Ⓑ

There is consistency across the results of our non-robust model, robust model and Bayesian model. To sum up, alcohol has an effect on ratings of the attractiveness of pictures of faces, but only when those pictures depict unattractive faces. When the faces are attractive alcohol has no significant effect on judgements of attraction. This pattern of results was evident from the interaction graph (Figure 14.16) so you'd want to include that in your write-up.

We report the details of the *F*-statistic and the degrees of freedom for each effect. For the effects of **alcohol** and the **alcohol** × **facetype** interaction, the model degrees of freedom were 2 ($df_M = 2$), but for the effect of **facetype** the degrees of freedom were 1 ($df_M = 1$). For all effects, the degrees of freedom for the residuals were 42 ($df_R = 42$). We can, therefore, report the three effects as follows:

- ✓ There was a significant main effect of the amount of alcohol consumed on ratings of the attractiveness of faces, $F(2, 42) = 6.04$, $p = 0.005$, $\omega^2 = 0.12$. Bonferroni *post hoc* tests revealed that the attractiveness ratings were significantly higher after a high dose than after a placebo drink ($p = 0.004$). The attractiveness ratings were not significantly different after a low dose compared to a placebo ($p = 0.231$) or a high dose compared to a low dose ($p = 0.312$).
- ✓ Attractive faces were rated significantly higher than unattractive faces, $F(1, 42) = 15.58$, $p < 0.001$, $\omega^2 = 0.17$.
- ✓ There was a significant interaction between the amount of alcohol consumed and the type of face of the person rated on attractiveness, $F(2, 42) = 8.51$, $p = 0.001$, $\omega^2 = 0.17$. This effect indicates that ratings of unattractive and attractive faces were affected differently by alcohol. Simple effects analysis revealed that ratings of attractive faces were significantly higher than of unattractive faces in the placebo group, $F(1, 42) = 24.15$, $p < 0.001$, and in the low-dose group, $F(1, 42) = 7.72$, $p = 0.008$, but not in the high-dose group, $F(1, 42) = 0.73$, $p = 0.398$.



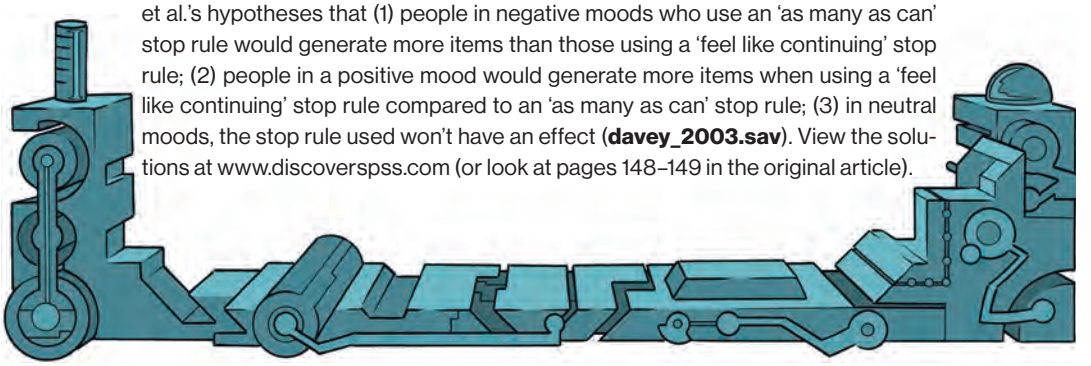
Labcoat Leni's Real Research 14.1
Don't forget your toothbrush Ⓑ

Davey, G. C. L. et al., (2003). *Journal of Behavior Therapy & Experimental Psychiatry, 34*, 141-160.

Many of us have experienced that feeling after we have left the house of wondering whether we remembered to lock the door, close the window, or remove the bodies from the fridge in case the police turn up. However, some people with obsessive compulsive disorder (OCD) check things so excessively that they might, for example, take hours to leave the house. One theory is that this checking behaviour is caused by the mood you are in (positive or negative) interacting with the rules you use to decide when to stop a task (do you continue until you feel like stopping, or until you have done the task as best as you can?) Davey et al. (2003) tested this hypothesis by getting people to think of as many things as they could that they should check before going on holiday (**checks**) after putting them into a negative, positive or neutral mood (**mood**). Within each mood group, half of the participants were instructed to generate as many items as they could,

whereas the remainder were asked to generate items for as long as they felt like continuing the task (**stop_rule**).

Plot an error bar chart and then conduct the appropriate analysis to test Davey et al.'s hypotheses that (1) people in negative moods who use an 'as many as can' stop rule would generate more items than those using a 'feel like continuing' stop rule; (2) people in a positive mood would generate more items when using a 'feel like continuing' stop rule compared to an 'as many as can' stop rule; (3) in neutral moods, the stop rule used won't have an effect (**davey_2003.sav**). View the solutions at www.discoverspss.com (or look at pages 148–149 in the original article).



14.15 Jane and Brian's story Ⓐ

After walking the streets alone with her thoughts, Jane arrived at Brian's apartment. It was going to seem weird that she was there: she barely knew his name, they'd never exchanged numbers, let alone addresses, but these things were easy enough to find out if you knew how to hack. She felt

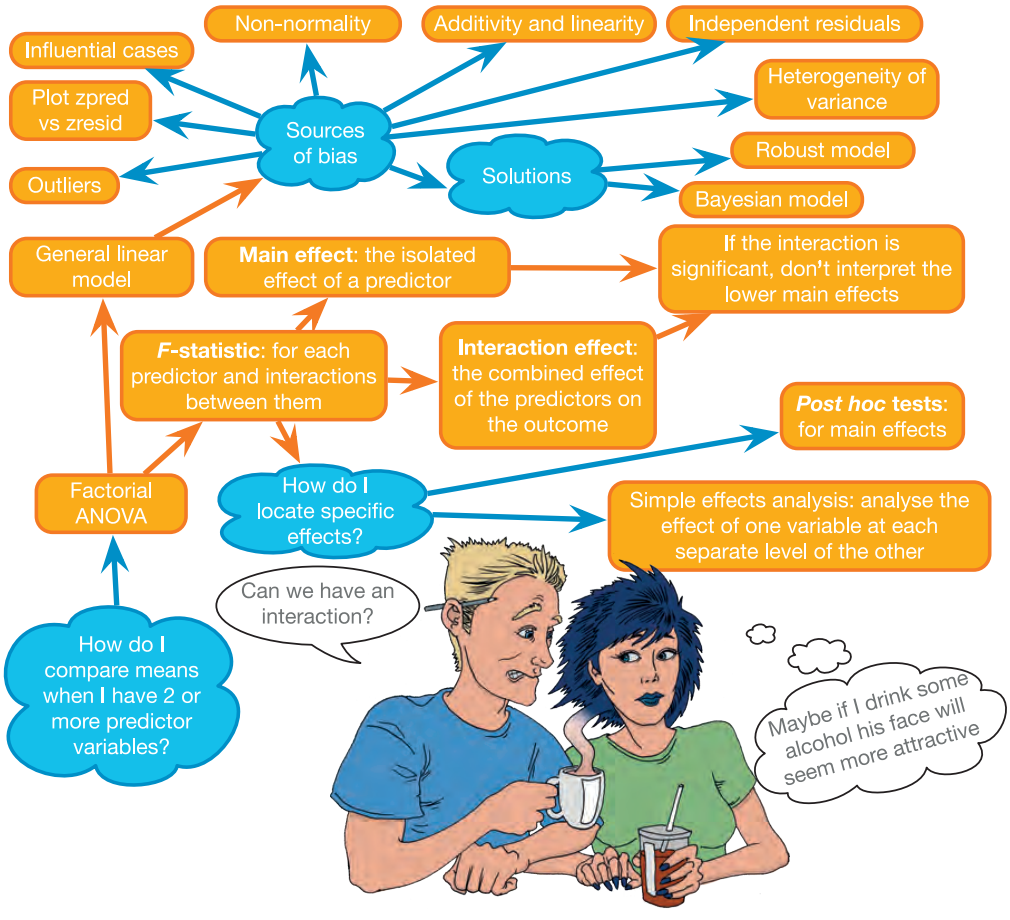


Figure 14.18 What Brian learnt from this chapter

anxious. At the third attempt she mustered the courage to press the buzzer. No answer. She buzzed again. And again.

'Hi!' said the puzzled voice through the intercom.

'It's Jane. I need to talk.'

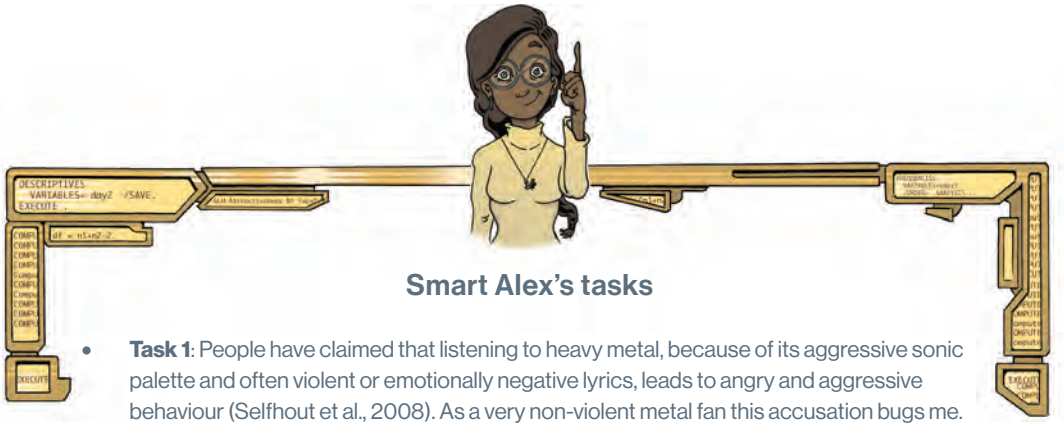
A minute later Brian opened the door, linked arms with her and walked. She had expected to be invited in, but this was cool too. They walked the neighbourhood in the moonlight. She needed to talk, but she didn't: their lack of intimacy muted her. Brian filled the silence, as he always did, with his latest statistical knowledge. When he ran out of things to impress her with he ended, as he usually did, with a question. She had no clever put-down this time, no enthusiasm for flirtatious banter and no one else to talk to. She told him about her mum, about her ambivalence towards her parents, about her upbringing, her brother. It was unfiltered and unguarded and made her feel vulnerable. Brian listened and suggested she go to her mother.

14.16 What next? Ⓑ

No sooner had I started my first band than it disintegrated. I went with drummer Mark to sing in a band called the Outlanders, who were much better musically but were not, if the truth be told, metal enough for me. They also sacked me after a very short period of time for not being able to sing like Bono (an insult at the time, but in retrospect …). So, that was two failed bands in very quick succession. You'd have thought that the message that perhaps singing wasn't the thing for me might have started to sink in. It hadn't and didn't for quite some time (still hasn't entirely). I needed a new plan, and one was hatched one evening while walking along a cliff top in Cornwall with drummer Mark…

14.17 Key terms that I've discovered

Beer-goggles effect
Factorial ANOVA
Independent factorial design
Interaction graph
Mixed design
Related factorial design
Simple effects analysis



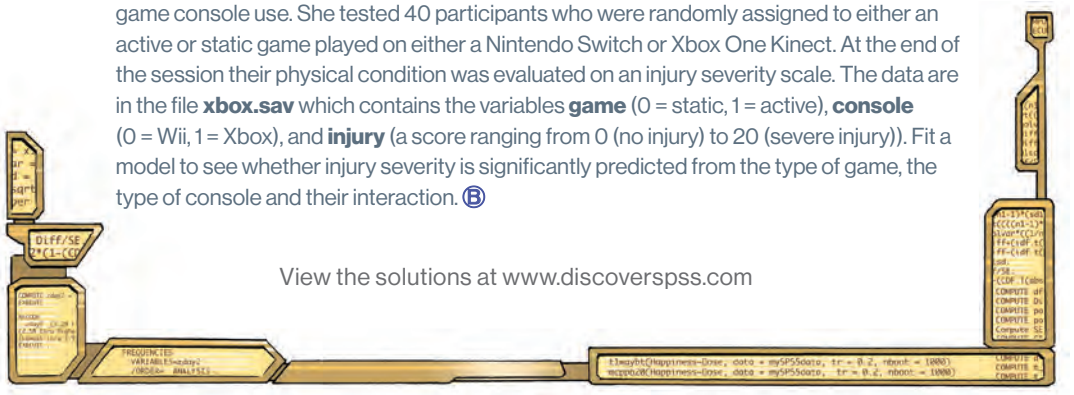
Smart Alex's tasks

- **Task 1**: People have claimed that listening to heavy metal, because of its aggressive sonic palette and often violent or emotionally negative lyrics, leads to angry and aggressive behaviour (Selfhout et al., 2008). As a very non-violent metal fan this accusation bugs me. Imagine I designed a study to test this possibility. I took groups of self-classifying metalheads and non-metalheads (**fan**) and assigned them randomly to listen to 15 minutes of either the sound of an angle grinder scraping a sheet of metal (control noise), metal music or pop music (**soundtrack**). Each person rated their **anger** on a scale ranging from

0 ('All you need is love, da, da, da-da-da') to 100 ('All you wanna do is drag me down, all I wanna do is stamp you out'). Fit a model to test my idea (**metal.sav**). Ⓑ

- **Task 2**: Compute omega squared for the effects in Task 1 and report the results of the analysis. Ⓒ
- **Task 3**: In Chapter 5 we used some data that related to male and female arousal levels when watching *The Notebook* or a documentary about notebooks (**notebook.sav**). Fit a model to test whether men and women differ in their reactions to different types of films. Ⓑ
- **Task 4**: Compute omega squared for the effects in Task 3 and report the results of the analysis. Ⓒ
- **Task 5**: In Chapter 4 we used some data that related to learning in men and women when either reinforcement or punishment was used in teaching (**teaching.sav**). Analyse these data to see whether men and women's learning differs according to the teaching method used. Ⓑ
- **Task 6**: At the start of this chapter I described a way of empirically researching whether I wrote better songs than my old bandmate Malcolm, and whether this depended on the type of song (a symphony or song about flies). The outcome variable was the number of screams elicited by audience members during the songs. Plot the data and fit a model to test my hypothesis that the type of song moderates which songwriter is preferred (**escape.sav**). Ⓑ
- **Task 7**: Compute omega squared for the effects in Task 6 and report the results of the analysis. Ⓒ
- **Task 8**: There are reports of increases in injuries related to playing games consoles. These injuries were attributed mainly to muscle and tendon strains. A researcher hypothesized that a stretching warm-up before playing would help lower injuries, and that athletes would be less susceptible to injuries because their regular activity makes them more flexible. She took 60 athletes and 60 non-athletes (**athlete**); half of them played on a Nintendo Switch and half watched others playing as a control (**switch**), and within these groups half did a 5-minute stretch routine before playing/watching whereas the other half did not (**stretch**). The outcome was a pain score out of 10 (where 0 is no pain and 10 is severe pain) after playing for 4 hours (**injury**). Fit a model to test whether athletes are less prone to injury, and whether the prevention programme worked (**switch.sav**). Ⓒ
- **Task 9**: A researcher was interested in what factors contributed to injuries resulting from game console use. She tested 40 participants who were randomly assigned to either an active or static game played on either a Nintendo Switch or Xbox One Kinect. At the end of the session their physical condition was evaluated on an injury severity scale. The data are in the file **xbox.sav** which contains the variables **game** (0 = static, 1 = active), **console** (0 = Wii, 1 = Xbox), and **injury** (a score ranging from 0 (no injury) to 20 (severe injury)). Fit a model to see whether injury severity is significantly predicted from the type of game, the type of console and their interaction. Ⓑ

View the solutions at www.discoverspss.com





GLM 4: REPEATED-MEASURES DESIGNS

15.1 What will this chapter tell me? 696
15.2 Introduction to repeated-measures designs 697
15.3 Emergency! The aliens are coming! 698
15.4 Repeated measures and the linear model 698
15.5 The ANOVA approach to repeated-measures designs 700
15.6 The *F*-statistic for repeated-measures designs 704
15.7 Assumptions in repeated-measures designs 709
15.8 One-way repeated-measures designs using SPSS 709
15.9 Output for one-way repeated-measures designs 714
15.10 Robust tests of one-way repeated-measures designs 721
15.11 Effect sizes for one-way repeated-measures designs 724
15.12 Reporting one-way repeated-measures designs 725
15.13 A scented factorial repeated-measures design 726
15.14 Factorial repeated-measures designs using SPSS 727
15.15 Interpreting factorial repeated-measures designs 733
15.16 Reporting the results from factorial repeated-measures designs 747
15.17 Jane and Brian's story 747
15.18 What next? 748
15.19 Key terms that I've discovered 748
Smart Alex's tasks 749

Letters A to D indicate increasing difficulty

15.1 What will this chapter tell me? Ⓑ

At the age of 15, I was on holiday with my friend Mark (the drummer) in Cornwall. I had a pretty decent mullet by this stage (nowadays I wish I had enough hair to grow a mullet) and a respectable collection of heavy metal T-shirts from going to various gigs. We were walking along the clifftops one evening at dusk reminiscing about Andromeda. We decided that the only thing we hadn't enjoyed about that band was Malcolm and that maybe we should re-form it with a different guitarist.[1] As I wondered who we could get to play guitar, Mark pointed out the blindingly obvious: I played guitar. So, when we got home Scansion was born.[2] As the singer, guitarist and songwriter I set about writing some songs. Lyrically speaking, flies were so 'last year', so I moved on to exploring existential angst (I have never really stopped). Obviously we failed totally in our musical ambitions otherwise I wouldn't be writing this book, but we did get reviewed in the music magazine *Kerrang!* They called us 'twee' in a live review, which is not really what you want to read as a metal band, but were slightly more flattering about one of our demos (only slightly though). Our career high point was playing at the famous Marquee Club in London (which, like most of the venues of my youth, has closed now, but not because of our playing there).[3] This was the biggest gig of our career, and we needed to play like we never had before. As it turned out, we did: I ran on stage and fell over, detuning my guitar beyond recognition and breaking the zip on my trousers in the process. I spent the whole gig out of tune and spread-eagle to prevent my trousers falling down. Like I said, I'd never played like *that* before. We used to get quite obsessed with comparing how we played at different gigs. I didn't know about statistics then (happy days), but if I had I would have realized that we could rate ourselves and compare the mean ratings for different gigs; because we would always be the ones rating the gigs, this would be a repeated-measures design. That's what this chapter is about. Hopefully it won't make our trousers fall down.



Figure 15.1 Scansion in the early days; I used to stare a lot (left to right: me, Mark Franklin and Mark Gatland)

1 Although this is what we thought at the time, in retrospect I feel bad because Malcolm was never anything other than a very nice, positive, enthusiastic guy. To be honest, at that age (and some would argue, beyond) I sometimes took myself too seriously.

2 Scansion is a term for the rhythm of poetry. We got the name by searching through a dictionary until we found a word that we liked. Originally we didn't think it was 'metal' enough, and we decided that any self-respecting heavy metal band needed to have a big spiky 'X' in their name. So, for the first couple of years we spelt it 'Scanxion'. I know.

3 In its day the Marquee Club started the careers of people like Jimi Hendrix, The Who, Iron Maiden and Led Zeppelin. Google it if you need to procrastinate.

15.2 Introduction to repeated-measures designs Ⓑ

So far in this book, when comparing means, we've concentrated on situations in which different entities contribute to different means. I have tended to focus on examples with different *people* taking part in different experimental conditions, but it could be different plants, companies, plots of land, viral strains, goats or even duck-billed platypuses (or whatever the plural is). I've completely ignored situations in which the same people (plants, goats, hamsters, seven-eyed green galactic leaders from space, or whatever) contribute to the different means. I've put it off long enough, and now I'm going to take you through the models we fit to repeated-measures data.

SELF TEST

What is a repeated-measures design? (Clue: it is described in Chapter 1.)



Figure 15.2 The confusion machine has created different ways to refer to repeated-measures designs

'Repeated measures' is a term used when the same entities participate in all conditions of an experiment or provide data at multiple time points. For example, you might test the effects of alcohol on enjoyment of a party. Some people can put away a lot of alcohol without really feeling the consequences, whereas others, like myself, have only to sniff a pint of lager and they start waving their arms and legs around shouting, 'Look at me, I'm Andy, King of the lost world of the Haddocks.' To control for these individual differences in tolerance to alcohol you can test the same people in all conditions of the experiment: participants could be given a questionnaire assessing their enjoyment of the party after they had consumed 1 pint, 2 pints, 3 pints and 4 pints of lager. There are lots of different ways to refer to this sort of design (Figure 15.2).

15.3 Emergency! The aliens are coming! Ⓐ

As you may know, or perhaps not, alien space lizards live among us. Mostly these extra-terrestrial guests live peacefully on our world carrying out useful functions in society such as warning of us of impending invasions and teaching statistics … I've said too much. It turns out that not all space lizards are the cute, stats-nerd types, and the threat of invasion from the evil ones is very real. When the alien invasion comes we'll need hyper-advanced technology to identify these crafty, shapeshifting ne'er-do-wells, and what could be more advanced than a spaniel's nose? Nothing, that's what. To this end, the top-secret government agency for Training Extra-terrestrial Reptile Detection (TERD) was assigned the task of conducting research to test the plausibility of training sniffer dogs to detect aliens. Eight of their best dogs (Milton, Woofy, Ramsey, Mr. Snifficus III, Willock, The Venerable Dr. Waggy, Lord Scenticle and Professor Nose) were recruited for a pilot study. During training, these dogs were rewarded for making vocalizations while sniffing alien space lizards (which they happened to have a few of at the local university Department of Statistics). After training, the dogs completed a test trial in which the eight dogs sniffed each of four entities, in random order, for 1 minute: an alien space lizard in its natural form, an alien space lizard who had shapeshifted into humanoid form, a human and a human mannequin. The number of vocalizations made during each 1-minute sniffing session was recorded. If the training has been successful, we'd expect the dogs to vocalize more when sniffing the aliens than when sniffing the control entities (the human and mannequin).

Table 15.1 shows the data: the eight rows show the different dogs, and the columns indicate the number of vocalizations they made during their 1-minute contact with each entity. Because they'll come in handy later, the mean number of vocalizations (and variance) for each dog is shown, as are the mean vocalizations for each entity that was sniffed. The overall variance in the number of vocalizations will, in part, be caused by the fact that the entities differ in their smell (that's the thing the experimenters have manipulated) and, in part, by individual differences in the dogs (e.g., by differences in their sensitivity to smell, their natural tendency to vocalize and their responsiveness to training).

Table 15.1 Data for the sniffer-dog example

Dog	Alien	Human	Mannequin	Shapeshifter	Mean	s^2
Milton	8	7	1	6	5.50	9.67
Woofy	9	5	2	5	5.25	8.25
Ramsey	6	2	3	8	4.75	7.58
Mr. Snifficus III	5	3	1	9	4.50	11.67
Willock	8	4	5	8	6.25	4.25
The Venerable Dr. Waggy	7	5	6	7	6.25	0.92
Lord Scenticle	10	2	7	2	5.25	15.58
Professor Nose	12	6	8	1	6.75	20.92
Mean	8.13	4.25	4.13	5.75		

15.4 Repeated measures and the linear model Ⓒ

Up until now we have considered everything as a variation of the general linear model. Repeated-measures designs can also be conceptualized in this way. First, imagine we wanted to conceptualize the experiment I've just described as a linear model in which different dogs sniffed different entities. To keep things simple, let's imagine there were just two entities sniffed: the alien and human. Thinking back to Chapters 10 and 12, we could write the model as:

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{vocalizations}_i &= b_0 + b_1 \text{entity}_i + \varepsilon_i \end{aligned} \tag{15.1}$$

This model doesn't account for the fact that the same people took part in all conditions. First off, in an independent design we have one observation for the outcome from each person (or in this case dog), so we predict the outcome for the individual (Y_i) based on the value of the predictor of that individual (X_i), but with repeated measures the individual has several values of the predictor so we predict the outcome from both the individual (i) and the specific value of the predictor that's of interest (g).[4] So we might write a simple model as:

$$\text{vocalizations}_{gi} = b_0 + b_1 \text{entity}_{gi} + \varepsilon_{gi} \tag{15.2}$$

which acknowledges that we predict the number of vocalizations (Y_{gi}) for entity g within dog i from the specific entity sniffed (entity_{gi}), and the error in prediction, ε_{gi}, to which both the dog (i) and the entity sniffed (g) contribute. All that's changed is the subscripts in the model, which acknowledge that levels of the treatment condition (g) occur within individuals (i).

Eq. 15.2 only really attaches some different subscripts to acknowledge the structure of the data. In terms of estimating the model, we would, at the very least, want to factor in the variability in individual dogs (because we know they'll differ in their smell sensitivity, training, etc.). We can do this by adding a variance term to the intercept. Remember that the intercept represents the number of vocalizations when the predictor is 0, so if we allow this parameter to vary across individuals, we're effectively modelling the possibility that different dogs will have different natural vocalization levels, different nose sensitivity, and different motivation levels for training. This is known as a random intercept model, and we'll look at these in detail in Chapter 21. This model is written as:

$$\begin{aligned} \text{vocalizations}_{gi} &= b_{0i} + b_1 X_{gi} + \varepsilon_{gi} \\ b_{0i} &= b_0 + u_{0i} \end{aligned} \tag{15.3}$$

All that's happened is the intercept has had an i added to the subscript to reflect that it is specific to an individual, and underneath we define the intercept as being made up of the group-level intercept (b_0), which is the intercept for all dogs, plus the deviation of the individual dog's intercept from the group-level intercept (u_{0i}). Put simply, u_{0i} reflects individual differences in overall vocalizations. The first line of Eq. 15.3 is the overall model, and the individual-level effects are incorporated into the second line.

We might also want to factor in the possibility that the experimental effect (the effect of different entities) varies across individuals. We do this by adding a variance term to the slope. Remember that the slope represents the effect that the different entities have on vocalizations. By allowing this parameter to vary across dogs, we model the possibility that the effect of the entity on vocalizations will be different in different dogs. This is known as a random slope model (again, see Chapter 21 for a fuller discussion). We write this model as:

$$\begin{aligned} \text{vocalizations}_{gi} &= b_{0i} + b_{1i} X_{gi} + \varepsilon_{gi} \\ b_{0i} &= b_0 + u_{0i} \\ b_{1i} &= b_1 + u_{1i} \end{aligned} \tag{15.4}$$

4 When using this model for repeated-measures designs people typically use subscripts i to represent different conditions and j to represent individuals, so my subscripts are a bit weird, but I think it's less confusing than referring to an earlier model (Eq. 15.1) in which i represents individuals but then have i represent something else in Eq. 15.2.

Compare it to Eq. 15.3: the main change is that the slope (b_1) has had an i added to the subscript to reflect that it is specific to an individual, and underneath we define it as being made up of the group-level slope (b_1) plus the deviation of the individual's slope from the group-level slope (u_{1i}). As before, u_{1i} reflects individual differences in the effect of entity on vocalizations. As in Eq. 15.3, the top of Eq. 15.4 is the overall model, and the individual-level effects are incorporated into the second and third lines.

We are getting ahead of ourselves a bit, but I want to give a flavour of how repeated measures can be incorporated into the general linear model. As we'll see in Chapter 21, these models get considerably more complex, but the take-home point is that we're again dealing with a linear model.

15.5 The ANOVA approach to repeated-measures designs Ⓑ

Although you can conceptualize repeated measures as a linear model, there are other ways too. The way that people typically handle repeated measures in IBM SPSS Statistics software is to use a **repeated-measures ANOVA** approach. Very crudely, this is a linear model like I have just described but with some very restrictive constraints. In a nutshell, we have seen that the standard linear model assumes that residuals are independent (not related to each other – see Section 6.7.3), but as you can see from the models above this assumption isn't true for repeated-measures designs: the residuals are affected by both between-participant factors (which should be independent) and within-participant factors (which won't be independent). There are (broadly) two solutions. One is to model this within-participant variability, which is what the models in the previous section do. The other is to apply additional assumptions that allow a simpler, more constrained model to be fitted. The later approach is the one that is historically popular.

15.5.1 The assumption of sphericity Ⓑ



The assumption that permits us to use a simpler model to analyse repeated-measures data is **sphericity**, which, trust me, is a pain to pronounce when you're giving statistics lectures at 9 a.m. on a Monday. Put crudely, sphericity is about assuming that the relationship between scores in pairs of treatment conditions is similar (i.e., the level of dependence between means is roughly equal).

The assumption of sphericity (denoted by ε and sometimes referred to as *circularity*) can be likened to the assumption of homogeneity of variance in between-group designs. It is a form of **compound symmetry**, which holds true when both the variances across conditions are equal (this is the same as the homogeneity of variance assumption in between-group designs) and the covariances between pairs of conditions are equal. So, we assume that the variation within conditions is similar and that no two conditions are any more dependent than any other two. Sphericity is a more general, less restrictive form of compound symmetry and refers to the equality of variances of the *differences* between treatment levels (Misconception Mutt 15.1).

The best way to explain what this means is by checking the assumption of sphericity by hand, which incidentally only a complete lunatic would do. Table 15.2 shows data from an experiment with three conditions and five people who have contributed scores to each condition. That's the first three columns. The second three show the differences between pairs of scores in all combinations of the treatment levels. I have calculated the variance of these differences in the bottom row. Sphericity is met when these variances are roughly equal. For these data, sphericity will hold when:

$$\text{variance}_{A-B} \approx \text{variance}_{A-C} \approx \text{variance}_{B-C} \quad (15.5)$$

Table 15.2 shows that there is some deviation from sphericity in these data because the variance of the differences between conditions *A* and *B* (15.7) is greater than the variance of the differences between both *A* and *C* (10.3) and *B* and *C* (10.7). However, these data have *local circularity* (or local sphericity) because two of the variances of differences are very similar, which means that sphericity can be assumed for any multiple comparisons involving these conditions (for a discussion of local circularity see Rouanet & Lépine, 1970). For the data in Table 15.2 the biggest difference in variances is between *A* and *B* (15.7) and *A* and *C* (10.3), but how do we know whether this difference is large enough to be a problem?

Table 15.2 Hypothetical data to illustrate the calculation of the variance of the differences between conditions

Condition A	Condition B	Condition C	*A* − *B*	*A* − *C*	*B* − *C*
10	12	8	−2	2	4
15	15	12	0	3	3
25	30	20	−5	5	10
35	30	28	5	7	2
30	27	20	3	10	7
		Variance:	15.7	10.3	10.7

15.5.2 Assessing the severity of departures from sphericity Ⓑ

Mauchly's test assesses the hypothesis that the variances of the differences between conditions are equal. If Mauchly's test statistic is significant (typically taken to mean that it has a *p*-value less than or equal to 0.05) it implies that there are significant differences between the variances of differences and, therefore, sphericity is not met. If Mauchly's test statistic is non-significant (typically taken to mean $p > 0.05$) then the implication is that the variances of differences are roughly equal and sphericity is met. However, like any significance test, Mauchly's test depends upon sample size and is probably best ignored (Misconception Mutt 15.1). Instead we can estimate the degree of sphericity using the **Greenhouse–Geisser estimate**, $\hat{\varepsilon}$ (Greenhouse & Geisser, 1959) or the **Huynh–Feldt estimate** (Huynh & Feldt, 1976), $\tilde{\varepsilon}$. The Greenhouse–Geisser estimate varies between $1/k - 1$ (where k is the number of repeated-measures conditions) and 1. For example, in a situation in which there are five conditions the lower limit of $\hat{\varepsilon}$ will be $1/(5 - 1)$, or 0.25 (known as the **lower-bound estimate** of sphericity). You can live a long, happy life oblivious to the precise calculation of these estimates so we won't go there (interested readers can consult Girden, 1992). We need know only that these estimates are used to correct for departures from sphericity and so are considerably more use than Mauchly's test (Misconception Mutt 15.1).

15.5.3 What's the effect of violating the assumption of sphericity? Ⓒ

Rouanet and Lépine (1970) provided a detailed account of the validity of the *F*-statistic under violations of sphericity (see also Mendoza et al., 1976). I summarized (Field, 1998) their findings in a very obscure newsletter that no one can ever access (see Oliver Twisted). The bottom line is that sphericity creates a loss of power and an *F*-statistic that doesn't have the distribution that it's supposed to have (i.e., an *F*-distribution). Lack of sphericity also causes some amusing complications for *post hoc* tests. If you don't want to worry about what these complications are then when sphericity is violated, the Bonferroni method is the most robust in terms of power and control of the Type I error rate. When sphericity is definitely not violated, Tukey's test can be used (Jane Superbrain Box 15.1).

Misconception Mutt 15.1
Sphericity and Mauchly's test

The Misconception Mutt was spinning around in circles. In the lecture he'd been taken to the lecturer kept rambling on about circularity, which had reminded the mutt how much fun it was to chase his tail. A ginger mist arose like a tornado from the epicentre of his spinning body, forming into a feline cloud above him. The cloud raised its eyebrows at him and muttered 'Stupid dogs!' The mutt stopped to throw the cat a look of disgust.

'What *are* you doing?' said the cat disapprovingly.

'I learnt about circularity today,' the dog wheezed, 'about how you need to look for sphericity when you compare two related means, and you use Mauchly's test, and if it's significant then sphericity can't be assumed but if it's non-significant then you can assume it ...'

The cat's paw reached towards the mutt's lips and gently closed them. The dog looked puzzled. 'Shhh...', said the cat, throwing him a patronizing look.

'Where to begin?' the cat whined. 'First, sphericity is not relevant if you're comparing only two means. The assumption is that the variances of difference scores between pairs of treatment levels are equal, and with only two conditions you have only one set of difference scores, and only one variance. Y*ou need at least three conditions for sphericity to be an issue.* Second, Mauchly's test is, essentially, pointless. In large samples a significant Mauchly's test could mean a small departure from sphericity that we have a lot of power to detect, and in small samples a non-significant Mauchly's could mean a large departure from sphericity that we didn't have the power to detect. Mauchly's test can't be interpreted without the context of the sample size. Also, if sphericity is violated then you can apply a correction proportionate to the extent of the violation (for example, the Greenhouse–Geisser correction) so you may as well ignore Mauchly's test and always apply the correction. When the data are spherical no correction is made, for small violations a small correction is made, and for large violations a large correction is made.'

The mutt started spinning furiously in the opposite direction, chasing his tail as though his life depended on it. As he span he created a reverse vortex that sucked the cat back into the ether. 'That'll learn him,' he thought to himself.



15.5.4 What do you do if you violate sphericity? Ⓑ

You might think that if your data violate the sphericity assumption then you need to book in to see a counsellor or something. You *can* do that, but a less costly option (emotionally and financially) is to adjust the degrees of freedom of any *F*-statistics affected. You can estimate sphericity in various ways (see above), resulting in a value that is 1 when your data are spherical and less than 1 when they are not.



Oliver Twisted
Please, Sir, can I have some more ... sphericity?

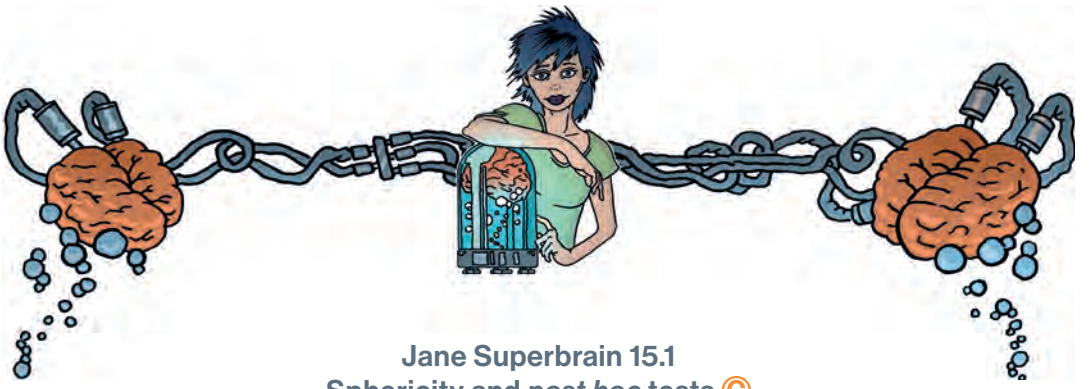
'Balls!' says Oliver. 'Balls are spherical, and I like them. Maybe I'd like sphericity too if only you could explain it to me in more detail.' Be careful what you wish for, Oliver. In my youth I wrote an article called 'A bluffer's guide to sphericity'. Occasionally people ask me for it, so I thought I might as well reproduce it at www.discoverspss.com.



You multiply the degrees of freedom for an affected F by this estimate. The result is that when you have sphericity the degrees of freedom don't change (because you multiply them by 1) but when you don't the degrees of freedom get smaller (because you multiply them by a value less than 1). The greater the violation of sphericity, the smaller the estimate gets, and the smaller the degrees of freedom become. Smaller degrees of freedom make the p-value associated with the F-statistic less significant. By adjusting the degrees of freedom by the extent to which the data are not spherical, we also make the F-statistic more conservative by the extent to which the data are not spherical. In doing so, the Type I error rate is controlled.

As I mentioned above, the degrees of freedom are adjusted using either the Greenhouse–Geisser or Huynh–Feldt estimates of sphericity. When the Greenhouse–Geisser estimate is greater than 0.75 the correction is too conservative (Huynh & Feldt, 1976) and this can also be true when the sphericity estimate is as high as 0.90 (Collier et al., 1967). However, the Huynh–Feldt estimate tends to overestimate sphericity (Maxwell & Delaney, 1990). Many authors recommend that when estimates of sphericity are greater than 0.75 the Huynh–Feldt estimate should be used, but when the Greenhouse–Geisser estimate of sphericity is less than 0.75 or nothing is known about sphericity the Greenhouse–Geisser correction should be used (Barcikowski & Robey, 1984; Girden, 1992; Huynh & Feldt, 1976). You can also take an average of the two estimates and adjust the df by this average (Stevens, 2002). We will see how these values are used in due course.

Another option when you have data that violate sphericity is to fit the kind of model described in Section 15.4, which is known as a multilevel model (see Chapter 21 for more detail). A third option is to use multivariate test statistics (MANOVA), because they do not assume sphericity (see O'Brien & Kaiser, 1985). SPSS produces multivariate test statistics (see also Chapter 17). However, there may be trade-offs in power between these univariate and multivariate tests (see Jane Superbrain Box 15.2).



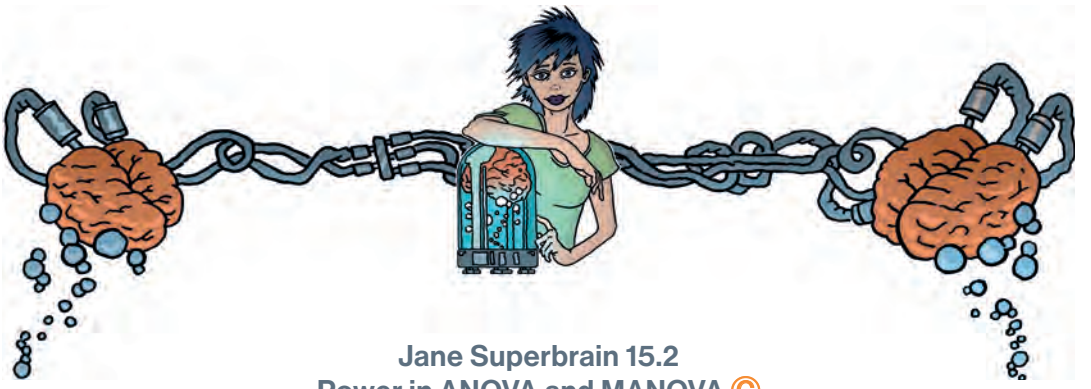
Jane Superbrain 15.1
Sphericity and *post hoc* tests ©

The violation of sphericity has implications for multiple comparisons. There is a more detailed summary of these online (see Oliver Twisted), but here are a few take-home messages. Boik (1981) recommends against using the *F*-statistic for repeated-measures contrasts because even very small departures from sphericity produce large biases. Maxwell (1980) compared the power and alpha levels for five *post hoc* tests: three variants of the Tukey procedure, the Bonferroni method and a multivariate approach (the Roy–Bose simultaneous confidence interval). The multivariate approach was always 'too conservative for practical use' (p. 277) and this was most extreme when n (the number of participants) is small relative to k (the number of conditions). All variants of Tukey's test inflated the Type I error rate unacceptably with increasing departures from sphericity. The Bonferroni method was extremely robust (although *slightly* conservative) and controlled Type I error rates well. In terms of power (or the Type II error rate), Tukey's honest significant difference was the most powerful under non-sphericity in tiny samples ($n = 8$), but this advantage was severely reduced in even slightly larger samples ($n = 15$). Keselman and Keselman (1988) extended Maxwell's work to unbalanced designs and concluded that Bonferroni was more powerful than the multivariate approach as the number of repeated treatment levels increases. A simple take-home message is that the Bonferroni method has much to recommend it.



15.6 The *F*-statistic for repeated-measures designs Ⓑ

In a repeated-measures design the effect of the experiment (the independent variable) is shown up in the within-participant variance (rather than in the between-group variance). Remember that in independent designs (Section 12.3.6) the within-participant variance is the residual sum of squares (SS_R); it is the variance created by individual differences in performance. When we carry out our experimental manipulation on the same entities, the within-participant variance will be made up of not just individual differences in performance but also the effect of our manipulation. Therefore, the main difference with a repeated-measures design is that we look for the experimental effect (the model sum of squares) within the individual rather than within the group. Figure 15.3 illustrates how the variance is partitioned in a repeated-measures design. The important thing to note is that the types of variances are the same as in independent designs: we have a total sum of squares (SS_T), a model sum of squares



Jane Superbrain 15.2
Power in ANOVA and MANOVA ©

There is a trade-off in test power between univariate and multivariate approaches. Univariate techniques are relatively powerless to detect small reliable changes between highly correlated conditions when other less correlated conditions are also present (Davidson, 1972). As the degree of violation of compound symmetry increases the power of multivariate tests also increases, whereas for univariate tests it decreases (Mendoza et al., 1974). However, multivariate approaches should probably not be used if n is less than $k + 10$, where k is the number repeated-measures levels (Maxwell & Delaney, 1990). To sum up, if you have a large violation of sphericity ($\varepsilon < 0.7$) and your sample size is greater than $k + 10$ then multivariate procedures are more powerful, but with small sample sizes or when sphericity holds ($\varepsilon > 0.7$) use the univariate approach (Stevens, 2002). It is also worth noting that the power of MANOVA varies as a function of the correlations between dependent variables (see Jane Superbrain Box 17.1) and so the relationship between treatment conditions must be considered.



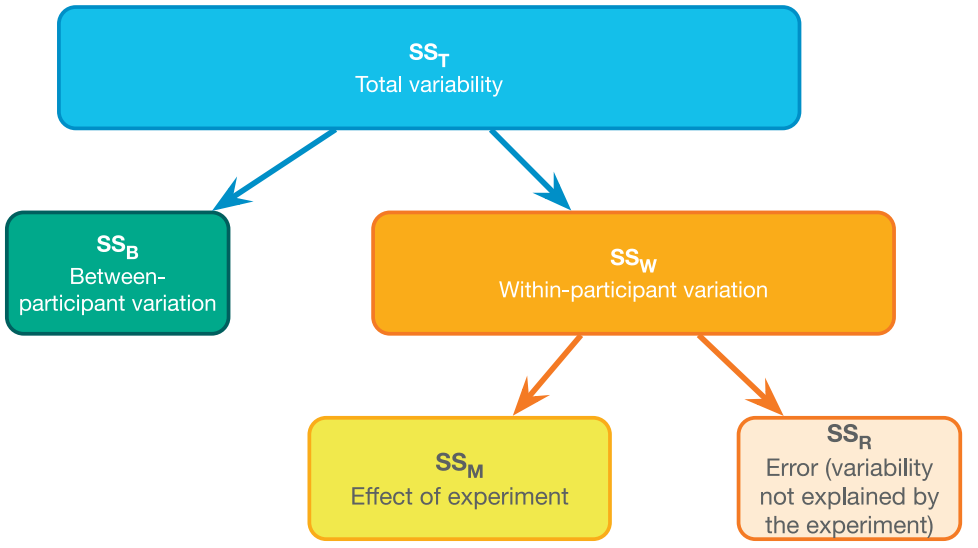


Figure 15.3 Partitioning variance for repeated-measures designs

(SS_M) and a residual sum of squares (SS_R). The *only* difference is from where those sums of squares come: in repeated-measures designs the model and residual sums of squares are both part of the within-participant variance. Let's have a look at an example.

15.6.1 The total sum of squares, SS_T (B)

Remember that for one-way independent designs SS_T is calculated as:

$$SS_T = s^2_{grand}(N-1) \tag{15.6}$$

In repeated-measures designs the total sum of squares is calculated in the same way. The grand variance in the equation is the variance of all scores when we ignore the group to which they belong. So, if we treated the data as one big group (as in Figure 15.4), the variance of these scores is 8.19 (try this on your calculator), and this is the grand variance. We used 32 scores to generate this value, so N is 32. Our sum of squares becomes:

$$\begin{aligned} SS_T &= s^2_{grand}(N-1) \\ &= 8.19(32-1) \\ &= 253.89 \end{aligned} \tag{15.7}$$

The degrees of freedom for this sum of squares, as for an independent design, is $N-1$, or 31.

8	7	1	6
9	5	2	5
6	2	3	8
5	3	1	9
8	4	5	8
7	5	6	7
10	2	7	2
12	6	8	1

Grand mean = 5.56

Grand variance = 8.19

Figure 15.4 Treating the data as a single group

15.6.2 The within-participant sum of squares, SS_W (B)

The crucial difference from an independent design is that in a repeated-measures design there is a within-participant variance component, which represents individual differences within participants. In independent designs these individual differences were quantified with the residual sum of squares (SS_R) using the following equation:

$$\begin{aligned} SS_R &= \sum_{g=1}^{k}\sum_{i=1}^{n}\left(x_{ig} - \bar{x}_g\right)^2 \\ &= \sum_{g=1}^{k} s_g^2(n_g - 1) \end{aligned} \tag{15.8}$$

In independent designs, because there are different participants within each condition, we calculated SS_R within each condition and added these values to get a total:

$$SS_R = s^2_{\text{group }1}(n_1 - 1) + s^2_{\text{group }2}(n_2 - 1) + s^2_{\text{group }3}(n_3 - 1) + \ldots + s^2_{\text{group }n}(n_n - 1) \tag{15.9}$$

In a repeated-measures design, because we've subjected participants to more than one experimental condition, we're interested in the variation not within a condition but *within an individual*. Therefore, we use the same equation but adapt it to look within individuals rather than groups. If we call this sum of squares SS_W (for within-participant SS), we adapt Eq. 15.9 to give:

$$\begin{aligned} SS_W = {} & s^2_{\text{individual }1}(n_1 - 1) + s^2_{\text{individual }2}(n_2 - 1) + s^2_{\text{individual }3}(n_3 - 1) + \ldots \\ & + s^2_{\text{individual }n}(n_n - 1) \end{aligned} \tag{15.10}$$

This equation translates as looking at the variation in each individual's scores and then adding these variances for all the individuals in the study. The *n*s represent the number of observations within the person (or, in our example, dog); that is, the number of experimental conditions, or in this case the number of entities sniffed. All the variances we need are in Table 15.1, so we can calculate SS_W as:

$$\begin{aligned} SS_W &= s^2_{\text{dog }1}(n_1 - 1) + s^2_{\text{dog }2}(n_2 - 1) + \ldots + s^2_{\text{dog }n}(n_n - 1) \\ &= 9.67(4 - 1) + 8.25(4 - 1) + 7.58(4 - 1) + 11.67(4 - 1) + 4.25(4 - 1) \\ &\quad + 0.92(4 - 1) + 15.58(4 - 1) + 20.92(4 - 1) \\ &= 29 + 24.75 + 22.75 + 35 + 12.75 + 2.75 + 46.75 + 62.75 \\ &= 236.50 \end{aligned} \tag{15.11}$$

The degrees of freedom for each individual dog are $n - 1$ (i.e., the number of conditions minus 1). To get the total degrees of freedom we add the *df*s for all participants. With eight participants (dogs) and four conditions (i.e., $n = 4$), there are 3 degrees of freedom for each dog and $8 \times 3 = 24$ degrees of freedom in total.

15.6.3 The model sum of squares, SS_M (B)

So far, we know that the total amount of variation within the vocalizations is 253.58 units. We also know that 236.50 of those units are explained by the variance created by individuals' (dogs) performances under different conditions. Some of this variation is the result of our experimental manipulation and some of this variation is due to unmeasured factors. The next step is to work out how much variance is explained by our manipulation (sniffing different entities) and how much is not.

In independent designs, we worked out how much variation could be explained by our experiment (the model SS) by looking at the means for each group and comparing these to the overall mean. We measured the variance resulting from the differences between group means and the overall mean (see Eq. 12.13). We do the same thing in a repeated-measures design:

$$SS_M = \sum_{g=1}^{k} n_g \left(\bar{x}_g - \bar{x}_{\text{grand}}\right)^2 \tag{15.12}$$

Using the means from Table 15.1, we get:

$$\begin{aligned} SS_M &= 8(8.13 - 5.56)^2 + 8(4.25 - 5.56)^2 + 8(4.13 - 5.56)^2 + 8(5.75 - 5.56)^2 \\ &= 8(2.57)^2 + 8(-1.31)^2 + 8(-1.44)^2 + 8(0.196)^2 \\ &= 83.13 \end{aligned} \tag{15.13}$$

With independent designs the model degrees of freedom are the number of conditions (k) minus 1. The same is true here: there were four conditions (entities), so the degrees of freedom will be 3.

15.6.4 The residual sum of squares, SS_R Ⓑ

We now know that there are 253.58 units of variation to be explained in our data, and that the variation across our conditions accounts for 236.50 units. Of these 236.50 units, our experimental manipulation can explain 83.13 units. The final sum of squares is the residual sum of squares (SS_R), which tells us how much of the variation cannot be explained by the model. This value is the amount of variation caused by extraneous factors outside experimental control. Knowing SS_W and SS_M already, the simplest way to calculate SS_R is to subtract SS_M from SS_W:

$$\begin{aligned} SS_R &= SS_W - SS_M \\ &= 236.50 - 83.13 \\ &= 153.37 \end{aligned} \tag{15.14}$$

The degrees of freedom are calculated in a similar way:

$$\begin{aligned} df_R &= df_W - df_M \\ &= 24 - 3 \\ &= 21 \end{aligned} \tag{15.15}$$

15.6.5 The mean squares Ⓑ

SS_M tells us how much variation the model (e.g., the experimental manipulation) explains and SS_R tells us how much variation is due to extraneous factors. Both these values are totals and depend on how many scores have contributed to them, so to make them comparable we convert them to the mean (or average) sum of squares (MS) by divided by the degrees of freedom:

$$\begin{aligned} MS_M &= \frac{SS_M}{df_M} = \frac{83.13}{3} = 27.71 \\ MS_R &= \frac{SS_R}{df_R} = \frac{153.37}{21} = 7.30 \end{aligned} \tag{15.16}$$

MS_M represents the average variation explained by the model (e.g., the average systematic variation), whereas MS_R is a gauge of the average variation explained by extraneous variables (the average unsystematic variation).

15.6.6 The *F*-statistic Ⓑ

The *F*-statistic is the ratio of the variation explained by the model and the variation explained by unsystematic factors. As for independent designs, it is the model mean sum of squares divided by the residual mean sum of squares:

$$F = \frac{MS_M}{MS_R} \tag{15.17}$$

So, as with independent designs, the *F*-statistic is the ratio of systematic to unsystematic variation: it is the experimental effect on performance relative to the effect of unmeasured factors.

For the sniffer-dog data, the *F*-statistic is 3.79:

$$F = \frac{MS_M}{MS_R} = \frac{27.71}{7.30} = 3.79 \tag{15.18}$$

This value is greater than 1, which indicates that the experimental manipulation had some effect above and beyond the effect of unmeasured factors. This value can be compared against a critical value

based on its degrees of freedom (df_M and df_R, which are 3 and 21 in this case), but more generally, it's possible to compute an exact *p*-value that is the probability of getting an *F* at least as big as the one we have observed if the null hypothesis is true.

15.6.7 The between-participant sum of squares Ⓑ

We've sort of forgotten about the between-participant variation in Figure 15.3 because we didn't need it to calculate the *F*-statistic. I will briefly mention what it represents. The easiest way to calculate this term is by subtraction:

$$SS_T = SS_B + SS_W$$
$$SS_B = SS_T - SS_W \quad (15.19)$$

We have already calculated SS_W and SS_T, so by rearranging the equation and substituting the values of these terms, we get:

$$SS_B = 253.89 - 236.50 = 17.39 \quad (15.20)$$

SS_B represents individual differences between cases. In this example different dogs will have different tendencies to vocalize. This variation is illustrated by the different means for the dogs in Table 15.1. For example, Professor Nose ($M = 6.75$), on average, tended to vocalize twice more than Mr. Snifficus III ($M = 4.50$). The between-participant sum of squares reflects these differences between individuals. In this case only 17.39 units of variation in the number of vocalizations is down to individual differences between dogs.

15.7 Assumptions in repeated-measures designs Ⓒ

Repeated-measures designs still use the linear model. so all of the sources of potential bias (and counteractive measures) discussed in Chapter 6 apply. Using the ANOVA approach, the assumption of independence is replaced by assumptions about the relationships between differences scores (sphericity). In the multilevel approach, sphericity isn't required, and you have a lot more flexibility to model different types of assumptions about residuals (see the section on covariance structures in Chapter 21).

If assumptions are not met, there is a robust variant of a one-way repeated-measures ANOVA that we'll cover in Section 15.10 (and you have the option of Friedman's ANOVA (see Chapter 6) although I'd use that as a last resort). The Bootstrap... button is noticeable by its absence in repeated-measures designs, which is a sad story. If you have a factorial design using repeated measures then you're stuffed.

15.8 One-way repeated-measures designs using SPSS Ⓑ

The general procedure for a one-way repeated-measures design (using ANOVA) is much the same as any other linear model, so, remind yourself of the general procedure in Chapter 9. Figure 15.5 shows a simpler overview that highlights some of the specific issues when using repeated measures.

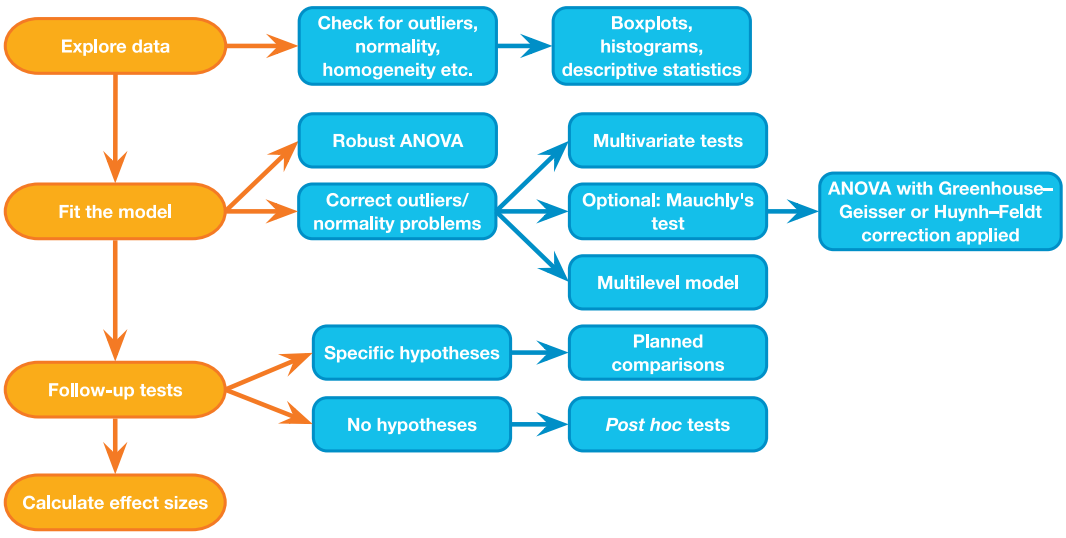


Figure 15.5 The process for analysing repeated-measures designs

15.8.1 The main analysis Ⓑ

The data for the example (**sniffer_dogs.sav**) can be entered into the data editor in the same format as Table 15.1 (although you don't need to include the columns labelled *Mean* or s^2). If you're entering the data by hand, create a variable called **alien** and use the labels dialog box to give this variable a full title of 'Alien (vocalizations in 1 min)'. In the next column, create a variable called **human**, with a full title of 'Human (vocalizations in 1 min)' and so on for variables called **mannequin** ('Mannequin (vocalizations in 1 min)') and **shapeshifter** ('Shapeshifter (vocalizations in 1 min)').

To compare means from a one-way repeated-measures design select *Analyze* > *General Linear Model* > Repeated Measures... . We use the *Define Factor(s)* dialog box to name our within-subject (repeated-measures) variable, which in this example is the type of entity sniffed, so replace the word *factor1* with the word *Entity* (Figure 15.6).[5] Next, we specify how many levels the variable had (i.e., how many conditions participants took part in). There were four different entities sniffed by each dog, so type '4' into the box labelled *Number of Levels*. Click Add to register this variable in the list of repeated-measures variables, where it appears as *Entity(4)* as in Figure 15.6. If your design has several repeated-measures variables you can add more factors to the list (see the next example). When you have finished creating repeated-measures factors, click Define to go to the main dialog box.

The main dialog box (Figure 15.7, left) has a space labelled *Within-Subjects Variables* that contains a list of four question marks followed by a number. These question marks are placeholders for the variables representing the four levels of the independent variable and need to be replaced with the variables corresponding to each level. For reasons that will become clear in the next section, we want to define the levels of the variable **Entity** in this order: mannequin, human, shapeshifter and alien. To do this, drag **mannequin** to the box labelled *Within-Subjects Variables* (or click), then drag **human** to the same box, and then **shapeshifter** and **alien.** The finished dialog box is shown in Figure 15.7 (right).

5 The name cannot have spaces in it.

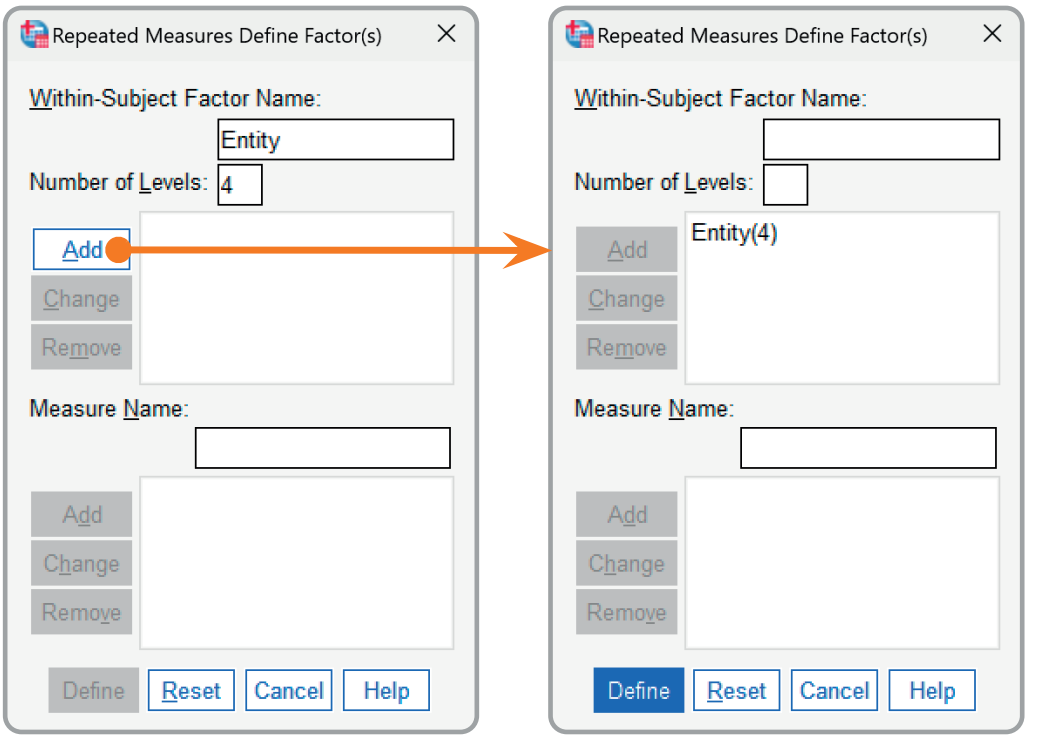


Figure 15.6 The *Define Factor(s)* dialog box for repeated-measures ANOVA

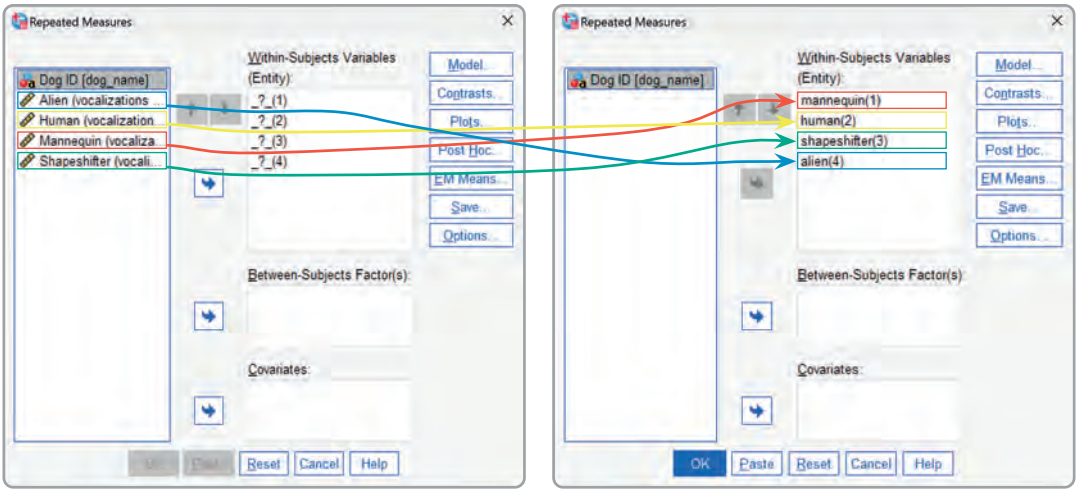


Figure 15.7 The main dialog box for repeated-measures designs (before and after completion)

15.8.2 Defining contrasts for repeated measures Ⓑ

Click Contrasts... to access the dialog box in Figure 15.8, which is used to specify one of the many standard contrasts that we have come across before (see Section 12.5.4). The default is a polynomial contrast, but you can change it by selecting a variable in the box labelled *Factors*, clicking Polynomial, selecting a contrast from the list and clicking Change. In this study you could argue that the mannequin is a control stimulus (because it's not a living organism), so it would make sense to perform a

simple contrast that compares every entity against the mannequin. If you choose a simple contrast then you can specify whether to compare each category against the first or last category. The first category would be the one entered as (1) in the main dialog box and, for these data, the last category would be the one entered as (4). For example, if we complete the main dialog box as in Figure 15.7 (right) then a simple contrast with ◉ First selected will compare each entity to the mannequin, but with ◉ Last selected each entity will be compared to the alien. Therefore, the order in which you enter variables in the main dialog box is important for the contrasts you choose.

Although a simple contrast would make sense, we'll use the *repeated* contrast, because we haven't looked at one before. A repeated contrast compares each category against the previous one, which can be useful in repeated-measures designs in which the levels of the independent variable have a meaningful order – for example, if you've measured the outcome variable at successive points in time, or administered increasing doses of a drug. In this example, there is a vaguely meaningful order to the entities that were sniffed in that we'd expect dogs trained to vocalize to aliens to respond least to a mannequin (because it's non-living), respond a bit more to humans (living but not alien), more still to shapeshifters (alien but their human appearance might confuse the dogs) and most to an alien in its natural form. By ordering the levels of **Entity** as mannequin, human, shapeshifter and alien in Figure 15.7 (right) a repeated contrast will compare human to mannequin, shapeshifter to human and alien to shapeshifter. So we see if each entity is associated with more vocalizations than the previous entity in the ordered list. When you have selected this contrast, click Continue.

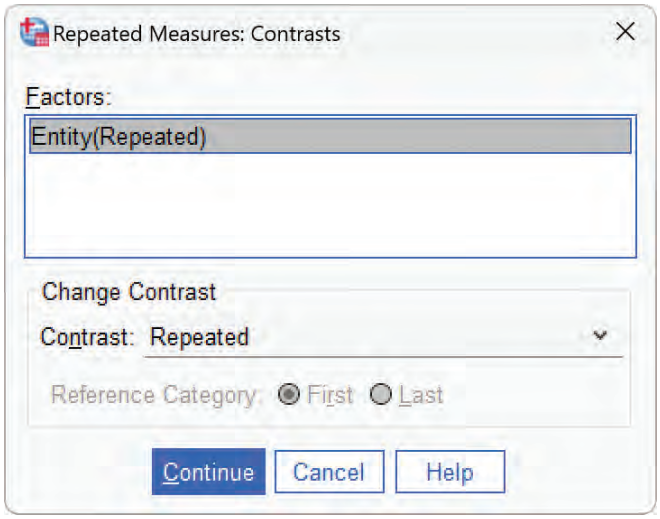


Figure 15.8 Repeated-measures contrasts

15.8.3 Custom contrasts ©

In Chapter 12 we looked at planned contrasts. Two of our entities are alien (the shapeshifter and alien) whereas the others are not (the human and mannequin). If the dog training has been successful, we would expect our dogs to vocalize more when sniffing an alien than when sniffing non-alien entities. We could test this hypothesis with a set of planned contrasts. Contrast 1 would compare vocalizations to alien entities against those to non-alien entities, contrast 2 would then compare the two alien entities (and ignore the human and mannequin) and contrast 3 would compare the non-alien entities (human and mannequin) while ignoring the alien and shapeshifter. The partitioning of variance is like that depicted in Figure 12.8.



The resulting contrast codes are in Table 15.3. If you follow the 'rules' in Section 12.5 you'll end up with codes of 2, 2, –2, –2 for contrast 1. I have divided these values by the number of groups (4) to get codes of 0.5, 0.5, –0.5, –0.5. I've done this so that the value for the contrast is the actual difference between means of the live and dead entities (rather than a multiple of it). Noting the group means in Table 15.1, my first contrast will yield a value of 2.75:

$$\frac{\bar{X}_{\text{Alien}} + \bar{X}_{\text{Shapeshifter}}}{2} - \frac{\bar{X}_{\text{Human}} + \bar{X}_{\text{Mannequin}}}{2} = \frac{8.13 + 5.75}{2} - \frac{4.25 + 4.13}{2} = 2.75 \quad (15.21)$$

whereas using codes of 2 and –2 will yield a value 4 times this value (4 × 2.75 = 11). This choice of coding doesn't affect the significance value at all, so in this sense you can use what codes you like, but it is handy to have the contrast value equal the actual difference between means because we can more easily interpret the confidence interval for the contrast (it will be the confidence interval for the difference between means, rather than the confidence interval for four times the difference).

Table 15.3 Contrast codes for the sniffer-dog data

	Alien	Shapeshifter	Human	Mannequin
Contrast 1 (alien vs. non-alien)	0.5	0.5	−0.5	−0.5
Contrast 2 (alien vs. shapeshifter)	1	−1	0	0
Contrast 3 (human vs. mannequin)	0	0	1	−1

To operationalize these contrasts you have to use the following syntax (**sniffer_contrasts.spwb**):

```
GLM alien human mannequin shapeshifter
  /WSFACTOR=Entity 4 Polynomial
  /WSDESIGN=Entity
  /MMATRIX =
    'Alien vs. non-alien' alien 0.5 shapeshifter 0.5 human -0.5 mannequin -0.5;
    'Alien vs. shapeshifter' alien 1 shapeshifter -1 human 0 mannequin 0;
    'Human vs. mannequin' alien 0 shapeshifter 0 human 1 mannequin -1.
```

The first three lines specify a basic repeated-measures model. The contrasts are defined within the /MMATRIX subcommand. I specify each contrast on a separate line (not essential, but it makes the syntax easier to read). I start with a name for the contrast in straight quotes, for example, 'Alien vs. non-alien' names the first contrast *Alien vs. non-alien*, which describes what the contrast tests. Then I list the variables that make up the levels of the predictor variable using the variable names from the data editor (*alien*, *shapeshifter*, *human* and *mannequin*) and after each one type its contrast code. Note that each contrast ends with a semicolon (which tells SPSS that the contrast specification is finished) except the last one, which ends with a period to tell SPSS that the entire *GLM* command is complete. Executing this syntax produces a table that is explained in Section 15.9.4.

15.8.4 *Post hoc* tests and additional options Ⓒ

Lack of sphericity creates entertaining complications for *post hoc* tests (see Jane Superbrain Box 15.1). When sphericity is definitely not violated, Tukey's test can be used, but if sphericity can't be assumed then the Games–Howell procedure is preferable. Because of these sphericity-related complications the standard *post hoc* tests for independent designs are not available for repeated-measures variables (the *post hoc* test dialog box does not list repeated-measured factors). The good news is that you can do some *post hoc* procedures by clicking EM Means... . To specify *post hoc* tests, drag the repeated-measures variable (in this case **Entity**) from the box labelled *Factor(s) and Factor Interactions* to the box labelled *Display Means for* (or click ➔). Then select ☑ Compare main effects to activate the LSD(none) drop-down menu (Figure 15.9). The default is to have no adjustment for multiple tests (Tukey LSD), which I don't recommend, but you can choose a Bonferroni or Šidák correction (recommended for the reasons already mentioned). I've selected Bonferroni. Click Continue .

Click Options... if you'd like to see things like descriptive statistics, parameter estimates and effect sizes. You can also change the level of significance for *post hoc* tests if you want to do a manual correction for multiple tests. Click Continue to return to the main dialog box, and OK to run the analysis.

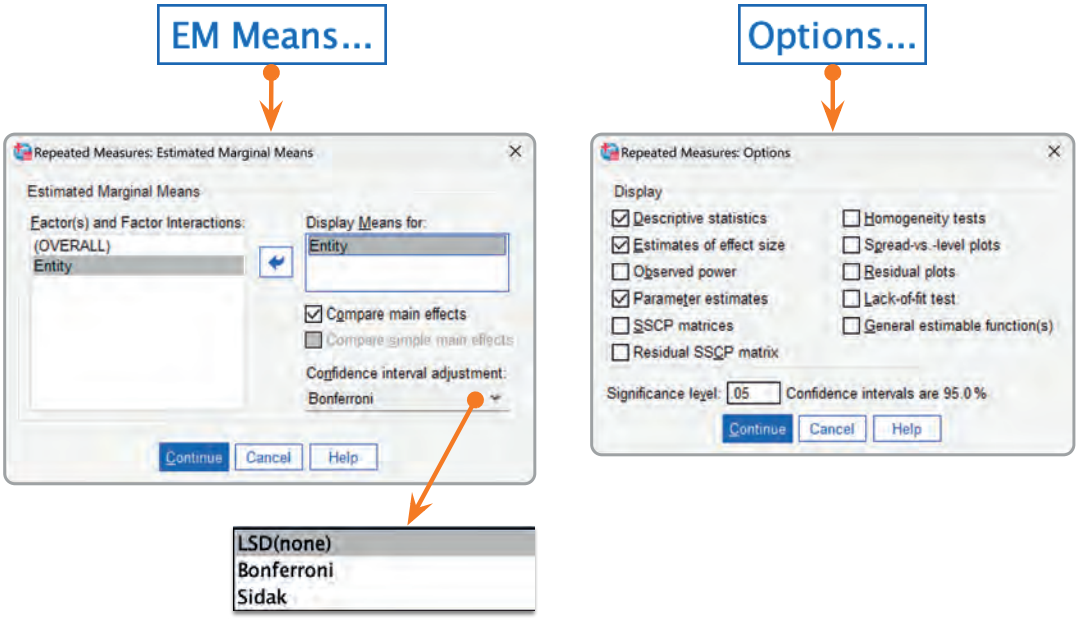


Figure 15.9 The *Options* dialog box

15.9 Output for one-way repeated-measures designs Ⓑ

15.9.1 Descriptive statistics Ⓐ

Output 15.1 shows two tables: the first (left) tells us the variables that represent each level of the predictor variable, which is a useful check that we entered the variables in the order that we intended to. The second table (right) shows that, on average, the number of vocalizations was largest after sniffing the alien and smallest after sniffing a mannequin. These means are useful for interpreting the main analysis.

Within-Subjects Factors

Measure: MEASURE_1

Entity	Dependent Variable
1	mannequin
2	human
3	shapeshifter
4	alien

Descriptive Statistics

	Mean	Std. Deviation	N
Mannequin (vocalisations in 1 min)	4.13	2.748	8
Human (vocalisations in 1 min)	4.25	1.832	8
Shapeshifter (vocalisations in 1 min)	5.75	2.915	8
Alien (vocalisations in 1 min)	8.13	2.232	8

Output 15.1

15.9.2 Correcting for sphericity Ⓑ

Output 15.2 shows Mauchly's test. The significance value (0.047) is less than the critical value of 0.05, which implies that the assumption of sphericity has been violated, but I suggested you ignore this test and routinely apply a correction for whatever deviation from sphericity is present in the data (Misconception Mutt 15.1). The more informative part of the table contains the Greenhouse–Geisser= ($\hat{\varepsilon} = 0.533$) and the Huynh–Feldt ($\tilde{\varepsilon} = 0.666$) estimates of sphericity.[6] If the data are perfectly spherical then these estimates will be 1. Therefore, both estimates indicate a departure from sphericity, so we may as well correct for it regardless of what Mauchly's test says. Jane Superbrain Box 15.3 explains how these estimates are used to correct the degrees of freedom for the *F*-statistic.

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

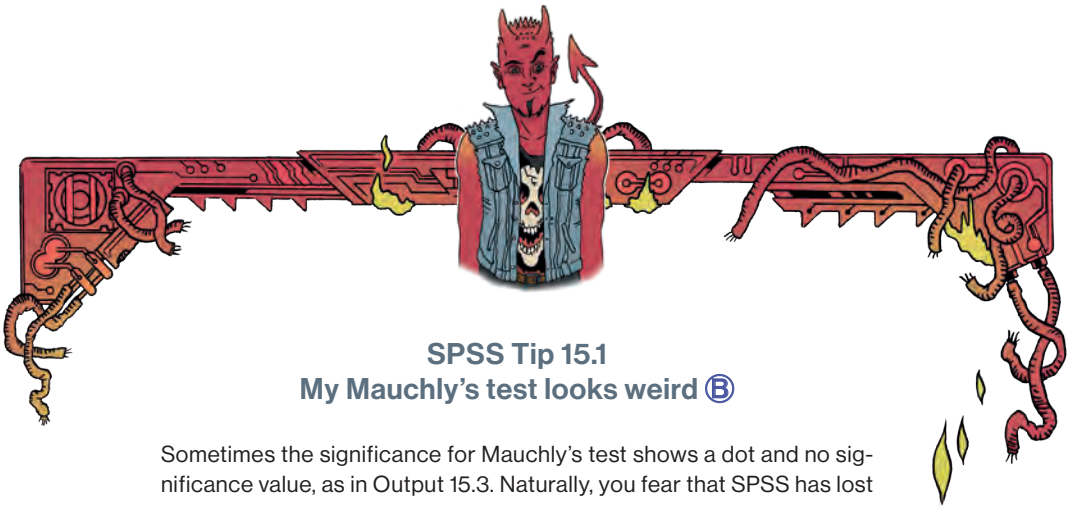
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b] Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Entity	.136	11.406	5	.047	.533	.666	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: Entity

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Output 15.2



SPSS Tip 15.1
My Mauchly's test looks weird Ⓑ

Sometimes the significance for Mauchly's test shows a dot and no significance value, as in Output 15.3. Naturally, you fear that SPSS has lost

6 The lowest possible value of the Greenhouse–Geisser estimate is $1/(k - 1)$, which with four conditions is $1/(4 - 1) = 0.33$. This is given as the lower-bound estimate in Output 15.2.

its senses and is going to break into your bedroom at night and print the equation for the Greenhouse–Geisser estimate on your face. Fear not, the reason for the dot is that you need at least three conditions for sphericity to be an issue (see Misconception Mutt 15.1). Therefore, if the repeated-measures variable has only two levels then sphericity is met, the estimates of sphericity are 1 (perfect sphericity), the chi-square is zero and has no degrees of freedom and so a *p*-value can't be computed. It would be a lot easier if SPSS printed in big letters 'Hooray! Hooray! Sphericity has gone away!' We can dream.

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b] Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Entity	1.000	.000	0	.	1.000	1.000	1.000

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: Entity

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Output 15.3

15.9.3 The *F*-statistic Ⓑ

Output 15.4 shows the summary information for the *F*-statistic that tests whether we can significantly predict the number of vocalizations from the group means (i.e., are the means significantly different?). Note that the values of the sum of squares for the repeated-measures effect of **Entity**, the model sum of squares (SS_M), the residual sum of squares (SS_R), the mean squares and *F*-statistic are the same as we calculated in Section 15.6. Assuming you've chosen an alpha of 0.05 as your criterion, the *p*-value associated with the *F*-statistic is significant because 0.026 is less than 0.05. This result implies that there was a significant difference in the number of vocalizations made when sniffing each entity. In other words, the sniffer-dog training has worked. The overall effect size is $\eta^2 = 0.35$, suggesting that the entities explain 35% of the variance in vocalizations, which is quite substantial.[7]

Although the result seems plausible, we have learnt that departures from sphericity make the *F*-statistic inaccurate against the assumed distribution, and Output 15.2 shows that these data are not spherical. Helpfully, alongside the *p*-value computed under the condition of sphericity, Output 15.4 also shows the *p*-value calculated using degrees of freedom adjusted by the three estimates of sphericity in Output 15.2 (Jane Superbrain Box 15.3).

7 Notice that despite SPSS labelling the effect size as partial eta squared (η_p^2) in Output 1.4, I have reported it as eta squared (η^2). In designs where there is a single predictor, partial eta squared and eta squared are the same thing, so although the SPSS is not incorrect to label it as partial eta squared, reporting it as such will lead people to assume there were other predictors that you haven't reported. Confusion will abound.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Entity	Sphericity Assumed	83.125	3	27.708	3.794	.026	.351
	Greenhouse-Geisser	83.125	1.599	52.001	3.794	.063	.351
	Huynh-Feldt	83.125	1.997	41.619	3.794	.048	.351
	Lower-bound	83.125	1.000	83.125	3.794	.092	.351
Error(Entity)	Sphericity Assumed	153.375	21	7.304			
	Greenhouse-Geisser	153.375	11.190	13.707			
	Huynh-Feldt	153.375	13.981	10.970			
	Lower-bound	153.375	7.000	21.911			

Output 15.4

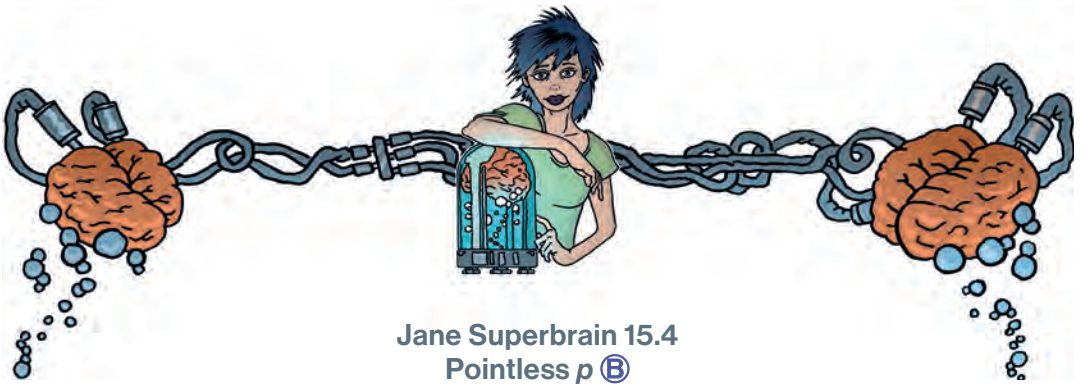


Jane Superbrain 15.3
Adjusting for sphericity Ⓑ

To adjust for the extent to which the data are not spherical, the degrees of freedom are multiplied by the estimates of sphericity in Output 15.2. The degrees of freedom for the model sum of squares were 3 and for the residual sum of squares they were 21 (Output 15.4). To get the Greenhouse–Geisser adjusted degrees of freedom, the original degrees of freedom are multiplied by 0.533 (the Greenhouse–Geisser estimate of sphericity), resulting in values of $3 \times 0.533 = 1.599$ for the model and $21 \times 0.533 = 11.19$ for the residual. The Huynh–Feldt correction is applied in the same way (see the Oliver Twisted on sphericity). The effect of reducing the degrees of freedom is to change the shape of the F-distribution that is used to obtain p, so the F-statistic is unchanged, but its p is based on a distribution with 1.599 and 11.19 degrees of freedom instead of 3 and 21. This increases the p-value associated with F.



The adjustments result in the observed F being non-significant when using the Greenhouse–Geisser correction (because $p > 0.05$) but significant using the Huynh–Feldt correction (because the probability value of 0.048 is just below the criterion value of 0.05). I noted earlier that the Greenhouse–Geisser correction is probably too strict and that the Huynh–Feldt correction is probably not strict enough, and we see this here because one of them takes the significance value above the conventional 0.05 threshold whereas the other doesn't. This leaves us with the puzzling dilemma of whether to accept this F-statistic as significant (see Jane Superbrain Box 15.4).



Jane Superbrain 15.4
Pointless *p* Ⓑ

In Section 3.2.2 I discussed significance testing leading to all-or-nothing thinking. These data illustrate this point beautifully: the two sphericity corrections lead to significance values just above (0.063) or just below (0.048) the 0.05 criterion. These significance values differ by only 0.015 but lead to completely opposite conclusions. The decision about 'significance' becomes rather arbitrary: choose one correction for a result that is 'significant' but choose another for non-significance. The conclusion of the study will largely be a function of researcher degrees of freedom (Section 3.3.2). The means and the effect size are unaffected by sphericity corrections, and so whether the *p* falls slightly above or slightly below 0.05 side-tracks us from the more important question of how big the effect was. We might be well advised to look at an effect size to see whether the effect is substantive, regardless of its significance.



A recommendation mentioned earlier is to use the Greenhouse–Geisser estimate unless it's greater than 0.75 (which it's not in this case). Some people also suggest taking an average of the two estimates (Stevens, 2002). In practical terms, rather than averaging the estimates, correcting the degrees of freedom manually and trying with an abacus or two to generate exact *p*-values, we could average the two *p*-values instead, which yields a $p = (0.063 + 0.048)/2 = 0.056$. In both cases the conclusion would be that there was no significant difference between means. Another option is to use multivariate test statistics (MANOVA), which do not assume sphericity (see O'Brien & Kaiser, 1985). Output 15.5 shows the multivariate test statistics (details of these test statistics can be found in Section 17.3), all of which are significant (because *p* is 0.002, which is less than the criterion value of 0.05). Based on multivariate tests we'd conclude that there are significant differences in the number of vocalizations to different entities when sniffed. It's easy to see how the decision rule applied to *p*-values can lead to results that don't replicate, conclusions that have been influenced by researcher degrees of freedom, and a lot of noise in the scientific literature.

15.9.4 Contrasts Ⓑ

It's worth remembering that, by some criteria, our main effect of the type of entity sniffed was not significant, and if this is the case then we would ignore contrasts. However, the multivariate tests provide some justification in interpreting the contrasts. Output 15.6 lists each contrast and its *F*-statistic, which compares the two chunks of variation within the contrast. Contrast 1 is labelled

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Entity	Pillai's Trace	.942	26.955[b]	3.000	5.000	.002	.942
	Wilks' Lambda	.058	26.955[b]	3.000	5.000	.002	.942
	Hotelling's Trace	16.173	26.955[b]	3.000	5.000	.002	.942
	Roy's Largest Root	16.173	26.955[b]	3.000	5.000	.002	.942

a. Design: Intercept
Within Subjects Design: Entity

b. Exact statistic

Output 15.5



Thinking back to the order in which we specified the levels of **Entity** (Section 15.8.1), what groups are being compared in each contrast?

Level 1 vs. Level 2, contrast 2 is labelled *Level 2 vs. Level 3* and contrast 3 is labelled *Level 3 vs. Level 4*. These labels reflect the fact that we selected a repeated contrast in Section 15.8.2, which compares each category of the predictor to the previous one.

Given the order in which we specified the levels of **Entity**, contrast 1 compares the mannequin and human, contrast 2 compares the shapeshifter to the human, and contrast 3 compares the alien to the shapeshifter. From Output 15.6 we can conclude that the numbers of vocalizations that dogs made were not significantly different when sniffing (1) a human compared to a mannequin, $F(1, 7) = 0.01$, $p = 0.920$, $\eta^2_p = 0.00$; (2) a shapeshifter compared to a human, $F(1, 7) = 1.26$, $p = 0.299$, $\eta^2_p = 0.15$; or (3) an alien compared to a shapeshifter, $F(1, 7) = 1.76$, $p = 0.227$, $\eta^2_p = 0.20$. Of course, these contrasts tell us nothing about whether vocalizations were different when sniffing aliens compared to mannequins and so on.

For those of you brave enough to attempt the custom contrasts in Section 15.8.3, Output 15.7 shows the tables of contrasts. Again, it's worth reiterating that you would not do both these contrasts and the ones we've just looked at: you would choose the ones that best reflect your hypotheses. For this particular example, I would personally have used the custom contrasts we're about to interpret. The top table in Output 15.7 is probably most useful because it includes the confidence interval for the contrast, but the bottom table reports the test statistics and degrees of freedom (note that the

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Entity	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Entity	Level 1 vs. Level 2	.125	1	.125	.011	.920	.002
	Level 2 vs. Level 3	18.000	1	18.000	1.260	.299	.153
	Level 3 vs. Level 4	45.125	1	45.125	1.756	.227	.201
Error(Entity)	Level 1 vs. Level 2	80.875	7	11.554			
	Level 2 vs. Level 3	100.000	7	14.286			
	Level 3 vs. Level 4	179.875	7	25.696			

Output 15.6

p-values are identical in the two tables). We can conclude from these tables that the numbers of vocalizations were significantly larger after sniffing aliens compared to non-aliens, $\bar{X}_{\text{Difference}} = 2.75$ [1.24, 4.27], $F(1, 7) = 18.41$, $p = 0.004$. However, there were no significant differences in vocalizations while sniffing aliens compared to shapeshifters, $\bar{X}_{\text{Difference}} = 2.38$ [–1.86, 6.61], $F(1, 7) = 1.76$, $p = 0.227$, or humans compared to mannequins, $\bar{X}_{\text{Difference}} = 0.13$ [–2.72, 2.97], $F(1, 7) = 0.01$, $p = 0.920$. The confidence intervals tell us (assuming this sample is one of the 95% that produces intervals containing the population value) that the difference in the number of vocalizations while sniffing aliens compared to non-aliens is likely to fall between 1.24 and 4.27, the difference while sniffing aliens compared to shapeshifters lies between –1.86 and 6.61, and the difference while sniffing humans compared to mannequins lies between –2.72 and 2.97.

Contrast Results (K Matrix)

Contrast[a]		Transformed Variable: Alien vs. non-alien	Alien vs. shapeshifter	Human vs. mannequin
L1	Contrast Estimate	2.750	2.375	.125
	Hypothesized Value	0	0	0
	Difference (Estimate – Hypothesized)	2.750	2.375	.125
	Std. Error	.641	1.792	1.202
	Sig.	.004	.227	.920
	95% Confidence Interval for Difference – Lower Bound	1.235	–1.863	–2.717
	95% Confidence Interval for Difference – Upper Bound	4.265	6.613	2.967

a. Estimable Function for Intercept

Univariate Test Results

Source	Transformed Variable	Sum of Squares	df	Mean Square	F	Sig.
Contrast	Alien vs. non-alien	60.500	1	60.500	18.413	.004
	Alien vs. shapeshifter	45.125	1	45.125	1.756	.227
	Human vs. mannequin	.125	1	.125	.011	.920
Error	Alien vs. non-alien	23.000	7	3.286		
	Alien vs. shapeshifter	179.875	7	25.696		
	Human vs. mannequin	80.875	7	11.554		

Output 15.7

15.9.5 *Post hoc* tests Ⓑ

Two health warnings. First, you wouldn't do contrasts and *post hoc* tests, you'd do one or the other. We're interpreting both purely for educational purposes. Second, we wouldn't interpret these post hoc tests if we had decided that the main effect of the entity sniffed was not significant. If you selected *post hoc* tests for the repeated-measures variable (see Section 15.8.3), then Output 15.8 is produced. The table should have a familiar feel from other post hoc tests we've looked at: it shows the difference between group means and their confidence intervals, the standard error and the significance value. The first two columns identify which levels of **Entity** are being compared, with 1 referring to level 1, 2 to level 2 and so on. We need to think back to how we specified the levels (Figure 15.7) to understand what's being compared in each row. Based on the significance values and the means (in Output 15.1) we can conclude that the number of vocalizations was significantly larger while sniffing an alien compared to both a human ($p = 0.012$) and a mannequin ($p = 0.006$), but not compared to a

shapeshifter (p = 1). The number of vocalizations while sniffing a human was not significantly different compared to while sniffing a mannequin or shapeshifter (both ps = 1). Finally, the number of vocalizations while sniffing a mannequin was not significantly different to when sniffing a shapeshifter (p = 1).

Pairwise Comparisons

Measure: MEASURE_1

(I) Entity	(J) Entity	Mean Difference (I–J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	-.125	1.202	1.000	-4.494	4.244
	3	-1.625	1.822	1.000	-8.249	4.999
	4	-4.000*	.732	.006	-6.661	-1.339
2	1	.125	1.202	1.000	-4.244	4.494
	3	-1.500	1.336	1.000	-6.359	3.359
	4	-3.875*	.811	.012	-6.825	-.925
3	1	1.625	1.822	1.000	-4.999	8.249
	2	1.500	1.336	1.000	-3.359	6.359
	4	-2.375	1.792	1.000	-8.891	4.141
4	1	4.000*	.732	.006	1.339	6.661
	2	3.875*	.811	.012	.925	6.825
	3	2.375	1.792	1.000	-4.141	8.891

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Output 15.8

15.10 Robust tests of one-way repeated-measures designs ©

A robust test of several dependent means with *post hoc* tests can be run using R via syntax in the workbook **robust_rm_anova.spwb** (Section 4.12). The syntax uses the *rmanova* and *rmmcp* functions from the *WRS2* package (Mair & Wilcox, 2019; Wilcox, 2017). These tests assume neither normality nor homogeneity of variance, so you can ignore these assumptions and plough ahead. The syntax is a little more complicated than in previous chapters because the data need to be restructured, so I have written a bunch of stuff to try to make that process as painless as possible. There's code that looks to see whether certain packages are installed and if they aren't it installs them, after which the core syntax is as follows (see SPSS Tip 15.2 for an explanation):

```
BEGIN PROGRAM R.
mySPSSdata = spssdata.GetDataFromSPSS(factorMode = "labels")
id <- "dog_name"
rm_factor <- c("alien", "human", "mannequin", "shapeshifter")
df <- pivot_longer(mySPSSdata, cols =all_of(rm_factor), names_to = "variable", values_to = "value")
names(df)[names(df) == id] <- "id"
rmanova(df$value, df$variable, df$id, tr = 0.2)
rmmcp(df$value, df$variable, df$id, tr = 0.2)
END PROGRAM.
```



Cramming Sam's Tips
One-way repeated-measures designs

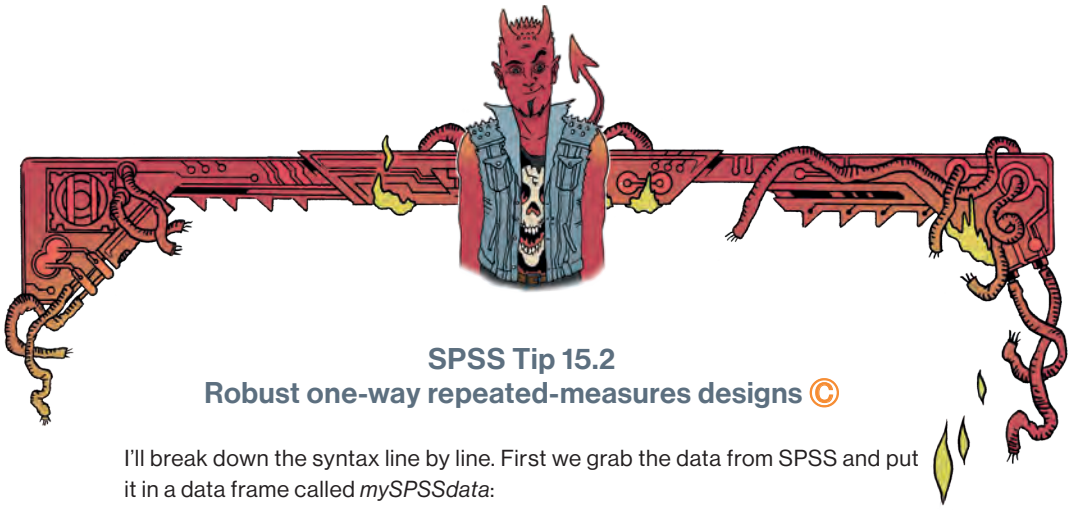
- One-way repeated-measures designs compare several means, when those means come from the same entities; for example, if you measured people's statistical ability each month over a year-long course.
- When you have three or more repeated-measures conditions there is an additional assumption: *sphericity*.
- You can test for sphericity using *Mauchly's test*, but it is better to always adjust for the departure from sphericity in the data.
- The table labelled *Tests of Within-Subjects Effects* shows the main *F*-statistic. Other things being equal, always read the row labelled *Greenhouse–Geisser* (or *Huynh–Feldt*, but you'll have to read this chapter to find out the relative merits of the two procedures). Assuming a criterion of 0.05 has been set before data were collected, if the value in the column labelled *Sig.* is less than or equal to 0.05 then the means of the conditions are significantly different.
- For contrasts and *post hoc* tests, look to the columns labelled *Sig.* to discover if your comparisons are significant (i.e., the value is less than 0.05).



Running this code (see SPSS Tip 15.2) should produce Output 15.9, which tells us that there was not a significant difference between means, $F_t(2.31, 11.55) = 2.75$, $p = 0.100$. The *post hoc* tests (which technically we should ignore because the overall test wasn't significant) show no significant difference between any groups (note that the column labelled *sig* shows FALSE for all comparisons, the *p*-values are all greater than the critical values (*p.crit*) and the confidence intervals all contain zero).

```
Test statistic: F = 2.7528
Degrees of freedom 1: 2.31
Degrees of freedom 2: 11.55
p-value: 0.1002
Call:
rmmcp(y = df$value, groups = df$variable, blocks = df$id, tr = 0.2)
                            psihat  ci.lower ci.upper p.value  p.crit   sig
alien vs. human            3.66667  -0.48300  7.81633 0.01360 0.01020 FALSE
alien vs. mannequin        4.00000  -0.35728  8.35728 0.01172 0.00851 FALSE
alien vs. shapeshifter     2.00000  -8.09920 12.09920 0.44148 0.01690 FALSE
human vs. mannequin        0.00000  -5.38802  5.38802 1.00000 0.05000 FALSE
human vs. shapeshifter    -1.83333  -9.23266  5.56599 0.34371 0.01270 FALSE
mannequin vs. shapeshifter -2.00000 -12.54827  8.54827 0.46001 0.02500 FALSE
```

Output 15.9



SPSS Tip 15.2
Robust one-way repeated-measures designs ©

I'll break down the syntax line by line. First we grab the data from SPSS and put it in a data frame called *mySPSSdata*:

```
mySPSSdata = spssdata.GetDataFromSPSS(factorMode = "labels")
```

Next, we define the name of the variable that represents the participant IDs. If you use this file on your own data, replace the word 'dog_name' with the (case-sensitive) name of your own ID variable. Make sure the name is in straight quotes and don't edit anything else:

```
id <- "dog_name"
```

Next, we list the variables that represent the levels of the repeated-measures predictor. If you use this file on your own data replace the words *alien*, *human*, etc. with the (case-sensitive) names from your own data file. Make sure the names are in straight quotes and don't edit anything else:

```
rm_factor <- c("alien", "human", "mannequin", "shapeshifter")
```

The next two lines convert your data frame from wide format to long (Section 4.6.1) and then rename the participant ID variable to be *id* (which is done to save you needing to edit the final two commands). If you've set up **id** and **rm_factor** correctly in the two lines above everything should go smoothly:

```
df <- pivot_longer(mySPSSdata, cols =all_of(rm_factor),
names_to = "variable", values_to = "value")

names(df)[names(df) == id] <- "id"
```

The final two lines run the robust test (*rmanova*) and *post hoc* tests (*rmmcp*). Because I've used generic variable names that were set up before we got to this point, you don't need to edit these functions unless you want to change the trim (*tr = 0.2*) to something other than the default of 20%:

```
rmanova(df$value, df$variable, df$id, tr = 0.2)

rmmcp(df$value, df$variable, df$id, tr = 0.2)
```

15.11 Effect sizes for one-way repeated-measures designs ©

As with independent designs the best measure of the overall effect size is omega squared (ω^2). However, just to make life even more complicated, the equations we've previously used for omega squared can't be used for repeated-measures designs: if you do use them the effect size will be overestimated. For the sake of simplicity some people do use the same equation but I'm going to hit simplicity in the face with Stingy the poison-ridden jellyfish, and embrace complexity like I do my children.

In repeated-measures designs, the equation for omega squared is (hang onto your hat):

$$\omega^2 = \frac{\left[\frac{k-1}{nk}\left(MS_M - MS_R\right)\right]}{MS_R + \frac{MS_B - MS_R}{k} + \left[\frac{k-1}{nk}\left(MS_M - MS_R\right)\right]} \tag{15.22}$$

I know what you're thinking and, no, I'm not having a laugh. Far from it. The equation isn't too bad if you break it down. First, there are some mean squares that we've come across before: the mean square for the model (MS_M) and the residual mean square (MS_R), both of which can be obtained from Output 15.4. There's also k, the number of conditions, which for these data would be 4 (there were four entities), and n, the number of participants that took part (in this case 8 dogs).

The main problem is MS_B. At the beginning of Section 15.3 (Figure 15.3) I mentioned that the total variation is broken down into a within-participant variation and a between-participant variation (SS_B), which we can calculate using Eq. 15.20. SPSS doesn't give us SS_W, but we know that it's made up of SS_M and SS_R, which we are given. By substituting these terms and rearranging the equation we get:

$$SS_T = SS_B + SS_M + SS_R$$

$$SS_B = SS_T - SS_M - SS_R \tag{15.23}$$

SPSS, which is clearly trying to hinder us at every step, also doesn't give us SS_T, and I'm afraid (unless I've missed something in the output) you're going to have to calculate it by hand (see Section 15.6.1). For these data it's 17.38 (Eq. 15.20). The next step is to convert this value to a mean squares by dividing by the degrees of freedom, which in this case are the number of dogs in the sample minus 1:

$$MS_B = \frac{SS_B}{df_B} = \frac{SS_B}{N-1} = \frac{17.38}{8-1} = 2.48 \tag{15.24}$$

Having done all this and probably died of boredom in the process, we now resurrect our corpses with renewed vigour for the effect size equation:

$$\begin{aligned}\omega^2 &= \frac{\left[\frac{4-1}{8\times 4}\left(27.71-7.30\right)\right]}{7.30 + \frac{2.48-7.30}{4} + \left[\frac{4-1}{8\times 4}\left(27.71-7.30\right)\right]} \\ &= \frac{1.91}{8.01} \\ &= 0.24\end{aligned} \tag{15.25}$$

I hope you agree it was worth the effort. Alternatively, live a happier life by coming to terms with the bias in the values of eta squared that SPSS includes in the output.

I've mentioned at various other points that it's more useful to have effect size measures for focused comparisons anyway (rather than an overall F-statistic), and so an alternative approach is to use our

old friend Cohen's $\hat{d}$. We could run a series of dependent *t*-tests between all pairs of entities using what we learnt in Section 10.9.3 (dependent *t*-test, see Figure 10.12) and select ☑ Estimate effect sizes and ◉ Corrected standard deviation of the difference to get the Cohen's $\hat{d}$ for every pair of entities. With a lot of pairs this might be more easily achieved using a syntax paragraph in your workbook. For the current data we'd run:

```
T-TEST PAIRS=alien alien alien human human mannequin WITH human mannequin
shapeshifter mannequin shapeshifter shapeshifter (PAIRED)
  /ES DISPLAY(TRUE) STANDARDIZER(CORRECTED_SD).
```

The first line creates the pairs we want with each variable after the word *PAIRS* matched with the corresponding variable after the word *WITH*. So, we're creating six pairs, in this order: alien–human, alien–mannequin, alien–shapeshifter, human–mannequin, human–shapeshifter and mannequin–shapeshifter. The line beginning /ES requests the corrected effect sizes. The resulting Output 15.10 indicates that the larger numbers of vocalizations when sniffing an alien compared to humans, $\hat{d}$ = 1.89 [0.37, 3.40], mannequins, $\hat{d}$ = 1.56 [0.37, 2.76], and shapeshifters, $\hat{d}$ = 0.92 [–0.82, 2.67], are all massive effects. The smaller number of vocalizations when sniffing a human compared to a shapeshifter, $\hat{d}$ = –0.62 [–1.99, 0.75], and a mannequin compared to a shapeshifter, $\hat{d}$ = –0.57 [–2.14, 0.99], are also substantial effects. Finally, the difference in the number of vocalizations when sniffing a human compared to a mannequin, $\hat{d}$ = 0.05 [–1.17, 1.27], is effectively zero. Note that all of the effects have very wide confidence intervals (because of the small sample size), indicating a lot of uncertainty around the true size of any of these effects.

Paired Samples Effect Sizes

			Standardizer[a]	Point Estimate	95% Confidence Interval Lower	Upper
Pair 1	Alien (vocalizations in 1 min) – Human (vocalizations in 1 min)	Cohen's d	2.054	1.887	.372	3.402
		Hedges' correction	2.312	1.676	.330	3.021
Pair 2	Alien (vocalizations in 1 min) – Mannequin (vocalizations in 1 min)	Cohen's d	2.558	1.564	.366	2.762
		Hedges' correction	2.880	1.389	.325	2.453
Pair 3	Alien (vocalizations in 1 min) – Shapeshifter (vocalizations in 1 min)	Cohen's d	2.574	.923	-.824	2.669
		Hedges' correction	2.899	.819	-.732	2.370
Pair 4	Human (vocalizations in 1 min) – Mannequin (vocalizations in 1 min)	Cohen's d	2.330	.054	-1.166	1.274
		Hedges' correction	2.624	.048	-1.036	1.131
Pair 5	Human (vocalizations in 1 min) – Shapeshifter (vocalizations in 1 min)	Cohen's d	2.412	-.622	-1.989	.746
		Hedges' correction	2.716	-.552	-1.767	.662
Pair 6	Mannequin (vocalizations in 1 min) – Shapeshifter (vocalizations in 1 min)	Cohen's d	2.832	-.574	-2.137	.990
		Hedges' correction	3.189	-.510	-1.899	.879

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation of the mean difference adjusted by the correlation between measures.
Hedges' correction uses the sample standard deviation of the mean difference adjusted by the correlation between measures, plus a correction factor.

Output 15.10

15.12 Reporting one-way repeated-measures designs Ⓑ

Reporting repeated-measures designs is much the same as independent designs, except that we need to pay attention to reporting the corrected degrees of freedom. The degrees of freedom used to assess the *F*-statistic are the degrees of freedom for the effect of the model (df_M = 1.60) and the degrees of freedom for the residuals of the model (df_R = 11.19). Therefore, we could report the main finding as:

- ✓ The Greenhouse–Geisser estimate of the departure from sphericity was $\varepsilon = 0.53$. The number of vocalizations was not significantly affected by the entity sniffed, $F(1.60, 11.19) = 3.79$, $p = 0.063$, $\omega^2 = 0.24$. In short, the sniffer-dog training appeared to be unsuccessful.

For Huynh–Feldt corrected values:

- ✓ The Huynh–Feldt estimate of the departure from sphericity was $\varepsilon = 0.67$. The number of vocalizations was significantly affected by the entity sniffed, $F(2, 13.98) = 3.79$, $p = 0.048$, $\omega^2 = 0.24$. In short, the sniffer-dog training appeared to be successful.

We could report multivariate tests. There are four different test statistics, but in most situations you should report Pillai's trace, V (see Chapter 17), and the associated F and its degrees of freedom (all from Output 15.5):

- ✓ The Greenhouse–Geisser estimate of sphericity showed a substantial deviation ($\varepsilon = 0.53$), therefore multivariate tests are reported. The sniffer-dog training appeared to be successful in that the number of vocalizations was significantly affected by the entity sniffed, $V = 0.94$, $F(3, 5) = 26.96$, $p = 0.002$, $\omega^2 = 0.24$.

Finally, robust tests:

- ✓ The Greenhouse–Geisser estimate of sphericity showed a substantial deviation ($\varepsilon = 0.53$). Robust tests of 20% trimmed means implemented with the *WRS2* package in R (Mair & Wilcox, 2019) showed that the sniffer-dog training was unsuccessful: the number of vocalizations was not significantly affected by the entity sniffed, F_t (2.31, 11.55) = 2.75, $p = 0.100$, $\omega^2 = 0.24$.

15.13 A scented factorial repeated-measures design Ⓑ

We have seen that between-group designs can be extended to incorporate multiple predictor/independent variables. The same is true for repeated-measures designs. Back to the alien invasion. Having intercepted the humans' pilot study on training sniffer dogs, the aliens were excited by our apparent inability to train sniffer dogs to detect them. Aliens are far too wedded to p-values in small samples. Nevertheless, they decided to move their invasion plan forward and, to make life even harder for the sniffer dogs, came up with a plan to mask their alien smell with other scents. After extensive research they agreed that the two most effective scents would be human pheromones (which they hoped would make them smell human-like) and fox pheromones (because they are a powerful, distracting smell for dogs). The aliens started smearing themselves with humans and foxes in preparation to invade. Meanwhile, TERD got wind of their plan and set about testing how effective it would be. They trained 50 sniffer dogs. Like the pilot study, these dogs were rewarded for making vocalizations while sniffing alien space lizards during a training phase. Then during a test phase the 50 dogs sniffed nine different entities for 1 minute each: three alien space lizards in their natural form, three shapeshifting alien space lizards in humanoid form and three actual humans. Within each type of entity, one had no masking scent, one was smeared in human pheromones and one wore fox pheromones. As in the previous experiment, the number of vocalizations made during each 1-minute sniff was recorded. The data are in **alien_scents.sav**.

15.14 Factorial repeated-measures designs using SPSS Ⓑ

Figure 15.10 shows a portion of the data and illustrates the experimental design. Note from the structure of the data that each row represents a single participant dog. Each dog participates in all conditions (in this case each dog sniffs all combinations of entities and scents) so, apart from the variable that identifies the dog (**dog_id**), there are nine variables that correspond to the various combinations of entity and scent. The variable names and value labels are in Table 15.4.

Table 15.4 Variable names and labels for the alien_scents.sav data

Variable name	Variable label
human_none	Vocalizations (human, no scent)
human_human	Vocalizations (human, human scent)
human_fox	Vocalizations (human, fox scent)
shapeshifter_none	Vocalizations (shapeshifter, no scent)
shapeshifter_human	Vocalizations (shapeshifter, human scent)
shapeshifter_fox	Vocalizations (shapeshifter, fox scent)
alien_none	Vocalizations (alien, no scent)
alien_human	Vocalizations (alien, human scent)
alien_fox	Vocalizations (alien, fox scent)

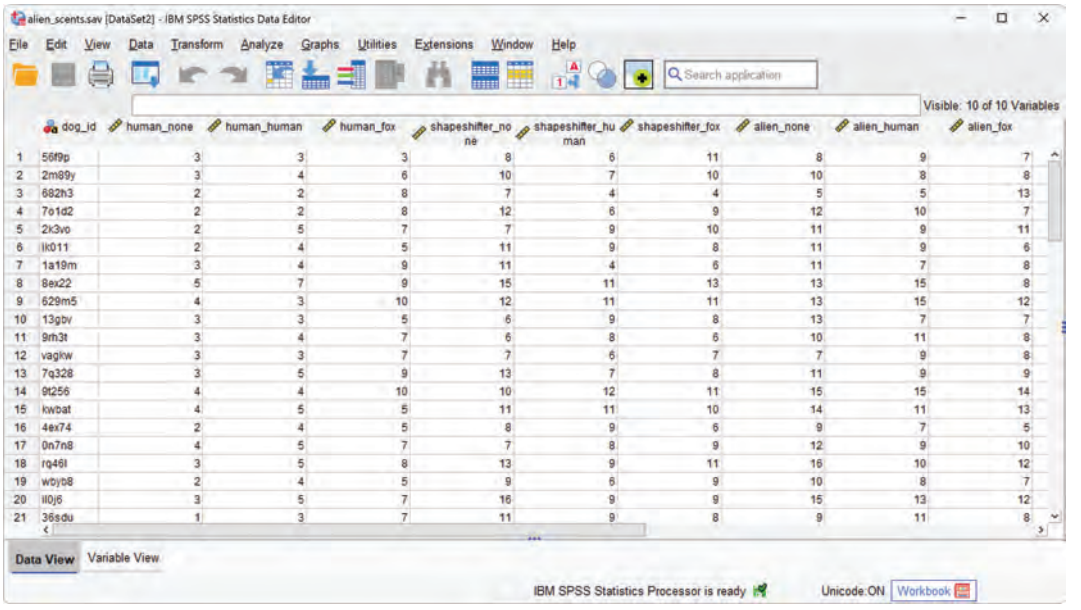


Figure 15.10 The first 21 cases of alien_scents.sav

Select *Analyze* > *General Linear Model* > Repeated Measures... to access the dialog boxes for a repeated-measures model. First, we define the repeated-measures predictors. In this case there are two: **Entity** (human, shapeshifter, alien) and **Scent** (none, human, fox). Type the word 'Entity' into the box labelled *Within-Subject Factor Name*, type '3' into the box labelled *Number of Levels*, and click Add (Figure 15.11, left). This variable appears in the list of variables as *Entity(3)*, which means we have defined a predictor called **Entity** that has three levels (Figure 15.11, middle). We repeat this process for the second predictor by typing 'Scent' into the box labelled *Within-Subject Factor Name*, typing '3' into the space labelled *Number of Levels*, and clicking Add (Figure 15.11, middle). The variable will appear in the list as *Scent(3)* (Figure 15.11, right). Click Define to go to the main dialog box.

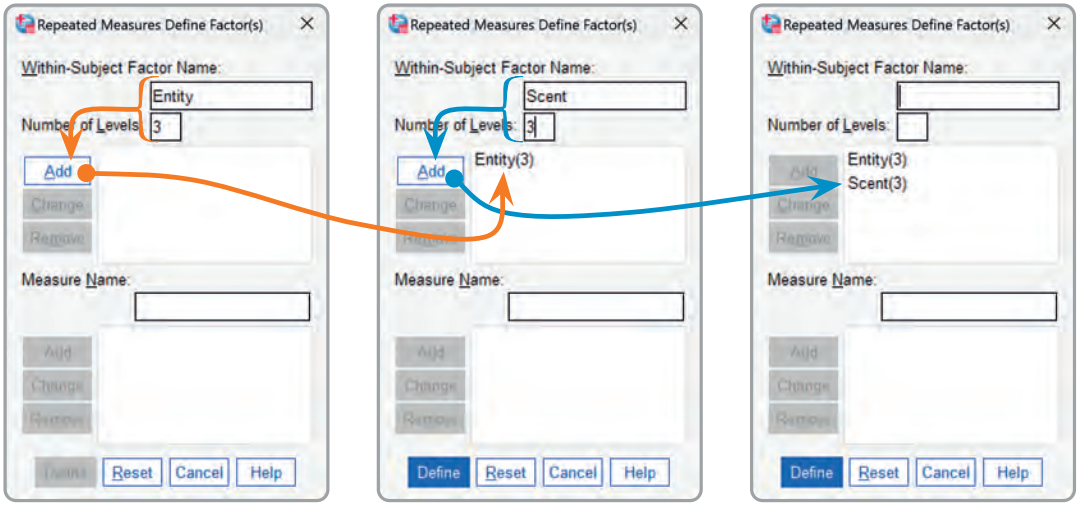


Figure 15.11 The *Define Factor(s)* dialog box for factorial repeated-measures designs

The main dialog box (Figure 15.12) is basically the same as in the previous example where we had one predictor variable. The main difference that there are nine question marks, after each of which there are two numbers in brackets, rather than one. These numbers relate to the levels of the two variables we have just created. To help us to make sense of these values at the top of the *Within-Subjects Variables* box, SPSS prints *(Entity, Scent)* which tells us that the first number in the brackets represents levels of the variable **Entity** and the second number in the brackets represents levels of the variable **Scent**. So, we have:

- _?_(1, 1) ⇒ variable representing the 1st level of **Entity** and the 1st level of **Scent**
- _?_(1, 2) ⇒ variable representing the 1st level of **Entity** and the 2nd level of **Scent**
- _?_(1, 3) ⇒ variable representing the 1st level of **Entity** and the 3rd level of **Scent**
- _?_(2, 1) ⇒ variable representing the 2nd level of **Entity** and the 1st level of **Scent**
- _?_(2, 2) ⇒ variable representing the 2nd level of **Entity** and the 2nd level of **Scent**
- _?_(2, 3) ⇒ variable representing the 2nd level of **Entity** and the 3rd level of **Scent**
- _?_(3, 1) ⇒ variable representing the 3rd level of **Entity** and the 1st level of **Scent**
- _?_(3, 2) ⇒ variable representing the 3rd level of **Entity** and the 2nd level of **Scent**
- _?_(3, 3) ⇒ variable representing the 3rd level of **Entity** and the 3rd level of **Scent**

We need to replace the question marks with variable names from the data editor (which are listed on the left-hand side of the dialog box). At this stage we need to think about which conditions to assign to which level of each variable. For example, if we entered **human_none** into the list first, then SPSS

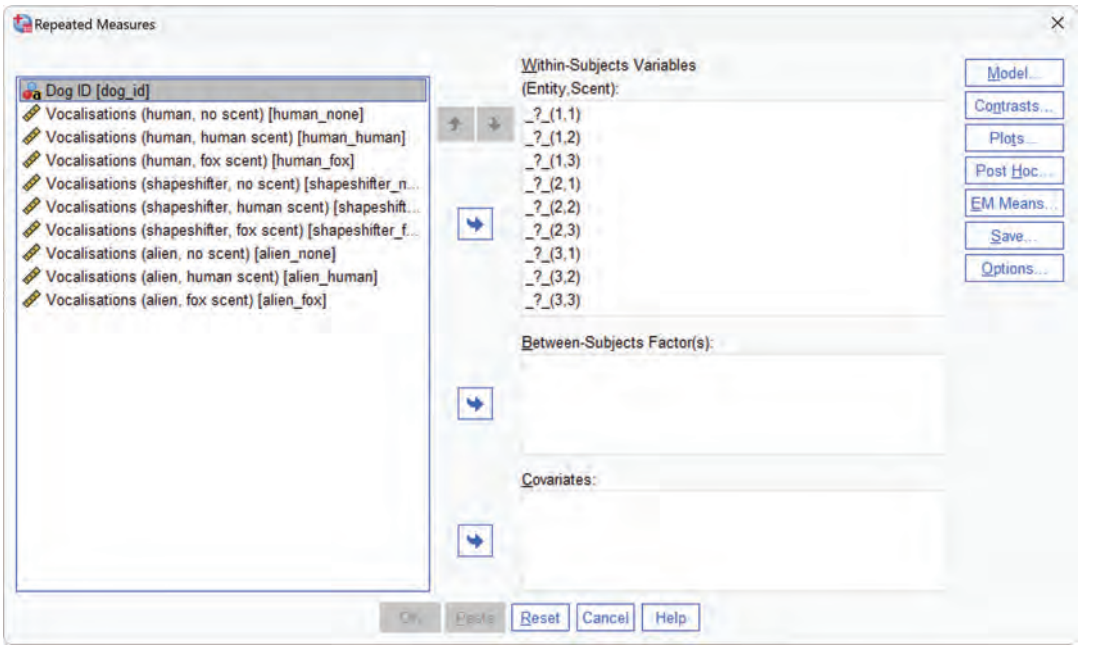


Figure 15.12 The main dialog box for factorial repeated-measures designs before completion

would treat human as the first level of **Entity** and no scent as the first level of the **Scent** variable. However, if we entered **shapeshifter_fox** into the list first, we're specifying shapeshifter as the first level of **Entity**, and fox scent as the first level of **Scent**. For this reason, we need to think about what contrasts we might want *before* assigning variables in this dialog box. The built-in contrasts we can choose from were described in Table 12.8.

The first variable, **Entity**, had three conditions, two of which involved alien entities. In a sense, the human acts as a control to whether the effects of the scents are specific to aliens. Therefore, we might want to compare the vocalizations to shapeshifters and aliens with those when sniffing humans. This comparison could be done by either specifying a simple contrast in which the shapeshifter and alien trials are compared to the human ones, or using a difference contrast in which both alien trials are compared to the human trials (contrast 1) before being compared to each other (contrast 2). We will use a simple contrast, but either way human should be entered as either the first or last level of **Entity**. For the **Scent** variable the trials where the entities did not wear a scent act as a baseline for the number of vocalizations we can expect when sniffing the different entities. Therefore, the no scent category is a sensible reference category in a simple contrast so, again, we'd want to enter it as either the first or last level.

Based on this discussion, it makes sense to have human as level 1 of **Entity** and none as level 1 of **Scent**. The remaining levels can be decided arbitrarily. I have chosen shapeshifter as level 2 and alien as level 3 of **Entity**, and human as level 2 and fox as level 3 of **Scent**. I would, therefore, assign the variables as in Figure 15.13. Coincidentally, this is the order in which variables are listed in the data editor. (It's not a coincidence: I thought ahead about what contrasts I would do and entered the variables in the appropriate order.) When these variables have been assigned, the dialog box looks like Figure 15.14.

15.14.1 Contrasts Ⓑ

Click Contrasts... to access the dialog box in Figure 15.15. In the previous section I described why it might be interesting to use the human and none conditions as baseline categories for the **Entity** and

human_none	→	_?_(1,1)
human_human	→	_?_(1,2)
human_fox	→	_?_(1,3)
shapeshifter_none	→	_?_(2,1)
shapeshifter_human	→	_?_(2,2)
shapeshifter_fox	→	_?_(2,3)
alien_none	→	_?_(3,1)
alien_human	→	_?_(3,2)
alien_fox	→	_?_(3,3)

Figure 15.13 Variable allocations for the alien scent data

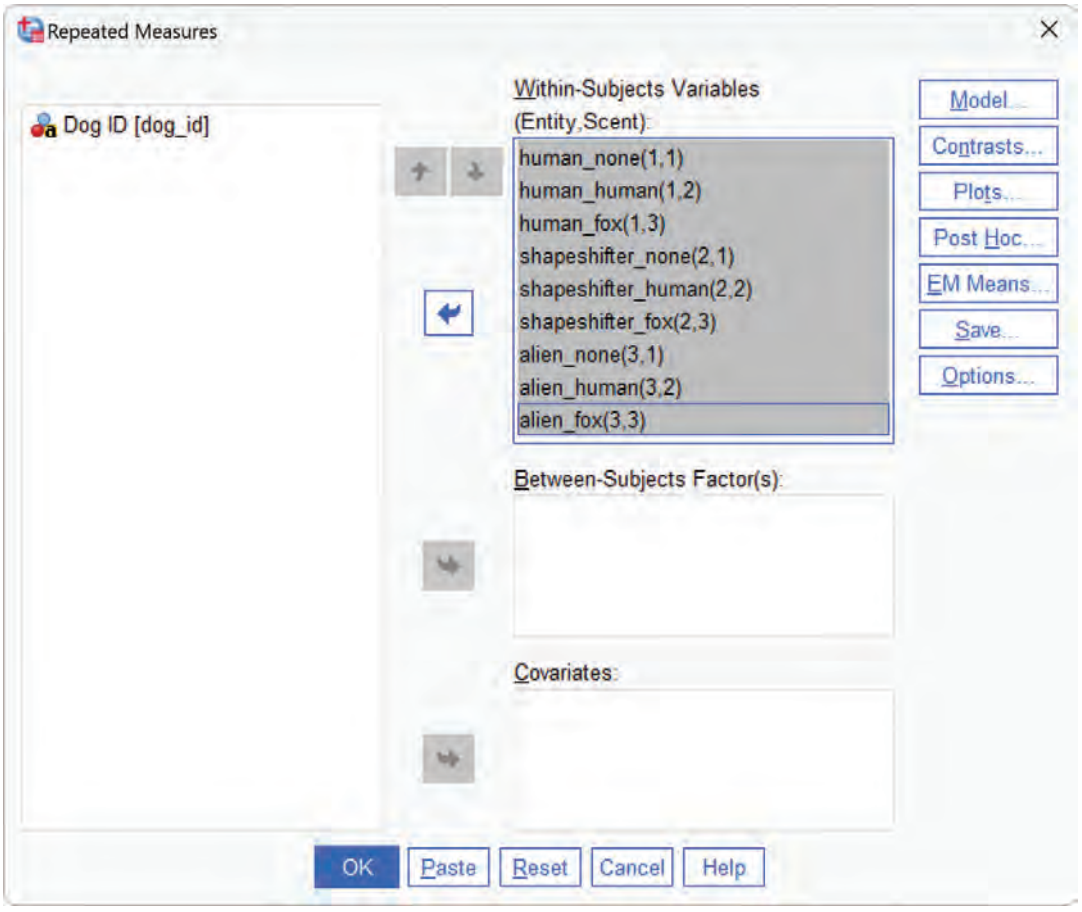


Figure 15.14 The main dialog box for factorial repeated-measures designs after completion

Scent variables, respectively. We've seen how to assign contrasts before (Sections 13.6.5 and 15.8.2), so use your SPSS skills to select a simple contrast for both **Entity** and **Scent**. We assigned variables such that the control category was the first one; therefore, we can leave the default reference category of ◉ First. Click Continue to return to the main dialog box.

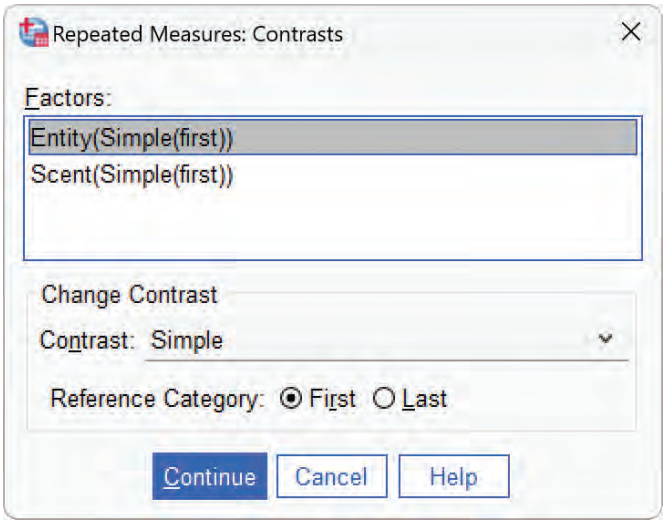


Figure 15.15 The *Contrasts* dialog box for factorial repeated-measures designs

15.14.2 Plotting interactions Ⓑ

In the previous example we ignored the *Plots* dialog box, but with two predictors it is a convenient way to plot the interaction (although better to plot these graphs before fitting the model). Click Plots..., drag **Entity** from the variables list to the space labelled *Horizontal Axis* (or click ➧), drag **Scent** to the space labelled *Separate Lines* and click Add. Choose a *Line chart*, ask for error bars that display 95% confidence intervals, and for reasons explained in the previous chapter, select *Y axis starts at 0* to scale the *y*-axis from zero (see Figure 15.16). Click Continue to return to the main dialog box.

15.14.3 Simple effects and other options Ⓑ

Figure 15.17 shows some of the other options that are familiar from previous examples. As in the previous example, *post hoc* tests are disabled because this design has only repeated-measures variables. To get some we can click EM Means..., drag (or click ➧) all variables in the box labelled *Factor(s) and Factor Interactions* to the box labelled *Display Means for:*, select ☑ Compare main effects and choose a correction from the drop-down menu (I chose Bonferroni). However, with factorial designs we are typically most interested in the interaction effect (in this case, does the effect of scent depend upon the entity wearing it?). As we have seen before, simple effects analysis is a useful way to break down an interaction (Section 14.5). By dragging the interaction into *Display Means for:*, the option to ☑ Compare simple main effects becomes available. So, select this option (Figure 15.17, left). Click Options... and select the same options as we have in previous examples (Figure 15.17, right). Click Continue to return to the main dialog box, then click OK to fit the model.

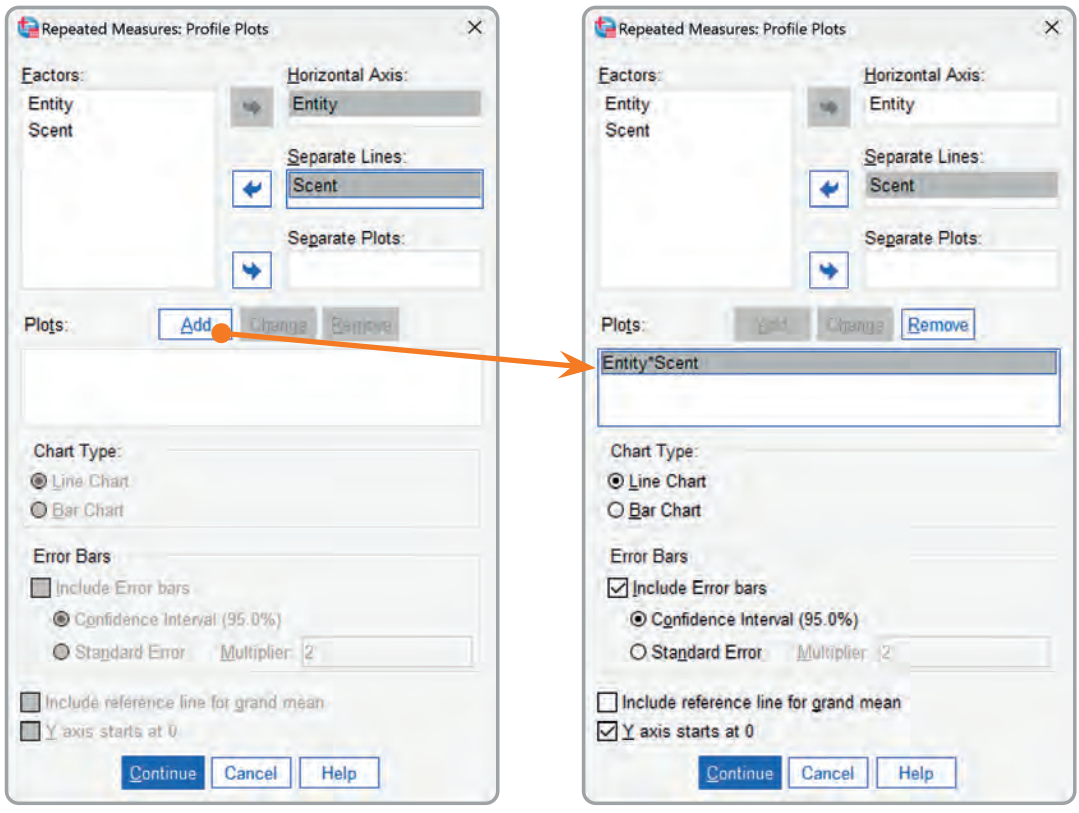


Figure 15.16 Defining profile plots in repeated-measures designs

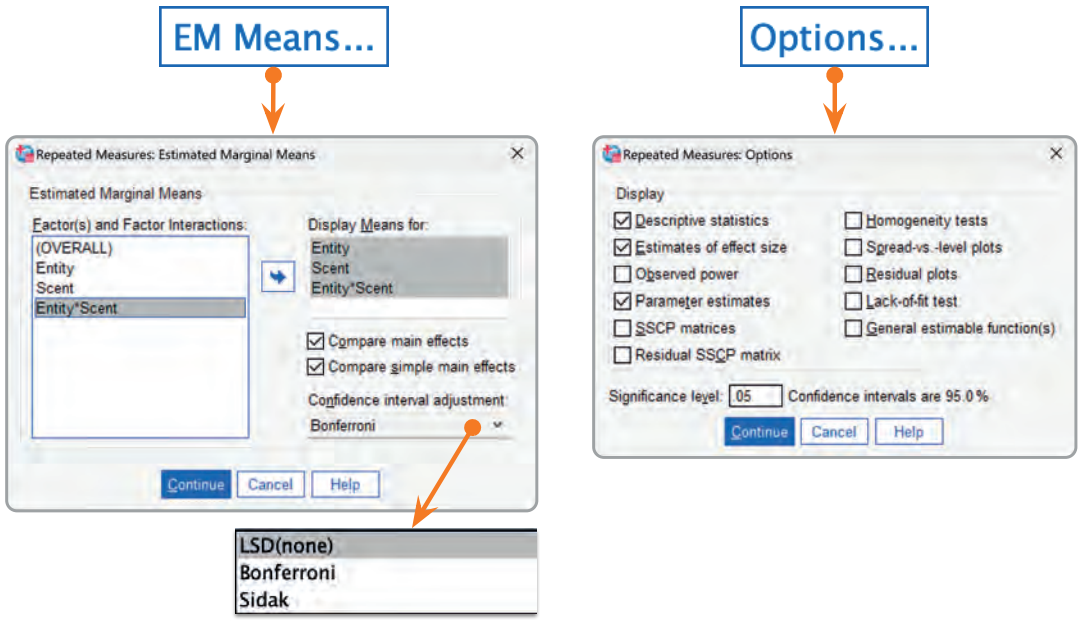


Figure 15.17 *Estimated Marginal Means* and *Options* dialog boxes for GLM repeated measures



Oditi's Lantern
Repeated-measures designs

'I, Oditi, believe that we are closer to achieving our mission of understanding the secrets hidden within the numbers. The Earth is a sphere, and I believe that if I am to dominate, erm, I mean understand the Earth then I must educate you about sphere-icity. Knowledge of sphericity will rewire your neural connections and give you insight into analysing repeated-measures designs. Stare into my lantern and feel your brain burn, but in a nice way.'



15.15 Interpreting factorial repeated-measures designs Ⓑ

Output 15.11 shows two tables. The first lists the variables from the data editor and the level of each predictor that they represent and is a useful way to verify that you entered the variables in the correct order for the comparisons that you want. The second table contains the means and standard deviations across the nine conditions. The names in this table are the variable labels that you entered in the data editor (refer to Table 15.4). The descriptive statistics tell us that the number of vocalizations tended to be lower (on average) when sniffing humans compared to either type of alien. When sniffing aliens the number of vocalizations was largest when they wore no scent.

Within-Subjects Factors

Measure: MEASURE_1

Entity	Scent	Dependent Variable
1	1	human_none
	2	human_human
	3	human_fox
2	1	shapeshifter_none
	2	shapeshifter_human
	3	shapeshifter_fox
3	1	alien_none
	2	alien_human
	3	alien_fox

Descriptive Statistics

	Mean	Std. Deviation	N
Vocalisations (human, no scent)	2.98	1.078	50
Vocalisations (human, human scent)	4.16	1.419	50
Vocalisations (human, fox scent)	7.32	1.984	50
Vocalisations (shapeshifter, no scent)	10.42	2.785	50
Vocalisations (shapeshifter, human scent)	8.78	2.468	50
Vocalisations (shapeshifter, fox scent)	8.84	2.427	50
Vocalisations (alien, no scent)	12.06	3.026	50
Vocalisations (alien, human scent)	9.98	2.752	50
Vocalisations (alien, fox scent)	9.18	2.413	50

Output 15.11

Output 15.12 shows the sphericity estimates for each of the three effects in the model (two main effects and one interaction). The Greenhouse–Geisser estimates are all very close to 1, so it would be reasonable not to correct the effects for sphericity, but given that I advised to routinely apply the correction, we will!

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b] Greenhouse-Geisser	Epsilon[b] Huynh-Feldt	Epsilon[b] Lower-bound
Entity	.988	.559	2	.756	.989	1.000	.500
Scent	.976	1.143	2	.565	.977	1.000	.500
Entity * Scent	.834	8.595	9	.476	.915	.998	.250

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: Entity + Scent + Entity * Scent

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Output 15.12

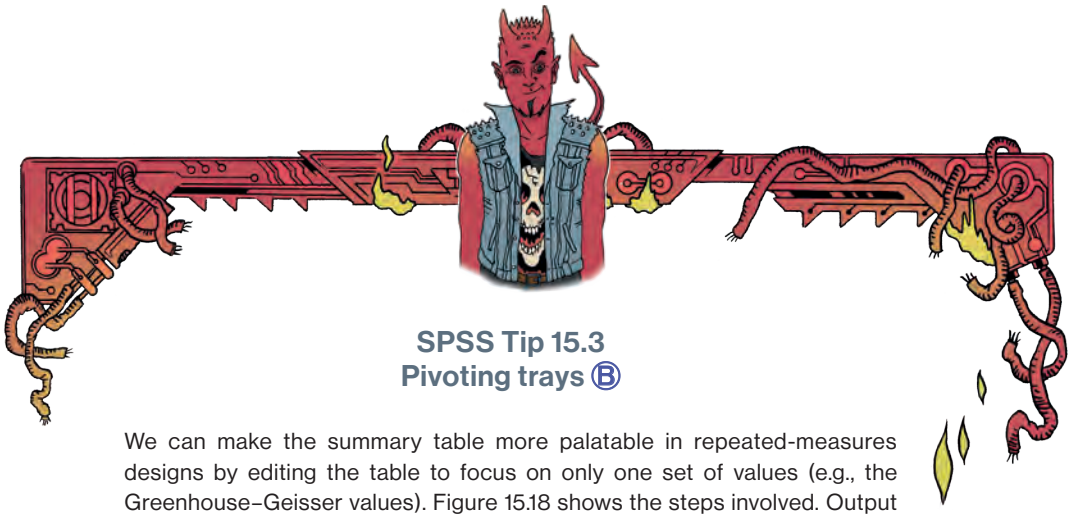
Output 15.13 shows the *F*-statistics (with corrections) and is split into sections that refer to each effect in the model and the associated error term. The table is quite mind-blowing, but we can stay calm by focusing on the information that we plan to use. For example, if you follow my advice to routinely report Greenhouse–Geisser corrected values then we can edit the table to show only these values (SPSS Tip 15.3). The significance values tell us that there is a significant main effect of the entity

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Entity	Sphericity Assumed	2641.258	2	1320.629	315.947	<.001	.866
	Greenhouse-Geisser	2641.258	1.977	1335.913	315.947	<.001	.866
	Huynh-Feldt	2641.258	2.000	1320.629	315.947	<.001	.866
	Lower-bound	2641.258	1.000	2641.258	315.947	<.001	.866
Error(Entity)	Sphericity Assumed	409.631	98	4.180			
	Greenhouse-Geisser	409.631	96.879	4.228			
	Huynh-Feldt	409.631	98.000	4.180			
	Lower-bound	409.631	49.000	8.360			
Scent	Sphericity Assumed	68.458	2	34.229	12.913	<.001	.209
	Greenhouse-Geisser	68.458	1.954	35.035	12.913	<.001	.209
	Huynh-Feldt	68.458	2.000	34.229	12.913	<.001	.209
	Lower-bound	68.458	1.000	68.458	12.913	<.001	.209
Error(Scent)	Sphericity Assumed	259.764	98	2.651			
	Greenhouse-Geisser	259.764	95.746	2.713			
	Huynh-Feldt	259.764	98.000	2.651			
	Lower-bound	259.764	49.000	5.301			
Entity * Scent	Sphericity Assumed	742.609	4	185.652	60.261	<.001	.552
	Greenhouse-Geisser	742.609	3.659	202.980	60.261	<.001	.552
	Huynh-Feldt	742.609	3.990	186.103	60.261	<.001	.552
	Lower-bound	742.609	1.000	742.609	60.261	<.001	.552
Error(Entity*Scent)	Sphericity Assumed	603.836	196	3.081			
	Greenhouse-Geisser	603.836	179.268	3.368			
	Huynh-Feldt	603.836	195.526	3.088			
	Lower-bound	603.836	49.000	12.323			

Output 15.13



SPSS Tip 15.3
Pivoting trays Ⓑ

We can make the summary table more palatable in repeated-measures designs by editing the table to focus on only one set of values (e.g., the Greenhouse–Geisser values). Figure 15.18 shows the steps involved. Output 15.14 shows the resulting summary table. Compare this output with Output 15.13: note that it is less headache-inducing because it contains only the Greenhouse–Geisser corrected values (the other information hasn't gone, it is hidden in the layers of the table and can be accessed by double-clicking on the table).

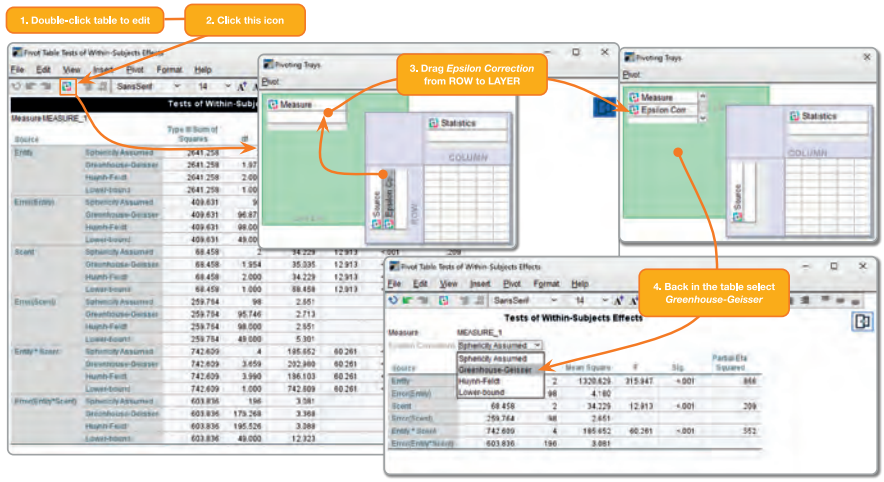


Figure 15.18 Using the pivot trays option to 'hide' parts of a summary table

Tests of Within-Subjects Effects

Measure: MEASURE_1
Greenhouse-Geisser

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Entity	2641.258	1.977	1335.913	315.947	<.001	.866
Error(Entity)	409.631	96.879	4.228			
Scent	68.458	1.954	35.035	12.913	<.001	.209
Error(Scent)	259.764	95.746	2.713			
Entity * Scent	742.609	3.659	202.980	60.261	<.001	.552
Error(Entity*Scent)	603.836	179.268	3.368			

Output 15.14

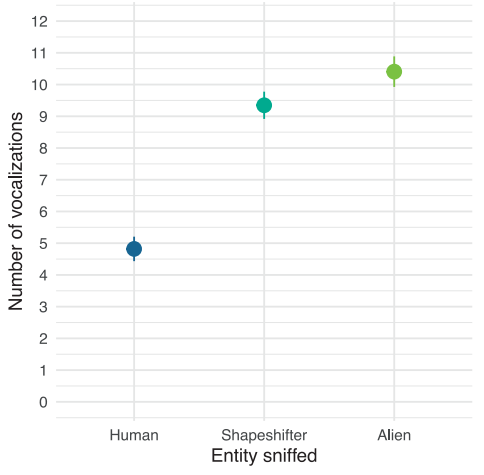
sniffed, a significant main effect of the type of scent used and a significant interaction between these two variables. For educational purposes, I will examine each of these effects in turn, but there's typically not a lot of value in interpreting main effects in the presence of a significant interaction effect.

15.15.1 The main effect of **Entity** Ⓑ

The type of entity sniffed was significant, which tells us that if we ignore the type of scent that was used, dogs vocalized more when sniffing some entities than others. In Section 15.14.3 we requested estimated marginal means for the effects in the model, and the means and standard errors for the main effect of **Entity** are shown in Figure 15.19. These means take the average of the means in Output 15.11 for a given entity. For example, for the human (ignoring type of scent) the mean is:

$$\begin{aligned}\bar{X}_{\text{Human}} &= \frac{\bar{X}_{\text{Human, no scent}} + \bar{X}_{\text{Human, human scent}} + \bar{X}_{\text{Human, fox scent}}}{3} \\ &= \frac{2.98 + 4.16 + 7.32}{3} \\ &= 4.82\end{aligned} \tag{15.26}$$

The levels of **Entity** are labelled 1, 2 and 3, so we must think back to the order in which we assigned variables to know which row of the table relates to which entity. We entered the human first and the alien last (as shown in Output 15.11). Figure 15.19 includes a plot of these means, which shows that dogs vocalized more when sniffing aliens and shapeshifters than when sniffing humans.



1. Entity

Measure: MEASURE_1

Entity	Mean	Std. Error	95% Confidence Interval Lower Bound	Upper Bound
1	4.820	.165	4.488	5.152
2	9.347	.291	8.763	9.931
3	10.407	.316	9.771	11.043

Figure 15.19 Output and graph of the main effect of **Entity**

Assuming that the interaction effect wasn't significant and this main effect was of interest, we'd break it down with either the simple contrast we specified or *post hoc* tests. We have done both, but bear in mind we'd typically use one or the other. Using the means (Figure 15.19) to determine the direction of the effect and the simple contrasts (Output 15.15), we can conclude that, compared to when sniffing humans, dogs vocalized significantly more when sniffing a shapeshifter, $F(1, 49) = 410.68$, $p < 0.001$, $\eta_p^2 = 0.89$, and an alien, $F(1, 49) = 543.12$, $p < 0.001$, $\eta_p^2 = 0.92$.

Output 15.16 shows the Bonferroni adjusted pairwise comparisons for the main effect of **Entity**. There is redundancy in this table in that every pair of entities appears twice, but with the order of means reversed. The highlighted rows show the positive differences, that is, where the

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Entity	Scent	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Entity	Level 2 vs. Level 1		1024.536	1	1024.536	410.678	<.001	.893
	Level 3 vs. Level 1		1560.542	1	1560.542	543.121	<.001	.917
Error(Entity)	Level 2 vs. Level 1		122.242	49	2.495			
	Level 3 vs. Level 1		140.791	49	2.873			
Scent		Level 2 vs. Level 1	35.842	1	35.842	21.176	<.001	.302
		Level 3 vs. Level 1	.080	1	.080	.039	.843	.001
Error(Scent)		Level 2 vs. Level 1	82.936	49	1.693			
		Level 3 vs. Level 1	99.476	49	2.030			
Entity * Scent	Level 2 vs. Level 1	Level 2 vs. Level 1	397.620	1	397.620	39.330	<.001	.445
		Level 3 vs. Level 1	1752.320	1	1752.320	136.795	<.001	.736
	Level 3 vs. Level 1	Level 2 vs. Level 1	531.380	1	531.380	49.726	<.001	.504
		Level 3 vs. Level 1	2606.420	1	2606.420	231.964	<.001	.826
Error(Entity*Scent)	Level 2 vs. Level 1	Level 2 vs. Level 1	495.380	49	10.110			
		Level 3 vs. Level 1	627.680	49	12.810			
	Level 3 vs. Level 1	Level 2 vs. Level 1	523.620	49	10.686			
		Level 3 vs. Level 1	550.580	49	11.236			

Main effect of **Entity**

Main effect of **Scent**

Entity × Scent interaction

Output 15.15

Pairwise Comparisons

Measure: MEASURE_1

(I) Entity	(J) Entity	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
1	2	-4.527*	.223	<.001	-5.080	-3.973
	3	-5.587*	.240	<.001	-6.181	-4.992
2	1	4.527*	.223	<.001	3.973	5.080
	3	-1.060*	.245	<.001	-1.666	-.454
3	1	5.587*	.240	<.001	4.992	6.181
	2	1.060*	.245	<.001	.454	1.666

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

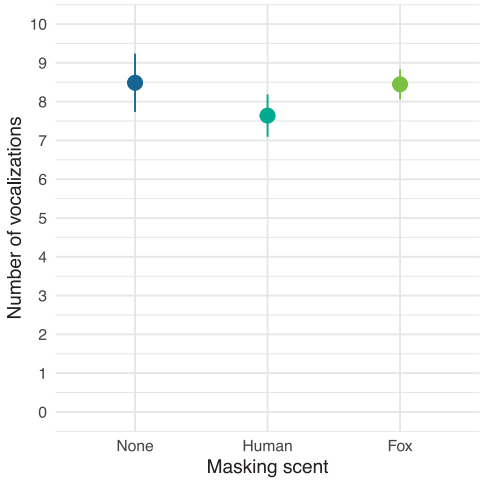
Output 15.16

entity in the first column has a larger mean than the one in the second. These effects tell us that when sniffing a shapeshifter dogs vocalized significantly more than when sniffing a human, $\bar{X}_{\text{Difference}} = 4.53$ [3.97, 5.08], $p < 0.001$. When sniffing an alien dogs vocalized significantly more than when sniffing both a human, $\bar{X}_{\text{Difference}} = 5.59$ [4.99, 6.18], $p < 0.001$, and a shapeshifter, $\bar{X}_{\text{Difference}} = 1.06$ [0.45, 1.67], $p < 0.001$.

15.15.2 The main effect of **Scent** Ⓑ

The type of scent also had a significant influence on participants' ratings of the entities (Output 15.13). This effect tells us that if we ignore the entity sniffed, the quantity of dogs' vocalizations differed according to the type of scent that was used. Figure 15.20 shows the means that we requested in Section 15.14.3. The levels of **Scent** are labelled 1, 2 and 3, so we need to again think back to how we assigned variables. We assigned no scent to the first level and fox scent to the last. Figure 15.20 includes a plot of these means (and their confidence intervals) and shows that the average numbers of vocalizations are quite similar for the different scents.

Using the means (Figure 15.20) to determine the direction of the effect and the simple contrasts (Output 15.15), we can conclude that, compared to when sniffing an entity wearing no scent, dogs vocalized significantly less when a human scent was used, $F(1, 49) = 21.18$, $p < 0.001$, $\eta_p^2 = 0.30$, but made a similar number of vocalizations when fox scent was used, $F(1, 49) = 0.04$, $p = 0.843$, $\eta_p^2 < 0.01$.



2. Scent

Measure: MEASURE_1

Scent	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	8.487	.264	7.956	9.018
2	7.640	.249	7.141	8.139
3	8.447	.245	7.955	8.939

Figure 15.20 Output and graph of the main effect of **Scent**

Output 15.17 shows the Bonferroni adjusted pairwise comparisons for the main effect of **Scent**. Again, there is redundancy in this table and I'll interpret the highlighted rows, which show the positive differences (i.e., where the scent in the first column has a larger mean than the one in the second). When no scent was used dogs vocalized significantly more than when a human scent was used, $\bar{X}_{\text{Difference}} = 0.85$ [0.39, 1.30], $p < 0.001$, but not when fox scent was used, $\bar{X}_{\text{Difference}} = 0.04$ [−0.46, 0.54], $p = 1$. When fox scent was used dogs vocalized significantly more than when a human scent was used, $\bar{X}_{\text{Difference}} = 0.81$ [0.37, 1.25], $p < 0.001$.

Pairwise Comparisons

Measure: MEASURE_1

(I) Scent	(J) Scent	Mean Difference (I–J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	95% Confidence Interval for Difference[b] Upper Bound
1	2	.847*	.184	<.001	.391	1.303
	3	.040	.202	1.000	-.460	.540
2	1	-.847*	.184	<.001	-1.303	-.391
	3	-.807*	.178	<.001	-1.247	-.366
3	1	-.040	.202	1.000	-.540	.460
	2	.807*	.178	<.001	.366	1.247

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Output 15.17

15.15.3 The **Entity** × **Scent** interaction Ⓑ

The type of scent used interacted significantly with the entity that was sniffed (Output 15.13). This effect means that the degree to which vocalizations were affected by the entity being sniffed depended on what scent was applied to them. Put the other way around, the degree to which vocalizations were affected by the scent depended on the entity to which it was applied. This significant interaction has implications for the main effects we've just discussed. For example, the main effect of **Entity** that we discussed in Section 15.15.1 ignores the type of scent applied to the entity. It compares means for human, shapeshifter and alien entities calculated as though different scents were applied. For example, we saw in Eq. 15.26 how to compute the mean vocalizations for human entities regardless of the scent they wore. The interaction tells us that the mean for humans will have been affected differently than the mean for, say, shapeshifters. In other words, the means tested by the main effect of entity

Estimates

Measure: MEASURE_1

Entity	Scent	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	2.980	.153	2.674	3.286
	2	4.160	.201	3.757	4.563
	3	7.320	.281	6.756	7.884
2	1	10.420	.394	9.628	11.212
	2	8.780	.349	8.078	9.482
	3	8.840	.343	8.150	9.530
3	1	12.060	.428	11.200	12.920
	2	9.980	.389	9.198	10.762
	3	9.180	.341	8.494	9.866

Output 15.18

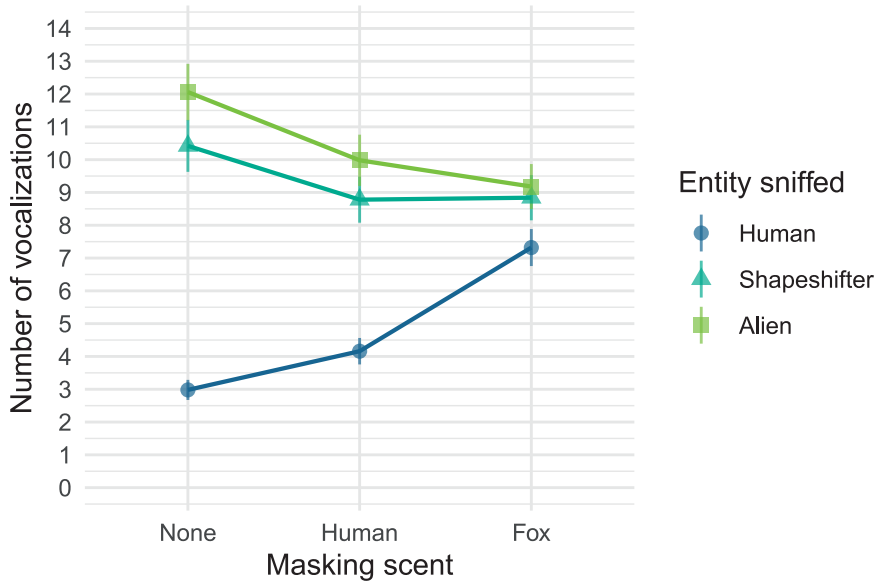


Figure 15.21 Interaction plot for the **Entity** × **Scent** interaction

will have been influenced by the scent worn, so this main effect cannot be interpreted unambiguously. The same applies to the main effect of the scent used. In short, the main effects are uninformative given the significant interaction effect.

Let's tease apart the interaction effect. As a starting point, let's look at the means that we requested in Section 15.14.3 and are shown in Output 15.18. This table is essentially the same as the initial descriptive statistics in Output 15.11, except that the standard errors are displayed rather than the standard deviations. Figure 15.21 plots these means. We can unpick this plot by thinking about simple effects. First, let's view it in terms of the effect of **Scent** within each entity. From this perspective we're comparing the lines (which show the trends across scents for each entity) and the plot shows:

- Vocalizations are always fewer when sniffing humans (the blue line and circles are all below the other lines and shapes).
- When sniffing humans, vocalizations increase a little when the human has human scent compared to no scent, but increase a lot when a fox scent is applied. Basically, the blue line slopes upwards, showing increasing vocalizations as we move from no scent to human scent to fox scent.
- When sniffing aliens (both shapeshifted and normal ones), vocalizations *decrease* quite a lot when a human scent is used compared to no scent, but there isn't much subsequent decrease when a fox scent is applied. Broadly speaking, the light and dark green lines slope downwards, showing decreasing vocalizations as we move from no scent to human scent to fox scent.

The corresponding simple effects of **Scent** within each entity are in Output 15.19, which shows that the effect of **Scent** is significant when sniffing humans, $F(2, 48) = 163.53$, $p < 0.001$, $\eta_p^2 = 0.87$, shapeshifters, $F(2, 48) = 10.76$, $p < 0.001$, $\eta_p^2 = 0.31$, and aliens, $F(2, 48) = 24.46$, $p < 0.001$, $\eta_p^2 = 0.51$.

Multivariate Tests

Entity		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
1	Pillai's trace	.872	163.530[a]	2.000	48.000	<.001	.872
	Wilks' lambda	.128	163.530[a]	2.000	48.000	<.001	.872
	Hotelling's trace	6.814	163.530[a]	2.000	48.000	<.001	.872
	Roy's largest root	6.814	163.530[a]	2.000	48.000	<.001	.872
2	Pillai's trace	.310	10.760[a]	2.000	48.000	<.001	.310
	Wilks' lambda	.690	10.760[a]	2.000	48.000	<.001	.310
	Hotelling's trace	.448	10.760[a]	2.000	48.000	<.001	.310
	Roy's largest root	.448	10.760[a]	2.000	48.000	<.001	.310
3	Pillai's trace	.505	24.459[a]	2.000	48.000	<.001	.505
	Wilks' lambda	.495	24.459[a]	2.000	48.000	<.001	.505
	Hotelling's trace	1.019	24.459[a]	2.000	48.000	<.001	.505
	Roy's largest root	1.019	24.459[a]	2.000	48.000	<.001	.505

Each F tests the multivariate simple effects of Scent within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

Output 15.19

Alternatively, we could look at Figure 15.21 in terms of the simple effect of **Entity** within each scent. From this perspective we're comparing the shapes at each category on the *x*-axis (which show the differences between entities within each scent). The plot shows:

- When no scent is used, there were a lot fewer vocalizations for humans than for either alien (the circle is a lot lower than the square and triangle), and there were fewer vocalizations for shapeshifted aliens than aliens in lizard form (the triangle is below the square).
- When human scent is used, the pattern is similar to no scent: there were a lot fewer vocalizations for humans than for either alien (the circle is a lot lower than the square and triangle), and there were fewer vocalizations for shapeshifters than normal aliens (the triangle is below the square).
- When fox scent is used, there were still fewer vocalizations for humans than for either alien (the circle is lower than the square and triangle) but the effect is much diminished compared to the other scents. The vocalizations for shapeshifters and normal aliens now seem really similar (the triangle and square are in a very similar location).

The corresponding simple effects of **Entity** within each scent are in Output 15.20, which shows that the effect of **Entity** is significant when no scent is used, $F(2, 48) = 383.12$, $p < 0.001$, $\eta_p^2 = 0.94$, when human scent is used, $F(2, 48) = 130.59$, $p < 0.001$, $\eta_p^2 = 0.85$, and when fox scent is used, $F(2, 48) = 18.71$, $p < 0.001$, $\eta_p^2 = 0.44$. Note that although the effects are all significant, the effect sizes show that the effect of **Entity** gets weaker as we move from no scent to human scent to fox scent.

Multivariate Tests

Scent		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
1	Pillai's trace	.941	383.120[a]	2.000	48.000	<.001	.941
	Wilks' lambda	.059	383.120[a]	2.000	48.000	<.001	.941
	Hotelling's trace	15.963	383.120[a]	2.000	48.000	<.001	.941
	Roy's largest root	15.963	383.120[a]	2.000	48.000	<.001	.941
2	Pillai's trace	.845	130.588[a]	2.000	48.000	<.001	.845
	Wilks' lambda	.155	130.588[a]	2.000	48.000	<.001	.845
	Hotelling's trace	5.441	130.588[a]	2.000	48.000	<.001	.845
	Roy's largest root	5.441	130.588[a]	2.000	48.000	<.001	.845
3	Pillai's trace	.438	18.709[a]	2.000	48.000	<.001	.438
	Wilks' lambda	.562	18.709[a]	2.000	48.000	<.001	.438
	Hotelling's trace	.780	18.709[a]	2.000	48.000	<.001	.438
	Roy's largest root	.780	18.709[a]	2.000	48.000	<.001	.438

Each F tests the multivariate simple effects of Entity within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

Output 15.20

Instead of using simple effects we could use contrasts. Remember we set up simple contrasts for our two predictors and these result in four contrasts for the interaction term, shown back in Output 15.15. To help us to interpret these contrasts, Figure 15.22 separates the interaction plot in Figure 15.21 into four panels, each of which relates to a contrast. The first contrast for the interaction looks at level 2 of **Entity** (shapeshifter) compared to level 1 (human), when a human scent (level 2) is used compared to no scent (level 1). This contrast is significant, $F(1, 49) = 39.33$, $p < 0.001$, $\eta_p^2 = 0.45$. Figure 15.22 (top left) shows this contrast: the significance means that the distance between the lines when no scent is used is greater than the distance between the lines when a human scent is used. In other words, the difference in the number of vocalizations (which reflect a difference when

dogs sniff shapeshifters compared to humans) is more pronounced when no scent is used compared to human scent. Basically, dogs' abilities to differentiate shapeshifters from humans are reduced by applying a human scent.

Contrast 2 looks at level 2 of **Entity** (shapeshifter) compared to level 1 (human), when a fox scent (level 3) is used compared to no scent (level 1). This contrast is significant, $F(1, 49) = 136.80$, $p < 0.001$, $\eta_p^2 = 0.74$. Figure 15.22 (top right) shows this contrast: the significance means that the distance between the lines when no scent is used is greater than the distance between the lines when fox scent is used. In other words, the difference in the number of vocalizations (which reflect a difference when dogs sniff shapeshifters compared to humans) is more pronounced when no scent is used compared to fox scent. Dogs' abilities to differentiate shapeshifters from humans are reduced by applying a fox scent.

Contrast 3 looks at level 3 of **Entity** (alien) compared to level 1 (human), when a human scent (level 2) is used compared to no scent (level 1). This contrast is significant, $F(1, 49) = 49.73$, $p < 0.001$, $\eta_p^2 = 0.50$. Figure 15.22 (bottom left) shows this contrast: the significance means that the distance between the lines when no scent is used is greater than the distance between the lines when a human scent is used. In other words, the difference in the number of vocalizations (which reflect a difference when dogs sniff an alien that looks like a lizard compared to humans) is more pronounced when no scent is used compared to human scent. Basically, dogs' abilities to differentiate aliens from humans are reduced by applying a human scent.

Contrast 4 looks at level 3 of **Entity** (alien) compared to level 1 (human), when a fox scent (level 3) is used compared to no scent (level 1). This contrast is significant, $F(1, 49) = 231.96$, $p < 0.001$, $\eta_p^2 = 0.83$. Figure 15.22 (bottom right) shows this contrast: the significance means that the distance between the lines when no scent is used is greater than the distance between the lines when fox scent is used. In other words, the difference in the number of vocalizations (which reflect a difference when dogs sniff aliens compared to humans) is more pronounced when no scent is used compared to fox scent. Dogs' abilities to differentiate aliens from humans are reduced by applying a fox scent.

These contrasts tell us nothing about the differences between the shapeshifter and alien entities or between the human scent and fox scent. However, the general pattern seems to be that applying scents does inhibit the dogs' abilities to differentiate both types of aliens from humans. However, the fact that the means are always lower for humans (the circles are always below the other shapes in Figure 15.22) and that their confidence intervals don't overlap with those for aliens suggests that although the dogs' abilities are dampened they can still successfully differentiate aliens (of both types) from humans.

A final option if we hadn't had specific hypotheses would be to look at *post hoc* tests for the interaction. There will be two tables of tests, one of which contains *post hoc* tests for **Scent** within each level of **Entity** (Output 15.21) while the other contains *post hoc* tests for **Entity** within each level of **Scent** (Output 15.22). As when we looked at the main effects, there is redundancy in these tables, and I've highlighted the rows that show unique effects. Again, I've highlighted the positive differences (i.e., where the group in the first column has a larger mean than the one in the second). Effects that are significant are highlighted in yellow, those that are non-significant are in lilac.

Starting with Output 15.21, within entity 1 (humans) all scents result in significantly different numbers of vocalizations. Looking back at Figure 15.21 this means that the circles are in significantly different locations: there were significantly more vocalizations when human and fox scent was used compared to no scent, and more vocalizations when fox scent was used than human scent. For humans, the scents did increase the vocalizations with fox scent having the largest impact. Within entity 2 (shapeshifters) both human and fox scent significantly decreased vocalizations compared to no scent, but fox and human scents did not differ. In Figure 15.21 this means that the triangles for fox and human scent are both lower than for no scent, but are not lower than each other. For

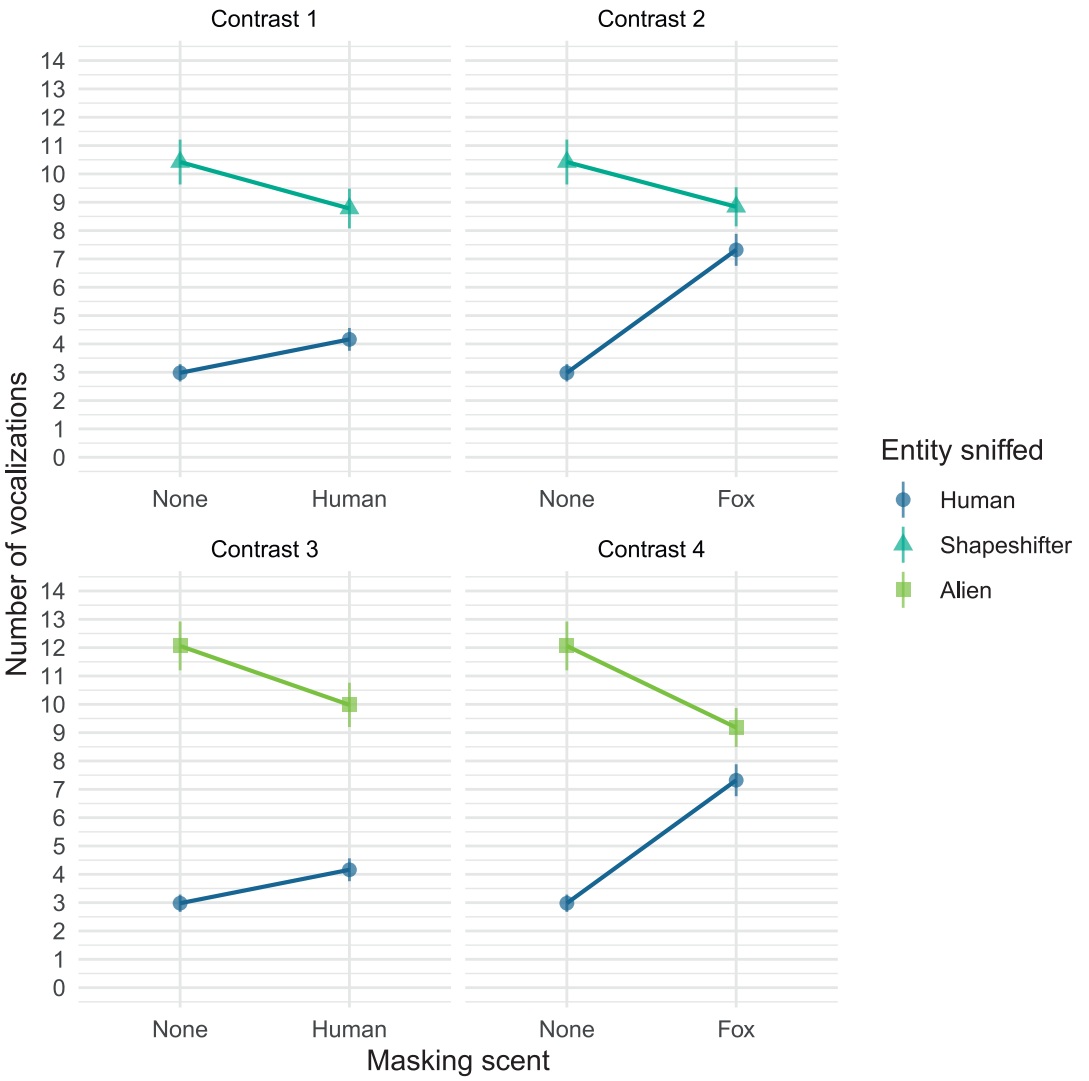


Figure 15.22 Graphs (generated using R, incidentally) illustrating the four contrasts in the sniffer dog analysis

shapeshifters, the scents did reduce the vocalizations compared to no scent, but neither human nor fox scent was more effective. Within entity 3 (aliens) the same pattern emerges: both human and fox scent significantly decreased vocalizations compared to no scent, but fox and human scents did not differ. In Figure 15.21 this means that the squares for fox and human scent are both lower than for no scent, but are not lower than each other. For aliens, the scents did reduce the vocalizations compared to no scent, but neither human nor fox scent was more effective.

Moving on to Output 15.22, within scent 1 (no scent) all entities have significantly different numbers of vocalizations. Looking at Figure 15.21 this means that the circle, triangle and square are in significantly different locations: there were significantly more vocalizations when aliens and shapeshifters were sniffed compared to humans, and more vocalizations when aliens were sniffed in their natural form than when they had shapeshifted. In short, the training of the sniffer dogs worked: they vocalize more when sniffing aliens. Within scent 2 (human scent) all entities have significantly

Pairwise Comparisons

Measure: MEASURE_1

Entity	(I) Scent	(J) Scent	Mean Difference (I–J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
1	1	2	-1.180*	.197	<.001	-1.669	-.691
		3	-4.340*	.242	<.001	-4.939	-3.741
	2	1	1.180*	.197	<.001	.691	1.669
		3	-3.160*	.289	<.001	-3.877	-2.443
	3	1	4.340*	.242	<.001	3.741	4.939
		2	3.160*	.289	<.001	2.443	3.877
2	1	2	1.640*	.383	<.001	.690	2.590
		3	1.580*	.391	<.001	.611	2.549
	2	1	-1.640*	.383	<.001	-2.590	-.690
		3	-.060	.354	1.000	-.937	.817
	3	1	-1.580*	.391	<.001	-2.549	-.611
		2	.060	.354	1.000	-.817	.937
3	1	2	2.080*	.378	<.001	1.143	3.017
		3	2.880*	.422	<.001	1.835	3.925
	2	1	-2.080*	.378	<.001	-3.017	-1.143
		3	.800	.363	.096	-.099	1.699
	3	1	-2.880*	.422	<.001	-3.925	-1.835
		2	-.800	.363	.096	-1.699	.099

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Output 15.21

different numbers of vocalizations. In Figure 15.21 the circle, triangle and square are in significantly different locations: the pattern is the same as when no scent was used. Within scent 3 (fox scent) significantly more vocalizations are made to aliens (of both sorts) than humans, but vocalizations are not significantly different when sniffing aliens that have shapeshifted compared to those that have not. In Figure 15.21 the circle is in a significantly different location than the triangle and square, but the triangle and square are in similar locations. Fox scent prevents dogs from discriminating aliens who have shapeshifted from those that have not, but they can still differentiate them from humans. The invasion will be thwarted by spaniels.

Interpreting interactions is complex, and some well-respected researchers struggle with them, so try not to feel disheartened if you find them hard. Try to be thorough and break down each effect systematically by using a plot and either contrasts, simple effects analysis or *post hoc* tests. On the subject of which, I have explained every conceivable approach to breaking down the interaction, but in real life doing this would result in a huge inflation of the Type I error rate. Although *post hoc* tests are corrected for the number of tests, this applies only within each table. So each of the 18 unique *post hoc* tests in Output 15.21 and Output 15.22 corrected only for the nine tests within the respective table. Similarly, the simple effects are not corrected at all, nor are the simple contrasts (because simple contrasts are not orthogonal). You want to try to break down the interaction with as few tests as possible, and correct them manually for the number of tests you end up conducting.

Pairwise Comparisons

Measure: MEASURE_1

Scent	(I) Entity	(J) Entity	Mean Difference (I–J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
1	1	2	-7.440*	.350	<.001	-8.308	-6.572
		3	-9.080*	.382	<.001	-10.027	-8.133
	2	1	7.440*	.350	<.001	6.572	8.308
		3	-1.640*	.433	.001	-2.714	-.566
	3	1	9.080*	.382	<.001	8.133	10.027
		2	1.640*	.433	.001	.566	2.714
2	1	2	-4.620*	.354	<.001	-5.499	-3.741
		3	-5.820*	.377	<.001	-6.753	-4.887
	2	1	4.620*	.354	<.001	3.741	5.499
		3	-1.200*	.343	.003	-2.050	-.350
	3	1	5.820*	.377	<.001	4.887	6.753
		2	1.200*	.343	.003	.350	2.050
3	1	2	-1.520*	.378	<.001	-2.458	-.582
		3	-1.860*	.318	<.001	-2.649	-1.071
	2	1	1.520*	.378	<.001	.582	2.458
		3	-.340	.394	1.000	-1.318	.638
	3	1	1.860*	.318	<.001	1.071	2.649
		2	.340	.394	1.000	-.638	1.318

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Output 15.22



Cramming Sam's Tips
Factorial repeated-measures designs

- Two-way repeated-measures designs compare means when there are two predictor/independent variables, and the same entities have been used in all conditions.
- You can test the assumption of *sphericity* when you have three or more repeated-measures conditions with *Mauchly's test*, but a better approach is to routinely interpret *F*-statistics that have been corrected for the amount by which the data are not spherical.
- The table labelled *Tests of Within-Subjects Effects* shows the *F*-statistics and their *p*-values. In a two-way design you will have a main effect of each variable and the interaction between

them. For *each* effect, read the row labelled *Greenhouse–Geisser* (you can also look at *Huynh–Feldt*, but you'll have to read this chapter to find out the relative merits of the two procedures). Assuming you have chosen 0.05 as your criterion, if the value in the column labelled *Sig.* is less than or equal to 0.05 then the effect is significant.

- Break down interactions with one of contrasts, simple effects or *post hoc* tests. Assuming you have chosen 0.05 as your criterion, if the values in the column labelled *Sig.* are less than or equal to 0.05 the test is significant.

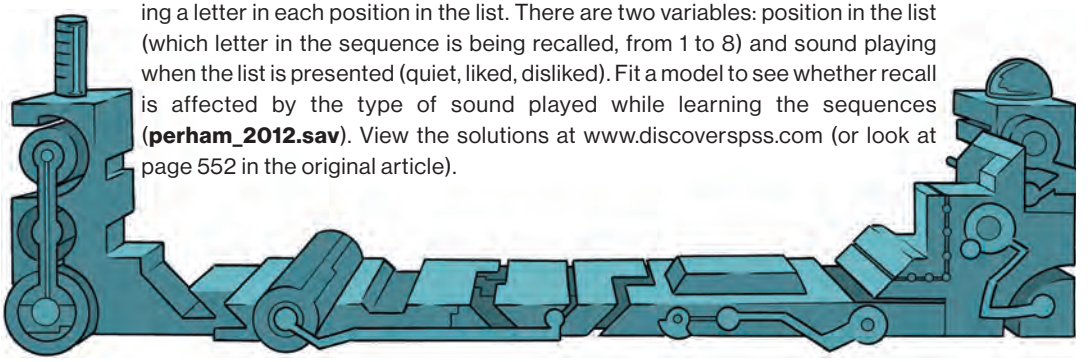




Labcoat Leni's Real Research 15.1
Are splattered cadavers distracting? Ⓑ

Perham, N., & Sykora, M. (2012). *Applied Cognitive Psychology*, *26*(4), 550–555.

In Chapter 10, I used the example of whether listening to my favourite music would interfere with people's ability to write an essay. It turns out that Nick Perham has tested this hypothesis (sort of). He was interested in the effects of liked and disliked music (compared to quiet) on people's ability to remember things. Twenty-five participants remembered lists of eight letters. Perham and Sykora (2012) manipulated the background noise while each list was presented: silence (the control), liked music or disliked music. They used music that they believed most participants would like (a popular song called 'From Paris to Berlin' by Infernal) and dislike (songs such as Repulsion's 'Acid Bath', 'Eaten Alive' and 'Splattered Cadavers' – in other words, the sort of thing I listen to, although I don't actually have any stuff by Repulsion). Participants recalled each list of eight letters, and the authors calculated the probability of correctly recalling a letter in each position in the list. There are two variables: position in the list (which letter in the sequence is being recalled, from 1 to 8) and sound playing when the list is presented (quiet, liked, disliked). Fit a model to see whether recall is affected by the type of sound played while learning the sequences (**perham_2012.sav**). View the solutions at www.discoverspss.com (or look at page 552 in the original article).



15.16 Reporting the results from factorial repeated-measures designs Ⓑ

We've got three effects to report and we need to report corrected degrees of freedom for each, and these effects might have different degrees of freedom. We can then report the three effects as follows:

- ✓ Unless otherwise stated, $p < 0.001$. For the main effect of entity, the Greenhouse–Geisser estimate of the departure from sphericity was $\varepsilon = 0.99$. This main effect was a significant, $F(1.98, 96.88) = 315.95$, $\eta_p^2 = 0.87$. Contrasts revealed that, compared to when sniffing humans, dogs vocalized significantly more when sniffing a shapeshifted alien, $F(1, 49) = 410.68$, $\eta_p^2 = 0.89$, and an alien in its natural form, $F(1, 49) = 543.12$, $\eta_p^2 = 0.92$.
- ✓ For the main effect of scent, the Greenhouse–Geisser estimate of the departure from sphericity was $\varepsilon = 0.98$. The type of scent also had a significant effect, $F(1.95, 95.75) = 12.91$, $\eta_p^2 = 0.21$. Contrasts revealed that, compared to when sniffing an entity wearing no scent, dogs vocalized significantly less when a human scent was used, $F(1, 49) = 21.18$, $\eta_p^2 = 0.30$, but made a similar number of vocalizations when fox scent was used, $F(1, 49) = 0.04$, $p = 0.843$, $\eta_p^2 = 0.00$.
- ✓ For the interaction, the Greenhouse–Geisser estimate of the departure from sphericity was $\varepsilon = 0.92$. There was a significant interaction effect of the entity sniffed and the type of scent used on the number of vocalizations, $F(3.66, 179.27) = 60.26$, $\eta_p^2 = 0.55$. To break down this interaction, contrasts compared all entities to their baseline (human) and all scents to their baseline (no scent). These contrasts revealed significant interactions when comparing (1) shapeshifters to humans, when a human scent was used compared to no scent, $F(1, 49) = 39.33$, $\eta_p^2 = 0.45$; (2) shapeshifters to humans, when a fox scent was used compared to no scent, $F(1, 49) = 136.80$, $\eta_p^2 = 0.74$; (3) aliens to humans, when a human scent was used compared to no scent, $F(1, 49) = 49.73$, $\eta_p^2 = 0.50$; and (4) aliens to humans, when a fox scent was used compared to no scent, $F(1, 49) = 231.96$, $\eta_p^2 = 0.83$. To sum up, it seems as though the sniffer-dog training was successful in that dogs could differentiate humans from both types of aliens and, although their ability was diminished by the use of human and (especially) fox scents, the scents did not prevent the dogs from vocalizing more when sniffing aliens.

15.17 Jane and Brian's story Ⓐ

The visit to the hospital had shaken Jane. Brian had offered to go with her, but the last thing she needed was a lecture from her mother on how distracting guys were to her studies. Seeing her mother unwell and her father so lost without her had been strange. Their emotional wall was so perfectly formed that it had never occurred to her that her parents might be deeply connected. Her dad was in pieces mentally, and her mum physically. They'd always seemed superhuman, like their intellect could defeat anything. She'd spent so many years trying to be what she thought they were, but seeing them bought home how little she knew them. She had become obsessed with intellect for the sake of intellect. But seeing her frail mum made her feel helpless. All the knowledge Jane had, and she didn't know how to help her. Maybe instead of trying any means to become cleverer, she should put her brainpower to some use in the world.

She was thinking this as she climbed the steps to the basement door, the taste of formaldehyde fresh on her lips. Was this what she became when she spent too much time alone? She locked the door. It felt like it might be for the last time. It had been a long day. Her legs took her to Brian's apartment. She went in, but was too exhausted for a deep conversation. 'Talk stats,' she said, and relaxed into the sofa as he recited what he'd recently read on repeated-measures designs.

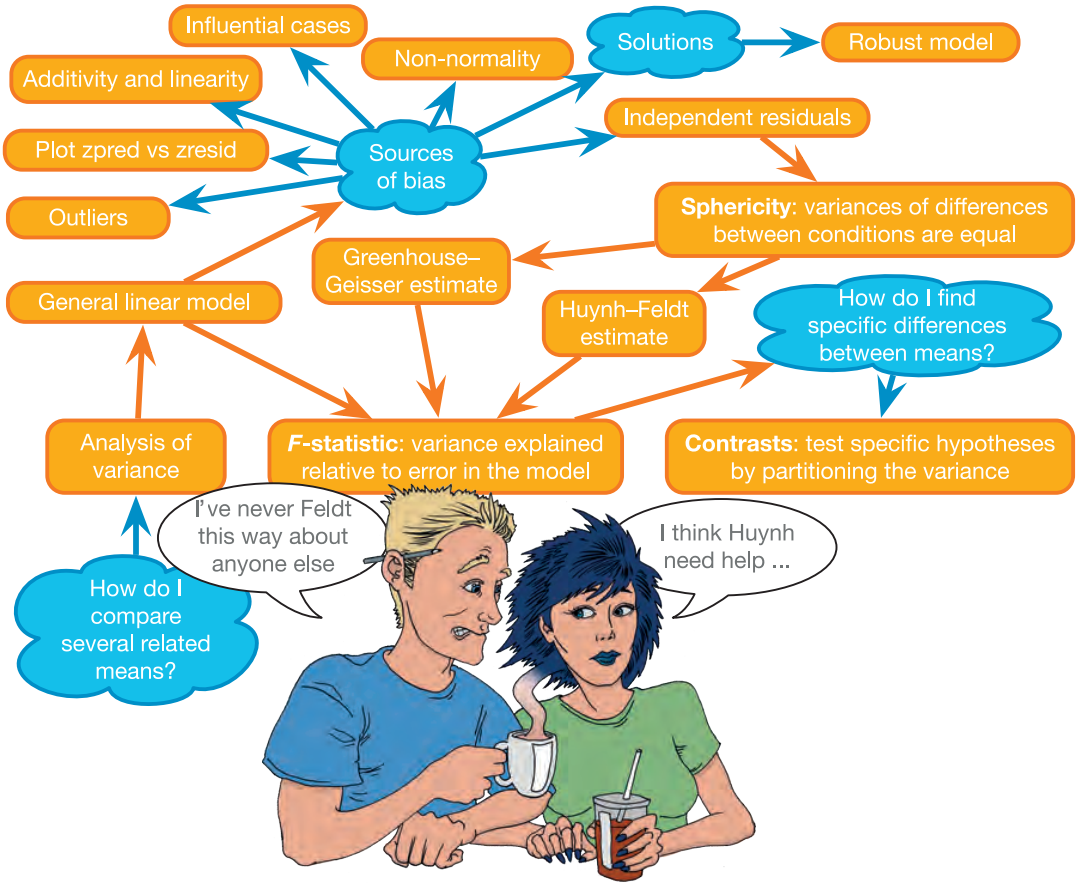


Figure 15.23 What Brian learnt from this chapter

15.18 What next? Ⓑ

By the age of 16 I had started my first 'serious' band. We stayed together for about 7 years (with the same line-up, and we're still friends now) before Mark (drummer) moved to Oxford, I moved to Brighton to do my PhD, and rehearsing became a mammoth feat of organization. We had a track on a CD, some radio play and transformed from a thrash metal band to a blend of Fugazi, Nirvana and metal. I never split my trousers during a gig again (although I did once split my head open). Why didn't we make it? Well, Mark is an astonishingly good drummer so it wasn't his fault, the other Mark was an extremely good bassist too, and so all logic points towards the weak link being me. This fact was especially unfortunate, given that I had three roles in the band (guitar, singing, songs) – my poor band mates never stood a chance.☺ I stopped playing music for quite a few years after we split. I still wrote songs (for personal consumption) but the three of us were such close friends that I couldn't bear the thought of playing with other people. At least not for a few years …

15.19 Key terms that I've discovered

Compound symmetry
Greenhouse–Geisser estimate
Huynh–Feldt estimate
Lower-bound estimate
Mauchly's test
Repeated-measures ANOVA
Sphericity



Smart Alex's tasks

- **Task 1**: It is common that lecturers obtain reputations for being 'hard' or 'light' markers, but there is often little to substantiate these reputations. A group of students investigated the consistency of marking by submitting the same essays to four different lecturers. The outcome was the percentage mark given by each lecturer and the predictor was the lecturer who marked the report (**tutor_marks.sav**). Compute the *F*-statistic for the effect of marker by hand. Ⓑ
- **Task 2**: Repeat the analysis for Task 1 using SPSS and interpret the results. Ⓑ
- **Task 3**: Calculate the effect sizes for the analysis in Task 1. Ⓑ
- **Task 4**: In the previous chapter we came across the beer-goggles effect. In that chapter, we saw that the beer-goggles effect was stronger for unattractive faces. We took a follow-up sample of 26 people and gave them doses of alcohol (0 pints, 2 pints, 4 pints and 6 pints of lager) over four different weeks. We asked them to rate a bunch of photos of unattractive faces in either dim or bright lighting. The outcome measure was the mean attractiveness rating (out of 100) of the faces and the predictors were the dose of alcohol and the lighting conditions (**goggles_lighting.sav**). Do alcohol dose and lighting interact to magnify the beer goggles effect? Ⓑ
- **Task 5**: Interpret the simple effect of effect of alcohol at different levels of lighting. Ⓒ
- **Task 6**: Early in my career I looked at the effect of giving children information about entities. In one study (Field, 2006), I used three novel entities (the quoll, quokka and cuscus) and children were told negative things about one of the entities, positive things about another, and given no information about the third (our control). After the information I asked the children to place their hands in three wooden boxes each of which they believed contained one of the aforementioned entities (**field_2006.sav**). Draw an error bar graph of the means interpret a Q-Q plot. Ⓐ
- **Task 7**: Log-transform the scores in Task 7, make a Q-Q plot of the transformed scores and interpret it. Ⓑ
- **Task 8**: Analyse the data in Task 7 with a robust model. Do children take longer to put their hands in a box that they believe contains an entity about which they have been told nasty things? Ⓑ

View the solutions at www.discoverspss.com



GLM 5: MIXED DESIGNS

16.1 What will this chapter tell me? 752
16.2 Mixed designs 753
16.3 Assumptions in mixed designs 753
16.4 A speed-dating example 754
16.5 Mixed designs using SPSS 756
16.6 Output for mixed factorial designs 762
16.7 Reporting the results of mixed designs 776
16.8 Jane and Brian's story 779
16.9 What next? 780
16.10 Key terms that I've discovered 781
Smart Alex's tasks 781

Letters A to D indicate increasing difficulty

16.1 What will this chapter tell me? Ⓐ

Most teenagers have anxiety and depression, but I probably had more than my fair share. The all-boys' grammar school that I attended had feasted on me like a parasitic leech. By the time I left school I was a terrified husk of a person. If I had my bandmates with me I had no real problem with playing guitar and 'singing' at people – I could even tolerate having my photo taken (Figure 16.1) – but interacting with strangers was another matter entirely. It still is. In the band I felt at ease, in the real world not so much.

Still, you can't hide away in your bedroom with your Iron Maiden albums for ever, and eventually I tried to integrate with society. Between the ages of 16 and 18 there was the sudden presence of girls in my social circle. I hadn't seen a girl since Clair Sparks, which was about 7 years earlier. Girls were particularly problematic for the teenage me because not only were you expected to talk to them, but what you said had to be impressive because then they might become your 'girlfriend', and I thought at the time that having a girlfriend would mean I felt good about myself. It turns out that it's more complicated than that. The main problem with talking to girls in 1990 (other than the incapacitating anxiety) was that hardly any of them liked to talk about Iron Maiden, and I didn't know what else to talk about. I still don't. Speed dating[1] didn't exist back then, but if it had I would have run a mile from it: it would have been a manifestation of hell on earth for me.



Figure 16.1 Me (far left) looking awkward with my handsome bandmates

I would have found it too chilling to contemplate entering a highly pressured social situation where I *had* to think of something witty and amusing to say or face

1 With the advent of Tinder, or whatever the latest dating app is, I suspect that speed dating isn't a thing anymore. To be honest, having become obsolete to the universe, I have no idea what is and is not a thing these days. Anyway, on a speed dating night the rough idea is that people wanting to find a romantic partner turn up at a venue where there will be lots of other people who also want a romantic partner. One-half of the group sit individually at small tables and the remainder choose a table, get 3 minutes to impress the other person at the table with tales of heteroscedastic data, and then a bell rings and they get up and move to the next table. Having worked around all the tables, the end of the evening is spent either locating the person who made you gooey inside or avoiding the person who was going on about heterosomethingorother.

3 minutes in awkward silence. I feel much the same about giving academic talks: there's only so much disappointment in people's eyes that I can take. I pretty much don't do them any more. Anyway, perhaps as self-help, or something, this chapter is about speed dating – oh, and mixed designs, but if I mention that you'll move swiftly on to the next chapter when the bell rings.

16.2 Mixed designs Ⓑ



If you thought that the previous chapter was bad, I'm about to throw an added complication into the mix. This chapter looks at situations where we combine repeated-measures and independent designs. As if this wasn't bad enough, I'm going to use this as an excuse to show you a design with three independent variables (at this point imagine me leaning back in my chair, and laughing like the Joker). When a design includes some independent variables that were measured using different entities and others that used repeated measures it is called a **mixed design**. It might be obvious that a mixed design requires at least two independent variables, but you can have more complex scenarios too (e.g., two independent measures and one repeated measure, one independent measure and two repeated measures, or even two of each). Because by adding independent variables we're adding predictors to the linear model, you can have virtually any number of independent variables if your sample size is big enough. However, as we shall see, interaction terms are very difficult to interpret with even three independent variables, so anything more than that isn't good for your mental health and is best avoided.[2]

We're still essentially using the linear model, which you might be sick of hearing about by now, so I'm going to dodge the theory (i.e., it's too complicated for me to understand) and assume that you can extend what we've learnt to the more complex situation of having three independent variables. Like in the last chapter, because there are repeated measures involved, people typically use an 'ANOVA-style' model. That is, rather than using a flexible multilevel model to incorporate scores within individuals, they use a simpler, less flexible model that requires certain assumptions (sphericity). Therefore, the variety of linear model we'll use is sometimes known as **mixed ANOVA**.

16.3 Assumptions in mixed designs Ⓑ

If you have read any of the previous chapters on comparing means you will be sick of me writing that we're using the linear model and so all the sources of potential bias (and counteractive measures) discussed in Chapter 6 apply. But, there you go, I've just written it again. Of course, because mixed designs include both repeated measures and independent measures you have the double whammy of concerning yourself with both homogeneity of variance *and* sphericity. It's enough to make you guzzle the ink from the octopus of inescapable despair. But don't: we'll apply the Greenhouse–Geisser correction and forget about sphericity.

The various other woes in Chapter 6 are more troublesome. As we saw in the previous chapter, the Bootstrap... button is absent for repeated-measures designs. 'What about non-parametric tests?' you might ask. You wouldn't be alone: if I had £1 (or whatever currency you fancy) for every time someone asked me what the non-parametric equivalent of mixed ANOVA was, then I'd have a

2 Fans of irony will enjoy the four-way ANOVAs that I conducted in various papers before I knew better (Field & Davey, 1999; Field & Moore, 2005).

shiny new drum kit. The short answer is, there isn't one. There is a robust model for mixed designs involving two independent variables that can be done using R (Wilcox, 2017), but it is not currently implemented in the *WRS2* package, which makes it tricky for me to demonstrate. So, the options are (1) learn R; (2) try out some of the stuff in Chapter 6 to shoehorn your data into a reasonable shape; or (3) stick an oxygen tank on your back and start swimming in the sea looking for that octopus ...

16.4 A speed-dating example Ⓑ

Lots of magazines go on about relationships. The big topic seems to be how to get a relationship in the first place, and within that lots of discussion of the relative importance of looks, personality and dating strategies. Scientists have looked at these issues too. For example, the top three most highly rated attributes of a partner in teenagers are reliability, honesty and kindness (Ha et al., 2010). (Shocked that knowledge of how to deal with heteroscedastic residuals wasn't in the list, but that's teenagers for you.) Beyond that, in the same study, those identifying as boys tended to rate attractiveness slightly higher than those identifying as girls, and girls rate a sense of humour more highly than boys (although both are ranked in the top 10 by boys and girls). With regard to dating strategies, some research suggests that if someone is committed to pursuing a relationship with a person who 'plays hard to get', they will find that person more desirable but less likeable (Dai et al., 2014).

Imagine a scientist designed a study to look at the interplay between looks, personality and dating strategies on evaluations of a date. She set up a speed-dating night with nine tables at which there sat a 'date'. All the dates were actors selected to vary in their attractiveness, which is subjective, but for the purpose of the study, the actors were split into three categories: highly attractive, average and unattractive), their charisma (high charisma, average charisma, writes statistics books), and also the strategy they were told to employ during the conversation (acting 'normally' or 'playing hard to get'). The dates were trained before the study so that they could act (1) charismatically to varying degrees; and (2) aloof or unobtainable (hard to get) or not. As such, across the nine dates/actors there were three 'attractive' people, one of whom acted charismatically, another acted in an average manner and a third acted like a bore, and likewise for the three average-looking dates and the three unattractive dates. Therefore, each participant attending a speed-dating night would be exposed to all combinations of attractiveness and charisma (these are repeated measures).[3] Upon arrival participants were randomly assigned a blue or red sticker. For the participants with the red sticker the actors played hard to get (unobtainable) and for those with a blue sticker they acted normally. Over the course of a few nights 20 people attended, spent 5 minutes with each of the nine 'dates' and then rated how much they'd like to have a proper date with the person as a percentage (100% = 'I'd pay a large sum of money for their phone number', 0% = 'I'd pay a large sum of money for a plane ticket to get me as far away from them as possible'). Having just read the paper by Benjamin et al. (2018), the researchers decide to use an alpha of 0.005 instead of the more common 0.05. Therefore, effects will be deemed significant if $p \leq 0.005$.

To be clear, each participant rated nine different people who varied in their attractiveness and charisma. These are two repeated-measures variables: **looks** (with three levels because the 'date' could be attractive, average or unattractive) and **charisma** (with three levels because the person could act with high charisma, with some charisma or like a bore). In addition the 'date' employed a 'hard to get' strategy for half of the participants and acted normally for the rest, so we can include **strategy** as a between-group variable. The data are in **speed_date.sav** (Table 16.1).

3 There was a set of nine male stooges and nine females, so that those attending could meet 'dates' of whichever gender interested them.

Table 16.1 Data from a fictional speed dating study

Charisma	High Charisma			Some Charisma			Bore		
Looks	Attractive	Average	Unattractive	Attractive	Average	Unattractive	Attractive	Average	Unattractive
Strategy									
Hard to get	86	84	67	88	69	50	97	48	47
	91	83	53	83	74	48	86	50	46
	89	88	48	99	70	48	90	45	48
	89	69	58	86	77	40	87	47	53
	80	81	57	88	71	50	82	50	45
	80	84	51	96	63	42	92	48	43
	89	85	61	87	79	44	86	50	45
	100	94	56	86	71	54	84	54	47
	90	74	54	92	71	58	78	38	45
	89	86	63	80	73	49	91	48	39
Normal	89	91	93	88	65	54	55	48	52
	84	90	85	95	70	60	50	44	45
	99	100	89	80	79	53	51	48	44
	86	89	83	86	74	58	52	48	47
	89	87	80	83	74	43	58	50	48
	80	81	79	86	59	47	51	47	40
	82	92	85	81	66	47	50	45	47
	97	69	87	95	72	51	45	48	46
	95	92	90	98	64	53	54	53	45
	95	93	96	79	66	46	52	39	47

16.5 Mixed designs using SPSS Ⓑ

The general procedure for mixed designs is the same as for any other linear model (see Chapter 9). Figure 16.2 shows a simpler overview that highlights some of the specific issues when using a mixed design.

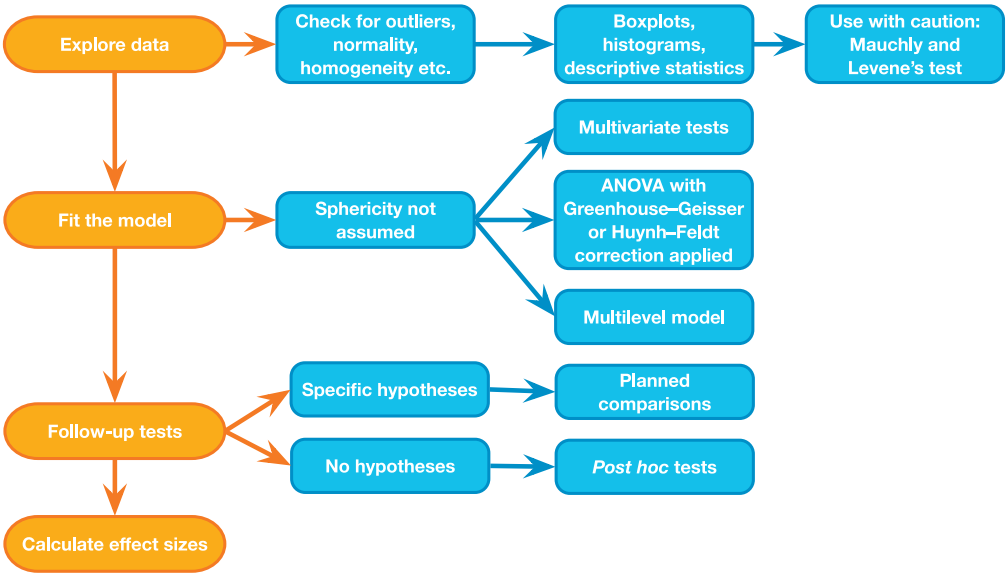


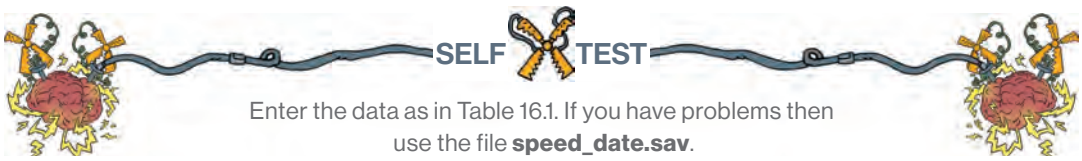
Figure 16.2 The process for analysing mixed designs

16.5.1 Entering data Ⓑ

We enter these data in the same way as the previous chapter. Remember that each row in the data editor represents a single participant and levels of repeated-measures variables are placed in columns. In this experiment there are nine experimental conditions and so the data need to be entered in nine columns (the format is identical to Table 16.1). You will also need to create a coding variable to enter values for the dating strategy employed by the 'dates'.



In the data editor create nine variables with the names and variable labels given in Figure 16.3. Create a variable **strategy** with value labels 0 = normal, 1 = hard to get.



Enter the data as in Table 16.1. If you have problems then use the file **speed_date.sav**.

Variable name	Variable label
att_high	Attractive and highly charismatic
av_high	Average and highly charismatic
ug_high	Ugly and highly charismatic
att_some	Attractive and some charisma
av_some	Average and some charisma
ug_some	Ugly and some charisma
att_none	Attractive and a dullard
av_none	Average and a dullard
ug_none	Ugly and a dullard

Figure 16.3 Variable names and labels

16.5.2 Fitting the model Ⓑ

Select *Analyze* > *General Linear Model* > Repeated Measures... to access the *Define Factor(s)* dialog box in Figure 16.4. As in the previous chapter we first name our repeated-measures variables and specify how many levels they have. We have two repeated measures: **Looks** (attractive, average and unattractive) and **Charisma** (high charisma, some charisma and bore). In the dialog box type 'Looks' into the box labelled *Within-Subject Factor Name* and '3' into the box labelled *Number of Levels*. Click Add to register this variable (it appears as *Looks(3)*), then type 'Charisma' into the box labelled *Within-Subject Factor Name* and '3' into the box labelled *Number of Levels*. Click Add and *Charisma(3)* will appear in the list (see Figure 16.4). Click Define to go to the main dialog box.

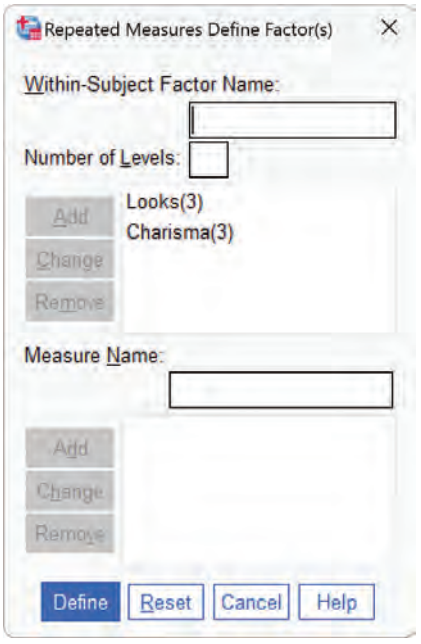


Figure 16.4 The dialog box for defining repeated measures

The main dialog box in Figure 16.5 looks the same as in the previous chapter. At the top of the *Within-Subjects Variables* box the two variables we just defined (**Looks** and **Charisma)** are listed and underneath is a series of question marks followed by numbers in parentheses. The numbers in parentheses represent the levels of the independent variables – see the previous chapter for a more detailed explanation. We have two repeated-measures independent variables and so there are two numbers in the brackets. The first number refers to levels of the first variable listed above the box (in this case **Looks**) and the second refers to levels of the second variable listed above the box (in this case **Charisma**). As with the other repeated-measures designs we've come across, we assign variables to the question marks. Before we do, we need to think about contrasts.

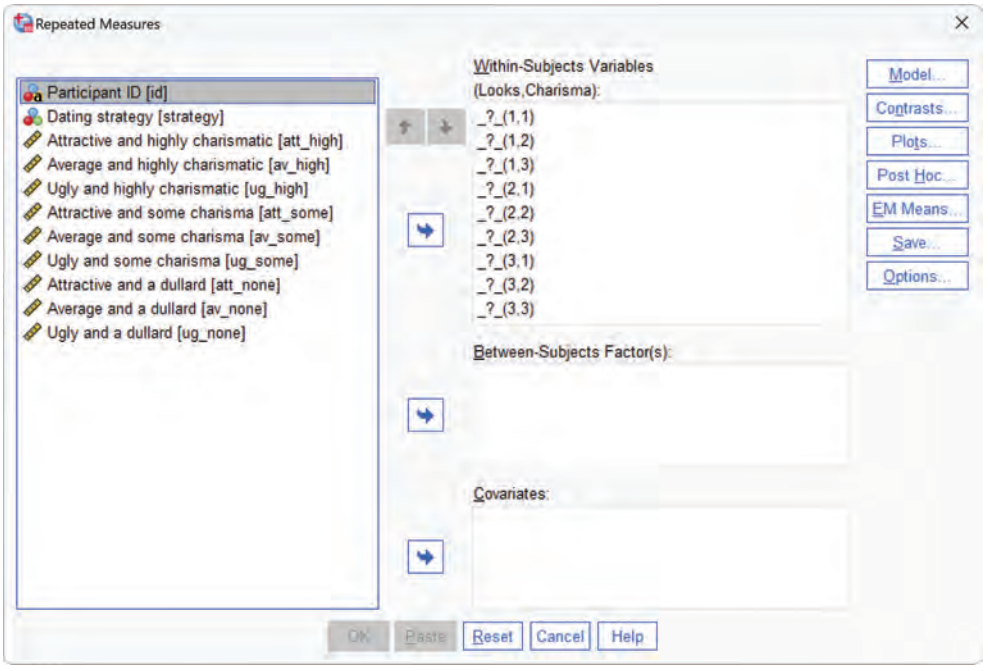


Figure 16.5 The main dialog box for mixed designs before completion

The first variable, **Looks**, has three conditions: attractive, average and unattractive. It makes sense to compare the attractive and unattractive conditions to the average because the average person represents the norm (although it wouldn't be wrong to, for example, compare attractive and average to unattractive). This comparison could be done using a simple contrast (see Table 12.8) if we assign 'average' to the first or last category. **Charisma** also has a category that represents the norm: some charisma. Again, we could use this as a control against which to compare our two extremes (high charisma and acting like a bore). As with **Looks**, we could use a simple contrast to compare everything against 'some charisma' if we assign this category to either the first or last level.

In the previous chapter we assigned our control category for each variable to the first category. To mix it up, for this example we'll assign the control category to the last category; in other words, we'll assign 'average' as level 3 of **Looks** and 'some charisma' as level 3 of **Charisma**. The remaining levels can be assigned arbitrarily but I assigned 'attractive' to level 1 and 'unattractive' to level 2 of **Looks**, and for **Charisma** I assigned 'high charisma' to level 1 and 'none' to level 2. These decisions mean that the variables should be entered as in Figure 16.6. I've deliberately made the order different from how the variables are listed in the data editor to mess with your head. It makes me feel better about the dating ineptitude that the teenage me endured.

att_high	→	_?_(1,1)
att_none	→	_?_(1,2)
att_some	→	_?_(1,3)
ug_high	→	_?_(2,1)
ug_none	→	_?_(2,2)
ug_some	→	_?_(2,3)
av_high	→	_?_(3,1)
av_none	→	_?_(3,2)
av_some	→	_?_(3,3)

Figure 16.6 Variable allocations for the speed-dating data

So far the procedure has been similar to other factorial repeated-measures designs. However, we have a mixed design, so we also need to specify any between-group variables as well. We do this by dragging **strategy** to the box labelled *Between-Subjects Factors* (or click →). The completed dialog box is shown in Figure 16.7.

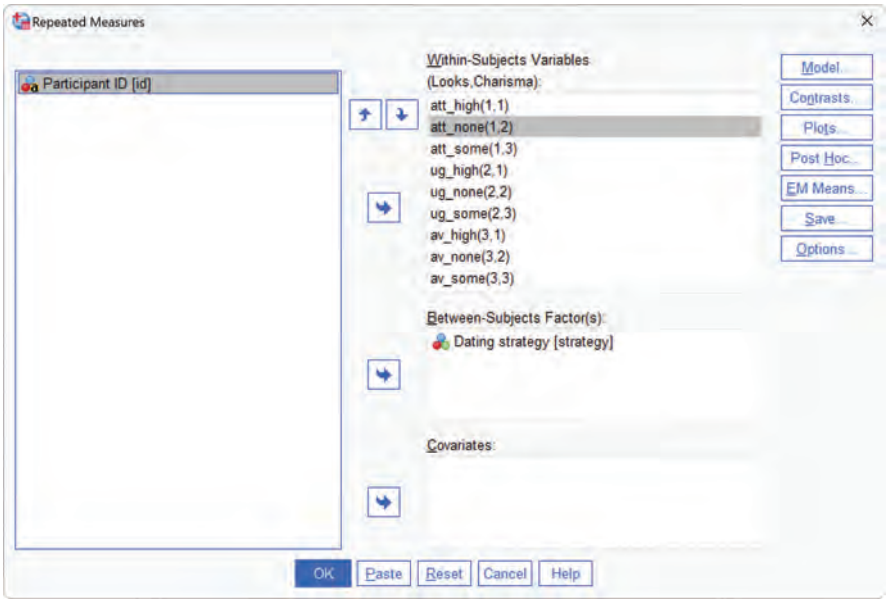


Figure 16.7 The main dialog box for mixed designs after completion

16.5.3 Other options Ⓑ

We've seen in previous chapters that we can enter custom contrast codes using syntax, but we'll use the built-in contrasts instead (see Table 12.8). In the previous section I described why it might be interesting to use the 'average' attractiveness and 'some charisma' as baseline categories for the **Looks** and **Charisma** variables, respectively. We used the *Contrasts* dialog box in Sections 13.6.5 and 15.8.2 and so we can use what we've learned to select a simple contrast for both **Looks** and **Charisma**. Start by clicking Contrasts... to activate the dialog box in Figure 16.8. In both cases, we specified the variables such that the control category was the last one; therefore, we can change the reference category to ◉ Last. Click Continue to return to the main dialog box.

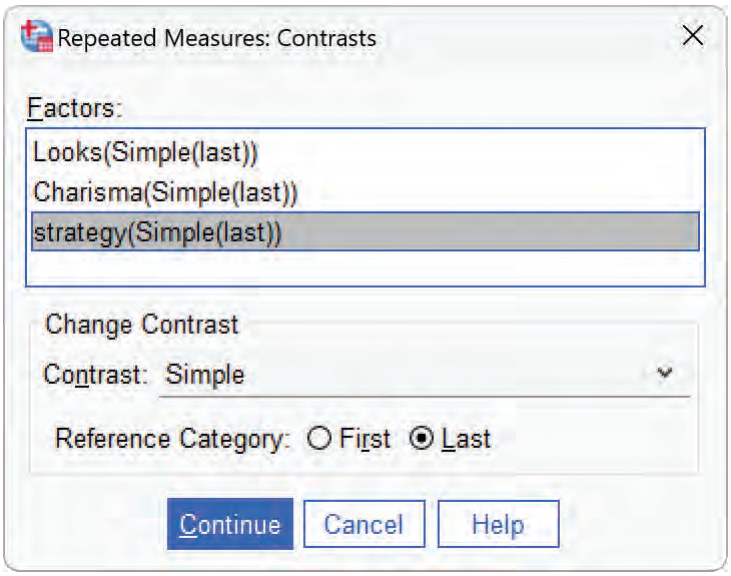


Figure 16.8 The *Contrasts* dialog box

We can plot the **Looks** × **Charisma** × **strategy** interaction effect by clicking Plots... to access the dialog box in Figure 16.9. Drag **Looks** to the slot labelled *Horizontal Axis*, **Charisma** to the slot labelled *Separate Line*, and **strategy** to the slot labelled *Separate Plots*. Click Add to register this plot. Specifying the plot in this way displays the interaction for **Looks** and **Charisma**, but produces separate versions for those whose dates used a 'hard to get' strategy and those whose dates did not. Choose a *Line Chart* and *Include Error bars* showing 95% confidence intervals. Select *Y axis starts at 0* to scale the plot at zero. You can also use this dialog box to plot the main effects and the various two-way interactions if you wish, but if the three-way interaction is significant these effects will be largely uninteresting.

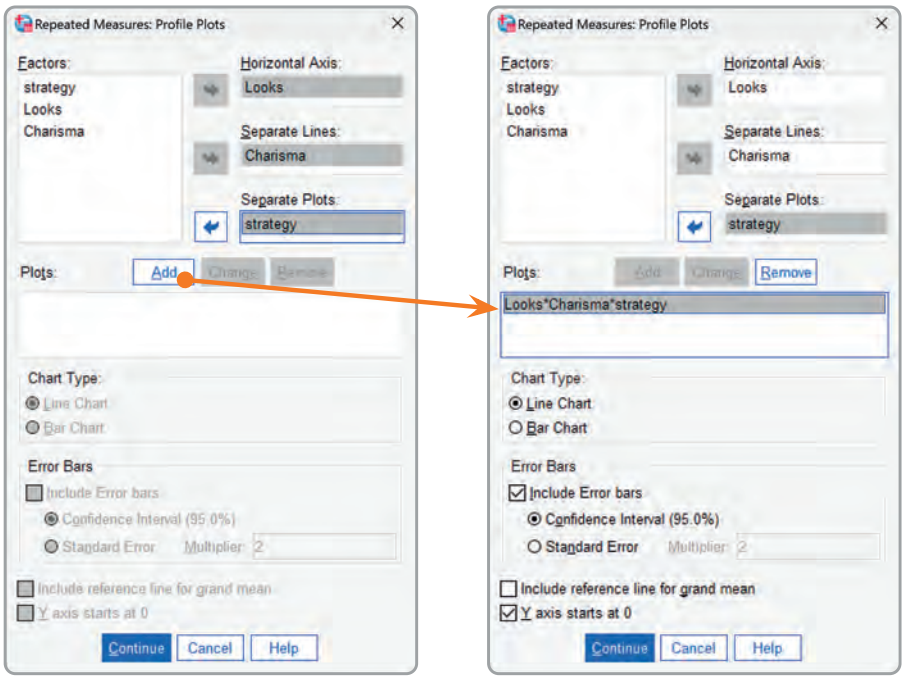


Figure 16.9 The *Profile Plots* dialog box for a three-way mixed design

As far as other options are concerned, select the same ones that were chosen for the example in the previous chapter (see Section 15.14.3): select estimated marginal means for the highest-order interaction (assuming this is the effect in which you're mainly interested), and you could select ☑ Compare simple main effects to help to break down this interaction instead of using contrasts (Figure 16.10, left). Also, remembering that the researchers are using an alpha of 0.005, we could change the significance level of confidence intervals to reflect this decision (Figure 16.10, right).

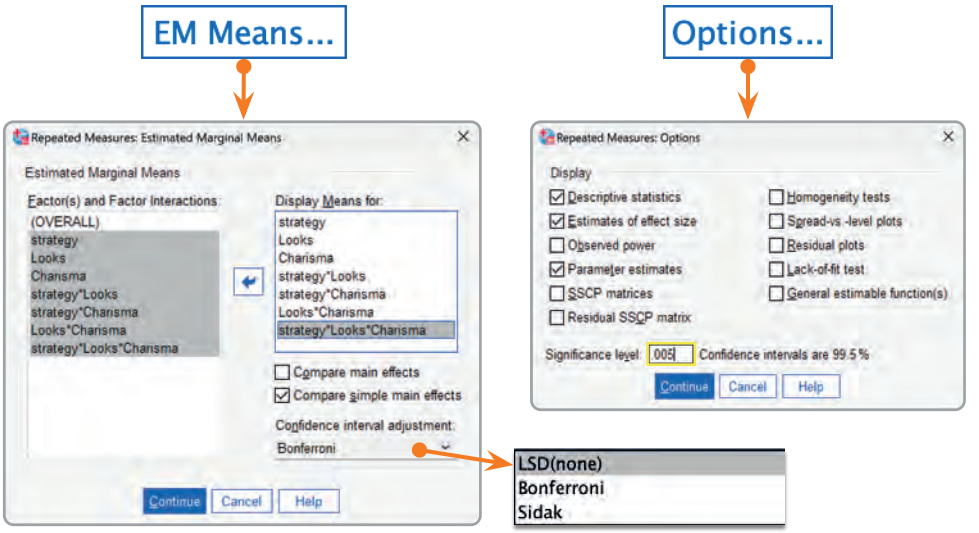


Figure 16.10 The *Estimated Marginal Means* dialog box for a three-way mixed design



Oditi's Lantern
Mixed designs

'I, Oditi, may be low on attractiveness to you but I am high on charisma. I'm high on a few other things too. Look deep into my charming eyes and you will find that you want to join the cult of undiscovered numerical truths. Our next lesson is mixed designs, so stare into my lantern and immerse yourself ever deeper in the cult. You will awaken with a strange love of three-way interactions, and a desire to do only as I say.'



16.6 Output for mixed factorial designs ©

Output 16.1 contains a table listing the repeated-measures variables from the data editor and the level of each independent variable that they represent. A second table contains descriptive statistics (mean and standard deviation) for each of the nine repeated-measures conditions split according to whether participants sat with dates who played hard to get or not.



Output 16.2 shows information about sphericity for each of the three repeated-measures effects in the model. Although I advised correcting for sphericity by default, the estimates show very little deviation from sphericity: the Huynh–Feldt estimates are all 1, which equates to spherical data, and the Greenhouse–Geisser estimates are very close to 1. Given that the Huynh–Feldt estimates show no

Within-Subjects Factors

Measure: MEASURE_1

Looks	Charisma	Dependent Variable
1	1	att_high
	2	att_none
	3	att_some
2	1	ug_high
	2	ug_none
	3	ug_some
3	1	av_high
	2	av_none
	3	av_some

Descriptive Statistics

	Dating strategy	Mean	Std. Deviation	N
Attractive and highly charismatic	Normal	89.60	6.637	10
	Hard to get	88.30	5.697	10
	Total	88.95	6.057	20
Attractive and a dullard	Normal	51.80	3.458	10
	Hard to get	87.30	5.438	10
	Total	69.55	18.743	20
Attractive and some charisma	Normal	87.10	6.806	10
	Hard to get	88.50	5.740	10
	Total	87.80	6.170	20
Ugly and highly charismatic	Normal	86.70	5.438	10
	Hard to get	56.80	5.731	10
	Total	71.75	16.274	20
Ugly and a dullard	Normal	46.10	3.071	10
	Hard to get	45.80	3.584	10
	Total	45.95	3.252	20
Ugly and some charisma	Normal	51.20	5.453	10
	Hard to get	48.30	5.376	10
	Total	49.75	5.476	20
Average and highly charismatic	Normal	88.40	8.329	10
	Hard to get	82.80	7.005	10
	Total	85.60	8.022	20
Average and a dullard	Normal	47.00	3.742	10
	Hard to get	47.80	4.185	10
	Total	47.40	3.885	20
Average and some charisma	Normal	68.90	5.953	10
	Hard to get	71.80	4.417	10
	Total	70.35	5.314	20

Output 16.1

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b] Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Looks	.960	.690	2	.708	.962	1.000	.500
Charisma	.929	1.246	2	.536	.934	1.000	.500
Looks * Charisma	.613	8.025	9	.534	.799	1.000	.250

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept + strategy
Within Subjects Design: Looks + Charisma + Looks * Charisma

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Output 16.2

deviation from sphericity it would be reasonable not to correct for it. However, the general advice to apply a correction still holds because it will correct to the degree that sphericity is violated, which in this case is a tiny amount. If you have more enthusiasm for Mauchly's test than I do you might note that all the values in the column labelled *Sig.* are above 0.005 (our criterion for significance for this study), indicating no significant departures from sphericity.

Output 16.3 contains the *F*-statistics. The version of this table that you will see will look a lot more hideous, but I have used the instructions in SPSS Tip 15.3 to hide the values that I don't want to see. The table is split into sections for each of the effects in the model and their associated error terms. The interactions between the between-groups variable of **strategy** and the repeated-measures effects are included in this table also. Remembering that the researchers decided to use an alpha of 0.005 in this study, significant effects are those with values in the column labelled *Sig.* that are below 0.005. Working down from the top of the table, we find significant effects of **Looks**, the **Looks** × **strategy** interaction, **Charisma**, the **Charisma** × **strategy** interaction, the **Looks** × **Charisma** interaction and the **Looks** × **Charisma** × **strategy** interaction. Everything, basically. You wouldn't normally be interested in main effects when there are significant interactions, you also wouldn't be interested in lower-order interactions (in this case the interactions involving two variables) when there is a significant high-order interaction (in this case the interaction involving all three predictors); however, in the interests of practice we'll interpret each effect in turn, starting with the main effect of **strategy**.



Tests of Within-Subjects Effects

Measure: MEASURE_1
Greenhouse-Geisser

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Looks	20779.633	1.923	10803.275	423.733	<.001	.959
Looks * strategy	3944.100	1.923	2050.527	80.427	<.001	.817
Error(Looks)	882.711	34.622	25.496			
Charisma	23233.600	1.868	12437.761	328.250	<.001	.948
Charisma * strategy	4420.133	1.868	2366.252	62.449	<.001	.776
Error(Charisma)	1274.044	33.624	37.891			
Looks * Charisma	4055.267	3.197	1268.295	36.633	<.001	.671
Looks * Charisma * strategy	2669.667	3.197	834.945	24.116	<.001	.573
Error(Looks*Charisma)	1992.622	57.554	34.622			

Output 16.3

16.6.1 The main effect of **strategy** Ⓑ

Before looking at the main effect of **strategy**, some people (but not me) would use Levene's test to check the assumption of homogeneity of variance (see Section 6.9.6).

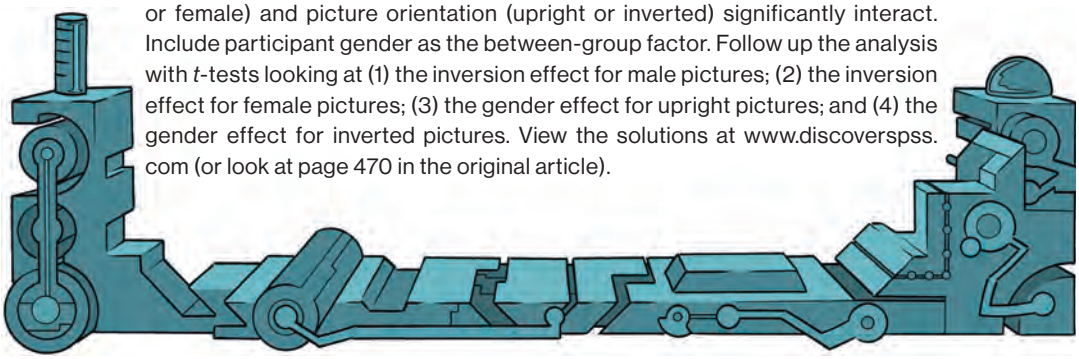




Labcoat Leni's Real Research 16.1
The objectification of desire ©

Bernard, P., et al. (2012). *Psychological Science, 23*(5), 469–471.

There is a concern that images that portray women as sexually desirable objectify them. Philippe Bernard and colleagues tested this idea in an inventive study that used the 'inversion effect', which describes the phenomenon that people find it harder to recognize upside-down (inverted) pictures than ones the right way up. It turns out that this 'inversion effect' occurs for pictures of humans, but not objects, so by seeing how easily people recognize inverted sexualized pictures of women, Bernard could see whether these pictures were processed more like people or objects (Bernard et al., 2012). Participants were presented with pictures of sexualized (i.e., not wearing many clothes) males and females, half of which were inverted (**inverted_women** and **inverted_men**) whereas the remainder were upright (**upright_women** and **upright_men**). Participants self-reported their **gender**. After each trial, participants were shown two pictures and asked to identify the one they had just seen. The outcome was the proportion of correctly identified pictures. An inversion effect is demonstrated by higher recognition scores for upright pictures than for inverted ones. If sexualized females are processed as objects you would expect an inversion effect for the male pictures but not the female ones. The data are in **bernard_2012.sav**. Fit a model to see whether picture gender (male or female) and picture orientation (upright or inverted) significantly interact. Include participant gender as the between-group factor. Follow up the analysis with *t*-tests looking at (1) the inversion effect for male pictures; (2) the inversion effect for female pictures; (3) the gender effect for upright pictures; and (4) the gender effect for inverted pictures. View the solutions at www.discoverspss.com (or look at page 470 in the original article).



Output 16.4 shows Levene's test for whether the variances were equivalent in the hard to get and normal conditions across all nine combined levels of the repeated-measures variables; because all significance values are greater than 0.005, variances are homogeneous for all levels of the repeated-measures variables.

The main effect of **strategy** is listed separately from the repeated-measures effects in Output 16.5. It has a non-significant effect on ratings of dates because the significance of 0.946 is greater than the cut-off of 0.005. This effect tells us that if we ignore all other variables, ratings were

equivalent regardless of whether the date adopted a hard to get persona or not. If you requested *Estimated Marginal Means* in the options (I'll assume you did from now on) you will output the table in Figure 16.11. I've also included a plot of these means. It is clear that, overall, ratings of dates playing hard to get were equivalent to dates who were not.

Levene's Test of Equality of Error Variances[a]

Based on Mean

	Levene Statistic	df1	df2	Sig.
Attractive and highly charismatic	1.131	1	18	.302
Attractive and a dullard	1.949	1	18	.180
Attractive and some charisma	.599	1	18	.449
Ugly and highly charismatic	.005	1	18	.945
Ugly and a dullard	.082	1	18	.778
Ugly and some charisma	.124	1	18	.729
Average and highly charismatic	.102	1	18	.753
Average and a dullard	.004	1	18	.950
Average and some charisma	1.763	1	18	.201

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + strategy
Within Subjects Design: Looks + Charisma + Looks * Charisma

Output 16.4

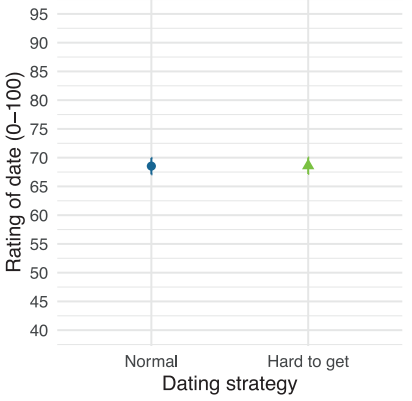
Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	94027.756	1	94027.756	20036.900	<.001	.999
strategy	.022	1	.022	.005	.946	.000
Error	84.469	18	4.693			

Output 16.5



1. Dating strategy

Measure: MEASURE_1

Dating strategy	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Normal	68.533	.685	67.094	69.973
Hard to get	68.600	.685	67.161	70.039

Figure 16.11 Means and plot of the main effect of **strategy**

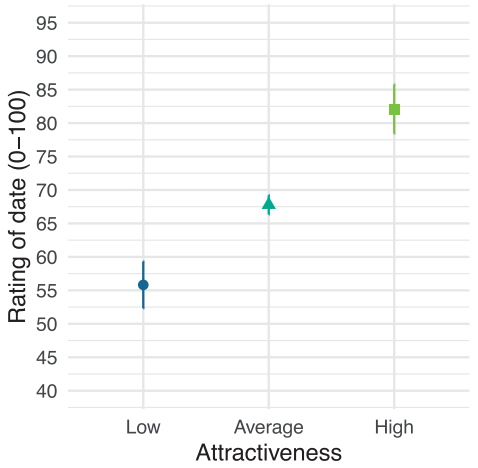
16.6.2 The main effect of **Looks** Ⓑ



Based on the previous section, on what you have learned in previous chapters, and on Output 16.3, can you interpret the main effect of **Looks**?

Output 16.3 showed a significant main effect of **Looks**, $F(1.92, 34.62) = 423.73$, $p < 0.001$, $\eta_p^2 = 0.96$, which means that if we ignore all other variables, ratings of attractive, average and unattractive dates differed. Figure 16.12 shows the *Estimated Marginal Means* and a plot of them. The levels of **Looks** are labelled as 1, 2 and 3, and it's down to you to remember how you entered the variables (or refer to Output 16.1). If you assigned variables as I did then level 1 is attractive, level 2 is unattractive and level 3 is average. From this table and plot you can see that as attractiveness increases, the mean rating increases too. This main effect seems to reflect that the raters were more likely to express a greater interest in going out with attractive people than with average or unattractive people. However, contrasts will help us to understand exactly what's going on.

Output 16.6 shows the contrasts that we requested. For the time being, just look at the row labelled *Looks*. Remember that we did a simple contrast, and so we get a contrast comparing level 1 to level 3, and then comparing level 2 to level 3; because of the order in which we entered the variables, these contrasts represent attractive compared to average (level 1 vs. level 3) and unattractive compared to average (level 2 vs. level 3). The values of F for each contrast, and their related significance values, tell us that the main effect of **Looks** represents the fact that attractive dates were rated significantly higher than average dates, $F(1, 18) = 226.99$, $p < 0.001$, $\eta_p^2 = 0.93$, and average dates were rated significantly higher than unattractive ones, $F(1, 18) = 160.07$, $p < 0.001$, $\eta_p^2 = 0.90$.



2. Looks

Measure: MEASURE_1

Looks	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	82.100	.652	80.729	83.471
2	55.817	.651	54.449	57.184
3	67.783	.820	66.061	69.505

Figure 16.12 Means and plot of the main effect of **Looks**

Tests of Within-Subjects Contrasts

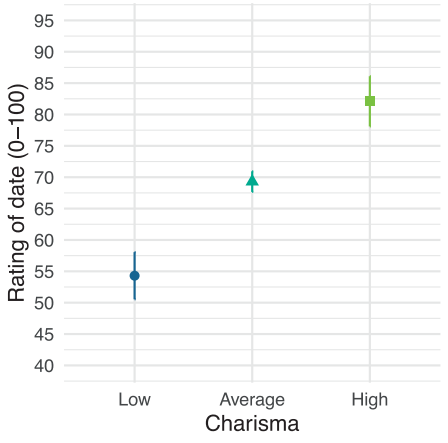
Measure: MEASURE_1

Source	Looks	Charisma	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Looks	Level 1 vs. Level 3		4099.339	1	4099.339	226.986	<.001	.927
	Level 2 vs. Level 3		2864.022	1	2864.022	160.067	<.001	.899
Looks * strategy	Level 1 vs. Level 3		781.250	1	781.250	43.259	<.001	.706
	Level 2 vs. Level 3		540.800	1	540.800	30.225	<.001	.627
Error(Looks)	Level 1 vs. Level 3		325.078	18	18.060			
	Level 2 vs. Level 3		322.067	18	17.893			
Charisma		Level 1 vs. Level 3	3276.800	1	3276.800	109.937	<.001	.859
		Level 2 vs. Level 3	4500.000	1	4500.000	227.941	<.001	.927
Charisma * strategy		Level 1 vs. Level 3	810.689	1	810.689	27.199	<.001	.602
		Level 2 vs. Level 3	665.089	1	665.089	33.689	<.001	.652
Error(Charisma)		Level 1 vs. Level 3	536.511	18	29.806			
		Level 2 vs. Level 3	355.356	18	19.742			
Looks * Charisma	Level 1 vs. Level 3	Level 1 vs. Level 3	3976.200	1	3976.200	21.944	<.001	.549
		Level 2 vs. Level 3	441.800	1	441.800	4.091	.058	.185
	Level 2 vs. Level 3	Level 1 vs. Level 3	911.250	1	911.250	6.231	.022	.257
		Level 2 vs. Level 3	7334.450	1	7334.450	88.598	<.001	.831
Looks * Charisma * strategy	Level 1 vs. Level 3	Level 1 vs. Level 3	168.200	1	168.200	.928	.348	.049
		Level 2 vs. Level 3	6552.200	1	6552.200	60.669	<.001	.771
	Level 2 vs. Level 3	Level 1 vs. Level 3	1711.250	1	1711.250	11.701	.003	.394
		Level 2 vs. Level 3	110.450	1	110.450	1.334	.263	.069
Error(Looks*Charisma)	Level 1 vs. Level 3	Level 1 vs. Level 3	3261.600	18	181.200			
		Level 2 vs. Level 3	1944.000	18	108.000			
	Level 2 vs. Level 3	Level 1 vs. Level 3	2632.500	18	146.250			
		Level 2 vs. Level 3	1490.100	18	82.783			

Output 16.6

16.6.3 The main effect of **Charisma** Ⓑ

In Output 16.3 there was a significant main effect of **Charisma**, $F(1.87, 33.62) = 328.25$, $p < 0.001$, $\eta_p^2 = 0.95$, which tells us that if we ignore all other variables, ratings for dates with high charisma, a little charisma and no charisma differed. Figure 16.13 shows the estimated marginal means from the output together with a plot. Again, the levels of **Charisma** are labelled as 1, 2 and 3. If you followed what I did then level 1 is high charisma, level 2 is bore and level 3 is some charisma. This main effect seems to reflect that as charisma increases, the mean rating of the date increases too: raters expressed a greater interest in going out with charismatic people than with average people or bores.



3. Charisma

Measure: MEASURE_1

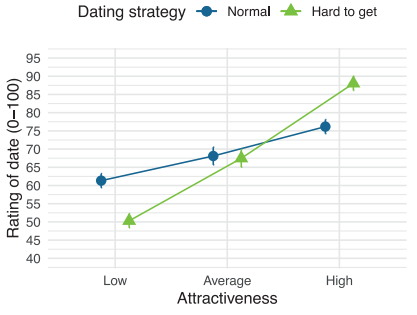
Charisma	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	82.100	1.010	79.978	84.222
2	54.300	.573	53.096	55.504
3	69.300	.732	67.763	70.837

Figure 16.13 Means and plot of the main effect of **Charisma**

We requested simple contrasts (the row labelled *Charisma* in Output 16.6), and because of the order that we entered variables these contrasts represent high charisma compared to some charisma (level 1 vs. level 3) and no charisma compared to some charisma (level 2 vs. level 3). These contrasts tell us that the main effect of **Charisma** is that highly charismatic dates were rated significantly higher than dates with some charisma, $F(1, 18) = 109.94$, $p < 0.001$, $\eta_p^2 = 0.68$, and dates with some charisma were rated significantly higher than bores, $F(1, 18) = 227.94$, $p < 0.001$, $\eta_p^2 = 0.93$.

16.6.4 The interaction between **strategy** and **Looks** Ⓑ

The strategy adopted significantly interacted with the attractiveness of the date, $F(1.92, 34.62) = 80.43$, $p < 0.001$, $\eta_p^2 = 0.82$ (Output 16.3). This effect tells us that the profile of ratings across dates of different attractiveness was different depending on whether or not they played hard to get. The estimated marginal means and interaction plot (you can obtain a version of this using the dialog box in Figure 16.9) are in Figure 16.14. The plot shows that for average-looking dates it doesn't make a difference whether they played hard to get (the blue circle and green triangle are in a similar location). For attractive dates, ratings were higher when the date played hard to get (green triangle) than when they didn't (blue circle) and for unattractive dates the opposite was true – ratings were lower when dates played hard to get. In short, playing hard to get only has an effect at the extremes of attractiveness. Another way to look at this is the slope of the lines: when dates played hard to get, the slope (green line connecting the triangles) is steeper than when they didn't (blue line that connects the circles), implying that attractiveness has a greater impact on ratings when dates play hard to get. This interaction can be clarified using the contrasts in Output 16.6.



4. Dating strategy * Looks

Measure: MEASURE_1

Dating strategy	Looks	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Normal	1	76.167	.923	74.228	78.105
	2	61.333	.921	59.399	63.267
	3	68.100	1.159	65.665	70.535
Hard to get	1	88.033	.923	86.095	89.972
	2	50.300	.921	48.366	52.234
	3	67.467	1.159	65.031	69.902

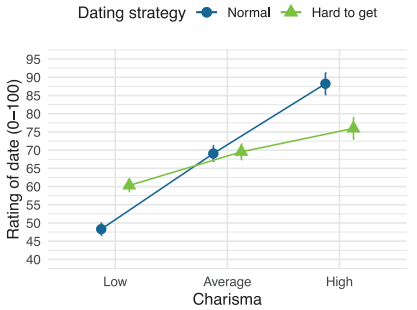
Figure 16.14 Means and plot of the **strategy** × **Looks** interaction

The first contrast for the interaction term looks at level 1 of **Looks** (attractive) compared to level 3 (average), comparing playing hard to get to normal. This contrast is highly significant, $F(1, 18) = 43.26$, $p < 0.001$, $\eta_p^2 = 0.71$, suggesting that the increased interest in attractive dates compared to average-looking dates found when dates played hard to get is significantly more than when they acted normally. So, in Figure 16.14 the slope of the green line (hard to get) between the triangles for attractive dates and average-looking dates is steeper than the comparable blue line (normal) that connects the circles. The preferences for attractive dates, compared to average-looking dates, are greater when they play hard to get than when they don't.

The second contrast, which compares playing hard to get to normal at level 2 of looks (unattractive) relative to level 3 (average) is also highly significant, $F(1, 18) = 30.23$, $p < 0.001$, $\eta_p^2 = 0.63$. This contrast tells us that the decreased interest in unattractive dates compared to average-looking dates found when dates played hard to get is significantly more than when they did not. In Figure 16.14 the slope of the green line between the triangles for unattractive and average dates is steeper than the corresponding blue line that connects the circles. The preferences for average-looking dates, compared to unattractive dates, are greater when they play hard to get than when they don't.

16.6.5 The interaction between **strategy** and **Charisma** Ⓑ

Output 16.3 showed that strategy significantly interacted with how charismatic the date was, $F(1.87, 33.62) = 62.45$, $p < 0.001$, $\eta_p^2 = 0.78$. This effect means that the profile of ratings across dates of different levels of charisma was influenced by the dating strategy employed. The estimated marginal means and plot in Figure 16.15 show almost the reverse pattern to the **strategy** × **Looks** interaction. For dates with normal amounts of charisma the dating strategy they adopted had little impact (the blue circle and green triangle coincide). However, highly charismatic dates were rated higher when they acted normally compared to when they played hard to get (the blue circle is above the green triangle). Conversely, boring dates were rated higher when they played hard to get compared to when they acted normally (the blue circle is below the green triangle). Although interest in dating increases as charisma increases, this increases is less pronounced when the dates play hard to get.



5. Dating strategy * Charisma

Measure: MEASURE_1

Dating strategy	Charisma	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Normal	1	88.233	1.428	85.233	91.234
	2	48.300	.810	46.598	50.002
	3	69.067	1.035	66.893	71.240
Hard to get	1	75.967	1.428	72.966	78.967
	2	60.300	.810	58.598	62.002
	3	69.533	1.035	67.360	71.707

Figure 16.15 Means and plot of the **strategy** × **Charisma** interaction

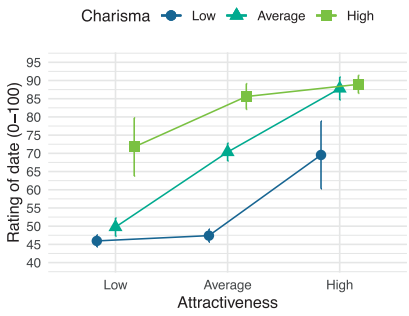
We can break this interaction down using the contrasts in Output 16.6. The first one, which looks at level 1 of **Charisma** (high charisma) compared to level 3 (some charisma), for playing hard to get relative to normal, is highly significant, $F(1, 18) = 27.20$, $p < 0.001$, $\eta_p^2 = 0.60$. This result tells us that the increased interest in highly charismatic dates compared to averagely charismatic dates found when dates acted normally is significantly more than when they played hard to get. In Figure 16.15 the slope of the green line (hard to get) between the triangles for high charisma and some charisma is steeper than the corresponding blue line between the circles (normal). The preferences for very charismatic dates, compared to averagely charismatic dates, are smaller when dates play hard to get.

The second contrast for the **Charisma** × **strategy** interaction looks at level 2 of **Charisma** (bore) compared to level 3 (some charisma), for playing hard to get relative to normal. This contrast is highly significant, $F(1, 18) = 33.69$, $p < 0.001$, $\eta_p^2 = 0.65$, suggesting that the decreased interest in boring dates compared to averagely charismatic dates found is significantly less when dates play hard to get than when they act normally. In Figure 16.15 the slope of the blue line (normal) between the circles for some charisma and bore is steeper than the corresponding green line/triangles (hard to get): the preferences for dates with some charisma over bore is greater when dates act normally than when they play hard to get.

16.6.6 The interaction between **Looks** and **Charisma** Ⓑ

There was a significant **Looks** × **Charisma** interaction, $F(3.20, 57.55) = 36.63$, $p < 0.001$, $\eta_p^2 = 0.67$ (Output 16.3). This effect tells us that the profiles of ratings across dates of different levels of charisma were different for attractive, average-looking and unattractive dates. We can unpick this interaction using the estimated marginal means, a plot (use the dialog box in Figure 16.9 to get a similar one), and contrasts. The plot (Figure 16.16) shows the mean ratings of dates of different levels of attractiveness when the date also had high levels of charisma (light green line/squares), some charisma (dark

green line/triangles) and low charisma (blue line/circles). Look first at the difference between average and highly attractive dates. The interest in highly charismatic dates doesn't change (the line connecting the squares is almost flat), but for dates with some charisma or no charisma ratings are higher for highly attractive dates than for average-looking dates (the line connecting the triangles and the line connecting the circles both slope up). If you have lots of charisma you can get away with being average-looking and people will still want to date you. Now look at the difference between average-looking and unattractive dates. A different pattern is observed: for dates with low charisma there is little difference between unattractive and average-looking people (the line connecting the circles is flat) but for those with average or high charisma, interest is higher if you're averagely attractive compared to unattractive (the other lines slope up). Basically, if you're boring you need to be really attractive before people want to date you, and if you're unattractive then having 'some' charisma won't help you much.



6. Looks * Charisma

Measure: MEASURE_1

Looks	Charisma	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	1	88.950	1.383	86.045	91.855
	2	69.550	1.019	67.409	71.691
	3	87.800	1.408	84.842	90.758
2	1	71.750	1.249	69.126	74.374
	2	45.950	.746	44.382	47.518
	3	49.750	1.211	47.206	52.294
3	1	85.600	1.721	81.985	89.215
	2	47.400	.888	45.535	49.265
	3	70.350	1.172	67.888	72.812

Figure 16.16 Means and plot of the **Looks** × **Charisma** interaction

The contrasts in Output 16.6 help to pick apart this interaction. Look at the row labelled *Looks * Charisma*. The first contrast investigates level 1 of looks (highly attractive) compared to level 3 (average-looking), for level 1 of **Charisma** (high charisma) relative to level 3 (average charisma). This contrast asks 'Is the difference between high charisma and average charisma the same for attractive people and average-looking people?' The best way to understand this contrast is to focus on the relevant bit of the interaction plot in Figure 16.16, which I have reproduced in Figure 16.17 (top left). Interest (as indicated by high ratings) in attractive dates was the same regardless of whether they had high or average charisma; however, for average-looking dates, there was more interest when that person had high charisma rather than some. The contrast is highly significant, $F(1, 18) = 21.94$, $p < 0.001$, $\eta_p^2 = 0.55$, and tells us that as attractiveness increases there is a significantly greater increase in interest when charisma is average compared to when it is high.

The second contrast asks the question 'Is the difference between low charisma and average charisma the same for attractive people and average-looking people?' It explores level 1 of **Looks** (highly attractive) compared to level 3 (average-looking), for level 2 of **Charisma** (low charisma) relative to level 3 (average charisma). We can again focus on the relevant part of the interaction plot (Figure 16.16) which is reproduced in Figure 16.17 (top right). This plot shows that interest in average-looking dates was higher when they had average charisma than when they were a bore (the triangle is higher than the circle); the same is also true for highly attractive dates. The two lines are fairly parallel, which is reflected in the non-significant contrast, $F(1, 18) = 4.09$, $p = 0.058$, $\eta_p^2 = 0.19$. It seems that as the attractiveness of dates increases there is an increase in interest both when charisma is average and when the date is a bore.

The third contrast investigates level 2 of **Looks** (low attractiveness) relative to level 3 (average-looking), comparing level 1 of **Charisma** (high charisma) to level 3 (average charisma). This contrast asks 'is the difference between high charisma and average charisma the same for people low on attractiveness and average-looking people?' The relevant part of the interaction plot is shown in

Figure 16.17 (bottom left). Interest in dating increases from unattractive dates to average-looking ones. Remembering that we're using a threshold of 0.005 (not the usual 0.05), this effect is statistically comparable in dates that are high and average in charisma, $F(1, 18) = 6.23$, $p = 0.022$, $\eta_p^2 = 0.26$. As dates' attractiveness increases from low to average, there is an increase in ratings that is comparable in those average and high in charisma.

The final contrast addresses the question 'Is the difference between low charisma and average charisma the same for people low on attractiveness and average-looking people?' It compares level 2 of **Looks** (low attractiveness) to level 3 (average-looking) at level 2 of **Charisma** (low charisma) relative to level 3 (average charisma). The relevant part of the interaction plot is shown in Figure 16.17 (bottom right). For unattractive dates the ratings are roughly the same regardless of the level of charisma, but for average-looking dates, ratings were higher when they had average charisma compared to when charisma was low. This contrast is significant, $F(1, 18) = 88.60$, $p < 0.001$, $\eta_p^2 = 0.83$.

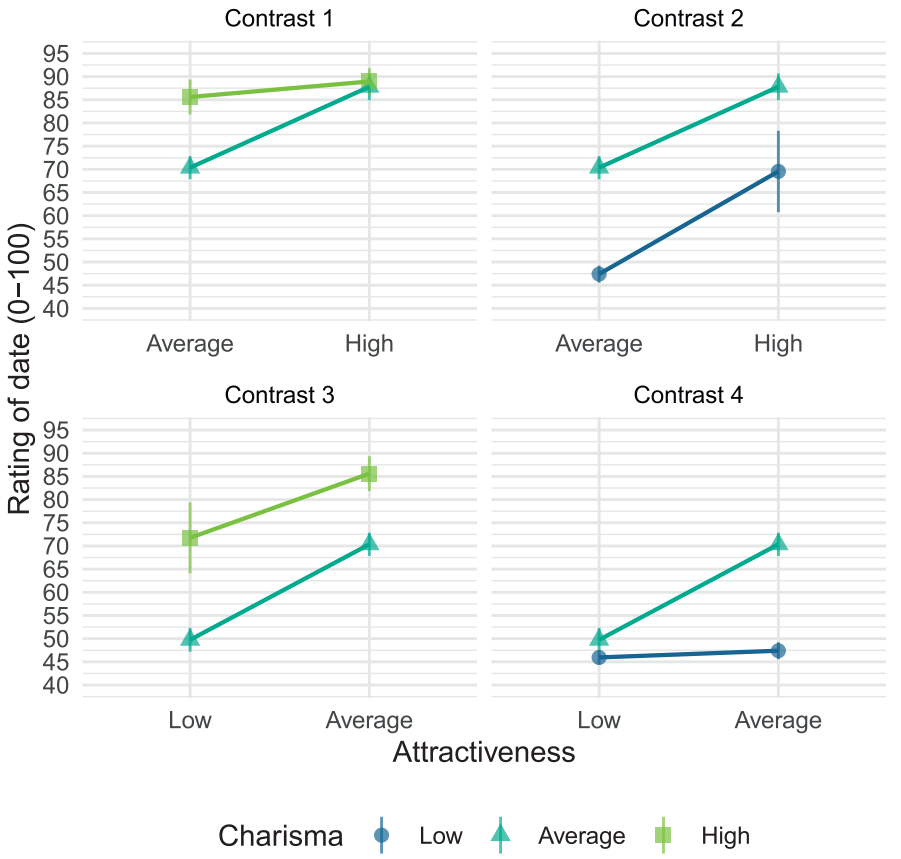


Figure 16.17 The **Looks** × **Charisma** interaction broken down into the four contrasts

16.6.7 The interaction between **Looks**, **Charisma** and **strategy** ©

The significant **Looks** × **Charisma** × **strategy** interaction, $F(3.20, 57.55) = 24.12$, $p < 0.001$, $\eta_p^2 = 0.57$ (Output 16.3), tells us whether the **Looks** × **Charisma** interaction described above is the same when dates played hard to get compared to when they didn't. The nature of this interaction is revealed in



Figure 16.18, which shows the **Looks** × **Charisma** interaction separately when dates played hard to get and acted normally (the means are in Output 16.7). The plot for dates who played hard to get shows that when dates are highly attractive, a high interest was expressed regardless of charisma levels (the square, triangle and circle are located in similar positions). At the opposite end of the attractiveness scale, when a date is unattractive, regardless of charisma, very little interest is expressed (ratings are all low). If the date plays hard to get, the only time charisma makes a difference is if the date is average-looking, in which case high charisma (square) boosts interest, being a bore (circle) reduces interest and having average charisma (triangle) leaves things somewhere in between. The take-home message is that playing hard to get works only if you're averagely attractive: if you're highly charismatic it will boost your appeal but it's a disastrous strategy if you're a bore.[4]

The picture when the date doesn't play hard to get (acts normally) is different. If someone has high levels of charisma then what they look like won't affect interest in them (the squares are connected by a relatively flat line). At the other extreme, if the date is a bore, then they will have very little interest expressed in them regardless of how attractive they are (the circles are connected by a relatively flat line). The only time attractiveness makes a difference is when someone has an average amount of charisma (the triangles), in which case being attractive boosts interest, and being unattractive reduces it. If you don't play hard to get and you are averagely attractive then you can influence others' interest in you with your charisma.

8. Dating strategy * Looks * Charisma

Estimates

Measure: MEASURE_1

Dating strategy	Looks	Charisma	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Normal	1	1	89.600	1.956	85.491	93.709
		2	51.800	1.441	48.773	54.827
		3	87.100	1.991	82.917	91.283
	2	1	86.700	1.767	82.989	90.411
		2	46.100	1.055	43.883	48.317
		3	51.200	1.712	47.603	54.797
	3	1	88.400	2.434	83.287	93.513
		2	47.000	1.255	44.363	49.637
		3	68.900	1.657	65.418	72.382
Hard to get	1	1	88.300	1.956	84.191	92.409
		2	87.300	1.441	84.273	90.327
		3	88.500	1.991	84.317	92.683
	2	1	56.800	1.767	53.089	60.511
		2	45.800	1.055	43.583	48.017
		3	48.300	1.712	44.703	51.897
	3	1	82.800	2.434	77.687	87.913
		2	47.800	1.255	45.163	50.437
		3	71.800	1.657	68.318	75.282

Output 16.7

4 These data are made up, so please don't base your life decisions on this example ☺

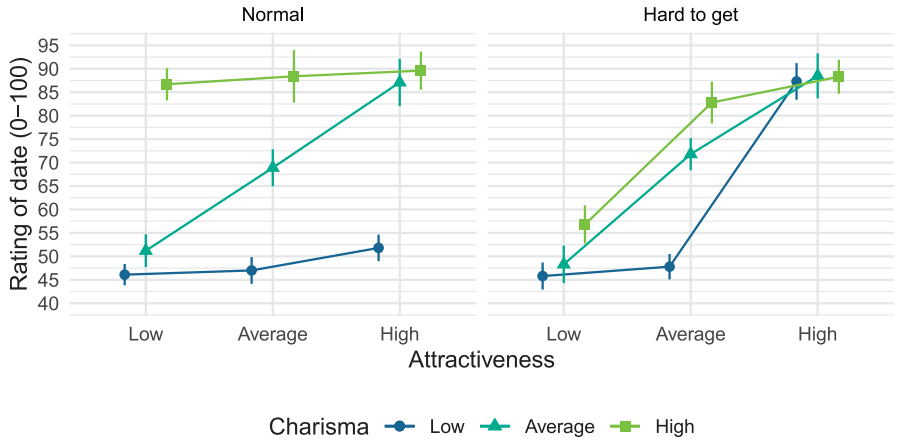


Figure 16.18 Graphs showing the **Looks** × **Charisma** interaction for different dating strategies

Again, we can use contrasts to break down this interaction (Output 16.6). Look at the highlighted rows. These contrasts are similar to those for the **Looks** × **Charisma** interaction, but they now also take into account the effect of dating strategy as well. The first contrast for the **Looks** × **Charisma** × **strategy** interaction explores level 1 of **Looks** (highly attractive) relative to level 3 (average-looking), when level 1 of **Charisma** (high charisma) is compared to level 3 (average charisma), when dates played hard to get relative to when they didn't, $F(1, 18) = 0.93$, $p = 0.348$, $\eta_p^2 = 0.05$. The relevant parts of Figure 16.18 are shown in the first column of Figure 16.19. It seems that interest in dating (as indicated by high ratings) attractive dates was the same regardless of whether they had high or average charisma (the squares and triangles are in a similar place). However, for average-looking dates, there was more interest when that person had high charisma rather than some charisma (the triangles are lower than the squares). The non-significance of this contrast indicates that this pattern of results is very similar when dates played hard to get and when they didn't.

The second contrast explores level 1 of **Looks** (attractive) relative to level 3 (average-looking), when level 2 of **Charisma** (low charisma) is compared to level 3 (average charisma), when dates played hard to get relative to when they didn't. The relevant means are shown in the second column of Figure 16.19. The contrast is significant, $F(1, 18) = 60.67$, $p < 0.001$, $\eta_p^2 = 0.77$, which reflects the fact that the pattern of means is different when dates played hard to get compared to when they didn't. First, if we look at average-looking dates, more interest was expressed when the date has some charisma than when they have none, and this is true whether or not dates played hard to get (the distance between the triangle and circle is about the same in the two dating strategy groups). So, the difference created by playing hard to get doesn't appear to be here. Now look at attractive dates. When dates played hard to get (bottom) the interest in the date is high regardless of their charisma (the circle and triangle meet). However, when dates acted normally (top) interest in dating an attractive person is much lower if they are a bore (the circle is much lower than the triangle).

Another way to look at it is that for dates with average charisma, the increase in interest as attractiveness goes up is about the same regardless of whether they played hard to get (the lines connecting the triangles have the same slope). However, for dates who are bores, the increase in interest if these dates are attractive rather than average-looking is much more dramatic if they play hard to get (the line connecting the circles is steeper in the hard to get group).

The third contrast was also significant, $F(1, 18) = 11.70$, $p = 0.003$, $\eta_p^2 = 0.39$. This contrast compares level 2 of **Looks** (low attractiveness) to level 3 (average-looking) in level 1 of **Charisma** (high charisma) relative to level 3 (average charisma), when dates played hard to get relative to when they didn't. The third column of Figure 16.19 shows the relevant means. First, let's look at when dates

played hard to get (bottom). As attractiveness increases, so does interest, regardless of levels of charisma (the slope of the line connecting the triangles is similar to the that of the line connecting the squares). So, regardless of charisma, there is a similar increase in interest as attractiveness increases. Now let's look at when the dates acted normally (top). The picture is quite different: when charisma is high, there is no change in interest as attractiveness increases (the line connecting the squares is flat); however, when charisma is average, interest in an average-looking date is higher than for an unattractive date (the line connecting the triangles slopes up).

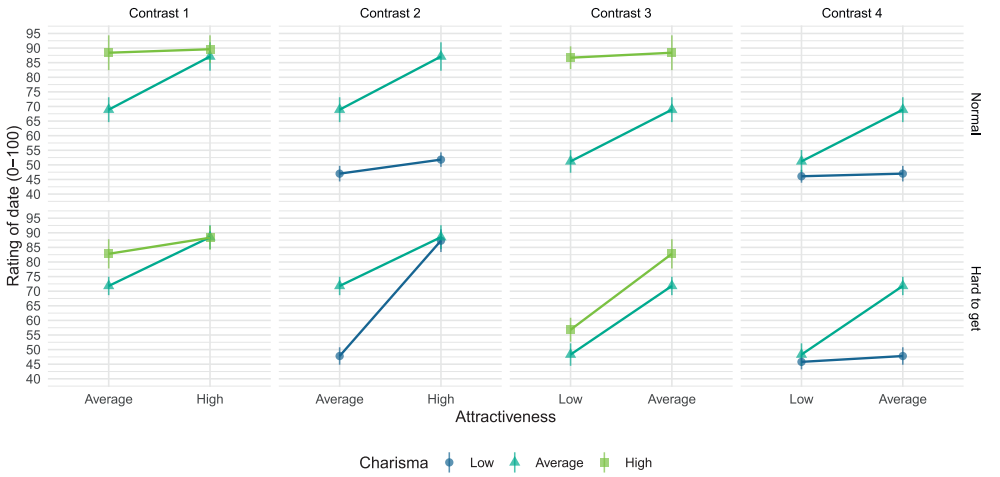


Figure 16.19 The **Looks** × **Charisma** × **strategy** interaction broken down into the four contrasts

Another way to look at it is that for dates with some charisma, the increase in interest as attractiveness increases is about the same regardless of whether dates play hard to get (the lines connecting the triangles have similar slopes). However, for dates who have high charisma, the increase in interest if these dates are average-looking rather than unattractive is much more dramatic when dates played hard to get than when they didn't (the line connecting the squares is steeper for dates that played hard to get).

The final contrast was not significant, $F(1, 18) = 1.33$, $p = 0.263$, $\eta_p^2 = 0.07$. This contrast looks at the effect of **strategy** when comparing level 2 of **Looks** (low attractiveness) to level 3 (average-looking), in level 2 of **Charisma** (low charisma) relative to level 3 (average charisma). The relevant means are displayed in the fourth column of Figure 16.19. Interest in unattractive dates was the same regardless of charisma levels (the circles and triangles are in the same place). Interest in average-looking dates was greatest when they had average charisma rather than when they were a bore (the triangles are higher than the circles). This pattern of results is very similar when dates played hard to get and when they didn't (hence the non-significant effect).

These contrasts tell us nothing about the differences between the high- and low-attractiveness conditions, or the high- and low-charisma conditions, because these were never compared. We could rerun the analysis and specify our contrasts differently to get these effects, or use *post hoc* tests instead of contrasts.

An alternative way to break down an interaction, which we have explored in other chapters, is using simple effects. As we have seen before, with two repeated-measures variables we get two tables. In this case one table will show the simple effects of **Charisma** within each level of **Looks**, and the other will show the simple effects of **Looks** within each level of **Charisma**. However, because we also have the effect of **strategy** to consider, these tables will show the simple effects of **Charisma** within each combination of the levels of **Looks** and **strategy**, and the simple effects of **Looks** within each

combination of the levels of **Charisma** and **strategy**. I won't look at both tables (because our plan was to use contrasts), but to give you a flavour of things we'll look at the simple effects of **Looks** within each combination of the levels of **Charisma** and **strategy** in Output 16.8. Remembering that we're using 0.005 as a criterion for significance in this example (rather than 0.05), we can see that when a normal dating strategy is used there is a significant effect of **Looks** (i.e., the mean ratings of unattractive, average-looking and highly attractive dates significantly differ) only at average levels of **Charisma**. Relating these findings to Figure 16.18 (left) we're saying that the squares (high charisma) are not in significantly different locations, $F(2, 17) = 0.86$, $\eta_p^2 = 0.09$, the circles (low charisma) are not in significantly different locations, $F(2, 17) = 6.13$, $\eta_p^2 = 0.42$, but the triangles (average charisma) *are* in significantly different locations, $F(2, 17) = 135.00$, $\eta_p^2 = 0.946$. In comparison, when the strategy is playing hard to get, there is a significant effect of **Looks** for *all* levels of **Charisma**. Again, relating these findings to Figure 16.18 (right) we're saying that the squares (high charisma) are in significantly different locations, $F(2, 17) = 109.58$, $\eta_p^2 = 0.93$, the circles (low charisma) are in significantly different locations, $F(2, 17) = 360.62$, $\eta_p^2 = 0.98$, and the triangles (average charisma) are in significantly different locations, $F(2, 17) = 179.92$, $\eta_p^2 = 0.96$.

Multivariate Tests

Dating strategy	Charisma		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Normal	1	Pillai's trace	.092	.856[a]	2.000	17.000	.442	.092
		Wilks' lambda	.908	.856[a]	2.000	17.000	.442	.092
		Hotelling's trace	.101	.856[a]	2.000	17.000	.442	.092
		Roy's largest root	.101	.856[a]	2.000	17.000	.442	.092
	2	Pillai's trace	.419	6.125[a]	2.000	17.000	.010	.419
		Wilks' lambda	.581	6.125[a]	2.000	17.000	.010	.419
		Hotelling's trace	.721	6.125[a]	2.000	17.000	.010	.419
		Roy's largest root	.721	6.125[a]	2.000	17.000	.010	.419
	3	Pillai's trace	.941	134.996[a]	2.000	17.000	<.001	.941
		Wilks' lambda	.059	134.996[a]	2.000	17.000	<.001	.941
		Hotelling's trace	15.882	134.996[a]	2.000	17.000	<.001	.941
		Roy's largest root	15.882	134.996[a]	2.000	17.000	<.001	.941
Hard to get	1	Pillai's trace	.928	109.581[a]	2.000	17.000	<.001	.928
		Wilks' lambda	.072	109.581[a]	2.000	17.000	<.001	.928
		Hotelling's trace	12.892	109.581[a]	2.000	17.000	<.001	.928
		Roy's largest root	12.892	109.581[a]	2.000	17.000	<.001	.928
	2	Pillai's trace	.977	360.620[a]	2.000	17.000	<.001	.977
		Wilks' lambda	.023	360.620[a]	2.000	17.000	<.001	.977
		Hotelling's trace	42.426	360.620[a]	2.000	17.000	<.001	.977
		Roy's largest root	42.426	360.620[a]	2.000	17.000	<.001	.977
	3	Pillai's trace	.955	179.915[a]	2.000	17.000	<.001	.955
		Wilks' lambda	.045	179.915[a]	2.000	17.000	<.001	.955
		Hotelling's trace	21.166	179.915[a]	2.000	17.000	<.001	.955
		Roy's largest root	21.166	179.915[a]	2.000	17.000	<.001	.955

Each F tests the multivariate simple effects of Looks within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. Exact statistic

Output 16.8

What will hopefully be clear from this chapter is that when more than two independent variables are used and you're comparing means it yields complex interaction effects that require a great deal of concentration to interpret. Imagine just how much your brain would throb if interpreting a four-way interaction. If faced with this particularly unpleasant scenario, my best advice is to take a systematic approach to interpretation and that making plots is a useful way to proceed. It is also advisable to think carefully about the most useful contrasts to use to answer the questions that your data were collected to test.



Cramming Sam's Tips
Mixed designs

- ✓ Mixed designs compare several means when there are two or more independent variables, and at least one of them has been measured using the same entities and at least one other has been measured using different entities.
- ✓ Correct for deviations from sphericity for the repeated-measures variable(s) by routinely interpreting the Greenhouse–Geisser corrected effects. (Some people do this only if Mauchly's test is significant, but this approach is problematic because the results of the test depend on the sample size.)
- ✓ The table labelled *Tests of Within-Subjects Effects* shows the *F*-statistic(s) for any repeated-measures variables and the interaction effects. For *each* effect, read the row labelled *Greenhouse–Geisser* or *Huynh–Feldt* (read the previous chapter to find out the relative merits of the two procedures). If the value in the *Sig.* column is less than or equal to your threshold for significance (typically 0.05, but we used 0.005 in this chapter) then the means are significantly different.
- ✓ The table labelled *Tests of Between-Subjects Effects* shows the *F*-statistic(s) for any between-group variables. If the value in the *Sig.* column is less than or equal to your threshold for significance then the means of the groups are significantly different.
- ✓ Break down the highest-order interaction using contrasts. These contrasts appear in the table labelled *Tests of Within-Subjects Contrasts*; look to the columns labelled *Sig.* to discover if your comparisons are significant (i.e. the value is less than or equal to your threshold for significance).
- ✓ Alternatively use simple effects analysis or *post hoc* tests to break down interactions.
- ✓ Look at the means – or, better still, plot them – to help you interpret the contrasts.



16.7 Reporting the results of mixed designs Ⓑ

As you've probably gathered, when you have more than two independent variables there's a hell of a lot of information to report. I've mentioned a few times that when interaction effects are significant there's no point in interpreting main effects, so you can save space by not reporting them; however, some journals will expect you to report them anyway. In any case, certainly reserve the most detail for the effects that are central to your main hypothesis.

Assuming we want to report all our effects, we could do it something like this (though not as a list!):

- ✓ A 0.005 criterion was used to determine significance (Benjamin et al., 2018). All effects are significant at $p < 0.001$ unless otherwise stated. There was a significant main effect of the

attractiveness of the date on interest expressed by participant, $F(1.92, 34.62) = 423.73$, $\eta_p^2 = 0.96$. Contrasts revealed that attractive dates were significantly more desirable than average-looking ones, $F(1, 18) = 226.99$, $\eta_p^2 = 0.93$, and unattractive dates were significantly less desirable than average-looking ones, $F(1, 18) = 160.07$, $\eta_p^2 = 0.90$.

- ✓ There was a significant main effect of the amount of charisma the date exhibited on the interest expressed in dating them, $F(1.87, 33.62) = 328.25$, $\eta_p^2 = 0.95$. Contrasts revealed that dates displaying high charisma were significantly more desirable than dates displaying average charisma, $F(1, 18) = 109.94$, $\eta_p^2 = 0.86$, and low-charisma dates were significantly less desirable than dates exhibiting some charisma , $F(1, 18) = 227.94$, $\eta_p^2 = 0.93$.
- ✓ There was no significant effect of strategy, indicating that overall ratings of dates who played hard to get was similar to those who acted normally, $F(1, 18) = 0.005$, $p = 0.946$, $\eta_p^2 < 0.001$.
- ✓ There was a significant interaction effect between the attractiveness of the date and the dating strategy of the date, $F(1.92, 34.62) = 80.43$, $\eta_p^2 = 0.82$. This effect indicates that the desirability of dates of different levels of attractiveness differed according to whether they played hard to get. Contrasts compared each level of attractiveness to average, across dating strategies. These contrasts revealed significant interactions when comparing ratings of dates who played hard to get and those who acted normally when comparing attractive to average-looking dates, $F(1, 18) = 43.26$, $\eta_p^2 = 0.71$, and to low attractiveness dates compared to average-looking dates $F(1, 18) = 30.23$, $\eta_p^2 = 0.63$. The interaction plot shows that although interest increased as attractiveness increased regardless of the date's strategy, this increase was more pronounced when they adopted a 'hard to get' strategy, suggesting that when charisma is ignored, attractiveness had a greater impact on ratings of the date when they adopted a 'hard to get' strategy.
- ✓ There was a significant interaction effect between the level of charisma of the date and the strategy of the date, $F(1.87, 33.62) = 62.45$, $\eta_p^2 = 0.78$, indicating that the desirability of dates of different levels of charisma differed according to whether they played hard to get. Contrasts compared each level of charisma to average charisma across dating strategies. These contrasts revealed significant interactions when comparing ratings of dates when they played hard to get relative to when they did not, in dates with average charisma compared to those with high charisma, $F(1, 18) = 27.20$, $\eta_p^2 = 0.60$, and no charisma, $F(1, 18) = 33.69$, $\eta_p^2 = 0.65$. The interaction plot reveals that interest increased as charisma increased, but this increase was more pronounced when the date did not play hard to get, suggesting charisma influenced dating preferences more when the dates acted normally than when they played hard to get.
- ✓ There was a significant charisma × attractiveness interaction, $F(3.20, 57.55) = 36.63$, $\eta_p^2 = 0.67$, indicating that the desirability of dates of different levels of charisma differed according to their attractiveness. Contrasts compared each level of charisma to the middle category of average charisma across each level of attractiveness compared to the category of average attractiveness. The first contrast revealed a significant interaction when comparing average-looking dates to highly attractive ones when the date had high charisma compared to average charisma, $F(1, 18) = 21.94$, $\eta_p^2 = 0.55$, and tells us that as attractiveness increased there was a greater increase in interest when charisma was average compared to when it was high. The second contrast, which compared average-looking dates to highly attractive ones when the date had average charisma compared to low, was not significant, $F(1, 18) = 4.09$, $p = 0.058$, $\eta_p^2 = 0.19$. This result suggests that as attractiveness increased the increase in interest in the date was similar for low- and average-charisma dates. The third contrast, which compared average-looking to unattractive dates when they had high compared to average charisma, was not significant, $F(1, 18) = 6.23$, $p = 0.022$, $\eta_p^2 = 0.26$. This result suggests that as attractiveness increased, the increase in interest in the date was similar for high- and average-charisma dates. The final contrast compared average-looking to unattractive dates for those with average charisma compared to low. This contrast was significant, $F(1, 18) = 88.60$, $\eta_p^2 = 0.83$, and suggests that as attractiveness increased the increase in interest in dates was significantly greater for dates with average charisma than those with low charisma.

- ✓ Finally, the looks × charisma × strategy interaction was significant, $F(3.20, 57.55) = 24.12$, $\eta_p^2 = 0.57$. This indicates that the looks × charisma interaction described previously was moderated by whether the date played hard to get. Contrasts were used to break down this interaction; these contrasts compared scores at each level of charisma relative to the middle category of average charisma across each level of attractiveness relative to the category of average attractiveness when dates played hard to get compared to when they did not. The first contrast revealed a non-significant effect of playing hard to get when comparing average-looking dates to highly attractive dates when the date had average charisma compared to high, $F(1, 18) = 0.93$, $p = 0.348$, $\eta_p^2 = 0.05$. This effect suggests that regardless of whether the date played hard to get, as the date's attractiveness increased there was a greater increase in interest when charisma was average compared to high. The second contrast investigated the effect of playing hard to get when comparing average-looking to highly attractive dates for low- relative to average-charisma dates, $F(1, 18) = 60.67$, $\eta_p^2 = 0.77$. This finding indicates that for dates with average charisma, the increase in interest as attractiveness increases was unaffected by whether the date played hard to get, but for low-charisma dates, the increase in interest if these dates were attractive rather than average-looking was more dramatic when they played hard to get. The third contrast looked at the effect of playing hard to get when comparing unattractive dates to average-looking dates when they had high compared to average charisma, $F(1, 18) = 11.70$, $p = 0.003$, $\eta_p^2 = 0.39$. For dates with average charisma, the increase in interest as attractiveness increased was unaffected by playing hard to get, but for dates who had high charisma, the increase in interest for average-looking dates compared to unattractive dates was greater when they played hard to get. The final contrast looked at the effect of playing hard to get when comparing average-looking to unattractive dates for those with average compared to low charisma, $F(1, 18) = 1.33$, $p = 0.263$, $\eta_p^2 = 0.07$. This effect suggests that regardless of whether dates played to get, as physical attractiveness increased, the increase in interest in dates with average charisma was significantly greater than for low-charisma dates.



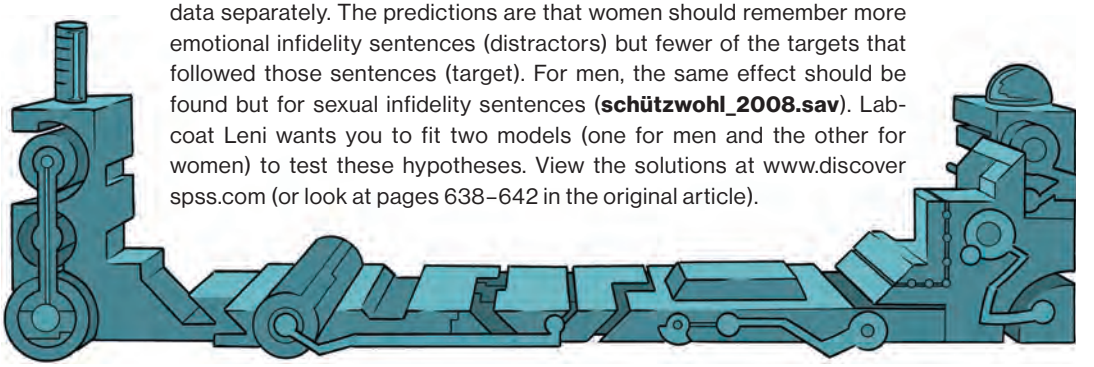
Labcoat Leni's Real Research 16.2
Keep the faith(ful)? ©

Schützwohl, A., (2008). *Personality and Individual Differences*, *44*, 633–644.

People can be jealous when they think that their partner is being unfaithful. An evolutionary view suggests that men and women have evolved distinctive types of jealousy: specifically, a woman's sexual infidelity deprives her mate of a reproductive opportunity and could burden him with years investing in a child that is not his. Conversely, a man's sexual infidelity does not burden his mate with unrelated children, but may divert his resources from his mate's progeny. This diversion of resources is signalled by emotional attachment to another female. Consequently, men's jealousy mechanism should have evolved to prevent a mate's *sexual* infidelity, whereas in women it has evolved to prevent emotional infidelity. If this is the case, women should be 'on the

lookout' for emotional infidelity, whereas men should be watching out for sexual infidelity.

Whether or not you buy into this theory, it can be tested. Achim Schützwohl exposed men and women to sentences on a computer screen (Schützwohl, 2008). At each trial, participants saw a target sentence that was emotionally neutral (e.g., 'The gas station is at the other side of the street'). However, before each of these targets, a distractor sentence was presented that could also be affectively neutral, or could indicate sexual infidelity (e.g., 'Your partner suddenly has difficulty becoming sexually aroused when he and you want to have sex') or emotional infidelity (e.g., 'Your partner doesn't say "I love you" to you anymore'). Schützwohl reasoned that if these distractor sentences grabbed a person's attention then (1) they would remember them, and (2) they would not remember the target sentence that came afterwards (because their attentional resources were focused on the distractor). These effects should show up only in people currently in a relationship. The outcome was the number of sentences that a participant could remember (out of 6), and the predictors were whether the person had a partner or not (**relationship**), whether the trial used a neutral distractor, an emotional infidelity distractor or a sexual infidelity distractor, and whether the sentence was a distractor or the target following a distractor. Schützwohl analysed the men's and women's data separately. The predictions are that women should remember more emotional infidelity sentences (distractors) but fewer of the targets that followed those sentences (target). For men, the same effect should be found but for sexual infidelity sentences (**schützwohl_2008.sav**). Labcoat Leni wants you to fit two models (one for men and the other for women) to test these hypotheses. View the solutions at www.discoverspss.com (or look at pages 638–642 in the original article).



16.8 Jane and Brian's story Ⓐ

Jane had seen a lot of Brian since her mother's illness. He was stable, reassuring and emotional. Their meetings had developed a predictable pattern: Jane would demand that he talked to her about anything he wanted to, he'd nervously recite whatever he'd been learning about as a safe emotion-free topic, and once his soothing words of statistics had calmed her, she'd slowly reveal a little more of her past. This routine had gone on for weeks. It was only now that it occurred to her to ask him about his own family. He smiled when she asked, but with tears in his eyes.

'My dad's awesome,' he said. 'He brought me up. He worked 50+ hours a week and still always had time for me. He was there getting me ready for school every morning, and there to pick me up. He never missed anything. *Anything*. The guy never slept for cramming in work while I was asleep so he could be there when I was awake. I don't know how he did it. He gave up his life to give me a happy childhood.'

'And your mum?'

'She died,' he said after a long pause. 'I was 10.'

Jane wasn't equipped for this kind of revelation. 'That must have been hard,' she said.

'The weird thing is I don't remember her. I've seen photos of us but they feel like someone else's life. My earliest memory is my dad telling me what had happened. I remember shaking, I remember him holding me, but everything before that day is gone.' The levee in his eyes broke.

Jane did something she'd never done in her adult life. She put her arm around another person and hugged him. She quickly felt awkward, not knowing whether it was weirder to keep holding him or to let go.

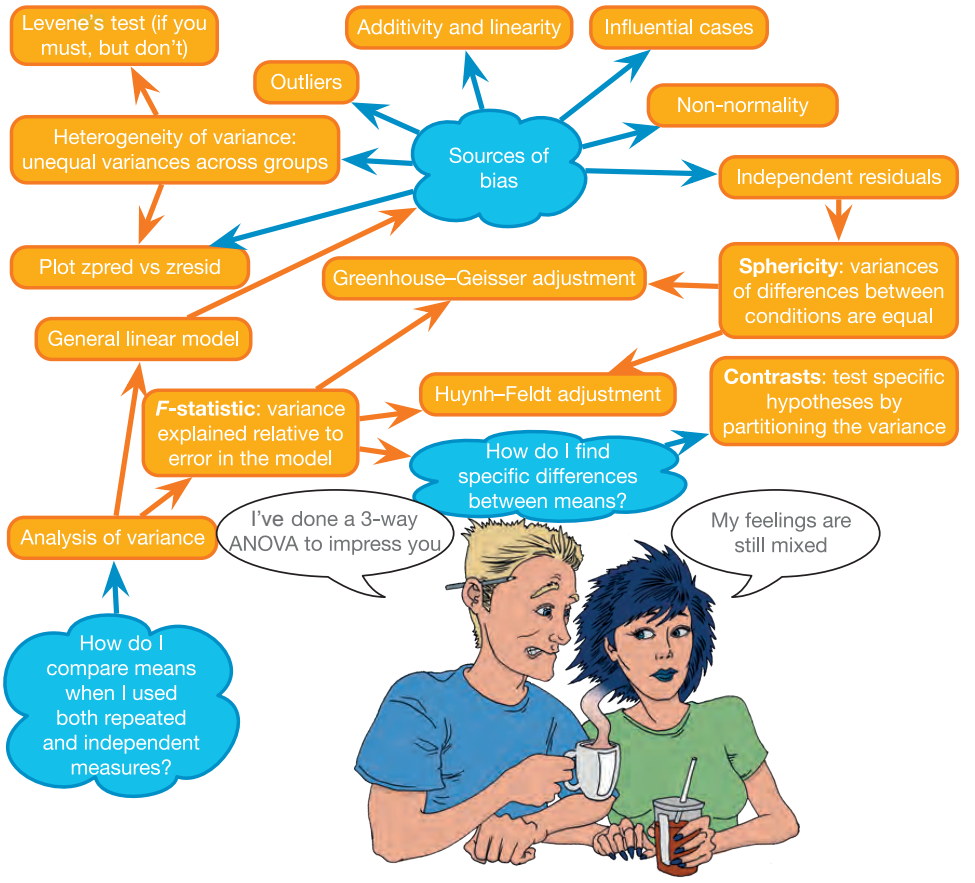


Figure 16.20 What Brian learnt from this chapter

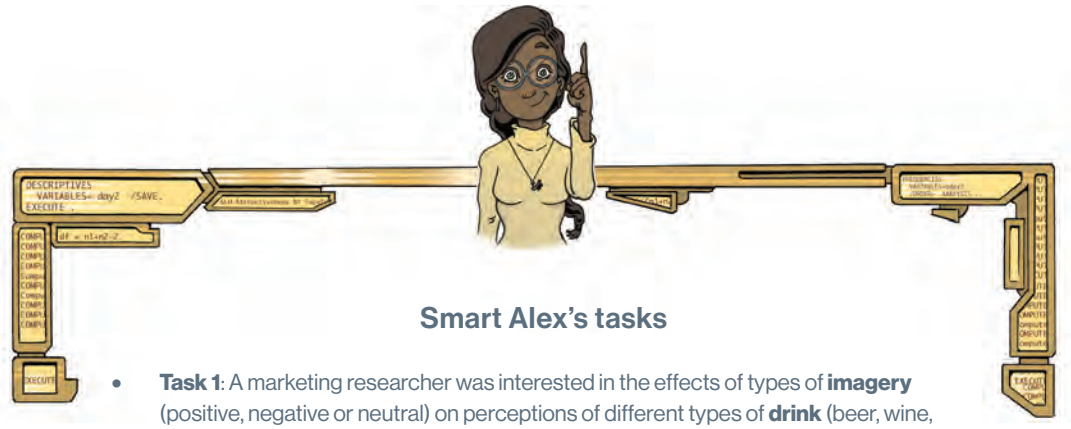
16.9 What next? Ⓑ

We've discovered in this chapter that if you are boring then it doesn't matter how attractive you are. Unless you play hard to get and are *really* attractive. This is why as a 16–18-year-old my life was so complicated, because I wouldn't know how to be hard to get, I didn't think I was attractive and where on earth do you discover your hidden charisma? Before you get out your tiny violins, I had one small bit of dating fortune. The girl (Nicola) I was particularly keen on at 16 was, as it turned out, keen on me too. I refused to believe this for quite some time. All our friends were getting so bored with us declaring our undying love for each other to them but not actually speaking to each other that they held an intervention. At a party one evening all of Nicola's friends had spent hours convincing me to ask her on a date, guaranteeing me that she would say 'yes'. I psyched myself up, I was going to do it, I was actually going to ask a girl out on a date. My whole life had been leading up to this moment,

I thought to myself, and I must not do anything to ruin it. By the time she arrived, I wasn't in a good way – my nerves had got the better of me, and at the time I coped with that by drinking alcohol. Too much alcohol. Later that evening, once I was back to a vaguely cohesive state, my friend Paul Spreckley (see Figure 11.1) physically carried Nicola from another room, put her next to me and said something to the effect of 'Andy, I'm going to sit here until you ask her out.' He had a long wait, but eventually, miraculously, the words came out of my mouth. What happened next is the topic for another book, not about statistics.

16.10 Key terms that I've discovered

Mixed ANOVA Mixed design

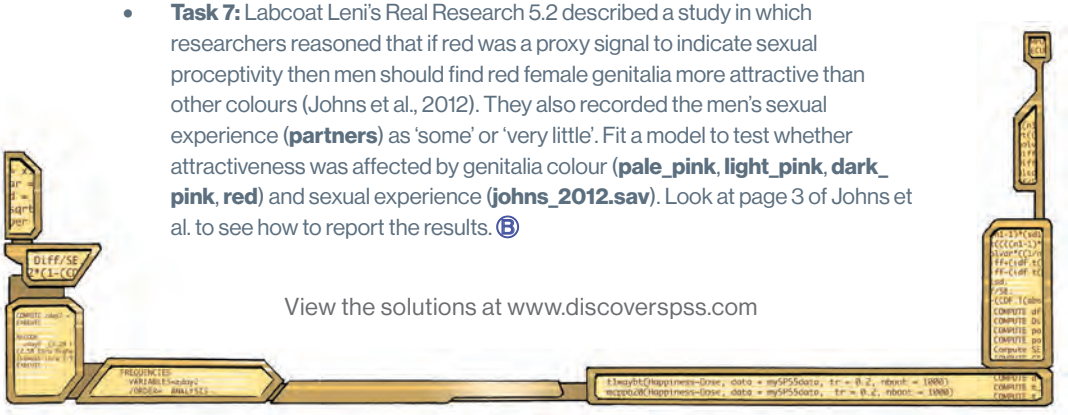


Smart Alex's tasks

- **Task 1**: A marketing researcher was interested in the effects of types of **imagery** (positive, negative or neutral) on perceptions of different types of **drink** (beer, wine, water). Participants viewed videos of different drink products in the context of positive, negative or neutral imagery and then rated the products on a scale from –100 (extremely dislike) through 0 (neutral) to 100 (extremely like). Those who identify as men and women might respond differently to the products, so participants self-reported their **gender** (a between-group variable). The data are in the file **mixed_attitude.sav**. Analyse the data to see whether there is a combined effect of **imagery**, **drink** and **gender** on ratings of the products. ©
- **Task 2**: Text messaging and Twitter encourage communication using abbreviated forms of words (if u no wat I mean). A researcher wanted to see the effect this had on children's understanding of grammar. One group of 25 children was encouraged to send text messages on their mobile phones over a 6-month period. A second group of 25 was forbidden from sending text messages for the same period (to ensure adherence, this group were given armbands that administered painful shocks in the presence of a phone signal). The outcome was a score on a grammatical test (as a percentage) that was measured both before and after the experiment. The data are in the file **text_messages.sav**. Does using text messages affect grammar? ©
- **Task 3**: A researcher hypothesized that reality TV show contestants start off with personality disorders that are exacerbated by being forced to spend time with people as attention-seeking as them (see Chapter 1). To test this hypothesis, she gave eight contestants a questionnaire measuring personality disorders before and after they entered the show. A second group of eight people were given the questionnaires at the same time; these people were short-listed to go on the show, but never did. The data are in **reality_tv.sav**. Does entering a reality TV competition give you a personality disorder? Ⓑ

- **Task 4:** Angry Birds is a video game in which you fire birds at pigs. A (fabricated) study was set up in which people played Angry Birds and a control game (Tetris) over a 2-year period (1 year per game). They were put in a pen of pigs for a day before the study, and after 1 month, 6 months and 12 months. Their violent acts towards the pigs were counted. Does playing Angry Birds make people more violent to pigs compared to a control game? (**angry_pigs.sav**) Ⓑ
- **Task 5:** A different study was conducted with the same design as in Task 4. The only difference was that the participant's violent acts in real life were monitored before the study, and after 1 month, 6 months and 12 months. Does playing Angry Birds make people more violent in general compared to a control game? (**angry_real.sav**) Ⓑ
- **Task 6:** My wife believes that she has received fewer friend requests from random men on Facebook since she changed her profile picture to a photo of us both. Imagine we took 40 women who had profiles on a social networking website; 17 of them had a relationship status of 'single' and the remaining 23 had their status as 'in a relationship' (**relationship_status**). We asked these women to set their profile picture to a photo of them on their own (**alone**) and to count how many friend request they got from men over 3 weeks, then to switch it to a photo of them with a man (**couple**) and record their friend requests from random men over 3 weeks. Fit a model to see if friend requests are affected by relationship status and type of profile picture (**profile_pic.sav**). Ⓑ
- **Task 7:** Labcoat Leni's Real Research 5.2 described a study in which researchers reasoned that if red was a proxy signal to indicate sexual proceptivity then men should find red female genitalia more attractive than other colours (Johns et al., 2012). They also recorded the men's sexual experience (**partners**) as 'some' or 'very little'. Fit a model to test whether attractiveness was affected by genitalia colour (**pale_pink**, **light_pink**, **dark_pink**, **red**) and sexual experience (**johns_2012.sav**). Look at page 3 of Johns et al. to see how to report the results. Ⓑ

View the solutions at www.discoverspss.com





MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

17.1 What will this chapter tell me? 784
17.2 Introducing MANOVA 785
17.3 The theory behind MANOVA 787
17.4 Practical issues when conducting MANOVA 800
17.5 MANOVA using SPSS 802
17.6 Interpreting MANOVA 804
17.7 Reporting results from MANOVA 810
17.8 Following up MANOVA with discriminant analysis 811
17.9 Interpreting discriminant analysis 814
17.10 Reporting results from discriminant analysis 817
17.11 The final interpretation 818
17.12 Jane and Brian's story 819
17.13 What next? 819
17.14 Key terms that I've discovered 821
Smart Alex's tasks 821

Letters A to D indicate increasing difficulty

17.1 What will this chapter tell me? Ⓑ

Having had what little confidence I had squeezed out of me by my first forays into dating and my band's unqualified failure to have an impact on the musical world, as I reached adulthood I decided that I could either kill myself or get a cat. I'd wanted to do both for years but when I was introduced to a little 4-week-old bundle of gingerness the choice was made. Fuzzy (as I named him) was born on 8 April 1996 and was my right-hand feline for over 20 years. Like the Cheshire Cat in Lewis Carroll's *Alice's Adventures in Wonderland*,[1] he used to vanish and reappear seemingly at will: I'd go to get clothes from my wardrobe and spot a ginger face peering out at me, I'd put my pants in the laundry basket and he'd look up at me from a pile of smelly socks, I'd go to have a bath and he'd be sitting in it, and I'd shut the bedroom door yet wake up to find him asleep next to me. His best vanishing act was when I moved house one time. He'd been locked up in his travel basket (which he hated) during the move, so once we were in our new house I thought I'd let him out as soon as possible. I found a quiet room, checked the doors and windows to make sure he couldn't escape, opened the basket, gave him a cuddle and left him to get to know his new territory. When I returned five minutes later, he was gone. The door had been shut, the windows closed and the walls were solid (I checked). He had literally vanished into thin air and he didn't even leave behind his smile.

We saw in Chapter 12 that there is a belief that having a pet is good for your mental health, and before his dramatic disappearance Fuzzy certainly was. If you wanted to test this idea you could compare people with pets against those without to see if they had better mental health. However, the term *mental health* covers a wide range of concepts including (to name a few) anxiety, depression, general distress and psychosis. As such, we might have several outcome measures, and the linear model we've looked at so far only deals with one. That is, until now, where we discover that it can mutate into MANOVA. Yes, it's as scary as it sounds.



Figure 17.1 Fuzzy doing some light reading

1 This is one of my favourite books from my childhood. For those who haven't read it, the Cheshire Cat is a big fat cat mainly remembered for vanishing and reappearing out of nowhere; on one occasion it vanished leaving only its smile behind.

17.2 Introducing MANOVA Ⓑ



Over Chapters 10–16, we have seen how the general linear model (GLM) can be used to detect group differences on a single outcome. However, there may be circumstances in which we are interested in several outcomes, and in these cases we use **multivariate analysis of variance** (or **MANOVA**). The principles of the linear model extend to MANOVA in that we can use MANOVA when there is one independent/predictor variable or several, we can look at interactions between outcome variables and we can do contrasts to see which groups differ. When we have only one outcome variable the model is known as **univariate** (meaning 'one variable') but when we include several outcome variables simultaneously the model is **multivariate** (meaning 'many variables'). There is a lengthy theory section explaining the workings of MANOVA, but for those of you who value the little time you have on Earth, accept that we're extending the linear model again and go straight to the sections on applying and interpreting MANOVA. This process leads us to another statistical tool known as *discriminant function analysis* (or *discriminant analysis* for short).



If we have scores on several outcome variables we could fit separate linear models (*F*-statistics) to each outcome (it is not unusual for researchers to do this). However, we learnt in Section 2.10.8 that when we carry out multiple tests on the same data the Type I errors mount up. For this reason, we shouldn't really fit separate linear models to each outcome variable. Also, if separate models are fitted to each outcome, then any relationship between these outcome variables is ignored – we lose this important information. By including all outcome variables in the model MANOVA factors in the relationship between them. Related to this point, separate models can tell us only whether groups differ along a single dimension, whereas MANOVA has the power to detect whether groups differ along a combination of dimensions.

For example, we might be able to distinguish people who are married, living together or single by their happiness. 'Happiness' is a complex construct, so we might want to measure their happiness with work, socially, sexually and within themselves (self-esteem). It might not be possible to distinguish people who are married, living together or single by only one aspect of happiness (which is what a univariate model tests), but these groups might be distinguished by *a combination* of their happiness across all four domains (which is what a MANOVA tests). In this sense MANOVA has greater potential power to detect an effect (see Jane Superbrain Box 17.1).

17.2.1 Choosing outcomes Ⓑ

MANOVA is probably looking like a pretty good way to measure hundreds of outcome variables and then sling them into an analysis without getting accused of *p*-hacking. Not so. It is a bad idea to lump outcome measures together in a MANOVA unless you have a good theoretical or empirical basis for doing so. The adage of 'garbage in, garbage out' applies here. Where there is a good theoretical basis for including some, but not all, of your outcome measures, then fit separate models: one for the outcomes being tested on a heuristic basis and one for the theoretically meaningful outcomes. The point here is not to include lots of outcome variables in a MANOVA just because you measured them.

17.2.2 An intrusive example Ⓑ

Obsessive compulsive disorder (OCD) is a mental health problem characterized by intrusive images or thoughts that the sufferer finds abhorrent – in my case the thought of someone fitting non-robust



Jane Superbrain 17.1
The power of MANOVA ©

In theory, MANOVA has greater power than ANOVA to detect effects because it takes account of the correlations between outcome variables (Huberty & Morris, 1989). However, the issue is complicated (when isn't it?). The evidence is contradictory, with some studies showing *diminishing* power as the correlation between outcome variables increases, whereas others show that power with high correlations between outcome variables is generally higher than for moderate correlations (Stevens, 1980). Cole et al. (1994) suggest that if you are expecting to find a large effect, MANOVA will have greater power if the measures are somewhat different (even negatively correlated) and if the group differences are in the same direction for each measure. If you have two outcome variables, one of which exhibits a large group difference and one of which exhibits a small or no group difference, then power will be increased if these variables are highly correlated. Although Cole et al.'s work is limited to the case where two groups are being compared, the take-home message is that if you are interested in how powerful the MANOVA is likely to be you should consider not only the correlation between outcome variables but also the size and pattern of group differences that you expect to get.



models, but more commonly thoughts with themes of aggression, sexuality or disease/contamination (Julien et al., 2007). These thoughts lead the sufferer to engage in activities to neutralize the unpleasantness of these thoughts (these activities can be mental, such as doing a MANOVA in my head to make me feel better about the non-robust models in the world, or physical, such as touching the floor 23 times so that you won't murder your statistics lecturer). A group of clinical psychologists were interested in the efficacy of two different interventions for OCD offered at their clinic: cognitive behaviour therapy (CBT) and behaviour therapy (BT).[2] A group who were awaiting treatment acted as a control (a no treatment condition, NT). Most psychopathologies have both behavioural and cognitive elements to them. For example, for someone with OCD who has intrusive thoughts about germs and contamination, the OCD might manifest itself both in the number of times they wash their hands (behaviour) and their *thoughts about* washing their hands (cognition). To gauge the success of therapy, it is not enough to look only at behavioural outcomes (such as whether obsessive behaviours are

2 A note for non-psychologists: to simplify somewhat, behaviour therapy targets the symptomatic behaviours, cognitive therapy targets symptomatic cognitions and CBT targets both.

reduced); we need to look at whether cognitions are changed too. Hence, the clinical psychologist measured two outcomes: the occurrence of obsession-related behaviours (**actions**) and the occurrence of obsession-related cognitions (**thoughts**) on a single day. The data are in Table 17.1 (and **ocd.sav**). Service users were randomly assigned to group 1 (CBT), group 2 (BT) or group 3 (NT), and within these groups all participants had both actions and thoughts measured.

Table 17.1 Data from ocd.sav

	DV 1: Actions			DV 2: Thoughts		
Group:	CBT (1)	BT (2)	NT (3)	CBT (1)	BT (2)	NT (3)
	5	4	4	14	14	13
	5	4	5	11	15	15
	4	1	5	16	13	14
	4	1	4	13	14	14
	5	4	6	12	15	13
	3	6	4	14	19	20
	7	5	7	12	13	13
	6	5	4	15	18	16
	6	2	6	16	14	14
	4	5	5	11	17	18
$\bar{X}$	4.90	3.70	5.00	13.40	15.20	15.00
s	1.20	1.77	1.05	1.90	2.10	2.36
s^2	1.43	3.12	1.11	3.60	4.40	5.56
	$\bar{X}_{grand\ (Actions)} = 4.53$			$\bar{X}_{grand\ (Thoughts)} = 14.53$		
	$S^2_{grand\ (Actions)} = 2.1195$			$S^2_{grand\ (Thoughts)} = 4.8782$		

17.3 The theory behind MANOVA ©

We have seen before that when we have a single outcome variable, we are often interested in computing an *F*-statistic that represents how much variance can be explained by the fact that certain scores appear in certain groups (which in experimental research represents our manipulation) relative to the error in prediction in the model (see Chapters 9 and 12). With multiple outcomes we basically want to do the same, but with the complication of having several outcome measures. To achieve this aim we need a multivariate analogue of the sums of squares that we used for univariate models; these are the sum of squares due to the model/grouping variable (the model sum of squares, SS_M), the error in prediction from the model (the residual sum of squares, SS_R) and of course the total amount of variation in the outcome(s) that needs to be explained (SS_T). It turns out that matrices are a good way to operationalize multivariate versions of these sums of squares (they're called **sum of squares and cross-products (SSCP) matrices**).

17.3.1 Matrices Ⓑ

Despite what Hollywood would have you believe, a **matrix** does not enable you to jump acrobatically through the air, Ninja style, as time seemingly slows to a point where you can gracefully contort to

avoid high-velocity objects. I have worked with matrices many times, but I have never (to my knowledge) stopped time and would certainly end up in a pool of my own innards if I ever tried to dodge a bullet. The mundane reality is that a matrix is a grid of numbers arranged in columns and rows (read the preface to explain some terminology that I'm about to use). As we shall see, an SSCP matrix is a grid of numbers where each row and column represent an outcome measure, the diagonal elements contain the sums of squares for each outcome, and the off-diagonal elements contain the cross-products between pairs of outcomes.

We end up with a matrix, **H**, that represents the systematic variance (or the model sum of squares for all variables) and is called the **hypothesis sum of squares and cross-products matrix**, or **hypothesis SSCP**. We also compute a matrix, **E**, that represents the unsystematic variance (the residual sums of squares for all variables) and is called the **error sum of squares and cross-products matrix**, or **error SSCP**. Finally, there is a matrix, **T**, that represents the total amount of variance present for each outcome variable (the total sums of squares for each outcome), which is called the **total sum of squares and cross-products matrix**, or **total SSCP**. It should be obvious why these matrices are referred to as sum of squares matrices, but why is there a reference to cross-products in their name?



Cross-products represent a total value for the combined error between two variables (in some sense they represent an unstandardized estimate of the total correlation between two variables). As such, whereas the sum of squares of a variable is the total squared difference between the observed values and the mean value, the cross-product is the total combined error *between* two variables. I mentioned earlier that MANOVA had the power to account for correlation between outcome variables, and it does this by using these cross-products. Later, I will show how these SSCP matrices are used in the same way as the sums of squares (SS_M, SS_R and SS_T) in univariate linear models to derive test statistics that are multivariate equivalents of the *F*-statistic (i.e., they represent the ratio of systematic to unsystematic variance in the model).

To begin with let's calculate univariate *F*-statistics on each of the two outcome variables in the OCD example (see Table 17.1). I will draw heavily on the assumption that you have read Chapter 12.

17.3.2 Univariate *F* for outcome 1 (**actions**) Ⓑ

There are three sums of squares that need to be calculated. First, we assess how much variability there is to be explained within the outcome (SS_T), and then break this variability down into that explained by the model (SS_M) and the error in prediction (SS_R). Referring to the equations in Chapter 12, we calculate each of these values as follows.

The total sum of squares, $SS_{T(Actions)}$, is obtained by calculating the difference between each of the 30 scores and the mean of those scores, then squaring these differences and adding them. Alternatively, you calculate the variance for the action scores (regardless of which group the score falls into) and multiply this value by the number of scores minus 1:

$$\begin{aligned} SS_T &= s^2_{grand}(N-1) \\ &= 2.1195(30-1) \\ &= 2.1195 \times 29 \\ &= 61.47 \end{aligned} \qquad (17.1)$$

The degrees of freedom will be $N - 1 = 29$.

The model sum of squares, $SS_{M(Actions)}$, is calculated by taking the difference between each group mean and the grand mean, squaring it, multiplying by the number of scores in the group and then adding the values for the groups together:

$$\begin{aligned} SS_M &= 10(4.90 - 4.53)^2 + 10(3.70 - 4.53)^2 + 10(5.00 - 4.53)^2 \\ &= 10(0.37)^2 + 10(-0.83)^2 + 10(0.47)^2 \\ &= 1.37 + 6.89 + 2.21 \\ &= 10.47 \end{aligned} \tag{17.2}$$

The degrees of freedom will be $k - 1 = 2$.

The residual sum of squares, $SS_{R(Actions)}$, is calculated by taking the difference between each score and the mean of the group from which it came. These differences are then squared and added together. Alternatively, multiply each group variance by the number of scores minus 1 and then add the results:

$$\begin{aligned} SS_R &= s^2_{CBT}(n_{CBT} - 1) + s^2_{BT}(n_{BT} - 1) + s^2_{NT}(n_{NT} - 1) \\ &= 1.433(10 - 1) + 3.122(10 - 1) + 1.111(10 - 1) \\ &= (1.433 \times 9) + (3.122 \times 9) + (1.111 \times 9) \\ &= 12.9 + 28.1 + 10 \\ &= 51 \end{aligned} \tag{17.3}$$

The degrees of freedom will be the sample size of each group minus 1 (e.g., 9) multiplied by the number of groups, $3 \times 9 = 27$.

The next step is to calculate the average sums of squares (the mean square) by dividing the sums of squares by their degrees of freedom:

$$\begin{aligned} MS_M &= \frac{SS_M}{df_M} = \frac{10.47}{2} = 5.235 \\ MS_R &= \frac{SS_R}{df_R} = \frac{51}{27} = 1.889 \end{aligned} \tag{17.4}$$

The *F*-statistic is the mean squares for the model divided by the mean squares for the error in the model:

$$F = \frac{MS_M}{MS_R} = \frac{5.235}{1.889} = 2.771 \tag{17.5}$$

17.3.3 Univariate *F* for outcome 2 (thoughts) Ⓑ

The three sums of squares for thoughts are calculated in the same way as for actions (the degrees of freedom for each is the same as above). For the total sum of squares, $SS_{T(Thoughts)}$, we have:

$$\begin{aligned} SS_T &= s^2_{grand}(n - 1) \\ &= 4.8782(30 - 1) \\ &= 4.8782 \times 29 \\ &= 141.47 \end{aligned} \tag{17.6}$$

The model sum of squares, $SS_{M(Thoughts)}$ is:

$$\begin{aligned} SS_M &= 10(13.40 - 14.53)^2 + 10(15.2 - 14.53)^2 + 10(15 - 14.53)^2 \\ &= 10(-1.13)^2 + 10(0.67)^2 + 10(0.47)^2 \\ &= 12.77 + 4.49 + 2.21 \\ &= 19.47 \end{aligned} \tag{17.7}$$

The residual sum of squares, $SS_{R(Thoughts)}$ is:

$$\begin{aligned} SS_R &= s^2_{CBT}(n_{CBT} - 1) + s^2_{BT}(n_{BT} - 1) + s^2_{NT}(n_{NT} - 1) \\ &= 3.6(10 - 1) + 4.4(10 - 1) + 5.56(10 - 1) \\ &= (3.6 \times 9) + (4.4 \times 9) + (5.56 \times 9) \\ &= 32.4 + 39.6 + 50 \\ &= 122 \end{aligned} \tag{17.8}$$

The mean sums of squares are the sums of squares divided by the degrees of freedom:

$$\begin{aligned} MS_M &= \frac{SS_M}{df_M} = \frac{19.47}{2} = 9.735 \\ MS_R &= \frac{SS_R}{df_R} = \frac{122}{27} = 4.519 \end{aligned} \tag{17.9}$$

The *F*-statistic is the mean squares for the model divided by the residual mean squares:

$$F = \frac{MS_M}{MS_R} = \frac{9.735}{4.519} = 2.154 \tag{17.10}$$

17.3.4 The relationship between outcomes: cross-products Ⓑ

OK, we have the sums of squares associated with each outcome variable, now let's look at the relationship between them. If we want a measure of the relationship that is comparable to a sum of squares then we need something that quantifies the *total* relationship (because sums of squares are totals). We saw in Chapter 8 that the cross-product does this job. There are three relevant cross-products that correspond to the three sums of squares that we just calculated: the total cross-product, the cross-product for the model and a residual cross-product. Let's look at the total cross-product (CP_T) first.

The cross-product is the difference between the scores and the mean for one variable multiplied by the difference between the scores and the mean for another variable. In the case of the total cross-product, the mean of interest is the grand mean for each outcome variable (see Table 17.2). We can apply the cross-product equation described in Chapter 8 to the two outcome variables:

$$CP_T = \sum_{i=1}^{n}\left(x_{i(Actions)} - \bar{X}_{grand(Actions)}\right)\left(x_{i(Thoughts)} - \bar{X}_{grand(Thoughts)}\right) \tag{17.11}$$

For each outcome variable you take each score and subtract from it the grand mean for that variable. This leaves you with two values per participant (one for each outcome variable) that are multiplied together to get the cross-product for each participant. The total can then be found by adding the cross-products of all participants. Table 17.2 illustrates this process.

The total cross-product is a gauge of the overall relationship between the two variables. We can decompose this total. First, we can quantify how the relationship between the outcome variables is influenced by our experimental manipulation using the model cross-product (CP_M):

Table 17.2 Calculation of the total cross-product

Group	Actions	Thoughts	Actions – $\bar{X}_{grand\ (Actions)}$ (D_1)	Thoughts – $\bar{X}_{grand\ (Thoughts)}$ (D_2)	$D_1 \times D_2$
CBT	5	14	0.47	−0.53	−0.25
	5	11	0.47	−3.53	−1.66
	4	16	−0.53	1.47	−0.78
	4	13	−0.53	−1.53	0.81
	5	12	0.47	−2.53	−1.19
	3	14	−1.53	−0.53	0.81
	7	12	2.47	−2.53	−6.25
	6	15	1.47	0.47	0.69
	6	16	1.47	1.47	2.16
	4	11	−0.53	−3.53	1.87
BT	4	14	−0.53	−0.53	0.28
	4	15	−0.53	0.47	−0.25
	1	13	−3.53	−1.53	5.40
	1	14	−3.53	−0.53	1.87
	4	15	−0.53	0.47	−0.25
	6	19	1.47	4.47	6.57
	5	13	0.47	−1.53	−0.72
	5	18	0.47	3.47	1.63
	2	14	−2.53	−0.53	1.34
	5	17	0.47	2.47	1.16
NT	4	13	−0.53	−1.53	0.81
	5	15	0.47	0.47	0.22
	5	14	0.47	−0.53	−0.25
	4	14	−0.53	−0.53	0.28
	6	13	1.47	−1.53	−2.25
	4	20	−0.53	5.47	−2.90
	7	13	2.47	−1.53	−3.78
	4	16	−0.53	1.47	−0.78
	6	14	1.47	−0.53	−0.78
	5	18	0.47	3.47	1.63
$\bar{X}_{grand}$	4.53	14.53			$CP_T = \sum D_1 \times D_2 = 5.47$

$$\mathrm{CP_M} = \sum_{g=1}^{k} n\left[\left(\bar{x}_{g(\mathrm{Actions})} - \bar{X}_{\mathrm{grand(Actions)}}\right)\left(\bar{x}_{g(\mathrm{Thoughts})} - \bar{X}_{\mathrm{grand(Thoughts)}}\right)\right] \tag{17.12}$$

The $\mathrm{CP_M}$ is calculated in a similar way to the model sum of squares. First, the difference between each group mean and the grand mean is calculated for each outcome variable. The cross-product is calculated by multiplying the differences found for each group. Each product is then multiplied by the number of scores within the group (as was done with the sum of squares). This principle is illustrated in Table 17.3.

Table 17.3 Calculating the model cross-product

	$\bar{X}_{group}$ Actions	$\bar{X}_{group} \times \bar{X}_{grand}$ (D_1)	$\bar{X}_{group}$ Thoughts	$\bar{X}_{group} \times \bar{X}_{grand}$ (D_2)	$D_1 \times D_2$	$N(D_1 \times D_2)$
CBT	4.9	0.37	13.4	−1.13	−0.418	−4.18
BT	3.7	−0.83	15.2	0.67	−0.556	−5.56
NT	5.0	0.47	15.0	0.47	0.221	2.21
$\bar{X}_{grand}$	4.53		14.53		$\mathrm{CP_M} = \sum N(D_1 \times D_2) = -7.53$	

Finally, we can quantify how the relationship between the two outcome variables is influenced by individual differences/unmeasured variables using the residual cross-product ($\mathrm{CP_R}$):

$$\mathrm{CP_R} = \sum_{i=1}^{n}\left(x_{i(\mathrm{Actions})} - \bar{X}_{\mathrm{group(Actions)}}\right)\left(x_{i(\mathrm{Thoughts})} - \bar{X}_{\mathrm{group(Thoughts)}}\right) \tag{17.13}$$

The $\mathrm{CP_R}$ is calculated in a similar way to the total cross-product except that the group means are used rather than the grand mean. So, to calculate each of the difference scores, we take each score and subtract from it the mean of the group to which it belongs (see Table 17.4). The residual cross-product can also be calculated by subtracting the model cross-product from the total cross-product:

$$\begin{aligned}\mathrm{CP_R} &= \mathrm{CP_T} - \mathrm{CP_M} \\ &= 5.47 - (-7.53) \\ &= 13\end{aligned} \tag{17.14}$$

Each of the cross-products tells us something important about the relationship between the two outcome variables. Although I have used a simple scenario to keep the maths relatively simple, these principles can be easily extended to more complex scenarios. For example, if we have measured three outcome variables then the cross-products between pairs of outcome variables are calculated (as they were in this example). As the complexity of the situation increases, so does the amount of calculation that needs to be done. At times such as these the benefit of software like SPSS becomes apparent!

17.3.5 The total SSCP matrix (**T**) ©

An SSCP matrix is an example of **square matrix**. With two outcome variables (as in this example) the SSCP matrices will be 2 × 2 matrices, with three outcome variables they'd be 3 × 3 matrices and so on. The total SSCP matrix, T, contains the total sums of squares for each outcome variable and the total cross-product between the outcome variables. You can think of the first column and first row as representing one outcome variable and the second column and row as representing the second outcome variable (Figure 17.2). We calculated the values in the matrix earlier in this section, and if we place them in the appropriate cells of the matrix we get:

Table 17.4 Calculation of CP_R

Group	Actions	Actions – $\bar{X}_{grand\ (Actions)}$ (D_1)	Thoughts	Actions – $\bar{X}_{group\ (Thoughts)}$ (D_2)	$D_1 \times D_2$
CBT	5	0.10	14	0.60	0.06
	5	0.10	11	−2.40	−0.24
	4	−0.90	16	2.60	−2.34
	4	−0.90	13	−0.40	0.36
	5	0.10	12	−1.40	−0.14
	3	−1.90	14	0.60	−1.14
	7	2.10	12	−1.40	−2.94
	6	1.10	15	1.60	1.76
	6	1.10	16	2.60	2.86
	4	−0.90	11	−2.40	2.16
$\bar{X}_{CBT}$	4.9		13.4		$\Sigma = 0.40$
BT	4	0.30	14	−1.20	−0.36
	4	0.30	15	−0.20	−0.06
	1	−2.70	13	−2.20	5.94
	1	−2.70	14	−1.20	3.24
	4	0.30	15	−0.20	−0.06
	6	2.30	19	3.80	8.74
	5	1.30	13	−2.20	−2.86
	5	1.30	18	2.80	3.64
	2	−1.70	14	−1.20	2.04
	5	1.30	17	1.80	2.34
$\bar{X}_{BT}$	3.7		15.2		$\Sigma = 22.60$
NT	4	−1.00	13	−2.00	2.00
	5	0.00	15	0	0.00
	5	0.00	14	−1.00	0.00
	4	−1.00	14	−1.00	1.00
	6	1.00	13	−2.00	−2.00
	4	−1.00	20	5.00	−5.00
	7	2.00	13	−2.00	−4.00
	4	−1.00	16	1.00	−1.00
	6	1.00	14	−1.00	−1.00
	5	0.00	18	3.00	0.00
$\bar{X}_{NT}$	5		15		$\Sigma = -10.00$
					$CP_R = \sum D_1 \times D_2 = 13$



Figure 17.2 The components of the total SSCP

$$\mathbf{T} = \begin{pmatrix} 61.47 & 5.47 \\ 5.47 & 141.47 \end{pmatrix} \tag{17.15}$$

From the values in the matrix (and what they represent) it makes sense that the total SSCP represents both the total amount of variation that exists within the outcome variables and the total co-dependence that exists between them. Also note that the off-diagonal components are the same value (they are the total cross-product), because this value is equally important for both outcome variables.

17.3.6 The residual SSCP matrix (**E**) ©

The residual (or error) sum of squares and cross-product matrix, **E**, contains the residual sums of squares for each outcome variable and the residual cross-product between the two outcome variables. This SSCP matrix is like the total SSCP except that the information relates to the error in the model (Figure 17.3). Placing the values that we calculated earlier in this section in the appropriate cells of the matrix, we get:

$$\mathbf{E} = \begin{pmatrix} 51 & 13 \\ 13 & 122 \end{pmatrix} \tag{17.16}$$

The residual SSCP represents both the unsystematic variation that exists for each outcome variable and the co-dependence between the outcome variables that is due to unmeasured factors. As with the total SSCP, the off-diagonal elements are the same (they are the residual cross-product).

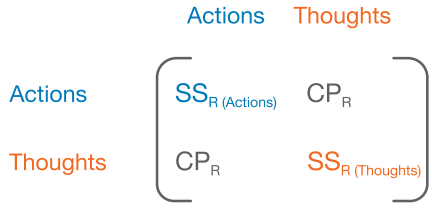


Figure 17.3 The components of the residual SSCP

17.3.7 The model SSCP matrix (**H**) ©

The model (or hypothesis) sum of squares and cross-product matrix, **H**, contains the model sums of squares for each outcome variable and the model cross-product between the two outcome variables (Figure 17.4). Placing the values that we calculated earlier in this section in the appropriate cells of the matrix, we get:

$$\mathbf{H} = \begin{pmatrix} 10.47 & -7.53 \\ -7.53 & 19.47 \end{pmatrix} \tag{17.17}$$

The model SSCP represents both the systematic variation that exists for each outcome variable and the co-dependence between the outcome variables that is due to the model (which in experimental research is the experimental manipulation).

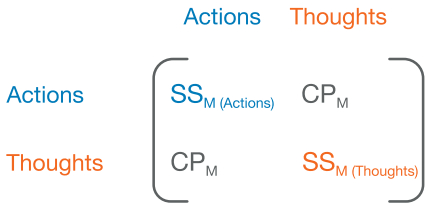


Figure 17.4 The components of the model SSCP

Matrices are additive, which means that you can add (or subtract) two matrices together by adding (or subtracting) corresponding elements. When we calculated univariate models we saw that the total sum of squares was the sum of the model sum of squares and the residual sum of squares (i.e., $SS_T = SS_M + SS_R$). The same is true in MANOVA, except that we add the SSCPs rather than single values:

$$\begin{aligned}\mathbf{T} &= \mathbf{H} + \mathbf{E} \\ &= \begin{pmatrix} 10.47 & -7.53 \\ -7.53 & 19.47 \end{pmatrix} + \begin{pmatrix} 51 & 13 \\ 13 & 122 \end{pmatrix} \\ &= \begin{pmatrix} 10.47 + 51 & -7.53 + 13 \\ -7.53 + 13 & 19.47 + 122 \end{pmatrix} \\ &= \begin{pmatrix} 61.47 & 5.47 \\ 5.47 & 141.47 \end{pmatrix}\end{aligned} \tag{17.18}$$

The demonstration that these matrices add up should (hopefully) reinforce the idea that the MANOVA calculations are conceptually the same as for univariate models – the difference is that matrices are used rather than single values.

17.3.8 $\mathbf{HE}^{-1}$: an analogue of *F* Ⓓ

The univariate *F* is the ratio of systematic variance to unsystematic variance (i.e., it is a function of SS_M divided by SS_R).[3] The conceptual equivalent would therefore be to divide the matrix **H** by the matrix **E**. The matrix equivalent to division is to multiply by what's known as the inverse of a matrix. So, an analogue of *F* would be to divide **H** by **E**, or in matrix terms to multiply **H** by the inverse of **E** (denoted by $\mathbf{E}^{-1}$). The resulting matrix is called $\mathbf{HE}^{-1}$ and it is a multivariate analogue of the univariate *F* (with respect to what it represents conceptually).

Calculating the inverse of a matrix is tricky, and there is no need for you to know how to do it to get a good conceptual grasp of MANOVA. If you're interested, Namboodiri (1984) is an accessible account of how to derive an inverse matrix, and having read it, you could look at Oliver Twisted. Everyone else can trust me that we end up with:

$$\begin{aligned}\mathbf{E}^{-1} &= \begin{pmatrix} 0.0202 & -0.0021 \\ -0.0021 & 0.0084 \end{pmatrix} \\ \mathbf{HE}^{-1} &= \begin{pmatrix} 0.2273 & -0.0852 \\ -0.1930 & 0.1794 \end{pmatrix}\end{aligned} \tag{17.19}$$

3 We use the mean squares, not the sum of squares, but the means squares are directly proportional to the sums of squares because they are the sums of squares divided by the degrees of freedom.

Remember that $\mathbf{HE}^{-1}$ represents the ratio of the systematic variance in the model to the unsystematic variance in the model, and so the resulting matrix is conceptually the same as the univariate *F*-statistic. There is another problem, though. In univariate models we get a single value for *F*, but, as Eq. 17.19 shows, when we divide matrices, we end up with several values (four in this case). In fact, $\mathbf{HE}^{-1}$ will always contain p^2 values, where p is the number of outcome variables. The problem is how to convert this matrix of values into a meaningful single value for which we can compute a *p*-value. This is the point at which we abandon any hope of understanding the maths and talk conceptually instead.

17.3.9 Discriminant function variates Ⓓ

The problem of having several values with which to assess statistical significance can be simplified by converting the outcome variables into underlying dimensions or factors (this process will be discussed in Chapter 18). Most of this book has looked at how we can use a linear model to predict an outcome variable. Linear models are made up of a combination of predictor variables, each of which makes a unique contribution to the model. We can do a similar thing here, except that we want to do the opposite: predict an independent variable from a set of outcome variables. It is possible to calculate underlying linear dimensions of the outcome variables known as *variates* (or sometimes called *components*). In this context, we want to use these linear variates to predict to which group a person belongs (i.e., whether they were given CBT, BT or no treatment). Because we are using them to discriminate groups of people/cases these variates are called *discriminant functions* or **discriminant function variates**.

That's the theory in relatively simple terms, but how do we discover these discriminant functions? Without going into too much detail, we use a mathematical procedure of maximization, such that the first discriminant function (V_1) is the linear combination of outcome variables that maximizes the differences between groups. It follows from this that the ratio of systematic to unsystematic variance (SS_M/SS_R) will be maximized for this first variate, but subsequent variates will have smaller values of this ratio. Remember that this ratio is an analogue of what the *F*-statistic represents in univariate models, and so in effect we obtain the maximum possible value of the *F*-statistic when we look at the first discriminant function. This variate can be described in terms of a linear model equation because it is a linear combination of the outcome variables:

$$\begin{aligned} y_i &= b_0 + b_1X_{1i} + b_2X_{2i} \\ V_{1i} &= b_0 + b_1\text{Outcome1}_{1i} + b_2\text{Outcome2}_{2i} \\ &= b_0 + b_1\text{Actions}_i + b_2\text{Thoughts}_i \end{aligned} \tag{17.20}$$

Eq. 17.20 shows the equation for two predictors and then extends this to show how a comparable form of this equation can describe discriminant functions. The *b*-values in the equation are weights that tell us something about the contribution of each outcome variable to the variate in question. In the linear models we have looked at in this book the values of *b* are estimated using the method of least squares. The values of *b* for the discriminant functions are obtained from the *eigenvectors* (see Jane Superbrain Box 9.3) of $\mathbf{HE}^{-1}$. We can ignore b_0 as well because it serves only to locate the variate in geometric space, which isn't necessary when we're using the variate to discriminate groups.

In a situation in which there are only two outcome variables and two groups to predict, there will be only one variate. This makes the scenario straightforward: by looking at the discriminant function of the outcome variables, rather than looking at the outcome variables themselves, we can obtain a single value of SS_M/SS_R for the discriminant function, and then assess this value for significance. However, in more complex cases where there are more than two outcome variables or three or more categories to predict (as is the case in our example) there will be more than one variate. The number of variates obtained will be the smaller of p (the number of outcome variables) and $k - 1$ (where k is the number of categories/groups to be predicted). In our example, both p and $k - 1$ are 2, so we will find two variates.

I mentioned earlier that the *b*-values that describe the variates are obtained by calculating the eigenvectors of the matrix $\mathbf{HE}^{-1}$, and in fact, there will be two eigenvectors derived from this matrix: one with the *b*-values for the first variate, and one with the *b*-values of the second variate. Conceptually speaking, eigenvectors are the vectors associated with a given matrix that are unchanged by transformation of that matrix to a diagonal matrix (refer to Jane Superbrain Box 9.3 for a visual explanation of eigenvectors and eigenvalues). In an **identity matrix** the off-diagonal elements are zero (see the preface of this book) so by changing $\mathbf{HE}^{-1}$ into an identity matrix we can eliminate all of the off-diagonal elements (thus reducing the number of values to be considered for significance testing). Therefore, by calculating the eigenvectors and eigenvalues, we still end up with values that represent the ratio of systematic to unsystematic variance (because they are unchanged by the transformation), but there are considerably fewer of them.

The calculation of eigenvectors is extremely complex (but if this kind of thing makes you tingle with excitement then read Namboodiri, 1984), so trust me that for the matrix $\mathbf{HE}^{-1}$ the eigenvectors obtained are:

$$\begin{aligned}\text{eigenvector}_1 &= \begin{pmatrix} 0.603 \\ -0.335 \end{pmatrix} \\ \text{eigenvector}_2 &= \begin{pmatrix} 0.425 \\ 0.339 \end{pmatrix}\end{aligned} \tag{17.21}$$

Substituting these values into the two equations for the variates (Eq. 17.20), and bearing in mind we can ignore b_0, we obtain the models described in the following equation:

$$\begin{aligned}V_{1i} &= 0.603\text{Actions}_i - 0.335\text{Thoughts}_i \\ V_{2i} &= 0.425\text{Actions}_i + 0.339\text{Thoughts}_i\end{aligned} \tag{17.22}$$

It is possible to use the equations for each variate to calculate a score for each person on the variate. For example, the first participant in the CBT group carried out 5 obsessive actions, and had 14 obsessive thoughts. Therefore, this participant's score on variate 1 would be –1.675:

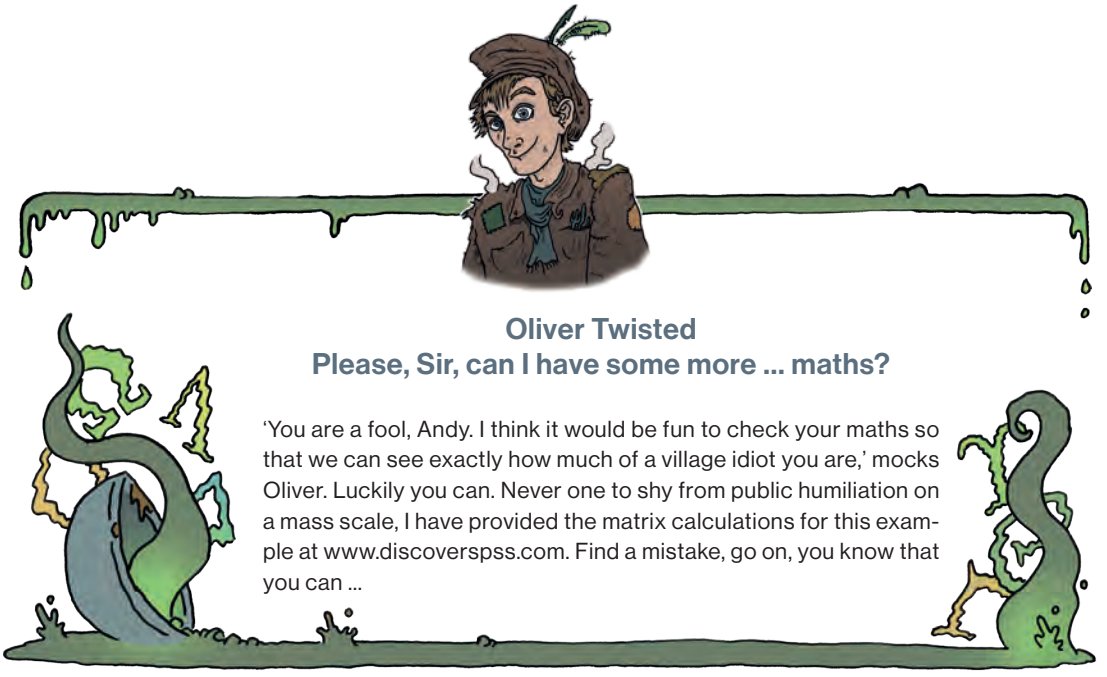
$$V_1 = (0.603 \times 5) - (0.335 \times 14) = -1.675 \tag{17.23}$$

Their score for variate 2 would be 6.87:

$$V_2 = (0.425 \times 5) + (0.339 \times 14) = 6.871 \tag{17.24}$$

If we calculated these variate scores for each participant and then calculated the SSCP matrices (e.g., **H**, **E**, **T** and $\mathbf{HE}^{-1}$) that we used previously, we would find that all of them have cross-products of zero. The reason for this is that the variates extracted from the data are orthogonal, which means that they are uncorrelated. In short, the variates extracted are independent dimensions constructed from a linear combination of the outcome variables that were measured.

This data reduction has a very useful property in that if we look at the matrix $\mathbf{HE}^{-1}$ calculated from the variate scores (rather than the outcome variables) we find that the off-diagonal elements (the cross-products) are zero. The diagonal elements of this matrix represent the ratio of the systematic variance to the unsystematic variance (i.e., SS_M/SS_R) for each of the underlying variates. So, in this example, this means that instead of having four values representing the ratio of systematic to unsystematic variance, we now have only two. This reduction may not seem a lot. However, in general if we have *p* outcome variables, then ordinarily we would end up with p^2 values representing the ratio of systematic to unsystematic variance; by using discriminant functions, we reduce this number back to *p*. For example, with four outcome variables we would end up with four values rather than 16.



Oliver Twisted
Please, Sir, can I have some more ... maths?

'You are a fool, Andy. I think it would be fun to check your maths so that we can see exactly how much of a village idiot you are,' mocks Oliver. Luckily you can. Never one to shy from public humiliation on a mass scale, I have provided the matrix calculations for this example at www.discoverspss.com. Find a mistake, go on, you know that you can ...

For the data in our example, the matrix $\mathbf{HE}^{-1}$ calculated from the variate scores is:

$$\mathbf{HE}^{-1}_{\text{variate}} = \begin{pmatrix} 0.335 & 0 \\ 0 & 0.073 \end{pmatrix} \tag{17.25}$$

It is evident from this matrix that we have two values to consider when assessing the significance of the group differences. It may seem like a complex procedure to reduce the data in this way: however, it transpires that the values along the diagonal of the matrix for the variates (namely 0.335 and 0.073) are the *eigenvalues* of the original $\mathbf{HE}^{-1}$ matrix. Therefore, these values can be calculated directly from the original scores without first forming the eigenvectors. If you have lost all sense of rationality and want to know how these eigenvalues are calculated, then see Oliver Twisted. These eigenvalues are conceptually equivalent to the univariate *F*-statistic, and so the final step is to assess how large these values are compared to what we would expect if there were no effect in the population. There are four common ways to do this.

17.3.10 The Pillai–Bartlett trace (*V*) Ⓓ

The **Pillai–Bartlett trace** (also known as Pillai's trace) is given by:

$$V = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} \tag{17.26}$$

in which λ represents the eigenvalues for each of the discriminant variates and s represents the number of variates. Pillai's trace is the sum of the proportion of explained variance on the discriminant functions. It is similar to R^2 (the ratio of SS_M/SS_T). For our data, Pillai's trace turns out to be 0.319, which can be transformed to a value that has an approximate *F*-distribution:

$$V = \frac{0.335}{1+0.335} + \frac{0.073}{1+0.073} = 0.319 \tag{17.27}$$

17.3.11 Hotelling's T^2 Ⓓ

The **Hotelling–Lawley trace** (also known as Hotelling's T^2; Figure 17.5) is the sum of the eigenvalues for each variate:

$$T = \sum_{i=1}^{s} \lambda_i \tag{17.28}$$

so for these data its value is 0.335 + 0.073 = 0.408. This test statistic is the sum of SS_M/SS_R for each of the variates and so it compares directly to the univariate *F*-statistic.



Figure 17.5 Harold Hotelling enjoying one of my favourite activities of drinking tea

17.3.12 Wilks's lambda (Λ) Ⓓ

Wilks's lambda is the product of the *unexplained* variance on each of the variates:

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \tag{17.29}$$

(the ∏ symbol is like the summation symbol (Σ) that we have encountered already except that it means *multiply* rather than add up). Wilks's lambda represents the ratio of error variance to total variance (SS_R/SS_T) for each variate. Large eigenvalues (which in themselves represent a large experimental effect) lead to small values of Wilks's lambda: hence statistical significance is found when Wilks's lambda is small. In this example Wilks's lambda is 0.698:

$$\Lambda = \frac{1}{1+0.335} \times \frac{1}{1+0.073} = 0.698 \tag{17.30}$$

17.3.13 Roy's largest root Ⓑ

Roy's largest root makes me think of some bearded statistician with a garden spade digging up an enormous parsnip (or similar root vegetable). It isn't a parsnip but, as the name suggests, it is the eigenvalue (or 'root') for the first variate:

$$\Theta = \lambda_{Largest} \tag{17.31}$$

Roy's largest root is basically the Hotelling–Lawley trace but for the first variate only; it represents the proportion of explained variance to unexplained variance (SS_M/SS_R) for the first discriminant function.[4] For the data in this example, the value of Roy's largest root is 0.335 (the eigenvalue for the first variate). This value is conceptually the same as the univariate *F*-statistic. It should be apparent, from what we have learnt about the maximizing properties of these discriminant variates, that Roy's root represents the maximum possible between-group difference given the data collected. Therefore, this statistic should in many cases be the most powerful.

4 This statistic is sometimes characterized as $\lambda_{largest}/(1 + \lambda_{largest})$, but this is not the statistic reported by SPSS.

17.4 Practical issues when conducting MANOVA ©

17.4.1 Assumptions and how to check them ©

MANOVA has similar assumptions to all the models in this book (see Chapter 6) but extended to the multivariate case:

- **Independence**: Residuals should be statistically independent.
- **Random sampling**: Data should be randomly sampled from the population of interest and measured at the interval level.
- **Multivariate normality**: In univariate models we assume that our residuals are normally distributed. In the case of MANOVA, we assume that the residuals have multivariate normality.
- **Homogeneity of covariance matrices**: In univariate models, it is assumed that the variances in each group are roughly equal (homogeneity of variance). In MANOVA we assume that this is true for each outcome variable, but also that the correlation between any two outcome variables is the same in all groups. This assumption is examined by testing whether the population **variance–covariance matrices** of the different groups in the analysis are equal.[5]

We can correct for bias in the usual ways; however, the assumption of multivariate normality cannot be tested using IBM SPSS Statistics software and so the only practical solution is to check the assumption of univariate normality of residuals for each outcome variable in turn (see Chapter 6). This solution is practical (because it is easy to implement) and useful (because univariate normality is a necessary condition for multivariate normality), but it does not *guarantee* multivariate normality.

The effect of violating the assumption of equality of covariance matrices is unclear, except that Hotelling's T^2 is robust in the two-group situation when sample sizes are equal (Hakstian et al., 1979). The assumption can be tested using **Box's test**, which should be non-significant if the matrices are similar. Box's test is notoriously susceptible to deviations from multivariate normality and so can be non-significant not because the matrices are similar, but because the assumption of multivariate normality is not tenable. Also, as with any significance test, in large samples Box's test could be significant even when covariance matrices are relatively similar.

If sample sizes are equal then people tend to disregard Box's test, because (1) it is unstable, and (2) in this situation we can assume that Hotelling's and Pillai's statistics are robust (see Section 17.4.3). However, if group sizes are different, then robustness cannot be assumed. The more outcome variables you have measured, and the greater the differences in sample sizes, the more distorted the probability values become. Tabachnick and Fidell (2012) suggest that if the larger samples produce greater variances and covariances then the probability values will be conservative (and so significant findings can be trusted). However, if it is the smaller samples that produce the larger variances and covariances then the probability values will be liberal and so significant differences should be treated with caution (although non-significant effects can be trusted). Therefore, the variance–covariance matrices for samples should be inspected to assess whether the printed probabilities for the multivariate test statistics are likely to be conservative or liberal. In the event that you cannot trust the printed probabilities, there is little you can do except equalize the samples by randomly deleting cases in the larger groups (although with this loss of information comes a loss of power, and of course your results will have been influenced by the process of deletion, so you should do a sensitivity analysis by comparing results when you delete different sets of random cases).

5 For those of you who read about SSCP matrices, if you think about the relationship between sums of squares and variance, and cross-products and correlations, it should be clear that a variance–covariance matrix is basically a standardized form of SSCP matrix.

17.4.2 What to do when assumptions are violated ©

SPSS doesn't offer a non-parametric version of MANOVA; however, some ideas have been put forward based on ranked data. There are some techniques that can be beneficial when multivariate normality or homogeneity of covariance matrices cannot be assumed (Zwick, 1985). In addition, there are robust methods for straightforward designs with multiple outcome variables, such as the Munzel–Brunner method, which can be implemented in R (Wilcox, 2017), although it can't be done using the *WRS2* package. Although you will see a Bootstrap... button in the dialog box for MANOVA it does not bootstrap the main tests and is, ultimately, disappointing.

17.4.3 Choosing a test statistic ©



Only when there is one underlying variate will the four test statistics necessarily be the same, which raises the question of which one is 'best'. As ever, when addressing this question we really need to know which has the most power, the least error and the greatest robustness to violations of test assumptions. Research investigating power (Olson, 1974, 1976, 1979; Stevens, 1980) suggests that (1) for small and moderate sample sizes the four statistics differ little; (2) if group differences are concentrated on the first variate then Roy's statistic should have the most power (because it takes account of only that first variate) followed by Hotelling's trace, Wilks's lambda and Pillai's trace; (3) when groups differ along more than one variate, this power order is reversed (i.e., Pillai's trace is most powerful and Roy's root is least); (4) unless sample sizes are large it's probably wise to use fewer than 10 outcome variables.

In terms of robustness, all four test statistics are relatively robust to violations of multivariate normality (although Roy's root is affected by platykurtic distributions – see Olson, 1976). Roy's root is not robust when the homogeneity of covariance matrix assumption is untenable (Stevens, 1979). Bray and Maxwell (1985) conclude that when sample sizes are equal the Pillai–Bartlett trace is the most robust to violations of assumptions, but when sample sizes are unequal this statistic is affected by violations of the assumption of equal covariance matrices. As a rule, with unequal group sizes, check the homogeneity of covariance matrices; if they seem homogeneous and if the assumption of multivariate normality is tenable, then assume that Pillai's trace is accurate.

17.4.4 Follow-up analysis ©

The traditional approach is to follow a significant MANOVA with separate univariate models (ANOVA) on each of the outcome variables. You might think that this approach is daft, given that I wrote earlier that multiple univariate models would inflate the Type I error rate. I'd agree that it's daft, but some people argue that the univariate *F*-statistics are 'protected' by the initial MANOVA (Bock, 1975). The logic is that the overall multivariate test protects against inflated Type I error rates because if that initial test is non-significant (i.e., the null hypothesis is true) then the subsequent univariate *F*-statistics are ignored (because any significant *F* must be a Type I error because the null hypothesis is true). This notion of protection is dubious because a significant MANOVA usually reflects a significant difference for some, but not all, of the outcome variables. This argument of 'protection' applies only to the outcome variables for which group differences genuinely exist (see Bray & Maxwell, 1985, pp. 40–41), not for all outcome variables that you've included in the model. Despite this limitation, people tend to interpret univariate *F*s on *all* outcome variables. Therefore, if you do use univariate *F*s then you ought to apply a Bonferroni correction (Harris, 1975).

The bigger problem I have with univariate *F*s is that they don't relate to what the multivariate tests look at. Remember that the multivariate test statistic quantifies the extent to which groups can be

differentiated by a *linear combination of the outcome variables*. Subsequent univariate *F*s look at the outcome variables as independent entities, not as a linear combination. Therefore, they make no sense as a follow-up strategy. An alternative that *is* consistent with the multivariate test statistics is discriminant analysis, which finds the linear combination(s) of the outcome variables that best *separates* (or discriminates) the groups. The major advantage of this approach over multiple univariate *F*s is that it reduces the outcome variables to a set of underlying dimensions thought to reflect substantive theoretical dimensions. As such, it is true to the ethos of MANOVA.

17.5 MANOVA using SPSS Ⓑ

In the remainder of this chapter we will use the OCD data (Section 17.2.2) to illustrate the application and interpretation of MANOVA. Either load the data in the file **ocd.sav**, or enter the data manually. If you enter the data manually you need three columns: one is a coding variable for the **group** variable (I used the codes CBT = 1, BT = 2, NT = 3), and in the remaining two columns enter the scores for each outcome variable. Figure 17.6 overviews the analysis procedure: basically, explore the data as you normally would, run the MANOVA, then follow up this analysis with a discriminant function analysis. Some of you will want to look at univariate ANOVAs, so I've included that in the diagram, but personally I'd avoid them.

17.5.1 The main analysis Ⓑ

Select *Analyze* > *General Linear Model* > Multivariate... to access the dialog box in Figure 17.7, which is very similar to the dialog box for factorial designs (Chapter 14) except that the box labelled *Dependent Variables* has room for several variables. Drag the two outcome variables (i.e., **actions** and **thoughts**) to the *Dependent Variables* box (or click ▶). Drag (or click ▶) **group** to the *Fixed Factor(s)* box. There is also a box for covariates. For this analysis there are none, but you can apply the principles of ANCOVA to the multivariate case and conduct multivariate analysis of covariance (MANCOVA). The buttons in this dialog box are pretty much the same as we have seen in the past few chapters.

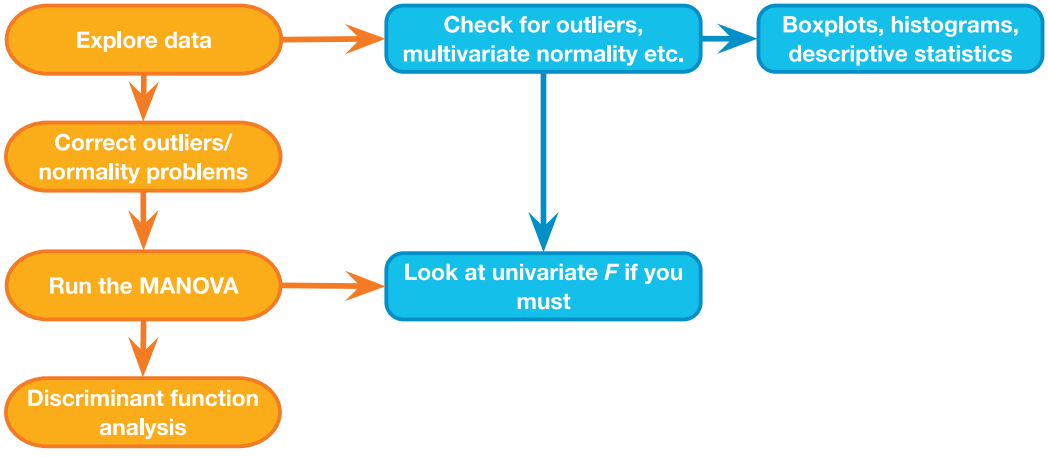


Figure 17.6 Overview of the general procedure for MANOVA

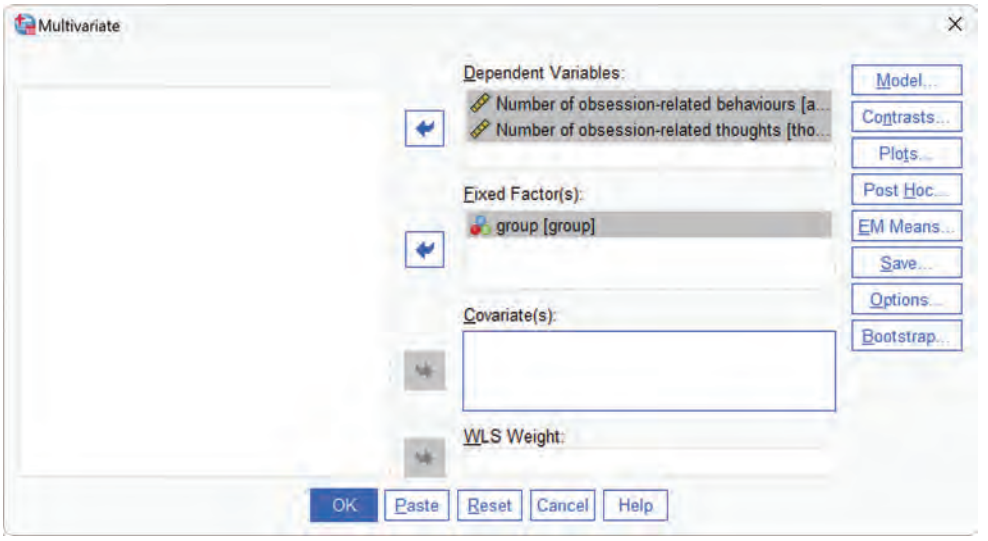


Figure 17.7 Main dialog box for MANOVA

17.5.2 Multiple comparisons in MANOVA Ⓑ

The default, but silly, way to follow up a MANOVA is to look at individual univariate *F*s for each outcome variable. You have the same options as we've seen for other univariate models (see Chapter 12). Click Contrasts... to open the dialog box (Figure 17.8) for specifying standard contrasts for the categorical predictors in the model. Refer back to Table 12.8, which describes what the standard contrasts compare. For this example it makes sense to use a *simple* contrast that compares each of the therapy groups to the no treatment control group. The no treatment control was coded as the last category (it had the largest code in the data editor), so select **Group**, change the contrast to *Simple* and select Last (see Figure 17.8).

Instead of a contrast, we could carry out *post hoc* tests that compare each group to all others by clicking Post Hoc... to access a dialog box that is the same as for factorial designs (see Figure 14.12). To choose a test see the discussion in Section 12.6. For the purposes of this example, I suggest selecting two of my usual recommendations: REGWQ and Games–Howell. Once you have selected *post hoc* tests return to the main dialog box.

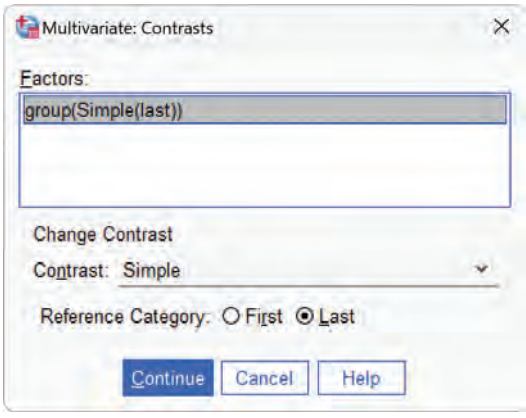


Figure 17.8 Contrasts for independent variable(s) in MANOVA

17.5.3 Additional options ©

Click Options... to access the dialog box in Figure 17.9, which is the same as the one we used for factorial designs (see Section 14.9.6). There are a few options that are worth mentioning and haven't been discussed before.

- *SSCP matrices*: If this option is selected, SPSS will produce the model, error and total SSCP matrices, which can be useful for understanding the computation of the MANOVA. If you skipped the theory section you might be happy not to select this option and live in blissful ignorance.
- *Residual SSCP matrix*: If this option is selected, SPSS produces the error SSCP matrix, the error variance–covariance matrix and the error correlation matrix. **Bartlett's test of sphericity** examines whether the variance–covariance matrix is proportional to an identity matrix (i.e., that the covariances are zero and the variances – the values along the diagonal – are roughly equal).

The remaining options are the same as for factorial designs and were described in Chapter 14).

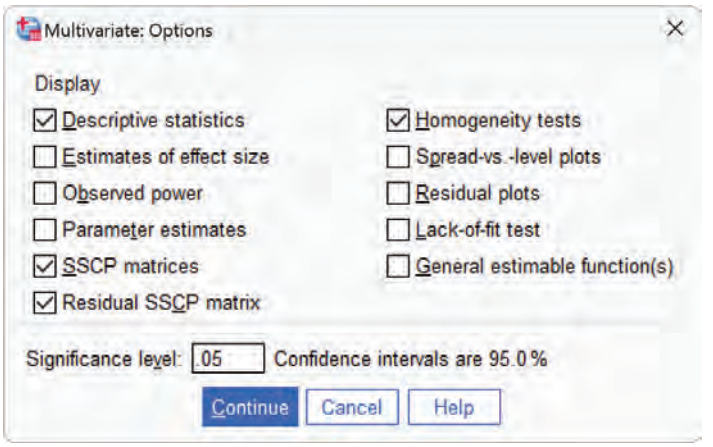


Figure 17.9 Additional options in MANOVA

17.6 Interpreting MANOVA ©

17.6.1 Preliminary analysis and testing assumptions ©

Output 17.1 contains the overall and group means and standard deviations for each outcome variable in turn. These values correspond to those in Table 17.1. It is clear from the means that participants had many more obsession-related thoughts than behaviours.

Descriptive Statistics

	group	Mean	Std. Deviation	N
Number of obsession-related behaviours	CBT	4.90	1.197	10
	BT	3.70	1.767	10
	No Treatment Control	5.00	1.054	10
	Total	4.53	1.456	30
Number of obsession-related thoughts	CBT	13.40	1.897	10
	BT	15.20	2.098	10
	No Treatment Control	15.00	2.357	10
	Total	14.53	2.209	30

Output 17.1



Oditi's Lantern
MANOVA

'I, Oditi, have seen the multivariate way. To understand the hidden truth of life we must embrace complexity as though it were our beloved partner. We probably shouldn't have sex with it though. To understand complex outcomes, such as how to change people's personalities so that they worship me ... erm, I mean, worship nature, and love, and the tulips that grow where the wild wind blows, we need MANOVA. Stare into my lantern and discover all about it.'



Output 17.2 shows Box's test of the assumption of equality of covariance matrices (see Section 17.4.1). This statistic should be *non-significant*, which it is because $p = 0.18$ is greater than 0.05; hence, the covariance matrices are roughly equal, as assumed. Bartlett's test of sphericity tests whether the assumption of sphericity has been met and is useful only in univariate repeated-measures designs, because MANOVA does not require this assumption. Basically, I included it to point out that you should ignore it. That's the kind of maverick I am.

Box's Test of Equality of Covariance Matrices[a]

Box's M	9.959
F	1.482
df1	6
df2	18168.923
Sig.	.180

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + group

Bartlett's Test of Sphericity[a]

Likelihood Ratio	.042
Approx. Chi-Square	5.511
df	2
Sig.	.064

Tests the null hypothesis that the residual covariance matrix is proportional to an identity matrix.

a. Design: Intercept + group

Output 17.2

17.6.2 MANOVA test statistics ©

Output 17.3 shows the test statistics for the intercept of the model (see, I told you, even MANOVA can be characterized as a linear model, although how this is done is beyond the scope of my brain) and for the group variable. The group effect tells us whether the therapies had different effects on the OCD clients. The four multivariate test statistics and their values correspond to those calculated in Sections

17.3.10–17.3.13. These values are transformed into an *F-statistic* (the degrees of freedom vary for each test statistic) and the *p*-value associated with this *F* is in the column labelled *Sig.* For these data, Pillai's trace ($p = 0.049$), Wilks's lambda ($p = 0.050$) and Roy's largest root ($p = 0.020$) all reach the criterion for significance (they are less than or equal to 0.05), but Hotelling's trace ($p = 0.051$) does not. This scenario is interesting, because the test statistic we choose determines whether we reject the null hypothesis that there are no between-group differences (have a guess how researcher degrees of freedom might have a negative role to play here – see Section 3.3.2). It again illustrates the pointlessness of having an all-or-nothing criterion for significance (see Section 3.2.2).

Enough ranting. Given what we know about the robustness of Pillai's trace when sample sizes are equal, we might trust the result of that test statistic, which indicates a significant difference. This example also highlights the additional power associated with Roy's root (note how this statistic is more significant than all others) when the test assumptions have been met and when the group differences are focused on one variate (which, as we will see later, they are).

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.983	745.230[b]	2.000	26.000	<.001
	Wilks' Lambda	.017	745.230[b]	2.000	26.000	<.001
	Hotelling's Trace	57.325	745.230[b]	2.000	26.000	<.001
	Roy's Largest Root	57.325	745.230[b]	2.000	26.000	<.001
group	Pillai's Trace	.318	2.557	4.000	54.000	.049
	Wilks' Lambda	.699	2.555[b]	4.000	52.000	.050
	Hotelling's Trace	.407	2.546	4.000	50.000	.051
	Roy's Largest Root	.335	4.520[c]	2.000	27.000	.020

a. Design: Intercept + group
b. Exact statistic
c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Output 17.3

We should probably conclude that the type of therapy employed had a significant effect on OCD. The nature of this effect is not clear from the multivariate test statistic: it tells us nothing about which groups differed from which, or about whether the effect of therapy was on the obsession-related thoughts, the obsession-related behaviours, or a combination of both. To determine the nature of the effect, a discriminant analysis would be helpful, but for some reason SPSS provides us with univariate tests instead. What a bilge rat.

17.6.3 Univariate test statistics Ⓑ

Output 17.4 shows a summary table of Levene's test of equality of variances for each of the outcome variables (I've edited the table so it shows only values based on the mean). These tests are the same as would be found if univariate tests had been conducted on each outcome variable. Levene's test should be non-significant for all outcome variables if the assumption of homogeneity of variance has been met (but see Section 6.7). If you buy into Levene's test being useful (ho hum), Output 17.4 shows that the assumption has been met, which strengthens the case for assuming that the multivariate test statistics are robust.

The second table shows the univariate *F*s for each outcome variable. In Sections 17.3.2 and 17.3.3, we computed various values for both actions and thoughts that are in this table: the model sum of squares (in the row labelled *Group*), the residual sum of squares (in the row labelled *Error*) and total sums of squares (in the row labelled *Corrected Total*). The *F-statistic*s for each outcome variable are *identical* to what we computed, which demonstrates that these tests are exactly what we'd get if we ran one-way ANOVAs on each outcome variable independently. As such, MANOVA offers *hypothetical* protection of inflated Type I error rates: there is no actual adjustment made to the values obtained.

Levene's Test of Equality of Error Variances[a]

Based on Mean

	Levene Statistic	df1	df2	Sig.
Number of obsession-related behaviours	1.828	2	27	.180
Number of obsession-related thoughts	.076	2	27	.927

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + group

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Number of obsession-related behaviours	10.467[a]	2	5.233	2.771	.080
	Number of obsession-related thoughts	19.467[b]	2	9.733	2.154	.136
Intercept	Number of obsession-related behaviours	616.533	1	616.533	326.400	<.001
	Number of obsession-related thoughts	6336.533	1	6336.533	1402.348	<.001
group	Number of obsession-related behaviours	10.467	2	5.233	2.771	.080
	Number of obsession-related thoughts	19.467	2	9.733	2.154	.136
Error	Number of obsession-related behaviours	51.000	27	1.889		
	Number of obsession-related thoughts	122.000	27	4.519		
Total	Number of obsession-related behaviours	678.000	30			
	Number of obsession-related thoughts	6478.000	30			
Corrected Total	Number of obsession-related behaviours	61.467	29			
	Number of obsession-related thoughts	141.467	29			

a. R Squared = .170 (Adjusted R Squared = .109)

b. R Squared = .138 (Adjusted R Squared = .074)

Output 17.4

Based on the *p*-values in Output 17.4 (in the column labelled *Sig.*) there was a non-significant difference between therapy groups in terms of both obsession-related thoughts ($p = 0.136$) and obsession-related behaviours ($p = 0.080$). Based on univariate tests then we should conclude that the type of therapy had no significant effect on the levels of OCD experienced by clients. Those of you who are still awake may have noticed something odd: the multivariate test statistics led us to conclude that therapy has had a significant impact on OCD, yet the univariate results indicate that therapy has not been successful.



Why might the univariate tests be non-significant when the multivariate tests were significant?

The reason for the anomaly is that the multivariate test takes account of the correlation between outcome variables and looks at whether groups can be distinguished by a *linear combination of the outcome variables*. This suggests that it is not thoughts or actions in themselves that distinguish the therapy groups, but some combination of them. The discriminant function analysis will provide more insight into this conclusion.

17.6.4 SSCP matrices ©

Output 17.5 and Output 17.6 will be produced if you selected the two options to display SSCP matrices (Section 17.5.3). Output 17.5 displays the model SSCP (**H**), which is labelled *Hypothesis Group* (I have shaded this matrix blue) and the error SSCP (**E**) which is labelled *Error* (shaded yellow). The values in the model and error matrices correspond to the values we calculated in

Between-Subjects SSCP Matrix

			Number of obsession-related behaviours	Number of obsession-related thoughts
Hypothesis	Intercept	Number of obsession-related behaviours	616.533	1976.533
		Number of obsession-related thoughts	1976.533	6336.533
	group	Number of obsession-related behaviours	10.467	-7.533
		Number of obsession-related thoughts	-7.533	19.467
Error		Number of obsession-related behaviours	51.000	13.000
		Number of obsession-related thoughts	13.000	122.000

Based on Type III Sum of Squares

Output 17.5

Sections 17.3.7 and 17.3.6, respectively. These matrices provide insight into the pattern of the data and especially in looking at the values of the cross-products to indicate the relationship between outcome variables. In this example, the sums of squares for the error SSCP matrix are substantially bigger than in the model (or group) SSCP matrix, whereas the absolute values of the cross-products are fairly similar. This pattern suggests that if the MANOVA is significant then it might be the relationship between outcome variables that is important rather than the individual outcome variables themselves.

Output 17.6 shows the residual SSCP matrix again, but this time it includes the variance–covariance matrix and the correlation matrix. These matrices are related. If you think back to Chapter 8, you might remember that the covariance is the average cross-product. Likewise, the variance is the average sum of squares. Hence, the variance–covariance matrix is the average form of the SSCP matrix. Similarly, we saw in Chapter 8 that the correlation was a standardized version of the covariance and so the correlation matrix represents the standardized form of the variance–covariance matrix. As with the SSCP matrix, these other matrices are useful for assessing the extent of the error in the model. The variance–covariance matrix is especially useful because Bartlett's test of sphericity is based on it. Bartlett's test examines whether this matrix is proportional to an identity matrix (see earlier). Therefore, Bartlett's test tests whether the diagonal elements of the variance–covariance matrix are equal (i.e., group variances are the same), and whether the off-diagonal elements are approximately zero (i.e., the outcome variables are not correlated). For these data, the variances are quite different (1.89 compared to 4.52) and the covariances slightly different from zero (0.48), so Bartlett's test has come out as nearly significant (see Output 17.2). Although this discussion is irrelevant to the multivariate tests, I hope that by expanding upon them here you can relate these ideas back to the issues of sphericity raised in Chapter 15, and see more clearly how this assumption is tested.

Residual SSCP Matrix

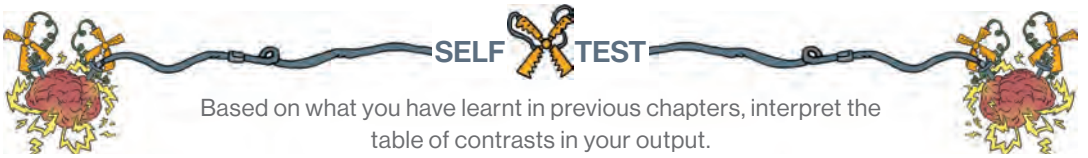
		Number of obsession-related behaviours	Number of obsession-related thoughts
Sum-of-Squares and Cross-Products	Number of obsession-related behaviours	51.000	13.000
	Number of obsession-related thoughts	13.000	122.000
Covariance	Number of obsession-related behaviours	1.889	.481
	Number of obsession-related thoughts	.481	4.519
Correlation	Number of obsession-related behaviours	1.000	.165
	Number of obsession-related thoughts	.165	1.000

Based on Type III Sum of Squares

Output 17.6

17.6.5 Contrasts ©

The univariate ANOVAs were non-significant so we should not interpret the contrasts that we requested. However, just for practice, try the self-test.



Based on what you have learnt in previous chapters, interpret the table of contrasts in your output.



Cramming Sam's Tips MANOVA

- MANOVA is used to test the difference between groups across several outcome variables/outcomes simultaneously.
- Box's test looks at the assumption of equal covariance matrices. This test can be ignored when sample sizes are equal because when they are, some MANOVA test statistics are robust to violations of this assumption. If group sizes differ this test should be inspected. If the value of *Sig.* is less than 0.001 then the results of the analysis should not be trusted (see Section 17.6.1).
- The table labelled *Multivariate Tests* gives us four test statistics (*Pillai's trace*, *Wilks's lambda*, *Hotelling's trace* and *Roy's largest root*). I recommend using Pillai's trace. If the value of *Sig.* for this statistic is less than or equal to 0.05 then the groups differ significantly with respect to a linear combination of the outcome variables.
- Univariate *F*-statistics can be used to follow up the MANOVA (a different *F*-statistic for each outcome variable). The results of these are listed in the table entitled *Tests of Between-Subjects Effects*. These *F*-statistics can in turn be followed up using contrasts. Personally I recommend *discriminant function analysis* over this approach.



17.7 Reporting results from MANOVA Ⓑ

Reporting a MANOVA is much like reporting a univariate comparison of group means. As you can see in Output 17.3, the multivariate tests are converted into approximate *F*s, and people often just report these *F*s in the usual way. Personally, I think the multivariate test statistic should be quoted as well. There are four different multivariate tests reported in Output 17.3; I'll report each one in turn (note that the degrees of freedom and value of *F* change), but in reality you would just report one of the four:

- ✓ Using Pillai's trace, there was a significant effect of therapy on the number of obsessive thoughts and behaviours, $V = 0.32$, $F(4, 54) = 2.56$, $p = 0.049$.
- ✓ Using Wilks's statistic, there was a significant effect of therapy on the number of obsessive thoughts and behaviours, $\Lambda = 0.70$, $F(4, 52) = 2.56$, $p = 0.05$.
- ✓ Using Hotelling's trace statistic, there was not a significant effect of therapy on the number of obsessive thoughts and behaviours, $T = 0.41$, $F(4, 50) = 2.55$, $p = 0.051$.
- ✓ Using Roy's largest root, there was a significant effect of therapy on the number of obsessive thoughts and behaviours, $\Theta = 0.35$, $F(2, 27) = 4.52$, $p = 0.02$.

We can also report the follow-up univariate *F*s in the usual way (see Output 17.3 and Output 17.4):

- ✓ Using Pillai's trace, there was a significant effect of therapy on the number of obsessive thoughts and behaviours, $V = 0.32$, $F(4, 54) = 2.56$, $p = 0.049$. However, separate univariate tests on the outcome variables revealed non-significant treatment effects on obsessive thoughts, $F(2, 27) = 2.15$, $p = 0.136$, and behaviours, $F(2, 27) = 2.77$, $p = 0.08$.



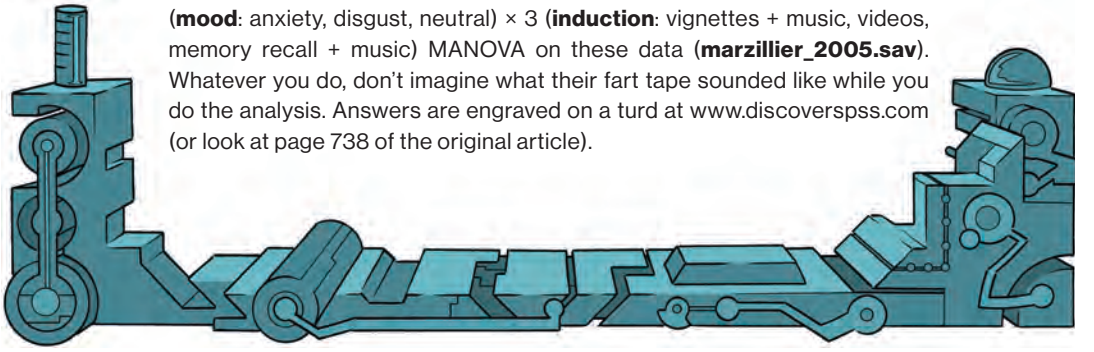
Labcoat Leni's Real Research 17.1
A lot of hot air! Ⓓ

Marzillier, S. L., & Davey, G. C. L. (2005). *Cognition and Emotion*, *19*, 729–750.

Have you ever wondered what researchers do in their spare time? Well, some of them spend it tracking down the sounds of people burping and farting! Anxious people are, typically, easily disgusted. Throughout this book I have talked about how you cannot infer causality from relationships between variables. This has been a bit of a conundrum for anxiety researchers: does anxiety cause feelings of disgust or does a low threshold for being disgusted cause anxiety? Two colleagues at Sussex addressed this by inducing feelings of anxiety, disgust, or a neutral mood. They looked at the effect of these induced moods on feelings of anxiety, sadness, happiness, anger, disgust and contempt. To induce these moods, they used three different types of

manipulation: vignettes (e.g., 'You're swimming in a dark lake and something brushes your leg' for anxiety and 'You go into a public toilet and find it has not been flushed. The bowl of the toilet is full of diarrhoea' for disgust), music (e.g., some scary music for anxiety, and a tape of burps, farts and vomiting for disgust), videos (e.g., a clip from *Silence of the Lambs* for anxiety and a scene from *Pink Flamingos* in which Divine eats dog faeces for disgust) and memory (remembering events from the past that had made the person anxious, disgusted or neutral).

Different people underwent anxious, disgust and neutral mood inductions. Within these groups, the induction was done using either vignettes and music, videos, or memory recall and music for different people. The outcome variables were the change (from before to after the induction) in six moods: anxiety, sadness, happiness, anger, disgust and contempt. Draw an error bar graph of the changes in moods in the different conditions, then conduct a 3 (**mood**: anxiety, disgust, neutral) × 3 (**induction**: vignettes + music, videos, memory recall + music) MANOVA on these data (**marzillier_2005.sav**). Whatever you do, don't imagine what their fart tape sounded like while you do the analysis. Answers are engraved on a turd at www.discoverspss.com (or look at page 738 of the original article).



17.8 Following up MANOVA with discriminant analysis ©

I mentioned earlier on that a significant MANOVA could be followed up using **discriminant analysis** (sometimes called discriminant function analysis). In my opinion this method is the best way to follow up a significant MANOVA because a MANOVA looks at whether groups differ along *a linear combination* of outcome variables, and discriminant analysis (unlike univariate *F*s) breaks down the linear combination in more detail. In discriminant analysis we look to see how we can best separate (or discriminate) a set of groups using several predictors (it is a little like logistic regression but where there are several groups rather than two).[6] In MANOVA we predicted a set of outcome measures from a grouping variable, whereas in discriminant function analysis we do the opposite by predicting a grouping variable from a set of outcome measures. The core underlying principles of these tests are the same: remember from the theory of MANOVA that it works by identifying linear variates that best differentiate the groups and these 'linear variates' are the 'functions' in discriminant function analysis.

Select *Analyze* > *Classify* > Discriminant... to access the dialog box in Figure 17.10. Drag **group** to the box labelled *Grouping Variable* (or click), then click Define Range... to activate a dialog box in which you specify the value of the highest and lowest coding values (1 and 3 in this case). Once you have specified the codes used for the grouping variable, drag **actions** and **thoughts** to the box labelled

6 I could just as easily describe discriminant analysis rather than logistic regression in Chapter 20 because they are different ways of achieving the same end result. However, logistic regression has far fewer restrictive assumptions and is generally more robust, which is why I have limited the coverage of discriminant analysis to this chapter.

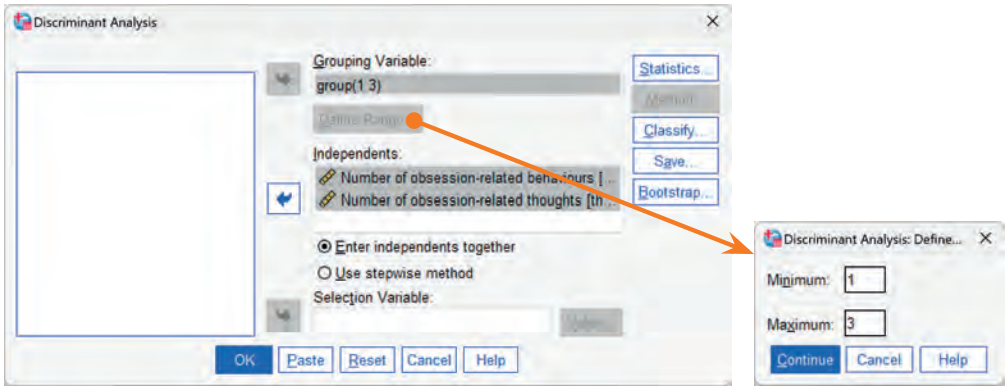


Figure 17.10 Main dialog box for discriminant analysis

Independents (or click ▸). There are two options that determine how predictors are entered into the model. Because in MANOVA the outcome variables are analysed simultaneously, we want to select the default option of ◉ Enter independents together.

Click Statistics... to activate the dialog box in Figure 17.11. This dialog box allows us to request group means, univariate ANOVAs (*F*s) and Box's test of equality of covariance matrices, all of which have already been provided in the MANOVA output (so don't ask for them again). Furthermore, we can ask for the within-group correlation and covariance matrices, which are the same as the residual correlation and covariance matrices seen in Output 17.6. There is also an option to display a *Separate-groups covariance* matrix, which can be useful for gaining insight into the relationships between outcome variables for each group (this matrix is something that the MANOVA procedure doesn't display, and I recommend selecting it). Finally, we can ask for a total covariance matrix, which displays covariances and variances of the outcome variables overall. Another useful option is to select *Unstandardized* function coefficients to produce the unstandardized *b*s for each variate (see Eq. 17.22). Click Continue to return to the main dialog box.

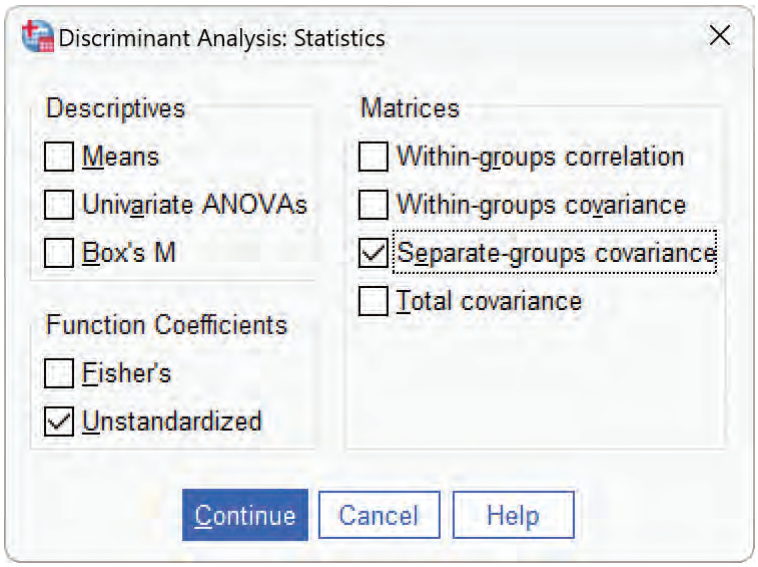


Figure 17.11 Statistics options for discriminant analysis

Click Classify... to access the dialog box in Figure 17.12, in which you select how prior probabilities are determined. If your group sizes are equal then leave the default setting alone; however, if you have an unbalanced design then it is beneficial to base prior probabilities on the observed group sizes. The default option for basing the analysis on the within-group covariance matrix is fine (because this is the matrix upon which the MANOVA is based). You should also request a combined-groups plot, which will plot the variate scores for each participant grouped according to the therapy they were given. The separate-groups plots show the same thing but using different plots for each of the groups; when the number of groups is small it is better to select a combined plot because it is easier to interpret. The remaining options are of little interest when using discriminant analysis to follow up MANOVA except for the *Summary table*, which provides an overall gauge of how well the discriminant variates classify the actual participants. Click Continue to return to the main dialog box.

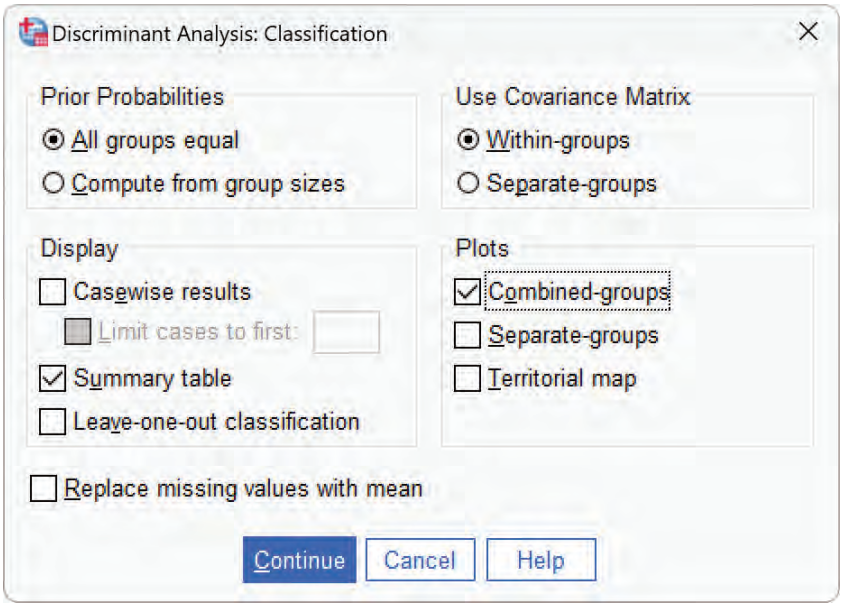


Figure 17.12 Discriminant analysis classification options

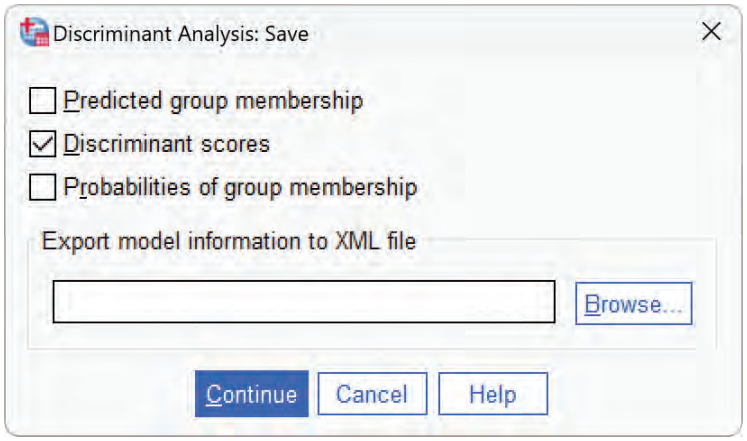


Figure 17.13 The save new variables dialog box in discriminant analysis

Click Save... to access the dialog box in Figure 17.13. There are three options available, two of which relate to the predicted group memberships and probabilities of group memberships from the model. The final option is to provide the **discriminant scores**. These are the scores for each person, on each variate, obtained from Eq. 17.22. These scores can be useful because the variates that the analysis identifies may represent underlying social or psychological constructs. If these constructs are identifiable, then it is useful for interpretation to know what a participant scores on each dimension.

17.9 Interpreting discriminant analysis Ⓓ

Output 17.7 shows the covariance matrices for separate groups (selected in Figure 17.11). These matrices are made up of the variances of each outcome variable for each group (in fact these values are shown in Table 17.1). The covariances are obtained by taking the cross-products between the outcome variables for each group (shown in Table 17.4 as 0.40, 22.6 and –10) and dividing each by 9, the degrees of freedom, $N - 1$ (where N is the number of observations). The values in this table give us some idea of how the relationship between outcome variables changes from group to group. For example, in the CBT group behaviours and thoughts have virtually no relationship because the covariance is almost zero. In the BT group thoughts and actions are positively related, so as the number of behaviours decreases, so does the number of thoughts. In the NT condition there is a negative relationship, so if the number of thoughts increases then the number of behaviours decreases. It is important to note that these matrices don't tell us about the substantive importance of the relationships because they are unstandardized (see Chapter 18), but they give a basic indication.

Covariance Matrices

group		Number of obsession-related behaviours	Number of obsession-related thoughts
CBT	Number of obsession-related behaviours	1.433	.044
	Number of obsession-related thoughts	.044	3.600
BT	Number of obsession-related behaviours	3.122	2.511
	Number of obsession-related thoughts	2.511	4.400
No Treatment Control	Number of obsession-related behaviours	1.111	-1.111
	Number of obsession-related thoughts	-1.111	5.556

Output 17.7

Output 17.8 shows the initial statistics from the discriminant analysis. The first table contains the eigenvalues for each variate – note that the values correspond to the values of the diagonal elements of the matrix $\mathbf{HE}^{-1}_{\text{variate}}$ in Eq. 17.25. These eigenvalues are converted into percentage of variance accounted for, and the first variate accounts for 82.2% of variance, whereas the second variate accounts for only 17.8%. This table also shows the canonical correlation, which we can square to use as an effect size (just like R^2 in a standard linear model).

The second table in Output 17.8 shows the significance tests of both variates (*1 through 2* in the table), and the significance after the first variate has been removed (*2* in the table). So, effectively we test the overall model, and then peel away variates one at a time to see whether what's left is significant. With two variates we get only two steps: the whole model, and then the model after the first variate is removed (leaving only the second variate). When both variates are tested in combination, Wilks's lambda has the same value (0.699), degrees of freedom (4) and significance value (0.05) as in the MANOVA (see Output 17.3). The important point to note from this table is that the two variates significantly discriminate the groups in combination ($p = 0.05$), but the second variate alone is non-significant ($p = 0.173$). Therefore, the group differences shown by the MANOVA can be explained in terms of *two* underlying dimensions in combination.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.335[a]	82.2	82.2	.501
2	.073[a]	17.8	100.0	.260

a. First 2 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.699	9.508	4	.050
2	.932	1.856	1	.173

Output 17.8

The tables in Output 17.9 are the most important for interpretation. The first shows the standardized discriminant function coefficients for the two variates. These values are standardized versions of the values in the eigenvectors calculated in Section 17.3.9. Recall that if the variates can be expressed in terms of a linear regression equation (see Eq. 17.20), the standardized discriminant function coefficients are equivalent to the standardized *b*-values in a linear model. The structure matrix shows the same information, but in a slightly different form. The values in this matrix are the canonical variate correlation coefficients. These values indicate the substantive nature of the variates. Bargman (1970) argues that when some outcome variables have high canonical variate correlations while others have low ones, the ones with high correlations contribute most to group separation. As such they represent the relative contribution of each outcome variable to group separation (see Bray & Maxwell, 1985, pp. 42–45). Hence, the coefficients in these tables tell us the relative contribution of each variable to the variates.

Standardized Canonical Discriminant Function Coefficients

	Function 1	Function 2
Number of obsession-related behaviours	.829	.584
Number of obsession-related thoughts	−.713	.721

Structure Matrix

	Function 1	Function 2
Number of obsession-related behaviours	.711*	.703
Number of obsession-related thoughts	−.576	.817*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

Output 17.9

If we look at variate 1 first, thoughts and behaviours have the opposite effect (behaviours have a positive relationship with this variate whereas thoughts have a negative relationship). Given that these values (in both tables) can vary between −1 and 1, we can also see that both relationships are strong (although behaviours have slightly larger contribution to the first variate). The first variate, then, could be seen as one that differentiates thoughts and behaviours (it affects thoughts and behaviours in opposite ways). Both thoughts and behaviours have a strong positive relationship with the second variate. This tells us that this variate represents something that affects thoughts and behaviours in similar ways. Remembering that ultimately these variates are used to differentiate groups, we could say that the first variate differentiates groups by some factor that affects thoughts and behaviours differently, whereas the second variate differentiates groups on some dimension that affects thoughts and behaviours in the same way.

Canonical Discriminant Function Coefficients

	Function 1	Function 2
Number of obsession-related behaviours	.603	.425
Number of obsession-related thoughts	-.335	.339
(Constant)	2.139	-6.857

Unstandardized coefficients

Functions at Group Centroids

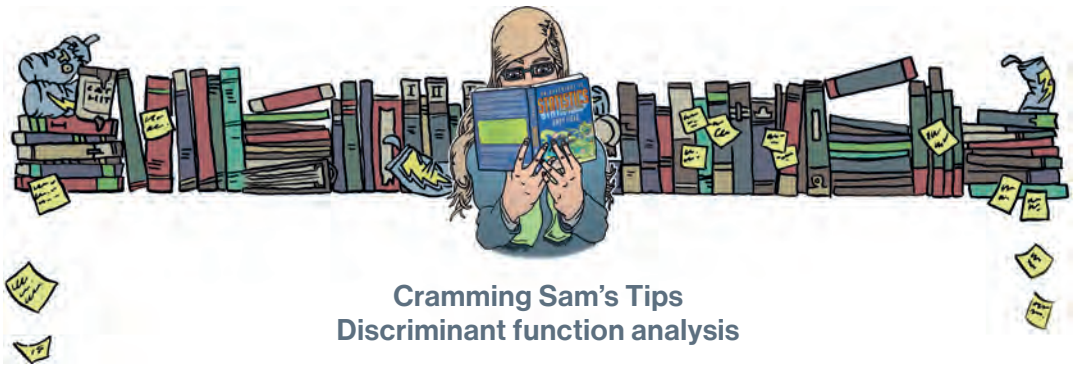
group	Function 1	Function 2
CBT	.601	-.229
BT	-.726	-.128
No Treatment Control	.125	.357

Unstandardized canonical discriminant functions evaluated at group means

Output 17.10

Output 17.10 tells us first the canonical discriminant function coefficients, which are the unstandardized versions of the standardized coefficients described above. These values are the values of b in Eq. 17.20, and you'll notice that these values correspond to the values in the eigenvectors derived in Section 17.3.9 and used in Eq. 17.22. The values are less useful than the standardized versions, but demonstrate from where the standardized versions come.

The centroids are the mean variate scores for each group. For interpretation look at the sign of the centroid (positive or negative). We can also use a combined-groups plot (selected using the dialog box in Figure 17.12. This plot shows the variate scores for each person, grouped according to the treatment



Cramming Sam's Tips
Discriminant function analysis

- Discriminant function analysis can be used after MANOVA to see how the outcome variables discriminate the groups.
- Discriminant function analysis identifies variates (combinations of the outcome variables). To find out how many variates are significant look at the tables labelled *Wilks's Lambda*: if the value of *Sig.* is less than or equal to 0.05 then the variate is significantly discriminating the groups.
- Once the significant variates have been identified, use the table labelled *Standardized Canonical Discriminant Function Coefficients* to find out how the outcome variables contribute to the variates. High scores indicate that an outcome variable is important for a variate, and variables with positive and negative coefficients are contributing to the variate in opposite ways.
- Finally, to find out which groups are discriminated by a variate look at the table labelled *Functions at Group Centroids:* for a given variate, groups with values opposite in sign are being discriminated by that variate.



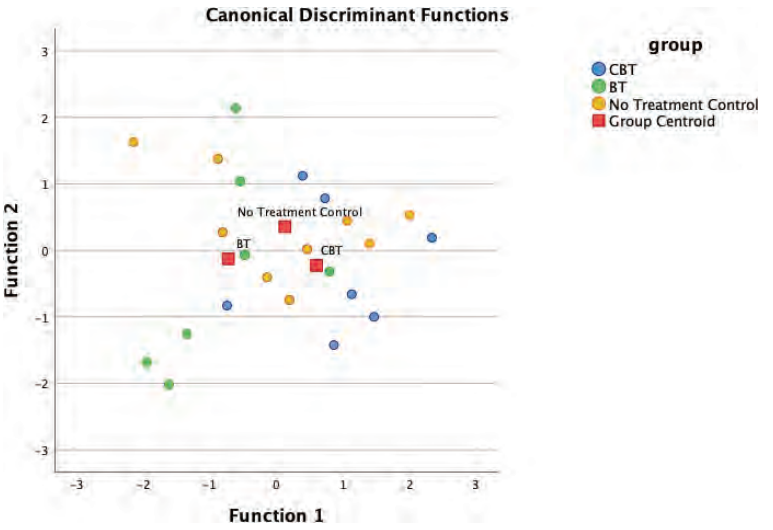


Figure 17.14 Combined-groups plot

to which that person was assigned. In addition, the group centroids from Output 17.10 are shown as red squares. The plot (Figure 17.14) and the tabulated values of the centroids (Output 17.10) tell us that variate 1 discriminates the BT group (look at the big squares labelled with the group initials) from the CBT (look at the horizontal distance between these centroids). The second variate differentiates the no treatment group from the two interventions (look at the vertical distances), but this difference is not as dramatic as for the first variate. Remember that the variates significantly discriminate the groups in combination (i.e., when both are considered).

17.10 Reporting results from discriminant analysis Ⓑ

The guiding principle in presenting data is to give readers enough information to be able to judge for themselves what the data mean. Personally, I would suggest reporting the percentage of variance explained (which gives the reader the same information as the eigenvalue but in a more palatable form) and the squared canonical correlation for each variate (this is the appropriate effect size measure for discriminant analysis). I would also report the chi-square significance tests of the variates. These values can be found in Output 17.8 (but remember to square the canonical correlation). It is probably also useful to quote the values in the structure matrix in Output 17.9 (which will tell the reader about how the outcome variables relate to the underlying variates). Finally, although I won't reproduce it below, you could consider including a (well-edited) copy of the combined-groups centroid plot (Figure 17.14), which will help readers to determine how the variates contribute to distinguishing your groups. We could, therefore, write something like this:

✓ The MANOVA was followed up with discriminant analysis, which revealed two discriminant functions. The first explained 82.2% of the variance, canonical $R^2 = 0.25$, whereas the second explained only 17.8%, canonical $R^2 = 0.07$. In combination these discriminant functions significantly differentiated the treatment groups, $\Lambda = 0.70$, $\chi^2(4) = 9.51$, $p = 0.05$, but removing the first function indicated that the second function did not significantly differentiate the treatment groups, $\Lambda = 0.93$, $\chi^2(1) = 1.86$, $p = 0.173$. The correlations between outcomes and the discriminant functions revealed that obsessive behaviours loaded highly onto both functions ($r = 0.71$ for the first function and $r = 0.70$ for the second); obsessive thoughts loaded more

highly on the second function ($r = 0.82$) than the first function ($r = -0.58$). The discriminant function plot showed that the first function discriminated the BT group from the CBT group, and the second function differentiated the no treatment group from the two interventions.

17.11 The final interpretation Ⓓ

How can we bring all of this information together to answer our research question: can therapy improve OCD and, if so, which therapy is best? The MANOVA tells us that therapy can have a significant effect on OCD symptoms, but the non-significant univariate *F*s suggested that this improvement is not simply in terms of either thoughts or behaviours. The discriminant analysis suggests that the group separation can be best explained in terms of one underlying dimension. The dimension is likely to be OCD itself (which we can realistically assume is made up of both thoughts and behaviours). Therapy doesn't necessarily change behaviours or thoughts *per se*, it influences the underlying dimension of OCD. So, the answer to the first question seems to be: yes, therapy can influence OCD in general.

The next question is more complex: which therapy is best? Figure 17.15 shows that for actions, BT reduces the number of obsessive behaviours, whereas CBT and NT do not. For thoughts, CBT reduces the number of obsessive thoughts more than BT and NT do (check the pattern of the blue squares). Figure 17.16 shows the relationships between thoughts and actions across the groups. In the BT group there is a positive relationship between thoughts and actions, so the more obsessive thoughts a person has, the more obsessive behaviours they carry out. In the CBT group there is no relationship at all (thoughts and actions vary quite independently). In the no treatment group there is a negative (and incidentally non-significant) relationship between thoughts and actions.

What we have discovered from the discriminant analysis is that BT and CBT can be differentiated from the control group based on variate 2, a variate that has a similar effect on both thoughts and behaviours. We could say then that BT and CBT are both better than a no treatment group at changing obsessive thoughts and behaviours. We also discovered that BT and CBT could be distinguished by variate 1, a variate that had the opposite effects on thoughts and behaviours. Combining this information with that in Figure 17.15, we could conclude that BT is better at changing behaviours and CBT is better at changing thoughts. So, the NT group can be distinguished from the CBT and BT groups using a variable that affects both thoughts and behaviours. Also, the CBT and BT groups can be distinguished by a variate that has opposite effects on thoughts and behaviours. So, some therapy is better than none, but the choice of CBT or BT depends on whether you think it's more important to target thoughts (CBT) or behaviours (BT).

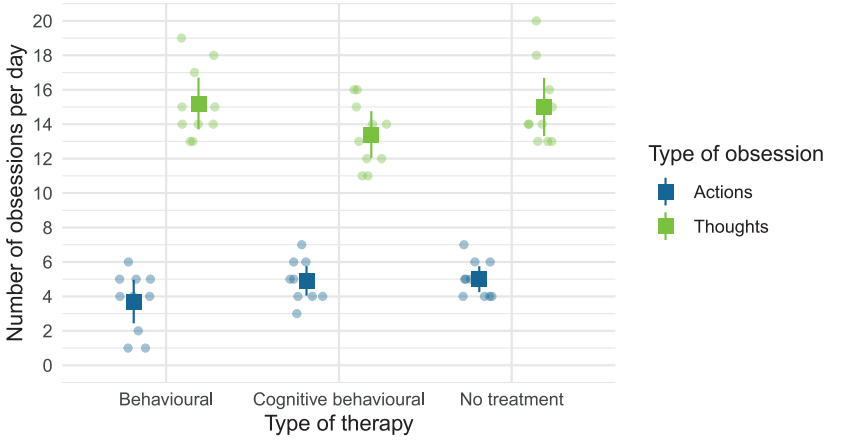


Figure 17.15 Plot showing the means and 95% confidence intervals between the outcome variables in each therapy group

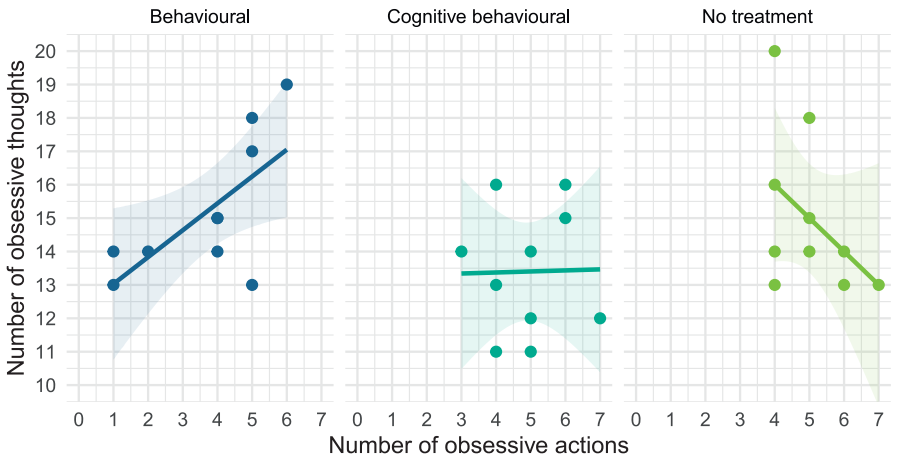


Figure 17.16 Plots showing the relationships between the outcome variables across each therapy group

17.12 Jane and Brian's story Ⓐ

Brian was confused. It had been 5 days since he'd opened up to Jane and he'd heard nothing since. He rarely told people about his mother, because it made them feel awkward. They acted differently. It was hard to say exactly how they acted, but they were sort of safe with him, as though scared of saying or doing something that might trigger him. It was easier not to talk about her. It was amazing how many friends you could have who never asked about your parents. After the recent weeks of talking to Jane about her family it seemed wrong not to tell her when she asked. It must've freaked her, though, because her phone was dead, and this was the longest, since Jane's mother got ill, that she hadn't contacted him.

Jane was exhausted. She emerged from the lab into the dark. She'd hit a wall and needed to sleep. Her battery had long died on her phone and she had no idea what time it was. She felt unsteady on her feet. A girl walking past asked if she was OK. She said Jane looked terrible. Jane asked her for the time. It was 9.30 p.m. As an afterthought she asked what day it was. 'Friday,' the girl replied, giving her the kind of look you'd give someone who didn't know what day it was. 'Holy shit!' Jane thought. She'd been working for 5 days. How was that possible? She panicked, realizing that she had vanished from Brian for so long without explanation. Why hadn't it occurred to her to call him? She guessed because there were things she didn't want him to know yet. Jane caught the train into the city. She stood in front of Brian's apartment a little scared to press the buzzer. Brian recoiled when he saw her. 'You look exhausted,' he said. She lay on his couch. 'Talk to me,' she demanded, and she drifted into sleep to the sound of his calming voice.

17.13 What next? Ⓑ

At the beginning of this chapter we discovered that pets can be therapeutic. I left the whereabouts of Fuzzy a mystery. After frantically searching the house, I went back to the room he had vanished from to check again whether there was a hole that he could have wriggled through. As I scuttled around on my hands and knees tapping the walls, a little ginger (and sooty) face popped out from the fireplace with a look as if to say 'have you lost something?' (see Figure 17.18). Yep, freaked out by the whole moving experience, he had done the only sensible thing and hidden up the chimney! Cats, you gotta love 'em!



Figure 17.17 What Brian learnt from this chapter

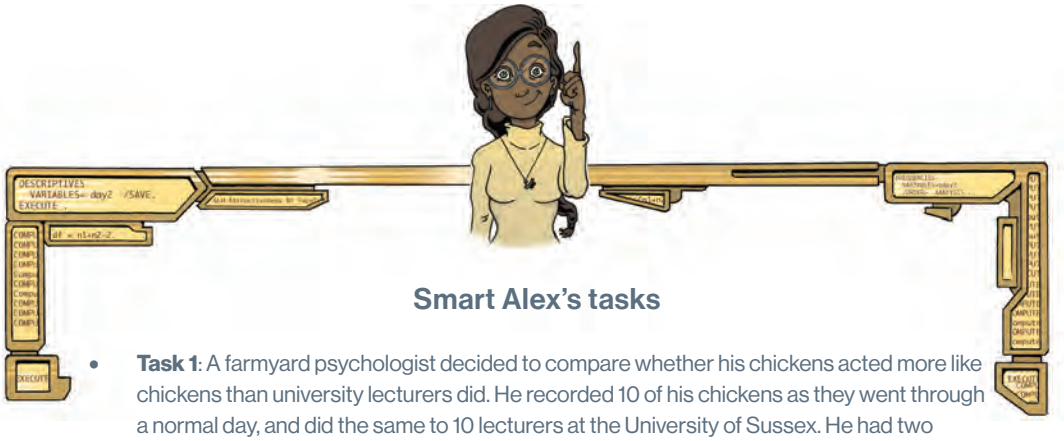
At some point everything ends (even my chapters), and for a few editions of this book I have been aware that the day would come when I had to edit my story about Fuzzy to be in the past tense. That day came. Fuzzy took on a whole other existence as 'the cat in the stat' (book). I started writing this book when he was a kitten, and he sat by my side, on my lap, or on the desk for 20 years. He saw me through the end of my PhD and four editions of this book (and a few others as well). I used to joke that he was immortal, but that was only to make myself feel better about the fact that he wasn't. He was just a soppy old ginger cat like lots of other soppy old ginger cats, except he was my soppy old ginger cat. I loved him because while people came and went from my life, he was the constant: the feline friend I could rely on for a cuddle. Even years on, it's weird not having him here while I type. So now, despite being widely known as 'a cat person' I don't have a cat because Fuzzy is irreplaceable. Ironically, I *do* have a crazy cocker spaniel (see Chapter 12) who *is* on the floor next to me. I like to think that when my back is turned, a ginger cloud emerges from the ether and gives him a hard time about stats. But not too hard a time.



Figure 17.18 Fuzzy hiding up a fireplace

17.14 Key terms that I've discovered

Bartlett's test of sphericity
Box's test
Discriminant analysis
Discriminant function variates
Discriminant scores
Error SSCP (**E**)
$\mathbf{HE}^{-1}$
Homogeneity of covariance matrices
Hotelling–Lawley trace (T^2)
Hypothesis SSCP (**H**)
Identity matrix
Matrix
Multivariate
Multivariate analysis of variance (or MANOVA)
Multivariate normality
Pillai–Bartlett trace (V)
Roy's largest root
Square matrix
Sum of squares and cross-products matrix (SSCP)
Total SSCP (**T**)
Univariate
Variance–covariance matrix
Wilks's lambda (Λ)



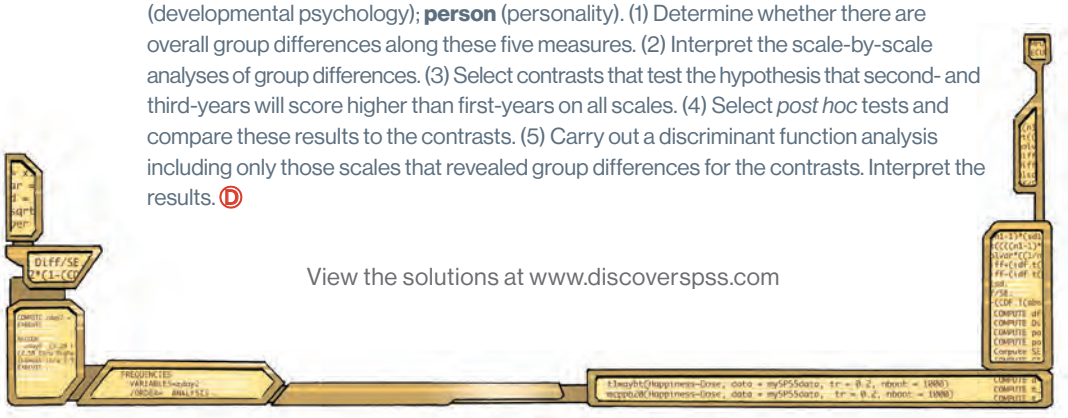
Smart Alex's tasks

- **Task 1**: A farmyard psychologist decided to compare whether his chickens acted more like chickens than university lecturers did. He recorded 10 of his chickens as they went through a normal day, and did the same to 10 lecturers at the University of Sussex. He had two outcome measures: the number of chicken noises they made and the quality of chicken noises (scored out of 10 by an independent farmyard noise expert). Use MANOVA and discriminant function analysis to find out whether these variables could be used to distinguish chickens from university lecturers (**chicken.sav**). ©
- **Task 2**: A news story claimed that children who lie would become successful citizens. I was intrigued because although the article cited a lot of well-conducted work by Dr. Khang Lee that shows that children lie, I couldn't find anything in that research that supported the journalist's claim that children who lie become successful citizens. Imagine a Huxleyesque parallel universe in which the government was daft enough to believe the contents of this

newspaper story and decided to implement a systematic programme of infant conditioning. Some infants were trained not to lie, others were bought up as normal and a final group was trained in the art of lying. Thirty years later, they collected data on how successful these children were as adults. They measured their **salary**, and two indices out of 10 (10 = as successful as it could possibly be, 0 = better luck in your next life) of how successful their **family** and **work** life was. Use MANOVA and discriminant function analysis to find out whether lying really does make you a better citizen (**lying.sav**). ©

- **Task 3**: I was interested in whether students' knowledge of different aspects of psychology improved throughout their degree (**psychology.sav**). I took a sample of first-years, second-years and third-years and gave them five tests (scored out of 15) representing different aspects of psychology: **exper** (experimental psychology such as cognitive and neuropsychology); **stats** (statistics); **social** (social psychology); **develop** (developmental psychology); **person** (personality). (1) Determine whether there are overall group differences along these five measures. (2) Interpret the scale-by-scale analyses of group differences. (3) Select contrasts that test the hypothesis that second- and third-years will score higher than first-years on all scales. (4) Select *post hoc* tests and compare these results to the contrasts. (5) Carry out a discriminant function analysis including only those scales that revealed group differences for the contrasts. Interpret the results. Ⓓ

View the solutions at www.discoverspss.com





EXPLORATORY FACTOR ANALYSIS

18.1 What will this chapter tell me? 824
18.2 When to use factor analysis 825
18.3 Factors and components 826
18.4 Discovering factors 832
18.5 An anxious example 841
18.6 Factor analysis using SPSS 847
18.7 Interpreting factor analysis 853
18.8 How to report factor analysis 866
18.9 Reliability analysis 868
18.10 Reliability analysis using SPSS 873
18.11 Interpreting reliability analysis 875
18.12 How to report reliability analysis 878
18.13 Jane and Brian's story 879
18.14 What next? 880
18.15 Key terms that I've discovered 880
Smart Alex's tasks 881

Letters A to D indicate increasing difficulty

18.1 What will this chapter tell me? Ⓐ

Having failed to become a rock star, I went to university and eventually ended up doing a PhD (in psychology) at the University of Sussex. Like many postgraduates, I taught to survive. I was allocated to second-year undergraduate statistics. I was very shy at the time, and I didn't have a clue about statistics, so standing in front of a room full of strangers and talking to them about ANOVA was about as appealing as dislocating my knees and running a marathon. I obsessively prepared for my first session so that it would go well; I created handouts, I invented examples, I rehearsed what I would say. I went in terrified but knowing that if preparation was any predictor of success then I would be OK. About half way through one of the students rose majestically from her chair. An aura of bright white light surrounded her and she appeared to me as though walking through dry ice. I guessed that she had been chosen by her peers to impart a message of gratitude for the hours of preparation I had done and the skill with which I was unclouding their brains of statistical mysteries. She stopped inches away from me. She looked into my eyes, and mine raced around the floor looking for the reassurance of my shoelaces. 'No one in this room has a rabbit[1] clue what you're going on about,' she spat before storming out. Scales have not been invented yet to measure how much I wished I'd run the dislocated-knees marathon that morning. To this day I have intrusive thoughts about students in my lectures walking zombie-like towards the front of the lecture theatre chanting 'No one knows what you're going on about' before devouring my brain in a rabid feeding frenzy.

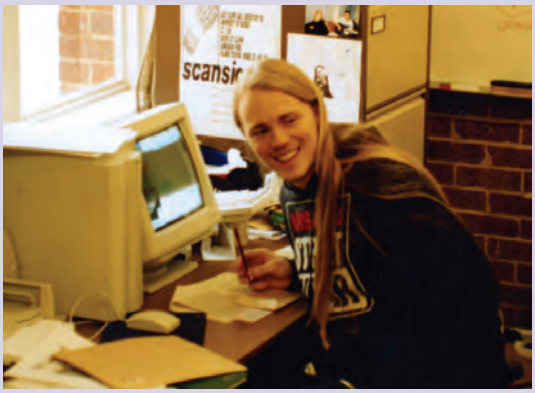


Figure 18.1 In my office during my PhD, probably preparing some teaching – I had quite long hair back then because it hadn't started falling out at that point

The aftermath of this humiliating experience was that I threw myself into trying to be the best teacher in the universe. I wrote detailed handouts and started using wacky examples. Based on these I was signed up by a publisher to write a book. This book. At the age of 23 I didn't realize that this was a terrible career move (really, textbooks take a long time to write and they are not at all valued compared to research articles). I also didn't realize the emotional toll it would take on me.

1 She didn't say 'rabbit', but she did say a word that describes what rabbits do a lot; it begins with an 'f' and my publishers think that it will offend you.

I soon discovered that writing a statistics book was like doing a factor analysis: in factor analysis we take a lot of information (variables) and a computer effortlessly reduces this mass of confusion into a simple message (fewer variables). A computer does this in a few seconds. Similarly, my younger self took a mass of information about statistics that I didn't understand and filtered it down into a simple message that I *could* understand: I became a living, breathing factor analysis … except that, unlike a computer, it took me 2 years and some considerable effort.

18.2 When to use factor analysis Ⓑ

In science we often need to measure something that cannot be accessed directly (a so-called **latent variable**). For example, management researchers might be interested in measuring 'burnout', which is when someone who has been working very hard on a project (a book, for example) for a prolonged period of time suddenly finds themself devoid of motivation, inspiration, and wants to curl up in a ball under his desk and cuddle a spaniel. You can't measure burnout directly: it has many facets. However, you can measure different aspects of burnout: you could get some idea of motivation, stress levels, whether the person has any new ideas and so on. Having done this, it would be helpful to know whether these facets reflect a single construct. Put another way, are these observable measures driven by the same underlying variable?

This chapter explores **factor analysis** and **principal component analysis (PCA)** – techniques for identifying clusters of variables. These techniques have three main uses: (1) to understand the structure of a set of variables (e.g., Spearman and Thurstone used factor analysis to try to understand the structure of the latent variable 'intelligence'); (2) to construct a questionnaire to measure an underlying variable (e.g., you might design a questionnaire to measure burnout); and (3) to reduce a data set to a more manageable size while retaining as much of the original information as possible (e.g., factor analysis can be used to solve the problem of multicollinearity that we discovered in Chapter 9 by combining variables that are collinear).

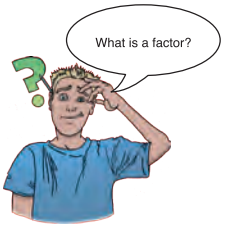
There are numerous examples of the use of factor analysis in science. Most readers will be familiar with the extroversion–introversion and neuroticism traits (Eysenck, 1953). Most other personality questionnaires are also based on factor analysis – notably Cattell's (1966a) 16 personality factors questionnaire – and these inventories are frequently used for recruiting purposes in industry (and even by some religious groups). Economists, for example, might also use factor analysis to see whether productivity, profits and workforce size can be reduced to an underlying dimension of company growth, and my friend Jeremy Miles told me of a biochemist who used it to analyse urine samples but I'm not sure whether he was taking the pi….

Both factor analysis and PCA aim to reduce a set of variables into a smaller set of dimensions (called 'factors' in factor analysis and 'components' in PCA). To non-statisticians, like me, the differences between a component and a factor are difficult to conceptualize (they are both linear models) and are hidden away in the maths behind the techniques.[2] However, there are important differences,

2 Principal component analysis is not the same as factor analysis. This doesn't stop frauds like me from discussing them as though they are. I tend to focus on the similarities between the techniques, which will reduce some statisticians (and psychologists) to tears. I'm banking on these people not needing to read this book, so I'll take my chances because I think it's easier for you if I give you a general sense of what the procedures do and not obsess too much about their differences. Once you have got the basics under your belt, feel free to obsess about their differences and complain to your friends about how awful the book by Field is …

which I'll discuss in due course. Most of the practical issues are the same regardless of whether you do factor analysis or PCA, so once the theory is over you can apply any advice I give to either factor analysis or PCA.

18.3 Factors and components Ⓑ



If we measure several variables, or ask people several questions about themselves, the correlation between each pair of variables (or questions) can be arranged in a table (just like the output from a correlation analysis as seen in Chapter 8). To scare you this table is sometimes called an *R*-matrix, or just **R** (because matrices tend to be denoted in bold). The preface goes over some matrix terminology if you need it. The diagonal elements of an *R*-matrix are all 1s because each variable correlates perfectly with itself. The off-diagonal elements are the correlation coefficients between pairs of variables, or questions.[3] Factor analysis attempts to achieve parsimony by explaining the maximum amount of *common variance* in a correlation matrix using the smallest number of explanatory constructs. These 'explanatory constructs' are known as latent variables (or *factors*) and they represent clusters of variables that correlate highly with each other. PCA differs in that it tries to explain the maximum amount of *total variance* (not just common variance) in a correlation matrix by transforming the original variables into linear components.

Imagine that we wanted to measure different aspects of what might make a person popular. We could administer several measures that we believe tap different aspects of popularity. So, we might measure a person's social skills (**Social skills**), their selfishness (**Selfish**), how interesting others find them (**Interest**), the proportion of time they spend talking about the other person during a conversation (**Talk (other)**), the proportion of time they spend talking about themself (**Talk (self)**) and their propensity to lie (**Liar**). We calculate the correlation coefficients for each pair of variables and create an *R*-matrix. Figure 18.2 shows this matrix. There appear to be two clusters of interrelating variables. First, the amount that someone talks about the other person during a conversation correlates highly with both the level of social skills and how interesting the other finds that person; social skills also correlate well with how interesting others perceive a person to be. The relationships between these three variables indicate that the better your social skills, the more interesting and talkative you are likely to be. Second, the amount that people talk about themselves within a conversation correlates well with how selfish they are and how much they lie. Being selfish also correlates highly with the degree to which a person tells lies. In short, selfish people are likely to lie and talk about themselves.

Factor analysis and PCA both aim to reduce this *R*-matrix into a smaller set of dimensions. In factor analysis these dimensions, or factors, are estimated from the data and are believed to reflect constructs that can't be measured directly. In this example, there appear to be two clusters that fit the bill. The first 'factor' seems to relate to general sociability, whereas the second 'factor' seems to relate to the way in which a person treats others socially (we might call it 'inconsideration'). It might, therefore, be assumed that popularity depends not only on your ability to socialize, but also on whether you are inconsiderate towards others. PCA, in contrast, transforms the data into a set of linear components; it does not estimate unmeasured variables, it just transforms measured ones. Strictly speaking, then, we shouldn't interpret components as unmeasured variables. Despite these differences, both techniques look for variables that correlate highly with a group of other variables, but do not correlate with variables outside of that group.

3 This matrix is called an *R*-matrix, or R, because it contains correlation coefficients and *r* usually denotes Pearson's correlation (see Chapter 8) – the *r* turns into a capital letter when it denotes a matrix.

	Talk (other)	Social skills	Interest	Talk (self)	Selfish	Liar
Talk (other)	1.000					
Social skills	0.772	1.000				
Interest	0.646	0.879	1.000			
Talk (self)	0.074	−0.120	0.054	1.000		
Selfish	−0.131	0.031	−0.101	0.441	1.000	
Liar	0.068	0.012	0.110	0.361	0.277	1.000

Factor 1 (Talk (other), Social skills, Interest)

Factor 2 (Talk (self), Selfish, Liar)

Figure 18.2 An *R*-matrix

18.3.1 Graphical representation Ⓑ

Factors and components can be visualized as the axis of a plot of the variables. The coordinates of variables along each axis represent the strength of relationship between that variable and each factor. In an ideal world a variable will have a large coordinate for one of the axes, and small coordinates for any others. This scenario indicates that this particular variable is related to only one factor. Variables that have large coordinates on the same axis are assumed to measure different aspects of some common underlying dimension. The coordinate of a variable along a classification axis is known as a **factor loading** (or *component loading*). The factor loading can be thought of as the Pearson correlation between a factor and a variable (see Jane Superbrain Box 18.1). Using what we know about interpreting correlation coefficients (see Section 8.4.2), we can square the factor loading for a variable to get a measure of its substantive importance to a factor.

Figure 18.3 shows such a plot for the popularity data (in which there were only two factors). Notice that for each factor, the axis ranges from –1 to +1, which are the outer limits of a correlation coefficient. The triangles represent the three variables that have high factor loadings (i.e., a strong relationship)

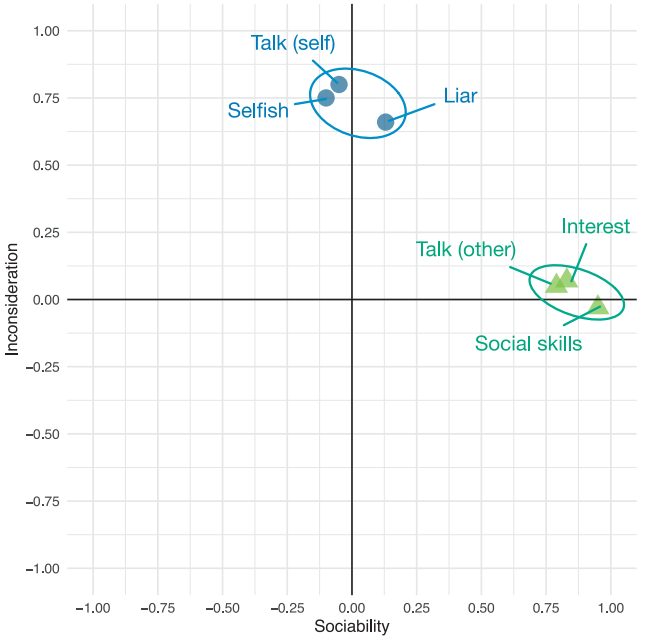


Figure 18.3 Example of a factor plot

with factor 1 (sociability: horizontal axis) but have a low correlation with factor 2 (inconsideration: vertical axis). Conversely, the circles represent variables that have high factor loadings with inconsideration but low loadings with sociability. This plot shows what we found in the *R*-matrix: selfishness, the amount a person talks about themself and their propensity to lie contribute to a factor which could be called inconsideration of others, and how much a person takes an interest in other people, how interesting they are and their level of social skills contribute to a second factor, sociability. Of course, if a third factor existed within these data it could be represented by a third axis (creating a 3-D plot). If more than three factors exist in a data set, then they cannot all be represented by a 2-D drawing.

18.3.2 Mathematical representation Ⓑ

Let's first look at components from PCA. The axes in Figure 18.3, which represent factors/components, are straight lines and any straight line can be described mathematically by a familiar equation.



The following equation reminds us of the equation describing a linear model:

$$Y_i = b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} \tag{18.1}$$

A component in PCA can be described in the same way:

$$\text{Component}_i = b_1\text{Variable}_{1i} + b_2\text{Variable}_{2i} + \ldots + b_n\text{Variable}_{ni} \tag{18.2}$$

You'll notice that there is no intercept in the equation because the lines intersect at zero (hence the intercept is zero), and there is also no error term because we are transforming the data from one form to another rather than making predictions beyond it (which would have associated error).

Sticking with our example of popularity, we found that there were two components: general sociability and inconsideration. We can, therefore, construct an equation that describes each factor in terms of the variables that have been measured:

$$\begin{aligned} Y_i &= b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} \\ \text{Sociability}_i &= b_{11}\text{Talk (other)}_i + b_{12}\text{Social Skills}_i + b_{13}\text{Interest}_i \\ &\quad + b_{14}\text{Talk (self)}_i + b_{15}\text{Selfish}_i + b_{16}\text{Liar}_i \\ \text{Inconsideration}_i &= b_{21}\text{Talk (other)}_i + b_{22}\text{Social Skills}_i + b_{23}\text{Interest}_i \\ &\quad + b_{24}\text{Talk (self)}_i + b_{25}\text{Selfish}_i + b_{26}\text{Liar}_i \end{aligned} \tag{18.3}$$

First, notice that the equations are identical in form: they both include all the variables that were measured. However, the values of *b* in the two equations will be different (depending on the relative importance of each variable to the component). In fact, we can replace each value of *b* with the coordinate of that variable on the plot in Figure 18.3 (i.e., replace the values of *b* with the factor loadings). The resulting equations are:

$$\begin{aligned} \text{Sociability}_i &= 0.87\text{Talk (other)}_i + 0.96\text{Social Skills}_i + 0.92\text{Interest}_i \\ &\quad + 0.00\text{Talk (self)}_i - 0.10\text{Selfish}_i + 0.09\text{Liar}_i \\ \text{Inconsideration}_i &= 0.01\text{Talk (other)}_i - 0.03\text{Social Skills}_i + 0.04\text{Interest}_i \\ &\quad + 0.82\text{Talk (self)}_i + 0.75\text{Selfish}_i + 0.70\text{Liar}_i \end{aligned} \tag{18.4}$$

For the **Sociability** component, the values of *b* are high for **Talk (other)**, **Social skills** and **Interest**. For the remaining variables (**Talk (self)**, **Selfish** and **Liar**) the values of *b* are very low (close to 0). This tells us that three of the variables are very important for that component (the ones with high values of *b*) and three are relatively unimportant (the ones with low values of *b*). The way the three variables clustered on the factor plot confirms this interpretation (Figure 18.3). The factor plot and these equations represent the same thing: the factor loadings in the plot are the *b*-values in these equations. For the second factor, **Inconsideration**, the opposite pattern can be seen: **Talk (self)**, **Selfish** and **Liar** have high *b*-values but the values for the remaining three variables are close to 0. Ideally, variables would have very high *b*-values for one component and very low *b*-values for all other components.

The factors in factor analysis are not represented in quite the same way as components. A factor is defined as:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \quad (18.5)$$

Variables = Variable Means + (Loadings × Common Factor) + Unique Factor

The Greek letters represent matrices or vectors containing numbers. If we put the Greek letters through Andy's magical translation machine then we can stop worrying about what the matrices contain and focus on what they represent. In factor analysis scores on the measured variables are predicted from the means of those variables plus a person's scores on the **common factor**s (i.e., factors that explain the correlations between variables) multiplied by their factor loadings, plus scores on any **unique factors** within the data (factors that cannot explain the correlations between variables).

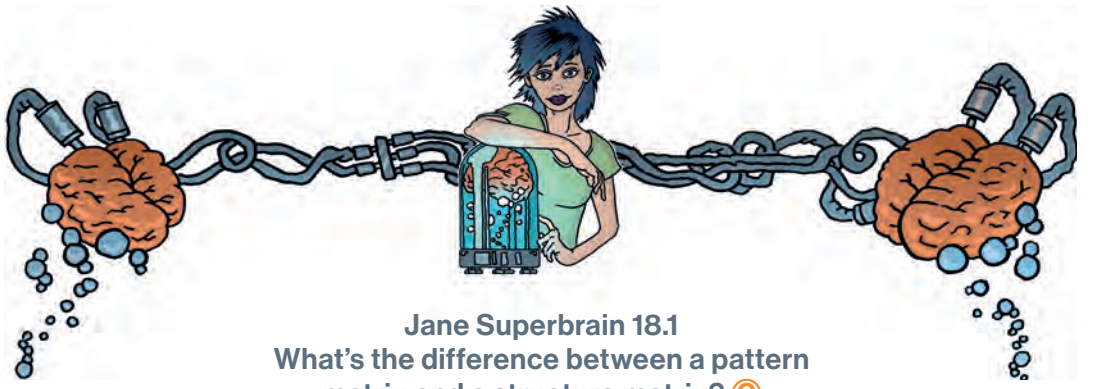
In a sense, the factor analysis model flips PCA on its head: in PCA we predict components from the measured variables, but in factor analysis we predict the measured variables from the underlying factors. Psychologists, for example, are usually interested in factors, because they're interested in how the stuff going on inside people's heads (the latent variables) affects how they answer the questions (the measured variables). The other big difference is that, unlike PCA, factor analysis contains an error term ($\boldsymbol{\delta}$ is made up of both scores on unique factors and measurement error). The fact that PCA assumes that there is no measurement error upsets a lot of people who use factor analysis.

Both factor analysis and PCA are linear models in which loadings are used as weights. In both cases, these loadings can be expressed as a matrix in which the columns represent each factor and the rows represent the loadings of each variable on each factor. For the popularity data this matrix would have two columns (one for each factor) and six rows (one for each variable). This matrix, $\boldsymbol{\Lambda}$, is:

$$\Lambda = \begin{pmatrix} 0.87 & 0.01 \\ 0.96 & -0.03 \\ 0.92 & 0.04 \\ 0.00 & 0.82 \\ -0.10 & 0.75 \\ 0.09 & 0.70 \end{pmatrix} \quad (18.6)$$

and is called the **factor matrix** or **component matrix** (if doing principal component analysis) – see Jane Superbrain Box 18.1 to find out about the different forms of this matrix. Try relating the elements to the loadings in Eq. 18.4 to give you an idea of what this matrix represents (in the case of PCA). For example, the top row represents the first variable, **Talk (other)**, which had a loading of 0.87 for the first factor (**Sociability**) and a loading of 0.01 for the second factor (**Inconsideration**).

The major assumption in factor analysis (but not PCA) is that these algebraic factors represent real-world dimensions, the nature of which must be *guessed at* by inspecting which variables load highly on the same factor. So, psychologists might believe that factors represent dimensions of the psyche, education researchers might believe they represent abilities and sociologists might believe



Jane Superbrain 18.1
What's the difference between a pattern matrix and a structure matrix? ©

The author has been a bit vague about factor loadings. Sometimes he's suggested that loadings can be thought of as the correlation between a variable and a given factor, then at other times he's described them as *b*-values (like in regression). The guy should make up his mind. However, his vagueness might not be the evidence of buffoonery that it seems. Broadly speaking, both correlation coefficients and *b*-values represent the relationship between a variable and linear model, so both descriptions allude to factor loadings expressing the relative contribution that a variable makes to a factor. If you understand that much, you'll be OK.

However, the factor loadings in a given model can be both correlation coefficients and *b*-values. In a few sections' time we'll discover that the interpretation of factor analysis is helped greatly by a technique known as *rotation*. There are two types: orthogonal and oblique rotation (see Section 18.4.6). When orthogonal rotation is used, the underlying factors are assumed to be independent, and the factor loading *is* the correlation between the factor and the variable, but it is also the *b*-value. Put another way, the values of the correlation coefficients are the same as the *b*-values. However, oblique rotation is used when the underlying factors are assumed to be related to each other, resulting in correlated factors. In these situations, the resulting correlations between variables and factors will differ from the corresponding *b*-values. In this case, there are, in effect, two different sets of factor loadings: the correlation coefficients between each variable and factor (contained in the factor **structure matrix**) and the *b*-values for each variable on each factor (contained in the factor **pattern matrix**). These coefficients can have quite different interpretations (see Graham et al., 2003).



they represent races or social classes. However, it is an extremely contentious point: some believe that the dimensions derived from factor analysis are real only in the statistical sense – and are real-world fictions.

18.3.3 Factor scores Ⓑ

Having discovered which factors exist, and estimated the equation that describes them, it should be possible to estimate a person's score on a factor, based on their scores for the constituent variables; these are known as **factor scores** (or *component scores* in PCA). For example, if we wanted to derive

a sociability score for someone after PCA, we could place their scores on the various measures into Eq. 18.4. This method is known as a *weighted average* and is rarely used because it is overly simplistic, but it is the easiest way to explain the principle. For example, imagine our six personality measures range from 1 to 10 and that someone scored the following: **Talk (other)** = 4, **Social skills** = 9, **Interest** = 8, **Talk (self)** = 6, **Selfish** = 8 and **Liar** = 6. We could plug these values into Eq. 18.4 to get a score for this person's sociability and their inconsideration:

$$\begin{aligned}
\text{Sociability}_i &= 0.87\text{Talk (other)}_i + 0.96\text{Social Skills}_i + 0.92\text{Interest}_i \\
&\quad + 0.00\text{Talk (self)}_i - 0.10\text{Selfish}_i + 0.09\text{Liar}_i \\
&= (0.87 \times 4) + (0.96 \times 9) + (0.92 \times 8) + (0.00 \times 6) - (0.10 \times 8) + (0.09 \times 6) \\
&= 19.22 \\
\text{Inconsideration}_i &= 0.01\text{Talk (other)}_i - 0.03\text{Social Skills}_i + 0.04\text{Interest}_i \\
&\quad + 0.82\text{Talk (self)}_i + 0.75\text{Selfish}_i + \mathbf{0}.70\text{Liar}_i \\
&= (0.01 \times 4) - (0.03 \times \mathbf{9}) + (0.04 \times 8) + (0.82 \times 6) + (0.75 \times 8) + (0.70 \times 6) \\
&= 15.21
\end{aligned} \tag{18.7}$$

The resulting scores of 19.22 and 15.21 reflect the degree to which this person is sociable and their inconsideration to others, respectively. This person scores higher on sociability than inconsideration. However, the scales of measurement used will influence the resulting scores, and if different variables use different measurement scales, then factor scores for different factors cannot be compared. As such, this method of calculating factor scores is poor, and more sophisticated methods are needed.

There are several techniques for calculating factor scores that use factor score coefficients as weights rather than the factor loadings. Factor score coefficients can be calculated in several ways. The simplest way is the regression method, in which the factor loadings are adjusted to take account of the initial correlations between variables; in doing so, differences in units of measurement and variable variances are stabilized.

To obtain the matrix of factor score coefficients ($\mathbf{B}$) we multiply the matrix of factor loadings by the inverse ($\mathbf{R}^{-1}$) of the original correlation or *R*-matrix (this is the same process that is used to estimate the *b*s in ordinary regression). You might remember from the previous chapter that multiplying by the inverse of a matrix is like division (see Section 17.3.9), so by multiplying the matrix of factor loadings by the inverse of the correlation matrix we are, conceptually speaking, dividing the factor loadings by the correlation coefficients. As such, the resulting factor score matrix represents the relationship between each variable and each factor, adjusting for the original relationships between pairs of variables. This matrix represents a purer measure of the *unique* relationship between variables and factors.

There are various methods for computing factor scores. One decision is whether you want the factor scores to be independent of each other (so that, for example, scores from factor 1 are uncorrelated with scores from factor 2). This in turn will likely depend on whether you believe that your factors should be independent or correlated (more on this in Section 18.4.6). Using the regression technique described above, the resulting factor scores have a mean of 0 and a variance equal to the squared multiple correlation between the estimated factor scores and the true factor values. Using the regression method, factor scores can correlate not only with factors other than the one on which they are based, but also with other factor *scores* from a different orthogonal factor. You might not want this, in which case consider either the Bartlett method or the **Anderson–Rubin method**. The Bartlett method produces scores that are unbiased and that correlate only with their own factor. The mean and standard deviation of the scores is the same as for the regression method. However, factor scores can still correlate with each other. The Anderson–Rubin method is a modification of the Bartlett method that produces factor scores that are uncorrelated and standardized (they have a mean of 0 and a standard deviation of 1). Tabachnick and Fidell (2012) conclude that the Anderson–Rubin method is best when uncorrelated scores are required but that the regression method is preferred in other circumstances simply because it is most easily understood.



Oliver Twisted
Please Sir, can I have some more ... matrix algebra?

'*The Matrix* – that was a good film,' enthuses Oliver. 'I want to dress in black and glide through the air as though time has stood still. Maybe the matrix of factor scores is as cool as the film.' I think you might be disappointed, Oliver, but we'll give it a shot. The matrix calculations of factor score coefficients for this example are available at www.discoverspss.com. Be afraid, be very afraid ...

You don't need to understand the maths behind any of the methods, although Grice (2001) provides a fairly non-technical discussion, but it is useful to understand what factor scores represent, namely, a composite score for each individual on a particular factor, and their uses. If the purpose of the factor analysis is to reduce a large set of data into a smaller subset of measurement variables, then the factor scores tell us an individual's score on this subset of measures. One use of factor scores is in analyses rather than the original data. For example, we could carry out a *t*-test to see whether extroverts are significantly more sociable than introverts using the factor scores for **Sociability**. A second use is in overcoming collinearity problems in linear models. If we have identified sources of multicollinearity in a linear model (see Section 9.9.3) then a solution is to reduce collinear predictors to a subset of uncorrelated factors using PCA and enter the component scores as predictors instead of the raw variable scores. By using uncorrelated component scores (e.g., by using the Anderson–Rubin method – see above) as predictors we can be confident that there will be no correlation between predictors – hence, no multicollinearity.

18.4 Discovering factors Ⓑ

Now you've grasped what factors and components are, let's delve into how to find or estimate these mythical beasts.

18.4.1 Choosing a method Ⓑ

There are several methods for unearthing factors, and the one you chose will depend on what you hope to do with the analysis (for a review see Tinsley & Tinsley, 1987). There are two things to consider: whether you want to generalize the findings from your sample to a population and whether you

are exploring your data or testing a specific hypothesis. This chapter describes techniques for exploring data using factor analysis. Testing hypotheses about the structures of latent variables and their relationships to each other is a big topic and needs a different software package (IBM SPSS AMOS, R or MPlus, to name a few), so I'm not going to cover it. For those interested in hypothesis testing techniques (known as **confirmatory factor analysis**) I recommend Brown (2015).

Assuming the goal is to explore, we need to consider whether we want to apply our findings to the sample collected (descriptive method) or to generalize our findings to a population (inferential methods). Factor analysis was originally developed to explore data to generate future hypotheses, and it was assumed that the technique would be applied to the entire population of interest. In other words, certain techniques assume that the sample used *is* the population and results cannot be extrapolated beyond that sample. Principal component analysis is an example, as are principal factors analysis (*principal axis factoring*) and image covariance analysis (*image factoring*). Of these, principal component analysis and principal factors analysis are the preferred methods and usually result in similar solutions (see Section 18.4.3). If you use one of these methods, you should restrict your conclusions to the sample collected. If you want to generalize the results, you would need to cross-validate the factor structure in a different sample.

A different approach assumes that participants are randomly selected but that the variables measured constitute the population of variables in which we're interested. By assuming this, it is possible to generalize from the sample to a larger population, but with the caveat that any findings hold true only for the set of variables measured (because we've assumed this set constitutes the entire population of variables). These techniques are based on the idea of minimizing the difference (or error) between the original correlation matrix and the version of it reproduced using the factor solution. In other words, it tries to find the factor solution that reproduces the original pattern of correlations with the minimum error. As with many other models in this book, 'minimum error' can be determined using ordinary least squares (so-called 'minres') and maximum likelihood (see Harman, 1976). Kaiser's **alpha factoring** (Kaiser & Caffrey, 1965) also falls into this group. The choice of method depends largely on what generalizations, if any, you want to make from your data.

18.4.2 Communality Ⓑ

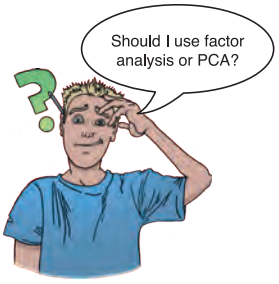
The idea of what variance is and how it is calculated should, by now, be an old friend with whom you enjoy tea and biscuits (if not, see Chapter 2). The total variance for a variable in the *R*-matrix will have two components: some of it will be shared with other variables or measures (**common variance**) and some of it will be specific to that measure (**unique variance**). We tend to use the term *unique variance* to refer to variance that can be reliably attributed to only one measure. However, there is also variance that is specific to one measure but not reliably so, known as error or **random variance**. The proportion of common variance present in a variable is known as the **communality**. As such, a variable that has no unique variance (or random variance) would have a communality of 1; a variable that shares none of its variance with any other variable would have a communality of 0.

Factor analysis tries to find common underlying dimensions within the data and so is primarily concerned with the common variance. In short, we want to find out how much of the variance in our data is common. This aim presents us with a logical impasse: to do the factor analysis we need to know the proportion of common variance in the data, but the only way to find out the extent of the common variance is by carrying out a factor analysis! There are two solutions. The first is to assume that all variance is common by assuming that the communality of every variable is 1. By making this assumption we transpose our original data into constituent linear components. This procedure is PCA. Remember that I said earlier that PCA assumes no measurement error? Well, by setting the communalities to 1, we assume that all variance is common and there is no random variance at all.

The second solution is used in factor analysis and is to estimate the amount of common variance by estimating communality values for each variable. There are various methods of estimating communalities.

For example, principal axis factoring uses the squared multiple correlation (SMC) of each variable with all others. So, for the popularity data, imagine you fit a linear model with one measure (**Selfish**) as the outcome and the other five measures as predictors: the resulting multiple R^2 (see Section 9.2.4) would act as an estimate of the communality for the variable **Selfish**. These estimates allow the factor analysis to be done. Once the underlying factors have been extracted, new communalities can be calculated that represent the multiple correlation between each variable and the factors extracted. Therefore, the communality is a measure of the proportion of variance explained by the extracted factors. Methods based on minimizing error (e.g. maximum likelihood) use the squared multiple correlation (SMC) as initial estimates of the communalities, then use these values to estimate the error between the original pattern of correlations and those reproduced by the model, and then adjust the estimates iteratively until this error is minimized.

18.4.3 Factor analysis or PCA? Ⓑ



Factor analysis and principal component analysis differ in the communality estimates that are used. As I have hinted before, factor analysis derives a mathematical model from which factors are estimated, whereas principal component analysis decomposes the original data into a set of linear variates (see Dunteman, 1989, Chapter 8, for more detail on the differences between the procedures). As such, only factor analysis can estimate the underlying factors and it relies on various assumptions for these estimates to be accurate. PCA is concerned only with establishing which linear components exist within the data and how a particular variable might contribute to a given component.

Based on an extensive literature review, Guadagnoli and Velicer (1988) concluded that the solutions generated from PCA differ little from those derived from factor-analytic techniques. In reality, with 30 or more variables and communalities greater than 0.7 for all variables, different solutions are unlikely; however, with fewer than 20 variables and any low communalities (less than 0.4) differences can occur (Stevens, 2002).

The flip side of this argument is eloquently described by Cliff (1987), who observed that proponents of factor analysis 'insist that components analysis is at best a common factor analysis with some error added and at worst an unrecognizable hodgepodge of things from which nothing can be determined' (p. 349). Indeed, feeling is strong on this issue, with some arguing that when PCA is used it should not be described as a factor analysis (oops!) and that you should not impute substantive meaning to the resulting components. Ultimately, as I hope to have made clear, they are doing slightly different things.

18.4.4 Theory behind PCA Ⓒ

The theory behind factor analysis involves a lot of matrix algebra. If you're the sort of person who likes matrix algebra then fair play to you and I recommend Mulaik (2009). Instead, for the benefit of those of us who enjoy the warm, fuzzy glow that a world without matrix algebra gives us, this section focuses on the theory of PCA.

Principal component analysis works in a very similar way to MANOVA and discriminant function analysis (see Chapter 17). In MANOVA, various sum of squares and cross-product matrices were calculated that contained information about the relationships between dependent variables. I mentioned before that these SSCP matrices can be converted to variance–covariance matrices, which represent the same information but in averaged form (i.e., taking account of the number of observations). I also pointed out that by dividing each element by the relevant standard deviation the variance–covariance matrices become standardized. The result is a correlation matrix. In PCA we usually deal with correlation matrices (though it is possible to analyse a variance–covariance matrix too) and my point is that this matrix represents the same information as an SSCP matrix in MANOVA.

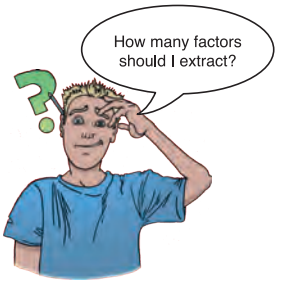
In MANOVA, because we were comparing groups we ended up looking at the variates or components of the SSCP matrix that represented the ratio of the model variance to the error variance. These variates were linear dimensions that separated the groups tested, and we saw that the dependent variables mapped onto these underlying components. In short, we looked at whether the groups could be separated by some linear combination of the dependent variables. These variates were found by calculating the eigenvectors of the SSCP. The number of variates obtained was the smaller of p (the number of dependent variables) and $k - 1$ (where k is the number of groups).

In PCA we do much the same thing but using the overall correlation matrix (because we're not interested in comparing groups of scores). So, and I'm simplifying things a little, we take a correlation matrix and calculate the variates. There are no groups of observations, and so the number of variates calculated will always equal the number of variables measured (p). The variates are described, as for MANOVA, by the eigenvectors associated with the correlation matrix. The elements of the eigenvectors are the weights of each variable on the variate. These values are the loadings described earlier (i.e., the b-values in Eq. 17.22). The largest eigenvalue associated with each of the eigenvectors provides a single indicator of the substantive importance of each component. The basic idea is that we retain components with relatively large eigenvalues and ignore those with relatively small eigenvalues.

Factor analysis works differently, but there are similarities. Rather than using the correlation matrix, factor analysis starts by estimating the communalities between variables using the SMC (as described earlier). It then replaces the diagonal of the correlation matrix (the 1s) with these estimates. Then the eigenvectors and associated eigenvalues of this matrix are computed. Again, these eigenvalues tell us about the substantive importance of the factors, and a decision is made about how many factors to retain based on them. Loadings and communalities are then estimated using only the retained factors.

18.4.5 Factor extraction: eigenvalues and the scree plot ©

In both PCA and factor analysis, not all factors/components are retained. The process of deciding how many factors to keep is called **extraction**. I mentioned above that eigenvalues associated with a variate indicate the substantive importance of that factor. Therefore, it is logical to retain only factors with large eigenvalues. This section looks at how we determine whether an eigenvalue is large enough to represent a meaningful factor.



Cattell (1966b) suggested plotting each eigenvalue (Y-axis) against the factor with which it is associated (X-axis). This plot is a **scree plot** (because it looks like a rock face with a pile of debris, or scree, at the bottom). I mentioned earlier that it is possible to obtain as many factors as there are variables and that each has an associated eigenvalue. By plotting the eigenvalues, the relative importance of each factor becomes apparent. Typically, there will be a few factors with quite high eigenvalues, and many factors with relatively low eigenvalues, and so this plot has a very characteristic shape: there is a sharp descent in the curve followed by a tailing off (see Figure 18.4). The point of inflexion is where the slope of the line changes dramatically, and Cattell (1966b) suggested using this point as the cut-off for retaining factors. In Figure 18.4, imagine drawing two straight lines (the red dashed lines): one summarizing the vertical part of the plot and the other summarizing the horizontal part. The point of inflexion is the data point at which these two lines meet. You retain only factors to the left of the point of inflexion (and do not include the factor at the point of inflexion itself),[4] so in both examples in Figure 18.4 we would extract two factors because the point of inflexion

4 Cattell (1996b) advised including the factor at the point of inflexion as well because it represents an error factor, or 'garbage can', as he put it. In practice the 'garbage can' factor is rarely retained but it's useful to reflect on Cattell's advice when looking for convergent evidence for how many factors to retain.

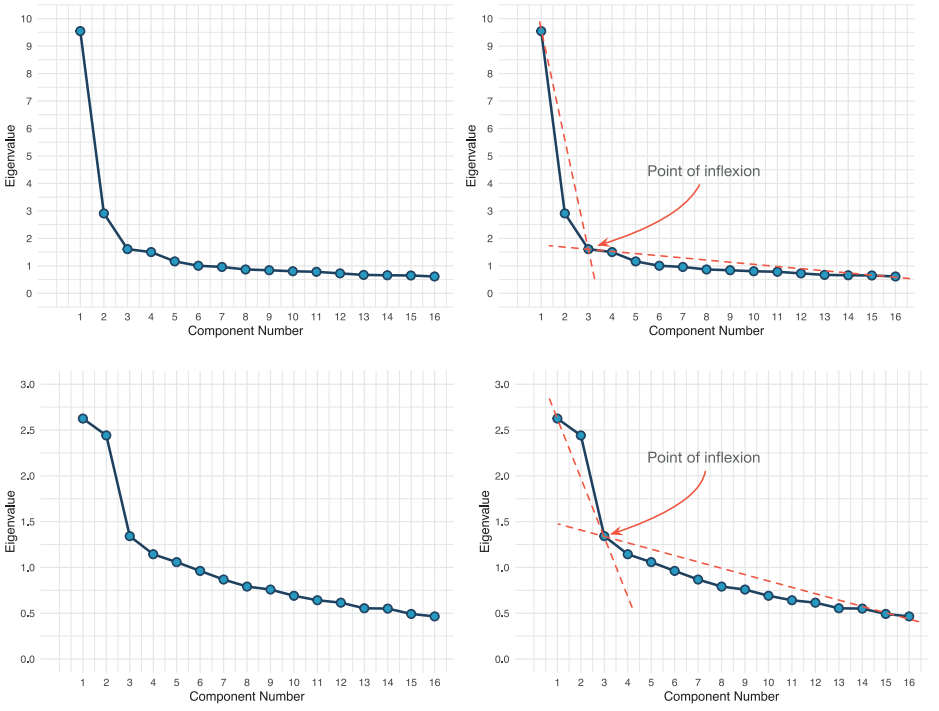


Figure 18.4 Examples of scree plots for data that probably have two underlying factors

occurs at the third data point (factor). With a sample of more than 200 participants, the scree plot provides a fairly reliable criterion for factor selection (Stevens, 2002).

An alternative to the scree plot is to use the eigenvalues themselves. The logic is that, because the eigenvalues represent the amount of variation explained by a factor, you set a criterion value that represents a substantial amount of variation and retain factors with eigenvalues above this criterion. There are two common criteria: **Kaiser's criterion** (Kaiser, 1960, 1970) is to retain factors with eigenvalues greater than 1 (followed by normal varimax rotation),[5] or one can take a more liberal value of 0.7 (Jolliffe, 1972, 1986). The difference between how many factors are retained using these two methods can be dramatic. Generally speaking, Kaiser's criterion overestimates the number of factors to retain (see Jane Superbrain Box 18.2), which means that Joliffe's criterion *really* overestimates. There is evidence that Kaiser's criterion is accurate when the number of variables is less than 30 and the resulting communalities (after extraction) are all greater than 0.7; it can also be accurate when the sample size exceeds 250 and the average communality is greater than or equal to 0.6.

A third approach is **parallel analysis** (Horn, 1965), which we can do using syntax and R (SPSS Tip 18.1). There are different ways to do a parallel analysis, but essentially each eigenvalue (which represents the size of the factor) is compared against an eigenvalue for the corresponding factor in many (e.g., 20) randomly generated data sets that have the same characteristics as the data being analysed. Factors with eigenvalues larger than their 'random' counterparts are retained on the basis that they explain more variance than a factor from data with no underlying factor structure. This is a bit like asking whether our observed factor is bigger than a non-existing factor. The eigenvalues used in the parallel analysis can be based on decomposing the correlation matrix as in PCA (PA_{PCA}) or on a common factor model (PA_{EFA}). Also, the reference eigenvalue from the random samples can be based on the mean eigenvalue in those samples, or the eigenvalue at a particular percentile across the

5 A colleague of Kaiser, Chester Harris, referred to this procedure as a 'little jiffy'.

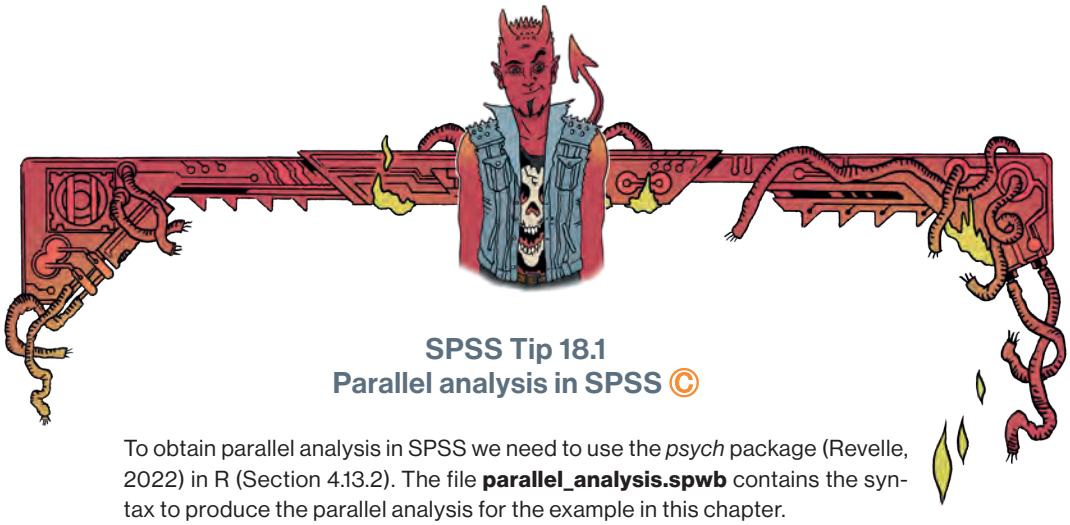
samples (typically the 95th percentile). When discussing the performance of parallel analysis I will refer to parallel analysis when the eigenvalues are based on the 95th percentile of the random samples.

If you're lucky, the methods above will give a consistent answer as to how many factors to retain. What do you do, though, when they disagree, which they often do? Of parallel analysis, the scree plot and Kaiser's criterion, Kaiser's criterion is, in general, worst and parallel analysis best (Zwick & Velicer, 1986), so it's not a bad idea to choose parallel analysis as a default and if it agrees with the screen plot then happy times. Needless to say, it isn't as simple as that. Auerswald and Moshagen (2019) compared various factor extraction criteria in a simulation study. We can summarize the relevant findings as follows:

- When there was one underlying factor to be detected, PA_{PCA} pretty much always correctly identified one factor even in samples sizes as low as 100, whereas PA_{EFA} was correct about 92% of the time. Kaiser's criterion using PCA (KC_{PCA}) had lower accuracy (around 79% correct) than when using eigenvalues based on EFA (KC_{EFA}), for which accuracy was around 94%.
- With multiple independent (so-called 'orthogonal') underlying factors, it matters how many indicators (measured variables) you have per factor and the sample size, but across all sample sizes parallel analysis outperformed Kaizer's criteria. PA_{PCA} and PA_{EFA} were close to 100% accurate even in samples size of 100 when there were at least eight indicator variables. With four indicator variables, these methods were close to 100% accurate in sample sizes greater than 200. Kaiser's criterion in stark contrast was poor. KC_{EFA} only got above 90% accuracy in sample sizes of at least 500 and with at least eight indicators and KC_{PCA} struggled to achieve 60% accuracy in nearly all of the situations studied.
- With multiple correlated underlying factors (arguably the most realistic situation) things get tricky. As the correlation between factors increases, accuracy drops for all methods. With only four indicator variables, PA_{PCA} and PA_{EFA} perform really badly (they underestimate the number of factors) even in huge samples. In contrast, KC_{PCA}, because of its tendency to overestimate factors compared to other methods, is more accurate provided that the correlation between factors is fairly low ($\rho \leq 0.5$) and the sample size is huge ($N = 1000$). PA_{PCA} and PA_{EFA} were close to 100% accurate in samples sizes of at least 200 when there were at least eight indicator variables and with a modest correlation between factors ($\rho = 0.25$). To achieve similar accuracy for more highly correlated factors ($\rho = 0.5$) a sample size of 500 was necessary. To be comparably accurate in these situations, KC_{PCA} required sample sizes of $N \geq 500$ and KC_{EFA} never managed more than 60% accuracy. With 12 indicator variables PA_{PCA} and PA_{EFA} were close to 100% accurate in samples sizes of at least 200 even with quite strong correlations between factors ($\rho = 0.5$), again KC_{PCA} needed larger samples to match this accuracy ($N = 1000$) and KC_{EFA} never got anywhere near. All methods fell apart spectacularly when factors were highly correlated ($\rho = 0.75$).

What should we make of all of this? Here are the take-homes: (1) if your factors are highly correlated all methods will be poor; (2) with eight or more indicators and a sample size of 200+ parallel analysis (PCA or EFA varieties) are going to do a pretty decent job; (3) the only time Kaiser's criterion will do a better job than parallel analysis is when you have small numbers of indicators (e.g., four) and correlated factors, but even then it'll do a *good* job only in large samples and using the PCA method. Auerswald and Moshagen (2019) compare some other methods to the ones discussed so it's well worth looking at their findings for more nuanced advice.

When thinking about factor extraction, you can also consider the communalities of the factors. Remember that communalities represent the common variance: if the values are 1 then all common variance is accounted for, and if the values are 0 then no common variance is accounted for. In both PCA and factor analysis we determine how many factors/components to extract and then re-estimate the communalities. The factors we retain will not explain all the variance in the data (because we have discarded some information), and so the communalities after extraction will always be less than 1. The factors retained do not map perfectly onto the original variables – they merely reflect the common

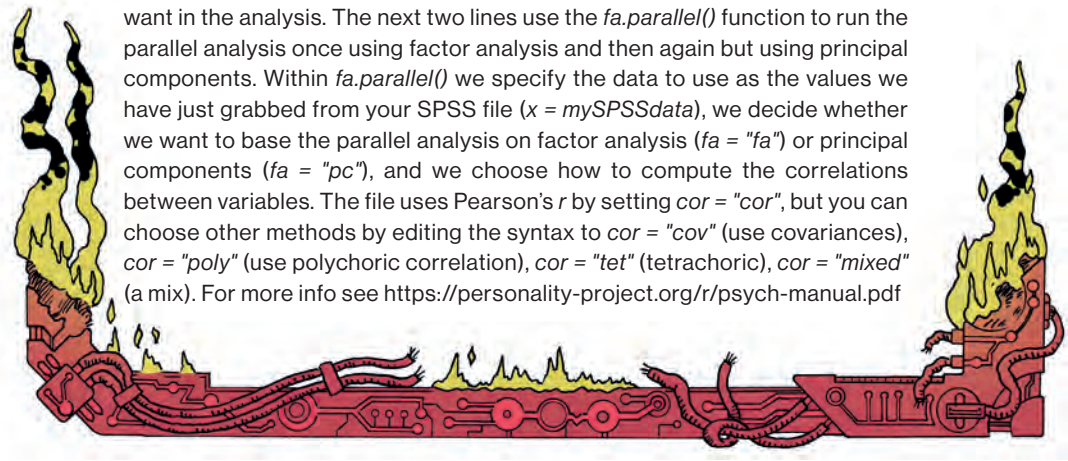


SPSS Tip 18.1
Parallel analysis in SPSS ©

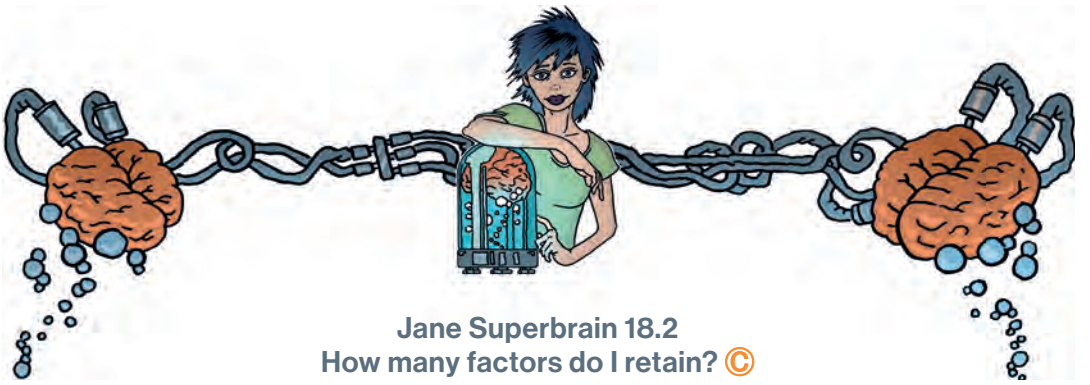
To obtain parallel analysis in SPSS we need to use the *psych* package (Revelle, 2022) in R (Section 4.13.2). The file **parallel_analysis.spwb** contains the syntax to produce the parallel analysis for the example in this chapter.

```
BEGIN PROGRAM R.
library(psych)
mySPSSdata =  spssdata.GetDataFromSPSS()
fa.parallel(x = mySPSSdata, cor = "cor", fa = "fa")
fa.parallel(x = mySPSSdata, cor = "cor", fa = "pc")
END PROGRAM.
```

The first line within the program loads the *psych* library. The next line grabs the currently active data and stores it as **mySPSSdata**. So, you'd need to have **saq.sav** open and active before running this syntax. For your own data, the SPSS file needs to include *only* the variables that you want to include in the parallel analysis. Therefore, if your file contains other variables you'll need to save a copy of the file where you have removed any variables that you don't want in the analysis. The next two lines use the *fa.parallel()* function to run the parallel analysis once using factor analysis and then again but using principal components. Within *fa.parallel()* we specify the data to use as the values we have just grabbed from your SPSS file (*x = mySPSSdata*), we decide whether we want to base the parallel analysis on factor analysis (*fa = "fa"*) or principal components (*fa = "pc"*), and we choose how to compute the correlations between variables. The file uses Pearson's *r* by setting *cor = "cor"*, but you can choose other methods by editing the syntax to *cor = "cov"* (use covariances), *cor = "poly"* (use polychoric correlation), *cor = "tet"* (tetrachoric), *cor = "mixed"* (a mix). For more info see https://personality-project.org/r/psych-manual.pdf



variance in the data. Because communalities represent a loss of information they are important statistics. The closer the communalities are to 1, the better our factors are at explaining the original data. It is logical that the greater the number of factors retained, the greater the communalities will be (because less information is discarded); therefore, the communalities are good indices of whether too few factors have been retained. In fact, with generalized least squares factor analysis and maximum likelihood factor analysis you can get a statistical measure of the goodness of fit of the factor solution (see the next chapter for more on goodness-of-fit tests). This basically measures the proportion of variance that the factor solution explains (so can be thought of as comparing communalities before



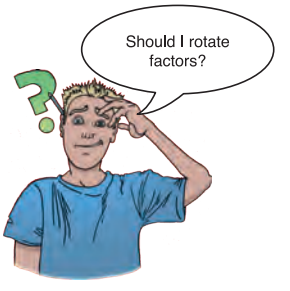
Jane Superbrain 18.2
How many factors do I retain? ©

A fundamental problem with Kaiser's criterion (Nunnally & Bernstein, 1994) is that an eigenvalue of 1 means different things in different analyses: with 100 variables it means that a factor explains 1% of the variance, but with 10 variables it means that a factor explains 10% of the variance. These two situations are very different and a single rule that covers both is inappropriate. An eigenvalue of 1 also means only that the factor explains as much variance as a variable, which rather defeats the original intention of the analysis to reduce variables down to 'more substantive' underlying factors. Consequently, Kaiser's criterion often overestimates the number of factors. Jolliffe's criterion is even worse (a factor explains less variance than a variable).



and after extraction). As a final word of advice, your decision on how many factors to extract will depend also on why you're doing the analysis; for example, if you're trying to overcome multicollinearity problems in regression, then it might be better to extract too many factors than too few.

18.4.6 Factor rotation ©



Once factors have been extracted, it is possible to calculate the degree to which variables load onto these factors (i.e., calculate the loadings for each variable on each factor). Generally, you will find that most variables have high loadings on the most important factor and small loadings on all other factors. This characteristic makes interpretation difficult, and so a technique called factor **rotation** is used to discriminate factors. If we visualize our factors as an axis along which variables can be plotted, then factor rotation effectively rotates these axes such that variables are loaded maximally to only one factor. Let's return to our example where we had measures of popularity that produced two factors (sociability and inconsideration). Earlier we looked at having just three measures loading highly onto each factor. Imagine now that we'd measured 20 variables, and it turned out that 10 seemed to reflect sociability and the other 10 inconsideration. Figure 18.5 shows two scenarios. As with the factor plot in Figure 18.3, the dashed lines represent the factors, and by looking at the coordinates we can see that the blue circles have high loadings for inconsideration (they are a long way

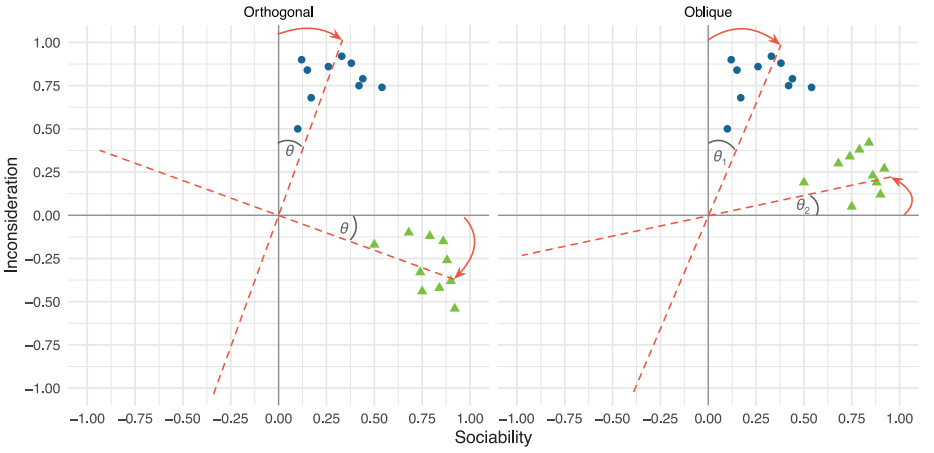


Figure 18.5 Schematic representations of factor rotation. The left-hand plot displays orthogonal rotation, whereas the right-hand plot displays oblique rotation (see text for more details). θ is the angle through which the axes are rotated

up this axis) and small to medium loadings for sociability (they are not very far along this axis). Conversely, the green triangles have high loadings for sociability and small-to-medium loadings for inconsideration. Factor rotation amounts to rotating the axes (the red dashed lines) to try to ensure that both clusters of variables are intersected by the factor to which they relate most. After rotation, the loadings of the variables are maximized on one factor (the factor that intersects the cluster) and minimized on the remaining factor(s). If an axis passes through a cluster of variables, then these variables will have a loading close to zero on the opposite axis. If this idea is confusing, then look at Figure 18.5 and think about the values of the coordinates before and after rotation (this is best achieved by turning the book when you look at the rotated axes).

There are two flavours of rotation. **Orthogonal rotation** is shown in the left-hand side of Figure 18.5 (left). We saw in Chapter 12 that the term *orthogonal* means unrelated, and in this context it means that we rotate factors while keeping them independent, or uncorrelated. Before rotation, all factors are independent (i.e., they have a correlation of zero), and orthogonal rotation ensures that the factors remain this way. That is why in Figure 18.5 (right) the axes remain perpendicular as they rotate. **Oblique rotation** allows factors to correlate, hence, the axes of the right-hand diagram of Figure 18.5 do not remain perpendicular: they can rotate by different amounts and in different directions if need be.

IBM SPSS Statistics software implements three methods of orthogonal rotation (**varimax**, **quartimax** and **equamax**) and two methods of oblique rotation (**direct oblimin** and **promax**). Quartimax rotation attempts to maximize the spread of factor loadings for a variable across all factors. Therefore, interpreting variables becomes easier but often results in lots of variables loading highly on a single factor. Varimax is the opposite in that it attempts to maximize the dispersion of loadings within factors. Therefore, it tries to load a smaller number of variables highly onto each factor, resulting in more interpretable clusters of factors. Equamax is a hybrid of the other two approaches and is reported to behave fairly erratically (see Tabachnick & Fidell, 2012). If you use orthogonal rotation then you should probably select varimax because it is a good general approach that simplifies the interpretation of factors.

The case with oblique rotations is more complex because correlation between factors is permitted. Direct oblimin determines the degree to which factors are allowed to correlate by the value of a constant called delta. The default value in SPSS is 0, and this ensures that high correlation between factors is not allowed (this is known as *direct quartimin rotation*). If you choose to set delta to a value greater than 0 (up to 0.8), then you can expect highly correlated factors; if you set delta less than 0 (down to –0.8) you can expect less correlated factors. The default setting of zero is sensible for most analyses, and I don't recommend changing it unless you know what you are doing (see Pedhazur and Schmelkin,

1991, p. 620). Promax is a faster procedure designed for very large data sets. If you use oblique rotation then, other things being equal, use direct oblimin.

The choice of orthogonal or oblique rotation depends on: (1) whether there is a good theoretical reason to suppose that the factors should correlate or be independent; and (2) how the variables cluster on the factors before rotation. On the first point, it is probably rare that you would measure a set of related variables and expect their underlying dimensions to be completely independent. For example, we wouldn't expect sociability to be completely independent of inconsideration (presumably inconsiderate people might not like other people much, which might make them antisocial). Therefore, on theoretical grounds, we would choose oblique rotation. There are strong grounds to believe that orthogonal rotations are a complete nonsense for naturalistic data, and certainly for any data involving humans (can you think of any psychological construct that is not in any way correlated with some other psychological construct?) As such, some argue that orthogonal rotations should never be used.

On the second point, Figure 18.5 demonstrates how the positioning of clusters will determine how successful a rotation will be (note the position of the green triangles). If an orthogonal rotation was carried out on the right-hand diagram it would be considerably less successful in maximizing loadings than the oblique rotation that is displayed.

A compromise is to run the analysis using both types of rotation. Pedhazur and Schmelkin (1991) suggest that if the oblique rotation demonstrates a negligible correlation between the extracted factors then it is reasonable to use the orthogonally rotated solution. If the oblique rotation reveals a correlated factor structure, then the orthogonally rotated solution should be discarded. We can check the relationships between factors using the **factor transformation matrix**, $\mathbf{\Lambda}$, which is used to convert the unrotated factor loadings into the rotated ones. Values in this matrix represent the angle through which the axes have been rotated, or the degree to which factors have been rotated.

18.4.7 Interpreting the factor structure Ⓑ

Once a factor structure has been found, it needs to be interpreted. Earlier I wrote that the loadings were a gauge of the substantive importance of a given variable to a given factor. Therefore, it makes sense that we use these values to place variables with factors. Every variable will have a loading on every factor, so what we're looking for is variables that load highly on a given factor. Once we've identified these variables, we look for a theme within them.

What do I mean by 'load highly'? It is possible to assess the statistical significance of a loading (after all, it is simply a correlation coefficient or *b*-value), but, as with all significance tests, the *p*-value depends on the sample size. Typically if you're using factor analysis you'll have a large sample, and in large samples even small loadings will be 'significant'. So, the significance of a loading is probably best ignored, especially given that the loading itself is an effect size. Stevens (2002) recommends interpreting factor loadings with an absolute value greater than 0.4 (the factor explains around 16% of the variance in the variable) although some researchers opt for the lower criterion of 0.3. There's no magic to these numbers, it's more about interpreting the patterns of the loadings.

18.5 An anxious example Ⓑ

Factor analysis is used frequently to develop questionnaires. I have noticed that a lot of students become very stressed about SPSS. Imagine that I wanted to design a questionnaire to measure a trait that I termed 'SPSS anxiety'. I devised a questionnaire to measure various aspects of students' anxiety towards learning SPSS, the SAQ (Figure 18.6). I generated questions based on interviews with anxious and non-anxious students and came up with 23 possible questions to include. Each question was a statement followed by a five-point Likert scale: 'strongly disagree', 'disagree', 'neither agree nor disagree',

The SPSS Anxiety Questionnaire (SAQ)

		SD	D	N	A	SA
1.	Statistics makes me cry	○	○	○	○	○
2.	My friends will think I'm stupid for not being able to cope with SPSS	○	○	○	○	○
3.	Standard deviations excite me	○	○	○	○	○
4.	I dream that Pearson is attacking me with correlation coefficients	○	○	○	○	○
5.	I don't understand statistics	○	○	○	○	○
6.	I have little experience of computers	○	○	○	○	○
7.	All computers hate me	○	○	○	○	○
8.	I have never been good at mathematics	○	○	○	○	○
9.	My friends are better at statistics than me	○	○	○	○	○
10.	Computers are useful only for playing games	○	○	○	○	○
11.	I did badly at mathematics at school	○	○	○	○	○
12.	People try to tell you that SPSS makes statistics easier to understand but it doesn't	○	○	○	○	○
13.	I worry that I will cause irreparable damage because of my incompetence with computers	○	○	○	○	○
14.	Computers have minds of their own and deliberately go wrong whenever I use them	○	○	○	○	○
15.	Computers are out to get me	○	○	○	○	○
16.	I weep openly at the mention of central tendency	○	○	○	○	○
17.	I slip into a coma whenever I see an equation	○	○	○	○	○
18.	SPSS always crashes when I try to use it	○	○	○	○	○
19.	Everybody looks at me when I use SPSS	○	○	○	○	○
20.	I can't sleep for thoughts of eigenvectors	○	○	○	○	○
21.	I wake up under my duvet thinking that I am trapped under a normal distribution	○	○	○	○	○
22.	My friends are better at SPSS than I am	○	○	○	○	○
23.	If I am good at statistics people will think I am a nerd	○	○	○	○	○

Figure 18.6 The SPSS anxiety questionnaire (SAQ)

'agree' and 'strongly agree' (SD, D, N, A and SA respectively). What's more, I wanted to know whether anxiety about SPSS could be broken down into specific forms of anxiety. In other words, what latent variables contribute to anxiety about SPSS?[6]

With a little help from a few lecturer friends I collected 2571 completed questionnaires. Load the data (**saq.sav**) into SPSS. Note that each question (variable) is represented by a different column: there

6 Such is my 'talent' (cough) for seamlessly blurring fact and fiction that I have had some people think that this is an actual bit of research that I've done. It's not, I made the example and data up.



Oliver Twisted
Please, Sir, can I have some more ... questionnaires?

'I'm going to design a questionnaire to measure one's propensity to pick a pocket or two,' says Oliver. 'But how would I go about doing it?' You'd read the useful information about the dos and don'ts of questionnaire design at www.discoverspss.com, that's how. Rate how useful it is on a Likert scale from 1 = not useful at all, to 5 = very useful.

are 23 variables labelled **question_01** to **question_23** and each has a label indicating the question. By labelling my variables I can be very clear about what each variable represents (this is the value of giving your variables full titles rather than just using restrictive column headings).

18.5.1 General procedure Ⓐ

Figure 18.7 shows the general procedure for conducting factor analysis or PCA. First screen the data. Then once you embark on the main analysis consider how many factors to retain and what rotation to use. If you are using the analysis to look at the factor structure of a questionnaire then follow up with a reliability analysis (see Section 18.9).

18.5.2 Adequacy of the data for factor analysis Ⓑ

There are measures of so-called 'sampling adequacy' such as the **Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy** (Kaiser, 1970; Kaiser & Rice, 1974).[7] Despite the name, the KMO is *not* a measure of sample *size* adequacy. Holding the other things constant, KMO improves as (1) the number of variables increases, (2) the number of underlying factors decreases, (3) the sample size increases and (4) the general level of correlation between variables increases (Kaiser, 1970). As such, the KMO is general measure of the adequacy of the data for factor analysis. The KMO can be calculated for individual and multiple variables and represents the ratio of the squared correlation between variables to the squared partial correlation between variables. The KMO statistic varies between 0 and 1. A value of 0 indicates that the sum of partial correlations is large relative to the sum of correlations, indicating diffusion in the pattern of correlations (hence, factor analysis is likely to be inappropriate). A value close to 1 indicates that patterns of correlations are relatively compact and so factor analysis should yield distinct and reliable factors. Kaiser and Rice (1974) provided appealing guidelines, especially if you like the letter M:

7 There are different versions of the KMO. SPSS implements the one in Kaiser and Rice (1974).

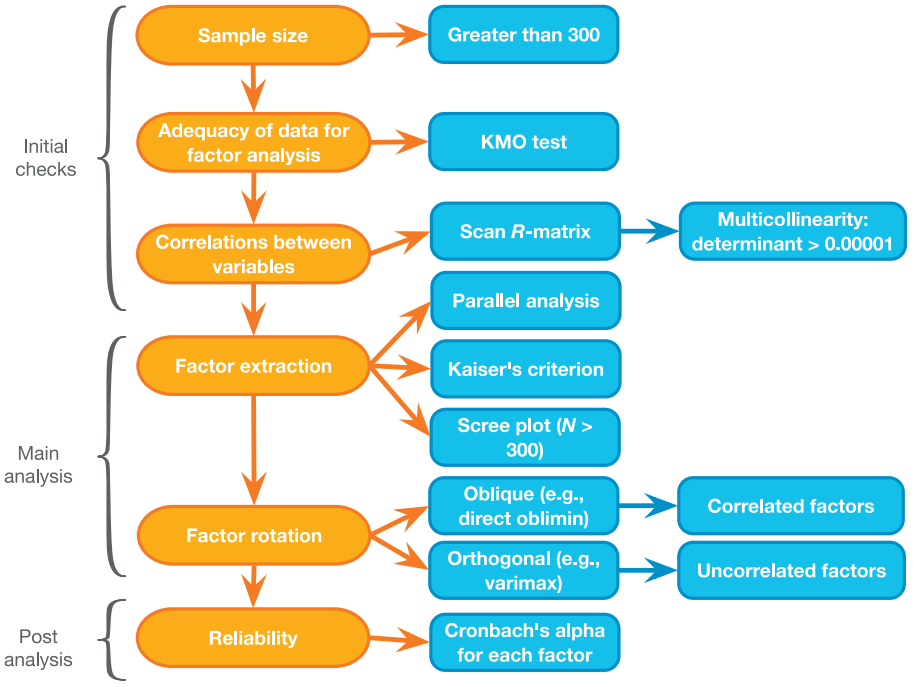


Figure 18.7 General procedure for factor analysis and PCA

- Marvellous: values in the 0.90s
- Meritorious: values in the 0.80s
- Middling: values in the 0.70s
- Mediocre: values in the 0.60s
- Miserable: values in the 0.50s
- Merde: values below 0.50. (They used the word 'unacceptable', but I don't like the fact that it doesn't start with the letter 'M' so I have changed it.)

To sum up, values smaller than 0.5 should lead you to collect more data or to rethink which variables to include.

18.5.3 Sample size Ⓑ

Correlation coefficients fluctuate from sample to sample, much more so in small samples than in large. Therefore, the reliability of factor analysis depends on sample size. Many 'rules of thumb' exist for the ratio of cases to variables; a common one is to have at least 10–15 participants per variable. Although I've heard this rule bandied about on numerous occasions, its empirical basis is unclear (although Nunnally, 1978 did recommend having 10 times as many participants as variables). Based on real data, Arrindell and van der Ende (1985) concluded that the cases-to-variables ratio made little difference to the stability of factor solutions.

What does matter is the overall sample size, factor loadings and communalities. Test parameters tend to be stable regardless of the cases-to-variables ratio (Kass & Tinsley, 1979), which is why Comrey and Lee (1992) class 100 as a poor sample size, 300 as good and 1000 as excellent. With respect to factor loadings, Guadagnoli and Velicer (1988) found that if a factor has four or more loadings greater than 0.6 then it is reliable regardless of sample size; factors with 10 or more loadings greater than 0.40 are reliable if the sample size is greater than 150; and factors with a few low loadings should not be interpreted unless the sample size is 300 or more. With respect to communalities, MacCallum et al.

(1999) showed that as communalities become lower the importance of sample size increases. With all communalities above 0.6, relatively small samples (less than 100) may be perfectly adequate. With communalities in the 0.5 range, samples between 100 and 200 can be good enough provided there are relatively few factors each with only a small number of indicator variables. In the worst scenario of low communalities (well below 0.5) and a larger number of underlying factors they recommend samples above 500.

What's clear from this work is that a sample of 300 or more will probably provide a stable factor solution, but that a wise researcher will measure enough variables to measure adequately all the factors that theoretically they would expect to find. With a sample of 2571 we shouldn't have cause to worry.

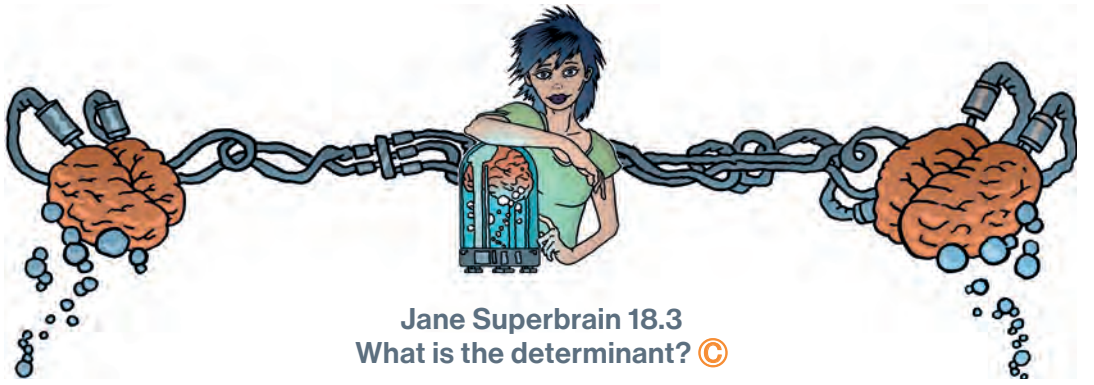
18.5.4 Correlations between variables ©

The 'garbage in, garbage out' adage applies particularly to factor analysis because a factor solution will usually be found to a set of variables, but will have little meaning if the variables put into the analysis are not sensible. A useful first step is to look at the correlations between variables. There are essentially two potential problems: (1) correlations that are not high enough; and (2) correlations that, are too high. In both cases the remedy is to remove variables from the analysis. The correlations between variables can be checked using the *correlate* procedure (see Chapter 8) to create a correlation matrix of all variables. This matrix can also be created as part of the factor analysis. We will look at each problem in turn.

If our test questions measure the same underlying dimension (or dimensions) then we would expect them to correlate with each other (because they are measuring the same thing). Even if questions measure different aspects of the same things (e.g., we could measure overall anxiety in terms of sub-components such as worry, intrusive thoughts and physiological arousal), there should still be moderate correlations between the variables relating to these sub-traits. We wouldn't expect to see variables that have small correlations with each other. We can visually scan the correlation matrix and look for correlations between about ±0.3 (you could use the *p*-values, but this approach isn't helpful because very small correlations will be significant in large samples and factor analysis typically employs large samples). This will narrow down the 'smallish' correlations in the matrix. Having done that, you're looking any variables have lots of correlations in this range, and in particular lots that are close to 0 (no relationship). Essentially, you're trying to identify variables that have very weak relations with the other variables. There's nothing magic about the value 0.3, it's a fairly arbitrary and inclusive heuristic to allow you to narrow your focus!

If you want an objective, but pointless, test of whether correlations (overall) are too small then you can test whether the correlation matrix resembles an identity matrix (see the Preface). This would mean that the off-diagonal components would be zero – in other words, the correlations between variables are all zero. This scenario is extreme. *Bartlett's test* (Section 17.5.3) tells us whether our correlation matrix is significantly different from an identity matrix. If it is significant then it means that the correlations between variables are (overall) significantly different from zero. The trouble is that because significance depends on sample size (see Section 2.10.11) and in factor analysis sample sizes are very large, Bartlett's test will nearly always be significant, even when the correlations between variables are very small indeed. As such, it's not a useful test (although in the unlikely event that it is non-significant you certainly have a big problem).

The opposite problem is when variables correlate too highly. Although mild multicollinearity is not a problem for factor analysis, it is important to avoid extreme multicollinearity (i.e., variables that are very highly correlated) and **singularity** (variables that are perfectly correlated). As with linear models, multicollinearity causes problems in factor analysis because it becomes impossible to determine the unique contribution to a factor of the variables that are highly correlated. Multicollinearity does not cause a problem for principal component analysis.



Jane Superbrain 18.3
What is the determinant? ©

The determinant of a matrix is a useful diagnostic tool in factor analysis, but the question of what it is is not easy to answer because it has a mathematical definition and not everyone is a mathematician. We can bypass the maths and think about the determinant conceptually. The way that I think of the determinant is as describing the 'area' of the data. In Jane Superbrain Box 9.3 we saw the diagrams reproduced in Figure 18.8. At the time I used these to describe eigenvectors and eigenvalues (which describe the shape of the data). The determinant is related to eigenvalues and eigenvectors, but instead of describing the height and width of the data it describes the overall area. So, in the left-hand diagram, the determinant of those data would represent the area inside the blue dashed circle. These variables have a low correlation so the determinant (area) is big; the maximum value is 1. In the right-hand diagram, the variables are perfectly correlated or singular, and the ellipse (blue dashed curve) has been squashed down to basically a straight line. In other words, the opposite sides of the ellipse will meet with no distance between them at all. Put another way, the area, or determinant, is 0. Therefore, the determinant tells us whether the correlation matrix is singular (determinant is 0), or if all variables are completely unrelated (determinant is 1), or somewhere in between.

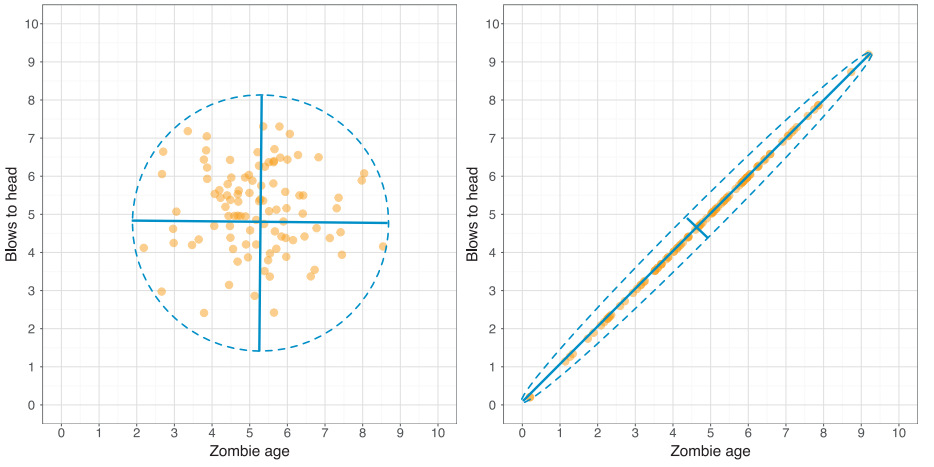


Figure 18.8 Data with a large (left) and small (right) determinant



Multicollinearity can be detected by looking at the determinant of the *R*-matrix, denoted |R| (see Jane Superbrain Box 18.3). One heuristic is that the determinant of the *R*-matrix should be greater than 0.00001, but again it's more about using this value in conjunction with a sensible scan of the correlation matrix. You could look through the correlation matrix for variables that correlate very highly ($r > 0.8$) and consider eliminating one of the variables (or more depending on the extent of the problem) before proceeding. The problem with a heuristic such as this is that the effect of two variables correlating with $r = 0.9$ might be less than the effect of, say, three variables that all correlate with $r = 0.6$. In other words, eliminating such highly correlating variables might not be getting at the cause of the multicollinearity (Rockwell, 1975). It may take trial and error to work out which variables are creating the problem.

18.5.5 The distribution of data Ⓑ

The assumption of normality is important if you wish to generalize the results of your analysis beyond the sample collected or do significance tests, but otherwise it's not. You can do factor analysis on non-continuous data; for example, if you had dichotomous variables, it's possible (using syntax) to do the factor analysis direct from the correlation matrix, but you should construct the correlation matrix from tetrachoric correlation coefficients. The only hassle is computing the correlations, but there's lots of help to be found at http://www.john-uebersax.com/stat/tetra.htm.

18.6 Factor analysis using SPSS Ⓑ

Access the main dialog box (Figure 18.9) by selecting *Analyze* > *Dimension Reduction* > Factor... Drag the variables you want to include in the analysis (or select them and click ▶) to the box labelled *Variables*. Remember to exclude any variables that were identified as problematic during the data screening.



Figure 18.9 Main dialog box for factor analysis

Click Descriptives... to access the dialog box in Figure 18.10. Checking ☑ Univariate descriptives produces means and standard deviations for each variable. Most of the other options relate to the correlation matrix of variables (the *R*-matrix described earlier): selecting ☑ Coefficients produces it, ☑ Significance levels includes the significance value of each correlation within it, and its ☑ Determinant is useful for testing for multicollinearity or singularity (see Section 18.5.4). Checking ☑ KMO and Bartlett's test of sphericity produces the Kaiser–Meyer–Olkin (see Section 18.5.2) measure of data adequacy and Bartlett's test (see Section 18.5.4). Checking ☑ Anti-image produces an anti-image matrix of covariances and correlations. This matrix contains the KMO statistic for each variable along the diagonal and the negatives of the partial correlations/covariances on the off-diagonals. The diagonal elements, like the KMO measure, should all be greater than 0.5 at a bare minimum if the sample is adequate for a given pair of variables. If any pair of variables has a value less than this, consider dropping one of them from the analysis. The off-diagonal elements should all be very small (close to zero) in a good model. Selecting ☑ Reproduced produces a correlation matrix based on the model (rather than the real data). Differences between the matrix based on the model and the matrix based on the observed data constitute the residuals of the model. These residuals appear in the lower table of the reproduced matrix and we want relatively few of these values to be greater than 0.05. To save us scanning this matrix, a summary of how many residuals lie above 0.05 is produced. Click Continue to return to the main dialog box.

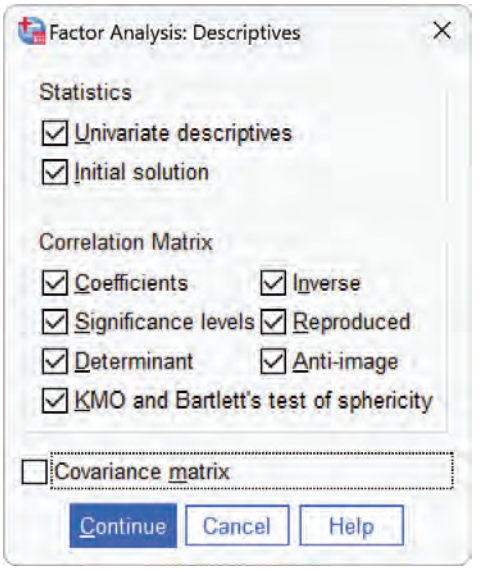
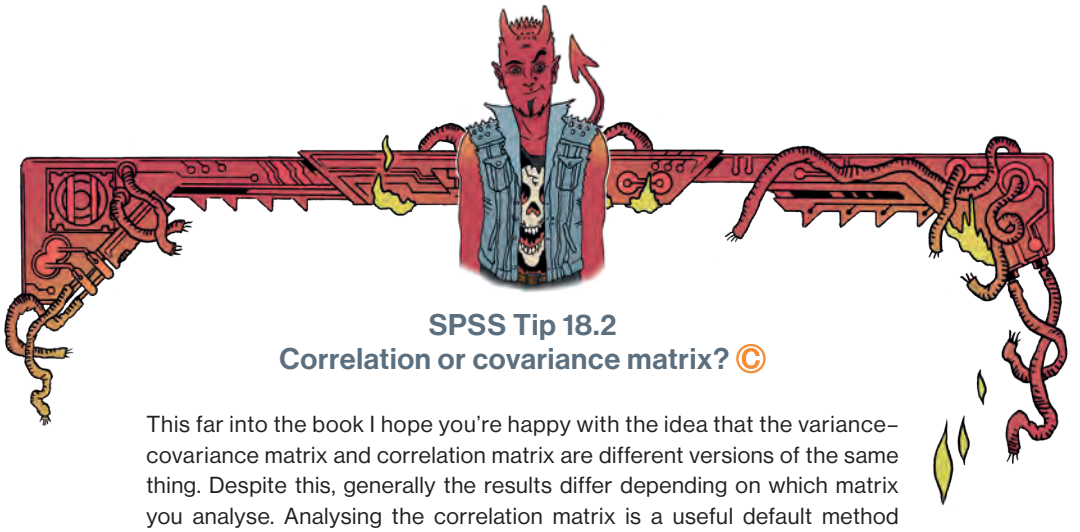


Figure 18.10 Descriptives in factor analysis

18.6.1 Factor extraction using SPSS Ⓑ

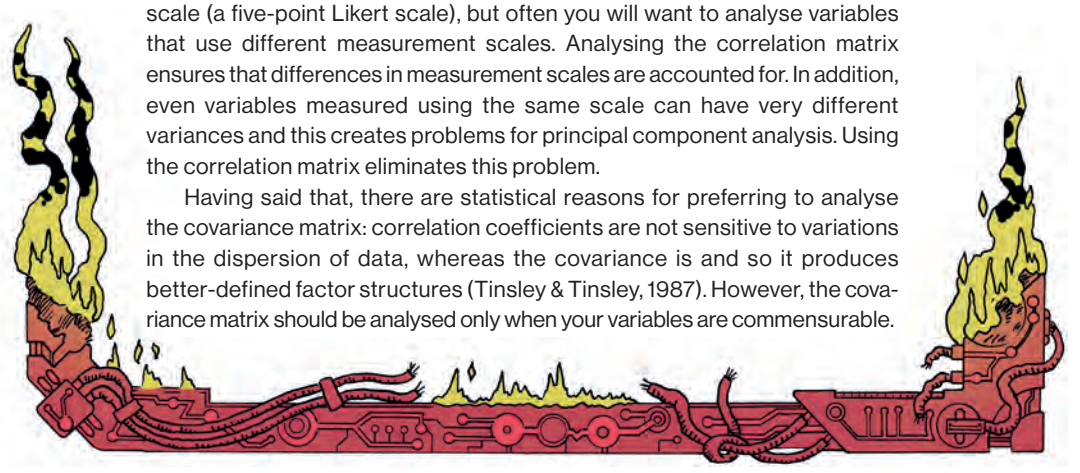
Click Extraction... to set the method of factor extraction (see Section 18.4.1). We will use *principal axis factoring* (Principal axis factoring) as in Figure 18.11. In the *Analyze* box we can choose between analysing the *Correlation matrix* and the *Covariance matrix* (SPSS Tip 18.2). We can choose to display the ☑ Unrotated factor solution and a ☑ Scree plot. The scree plot is a useful way to establish how many factors to retain, and the unrotated factor solution is useful in assessing how much rotation improves interpretation of the factor solution. If the rotated solution is little better than the unrotated one then it is possible that an inappropriate (or suboptimal) rotation method has been used.



SPSS Tip 18.2
Correlation or covariance matrix? ©

This far into the book I hope you're happy with the idea that the variance–covariance matrix and correlation matrix are different versions of the same thing. Despite this, generally the results differ depending on which matrix you analyse. Analysing the correlation matrix is a useful default method because it takes the standardized form of the matrix; therefore, if variables have been measured using different scales this will not affect the solution. In this example, all variables have been measured using the same measurement scale (a five-point Likert scale), but often you will want to analyse variables that use different measurement scales. Analysing the correlation matrix ensures that differences in measurement scales are accounted for. In addition, even variables measured using the same scale can have very different variances and this creates problems for principal component analysis. Using the correlation matrix eliminates this problem.

Having said that, there are statistical reasons for preferring to analyse the covariance matrix: correlation coefficients are not sensitive to variations in the dispersion of data, whereas the covariance is and so it produces better-defined factor structures (Tinsley & Tinsley, 1987). However, the covariance matrix should be analysed only when your variables are commensurable.



The *Extract* box provides options pertaining to the retention of factors. You have the choice of either extracting factors ◉ Based on Eigenvalue greater than a user-specified value (by default Kaiser's criterion of 1, but you can change this value) or retaining a ◉ Fixed number of factors. It is probably best to run a primary analysis ◉ Based on Eigenvalue greater than 1, select a scree plot and compare the results. If looking at the scree plot and the eigenvalues over 1 leads you to retain the same number of factors, happy days. Ideally run a parallel analysis too (SPSS Tip 18.1). If the various criteria give different results, examine the communalities and decide for yourself which of the two criteria to believe. If you decide to use the scree plot then you may need to redo the analysis specifying the number of factors to extract by selecting ◉ Fixed number of factors and typing the appropriate number in the space provided (e.g., 4). Click Continue.

18.6.2 Rotation Ⓑ

The interpretability of factors can be improved through rotation (Section 18.4.6), so, click Rotation... to set the rotation method (Figure 18.12). I discussed the rotation options earlier. In this chapter I'll show the output for both an orthogonal (◉ Varimax) and oblique (◉ Direct Oblimin) so we can compare

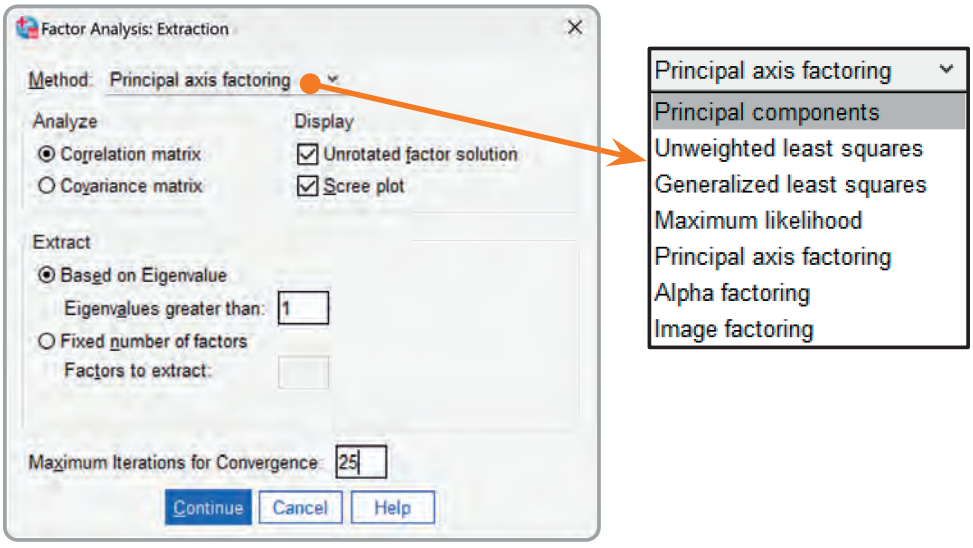


Figure 18.11 Dialog box for factor extraction

them. For now, chose one of the two (and maybe go back and rerun with the other so you can follow my interpretation).

The dialog box also has options for displaying the ☑ Rotated solution and a *Loading plot*. The rotated solution is displayed by default and is essential for interpreting the final rotated analysis. The loading plot will provide a graphical display of each variable plotted against the extracted factors up to a maximum of three factors (four- or five-dimensional plots are not yet possible). This plot is like Figure 18.3 and uses the factor loading of each variable for each factor. With two factors these plots are interpretable: you hope to see one group of variables clustered close to the *X*-axis and a different group clustered around the *Y*-axis. If all variables are clustered between the axes, then the rotation has been relatively unsuccessful in maximizing the loading of a variable onto a single factor. With three factors these plots will strain even the most dedicated visual system, so unless you have only two factors I would probably avoid them.

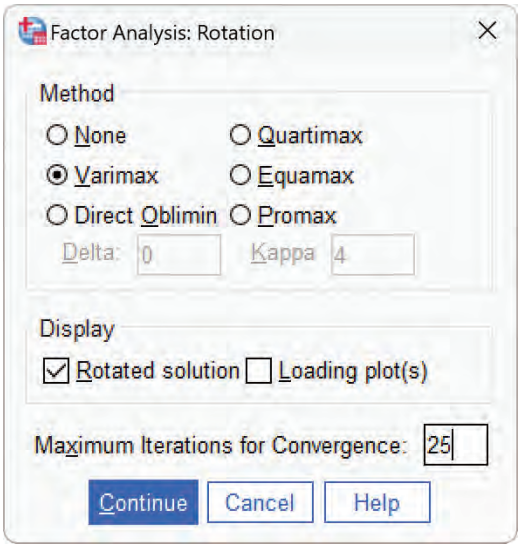


Figure 18.12 *Factor Analysis*: *Rotation* dialog box

A final option is to set the *Maximum Iterations for Convergence* (see SPSS Tip 20.1), which specifies the number of times that the computer will search for an optimal solution. In most circumstances the default of 25 is adequate; however, if you get an error message about convergence then increase this value. Click Continue.

18.6.3 Scores Ⓑ

Chose a method to compute factor scores (Figure 18.13) by clicking Scores... . This option will save factor scores (see Section 18.3.3) for each case in the data editor. A new column is created for each factor extracted and the factor score for each case is placed within that column. These scores can be used for further analysis, or to identify groups of participants who score high on particular factors. If you want to ensure that factor scores are uncorrelated then select ◉ Anderson-Rubin; if correlations between factor scores are acceptable then choose ◉ Regression. As a final option, you can produce the factor score coefficient matrix, which, realistically, we don't need to see. Click Continue.

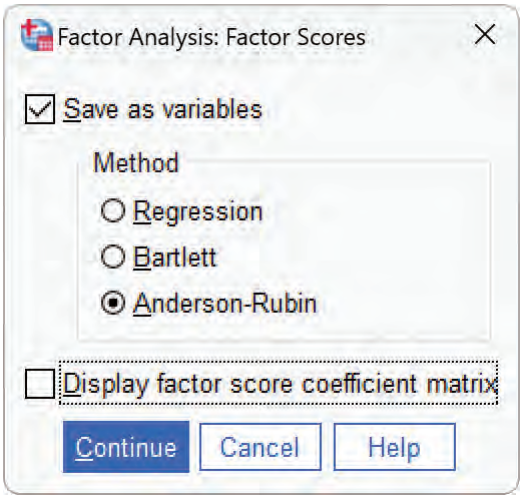


Figure 18.13 *Factor Analysis: Factor Scores* dialog box

18.6.4 Options Ⓑ

Click Options... to set the remaining options (Figure 18.14). Missing data are a problem for factor analysis just like most other procedures, and the options reflect the usual choices that were explained in SPSS Tip 6.1. If data are missing at random, and there's not too much missing data, there really is no excuse (other than that I don't cover it in this book) not to use methods more sophisticated than replacing values with the mean (Enders, 2022). If you exclude cases pairwise your estimates can go all over the place, so definitely don't do that. The final two options relate to how coefficients are displayed. By default, variables are listed in the order in which they are entered into the data editor. If you select ☑ Sorted by size then variables will be ordered by their factor loadings, which can be useful for interpreting factors. The sorting is done intelligently inasmuch as variables that load highly on the same factor are displayed together. The second option is to ☑ Suppress small coefficients (by default absolute values below 0.1 are suppressed). This option hides factor loadings within ±0.1 in the output. Again, this option is useful for interpretation. The default value is probably sensible, but on the first run-through I recommend changing it to 0.3 to make interpretation simpler. We know that a loading of 0.4 is substantial, but so I don't throw out the baby with the bathwater I tend to set the value at 0.3 as in Figure 18.14. Click Continue.

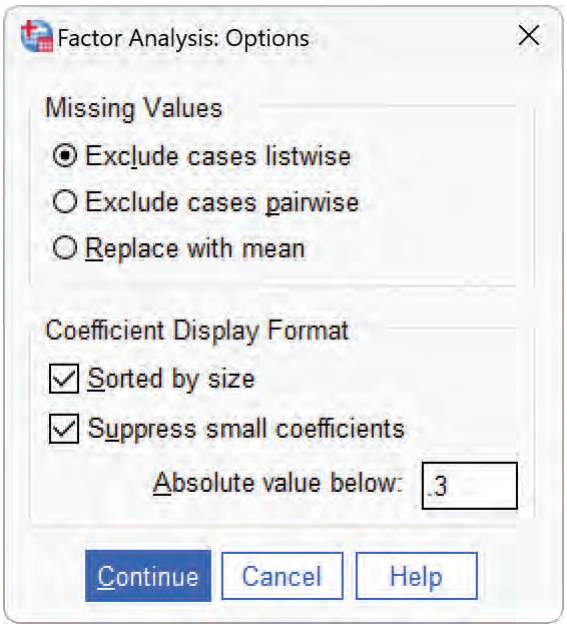
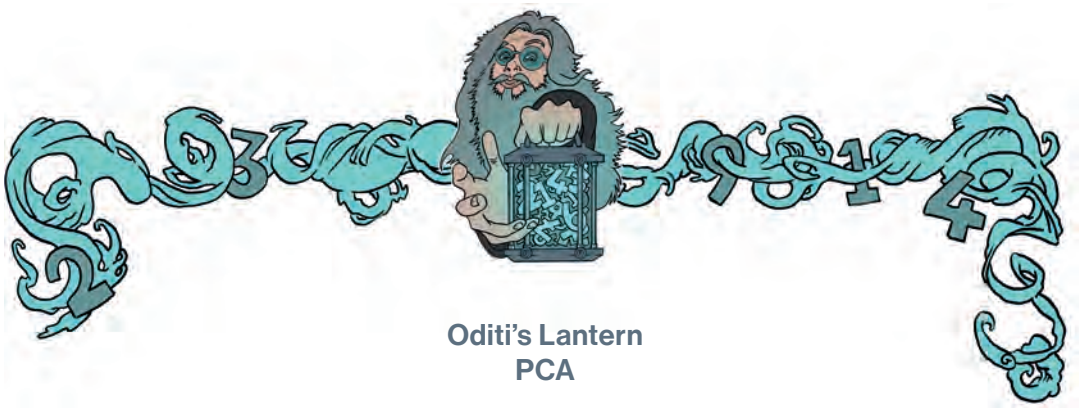


Figure 18.14 *Factor Analysis*: *Options* dialog box



Oditi's Lantern
PCA

'I, Oditi, feel that we are getting closer to finding the hidden truths behind the numbers. Factor analysis allows us to estimate variables "hidden" within the data. This technique is the very essence of the cult of undiscovered numerical truths. Once we have mastered this tool we can find out what people are really thinking, even if they don't know they're thinking it. We might find that they think I'm the kind saviour of cute furry gerbils, but that underneath they know the truth ... stare into my lantern to discover factor analysis.'



18.7 Interpreting factor analysis Ⓑ

Select the same options as I have in the screen shots and run a factor analysis with orthogonal rotation.



Having done this, select the ◉ Direct Oblimin option in Figure 18.12 and repeat the analysis. You should obtain two outputs identical in all respects except that one used an orthogonal rotation and the other an oblique.

Everything should go smoothly with the example data, but with your own data you might be unlucky enough to get an error message about a 'non-positive definite matrix' (see SPSS Tip 18.3). A 'non-positive definite matrix' sounds a bit like a collection of depressed numbers that lack certainty about their lives. In some respects it is.

To make the larger outputs vaguely legible, I have switched my SPSS preferences so that each variable is referred to by its name in the data editor (e.g., **question_12**) rather than the variable label (the question itself). Your output will show the variable label, so when using my output refer to Figure 18.6 if you need to know the question content.

18.7.1 Preliminary analysis Ⓑ

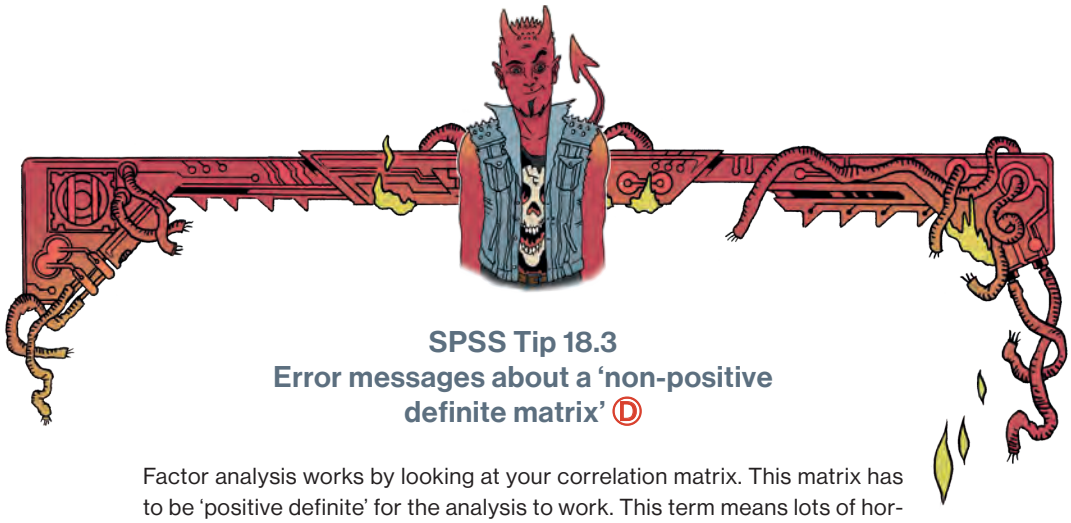
The first body of output concerns data screening, assumption testing and sampling adequacy. You'll find several large tables (or matrices) that tell us interesting things about our data. If you selected ☑ Univariate descriptives (Figure 18.10) then you'll get a table of descriptive statistics for each variable (the mean, standard deviation and number of cases). I haven't reproduced it here because I'm confident that you can interpret it by now. The table is a useful way to determine the extent of missing data.

Output 18.1 shows the *R*-matrix (i.e., the correlation matrix):[8] the top half contains the Pearson correlation coefficient between all pairs of questions and the bottom half contains the one-tailed *p*-values of these coefficients. Note correlations less than 0.3 and greater than 0.9. Variables with very few correlations above 0.3 might not 'fit' with the pool of items, and variables with correlations greater than 0.9 might be collinear. You can also check the determinant of the correlation matrix and, if necessary, eliminate variables that you think are causing the problem. The determinant is listed at the bottom of the matrix (blink and you'll miss it). For these data its value is 0.004, which is enough above zero to suggest no multicollinearity problems (see Section 18.5.4). More important, however, a scan of Output 18.1 reveals that the largest correlations are below any value that might give us pause for thought. All questions in the SAQ correlate reasonably well with all others and none of the correlation coefficients are excessively large; therefore, we won't eliminate any questions.

Output 18.2 shows the Kaiser–Meyer–Olkin measure of sampling adequacy and Bartlett's test of sphericity. The KMO statistic is 0.92, which is well above the minimum criterion of 0.5 and falls into the range of 'marvellous' (see Section 18.5.2), so we might take comfort that the data are probably adequate for factor analysis. The KMO values for individual variables are produced on the diagonal of the anti-image correlation matrix in Output 18.3 (I have highlighted these cells).[9] We

8 To save space only columns for the first and last five questions in the questionnaire are included here.

9 In your output the anti-image correlation appears with the covariance matrices. These matrices provide similar information (remember the relationship between covariance and correlation) but the anti-image correlation matrix is the most informative.

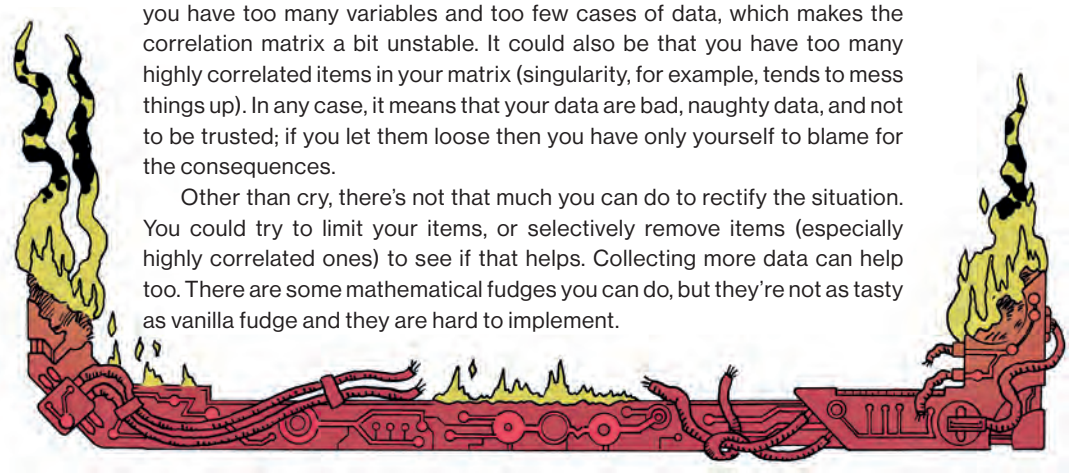


SPSS Tip 18.3
Error messages about a 'non-positive definite matrix' Ⓓ

Factor analysis works by looking at your correlation matrix. This matrix has to be 'positive definite' for the analysis to work. This term means lots of horrible things mathematically (e.g., the eigenvalues and determinant of the matrix are positive), but in more basic terms, factors are like lines floating in space, and eigenvalues measure the length of those lines. If your eigenvalue is negative then it means that the length of your line/factor is negative too. It's a bit like me asking you how tall you are, and you responding 'I'm minus 175 cm tall'. That would be nonsense. If a factor has negative length, that's nonsense too. When SPSS decomposes the correlation matrix to look for factors, if it comes across a negative eigenvalue it starts thinking 'oh dear, I've entered some weird parallel universe where the usual rules of maths no longer apply and things can have negative lengths, and this probably means that time runs backwards, my mum is my dad, my sister is a dog, my head is a fish and my toe is a frog called Gerald'. It does the sensible thing and decides not to proceed. Things like the KMO test and the determinant rely on a positive definite matrix too: they can't be computed without one.

The most likely reason for having a non-positive definite *R*-matrix is that you have too many variables and too few cases of data, which makes the correlation matrix a bit unstable. It could also be that you have too many highly correlated items in your matrix (singularity, for example, tends to mess things up). In any case, it means that your data are bad, naughty data, and not to be trusted; if you let them loose then you have only yourself to blame for the consequences.

Other than cry, there's not that much you can do to rectify the situation. You could try to limit your items, or selectively remove items (especially highly correlated ones) to see if that helps. Collecting more data can help too. There are some mathematical fudges you can do, but they're not as tasty as vanilla fudge and they are hard to implement.



should check that the diagonal elements of the anti-image matrix are above the bare minimum of 0.5 (and preferably higher). For these data all values are well above 0.5, which is good news. If you find variables with values below 0.5 then consider excluding them from the analysis (or run the analysis with and without that variable and note the difference). Removal of a variable affects the KMO statistics, so if you remove a variable be sure to re-examine the new anti-image correlation

Correlation Matrix[a]

		question_01	question_02	question_03	question_04	question_05	...	question_19	question_20	question_21	question_22	question_23
Correlation	question_01	1.000	.107	-.179	.217	.223	...	.107	.207	.227	.123	.077
	question_02	.107	1.000	-.120	.124	.279	...	.390	.128	.161	.339	.391
	question_03	-.179	-.120	1.000	-.194	-.255	...	-.117	-.234	-.260	-.118	-.052
	question_04	.217	.124	-.194	1.000	.319	...	.141	.284	.320	.157	.083
	question_05	.223	.279	-.255	.319	1.000	...	.292	.341	.371	.271	.169
	question_06	.159	.338	-.166	.234	.513	...	.377	.226	.264	.326	.172
	question_07	.106	.256	-.126	.166	.364	...	.256	.140	.197	.225	.160
	question_08	.222	.178	-.218	.248	.357	...	.210	.262	.304	.225	.142
	question_09	.084	.391	-.081	.098	.161	...	.465	.100	.167	.427	.547
	question_10	.112	.184	-.102	.177	.312	...	.211	.179	.162	.156	.070
	question_11	.189	.152	-.190	.222	.311	...	.194	.270	.286	.202	.128
	question_12	.150	.090	-.148	.196	.187	...	.067	.123	.164	.080	.053
	question_13	.109	.230	-.131	.164	.351	...	.260	.162	.174	.230	.140
	question_14	.115	.198	-.088	.147	.267	...	.221	.148	.167	.220	.127
	question_15	.110	.187	-.148	.184	.327	...	.230	.158	.165	.237	.150
	question_16	.176	.097	-.204	.271	.240	...	.136	.232	.264	.109	.100
	question_17	.166	.174	-.175	.237	.301	...	.197	.219	.258	.207	.137
	question_18	.119	.264	-.152	.171	.382	...	.300	.179	.193	.268	.130
	question_19	.107	.390	-.117	.141	.292	...	1.000	.164	.184	.418	.440
	question_20	.207	.128	-.234	.284	.341	...	.164	1.000	.349	.168	.114
	question_21	.227	.161	-.260	.320	.371	...	.184	.349	1.000	.178	.142
	question_22	.123	.339	-.118	.157	.271	...	.418	.168	.178	1.000	.401
	question_23	.077	.391	-.052	.083	.169	...	.440	.114	.142	.401	1.000
Sig. (1-tailed)	question_01		<.001	<.001	<.001	<.001	...	<.001	<.001	<.001	<.001	<.001
	question_02	.000		.000	.000	.000	...	.000	.000	.000	.000	.000
	question_03	.000	.000		.000	.000	...	.000	.000	.000	.000	.004
	question_04	.000	.000	.000		.000	...	.000	.000	.000	.000	.000
	question_05	.000	.000	.000	.000		...	.000	.000	.000	.000	.000
	question_06	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_07	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_08	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_09	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_10	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_11	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_12	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.003
	question_13	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_14	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_15	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_16	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_17	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_18	.000	.000	.000	.000	.000	...	.000	.000	.000	.000	.000
	question_19	.000	.000	.000	.000	.000	...		.000	.000	.000	.000
	question_20	.000	.000	.000	.000	.000	...	.000		.000	.000	.000
	question_21	.000	.000	.000	.000	.000	...	.000	.000		.000	.000
	question_22	.000	.000	.000	.000	.000	...	.000	.000	.000		.000
	question_23	.000	.000	.004	.000	.000	...	.000	.000	.000	.000	

a. Determinant = .004

Output 18.1

matrix. As for the rest of the anti-image correlation matrix, the off-diagonal elements represent the partial correlations between variables. We want these correlations to be very small (the smaller, the better). So, as a final check you can look through to see that the off-diagonal elements are small (they should be for these data).

Bartlett's measure (Output 18.2) tests the null hypothesis that the original correlation matrix is an identity matrix. We want this test to be *significant* (see Section 18.5.4). Given the large sample sizes usually used in factor analysis, this test will almost certainly be significant, and it is ($p < 0.001$). In short, we can't read a lot into this test: a non-significant test would certainly indicate a massive problem, but this significant value only really tells us that we don't have a massive problem. We could have non-massive problems though. On the plus side, you've learnt how to conduct a largely pointless test.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.916
Bartlett's Test of Sphericity	Approx. Chi-Square	14431.741
	df	253
	Sig.	<.001

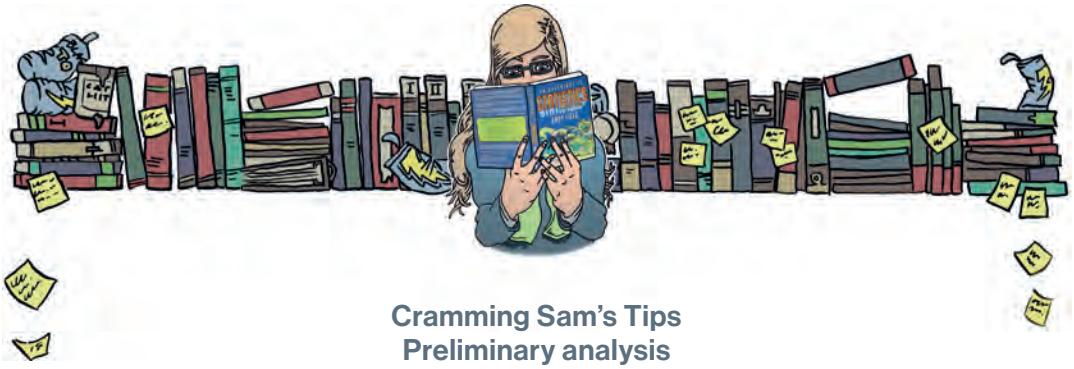
Output 18.2

Anti-image Matrices

Anti-image Correlation

	question_01	question_02	question_03	question_04	question_05	...	question_19	question_20	question_21	question_22	question_23
question_01	.941	-.014	.064	-.077	-.044	...	.005	-.062	-.067	-.019	-.002
question_02	-.014	.935	.032	.007	-.055	...	-.102	.010	.006	-.063	-.169
question_03	.064	.032	.928	.028	.070	...	.008	.086	.100	.009	-.042
question_04	-.077	.007	.028	.925	-.086	...	.007	-.101	-.126	-.024	.019
question_05	-.044	-.055	.070	-.086	.944	...	-.029	-.123	-.128	-.025	-.012
question_06	.000	-.091	-.022	-.012	-.188	...	-.110	.008	-.023	-.060	.057
question_07	.012	-.055	.002	-.003	-.088	...	-.014	.037	-.023	-.005	-.023
question_08	-.061	.011	.044	-.013	-.052	...	.008	-.019	-.042	-.018	-.005
question_09	.000	-.146	.007	-.002	.045	...	-.212	.029	-.057	-.185	-.346
question_10	-.013	-.023	-.016	-.037	-.063	...	-.039	-.049	.007	.021	.049
question_11	-.025	.023	.018	-.011	-.020	...	-.001	-.080	-.054	-.007	.016
question_12	-.070	-.025	.063	-.101	-.068	...	.017	-.002	-.040	-.005	-.001
question_13	.001	-.025	.015	-.006	-.078	...	-.033	-.004	.009	-.023	.002
question_14	-.022	-.032	-.022	-.016	-.018	...	-.035	-.016	-.023	-.058	-.005
question_15	2.140E-5	.021	.042	-.043	-.066	...	-.009	-.002	.017	-.052	-.032
question_16	-.054	.013	.084	-.132	-.044	...	-.031	-.073	-.096	.011	-.034
question_17	-.005	-.024	.014	-.063	-.026	...	-.015	-.007	-.034	-.029	-.016
question_18	.009	-.052	.021	.012	-.068	...	-.072	-.006	.008	-.057	.036
question_19	.005	-.102	.008	.007	-.029	...	.925	-.017	.004	-.131	-.180
question_20	-.062	.010	.086	-.101	-.123	...	-.017	.923	-.163	-.029	-.025
question_21	-.067	.006	.100	-.126	-.128	...	.004	-.163	.927	.008	-.016
question_22	-.019	-.063	.009	-.024	-.025	...	-.131	-.029	.008	.936	-.146
question_23	-.002	-.169	-.042	.019	-.012	...	-.180	-.025	-.016	-.146	.835

Output 18.3



Cramming Sam's Tips
Preliminary analysis

- Scan the correlation matrix for variables that have very small correlations with most other variables, or correlate very highly ($r = 0.9$) with one or more other variables.
- In factor analysis, check that the determinant of this matrix is bigger than 0.00001; if it is then multicollinearity isn't a problem. You don't need to worry about this for principal component analysis.
- In the table labelled *KMO and Bartlett's Test* the KMO statistic should be greater than 0.5 as a bare minimum; if it isn't, collect more data or rethink the variables you're including. You should check the KMO statistic for individual variables by looking at the diagonal of the anti-image matrix. These values should also be above 0.5 (this is useful for identifying problematic variables if the overall KMO is unsatisfactory).
- Bartlett's test of sphericity will usually be significant (the value of *Sig.* will be less than 0.05); if it's not, you've got a disaster on your hands.



18.7.2 Factor extraction Ⓑ

The first part of the factor extraction process is to determine the linear components within the variables – the eigenvectors (see Section 18.4.4). There are as many components (eigenvectors) in the *R*-matrix as there are variables, but most will be unimportant. To determine which vectors to retain we apply criteria based on the magnitude of the associated eigenvalues. By default, Kaiser's criterion of retaining factors with eigenvalues greater than 1 is used (see Figure 18.11).

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.858	25.470	25.470	5.271	22.918	22.918	2.788	12.121	12.121
2	1.988	8.642	34.113	1.434	6.234	29.152	2.104	9.149	21.270
3	1.562	6.790	40.902	.975	4.237	33.389	1.954	8.494	29.764
4	1.249	5.429	46.331	.664	2.887	36.276	1.498	6.512	36.276
5	.904	3.931	50.263						
6	.851	3.700	53.963						
7	.840	3.652	57.615						
8	.795	3.456	61.071						
9	.787	3.421	64.492						
10	.764	3.321	67.812						
11	.731	3.180	70.992						
12	.704	3.059	74.052						
13	.676	2.941	76.993						
14	.655	2.847	79.840						
15	.639	2.780	82.619						
16	.625	2.717	85.337						
17	.574	2.494	87.831						
18	.542	2.357	90.187						
19	.528	2.295	92.483						
20	.508	2.210	94.692						
21	.435	1.893	96.586						
22	.402	1.749	98.335						
23	.383	1.665	100.000						

Extraction Method: Principal Axis Factoring.

Output 18.4

Output 18.4 lists the eigenvalues associated with each factor before extraction, after extraction and after rotation. Before extraction, 23 factors are identified (there should be as many eigenvectors as there are variables and so there will be as many factors as variables – see Section 18.4.4). The eigenvalues associated with each factor represent the variance explained by that particular factor and can be translated into the percentage of variance explained (e.g., factor 1 explains 25.47% of total variance). The first few factors explain relatively large amounts of variance (especially factor 1) whereas subsequent factors explain only small amounts. All factors with eigenvalues greater than 1 are then extracted, leaving us with four factors. The eigenvalues associated with these factors (and the percentage of variance explained) are displayed under the heading of *Extraction Sums of Squared Loadings*. In the part of the table labelled *Rotation Sums of Squared Loadings*, the eigenvalues of the factors after rotation are displayed. Rotation has the effect of optimizing the factor structure, and one consequence for these data is that the relative importance of the four factors is equalized a bit. Before rotation, factor 1 accounted for considerably more variance than the remaining three (22.92% compared to 6.23%, 4.24% and 2.89%), but after rotation it accounts for only 12.12% of variance (compared to 9.15%, 8.49% and 6.51%).

Output 18.5 (left) shows the table of communalities before and after extraction. Remember that the communality is the proportion of common variance within a variable (see Section 18.4.1). Factor analysis starts by estimating the variance that is common; therefore, before extraction the communalities are a kind of best guess. Once factors have been extracted, we can better estimate how much variance is common. The communalities in the column labelled *Extraction* reflect this common variance. So, for example, we can say that 15.9% of the variance associated with question 1 is common, or shared, variance. Another way to look at these communalities is in terms of the proportion of variance explained by the underlying factors. Remember that after extraction we have discarded some factors (in this case we've retained only four), so the communalities after extraction represent the amount of variance in each variable that can be explained by the retained factors.

Communalities

	Initial	Extraction
question_01	.121	.159
question_02	.293	.339
question_03	.142	.181
question_04	.219	.294
question_05	.422	.493
question_06	.519	.665
question_07	.271	.318
question_08	.474	.682
question_09	.407	.566
question_10	.204	.236
question_11	.394	.498
question_12	.079	.102
question_13	.265	.310
question_14	.179	.200
question_15	.225	.256
question_16	.153	.211
question_17	.356	.439
question_18	.329	.387
question_19	.379	.453
question_20	.227	.287
question_21	.274	.361
question_22	.318	.372
question_23	.386	.534

Extraction Method: Principal Axis Factoring.

Factor Matrix[a]

	Factor			
	1	2	3	4
question_06	.708		-.401	
question_05	.659			
question_08	.601	-.314		-.378
question_18	.550			
question_19	.544	.394		
question_11	.530			
question_17	.509			
question_22	.507	.328		
question_07	.505			
question_13	.488			
question_21	.480			
question_02	.477	.334		
question_15	.465			
question_20	.423			
question_10	.420			
question_04	.416			
question_14	.413			
question_16	.329			
question_03	-.326			
question_01	.307			
question_12				
question_09	.436	.547		
question_23	.408	.543		

Extraction Method: Principal Axis Factoring.

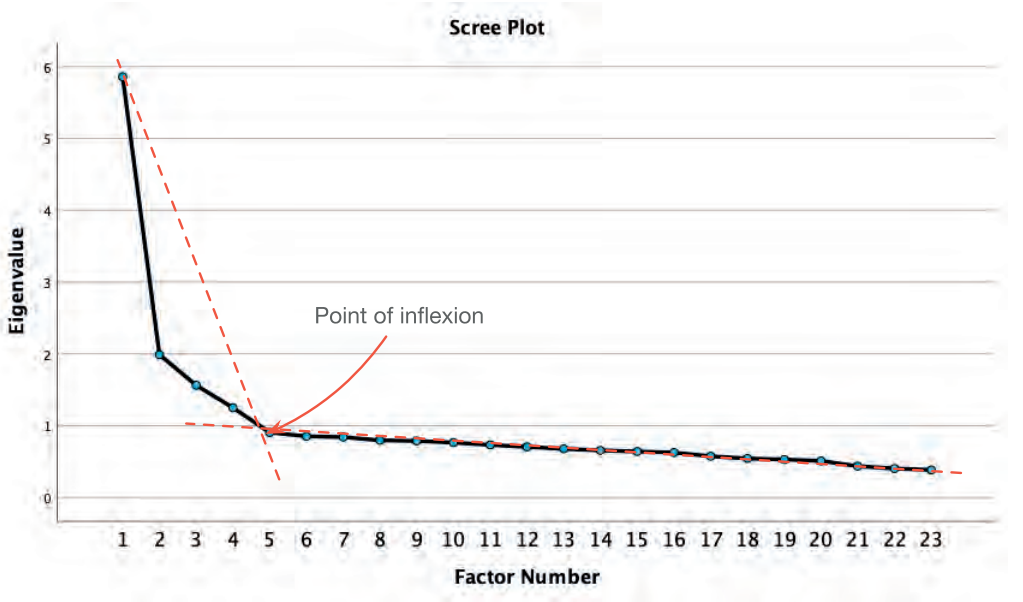
a. 4 factors extracted. 13 iterations required.

Output 18.5

Output 18.5 (right) shows the factor matrix before rotation. This matrix contains the loadings of each variable onto each factor. Because we requested that loadings less than 0.3 be suppressed (see Figure 18.14) there are blank spaces for many of the loadings. This matrix is not particularly important for interpretation, but it is interesting to note that before rotation most variables load highly onto the first factor (that is why this factor accounts for most of the variance in Output 18.4).

Because you should never let a computer make important decisions for you, we need to think about the number of factors to extract (Section 18.4.5). By Kaiser's criterion we should extract four factors (which is what has been done). This criterion is accurate when there are less than 30 variables and communalities after extraction are greater than 0.7 or when the sample size exceeds 250 and the average communality is greater than 0.6. For these data, no communalities exceed 0.7 (Output 18.5), and the average communality is quite low: adding up the communalities and dividing by how many there are gives us 6.636/23 = 0.288. Both criteria suggest Kaiser's rule might be inappropriate for these data. Using Jolliffe's criterion (retain factors with eigenvalues greater than 0.7) we'd end up with 12 factors (see Output 18.4) most of which equate to small portions of variance, so I think that would be a silly idea. The scree plot (Output 18.6) seems fairly clear in that it levels off at the fifth factor suggesting that we should retain four.

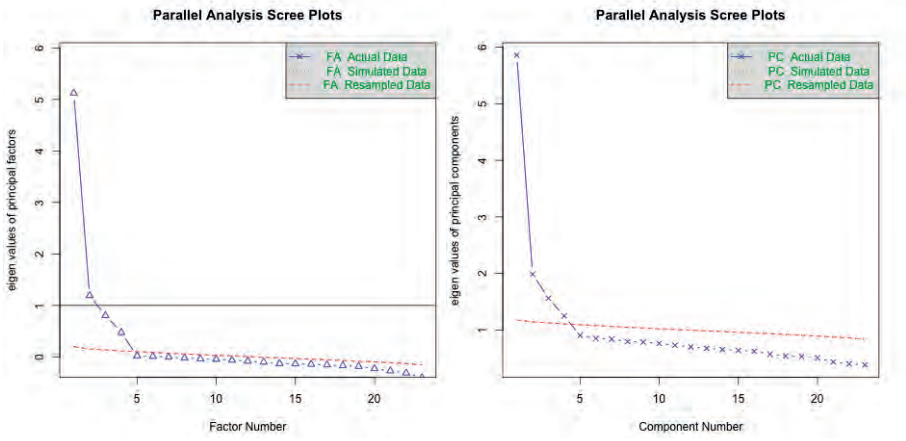
For those of you who had a go at parallel analysis (SPSS Tip 18.1) Output 18.7 shows the plots based on factor analysis eigenvalues (left) and eigenvalues of the principal components (right). Both plots are interpreted in the same way. Each shows a scree plot in which each factor/component is plotted on the *x*-axis with its corresponding eigenvalue is on the *y*-axis. The observed eigenvalues for each factor are shown as blue triangles (crosses for the components) connected by a blue line to form the characteristic 'scree' shape. The red dotted line shows the corresponding line from the randomly



Output 18.6

generated data sets. Given we want to keep factors that have eigenvalues larger than we'd expect in data that have no underlying factors, we want to keep factors that have eigenvalues (triangles or crosses) that fall above the red line. In this case, in the left-hand plot, the first four factors have eigenvalues larger than their counterparts from random data (the first four triangles fall above the red line). Similarly, in the right-hand plot, the first four components have eigenvalues larger than their counterparts from random data (the first four crosses fall above the red line). Both methods agree that there are four underlying factors/components. Note that your output also contains a handy text message confirming this fact for those who can't interpret the plot.

We are in the fortunate position that Kaiser's criterion, the scree plot and parallel analysis all agree on four factors. When the methods disagree, I'd be inclined to follow parallel analysis and use the detailed guidelines earlier in the chapter.



Output 18.7

Output 18.8 shows an edited version of the reproduced correlation matrix. The top half of this matrix (labelled *Reproduced Correlations*) contains the correlation coefficients between the questions based on the factor model. The diagonal of this matrix contains the communalities after extraction for each variable (you can check the values against Output 18.5). The correlations in the reproduced matrix differ from those in the *R*-matrix because they stem from the model rather than the observed data. If the model were a perfect fit to the data then we would expect the reproduced correlation coefficients to be the same as the original correlation coefficients. Therefore, to assess the fit of the model we can look at the differences between the observed correlations and the correlations based on the model. For example, if we take the correlation between questions 1 and 2, the correlation based on the observed data is 0.107 (see the blue cells in Output 18.1), and based on the model is 0.099 (highlighted in blue in Output 18.8). The difference is:

$$\begin{aligned} \text{residual} &= r_{\text{observed}} - r_{\text{from model}} \\ \text{residual}_{Q_1Q_2} &= 0.107 - 0.099 \\ &= 0.008 \end{aligned} \tag{18.8}$$

This value (0.008) is shown in the lower half of the reproduced matrix (labelled *Residual*) for questions 1 and 2 (again highlighted in blue). More generally, the lower half of the reproduced matrix contains the differences between the observed correlation coefficients and the ones predicted from the model for all pairs of variables. For a good model these values will all be small: ideally, we want most values to be less than 0.05. The footnote to the matrix helpfully states how many residuals have an absolute value greater than 0.05. For these data there are 0 residuals (0%) that are greater than 0.05. There are no hard-and-fast rules about what proportion of residuals should be below 0.05; however, if more than 50% are greater than 0.05 you probably have grounds for concern; 0% (as we have here) is ideal.

Reproduced Correlations

		question_01	question_02	question_03	question_04	question_05	...	question_19	question_20	question_21	question_22	question_23
Reproduced Correlation	question_01	.159	.099	-.169	.214	.232	...	.112	.213	.239	.114	.081
	question_02	.099	.339	-.102	.132	.267	...	.392	.139	.169	.352	.379
	question_03	-.169	-.102	.181	-.230	-.254	...	-.114	-.228	-.254	-.116	-.073
	question_04	.214	.132	-.230	.294	.329	...	.145	.290	.323	.146	.088
	question_05	.232	.267	-.254	.329	.493	...	.296	.324	.359	.275	.160
	question_06	.156	.333	-.176	.236	.507	...	.371	.230	.253	.328	.180
	question_07	.114	.244	-.127	.169	.354	...	.273	.166	.184	.243	.148
	question_08	.211	.176	-.214	.255	.354	...	.212	.276	.311	.227	.138
	question_09	.086	.393	-.078	.094	.173	...	.463	.113	.151	.423	.549
	question_10	.110	.182	-.124	.165	.317	...	.200	.160	.174	.177	.081
	question_11	.191	.163	-.193	.232	.314	...	.195	.249	.282	.207	.135
	question_12	.123	.073	-.133	.170	.179	...	.080	.167	.186	.079	.056
	question_13	.112	.237	-.126	.169	.351	...	.264	.165	.182	.233	.138
	question_14	.097	.197	-.107	.141	.285	...	.222	.140	.156	.199	.127
	question_15	.113	.218	-.125	.165	.327	...	.245	.163	.181	.219	.134
	question_16	.181	.107	-.194	.247	.259	...	.119	.244	.273	.121	.087
	question_17	.175	.163	-.178	.213	.301	...	.194	.229	.259	.203	.133
	question_18	.122	.256	-.137	.181	.387	...	.286	.179	.197	.255	.142
	question_19	.112	.392	-.114	.145	.296	...	.453	.156	.190	.408	.446
	question_20	.213	.139	-.228	.290	.324	...	.156	.287	.321	.157	.106
	question_21	.239	.169	-.254	.323	.359	...	.190	.321	.361	.190	.143
	question_22	.114	.352	-.116	.146	.275	...	.408	.157	.190	.372	.406
	question_23	.081	.379	-.073	.088	.160	...	.446	.106	.143	.406	.534
Residual[b]	question_01		.008	-.010	.003	-.009	...	-.005	-.006	-.012	.009	-.003
	question_02	.008		-.018	-.008	.012	...	-.001	-.011	-.008	-.013	.013
	question_03	-.010	-.018		.036	-.001	...	-.003	-.006	-.005	-.002	.021
	question_04	.003	-.008	.036		-.010	...	-.004	-.006	-.003	.011	-.005
	question_05	-.009	.012	-.001	-.010		...	-.004	.017	.012	-.004	.009
	question_06	.003	.004	.011	-.001	.006	...	.006	-.005	.011	-.002	-.008
	question_07	-.008	.013	.001	-.002	.010	...	-.017	-.026	.013	-.018	.012
	question_08	.010	.002	-.004	-.008	.004	...	-.002	-.014	-.007	-.003	.003
	question_09	-.003	-.002	-.003	.004	-.011	...	.002	-.013	.016	.005	-.002
	question_10	.002	.002	.022	.012	-.004	...	.011	.019	-.012	-.022	-.012
	question_11	-.002	-.010	.004	-.010	-.003	...	-.001	.021	.005	-.005	-.007
	question_12	.027	.017	-.015	.025	.008	...	-.013	-.044	-.022	.000	-.002
	question_13	-.003	-.007	-.005	-.005	.000	...	-.004	-.003	-.007	-.003	.002
	question_14	.018	.000	.020	.006	-.018	...	-.001	.008	.011	.021	.000
	question_15	-.003	-.032	-.023	.019	.001	...	-.014	-.005	-.015	.018	.016
	question_16	-.006	-.010	-.010	.024	-.019	...	.017	-.012	-.009	-.011	.013
	question_17	-.009	.011	.003	.024	-9.220E-5	...	.003	-.010	-.001	.004	.004
	question_18	-.003	.008	-.015	-.010	-.005	...	.014	.001	-.004	.013	-.012
	question_19	-.005	-.001	-.003	-.004	-.004	...		.008	-.006	.010	-.006
	question_20	-.006	-.011	-.006	-.006	.017	...	.008		.027	.011	.008
	question_21	-.012	-.008	-.005	-.003	.012	...	-.006	.027		-.012	.000
	question_22	.009	-.013	-.002	.011	-.004	...	.010	.011	-.012		-.005
	question_23	-.003	.013	.021	-.005	.009	...	-.006	.008	.000	-.005	

Extraction Method: Principal Axis Factoring.

b. Residuals are computed between observed and reproduced correlations. There are 0 (0.0%) nonredundant residuals with absolute values greater than 0.05.

Output 18.8



Cramming Sam's Tips
Factor extraction

- To decide how many factors to extract, ideally use parallel analysis. Retain factors/components that have eigenvalues greater than the corresponding eigenvalue from simulated data without a factor structure. To identify these, use the scree plot. Starting at the left-hand side, count the number of triangles (factors) or stars (components) that are above the red dotted line. That's the number of factors you extract. (Text output will also tell you how many.)
- You can also look at the table labelled *Communalities* and the column labelled *Extraction*. If these values are all 0.7 or above and you have less than 30 variables then the SPSS default (Kaiser's criterion) for extracting factors is fine. Likewise, if your sample size exceeds 250 and the average of the communalities is 0.6 or greater. With 200 or more participants the scree plot can be used.
- Check the bottom of the table labelled *Reproduced Correlations* for the percentage of 'non-redundant residuals with absolute values greater than 0.05'. This percentage should be less than 50%, and the smaller it is, the better.



18.7.3 Orthogonal rotation (varimax) Ⓑ

The first analysis I asked you to run was using an orthogonal rotation, but I also asked you to rerun the analysis using oblique rotation. The results of both analyses will be presented to highlight the differences. This comparison will also be a useful way to show the circumstances in which one type of rotation might be preferable to another.

Output 18.9 shows the rotated factor matrix (called the rotated component matrix in PCA), which is a matrix of the factor loadings for each variable on each factor. This matrix contains the same information as the factor matrix in Output 18.5, but calculated *after* rotation. Factor loadings less than 0.3 have not been displayed because we asked for these to be suppressed, and the variables are listed in the order of size of their factor loadings because we asked for the output to be sorted by size (Figure 18.14). For all other parts of the output I suppressed the variable labels (to save space) but at this point my outputs show the variable labels because it makes interpretation easier.

Before rotation (Output 18.5) most variables loaded highly on the first factor and the remaining factors didn't really get a look-in. The rotation of the factor structure has clarified things considerably: there are four factors and most variables load very highly on only one factor.[10] In cases where a

10 The suppression of loadings less than 0.3 and ordering variables by their loading size makes this pattern easy to see.

Rotated Factor Matrix[a]

	Factor			
	1	2	3	4
I have little experience of computers	.768			
SPSS always crashes when I try to use it	.577			
I worry that I will cause irreparable damage because of my incompetence with computers	.519			
All computers hate me	.518			
I don't understand statistics	.507		.446	
Computers are out to get me	.454			
Computers are useful only for playing games	.450			
Computers have minds of their own and deliberately go wrong whenever I use them	.399			
My friends are better at statistics than me		.741		
If I'm good at statistics my friends will think I'm a nerd		.723		
Everybody looks at me when I use SPSS	.323	.575		
My friends are better at SPSS than I am		.520		
My friends will think I'm stupid for not being able to cope with SPSS		.486		
I wake up under my duvet thinking that I am trapped under a normal distribution			.546	
I dream that Pearson is attacking me with correlation coefficients			.503	
I can't sleep for thoughts of eigenvectors			.490	
I weep openly at the mention of central tendency			.441	
Standard deviations excite me			-.393	
Statistics makes me cry			.365	
People try to tell you that SPSS makes statistics easier to understand but it doesn't			.310	
I have never been good at mathematics				.743
I did badly at mathematics at school				.621
I slip into a coma whenever I see an equation				.580

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Output 18.9

variable loads highly onto more than one factor the loading is typically higher for one factor than another. For example, 'Everybody looks at me when I use SPSS' loads highly on both factor 1 and 2, but the loading for factor 2 (0.575) is greater than for factor 1 (0.323), so it makes sense to think of it as making a bigger contribution to factor 2 than to factor 1. Remember that every variable has a loading on every factor, it just appears as though they don't in Output 18.9 because we asked that they not be printed if they were lower than 0.3. It's not the case that a variable loads on one factor but not another (although people often use that turn of phrase); variables load on all factors, but to interpret the factors we assign variables to a factor based on their having a greater loading on that factor than on the others. If a variable has similar-sized loadings across two or more factors this could be because the factors reflect related constructs (and if you expect factors to correlate this can be fine) or that it is not a good item for distinguishing those constructs.

The next step is to look at the content of questions that load highly on the same factor to try to identify common themes. If the mathematical factors represent some real-world construct then common themes among highly loading questions can help us identify what that construct might be. The questions that load highly on factor 1 seem to relate to using computers or SPSS. Therefore, we might label this factor *fear of computers*. The questions that load highly on factor 2 contain some component of social evaluation and might reflect *peer evaluation*. The questions that load highly on factor 3 seem to relate to different aspects of statistics; therefore, we might label this factor *fear of statistics*. The three questions that load highly on factor 4 relate to mathematics, and we might label this factor *fear of mathematics*.

This analysis seems to reveal that the questionnaire is composed of four subscales:

- Factor 1: fear of computers
- Factor 2: fear of negative peer evaluation
- Factor 3: fear of statistics,
- Factor 4: fear of mathematics

There are two possibilities here. The first is that the SAQ failed to measure what it set out to (namely, SPSS anxiety) but instead measures related constructs. The second is that these four constructs are sub-components of SPSS anxiety; however, the factor analysis does not indicate which of these possibilities is true.

18.7.4 Oblique rotation (direct oblimin) Ⓑ

With oblique rotation the factor matrix is split into two: the *pattern matrix* and the *structure matrix* (see Jane Superbrain Box 18.1). For orthogonal rotation these matrices are the same. The pattern matrix contains the factor loadings and is comparable to the factor matrix that we interpreted for the orthogonal rotation. The structure matrix adjusts for the relationship between factors; it is the product of the pattern matrix and the matrix containing the correlation coefficients between factors). Most researchers interpret the pattern matrix, because it is usually simpler, but there are situations in which values in the pattern matrix are suppressed because of relationships between the factors. Therefore, the structure matrix is a useful double-check and Graham et al. (2003) recommend reporting both (with some useful examples of why this can be important).

Pattern Matrix[a]

	Factor			
	1	2	3	4
I have little experience of computers	.818			
SPSS always crashes when I try to use it	.604			
I worry that I will cause irreparable damage because of my incompetence with computers	.550			
All computers hate me	.537			
Computers are useful only for playing games	.478			
Computers are out to get me	.460			
I don't understand statistics	.447			.354
Computers have minds of their own and deliberately go wrong whenever I use them	.401			
My friends are better at statistics than me		.784		
If I'm good at statistics my friends will think I'm a nerd		.769		
Everybody looks at me when I use SPSS		.550		
My friends are better at SPSS than I am		.497		
My friends will think I'm stupid for not being able to cope with SPSS		.460		
I have never been good at mathematics			.821	
I did badly at mathematics at school			.678	
I slip into a coma whenever I see an equation			.632	
I wake up under my duvet thinking that I am trapped under a normal distribution				.550
I dream that Pearson is attacking me with correlation coefficients				.519
I can't sleep for thoughts of eigenvectors				.494
I weep openly at the mention of central tendency				.472
Standard deviations excite me				-.401
Statistics makes me cry				.368
People try to tell you that SPSS makes statistics easier to understand but it doesn't				.345

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 10 iterations.

Output 18.10

Structure Matrix

	Factor			
	1	2	3	4
I have little experience of computers	.815	.350	.397	.389
I don't understand statistics	.628		.422	.583
SPSS always crashes when I try to use it	.620		.336	
All computers hate me	.562			
I worry that I will cause irreparable damage because of my incompetence with computers	.556			
Computers are out to get me	.503			
Computers are useful only for playing games	.481			
Computers have minds of their own and deliberately go wrong whenever I use them	.443			
My friends are better at statistics than me		.746		
If I'm good at statistics my friends will think I'm a nerd		.724		
Everybody looks at me when I use SPSS	.454	.643		
My friends are better at SPSS than I am	.402	.583		
My friends will think I'm stupid for not being able to cope with SPSS	.409	.550		
I have never been good at mathematics	.402		.826	.452
I did badly at mathematics at school	.350		.705	.413
I slip into a coma whenever I see an equation	.351		.661	.378
I wake up under my duvet thinking that I am trapped under a normal distribution	.319		.374	.596
I dream that Pearson is attacking me with correlation coefficients			.305	.541
I can't sleep for thoughts of eigenvectors			.331	.533
I weep openly at the mention of central tendency				.459
Standard deviations excite me				-.424
Statistics makes me cry				.395
People try to tell you that SPSS makes statistics easier to understand but it doesn't				.317

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.

Output 18.11

The same four factors as for the orthogonal rotation seem to have emerged from the pattern matrix in Output 18.10:

- Factor 1: fear of computers
- Factor 2: fear of negative peer evaluation
- Factor 3: fear of mathematics
- Factor 4: fear of statistics

Note that the factors are in a different order (factor 3 after orthogonal rotation is factor 4 after oblique rotation, and vice versa). The structure matrix (Output 18.11) differs in that shared variance is not ignored. The picture becomes more complicated because variables load highly on more than one factor. In particular, there is overlap between factors 1, 3 and 4 (which makes sense because they relate to fear of computing, mathematics and statistics, and you might expect overlap between those fears) and also factors 1 and 2 (which also makes sense because the questions that overlap mention SPSS so capture both peer evaluation and computing). This example highlights why the pattern matrix is preferable for interpretative reasons: because it contains information about the *unique* contribution of a variable to a factor.

Output 18.12 is a matrix containing the correlation coefficients between factors. As predicted from the structure matrix, factors 3 and 4 (fear of mathematics and statistics) are strongly correlated ($r = 0.541$), and factor 1 (fear of computing) is strongly correlated with all factors (rs = 0.428, 0.478 and 0.494). The correlations between factor 2 (peer evaluation) and the factors representing fear of mathematics and statistics are weaker, but nevertheless substantial (in the region of 0.3). To sum up, the latent constructs represented by the factors are related. If the constructs were independent, oblique rotation should produce an identical solution to an orthogonal rotation and the factor correlation matrix would be an identity matrix (i.e., all factors would have correlation coefficients of 0). Therefore, this matrix is useful for assessing how reasonable it is to assume independence between factors; for these data it appears that we cannot assume independence and so the obliquely rotated solution is a more reasonable representation of reality.



Cramming Sam's Tips
Interpretation

- If you've conduced orthogonal rotation then look at the table labelled *Rotated Factor Matrix*. For each variable, note the factor/component for which the variable has the highest loading (above about 0.3–0.4 when you ignore the plus or minus sign). Try to make sense of what the factors represent by looking for common themes in the items that load highly on the same factor.
- If you've conducted oblique rotation then do the same as above but for the *Pattern Matrix*. Double-check what you find by doing the same for the *Structure Matrix*.



On a theoretical level the dependence between factors does not cause concern; we might expect a strong relationship between fear of mathematics, fear of statistics and fear of computing. Generally, those with less background (or educational opportunities) in mathematics and computing tend to be the ones who struggle with statistics. However, we would not necessarily expect these constructs to correlate strongly with fear of peer evaluation (because this construct is more socially based) and this is the factor that correlates least with all others – on a theoretical level, things make sense.

Factor Correlation Matrix

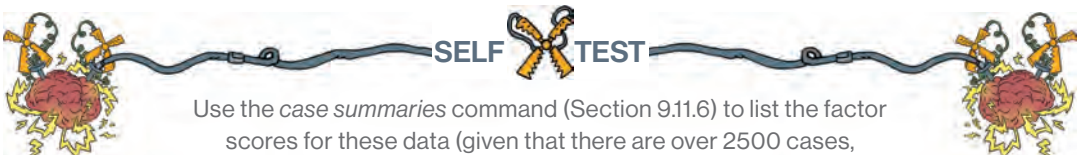
Factor	1	2	3	4
1	1.000	.428	.478	.494
2	.428	1.000	.289	.282
3	.478	.289	1.000	.541
4	.494	.282	.541	1.000

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser ...

Output 18.12

18.7.5 Factor scores Ⓑ

Having reached a suitable solution and rotated it, we can look at the factor scores. SPSS will display the component score matrix *B* (see Section 18.3.3) from which the factor scores are calculated. I haven't reproduced this table here because I can't think of a reason why most people would want to look at it. We asked for scores to be calculated based on the Anderson–Rubin method, and these scores will be in the data editor in columns labelled *FAC1_1*, *FAC2_1*, *FAC3_1* and *FAC4_1* for each factor, respectively. If you asked for factor scores in the subsequent obliquely rotated solution then these will appear in the data editor in four columns labelled *FAC2_1* and so on.



Use the *case summaries* command (Section 9.11.6) to list the factor scores for these data (given that there are over 2500 cases, restrict the output to the first 10).

Output 18.13 shows the factor scores for the first 10 participants after the oblique rotation (because we have decided that this is the better model to use). In using the oblique solution, remember that factor 3 relates to fear of mathematics and factor 4 to fear of statistics. Participant 1 has fairly low negative scores on all factors, suggesting general low anxiety. Participant 5 has negative scores on factors 1 and 2 but positive scores on factors 3 and 4 suggesting fear of mathematics and statistics (especially mathematics), low anxiety about computing and peer evaluation. Participant 10 has a large negative scores on factor 4 but positive scores on the other factors (and large for factors 1 and 3) suggesting a lot of fear of mathematics and computing (and to a lesser extent peer evaluation) but low fear on statistics.

Factor scores can be used in this way to assess the relative fear of one person compared to another, or we could add the scores up to obtain a single score for each participant (which we might assume represents SPSS anxiety as a whole). We can also use factor scores in regression when groups of predictors correlate so highly that there is multicollinearity. However, normally people do not use factor scores themselves but instead sum scores on items that they have decided load onto the same factor (e.g., create a score for statistics anxiety by adding up a person's scores on items 1, 3, 4, 5, 12, 16, 20 and 21).

Case Summaries[a]

	Case Number	A-R factor score 1 for analysis 1	A-R factor score 2 for analysis 1	A-R factor score 3 for analysis 1	A-R factor score 4 for analysis 1
1	1	-.37826	-.97052	-.95492	-.36812
2	2	1.00126	1.13802	.41100	-.74140
3	3	.80731	-.52026	.19187	.22266
4	4	.85789	.23736	-1.31382	.61144
5	5	-1.12545	-.45627	1.61115	.24798
6	6	-1.12276	-.06630	-1.99204	.68197
7	7	.37920	-.46823	-.75412	-.89562
8	8	1.13439	-1.31482	.24656	-.42413
9	9	.47483	2.28054	-.29487	-2.01469
10	10	1.79792	.97815	1.53262	-2.67838
Total N		10	10	10	10

a. Limited to first 10 cases.

Output 18.13

18.7.6 Summary Ⓑ

To sum up, the analyses revealed four underlying subscales in our questionnaire that may, or may not, relate to genuine sub-components of SPSS anxiety. It also seems as though an obliquely rotated solution was preferred due to the relationships between factors. The use of factor analysis is purely exploratory; it should be used only to guide future hypotheses, or to inform researchers about patterns within data sets. A great many decisions are left to the researcher using factor analysis, and so researcher degrees of freedom (Section 3.3.2) come into play a lot: try to make informed, impartial decisions and resist the lure of decisions that give you the outcomes you want to get.

18.8 How to report factor analysis Ⓐ

When reporting factor analysis, provide readers with enough information to make an informed opinion about what you've done. We need to be clear about our criteria for extracting factors and the method of rotation used. We should produce a table of the rotated factor loadings of all items and flag (in bold) values above a criterion level (I would personally choose 0.40). Report the percentage of variance that each factor explains and possibly the eigenvalue too. Table 18.1 shows an example of such a table for the SAQ data (oblique rotation); note that I have reported the sample size in the title.

A table of factor loadings and a description of the analysis are a bare minimum. You should provide some information on the data adequacy. Ideally, deposit the table of correlations on an open science repository such as the Open Science Framework (https://osf.io/) so that someone can reproduce your analysis. For example:

- ✓ A principal axis factor analysis was conducted on the 23 items with oblique rotation (direct oblimin). The Kaiser–Meyer–Olkin measure verified the data were likely suitable for factor analysis, KMO = 0.92 ('marvellous' according to Kaiser & Rice, 1974), and all KMO values for individual items were greater than 0.84, which is well above the acceptable limit of 0.5 (Kaiser & Rice, 1974). A parallel analysis revealed four factors with eigenvalues greater than those at the

Table 18.1 Summary of exploratory factor analysis results for the SPSS anxiety questionnaire ($N = 2571$)

	Rotated factor loadings			
Item	Fear of Computers	Peer Evaluation	Fear of Maths	Fear of Statistics
I have little experience of computers	**0.82**	0.00	0.02	−0.03
SPSS always crashes when I try to use it	**0.60**	0.01	0.06	−0.04
I worry that I will cause irreparable damage because of my incompetence with computers	**0.55**	0.03	−0.03	0.02
All computers hate me	**0.54**	0.04	0.03	−0.01
Computers are useful only for playing games	**0.48**	−0.04	−0.02	0.07
Computers are out to get me	**0.46**	0.03	0.04	0.02
I don't understand statistics	**0.45**	−0.01	0.02	0.35
Computers have minds of their own and deliberately go wrong whenever I use them	**0.40**	0.04	0.06	0.00
My friends are better at statistics than me	−0.12	**0.78**	0.04	0.01
If I'm good at statistics my friends will think I'm a nerd	−0.12	**0.77**	0.00	0.02
Everybody looks at me when I use SPSS	0.22	**0.55**	−0.01	0.00
My friends are better at SPSS than I am	0.15	**0.50**	0.05	0.03
My friends will think I'm stupid for not being able to cope with SPSS	0.22	**0.46**	−0.03	0.02
I have never been good at mathematics	0.01	−0.01	**0.82**	0.01
I did badly at mathematics at school	0.00	0.01	**0.68**	0.04
I slip into a coma whenever I see an equation	0.04	0.01	**0.63**	0.01
I wake up under my duvet thinking that I am trapped under a normal distribution	0.00	0.05	0.06	**0.55**
I dream that Pearson is attacking me with correlation coefficients	0.05	−0.02	0.01	**0.52**
I can't sleep for thoughts of eigenvectors	0.02	0.00	0.05	**0.49**
I weep openly at the mention of central tendency	−0.03	0.01	−0.01	**0.47**
Standard deviations excite me	−0.01	0.01	−0.04	**−0.40**
Statistics makes me cry	−0.02	0.01	0.06	0.37
People try to tell you that SPSS makes statistics easier to understand but it doesn't	0.00	0.00	−0.06	0.35
Eigenvalues (before extraction)	5.86	1.99	1.56	1.25
% of variance (before extraction)	25.47	8.64	6.79	5.43
α	0.77	0.79	0.77	0.71

Note: Factor loadings over 0.40 appear in bold.

95th percentile of 20 random data sets with no underlying factor structure. In combination, these factors explained over 40% of the variance. Table 18.1 shows the factor loadings after rotation. The items that cluster on the same factor suggest that factor 1 represents a fear of computing, factor 2 represents peer evaluation concerns, factor 3 a fear of mathematics and factor 4 a fear of statistics.

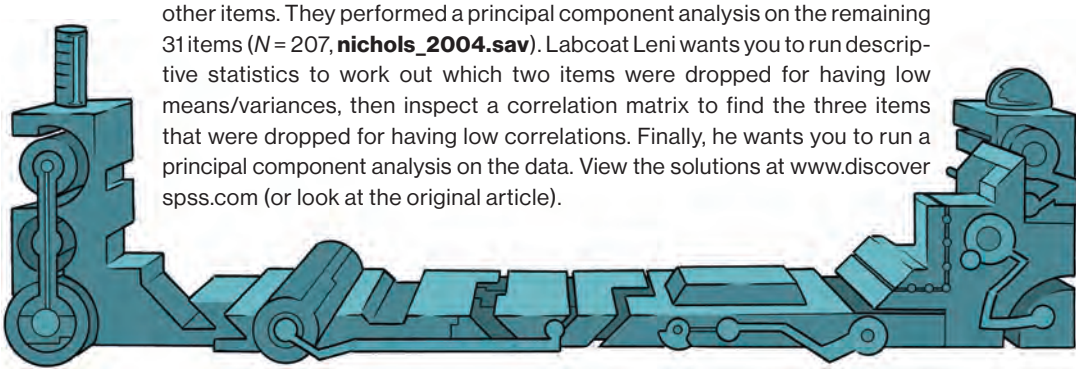


Labcoat Leni's Real Research 18.1 Worldwide addiction? Ⓑ

Nichols, L. A., & Nicki, R. (2004). *Psychology of Addictive Behaviors*, 18(4), 381–384.

The increasing popularity (and usefulness) of the Internet has led to the serious problem of Internet addiction. To research this construct it's helpful to be able to measure it, so Laura Nichols and Richard Nicki developed the Internet Addiction Scale (Nichols & Nicki, 2004). This 36-item questionnaire contains items such as 'I have stayed on the Internet longer than I intended to' and 'My grades/work have suffered because of my Internet use' to which responses are made on a five-point scale (never, rarely, sometimes, frequently, always). (Incidentally, while researching this topic I encountered a deeply ironic Internet addiction recovery website that offered enough reasons to stay online for a week.)

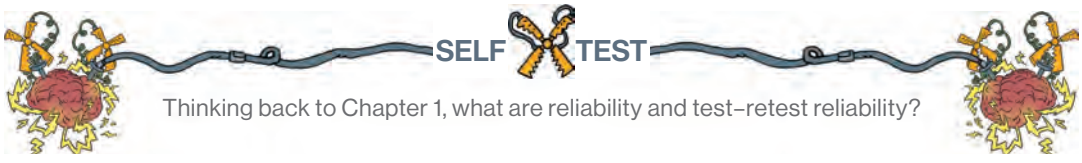
The authors dropped two items because they had low means and variances, and dropped three others because of relatively low correlations with other items. They performed a principal component analysis on the remaining 31 items ($N = 207$, **nichols_2004.sav**). Labcoat Leni wants you to run descriptive statistics to work out which two items were dropped for having low means/variances, then inspect a correlation matrix to find the three items that were dropped for having low correlations. Finally, he wants you to run a principal component analysis on the data. View the solutions at www.discoverspss.com (or look at the original article).



18.9 Reliability analysis Ⓑ

18.9.1 What is reliability? Ⓒ

If you're using factor analysis to validate a questionnaire, it is useful to check the reliability of your scale within the sample/context that you have used it.



Thinking back to Chapter 1, what are reliability and test–retest reliability?



Reliability means that scores on a measure (or in this case questionnaire) should consistently reflect the construct that it is measuring. One way to think of reliability is that, other things being equal, a person should get the same score on a questionnaire if they complete it at two different points in time (we have already discovered that this is called test–retest reliability). So, someone who is terrified of SPSS and who scores highly on our SAQ should score similarly highly if we tested them a month later (assuming they hadn't completed some kind of SPSS-anxiety therapy in that month). Another way to look at reliability is to say that two people who are the same in terms of the construct being measured should get the same score. So, if we took two people who were equally SPSS-anxious, they should get more or less identical scores on the SAQ. Likewise, if we took two people who loved SPSS, they should get equally low scores. The SAQ wouldn't be an accurate measure of SPSS anxiety if we took someone who loved SPSS and someone who was terrified of it and they got the same score! In statistical terms, the usual way to look at reliability is based on the idea that individual items (or sets of items) should produce results consistent with the overall questionnaire. So, if we take someone scared of SPSS, then their overall score on the SAQ will be high; if the SAQ is reliable then if we randomly select some items from it the person's score on those items should also be high.

Reliability isn't a property of a scale or measure, it is a function of the measure, the people responding to the measure and the context in which the measure is taken. So, if a measure shows reliability in one sample or in one context it doesn't mean it will be reliable in a different sample or context. Let's formalize the discussion a bit. A scale or questionnaire is supposed to measure some underlying construct (in our case SPSS anxiety). If lots of people complete our measure, their scale scores will differ. This variability in scores, we hope, reflects variability in the construct itself but it will also reflect measurement error. For example, variability in our SPSS anxiety scores will partly reflect genuine variability in 'true' SPSS anxiety but, because we are not measuring SPSS anxiety directly, will also reflect measurement error (both from the scale itself and the context in which the scale is applied). As an equation we could say that observed scores (X) are made up of the unobserved true score on the construct (T) and measurement error (E):

$$X = T + E$$
$$\text{observed scores} = \text{true scores} + \text{measurement error} \tag{18.9}$$

It follows then that variance in observed scores (σ_X^2) is likewise made up of the variance in true scores (σ_T^2) and the variance in error (σ_E^2):

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{18.10}$$

When we ask how reliable our observed scores are were asking how closely the observed scores match the true scores. In other words, how much variance does the observed scores share with the unobserved 'true' scores on the construct of interest. We can express the reliability (ρ_{XT}^2) as:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} \tag{18.11}$$

An alternative way to express this equation is by substituting the value of σ_X^2 from Eq. 18.10 to give:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \tag{18.12}$$

Whichever equation you use, it expresses the variability in true scores as a proportion of the variability in observed scores. In other words, it quantifies the proportion of variability in observed scores that is attributable to the construct that those scores are supposed to measure. Another way to think about the reliability equation is that it is the correlation between the true scores and the observed scores squared. Hopefully it makes intuitive sense that reliability implies scores that (1) correlate highly with the true scores on the construct being measured; and (2) share a decent amount of variance with the construct being measured. Reliability quantifies these two related things.

18.9.2 Cronbach's alpha ©

The problem with reliability as we have just defined it is that to compute it we need to know the variance of true scores, but the 'true scores' are unknown and unobservable. So, we need to know the variance of something that we can't measure directly. One impractical option is to find a different test that measures the same construct (known as a *parallel test*). If these two tests are administered to the same people then it's reasonable to assume that the true scores are very similar for the two tests. In fact we can assume that the true scores differ only by a constant value that shifts scores for all respondents. In other words, $T_{\text{Test 1}} = T_{\text{Test 2}} + c$. In other words, the true, unobservable, scores are perfectly correlated for the two tests although the scores themselves might be higher or lower on one test by a fixed amount (because, for example, it contains fewer items). This assumption is known as *essential tau-equivalence*, and it allows us to estimate the variance in true scores as being the covariance between scores on the two tests (see M. B. Miller, 1995, for a clear explanation). Using this estimate of the true score variance, it turns out that the reliability is equal to the correlation between the scores on the two parallel tests (M. B. Miller, 1995; Novick, 1966).

As I said, though, it's impractical to find parallel tests and it means doubling up measures and increasing the burden on responders. A more practical version of the parallel test idea is to use **split-half reliability**. This method splits the scale into two randomly selected sets of items, scores for each participant are computed on the two halves and the reliability of the scale is estimated as the correlation between the scores on the two halves. In effect, one half of the scale acts as a parallel test for the other half. This idea is intuitive: if a scale is reliable a person's score on one half of the scale should be the same as (or similar to) their score on the other half. Across several participants, scores from the two halves of the scale should correlate very highly.

The problem with the split-half method is that there are several ways in which a set of data can be randomly split into two and so the results could be a product of the way in which the data were split. To overcome this problem, Cronbach (1951) came up with a measure that is loosely equivalent to creating two sets of items in every way possible and computing the correlation coefficient for each split. The average of these values is equivalent to **Cronbach's alpha**, α, which is the most common measure of scale reliability:[11]

$$\alpha = \frac{N^2\overline{\text{cov}}}{\sum s^2_{\text{item}} + \sum \text{cov}_{\text{item}}} \tag{18.13}$$

11 Although this is the easiest way to conceptualize Cronbach's α, whether it is exactly equal to the average of all possible split-half reliabilities depends on exactly how you calculate the split-half reliability (see the glossary for computational details). If you use the Spearman–Brown formula, which takes no account of item standard deviations, then Cronbach's α will be equal to the average split-half reliability only when the item standard deviations are equal; otherwise α will be smaller than the average. However, if you use a formula for split-half reliability that does account for item standard deviations (such as Flanagan, 1937; Rulon, 1939) then α will always equal the average split-half reliability (Cortina, 1993).

This equation is less complicated than it looks. For each item on our scale we can calculate two things: the variance within the item, and the covariance between a particular item and any other item on the scale. Put another way, we can construct a variance–covariance matrix of all items. In this matrix the diagonal elements will be the variance within a particular item, and the off-diagonal elements will be covariances between pairs of items. The top half of the equation is the number of items (N) squared multiplied by the average covariance between items (the average of the off-diagonal elements in the variance–covariance matrix). The bottom half is the sum of all the item variances and item covariances (i.e., the sum of everything in the variance–covariance matrix).

There is a standardized version of the coefficient too, which essentially uses the same equation except that correlations are used rather than covariances, and the bottom half of the equation uses the sum of the elements in the correlation matrix of items (including the 1s that appear on the diagonal of that matrix). The normal alpha is appropriate when items on a scale are summed to produce a single score for that scale (the standardized alpha is not appropriate in these cases). The standardized alpha is useful, though, when items on a scale are standardized before being summed.

18.9.3 Problems with Cronbach's α Ⓑ

Cronbach's α is the Imagine Dragons of reliability measures: people who know what they're talking about consider it to be terrible, but it's hugely popular.[12] Like, popular enough that I feel obliged to waste your time telling you about it because if you read any scientific paper that uses any kind of self-report scale, the authors *will* report Cronbach's α. It's worth knowing about so that you can be better than other scientists and report something different. The problems with α are widely known and there are some fabulous papers with more detailed coverage, should you need it (e.g., Cortina, 1993; Dunn et al., 2014; McNeish, 2018; M. B. Miller, 1995; Pedhazur & Schmelkin, 1991; Revelle & Condon, 2019; Revelle & Zinbarg, 2009).

We saw earlier that the problem with estimating reliability is that we don't know the variance of true scores (because these scores cannot be measured directly) and to estimate it we need to make the assumption of essential tau-equivalence. In the case of Cronbach's alpha, because it (effectively) splits the scale items into all possible combinations, essential tau-equivalence must be assumed for every pair of items. Mathematically, this assumption is the same as saying that all items load equally onto one factor (M. B. Miller, 1995). The example in this chapter shows that in reality loadings will differ. In short, the assumption that all items load equally onto one factor is restrictive and highly unrealistic.

In the absence of essential tau-equivalence, α underestimates test reliability. A positive view of this fact is that α is a 'lower bound' or 'errs on the side of caution' and that your scale will be more reliable than you think it is. However, α is also affected by other things in unpredictable ways. For one thing, α depends on the number of items on the scale. You'll notice that the top half of the equation for α includes the number of items squared. Therefore, as the number of items on the scale increases, α will increase. Therefore, it's possible to get a large value of α because you have a lot of items on the scale, and not because your scale is reliable. For example, Cortina (1993) reports data from two scales, both of which have $\alpha = 0.8$. The first scale has only three items, and the average correlation between items was a respectable 0.57; however, the second scale has 10 items with an average correlation between these items of a less respectable 0.28. Clearly the internal consistency of these scales differs, but according to Cronbach's α they are both equally reliable. Sampling consistency (whether samples are made up of people from the same or different underlying populations) also biases alpha, but sometimes leading to overestimations and sometimes underestimations of true reliability (Waller, 2008).

12 The hive mind of Twitter suggested I use the Imagine Dragons in this metaphor, but if they happen to be your favourite band take comfort from the fact that a middle-aged, obsolete, stats bloke dissing your favourite band is undoubtedly a badge of honour.

There's also the issue of 'unidimensionality', or the extent to which the scale measures one underlying factor or construct. It seems perfectly reasonable – if we're developing scales to measure a construct – to expect items on the scale to represent a single construct. Reality has other ideas. Reflecting on the example in this chapter, our SPSS-anxiety scale seemed to be composed of four related constructs, not one. This situation is common and shows that a key assumption of Cronbach's alpha is unrealistic. Some people deal with this inconvenient truth by turning it on its head and suggesting that alpha *measures* unidimensionality, with a low alpha indicating that the scale does not tap one underlying construct. However, data sets with the same α can have very different factor structures. Grayson (2004) showed that $\alpha = 0.8$ can be achieved in a scale with one underlying factor, with two moderately correlated factors and with two uncorrelated factors. Cortina (1993) has also shown that with more than 12 items, and fairly high correlations between items ($r > 0.5$), α can reach values around and above 0.7 (0.65 to 0.84) when the data have two or three underlying dimensions. These results show that α should not be used as a measure of 'unidimensionality' (Lucke, 2005; M. B. Miller, 1995; Revelle & Condon, 2019). Even Cronbach (1951) suggested that if several factors exist then the formula should be applied separately to items relating to different factors. In other words, if your questionnaire has subscales, α should be computed on these subscales items separately and not for the overall collection of items.

18.9.4 Interpreting Cronbach's α (more cautionary tales ...) Ⓑ

Summing up, Cronbach's α doesn't measure unidimensionality, is based on highly unrealistic assumptions, and is affected by things other than reliability (sampling consistency, the number of items etc.). If that doesn't put you off, then a value of 0.7 to 0.8 is widely touted as acceptable and values substantially lower indicate an unreliable scale. Kline (1999) notes that although the generally accepted value of 0.8 is appropriate for cognitive tests such as intelligence tests, for ability tests a cut-off point of 0.7 is more suitable. He goes on to say that when dealing with psychological constructs values below even 0.7 can, realistically, be expected because of the diversity of the constructs being measured. Some even suggest that in the early stages of research values as low as 0.5 will suffice (Nunnally, 1978). However, given that α is affected by things other than reliability, it doesn't make sense to use general cut-offs; instead you should interpret it within the contexts of both how it was measured and the research you're doing (Pedhazur & Schmelkin, 1991).

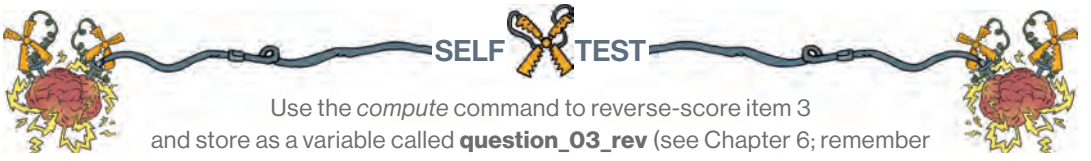


A final warning is about items that have a reverse phrasing. For example, in the SAQ there is one item (question 3) that was phrased the opposite way around to all other items. The item was 'standard deviations excite me'. Compare this to any other item and you'll see it requires the opposite response. For example, item 1 is 'statistics makes me cry'. If you don't like statistics then you'll strongly agree with this statement and so will get a score of 5 on our scale. For item 3, if you hate statistics then standard deviations are unlikely to excite you so you'll strongly disagree and get a score of 1 on the scale. These reverse-phrased items are important for reducing response bias because participants need to pay attention to the questions. For factor analysis, this reverse phrasing doesn't matter; all that happens is you get a negative factor loading for any reversed items (in fact, you'll see that item 3 has a negative factor loading in Output 18.10). However, these reverse-scored items will affect alpha. To see why, think about the equation for Cronbach's α. The top half incorporates the *average* covariance between items. If an item is reverse-phrased then it will have a negative relationship with other items, hence the covariances between this item and other items will be negative. The average covariance is the sum of covariances divided by the number of covariances, and by including a bunch of negative values we reduce the sum of covariances, and hence we also reduce Cronbach's α, because the top half of the equation gets smaller. In extreme cases, it is even possible to get a negative value for Cronbach's α,

simply because the magnitude of negative covariances is bigger than the magnitude of positive ones. A negative Cronbach's α doesn't make sense, but it happens, usually because of reverse-phrased items. If you have reverse-phrased items then reverse the way in which they're scored before you conduct reliability analysis.

If you have reverse-phrased items then reverse the way in which they're scored before you conduct reliability analysis. This is quite easy. To take our SAQ data, items are scored as 1 = strongly disagree, 2 = disagree, 3 = neither, 4 = agree and 5 = strongly agree. For all but one item the statement is phrased such that agreement indicates statistics anxiety, but for item 3 (standard deviations excite me), *disagreement* indicates statistics anxiety. To reflect this numerically, we reverse the scale such that 1 = strongly agree, 2 = agree, 3 = neither, 4 = disagree and 5 = strongly disagree. In doing so, a statistics-anxious person gets 5 on this item (because they'd strongly disagree with it), which is consistent with 'strongly agreeing' to an item like 'statistics makes me cry'.

To reverse the scoring find the maximum value of your response scale (in this case 5) and add 1 to it (6 in this case). For each person, subtract from this value the score they got. For example, someone who scored 5 originally now scores 6 – 5 = 1, and someone who scored 1 originally now gets 6 – 1 = 5. Someone in the middle of the scale with a score of 3 will still get 6–3 = 3. We can get SPSS to do it for all participants simultaneously.



Use the *compute* command to reverse-score item 3 and store as a variable called **question_03_rev** (see Chapter 6; remember that you are changing the variable to 6 minus its original value).

18.10 Reliability analysis using SPSS Ⓑ

Let's test the reliability of the SAQ (**saq.sav**). You should have reverse-scored item 3 (see above), but if you can't be bothered, load up the file **saq_q3_reversed.sav**, which contains the reversed scores in a variable called **question_03_rev**. Remember that I said we should conduct reliability analysis on individual subscales. Using the results from the oblique rotation (Output 18.10) we have these four subscales:

1 Subscale 1 (*fear of computing*): questions 6, 7, 10, 13, 14, 15, 18
2 Subscale 2 (*peer evaluation*): questions 2, 9, 19, 22, 23
3 Subscale 3 (*fear of mathematics*): questions 8, 11, 17
4 Subscale 4 (*fear of statistics*): questions 1, 3, 4, 5, 12, 16, 20, 21

To conduct each reliability analysis select *Analyze* > *Scale* > Reliability Analysis... to access the dialog box in Figure 18.15. Drag any variables that you want to analyse to the box labelled *Items* (or click ▶). Remember that you can select several items at the same time by holding down *Ctrl* (*Cmd* on a Mac) while you select the variables. To begin with, let's select the items from the fear of statistics subscale: questions 1, 3, 4, 5, 12, 16, 20 and 21 (Figure 18.15). Remember to use **question_03_rev** for item 3.

There are several reliability statistics available. The default option is Cronbach's α, which is fine for our purpose, but you can change the method (e.g., to the split-half method) by clicking on Alpha to reveal a drop-down list of possibilities. Also, it's a good idea to type the name of the scale (in this case 'Fear of Statistics') into the box labelled *Scale label* because doing so adds a header to the output with whatever you type in this box: typing a sensible name here will make your output easier to follow.

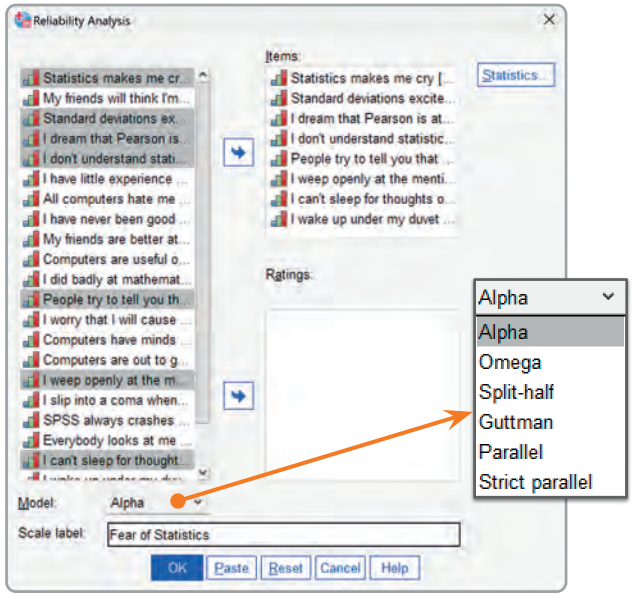


Figure 18.15 Main dialog box for reliability analysis

Click Statistics... to access the dialog box in Figure 18.16. One of most useful options for questionnaire reliability is *Scale if item deleted*, which tells us what the value of α would be if each item were deleted. If our questionnaire is reliable then we would not expect any one item to greatly affect the overall reliability. In other words, no item should cause a substantial change in α if it is removed. If an item's removal changes alpha substantially then you should review your items (and that probably means going back to square one with the factor analysis).

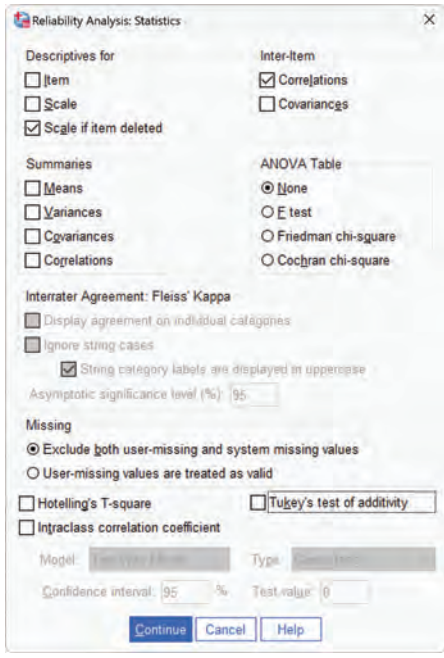


Figure 18.16 Statistics for reliability analysis

The inter-item correlations and covariances (and summaries) provide us with correlation coefficients and averages for items on our scale. We should already have these values from our factor analysis, so there is little point in selecting these options. Options like the *F-test, Friedman chi-square* (if your data are ranked), *Cochran chi-square* (if your data are dichotomous) and *Hotelling's T-square* compare the central tendency of different items on the questionnaire. These tests might be useful to check that items have similar distributional properties (i.e., the same average value), but, given the large sample sizes you ought to be using for factor analysis, they will inevitably produce significant results even when only small differences exist between the questionnaire items.

You can also request an **intraclass correlation coefficient (ICC)**. The correlation coefficients that we encountered earlier in this book measure the relation between variables that measure different things. For example, the correlation between listening to Deathspell Omega and Satanism involves two classes of measures: the type of music a person likes and their religious beliefs. Intraclass correlations measure the relationship between two variables that measure the same thing (i.e., variables within the same class). Two common uses are in comparing paired data (such as twins) on the same measure, and assessing the consistency between judges' ratings of a set of objects (hence why it is found in the reliability statistics). If you'd like to know more, see Section 21.2.

Use the simple set of options in Figure 18.16 to run a basic reliability analysis. Click Continue to return to the main dialog box and OK to run the analysis.

18.11 Interpreting reliability analysis Ⓑ

Output 18.14 shows the results for the fear of statistics subscale. The value of Cronbach's α is presented in a small table and indicates the overall reliability of the scale. Bearing in mind what we've already noted about effects from the number of items, and how daft it is to apply general rules, we're looking for values in the region of about 0.7–0.8. In this case α is 0.707, which is in the region indicated by Kline (1999), and probably indicates good reliability.

Scale: Fear of Statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.707	.707	8

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Statistics makes me cry	23.3971	16.868	.332	.111	.693
Standard deviations excite me (reversed)	24.3617	16.664	.357	.131	.687
I dream that Pearson is attacking me with correlation coefficients	23.4096	16.004	.445	.206	.668
I don't understand statistics	23.4154	15.704	.483	.246	.660
People try to tell you that SPSS makes statistics easier to understand but it doesn't	23.3862	17.390	.266	.076	.706
I weep openly at the mention of central tendency	23.4072	16.568	.375	.145	.683
I can't sleep for thoughts of eigenvectors	23.3820	16.042	.437	.212	.670
I wake up under my duvet thinking that I am trapped under a normal distribution	23.4049	15.783	.488	.253	.659

Output 18.14

In the table labelled *Item-Total Statistics* the column labelled *Corrected Item-Total Correlation* shows the correlations between each item and the total score from the questionnaire. In a reliable scale all items should correlate well with the total. So, we're looking for items that don't correlate well with the overall score from the scale: if any of these values are less than about 0.3 then we've got problems, because it means that a particular item does not correlate very well with the scale overall. Items with low correlations may have to be dropped. For these data, all but one item have item–total correlations above 0.3, which is encouraging.

The values in the column labelled *Cronbach's Alpha if Item Deleted* are the values of the overall α if that item isn't included in the calculation. As such, they reflect the change in Cronbach's α that would be seen if an item were deleted. The overall α is 0.707, and so all values in this column should be around that same value. We're mainly looking for values of alpha greater than the overall α. If you think about it, if the deletion of an item increases Cronbach's α then this means that the deletion of that item improves reliability. Therefore, any items that have values of α in the column labelled *'Cronbach's Alpha if Item Deleted'* greater than the overall α may need to be deleted from the scale to improve its reliability. None of the items here would increase alpha if they were deleted, which is good news. It's worth noting that if you have doubts about items at this stage then you should rerun your factor analysis as well to make sure that the deletion of the item has not affected the factor structure.

Scale: Fear of Statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.537	.537	8

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Statistics makes me cry	24.36	11.794	.294	.111	.489
Standard deviations excite me	24.36	16.664	-.357	.131	.687
I dream that Pearson is attacking me with correlation coefficients	24.37	10.994	.421	.206	.441
I don't understand statistics	24.38	10.939	.426	.246	.439
People try to tell you that SPSS makes statistics easier to understand but it doesn't	24.35	12.191	.234	.076	.510
I weep openly at the mention of central tendency	24.37	11.589	.329	.145	.476
I can't sleep for thoughts of eigenvectors	24.34	11.195	.384	.212	.455
I wake up under my duvet thinking that I am trapped under a normal distribution	24.37	11.015	.429	.253	.439

Output 18.15

To illustrate the importance of reverse-scoring items before running reliability analysis, Output 18.15 shows the reliability analysis for the fear of statistics subscale, but done on the original data (i.e., without item 3 being reverse-scored). Note that (1) the overall α is considerably lower (0.537 rather than 0.707); (2) item 3 has a negative item–total correlation (which is a good way to spot if you have a potential reverse-scored item in the data that haven't been reverse-scored); and (3) the α if item deleted is 0.687. That is, if this item were deleted then the reliability goes up from about from 0.54 to about 0.69. This example, I hope, illustrates that failing to reverse-score items that have been phrased oppositely to other items on the scale will mess up your reliability analysis.



Let's now look at our subscale of fear of computing, Output 18.16 shows an overall α of 0.769, which is pretty good. The values in the column labelled *Corrected Item-Total Correlation* are again all above 0.3, which is also good. The values in the column labelled *Cronbach's Alpha if Item Deleted* show that none of the items would increase the reliability if they were deleted. This indicates that all items are positively contributing to the overall reliability.

Scale: Fear of computing

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.769	.769	7

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
I have little experience of computers	20.96	14.120	.677	.467	.699
All computers hate me	20.94	15.327	.489	.252	.740
Computers are useful only for playing games	20.98	15.842	.418	.188	.755
I worry that I will cause irreparable damage because of my incompetence with computers	20.96	15.450	.484	.251	.741
Computers have minds of their own and deliberately go wrong whenever I use them	20.94	15.992	.393	.161	.760
Computers are out to get me	20.92	15.685	.444	.205	.749
SPSS always crashes when I try to use it	20.96	15.202	.526	.304	.732

Output 18.16

Scale: Peer evaluation

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.785	.784	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
My friends will think I'm stupid for not being able to cope with SPSS	13.93	9.026	.493	.245	.766
My friends are better at statistics than me	13.88	8.361	.618	.399	.725
Everybody looks at me when I use SPSS	13.92	8.536	.571	.327	.741
My friends are better at SPSS than I am	13.91	8.960	.522	.275	.757
If I'm good at statistics my friends will think I'm a nerd	13.91	8.494	.598	.376	.732

Output 18.17

Scale: Fear of Mathematics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.773	.773	3

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
I have never been good at mathematics	6.94	2.816	.660	.437	.635
I did badly at mathematics at school	6.96	2.968	.595	.369	.708
I slip into a coma whenever I see an equation	6.95	3.068	.569	.333	.736

Output 18.18

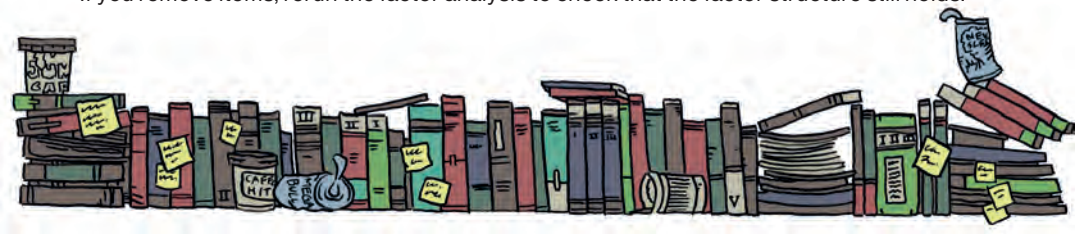
Moving onto our subscale of peer evaluation (Output 18.17). The overall α is 0.785, which is fine, the values in the column labelled *Corrected Item-Total Correlation* are again well above 0.3, and the values in the column labelled *Cronbach's Alpha if Item Deleted* show that none of the items would increase the reliability if they were deleted.

Finally, the fear of maths subscale (Output 18.18) shows an overall reliability of 0.773, which indicates good reliability. The values in the column labelled *Corrected Item-Total Correlation* are all well above 0.3, which is good, and the values in the column labelled *Cronbach's Alpha if Item Deleted* indicate that none of the items here would increase the reliability if they were deleted because all values in this column are less than the overall reliability value. Basically, it's all good.



Cramming Sam's Tips
Reliability

- Reliability analysis is used to measure the consistency of a measure.
- Remember to reverse-score any items that were reverse-phrased on the original questionnaire before you run reliability analysis.
- Run separate reliability analyses for all subscales of your questionnaire.
- Cronbach's α indicates the overall reliability of a questionnaire and values around 0.8 are good (or 0.7 for ability tests and the like).
- The *Cronbach's Alpha if Item Deleted* column tells you whether removing an item will improve the overall reliability: values greater than the overall reliability indicate that removing that item will improve the overall reliability of the scale. Look for items that dramatically increase the value of α and remove them.
- If you remove items, rerun the factor analysis to check that the factor structure still holds.



18.12 How to report reliability analysis Ⓑ

Report the reliabilities in the text using the symbol α and remember that if you follow APA practice (which I'm not doing) because Cronbach's α can't be larger than 1 then drop the zero before the decimal:

✓ The fear of computers ($\alpha = 0.77$), fear of peer evaluation ($\alpha = 0.79$), fear of mathematics ($\alpha = 0.77$), and fear of statistics ($\alpha = 0.71$), subscales of the SAQ all had high reliabilities.

However, the most common way to report reliability analysis when it follows a factor analysis is to report the values of Cronbach's α as part of the table of factor loadings. For example, in Table 18.1 notice that in the last row of the table I quoted the value of Cronbach's α for each subscale in turn.

18.13 Jane and Brian's story Ⓐ

Brian put his duvet over Jane, wedged a pillow under her head, turned the light out and went to his bedroom. He dug out a sleeping bag to replace the duvet that was keeping Jane warm on the sofa. He found it hard to sleep. Just when he thought he'd scared Jane off, she turns up looking and sounding exhausted. With no explanation for the past five days, she passes out on his sofa. She was an enigma, that was for sure. A brilliant, dazzling, enigma. Brian's head told him to stay away, but it was too late: he was strapped in for whatever the ride was.

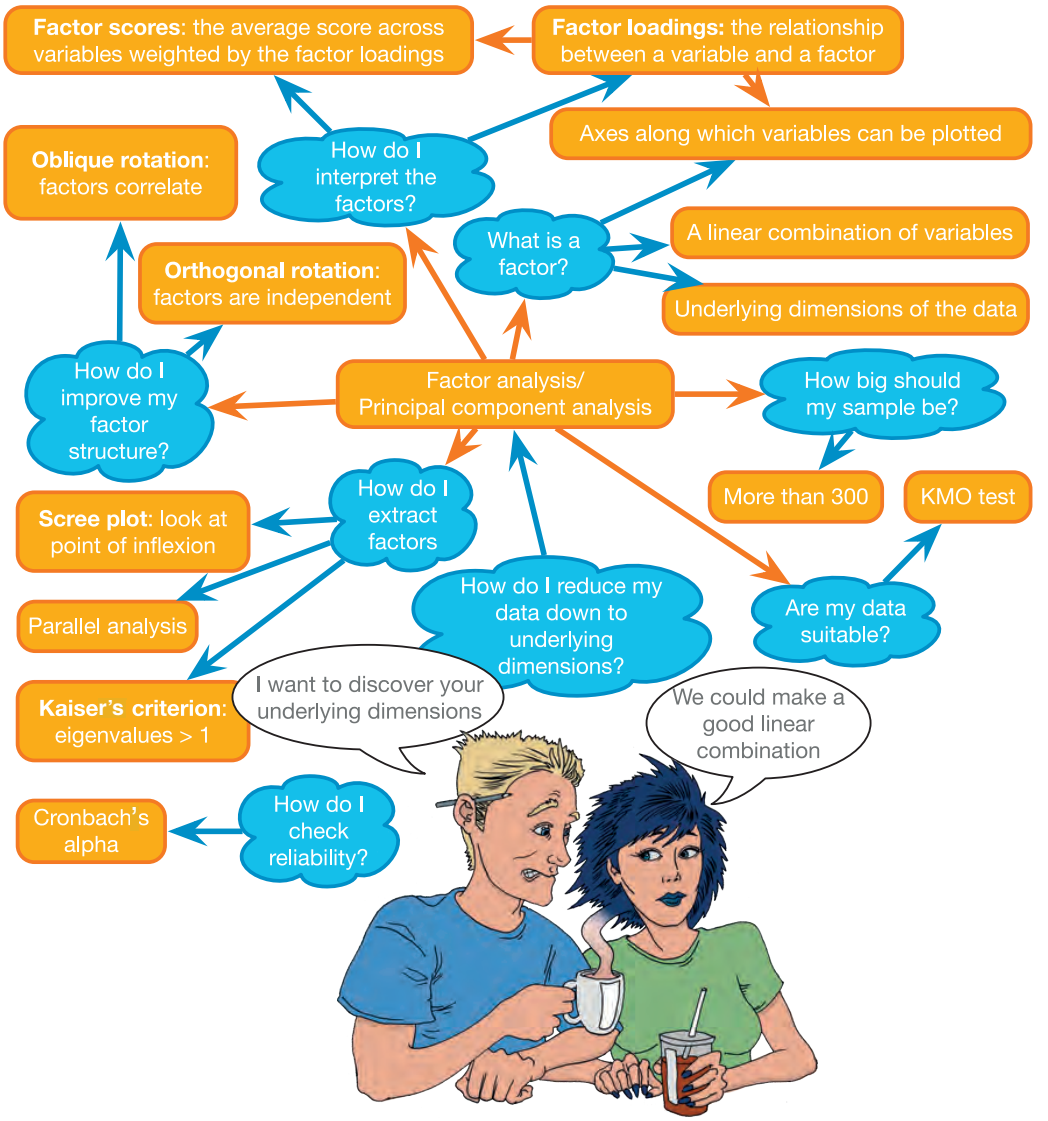


Figure 18.17 What Brian learnt from this chapter

Brian woke too soon. He felt disorientated. Why was he in a sleeping bag? Where was his duvet? As the previous night came back to him, he raced clumsily into the lounge. Neatly folded on the sofa was his duvet, with the pillow on top. Jane had gone. He went to the bathroom. The shower was wet, his towel neatly folded on the floor. It smelt of his coconut shampoo. Back in the lounge, he noticed cereals spilled on the breakfast bar, but as he got closer he realized they were carefully arranged into a word. 'Wait,' it said, with a Rice Krispie smile.

He took a shower and emerged to find Jane at the breakfast bar with two coffees and bagels. She looked down at the towel covering his waist, and with a knowing smile, said, 'Put some clothes on, breakfast is getting cold.'

18.14 What next? Ⓑ

At the age of 23 I took it upon myself to become a living homage to the digestive system. I furiously devoured articles and books on statistics (some of them I even understood), I mentally chewed over them, I broke them down with the stomach acid of my intellect, I stripped them of their goodness and nutrients, I compacted them down, and after about two years I forced the remnants of those intellectual meals out of me in the form of a book. I was mentally exhausted at the end of it; 'It's a good job I'll never have to do that again,' I thought.

18.15 Key terms that I've discovered

Alpha factoring
Anderson–Rubin method
Common factor
Common variance
Communality
Component matrix
Confirmatory factor analysis
Cronbach's α
Direct oblimin
Equamax
Extraction
Factor analysis
Factor loading
Factor matrix
Factor scores
Factor transformation matrix, Λ
Intraclass correlation coefficient (ICC)
Kaiser's criterion
Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy
Latent variable
Oblique rotation
Orthogonal rotation
Parallel analysis
Pattern matrix
Principal component analysis (PCA)
Promax
Quartimax
Random variance
Rotation
Scree plot
Singularity
Split-half reliability
Structure matrix
Unique factor
Unique variance
Varimax



Smart Alex's tasks

- **Task 1**: Rerun the analysis in this chapter using principal component analysis and compare the results to those in the chapter. (Set the iterations to convergence to 30.) Ⓑ

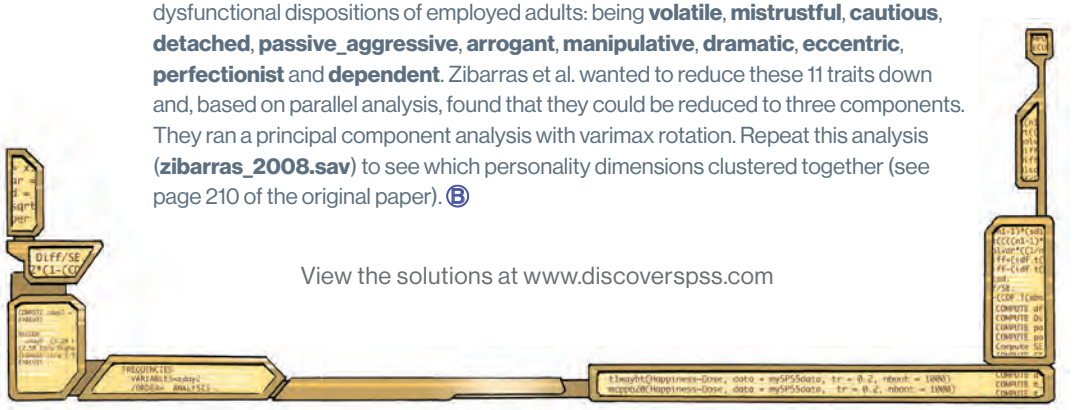
Teaching of Statistics for Scientific Experiments — Revised (TOSSE-R)

		SD	D	N	A	SA
1.	I once woke up in a vegetable patch hugging a turnip that I'd mistakenly dug up thinking it was Roy's largest root	○	○	○	○	○
2.	If I had a big gun I'd shoot all the students I have to teach	○	○	○	○	○
3.	I memorize probability values for the *F*-distribution	○	○	○	○	○
4.	I worship at the shrine of Pearson	○	○	○	○	○
5.	I still live with my mother and have little personal hygiene	○	○	○	○	○
6.	Teaching others makes me want to swallow a large bottle of bleach because the pain of my burning oesophagus would be light relief in comparison	○	○	○	○	○
7.	Helping others to understand sums of squares is a great feeling	○	○	○	○	○
8.	I like control conditions	○	○	○	○	○
9.	I calculate 3 ANOVAs in my head before getting out of bed	○	○	○	○	○
10.	I could spend all day explaining statistics to people	○	○	○	○	○
11.	I like it when I've helped people to understand factor rotation	○	○	○	○	○
12.	People fall asleep as soon as I open my mouth to speak	○	○	○	○	○
13.	Designing experiments is fun	○	○	○	○	○
14.	I'd rather think about appropriate dependent variables than go to the pub	○	○	○	○	○
15.	I soil my pants with excitement at the mere mention of factor analysis	○	○	○	○	○
16.	Thinking about whether to use repeated or independent measures thrills me	○	○	○	○	○
17.	I enjoy sitting in the park contemplating whether to use participant observation in my next experiment	○	○	○	○	○
18.	Standing in front of 300 people in no way makes me lose control of my bowels	○	○	○	○	○
19.	I like to help students	○	○	○	○	○
20.	Passing on knowledge is the greatest gift you can bestow on an individual	○	○	○	○	○
21.	Thinking about Bonferroni corrections gives me a tingly feeling in my groin	○	○	○	○	○
22.	I quiver with excitement when thinking about designing my next experiment	○	○	○	○	○
22.	I often spend my spare time talking to the pigeons ... and even they die of boredom	○	○	○	○	○
23.	I tried to build myself a time machine so that I could go back to the 1930s and follow Fisher around on my hands and knees licking the floor on which he'd just trodden	○	○	○	○	○
25.	I love teaching	○	○	○	○	○
26.	I spend lots of time helping students	○	○	○	○	○
27.	I love teaching because students have to pretend to like me or they'll get bad marks	○	○	○	○	○
28.	My cat is my only friend	○	○	○	○	○

Figure 18.18 The TOSSE-R questionnaire

- **Task 2**: The University of Sussex constantly seeks to employ the best people possible as lecturers. They wanted to revise the 'Teaching of Statistics for Scientific Experiments' (TOSSE) questionnaire, which is based on Bland's theory which says that good research methods lecturers should have: (1) a profound love of statistics; (2) an enthusiasm for experimental design; (3) a love of teaching; and (4) a complete absence of normal interpersonal skills. These characteristics should be related (i.e., correlated). The University revised this questionnaire to become the 'Teaching of Statistics for Scientific Experiments – Revised' (TOSSE–R; Figure 18.18). They gave this questionnaire to 239 research methods lecturers to see if it supported Bland's theory. Conduct a factor analysis (with appropriate rotation) and interpret the factor structure (**tosser.sav**). Ⓑ
- **Task 3:** Dr Sian Williams (University of Brighton) devised a questionnaire to measure organizational ability. She predicted five factors to do with organizational ability: (1) preference for organization; (2) goal achievement; (3) planning approach; (4) acceptance of delays; and (5) preference for routine. These dimensions are theoretically independent. Williams's questionnaire contains 28 items using a seven-point Likert scale (1 = strongly disagree, 4 = neither, 7 = strongly agree). She gave it to 239 people. Run a principal component analysis on the data in **williams.sav**. Ⓑ
- **Task 4**: Zibarras et al. (2008) looked at the relationship between personality and creativity. They used the Hogan Development Survey (HDS), which measures 11 dysfunctional dispositions of employed adults: being **volatile**, **mistrustful**, **cautious**, **detached**, **passive_aggressive**, **arrogant**, **manipulative**, **dramatic**, **eccentric**, **perfectionist** and **dependent**. Zibarras et al. wanted to reduce these 11 traits down and, based on parallel analysis, found that they could be reduced to three components. They ran a principal component analysis with varimax rotation. Repeat this analysis (**zibarras_2008.sav**) to see which personality dimensions clustered together (see page 210 of the original paper). Ⓑ

View the solutions at www.discoverspss.com





CATEGORICAL OUTCOMES: CHI-SQUARE AND LOGLINEAR ANALYSIS

19.1 What will this chapter tell me? 884
19.2 Analysing categorical data 885
19.3 Associations between two categorical variables 885
19.4 Associations between several categorical variables: loglinear analysis 895
19.5 Assumptions when analysing categorical data 897
19.6 General procedure for analysing categorical outcomes 898
19.7 Doing chi-square using SPSS 899
19.8 Interpreting the chi-square test 902
19.9 Loglinear analysis using SPSS 910
19.10 Interpreting loglinear analysis 913
19.11 Reporting the results of loglinear analysis 919
19.12 Jane and Brian's story 920
19.13 What next? 921
19.14 Key terms that I've discovered 922
Smart Alex's tasks 922

Letters A to D indicate increasing difficulty

19.1 What will this chapter tell me? Ⓐ

We discovered in the previous chapter that I wrote a book. This book. There are a lot of good things about writing books. For one thing, your parents are impressed. They're not *that* impressed, because they think that a successful book sells as many copies as *Harry Potter* and that people should queue outside bookshops for the latest enthralling instalment of *Discovering Statistics* … My parents are, consequently, quite baffled about how this book is perceived as successful, yet I don't get invited to dinner by the King. Nevertheless, given that my family don't really understand what I do, books are tangible proof that I do *something*. The size of this book and the fact it has equations in it is a bonus because it makes me look cleverer than I actually am. But not as clever as 'the clever one'.☺ The downside to writing books is the immeasurable mental anguish. In England we don't talk about our emotions, because we fear that if they get out into the open, civilization will collapse. I'm going to buck my national trend and reveal that the writing process for the second edition of this book was so stressful that I came within one of Fuzzy's whiskers of a total meltdown. It took me two years to recover, just in time to start thinking about the third edition. The pain is worth it when people tell me that they found the book vaguely useful. Of course, the publishers focus less on warm fuzzy feelings of altruism and more on sales figures and comparisons with other books. They collate data on sales figures for this book and its competitors in different 'markets' (you are not a person, you are a 'consumer', and you don't live in a country, you live in a 'market') and they gibber and twitch at their consoles creating pink frequency distributions (with 3-D effects) of these values. The data they get are frequency data (the number of books sold in a certain market or discipline). If they wanted to compare sales of this book to its competitors, in different countries, they would need to read this chapter because it's all about analysing data for which we know only the frequency with which events occur. They won't read this chapter, but they should …



Figure 19.1 Midway through writing the second edition of this book, things became a little strange

19.2 Analysing categorical data Ⓐ

So far we have looked at fitting linear models with categorical predictor variables, but always predicting a continuous outcome variable. Sometimes we want to predict categorical outcome variables. In other words, we want to predict into which category an entity falls. For example, we might want to predict whether someone is pregnant or not, for which political party a person will vote, whether a tumour is benign or malignant, whether a sports team will win, lose or draw. In these examples, an entity can fall into only one category, for example a woman is either pregnant or not and a team can't win *and* lose the same match. This chapter and the next deal with statistical models for categorical outcomes. We'll begin with modelling associations between categorical variables, then look at predicting categorical outcomes from categorical predictors, then in the next chapter we'll move on to look at predicting categorical outcomes from both categorical and continuous predictor variables.

19.3 Associations between two categorical variables Ⓐ

Let's first look at the simplest situation of quantifying the relationship between two categorical variables. With categorical variables we can't use the mean or any similar statistic because the mean of a categorical variable is meaningless: the numeric values you attach to different categories are arbitrary, and the mean of those numeric values will depend on how many members each category has. Therefore, when we've measured only categorical variables, we analyse the number of things that fall into each combination of categories (i.e., the frequencies).

19.3.1 A furry example Ⓐ

A researcher was interested in whether animals could be trained to dance. He took 200 cats and tried to train them to dance by giving them either food or affection as a reward for dance-like behaviour. At the end of the week he counted how many animals could dance and how many could not. There are two categorical variables here: **training** (the animal was trained using either food or affection, not both) and **dance** (the animal either learned to dance or it did not). By combining categories, we end up with four different categories. All we then need to do is to count how many cats fall into each category. Table 19.1 shows a contingency table (Section 3.7.5) of these data.

Table 19.1 Contingency table showing how many cats will dance after being trained with different rewards

		Training		
		Food as reward	Affection as reward	Total
Could they dance?	Yes	28	48	76
	No	10	114	124
	Total	38	162	200

19.3.2 Pearson's chi-square test Ⓐ

To see whether there's a relationship between two categorical variables (i.e., does the number of cats that dance relate to the type of training used?) we can use Pearson's **chi-square test** (Fisher, 1922; K.

Pearson, 1900). This statistic is based on the idea of comparing the frequencies you observe in certain categories to the frequencies you might expect to get in those categories by chance. We saw in Chapter 2, (Eq. 2.17) that if we want to calculate the fit (or total error) of a model we add up the squared differences between the observed values of the outcome and the predicted values that come from the model:

$$\text{Total error} = \sum_{i=1}^{n}(\text{observed}_i - \text{model}_i)^2 \tag{19.1}$$

This equation was the basis of our sums of squares in the linear model. We use essentially the same equation when variables are categorical. There is a slight variation in that we divide by the model scores as well, which is much the same process as dividing the sum of squares by the degrees of freedom to get the mean squares: basically, it standardizes the deviation for each observation. If we add these standardized deviations together the resulting statistic is Pearson's chi-square (χ^2) given by:

$$\chi^2 = \sum \frac{(\text{observed}_{ij} - \text{model}_{ij})^2}{\text{model}_{ij}} \tag{19.2}$$

in which i represents the rows in the contingency table and j represents the columns.

The observed data are the frequencies in Table 19.1, but what is the model? When we predict a continuous outcome from categorical predictors (e.g., the linear model) the model we use is group means, but we can't work with means when we have a categorical outcome variable (see above) so we work with frequencies instead. We use 'expected frequencies'. A simple way to estimate the expected frequencies would be to say 'we've got 200 cats in total, and four categories, so the expected value is 200/4 = 50'. This approach would be fine if, for example, we had the same number of cats that had affection as a reward as we did cats that had food as a reward, but we didn't: 38 got food and 162 got affection as a reward. Likewise, there are not equal numbers that could and couldn't dance. To adjust for these inequalities, we calculate expected frequencies for each cell in the table using the column and row totals for that cell. By doing so we factor in the total number of observations that *could have* contributed to that cell. The following equation, in which n is the total number of observations (in this case 200), shows this process:

$$\text{model}_{ij} = E_{ij} = \frac{\text{Row total}_i \times \text{Column total}_j}{n} \tag{19.3}$$

We can calculate these expected frequencies for the four cells within our table as follows (where row total and column total are abbreviated to RT and CT, respectively):

$$\begin{aligned}
\text{model}_{\text{Food, Yes}} &= \frac{\text{RT}_{\text{Yes}} \times \text{CT}_{\text{Food}}}{n} = \frac{76 \times 38}{200} = 14.44 \\
\text{model}_{\text{Food, No}} &= \frac{\text{RT}_{\text{No}} \times \text{CT}_{\text{Food}}}{n} = \frac{124 \times 38}{200} = 23.56 \\
\text{model}_{\text{Affection, Yes}} &= \frac{\text{RT}_{\text{Yes}} \times \text{CT}_{\text{Affection}}}{n} = \frac{76 \times 162}{200} = 61.56 \\
\text{model}_{\text{Affection, No}} &= \frac{\text{RT}_{\text{No}} \times \text{CT}_{\text{Affection}}}{n} = \frac{124 \times 162}{200} = 100.44
\end{aligned} \tag{19.4}$$

These are the model values that we put into Eq. 19.2.

We now have the model values, and the observed values are in Table 19.1. All we need to do is take each value in each cell of Table 19.1, subtract from it the corresponding model value, square the result and then divide by the corresponding model value. Once we've done this for each cell we add them up:

$$\begin{aligned}\chi^2 &= \frac{(28-14.44)^2}{14.44} + \frac{(10-23.56)^2}{23.56} + \frac{(48-61.56)^2}{61.56} + \frac{(114-100.44)^2}{100.44} \\ &= \frac{13.56^2}{14.44} + \frac{-13.56^2}{23.56} + \frac{-13.56^2}{61.56} + \frac{13.56^2}{100.44} \\ &= 12.73 + 7.80 + 2.99 + 1.83 \\ &= 25.35\end{aligned} \tag{19.5}$$

This statistic has (approximately) a distribution with known properties called the **chi-square distribution**, which has a shape determined by the degrees of freedom which are $(r - 1)(c - 1)$, in which r is the number of rows and c is the number of columns. Another way to think of it is the number of levels of each variable minus one multiplied together. In this example we get $df = (2 - 1)(2 - 1) = 1$.

In the days when people did these things by hand, they would get their pet diplodocus to find a critical value for the chi-square distribution with (for these data) $df = 1$. If the value of the observed chi-square statistic was equal to or bigger than this critical value they would conclude that there was a significant relationship between the two variables, rub some sticks together to make a fire and invite their friends over to celebrate. For those of you still living in caves, critical values are in the Appendix; and for $df = 1$ the critical values are 3.84 ($p = 0.05$) and 6.63 ($p = 0.01$), and because the observed chi-square is bigger than these values it is significant at $p < 0.01$. For the rest of us, we can get a computer to compute the precise probability of obtaining a chi-square statistic at least as big as (in this case) 25.35 if there were no association between the variables in the population.

19.3.3 Fisher's exact test Ⓐ

The chi-square statistic has a sampling distribution that is only *approximately* a chi-square distribution. The larger the sample is, the better this approximation becomes, and in large samples the approximation is good enough to not worry about the fact that it is an approximation. In small samples, the approximation is not good enough, making significance tests of the chi-square statistic inaccurate. This is why you'll often read about the chi-square test needing expected frequencies in each cell to be greater than 5 (see Section 19.5). When the expected frequencies are greater than 5, the sampling distribution is probably close enough to a chi-square distribution for us not to worry. However, when the expected frequencies are too low, it probably means that the sampling distribution of the test statistic is too deviant from a chi-square distribution to be accurate.

Fisher came up with a solution to this problem called **Fisher's exact test** (Fisher, 1922). It's not a test as such, it's a way to compute the exact probability of the chi-square statistic in small samples. This procedure is normally used on 2 × 2 contingency tables (i.e., two variables each with two options) and with small samples. It *can* be used on larger contingency tables and with large samples, but there's no point because it was designed to overcome the problem of small samples, and in larger contingency tables your computer might have a meltdown trying to compute it.

19.3.4 The likelihood ratio Ⓑ

An alternative to Pearson's chi-square is the likelihood ratio statistic, which is based on maximum likelihood theory. We discussed maximum likelihood in Section 2.7. The general idea is to (1) collect data and create a model for which the probability of obtaining the observed set of data is maximized; (2) compare this model to the probability of obtaining those data under the null hypothesis. The resulting statistic is based on comparing observed frequencies with those predicted by the model. The computation is:

$$L\chi^2 = 2\sum \text{observed}_{ij} \ln\left(\frac{\text{observed}_{ij}}{\text{model}_{ij}}\right) \tag{19.6}$$

in which *i* and *j* are the rows and columns of the contingency table and *ln* is the natural logarithm (a standard mathematical function that we came across in Chapter 6). Using the same model and observed values as in the previous section, we get a likelihood ratio of 24.94:

$$\begin{aligned} L\chi^2 &= 2\left[28 \times \ln\left(\frac{28}{14.44}\right) + 10 \times \ln\left(\frac{10}{23.56}\right) + 48 \times \ln\left(\frac{48}{61.56}\right) + 114 \times \ln\left(\frac{114}{100.44}\right)\right] \\ &= 2\left[28 \times 0.662 + 10 \times -0.857 + 48 \times -0.249 + 114 \times 0.127\right] \\ &= 2\left[18.54 - 8.57 - 11.94 + 14.44\right] \\ &= 24.94 \end{aligned} \tag{19.7}$$

As with Pearson's chi-square, this statistic has a chi-square distribution with the same degrees of freedom (in this case 1). We would use the same critical values as before and again conclude that the test statistic is significant because the observed value of 24.94 is bigger than the critical values of 3.84 ($p = 0.05$) and 6.63 ($p = 0.01$). A computer would give us a precise *p*-value. For large samples the likelihood ratio will be roughly the same as Pearson's chi-square, but is preferred when samples are small.

19.3.5 Yates's correction Ⓑ

When you have a 2 × 2 contingency table (i.e., two categorical variables each with two categories), Pearson's chi-square tends to produce significance values that are too small (it tends to make a Type I error). Yates suggested a correction to the Pearson formula (usually referred to as **Yates's continuity correction**). The basic idea is that when you calculate the deviation from the model (the $\text{observed}_{ij} - \text{model}_{ij}$ in Eq. 19.2) you subtract 0.5 from the absolute value of this deviation before you square it. Put simply, you calculate the deviation, ignore whether it is positive or negative, subtract 0.5 from it and then square it. With Yates's correction applied the Pearson's equation becomes:

$$\chi^2 = \sum \frac{\left(\left|\text{observed}_{ij} - \text{model}_{ij}\right| - 0.5\right)^2}{\text{model}_{ij}} \tag{19.8}$$

For the data in our example this translates into a value of 23.52:

$$\begin{aligned} \chi^2 &= \frac{(13.56 - 0.5)^2}{14.44} + \frac{(13.56 - 0.5)^2}{23.56} + \frac{(13.56 - 0.5)^2}{61.56} + \frac{(13.56 - 0.5)^2}{100.44} \\ &= 11.81 + 7.24 + 2.77 + 1.70 \\ &= 23.52 \end{aligned} \tag{19.9}$$

Note that the correction lowers the value of the chi-square statistic and, therefore, makes it less significant. There is a fair bit of evidence that this adjustment overcorrects and produces chi-square values that are too small. Howell (2012) provides an excellent discussion, if you're interested; all I will say is that although the correction is worth knowing about, it's probably best ignored.

19.3.6 Other measures of association Ⓐ

There are measures of the strength of association that modify the chi-square statistic to take account of sample size and degrees of freedom and try to restrict the range of the test statistic from 0 to 1 (to make them similar to the correlation coefficient described in Chapter 8). Three related measures are:

- *Phi*: This statistic is accurate for 2 × 2 contingency tables. However, for tables with more than two dimensions the value of phi may not lie between 0 and 1 because the chi-square value can exceed the sample size. Therefore, Pearson suggested the use of the contingency coefficient.
- *Contingency coefficient*: This coefficient ensures a value between 0 and 1 but, unfortunately, it seldom reaches its upper limit of 1 and for this reason Cramér devised an alternative denoted by *V*.
- Cramér's *V*: When both variables have only two categories, phi and Cramér's *V* are identical. However, when variables have more than two categories Cramér's statistic can attain its maximum of 1 (unlike the other two) and so it is the most useful.

19.3.7 The chi-square test as a linear model Ⓓ

As with all the models in this book, the chi-square test can be conceptualized as a general linear model. The general linear model is expressed as:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} + \varepsilon_i \quad (19.10)$$

which is an equation we have seen several times throughout this book (e.g., Chapter 9). We've also seen that this linear model is perfectly capable of accommodating categorical predictor variables. For example, in Chapter 12 when we wanted to compare the means of puppy therapy groups we included these groups using categorical dummy variables:

$$\text{Happiness}_i = b_0 + b_1\text{Long}_i + b_2\text{Short}_i + \varepsilon_i \quad (19.11)$$

With a categorical outcome variable we can use essentially the same model. Let's see how. In our dancing cat example we have two categorical variables: training (food or affection) and dance (yes they did dance or no they didn't dance). Both variables have two categories and we can represent each one with a single dummy variable (see Section 12.3.1) in which one category is coded as 0 and the other as 1. Let's code training as 0 for food and 1 for affection, and dancing as 1 for yes and 0 for no (see Table 19.2).

Table 19.2 Coding scheme for dancing cats

Training	Dance	Dummy (Training)	Dummy (Dance)	Interaction	Frequency
Food	No	0	0	0	10
Food	Yes	0	1	0	28
Affection	No	1	0	0	114
Affection	Yes	1	1	1	48

This situation is like the factorial design that we looked at in Section 14.3: in that example, we also had two variables as predictors and the general linear model became:

$$\text{Outcome}_i = b_0 + b_1\text{A}_i + b_2\text{B}_i + b_3\text{AB}_i + \varepsilon_i \quad (19.12)$$

in which *A* represents the first variable, *B* represents the second and *AB* represents the interaction between the two variables (think back to Eq. 14.4). Therefore, we can estimate the linear model from our sample using the dummy variables in Table 19.2 that is like the one we used for factorial designs:

$$Y_{ij} = b_0 + \hat{b}_0 + \hat{b}_1\text{training}_i + \hat{b}_2\text{dance}_j + \hat{b}_3\text{interaction}_{ij} + e_{ij} \quad (19.13)$$

The interaction term will be the training variable multiplied by the dance variable (look at Section 11.3.4 and if it doesn't make sense look back to Section 14.3 because the coding is the same as this

example). The next issue is that because the *outcome variable* is categorical the assumption of linearity will be broken. We can fix this by transforming the outcome variable to log values (which also affects the error term).[1] The model becomes:[2]

$$\ln(Y_{ij}) = \hat{b}_0 + \hat{b}_1\text{training}_i + \hat{b}_2\text{dance}_j + \hat{b}_3\text{interaction}_{ij} + \ln(e_{ij}) \quad (19.14)$$

The predicted values of the outcome are described as (note the error term is zero):

$$\ln(\hat{Y}_{ij}) = \hat{b}_0 + \hat{b}_1\text{training}_i + \hat{b}_2\text{dance}_j + \hat{b}_3\text{interaction}_{ij} \quad (19.15)$$

The training and dance variables and the interaction can take the values 0 and 1, depending on which combination of categories we're looking at (Table 19.2). Therefore, to work out what the *b*-values represent in this model we can do the same as we did for other linear models and look at what happens when we replace training and dance with different values of 0 and 1.

To begin with, let's see what happens when we look at when training and dance are both zero. This situation represents the category of cats that got food reward and didn't dance. When we had a continuous outcome predicted from categorical predictors, the predicted or expected value of the outcome was the observed mean of the category to which an entity belonged (see, for example, Sections 10.4 and 12.3.2). With a categorical outcome, the predicted value ($\hat{Y}_{ij}$) is the observed frequency in the category, which we'll represent as O_{ij}. Therefore, $\ln(\hat{Y}_{ij}) = \ln(O_{ij})$ and

$$\ln(O_{ij}) = \hat{b}_0 + \hat{b}_1\text{training}_i + \hat{b}_2\text{dance}_j + \hat{b}_3\text{interaction}_{ij} \quad (19.16)$$

In Table 19.1 there were $O_{ij} = 10$ cats that had food for a reward and didn't dance, and the training, dance and interaction variables are all 0 for this combination of categories. Placing these values into the equation gives us:

$$\begin{aligned} \ln(O_{\text{Food, No}}) &= \hat{b}_0 + (\hat{b}_1 \times 0) + (\hat{b}_2 \times 0) + (\hat{b}_3 \times 0) \\ &= \hat{b}_0 \\ \ln(10) &= \hat{b}_0 \\ \hat{b}_0 &= 2.303 \end{aligned} \quad (19.17)$$

Therefore, $\hat{b}_0$ in the model represents the log of the observed value when all categories are zero: it's the log of the observed value of the base category (in this case cats that got food and didn't dance).

Now, let's see what happens when we look at cats that had affection as a reward and didn't dance. In this case, the training variable is 1 and the dance variable and the interaction are still 0. Also, our predicted value of the outcome is the observed number of cats that received affection and didn't dance (from Table 19.1 the value is 114). The equation becomes:

$$\begin{aligned} \ln(O_{\text{Affection, No}}) &= \hat{b}_0 + (\hat{b}_1 \times 1) + (\hat{b}_2 \times 0) + (\hat{b}_3 \times 0) \\ &= \hat{b}_0 + \hat{b}_1 \\ \hat{b}_1 &= \ln(O_{\text{Affection, No}}) - \hat{b}_0 \end{aligned} \quad (19.18)$$

Remembering that $\hat{b}_0$ is the expected value for cats that had food and didn't dance, we get:

1 The parameter estimates will also be in log units as we shall see.

2 The convention is to denote b_0 as θ and the *b*-values as λ, but I think these notational changes serve only to confuse people so I'm sticking with *b* because I want to emphasize the similarities to other linear models in this book.

$$\begin{aligned}\hat{b}_1 &= \ln(O_{\text{Affection, No}}) - \ln(O_{\text{Food, No}}) \\ &= \ln(114) - \ln(10) \\ &= 4.736 - 2.303 \\ &= 2.433\end{aligned} \tag{19.19}$$

The important thing is that $\hat{b}_1$ is the difference between the log of the observed number of cats that received affection and didn't dance, and the log of the observed number of cats that received food and didn't dance. Put another way, within the group of cats that didn't dance it represents the difference between those trained using food and those trained using affection.

When we look at cats that had food as a reward and danced the training variable is 0, the dance variable is 1 and the interaction is 0. Our expected value is the observed number of cats that received food and danced (from Table 19.1 the value is 28). The equation becomes:

$$\begin{aligned}\ln(O_{\text{Food, Yes}}) &= \hat{b}_0 + (\hat{b}_1 \times 0) + (\hat{b}_2 \times 1) + (\hat{b}_3 \times 0) \\ &= \hat{b}_0 + \hat{b}_2 \\ \hat{b}_2 &= \ln(O_{\text{Food, Yes}}) - \hat{b}_0\end{aligned} \tag{19.20}$$

Let's replace $\hat{b}_0$ with what we know it represents (the expected value for cats that had food and didn't dance):

$$\begin{aligned}\hat{b}_2 &= \ln(O_{\text{Food, Yes}}) - \ln(O_{\text{Food, No}}) \\ &= \ln(28) - \ln(10) \\ &= 3.332 - 2.303 \\ &= 1.029\end{aligned} \tag{19.21}$$

Therefore, $\hat{b}_2$ is the difference between the log of the observed frequency for cats that received food and danced, and the log of the observed frequency for cats that received food and didn't dance. Put another way, within the group of cats that received food as a reward it represents the difference between cats that didn't dance and those that did.

Finally, let's look at cats that had affection and danced. The training and dance variables are both 1 and the interaction (which is the value of training multiplied by the value of dance) is also 1. We can replace $\hat{b}_0$, $\hat{b}_1$ and $\hat{b}_2$ with what we now know they represent. The expected value is the log of the observed frequency for cats that received affection and danced (this value is 48 – see Table 19.1). Therefore (using the shorthand of A for affection, F for food, Y for yes and N for no), the equation becomes:

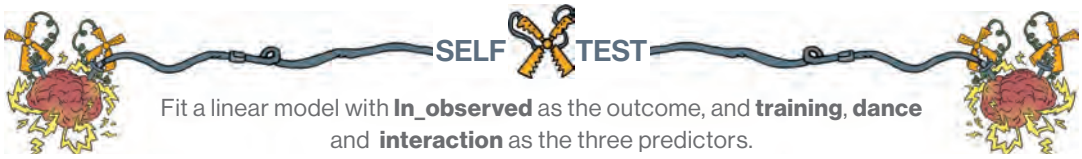
$$\begin{aligned}\ln(O_{A,Y}) &= \hat{b}_0 + (\hat{b}_1 \times 1) + (\hat{b}_2 \times 1) + (\hat{b}_3 \times 1) \\ &= \hat{b}_0 + \hat{b}_1 + \hat{b}_2 + \hat{b}_3 \\ &= \ln(O_{F,N}) + [\ln(O_{A,N}) - \ln(O_{F,N})] + [\ln(O_{F,Y}) - \ln(O_{F,N})] + \hat{b}_3 \\ &= \ln(O_{A,N}) + \ln(O_{F,Y}) - \ln(O_{F,N}) + \hat{b}_3 \\ \hat{b}_3 &= \ln(O_{A,Y}) - \ln(O_{F,Y}) + \ln(O_{F,N}) - \ln(O_{A,N}) \\ &= \ln(48) - \ln(28) + \ln(10) - \ln(114) \\ &= -1.895\end{aligned} \tag{19.22}$$

which shows that $\hat{b}_3$ compares the difference between affection and food when the cats didn't dance to the difference between affection and food when the cats *did* dance. Put another way, it compares the effect of training when cats didn't dance to the effect of training when they did dance.

Putting all these *b*-values together, we get the following model:

$$\ln(O_{ij}) = 2.303 + 2.433\text{training}_i + 1.029\text{dance}_j - 1.895\text{interaction}_{ij} \tag{19.23}$$

The important thing to take from this is that everything is the same as in factorial designs except that we dealt with log-transformed values (compare this section to Section 14.3 to see just how similar everything is). In case you don't believe that the chi-square test works as a general linear model, load **cat_reg.sav**, which contains the variables **dance** (0 = no, 1 = yes) and **training** (0 = food, 1 = affection) and the interaction (**interaction**). There is a variable called **observed** that contains the observed frequencies in Table 19.1 for each combination of **dance** and **training** and a variable called **ln_observed**, which is the natural logarithm of these observed frequencies (remember that throughout this section we've dealt with the log observed values).



Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.303	.000		72046662.8	<.001
	Type of reward	2.434	.000	1.385	73011512.2	<.001
	Did they dance?	1.030	.000	.725	27654265.1	<.001
	Interaction	-1.895	.000	-1.174	-46106002.7	<.001

a. Dependent Variable: Log(observed frequencies)

Output 19.1

Output 19.1 shows the resulting coefficients. Note that the constant, $\hat{b}_1$, is 2.303 as calculated above, the *b*-value for type of training, $\hat{b}_1$, is 2.434 and for dance, $\hat{b}_2$, is 1.030, both of which are within rounding error of what was calculated above. Finally, the coefficient for the interaction, $\hat{b}_3$, is –1.895 as predicted. One weird thing is that the standard errors are all zero: there is *no* error whatsoever in this model. This lack of error is because the various combinations of coding variables completely explain the observed values. This is known as a **saturated model** and I will return to it later, so make a mental note of it.

This is all very well, but the heading of this section did rather imply that I would show you how the chi-square test can be conceptualized as a linear model. Here goes. The chi-square test looks at whether two variables are independent. If variables are independent then their interaction will be zero, therefore, the interaction quantifies the extent to which the variables are dependent. If we remove the interaction term and the model is still a good fit to the data it suggests that the interaction effect is not contributing to the fit to the model, which implies that the dependency between the predictors is small. Put another way, if the interaction doesn't contribute to the model, the predictors are independent. Conversely, if removing the interaction results in a poor fit to the data, it suggests that the interaction term is contributing a lot to the model, which implies the predictors are dependent. If we take the interaction out of the saturated model, and the model becomes:

$$\ln(Y_{ij}) = \hat{b}_0 + \hat{b}_1 \text{training}_i + \hat{b}_2\ \text{dance}_j + \ln(e_{ij}) \qquad (19.24)$$

in which $\ln(Y_{ij})$ is the observed number of cats in each category, which we denoted earlier by $\ln(O_{ij})$. The predicted number of cats, $\ln(\hat{Y}_{ij})$, is

$$\ln(\hat{Y}_{ij}) = \hat{b}_0 + \hat{b}_1 \text{training}_i + \hat{b}_2 \text{ dance}_j \tag{19.25}$$

With this new model, we cannot predict the observed values perfectly (like the saturated model) because we've lost information (namely, the interaction term). Therefore, the outcome from the model changes, and the *b*-values too. We saw earlier that the chi-square test is based on 'expected frequencies'. Our predicted values from the model are, therefore, the log of these expected values:

$$\ln(E_{ij}) = \hat{b}_0 + \hat{b}_1 \text{training}_i + \hat{b}_2 \text{ dance}_j \tag{19.26}$$

We already computed the expected values for this example in Eq. 19.4. We can recalculate the beta values based on these expected values. For cats that had food reward and didn't dance, the training and dance variables will be 0 and the equation becomes:

$$\begin{aligned} \ln(E_{\text{Food, No}}) &= \hat{b}_0 + (\hat{b}_1 \times 0) + (\hat{b}_2 \times 0) \\ &= \hat{b}_0 \\ \hat{b}_0 &= \ln(23.56) \\ &= 3.16 \end{aligned} \tag{19.27}$$

Therefore, $\hat{b}_0$ represents the log of the expected value when all categories are zero.

When we look at cats that had affection as a reward and didn't dance, the training variable is 1 and the dance variable is still 0. Also, the predicted value is the expected number for cats that received affection and didn't dance:

$$\begin{aligned} \ln(E_{\text{Affection, No}}) &= \hat{b}_0 + (\hat{b}_1 \times 1) + (\hat{b}_2 \times 0) \\ &= \hat{b}_0 + \hat{b}_1 \\ \hat{b}_1 &= \ln(E_{\text{Affection, No}}) - \hat{b}_0 \\ &= \ln(E_{\text{Affection, No}}) - \ln(E_{\text{Food, No}}) \\ &= \ln(100.44) - \ln(23.56) \\ &= 1.45 \end{aligned} \tag{19.28}$$

The important thing is that $\hat{b}_1$ is the difference between the log of the expected frequency for cats that received *affection* and didn't dance and the log of the expected values for cats that received *food* and didn't dance. This value is the same as the column marginal, that is, the difference between the total number of cats getting affection and the total number of cats getting food: ln(162) – ln(38) = 1.45. Put simply, it represents the main effect of the type of training.

When we look at cats that had food as a reward and danced, the training variable is 0 and the dance variable is 1. Our predicted outcome is the expected number of cats that received food and danced:

$$\begin{aligned} \ln(E_{\text{Food, Yes}}) &= \hat{b}_0 + (\hat{b}_1 \times 0) + (\hat{b}_2 \times 1) \\ &= \hat{b}_0 + \hat{b}_2 \\ \hat{b}_2 &= \ln(E_{\text{Food, Yes}}) - \hat{b}_0 \\ &= \ln(E_{\text{Food, Yes}}) - \ln(E_{\text{Food, No}}) \\ &= \ln(14.44) - \ln(23.56) \\ &= -0.49 \end{aligned} \tag{19.29}$$

Therefore, $\hat{b}_2$ is the difference between the log of the expected frequencies for cats that received food and did or didn't dance. In fact, the value is the same as the row marginal, that is, the difference between the total number of cats that did and didn't dance: ln(76) – ln(124) = –0.49. In simpler terms, it is the main effect of whether the cat danced.

We can double-check all of this by looking at the final cell (cats that had affection and danced):

$$\begin{aligned}\ln(E_{\text{Affection, Yes}}) &= \hat{b}_0 + (\hat{b}_1 \times 1) + (\hat{b}_2 \times 1) \\ &= \hat{b}_0 + \hat{b}_1 + \hat{b}_2 \\ \ln(61.56) &= 3.16 + 1.45 - 0.49 \\ 4.12 &= 4.12\end{aligned} \tag{19.30}$$

Therefore, the predicted values from the model (using Eq. 19.25) are

$$\begin{aligned}\ln(\hat{Y}_{ij}) &= \hat{b}_0 + \hat{b}_1\text{training}_i + \hat{b}_2\text{dance}_j \\ &= 3.16 + 1.45\text{training} - 0.49\text{dance}\end{aligned} \tag{19.31}$$

Placing this into Eq. 19.24 gives us the final chi-square as:

$$\begin{aligned}\ln(Y_{ij}) &= \hat{b}_0 + \hat{b}_1\text{training}_i + \hat{b}_2\text{dance}_j - \ln(e_{ij}) \\ &= \ln(\hat{Y}_{ij}) - \ln(e_{ij})\end{aligned} \tag{19.32}$$

which is the same as the equation we saw in Section 6.2.4. We can rearrange this equation to get residuals (the error term):

$$\ln(e_{ij}) = \ln(Y_{ij}) - \ln(\hat{Y}_{ij}) \tag{19.33}$$

Remembering that predicted values $\ln(\hat{Y}_{ij})$ are the expected frequencies, and the outcome is the observed number of cats in each category, we can express the residuals as the difference between the observed and expected cell frequencies:

$$\ln(e_{ij}) = \ln(O_{ij}) - \ln(E_{ij}) \tag{19.34}$$



Fit another linear model using **cat_reg.sav**. This time the outcome is the log of expected frequencies (**ln_expected**) and **training** and **dance** are the predictors (the interaction is not included).

The resulting Output 19.2 confirms that the parameter estimates match the values that we just calculated (and summarized in Eq. 19.31). This section demonstrates how chi-square can be thought of as a linear model in which the beta values tell us something about the relative differences in frequencies across categories of our two variables. The take-home point is simply that even with categorical outcomes you're using the same model that you have been learning about throughout this book.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3.160	.000		366695255	<.001
	Type of reward	1.450	.000	.824	173423149	<.001
	Did they dance?	-.490	.000	-.344	-72443567.9	<.001

a. Dependent Variable: Log(expected frequency)

Output 19.2

19.4 Associations between several categorical variables: loglinear analysis ©

Often we want to analyse more complex contingency tables in which there are three or more variables. For example, suppose we took the example we've just used but also collected data from a sample of 70 dogs. We might want to compare the behaviour in dogs to that in cats. We would have three variables: **animal** (dog or cat), **training** (food as reward or affection as reward) and **dance** (did they dance or not?). This couldn't be analysed with the Pearson chi-square and instead has to be analysed with a technique called **loglinear analysis.**

In the previous section, after nearly reducing my brain to even more of a rotting vegetable than it already is trying to explain how categorical data analysis is another form of the linear model, I fitted an ordinary regression to prove that I wasn't talking gibberish. At the time I rather glibly said 'oh, by the way, there's no error in the model, that's odd, isn't it?' and sort of passed this off by telling you that it was a 'saturated' model and not to worry too much about it because I'd explain it all later just as soon as I'd worked out what the hell was going on. That was a good avoidance tactic while it lasted, but I now have to explain what I was going on about.

To begin with, I hope you're now happy with the idea that categorical data can be expressed in the form of a linear model if we use log values (this, incidentally, is why the technique we're discussing is called *log*linear analysis).[3] From what you hopefully already know about linear models generally, you should also be cosily tucked up in bed with the idea that we can extend any linear model to include any number of predictors and any resulting interaction terms between predictors. If we can represent a simple two-variable categorical analysis as a linear model, then it shouldn't amaze you to discover that if we have more than two variables the model extends to include new categorical predictors and their interactions with existing predictors. Each predictor will have a parameter (b). This is all you really need to know. By thinking in terms of a linear model, it then becomes conceptually very easy to understand how the chi-square model expands to incorporate new predictor variables. For example, if we have three predictors (A, B and C) in a linear model we end up with three two-way interactions (AB, AC, BC) and one three-way interaction (ABC):

$$\text{outcome}_{ijk} = b_0 + b_1A_i + b_2B_j + b_3C_k + b_4AB_{ij} + b_5AC_{ik} + b_6BC_{jk} + b_7ABC_{ijk} + \varepsilon_{ij} \qquad (19.35)$$

If the outcome is categorical we get an identical model, but with an outcome expressed as logs:

$$\ln(O_{ijk}) = b_0 + b_1A_i + b_2B_j + b_3C_k + b_4AB_{ij} + b_5AC_{ik} + b_6BC_{jk} + b_7ABC_{ijk} + \ln(\varepsilon_{ij}) \qquad (19.36)$$

The calculation of b-values and expected values from the model becomes considerably more cumbersome and confusing than with just one predictor and one outcome, but that's why we invented computers – so that we don't have to worry about it. Just imagine that a puppy called Tobin collects numbers into a pile and then circles them while chasing his tail. As the numbers get sucked into the vortex that his spinning creates, a value eventually flies out like a leaf floating on the wind and lands on the grass beside him. Loglinear analysis works on these principles. Sort of.

As we saw in the two-variable case, when our outcome is categorical and we include all the available terms (main effects and interactions) we get no error: our predictors perfectly predict the outcome (the expected values). The model is saturated. If we start with this model, we get no error. The job of loglinear analysis is to try to fit a simpler model without any substantial loss of predictive power. Therefore, loglinear analysis typically works on a principle of backward elimination (yes, the same kind of backward elimination as in Section 9.9.1). We begin with the saturated model, remove a

3 The loglinear model is actually one of a more general set of models known as generalized linear models – specifically we're fitting a Poisson generalized linear model.

predictor from the model, re-estimate the model and use it to predict our outcome (calculate expected frequencies, just like the chi-square test) and see how well it fits the data (i.e., are the expected frequencies close to the observed frequencies?). If the fit of the new model is not very different from the more complex model, then we abandon the complex model in favour of the new, simpler, one. Put another way, we assume the term we removed was not having a significant impact on the ability of our model to predict the observed outcome.

We don't remove terms randomly, we do it hierarchically. So, we start with the saturated model, remove the highest-order interaction, and assess the effect that this has. If removing the highest-order interaction term has no substantial impact on the model then we get rid of it and move on to remove the next highest-order interactions. If removing these interactions has no effect then we remove them and so on down to the main effects. We carry on until we find an effect that *does* affect the fit of the model when it is removed.

To put this in more concrete terms, at the beginning of this section I asked you to imagine we'd extended our training and dancing example to incorporate a sample of dogs. So, we now have three variables: **animal** (dog or cat), **training** (food or affection) and **dance** (did they dance or not?). This model has three main effects:

- **animal**
- **training**
- **dance**

three interactions involving two variables:

- **animal** × **training**
- **animal** × **dance**
- **training** × **dance**

and one interaction involving all three variables:

- **animal** × **training** × **dance**

When I talk about backward elimination I mean that loglinear analysis starts by including all these effects. The highest-order interaction (in this case the three-way interaction of **animal** × **training** × **dance**) is removed. A new model is estimated without this interaction, and expected frequencies are computed from this new model. These expected frequencies (or model frequencies) are compared to the observed frequencies using the standard equation for the likelihood ratio statistic (see Section 19.3.4). If the new model significantly changes the likelihood ratio statistic, then removing the interaction term has a significant effect on the fit of the model, which tells us that this effect is statistically important. If this is the case then we stop here and conclude that we have a significant three-way interaction. We won't test any other effects because all lower-order effects are subsumed within higher-order effects. If, however, removing the three-way interaction doesn't significantly affect the fit of the model then we move on to lower-order interactions. Therefore, we look at the **animal** × **training**, **animal** × **dance** and **training** × **dance** interactions in turn and construct models in which these terms are not present. For each model expected frequencies are computed and compared to the observed frequencies using a likelihood ratio statistic.[4] Again, if any one of these models results in a significant change in the likelihood ratio then the term is retained and we won't move on to look at any main effects involved in that interaction (so if the **animal** × **training** interaction is significant the computer won't look at the main

4 It's worth mentioning that for every model, the computation of expected values differs, and as the designs get more complex, the computation gets increasingly tedious and incomprehensible (at least to me); however, you don't need to know the calculations to get a feel for what is going on.

effects of **animal** or **training**). However, if the likelihood ratio is unchanged then the offending interaction term is removed and the computer moves on to look at main effects.

I mentioned that the likelihood ratio statistic (Section 19.3.4) is used to assess each model. This equation (Eq. 19.6) can be adapted to fit any model: the observed values are the same throughout, and the model frequencies are the expected frequencies from the model being tested. For the saturated model, this statistic will always be 0 (because the observed and model frequencies are the same, so the ratio of observed to model frequencies will be 1, and ln(1) = 0), but in other situations it will provide a measure of how well the model fits the observed frequencies. To test whether a new model has changed the likelihood ratio, we take the likelihood ratio for a model and subtract from it the likelihood statistic for the previous model (provided the models are hierarchically structured):

$$L\chi^2_{\text{Change}} = L\chi^2_{\text{Current model}} - L\chi^2_{\text{Previous model}} \tag{19.37}$$

I've tried to give you a flavour of how loglinear analysis works without getting bogged down in the nitty-gritty of the calculations. The curious among you might want to know exactly how everything is calculated, and to these people I have two things to say: 'I don't know' and 'I know a really good place where you can buy a straitjacket'. Tabachnick and Fidell (2012) has a wonderfully detailed and lucid chapter on the subject, unlike me.

19.5 Assumptions when analysing categorical data Ⓐ

The chi-square test does not rely on the assumptions discussed in Chapter 6 (e.g., categorical variables cannot have a normal sampling distribution because they aren't continuous). However, it has two important assumptions relating to (1) independence and (2) expected frequencies.

19.5.1 Independence Ⓐ

The general linear model makes an assumption about the independence of residuals, and the chi-square test, being a linear model of sorts, is no exception. For the chi-square test to be meaningful each person, item or entity must contribute to only one cell of the contingency table. Therefore, you cannot use a chi-square test on a repeated-measures design (e.g., if we had trained some cats with food to see if they would dance and then trained the same cats with affection to see if they would dance we couldn't analyse the resulting data with Pearson's chi-square test). If you find yourself in this situation you need to get yourself a good book about generalized linear mixed models because you effectively need to fit a variant of a multilevel model (Chapter 21) for categorical outcomes.

19.5.2 Expected frequencies Ⓐ

With 2 × 2 contingency tables (i.e., two categorical variables both with two categories) no expected values should be below 5. In larger tables, and when looking at associations between three or more categorical variables (loglinear analysis), the rule is that all expected counts should be greater than 1 and no more than 20% of expected counts should be less than 5. Howell (2012) gives a nice explanation of why violating this assumption creates problems. If this assumption is broken the result is a radical reduction in test power – so dramatic in fact that it may not be worth bothering with the analysis at all.

In terms of remedies, if you're looking at associations between only two variables then consider using Fisher's exact test (Section 19.3.3). With three or more variables (i.e., loglinear analysis) your options are to: (1) collapse the data across one of the variables (preferably the one you least expect to have an effect); (2) collapse levels of one of the variables; (3) collect more data; or (4) accept the loss of power. If you want to collapse data across one of the variables then:

1 The highest-order interaction should be non-significant.
2 At least one of the lower-order interaction terms involving the variable to be deleted should be non-significant.

Let's think about our loglinear example in which we're looking at the relationship between training (food vs. affection), whether the animal danced (yes vs. no), and the species of animal (cats vs. dogs). Say we wanted to delete the animal variable; then for this to be valid, the **animal** × **training** × **dance** variable should be non-significant, and either the **animal** × **training** or the **animal** × **dance** interaction should also be non-significant.

You can also collapse categories within a variable. So, if you had a season variable relating to spring, summer, autumn and winter, and you had very few observations in winter, you could consider reducing the variable to three categories: spring, summer, autumn/winter perhaps. However, you should combine only categories for which it makes theoretical sense.

Finally, some people overcome the problem by adding a constant to all cells of the table, but there really is no point in doing this because it doesn't address the issue of power.

19.5.3 More doom and gloom Ⓐ

Finally, although it's not an assumption, it seems fitting to mention in a section in which a gloomy and foreboding tone is being used that proportionately small differences in cell frequencies can result in statistically significant associations between variables if the sample is large enough (although it might need to be very large indeed). Therefore, we must look at row and column percentages to interpret the significant effects that we get. These percentages will reflect the patterns of data far better than the frequencies themselves (because these frequencies will be dependent on the sample sizes in different categories).

19.6 General procedure for analysing categorical outcomes Ⓐ

Figure 19.2 shows a general procedure for analysing data when you want to fit models that have both an outcome and predictor(s) that are categorical. Essentially you first look at a contingency table and

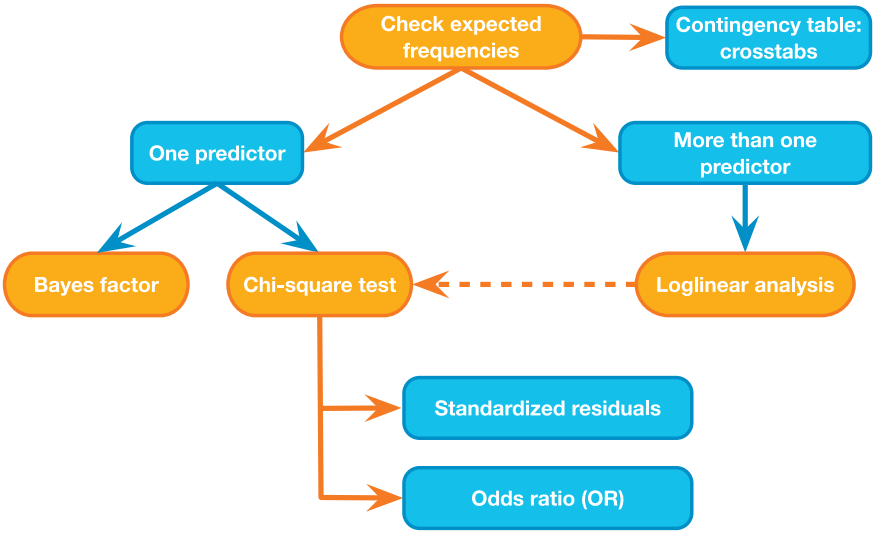


Figure 19.2 The general process for fitting models in which both predictors and the outcome are categorical

check the expected frequencies. If you have one predictor then head straight to a chi-square test or Bayes factor, but if you have more than one predictor first do a loglinear analysis (Section 19.9) and then follow up any significant effects with one or more chi-square tests. After a chi-square test it's useful to inspect the standardized residuals and compute an odds ratio, which is an effect size quantifying the relationship between variables.

19.7 Doing chi-square using SPSS Ⓐ

To begin with, let's imagine we're looking at the data from only cats. We want to input data about whether the 200 cats danced and what type of training they had. There are two ways to do this.

19.7.1 Entering raw scores Ⓐ

The first way is to do what we normally do: enter each cat's data as a row of the data. You would create two coding variables (**training** and **dance**) and, in keeping with Table 19.2, **training** could be coded 0 to represent a food reward and 1 to represent affection, and **dance** could be coded 1 to represent an animal that danced and 0 to represent one that did not. For each animal, you put the appropriate numeric code into each column. For example, a cat that was trained with food that did not dance would have 0 in the training column and 0 in the dance column as in **cats.sav**. Note that there are 200 cats and, therefore, 200 rows of data.

19.7.2 Entering frequencies and weighting cases Ⓐ

The second way to enter data is to create the same coding variables as before, but have a third variable that represents the number of animals that fell into each combination of categories. We could call this variable **frequency**. Instead of having 200 rows, each one representing a different animal, we have one row representing each combination of categories and in the **frequency** variable we enter the number of animals that fell into this category combination. Figure 19.3 shows the data set up in this way: the first row represents cats that had food as a reward and then danced and the value in **frequency** tells us that there were 28 cats that had food as a reward and then danced. Extending this principle, we can see that when affection was used as a reward 114 cats did not dance. Data entered in this way are in the file **cats_weight.sav**.

This method of data entry saves a lot of time, but if you use this method to enter data you must tell the computer that the variable **frequency** represents the number of cases that fell into a particular combination of categories. To do this, select *Data* > Weight Cases... to access the dialog box in Figure 19.4, select Weight cases by and drag the variable in which the number of cases is specified (in this case **frequency**) to the box labelled *Frequency Variable* (or click). IBM SPSS Statistics software will now weight each category combination by the number in the column labelled **frequency**. For example, SPSS will act as though there are 28 rows of data that have the category combination 0, 1 (representing cats trained with food and that danced).

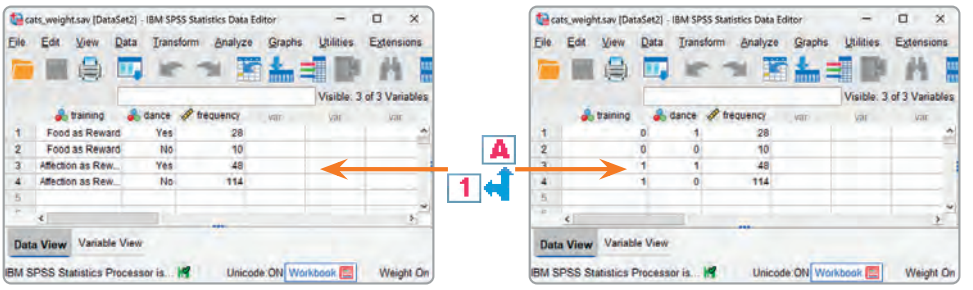


Figure 19.3 Data entry using weighted cases

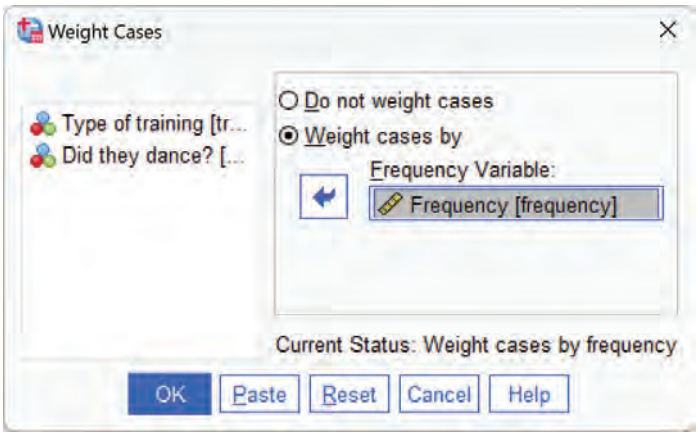


Figure 19.4 The dialog box for the *weight cases* command

19.7.3 Specifying a chi-square test Ⓐ

The first step in Figure 19.2 is to create a contingency table using the *crosstabs* command, check the expected frequencies and then do the chi-square test. We can do these steps simultaneously. Select *Analyze* > *Descriptive Statistics* > Crosstabs... to access the dialog box in Figure 19.5 (the variable **frequency** is shown in the diagram because I ran the analysis on the **cats_weight.sav** data). First, drag

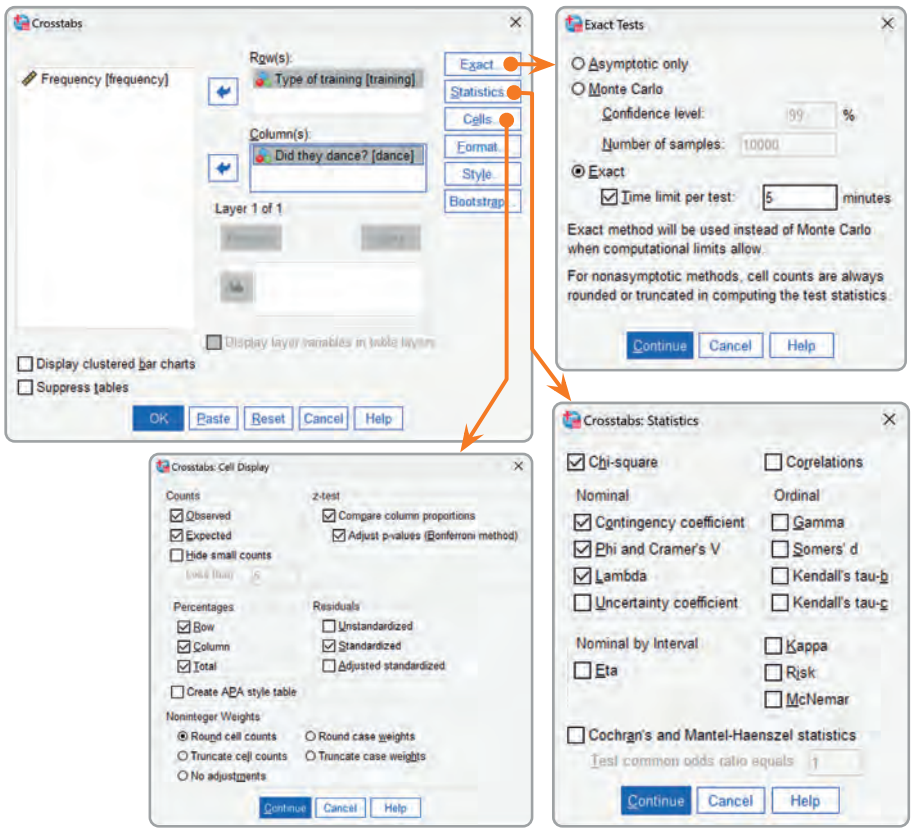
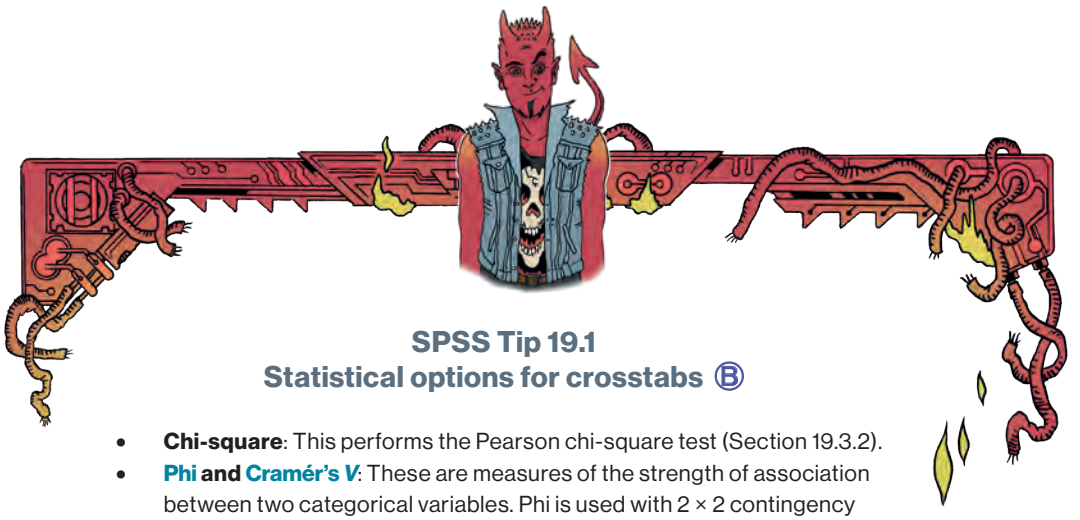
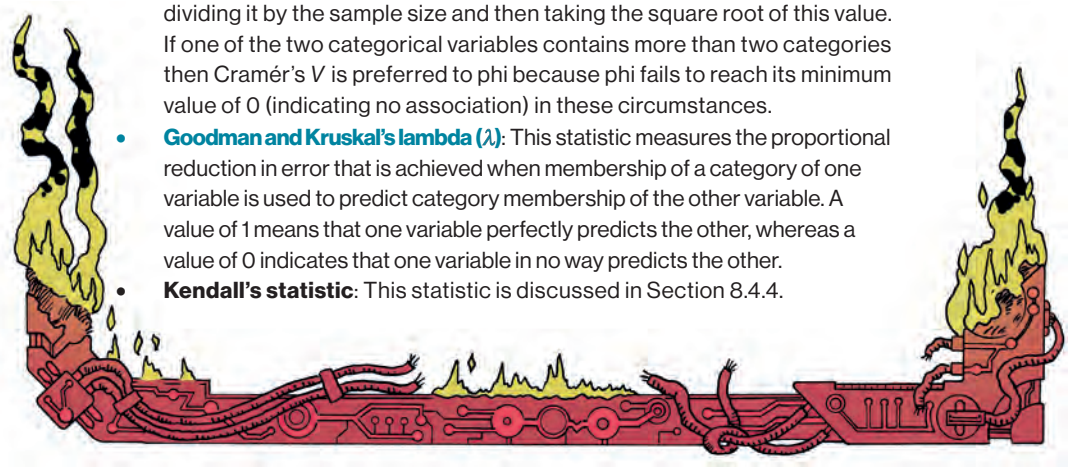


Figure 19.5 Dialog boxes for the *crosstabs* command



SPSS Tip 19.1
Statistical options for crosstabs Ⓑ

- **Chi-square**: This performs the Pearson chi-square test (Section 19.3.2).
- **Phi and Cramér's *V***: These are measures of the strength of association between two categorical variables. Phi is used with 2 × 2 contingency tables (i.e., you have two categorical variables and each variable has only two categories). Phi is calculated by taking the chi-square value and dividing it by the sample size and then taking the square root of this value. If one of the two categorical variables contains more than two categories then Cramér's *V* is preferred to phi because phi fails to reach its minimum value of 0 (indicating no association) in these circumstances.
- **Goodman and Kruskal's lambda (*λ*)**: This statistic measures the proportional reduction in error that is achieved when membership of a category of one variable is used to predict category membership of the other variable. A value of 1 means that one variable perfectly predicts the other, whereas a value of 0 indicates that one variable in no way predicts the other.
- **Kendall's statistic**: This statistic is discussed in Section 8.4.4.



one of the variables of interest (I chose **training**) to the box labelled *Row(s)* (or select it and click ▸). Next, drag the other variable of interest (**dance**) to the box labelled *Column(s)* (or click ▸). It is also possible to select a layer variable (i.e., to split the rows of the table into further categories). If you had a third categorical variable (as we will later in this chapter) you could split the contingency table by this variable (so layers of the table represent different categories of this third variable).

Click Statistics... to specify various statistical tests (see SPSS Tip 19.1). Select the chi-square test, the contingency coefficient, phi and lambda. Click Cells... to specify what values are displayed in the crosstabulation table. Select *Expected* because we use these to check the assumptions about the expected frequencies (Section 19.5). It is also useful to have a look at the row, column and total percentages because these values are usually more easily interpreted than the raw frequencies. There are two options that are useful for breaking down a significant effect (should we get one): (1) a *z*-test that compares cell counts across columns of the contingency table (☑ Compare column proportions), and it's a good idea to ☑ Adjust p-values (Bonferroni method) because there will be multiple tests; and (2) ☑ Standardized residuals. Click Exact... in the main dialog box to compute Fisher's exact test (Section 19.3.3) if your sample is small or if your expected frequencies are too low (see Section 19.5). Even though we don't need it for these data, I have selected the *Exact* test option to show you how it is used. Click Continue to return to the main dialog box and OK to run the analysis.



Oditi's Lantern
Dancing cats

'I, Oditi, want my followers to harness the power of dancing cats. It is a well established fact that a dancing cat creates more energy than nuclear fusion. To solve the mysteries of statistics, we must power thousands of computers, and the only way to generate that kind of power is a stadium of dancing cats. So that you can identify a dancing cat, I have prepared a video of one ... it also shows you how to do the chi-square test. Stare into my lantern and be amazed.'



19.8 Interpreting the chi-square test Ⓐ



The contingency table (Output 19.3) contains the number of cases that fall into each combination of categories. We can see that in total 76 cats danced (38% of the total) and of these 28 were trained using food (36.8% of the total that danced) and 48 were trained with affection (63.2% of the total that danced). Further, 124 cats didn't dance at all (62% of the total) and of those that didn't dance, 10 were trained using food as a reward (8.1% of the total that didn't dance) and a massive 114 were trained using affection (91.9% of the total that didn't dance). The numbers of cats can be read from the rows labelled *Count* and the percentages are read from the rows labelled *% within Did they dance?* We can also look at the percentages within the training categories by looking at the rows labelled *% within Type of training*. This tells us, for example, that of those trained with food as a reward, 73.7% danced and 26.3% did not. Similarly, for those trained with affection, only 29.6% danced compared to 70.4% that didn't. In summary, when food was used as a reward most cats would dance, but when affection was used most cats refused to dance.

First, let's check the expected frequencies (Section 19.5). We have a 2 × 2 table, so all expected frequencies need to be greater than 5. Looking at the expected counts in the contingency table (which incidentally are the same as we calculated earlier), we see that the smallest expected count is 14.4 (for cats that were trained with food and did dance). This value exceeds 5 and so the assumption has been met. If you found an expected count lower than 5, calculate a Bayes factor instead.

If you're wondering why the counts have subscript letters it's because we selected *Compare column proportions* in Figure 19.5. These subscripts tell us the results of the *z*-test that we asked for: columns with different subscripts have significantly different column proportions. It's not immediately obvious what's

being tested here, and to be honest with you it took me a while to fathom it out because I could interpret the SPSS help files in different ways (perhaps it's just me). I got there in the end (I think). We need to look within rows of the table. So, for *Food as Reward* the columns have different subscripts (the count of 10 has a subscript letter *a* and the count of 28 has a subscript letter *b*), which means that proportions within the column variable (i.e., *Did they dance?*) are significantly different. The *z*-test compares the *proportion* of the total frequency of the first column that falls into the first row against the *proportion* of the total frequency of the second column that falls into the first row. The different subscripts tell us that these proportions are significantly different. So, the proportion of cats that danced after food (36.8%) was significantly more than the proportion that *didn't* dance after food (8.1%). The test compares the proportions and *not the counts themselves*, so it is not the case that the count of 28 is different from the count of 10 (in this example). The self-test uses an example to illustrate this point.

Moving on to the row labelled *Affection as Reward*, the count of 114 has a subscript letter *a* and the count of 48 has a subscript letter *b*; as before, the fact they have different letters tells us that the column proportions are significantly different: in other words, 91.9% is significantly different from 63.2%. The proportion of cats that danced after affection was significant less than the proportion that didn't dance after affection.

Type of training * Did they dance? Crosstabulation

			Did they dance? No	Did they dance? Yes	Total
Type of training	Food as reward	Count	10a	28b	38
		Expected Count	23.6	14.4	38.0
		% within Type of training	26.3%	73.7%	100.0%
		% within Did they dance?	8.1%	36.8%	19.0%
		% of Total	5.0%	14.0%	19.0%
		Standardized Residual	-2.8	3.6	
	Affection as reward	Count	114a	48b	162
		Expected Count	100.4	61.6	162.0
		% within Type of training	70.4%	29.6%	100.0%
		% within Did they dance?	91.9%	63.2%	81.0%
		% of Total	57.0%	24.0%	81.0%
		Standardized Residual	1.4	-1.7	
Total		Count	124	76	200
		Expected Count	124.0	76.0	200.0
		% within Type of training	62.0%	38.0%	100.0%
		% within Did they dance?	100.0%	100.0%	100.0%
		% of Total	62.0%	38.0%	100.0%

Each subscript letter denotes a subset of Did they dance? categories whose column proportions do not differ significantly from each other at the .05 level.

Output 19.3



Using the **cats_weight.sav** data, change the frequency of cats that had food as reward and didn't dance from 10 to 28. Redo the chi-square test and select and interpret *z*-tests (☑ Compare column proportions). Is there anything about the results that seems strange?

Output 19.4 shows the chi-square statistic and its significance value. The value of the chi-square statistic is given in the table (together with the degrees of freedom), as is the significance value. The value of the chi-square statistic is 25.356, which is within rounding error of what we calculated in

Section 19.3.2, and this value is highly significant ($p < 0.001$), indicating that the type of training was significantly associated with whether an animal danced.

The table also includes other statistics that you requested in Figure 19.5. *Continuity Correction* is Yates's continuity corrected chi-square (Section 19.3.5), which matches the value we calculated earlier (23.52). This test is probably best ignored. The *Likelihood Ratio* (Section 19.3.4), which we'd prefer to the chi-square test if the sample were small, is within rounding error of the value we calculated (24.93) and is also highly significant ($p < 0.001$). There are several footnotes. The first is a summary of expected counts in case you forgot to check this yourself. We're told that there were no expected frequencies less than 5 so the chi-square statistic should be accurate.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	25.356[a]	1	<.001	<.001	<.001	
Continuity Correction[b]	23.520	1	<.001			
Likelihood Ratio	24.932	1	<.001	<.001	<.001	
Fisher's Exact Test				<.001	<.001	
Linear-by-Linear Association	25.229[c]	1	<.001	<.001	<.001	.000
N of Valid Cases	200					

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 14.44.
b. Computed only for a 2x2 table
c. The standardized statistic is -5.023.

Output 19.4

Output 19.5 contains the measures of association discussed in Section 19.3.6 (if requested). Cramér's statistic is 0.36 out of a possible maximum value of 1, which represents a reasonably strong association between the type of training and whether the cats danced or not (think of it like a correlation coefficient). This value is highly significant ($p < 0.001$), indicating that a value of the test statistic that is at least this big if the null hypothesis were true is unlikely to have happened if there were no association in the population. These results confirm what the chi-square test already told us but also give us an estimate of the effect size.

Symmetric Measures

		Value	Approximate Significance	Exact Significance
Nominal by Nominal	Phi	-.356	<.001	<.001
	Cramer's V	.356	<.001	<.001
	Contingency Coefficient	.335	<.001	<.001
N of Valid Cases		200		

Output 19.5

19.8.1 Using standardized residuals Ⓑ

In a 2 × 2 contingency table like the one in this example, the nature of a significant association can be clear from just the cell percentages or counts. In larger contingency tables this may not be the case and you need a finer-grained investigation of the contingency table. You can think of a significant chi-square test in much the same way as a significant interaction in a linear model: it is an effect that needs to be broken down further. We have already looked at the *z*-tests in the contingency table, but we can also use the standardized residual.

Like in any linear model, the residual is the error between what the model predicts (the expected frequency) and the observed data (the observed frequency):

$$\text{residual}_{ij} = \text{observed}_{ij} - \text{model}_{ij} \tag{19.38}$$

in which *i* and *j* represent the two variables (i.e., the rows and columns in the contingency table). Eq. 19.38 is a friendlier way of expressing Eq. 19.33 from earlier in the chapter. This residual is the same conceptually as every other residual or deviation in this book (compare Eq. 19.38 to, for example, Eq. 2.17).

To standardize this equation, we divide by the square root of the expected frequency:

$$\text{standardized residual} = \frac{\text{observed}_{ij} - \text{model}_{ij}}{\sqrt{\text{model}_{ij}}} \tag{19.39}$$

Does this equation look familiar? It's basically part of Eq. 19.2. The only difference is that rather than looking at squared deviations, we're looking at the pure deviations. Remember that the rationale for squaring deviations is to make them positive so that they don't cancel out when we add them to get the chi-square statistic. If we're not planning to add up the deviations/residuals then we don't need to square them; in fact the direction of the value (plus or minus) is useful information about whether the model over- or underestimates. There are two important things about these standardized residuals:

1. Given that the chi-square statistic is the sum of these standardized residuals (sort of), if we want to decompose what contributes to the overall association that the chi-square statistic measures, then looking at the individual standardized residuals is a good idea because they have a direct relationship with the test statistic.
2. These standardized residuals behave like any other (see Section 6.4): each one is a *z*-score. This is very useful because by looking at a standardized residual we can assess its significance (see Section 1.7.6.): if the value lies outside of ±1.96 then it is significant at $p \leq 0.05$, if it lies outside ±2.58 then it is significant at $p \leq 0.01$, and if it lies outside ±3.29 then it is significant at $p \leq 0.001$.

Because we selected ☑ Standardized in Figure 19.5 the standardized residuals are in Output 19.3. There are four residuals: one for each combination of the type of training and whether the cats danced. When food was used as a reward the standardized residual was significant[5] for both those that danced ($z = 3.6$) and those that didn't dance ($z = -2.8$). The plus or minus sign tells us something about the direction of the effect, as do the counts and expected counts within the cells. We can interpret these standardized residuals as follows. When food was used as a reward significantly more cats than expected danced, and significantly fewer cats than expected did not dance. When affection was used as a reward the standardized residual was not significant[6] for both those that danced ($z = -1.7$) and those that didn't dance ($z = 1.4$). This tells us that when affection was used a reward as many cats as expected danced and did not dance. In a nutshell, the cells for when food was used as a reward both significantly contribute to the overall chi-square statistic: the association between the type of reward and dancing is mainly driven by when food is a reward.

19.8.2 Summary Ⓐ

The highly significant result indicates that there is an association between the type of training and whether the cat danced or not. In other words, the pattern of responses (i.e., the proportion of cats that danced compared to the proportion that did not) in the two training conditions is significantly different. We saw from the earlier *z*-tests that of the cats trained with food a significantly greater proportion danced, and conversely of those trained with affection a significantly greater proportion didn't dance. From the standardized residuals we know that when food is used as a reward, more cats

5 Because both values are larger than 1.96 (when you ignore the minus sign).

6 Because both values are smaller than 1.96 (when you ignore the minus sign).

danced (and fewer did not dance) than expected: about 74% of cats learn to dance and 26% do not. When affection is used, the opposite is true (about 70% refuse to dance and 30% do dance), which is consistent with expected frequencies. We can conclude that the type of training used significantly influences the cats: they will dance for food but not for love. Having lived with a lovely cat for many years, this supports my cynical view that they will do nothing unless there is a bowl of cat food waiting for them at the end of it!



Labcoat Leni's Real Research 19.1
The impact of sexualized images on women's self-evaluations Ⓐ

Daniels, E. A. (2012). *Journal of Applied Developmental Psychology*, 33, 79–90.

Women (and increasingly men) are bombarded with 'idealized' images in the media and there is a growing concern about how these images affect our perceptions of ourselves. Daniels (2012) showed young women images of successful female athletes (e.g., Anna Kournikova) in which they were either playing sport (performance athlete images) or posing in bathing suits (sexualized images). Participants completed a short writing exercise after viewing these images. Each participant saw only one type of image, but several examples. Daniels then coded these written exercises and identified themes, one of which was whether women self-objectified (i.e., commented on their own appearance/attractiveness). Daniels hypothesized that women who viewed the sexualized images (n = 140) would self-objectify (i.e., this theme would be present in what they wrote) more than those who viewed the performance athlete pictures (n = 117, despite what the Participants section of the paper implies). Labcoat Leni wants you to enter the data from Daniels's study (Table 19.3) and test her hypothesis that there is a significant association between the type of image viewed, and whether or not the women self-objectified (**daniels_2012.sav**). View the solutions at www.discoverspss.com, or on page 85 of Daniels's paper.

Table 19.3 some data from Daniels (2012)

	Theme present	Theme absent	Total
Performance athletes	20	97	117
Sexualized athletes	56	84	140

19.8.3 Bayesian test of association between two categorical variables ©

To compute a Bayes factor for a contingency table select *Analyze* > *Bayesian Statistics* > *Loglinear Models* (Figure 19.6). Drag one variable (I chose **training**) to the box labelled *Row variable* (or click), and drag the other (**dance**) to the box labelled *Column variable* (or click). Click Bayes Factor... to select a sampling plan; the most appropriate one for this example is *Multinomial Model* with Row Sum as *Fixed Margins* (SPSS Tip 19.1). In the main dialog box click OK to run the analysis.

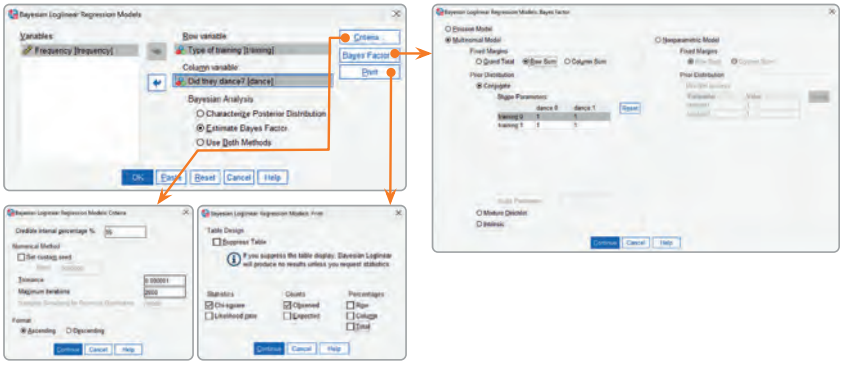


Figure 19.6 Dialog box for a Bayesian test of independence of two categorical variables

Output 19.6 shows the Bayes factor, which appears to be 0. However, if you double-click the table to edit it, and double-click the cell containing the Bayes factor value you will see that its value is 0.000021. This value suggests that the data are 0.000021 more likely under the null than under the alternative hypothesis. We can flip the interpretation by dividing 1 by this value: 1/0.000021 = 47,619. The data are more than 47,000 times more likely under the alternative than the null. In other words, we should strongly change our prior beliefs towards the alternative hypothesis: there is extremely strong evidence that dancing is associated with the type of reward.

Test of Independence[a]

	Value	df	Asymptotic Sig.(2-sided)	Exact Sig.(2-sided)	Exact Sig.(1-sided)
Bayes Factor	.000[b]				
Pearson Chi-Square	25.356[c]	1	.000		
Continuity Correction	23.520	1	.000		
Fisher's Exact Test				.000	.000

a. The row sums are fixed in the contingency table.
b. This analysis tests independence versus association, and assumes a multinomial model and conjugate priors.
c. 0 cells(.0%) have expected count less than 5. The minimum expected count is 14.440.

Double-click this cell

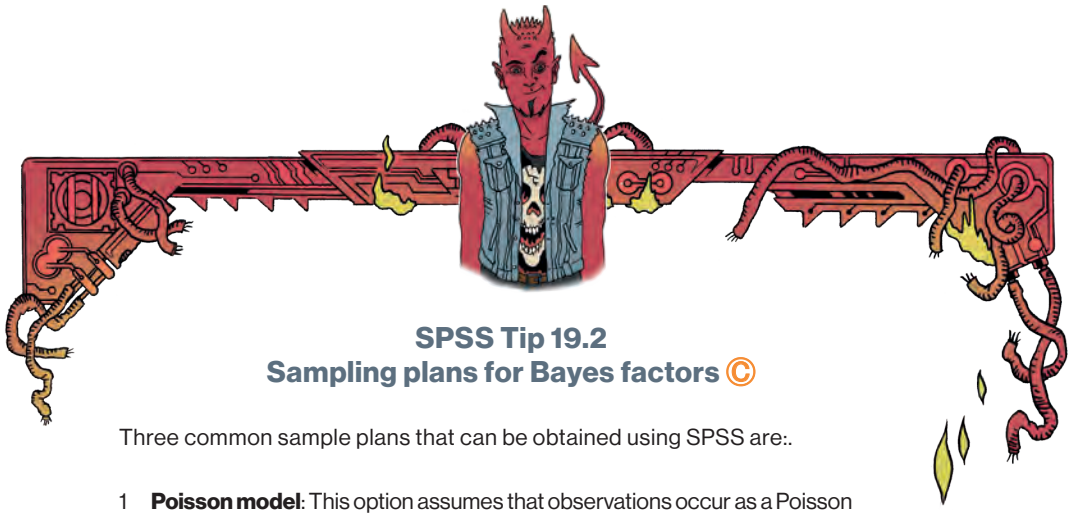
Test of Independence[a]

	Value	df	Asymptotic Sig.(2-sided)	Exact Sig.(2-sided)	Exact Sig.(1-sided)
Bayes Factor	.000[b]				
Pearson Chi-Square	25.356[c]				
Continuity Correction	23.520				
Fisher's Exact Test				.000	.000

0.000020966294983
b. This analysis tests independence versus association, and assumes a multinomial model and conjugate priors.

a. The row sums are fixed in the contingency table.
b. This analysis tests independence versus association, and assumes a multinomial model and conjugate priors.
c. 0 cells(.0%) have expected count less than 5. The minimum expected count is 14.440.

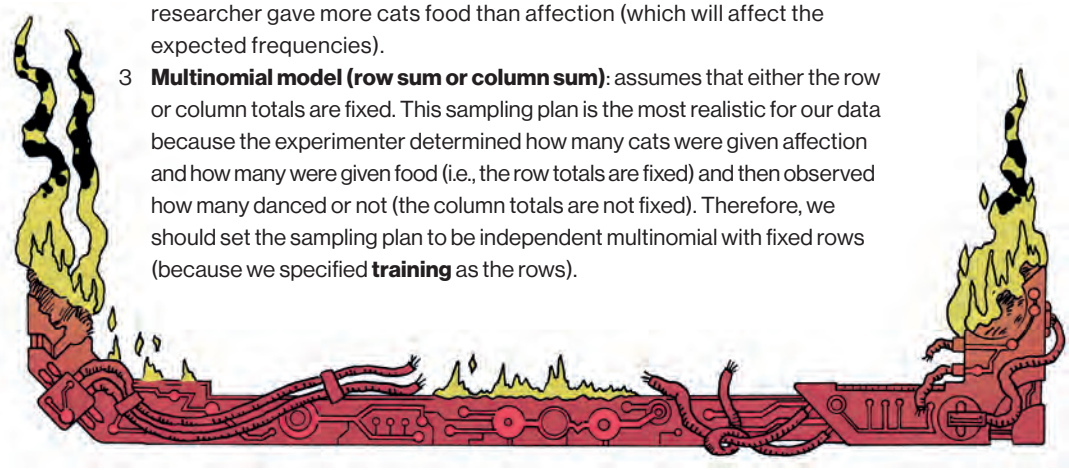
Output 19.6



SPSS Tip 19.2
Sampling plans for Bayes factors ©

Three common sample plans that can be obtained using SPSS are:.

1 **Poisson model**: This option assumes that observations occur as a Poisson process in which the total sample size is not fixed. In effect, we assume that the four cells of the contingency table are independent Poisson random variables. It's like assuming the four cells represent different groups of cats who materialized randomly in our study. This model would be unrealistic because the researcher determined which cats had affection and which had reward – these were not pre-existing groups.
2 **Multinomial model (grand total)**: This option assumes that total *N* is fixed, and observations are assigned to cells with fixed probability. This is more realistic that the Poisson sampling plan because our researcher chose to study 200 cats (so the total *N* is fixed). However, it is not realistic to assume that cells are assigned with fixed probability because the researcher gave more cats food than affection (which will affect the expected frequencies).
3 **Multinomial model (row sum or column sum)**: assumes that either the row or column totals are fixed. This sampling plan is the most realistic for our data because the experimenter determined how many cats were given affection and how many were given food (i.e., the row totals are fixed) and then observed how many danced or not (the column totals are not fixed). Therefore, we should set the sampling plan to be independent multinomial with fixed rows (because we specified **training** as the rows).



19.8.4 Calculating an effect size Ⓑ

Cramér's *V* is an adequate effect size (it is constrained to fall between 0 and 1 and is, therefore, easily interpretable), but a more common and useful measure of effect size for categorical data is the odds ratio (Section 3.7.5). Odds ratios are most interpretable in 2 × 2 contingency tables and are not as useful for larger contingency tables. I've said many times already that effect sizes are most useful

to summarize a focused comparison, and a 2 × 2 contingency table is the categorical data equivalent of a focused comparison.

To calculate the odds ratio for the type of reward we'd first calculate the odds that a cat danced given that they had food as a reward, which is the number of cats that were given food and danced, divided by the number of cats given food that didn't dance:

$$\begin{aligned}\text{odds}_{\text{dancing after food}} &= \frac{\text{number that had food and danced}}{\text{number that had food but didn't dance}} \\ &= \frac{28}{10} \\ &= 2.8\end{aligned} \tag{19.40}$$

Next we calculate the odds that a cat danced given that they had affection as a reward, which is the number of cats that were given affection and danced, divided by the number of cats given affection that didn't dance:

$$\begin{aligned}\text{odds}_{\text{dancing after affection}} &= \frac{\text{number that had affection and danced}}{\text{number that had affection but didn't dance}} \\ &= \frac{48}{114} \\ &= 0.421\end{aligned} \tag{19.41}$$

The odds ratio for the type of reward is the odds of dancing after food divided by the odds of dancing after affection:

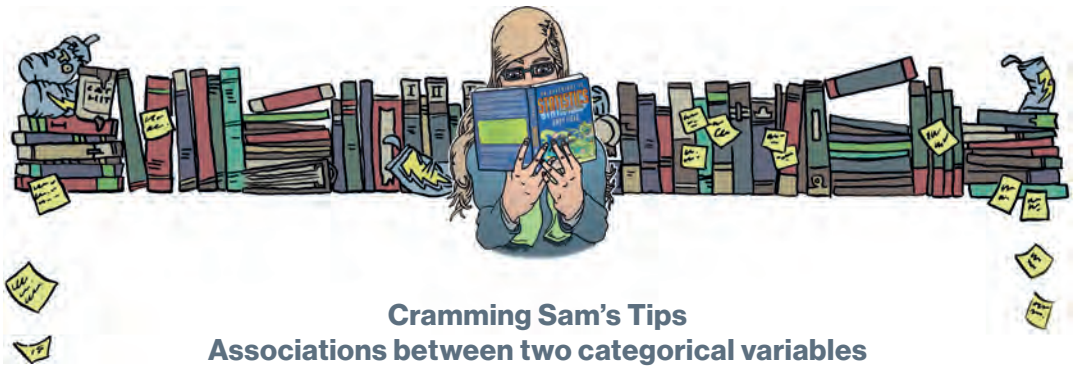
$$\begin{aligned}\text{odds Ratio} &= \frac{\text{odds}_{\text{dancing after food}}}{\text{odds}_{\text{dancing after affection}}} \\ &= \frac{2.8}{0.421} \\ &= 6.65\end{aligned} \tag{19.42}$$

It tells us that if a cat was trained with food the odds of their dancing were 6.65 times higher than if they had been trained with affection. The odds ratio is an extremely elegant and easily understood metric for expressing this type of effect.

19.8.5 Reporting the results of a chi-square test Ⓐ

When reporting Pearson's chi-square we report the value of the test statistic with its associated degrees of freedom and the significance value. The test statistic, as we've seen, is denoted by χ^2. The output tells us that the value of χ^2 was 25.36, that the degrees of freedom on which this was based were 1, and that it was significant at $p < 0.001$ (too small to report the exact p-value). It's also useful to reproduce the contingency table, and my vote would go to quoting the odds ratio and Bayes factor too. As such, we could report:

✓ There was a significant association between the type of training and whether cats would dance, $\chi^2(1) = 25.36$, $p < 0.001$. The Bayes factor strongly supported the alternative hypothesis, $BF_{10} = 47{,}619$. The odds ratio showed that the odds of cats dancing were 6.65 times higher if they were trained with food than if trained with affection.



Cramming Sam's Tips
Associations between two categorical variables

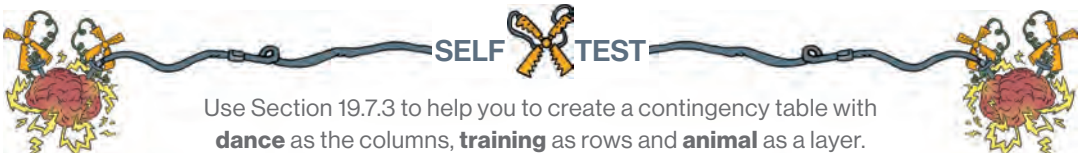
- To test the relationship between two categorical variables use *Pearson's chi-square test* or the *likelihood ratio statistic.*
- Look at the table labelled *Chi-Square Tests*. If the *Exact Sig.* value is less than or equal to 0.05 for the row labelled *Pearson Chi-Square* then there is a significant relationship between your two variables.
- Check underneath this table to make sure that no expected frequencies are less than 5.
- Look at the contingency table to work out what the relationship between the variables is: look out for significant standardized residuals (values outside of ±1.96), and columns that have different letters as subscripts (this indicates a significant difference).
- Calculate the *odds ratio*.
- The Bayes factor reported by SPSS tells you how much more likely the data are under the null hypothesis than the alternative. Divide 1 by this value to see how much more likely the data are under the alternative hypothesis than the null. Values greater than 1 indicate that your belief should change towards the alternative hypothesis, with values greater than 3 starting to indicate a change in beliefs that has substance.
- Report the χ^2 statistic, the degrees of freedom, the significance value and odds ratio. Also report the contingency table.



19.9 Loglinear analysis using SPSS Ⓑ

19.9.1 Initial considerations Ⓑ

Data are entered for loglinear analysis in the same way as for the chi-square test (see Sections 19.7.1 and 19.7.2). Let's extend the previous example to include dogs as well as cats. The data are in the file **animal_dance.sav**; open this file. There are three variables (**animal**, **training** and **dance**) that each have codes to represent the different categories within these variables. The process for fitting the model is outlined in Figure 19.2 – it's the same as for the chi-square test. First, let's check the expected frequencies in the contingency table (Section 19.5.2) using the *crosstabs* command.



Use Section 19.7.3 to help you to create a contingency table with **dance** as the columns, **training** as rows and **animal** as a layer.



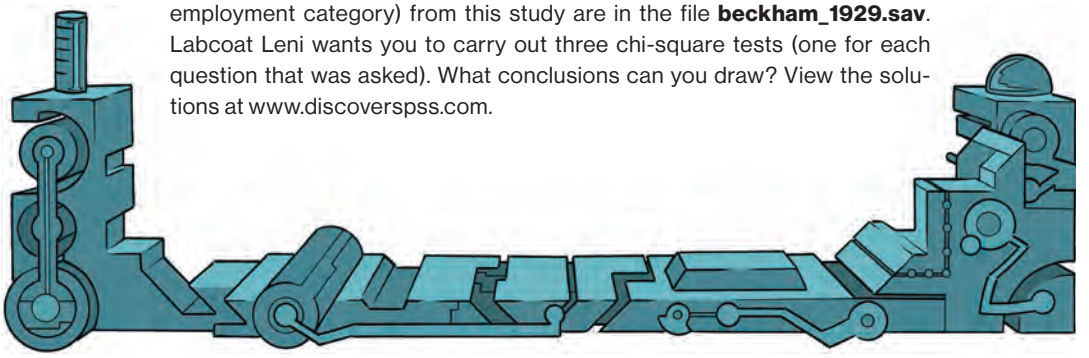
Labcoat Leni's Real Research 19.2
Is the Black American happy? Ⓐ

Beckham, A. S. (1929). *Journal of Abnormal and Social Psychology*, *24*, 186–190.

During my psychology degree I spent a lot of time reading about the civil rights movement in the USA. Instead of reading psychology, I read about Malcolm X and Martin Luther King Jr. For this reason I find Beckham's (1929) study of Black Americans a fascinating historical piece of research. Beckham was a Black American who founded the psychology laboratory at Howard University, Washington, DC, and his wife Ruth was the first Black woman ever to be awarded a PhD (also in psychology) at the University of Minnesota. To put some context on the study, it was published 36 years before the Jim Crow laws were finally overthrown by the Civil Rights Act of 1964, and at a time when Black Americans were segregated, openly discriminated against and victims of the most abominable violations of civil liberties and human rights (I recommend James Baldwin's superb *The Fire Next Time* for an insight into the times). The language of the study and the data from it are an uncomfortable reminder of the era in which it was conducted.

Beckham sought to measure the psychological state of 3443 Black Americans with three questions. He asked them to answer yes or no to whether they thought Black Americans were happy, whether they personally were happy as a Black American, and whether Black Americans *should* be happy. Beckham did no formal statistical analysis of his data (Fisher's article containing the popularized version of the chi-square test was published only 7 years earlier in a statistics journal that would not have been read by psychologists). I love this study, because it demonstrates that you do not need elaborate methods to answer important and far-reaching questions; with just three questions, Beckham told the world an enormous amount about very real and important psychological and sociological phenomena.

The frequency data (number of yes and no responses within each employment category) from this study are in the file **beckham_1929.sav**. Labcoat Leni wants you to carry out three chi-square tests (one for each question that was asked). What conclusions can you draw? View the solutions at www.discoverspss.com.



Type of training * Did they dance? * Animal Crosstabulation

Animal				Did they dance? No	Did they dance? Yes	Total
Cat	Type of training	Food as reward	Count	10a	28b	38
			Expected Count	23.6	14.4	38.0
			% within Type of training	26.3%	73.7%	100.0%
			% within Did they dance?	8.1%	36.8%	19.0%
			Standardized Residual	-2.8	3.6	
		Affection as reward	Count	114a	48b	162
			Expected Count	100.4	61.6	162.0
			% within Type of training	70.4%	29.6%	100.0%
			% within Did they dance?	91.9%	63.2%	81.0%
			Standardized Residual	1.4	-1.7	
	Total		Count	124	76	200
			Expected Count	124.0	76.0	200.0
			% within Type of training	62.0%	38.0%	100.0%
			% within Did they dance?	100.0%	100.0%	100.0%
Dog	Type of training	Food as reward	Count	14a	20b	34
			Expected Count	10.2	23.8	34.0
			% within Type of training	41.2%	58.8%	100.0%
			% within Did they dance?	66.7%	40.8%	48.6%
			Standardized Residual	1.2	-.8	
		Affection as reward	Count	7a	29b	36
			Expected Count	10.8	25.2	36.0
			% within Type of training	19.4%	80.6%	100.0%
			% within Did they dance?	33.3%	59.2%	51.4%
			Standardized Residual	-1.2	.8	
	Total		Count	21	49	70
			Expected Count	21.0	49.0	70.0
			% within Type of training	30.0%	70.0%	100.0%
			% within Did they dance?	100.0%	100.0%	100.0%

Each subscript letter denotes a subset of Did they dance? categories whose column proportions do not differ significantly from each other at the .05 level.

Output 19.7

The contingency table (Output 19.7) contains the number of cases that fall into each combination of categories (I've edited out the overall totals). The top half of this table is the same as Output 19.3 because the data are the same (we've just added some dogs), so look back in this chapter for a summary of what this tells us. For the dogs, 49 danced (70% of the total) and of these 20 were trained using food (40.8% of the total that danced) and 29 were trained with affection (59.2% of the total that danced). Twenty-one dogs didn't dance at all (30% of the total) and of those that didn't dance, 14 were trained using food as a reward (66.7% of the total that didn't dance) and 7 were trained using affection (33.3% of the total that didn't dance). In summary, dogs (70%) seem more willing to dance than cats (38%), and they're not too worried what training method is used: about half of those that danced were trained with affection and about half with food.

For loglinear analysis there should be no expected counts less than 1, and no more than 20% less than 5 (Section 19.5.2). The smallest expected count in the contingency table is 10.2 (for dogs that were trained with food but didn't dance), which exceeds 5 and so the assumption has been met.

19.9.2 The main analysis Ⓑ

Having established that the assumptions have been met we select *Analyze* > *Loglinear* > Model Selection... to access the dialog box in Figure 19.7. Drag any variables that you want to include in the analysis to

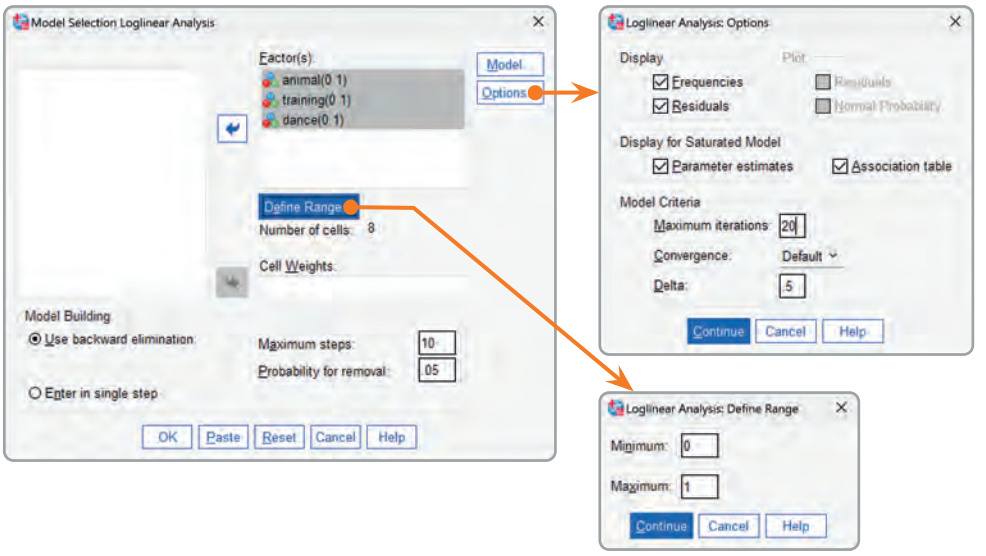


Figure 19.7 Main dialog box for loglinear analysis

the box labelled *Factor(s)* (or click ▶).[7] We have to specify the codes that we've used to define our categorical variables by selecting one or more variables in the *Factor(s)* box and clicking Define Range... to activate a dialog box in which you type the value of the minimum and maximum code that you used for the selected variables. Because all three variables in this example have the same codes (they all have two categories and were coded with 0 and 1) we can select all three, click Define Range..., type '0' in the *Minimum* box and '1' in the *Maximum* box, then click Continue.

The default options in the main box are fine. The default method is backward elimination (as I've described elsewhere). Alternatively, you can select *Enter in a single step*, which is a non-hierarchical method (in which all effects are entered and evaluated, like forced entry in the linear model). In loglinear analysis the combined effects take precedence over lower-order effects and so there is little to recommend non-hierarchical methods.

Click Model... to open a dialog box like we've seen before (e.g., Figure 13.9). By default the saturated model is fitted, and unless you have a very good reason for not fitting it, leave well alone. Click Options... to open the dialog box in Figure 19.7. The default options are fine but I would also select ☑ Parameter estimates to produce a table of parameter estimates for each effect (a *z*-score and associated confidence interval), and ☑ Association table to produce chi-square statistics for all of the effects in the model. This table can be useful in some situations, but as I've said before, if the higher-order interactions are significant then we shouldn't really be interested in the lower-order effects because they're confounded with the higher-order effects. Click Continue to return to the main dialog box and click OK to run the analysis.

19.10 Interpreting loglinear analysis ©

Output 19.8 contains three tables. The first tells us that we have 270 cases (remember that we had 200 cats and 70 dogs, and this is a useful check that no cats or dogs have been lost – they do tend to wander off). To begin, the saturated model is fitted (all terms are in the model including the highest-order

7 Remember that you can select several at the same time by holding down the *Ctrl* key, or *Cmd* on a Mac.

interaction, in this case the **animal** × **training** × **dance** interaction). The second table gives us the observed and expected counts for each of the combinations of categories in our model. These values should be the same as the original contingency table, except that each cell has 0.5 added to it (this value is the default and is fine, but if you want to change it you can do so by changing *Delta* in Figure 19.7). The final table contains two goodness-of-fit statistics: Pearson's chi-square and the likelihood ratio. These statistics are testing the hypothesis that the frequencies predicted by the model (the expected frequencies) are significantly different from the observed frequencies in the data. If our model is a good fit to the data then the observed and expected frequencies should be very similar (i.e., not significantly different). A significant result means that our model predictions are significantly different from our data (i.e., the model is a poor fit). In large samples these statistics should give the same results, but the likelihood ratio statistic is preferred in small samples. Both statistics are 0 and yield a probability value, *p*, of '.', which is a rather confusing way of saying that the probability cannot be computed. The reason why is that at this stage the model predicts the data *perfectly* (I explained why in Section 19.3.7). The next question is what bits of the model we can remove without significantly affecting the fit.

Output 19.9 tells us about the effects of removing parts of the model. The part of the table labelled *K-Way and Higher-Order Effects* has rows showing likelihood ratio and Pearson chi-square statistics when $K = 1$, 2 and 3 (as we go down the rows of the table). The first row ($K = 1$) tells us whether removing the one-way effects (i.e., the main effects of **animal**, **training** and **dance**) and any higher-order effects will significantly affect the fit of the model. There are lots of higher-order effects here – there are the two-way interactions and the three-way interaction – and so this is basically testing whether

Data Information

		N
Cases	Valid	270
	Out of Range[a]	0
	Missing	0
	Weighted Valid	270
Categories	Animal	2
	Type of training	2
	Did they dance?	2

a. Cases rejected because of out of range factor values.

Design 1

Cell Counts and Residuals

			Observed		Expected			
Animal	Type of training	Did they dance?	Count[a]	%	Count	%	Residuals	Std. Residuals
Cat	Food as reward	No	10.500	3.9%	10.500	3.9%	.000	.000
		Yes	28.500	10.6%	28.500	10.6%	.000	.000
	Affection as reward	No	114.500	42.4%	114.500	42.4%	.000	.000
		Yes	48.500	18.0%	48.500	18.0%	.000	.000
Dog	Food as reward	No	14.500	5.4%	14.500	5.4%	.000	.000
		Yes	20.500	7.6%	20.500	7.6%	.000	.000
	Affection as reward	No	7.500	2.8%	7.500	2.8%	.000	.000
		Yes	29.500	10.9%	29.500	10.9%	.000	.000

a. For saturated models, .500 has been added to all observed cells.

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	.000	0	.
Pearson	.000	0	.

Output 19.8

if we remove everything from the model there will be a significant effect on the fit of the model. This effect is highly significant. If this test was non-significant (if the values of *Sig.* were above 0.05) then this would tell you that removing everything from your model would not affect the fit (overall the combined effect of your variables and interactions is not significant). The next row ($K = 2$) tells us whether removing the two-way interactions (i.e., the **animal** × **training**, **animal** × **dance** and **training** × **dance** interactions) and any higher-order effects (i.e. the three-way interaction) will affect the model. This test is also highly significant, indicating that if we removed the two-way interactions and the three-way interaction then this would have a significant detrimental effect on the model. The final row ($K = 3$) tests whether removing the three-way effect *and* higher-order effects will significantly affect the fit of the model. The three-way interaction *is* the highest order effect that we have, so this test evaluates removal of three-way interaction (i.e., the **animal** × **training** × **dance** interaction). Both chi-square and likelihood ratio tests agree that removing this interaction will significantly affect the fit of the model (because the probability value is less than 0.05).

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio Chi-Square	Likelihood Ratio Sig.	Pearson Chi-Square	Pearson Sig.	Number of Iterations
K-way and Higher Order Effects[a]	1	7	200.163	<.001	253.556	<.001	<.001
	2	4	72.267	<.001	67.174	<.001	2
	3	1	20.305	<.001	20.778	<.001	4
K-way Effects[b]	1	3	127.896	<.001	186.382	<.001	<.001
	2	3	51.962	<.001	46.396	<.001	<.001
	3	1	20.305	<.001	20.778	<.001	<.001

a. Tests that k-way and higher order effects are zero.
b. Tests that k-way effects are zero.

Output 19.9

The bottom of the table (*K-way Effects*) expresses the same thing but without including the higher-order effects. The first row ($K = 1$) tests whether removing the main effects (the one-way effects of **animal**, **training** and **dance**) has a significant detrimental effect on the model, and it does (because the *p*-value is less than 0.05). The second row ($K = 2$) tests whether removing the two-way interactions (**animal** × **training**, **animal** × **dance** and **training** × **dance**) has a significant detrimental effect on the model, and again it does ($p < 0.001$). This finding tells us that one or more of these two-way interactions is a significant predictor. The final row ($K = 3$) tests whether removing the three-way interaction (**animal** × **training** × **dance**) has a detrimental effect on the model. It does ($p < 0.001$), suggesting that this interaction is a significant predictor of the data. The results in this row are identical to the final row of the top half of the table (the *K-way and Higher-Order Effects*) because it is the highest-order effect and so in the top part of the table there were no higher-order effects to include.

In a nutshell, Output 19.9 tells us is that the three-way interaction is significant: removing it from the model has a significant effect on how well the model fits the data. We also know that removing all two-way interactions has a significant effect on the model, but remember that loglinear analysis should be done hierarchically and so these two-way interactions aren't of interest to us because the three-way interaction is significant (we'd look only at these effects if the three-way interaction were non-significant).

Output 19.10 shows the *Partial Associations* table (which you might have selected by ticking ☑ Association table in Output 19.10). This table breaks down the model into specific components. For example, the previous output told us that removing all of the two-way interactions significantly affects the model fit, but we don't know which of the two-way interactions specifically make a difference; this table tells us. The Pearson chi-square tests are significant for all three interactions (*Sig.* is less than 0.05). Likewise, the previous output told us that removing the main effects significantly affected the fit of the model, and Output 19.10 shows specifically that the main effects of **animal** and **training** are both significant ($p < 0.001$), but the main effect of **dance** is not ($p = 0.223$). We should ignore these effects, though, because they are confounded with the higher-order interaction of **animal** × **training** × **dance**.

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
animal*training	1	13.760	<.001	2
animal*dance	1	13.748	<.001	2
training*dance	1	8.611	.003	2
animal	1	65.268	<.001	2
training	1	61.145	<.001	2
dance	1	1.483	.223	2

Output 19.10

Parameter Estimates

Effect	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
animal*training*dance	1	-.360	.083	-4.320	<.001	-.523	-.197
animal*training	1	-.402	.083	-4.823	<.001	-.565	-.239
animal*dance	1	.197	.083	2.364	.018	.034	.360
training*dance	1	-.104	.083	-1.251	.211	-.268	.059
animal	1	.404	.083	4.843	<.001	.240	.567
training	1	-.328	.083	-3.937	<.001	-.492	-.165
dance	1	-.232	.083	-2.782	.005	-.395	-.069

Output 19.11

Output 19.11 shows the ☑ Parameter estimates (selected in Figure 19.7) of the model. It tests each effect in the model with a *z*-score and gives us confidence intervals. If you ignore the plus or minus sign, the bigger the *z*, the more significant the effect; therefore, the value of *z* gives us a useful comparison between effects. The main effect of **animal** is the most important effect in the model ($z = 4.84$) followed by the **animal** × **training** interaction ($z = -4.82$) and then the **animal** × **training** × **dance** interaction ($z = -4.32$) and so on. However, it's worth reiterating that in this case we don't need to concern ourselves with anything other than the three-way interaction.



Output 19.12 deals with the backward elimination. We begin with the highest-order effect (in this case the **animal** × **training** × **dance** interaction); we remove it from the model, see what effect this has, and if it doesn't have a significant effect then we move on to the next highest effects (in this case the two-way interactions). However, we've already seen that removing the three-way interaction will have a significant effect, and this is confirmed at this stage by the table labelled *Step Summary*. Therefore, the analysis stops here: the three-way interaction is not removed and this final model is evaluated using the likelihood ratio statistic. We're looking for a non-significant test statistic, which indicates that the expected values generated by the model are not significantly different from the observed data (put another way, the model is a good fit to the data). In this case the model is a perfect fit to the data.[8]

8 The fact that the analysis has stopped here is unhelpful because I can't show you how it would proceed in the event of a non-significant three-way interaction. However, it does keep things simple, and if you're interested in exploring loglinear analysis further, there is a task at the end of the chapter that shows you what happens when the highest-order interaction is not significant.

Backward Elimination Statistics

Step Summary

Step[a]		Effects	Chi-Square[c]	df	Sig.	Number of Iterations
0	Generating Class[b]	animal*training*dance	.000	0	.	
	Deleted Effect 1	animal*training*dance	20.305	1	<.001	4
1	Generating Class[b]	animal*training*dance	.000	0	.	

a. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than .050.

b. Statistics are displayed for the best model at each step after step 0.

c. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

Cell Counts and Residuals

			Observed		Expected			
Animal	Type of training	Did they dance?	Count	%	Count	%	Residuals	Std. Residuals
Cat	Food as reward	No	10.000	3.7%	10.000	3.7%	.000	.000
		Yes	28.000	10.4%	28.000	10.4%	.000	.000
	Affection as reward	No	114.000	42.2%	114.000	42.2%	.000	.000
		Yes	48.000	17.8%	48.000	17.8%	.000	.000
Dog	Food as reward	No	14.000	5.2%	14.000	5.2%	.000	.000
		Yes	20.000	7.4%	20.000	7.4%	.000	.000
	Affection as reward	No	7.000	2.6%	7.000	2.6%	.000	.000
		Yes	29.000	10.7%	29.000	10.7%	.000	.000

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	.000	0	.
Pearson	.000	0	.

Output 19.12

19.10.1 Following up loglinear analysis Ⓑ

An alternative way to interpret a three-way interaction is to conduct chi-square analysis at different levels of one of your variables. For example, to interpret our **animal** × **training** × **dance** interaction, we could perform a chi-square test on **training** and **dance** separately for dogs and cats (the analysis for cats will be the same as earlier in the chapter) and compare the results in the different animals.



Use the *split file* command (see Section 6.9.6) to run a chi-square test on **dance** and **training** for dogs and cats.

The results and interpretation for cats are in Output 19.4 and for dogs is shown in Output 19.13. For dogs there is still a significant relationship between the types of training and whether they danced but it is weaker (the chi-square is 3.93, compared to 25.2 for cats).[9] This finding seems to suggest that dogs are more likely to dance if given affection than if given food, the opposite of cats.

9 The chi-square statistic depends on the sample size, so really you need to calculate effect sizes and compare them to make this kind of statement (unless you had equal numbers of dogs and cats).

Chi-Square Tests[a]

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	3.932[b]	1	.047	.068	.042	
Continuity Correction[c]	2.966	1	.085			
Likelihood Ratio	3.984	1	.046	.068	.042	
Fisher's Exact Test				.068	.042	
Linear-by-Linear Association	3.876[d]	1	.049	.068	.042	.030
N of Valid Cases	70					

a. Animal = Dog
b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.20.
c. Computed only for a 2x2 table
d. The standardized statistic is 1.969.

Output 19.13

19.10.2 Interpreting the interaction Ⓑ

Let's piece this together to make sense of the three-way interaction. Let's plot the frequencies across all the different categories. Figure 19.8 shows something called a mosaic plot, in which each coloured area represents the proportion of cases. Look at affection as a reward: of those that didn't dance the vast majority were cats (blue area) and relatively few were dogs (the green strip), but of those that did dance it was slightly more cats than dogs (top right). Now, look at food as a reward: for those that didn't dance the areas for both animals are relatively small and similar in size, indicating that few animals didn't dance after food and this was true for cats and dogs. For those that did dance after food the area for cats is a little bigger than for dogs. The take-home is that cats are sensible creatures that do stupid stuff only when there's something tasty in it for them (i.e., food), whereas dogs will dance for a cuddle as well as for food.☺

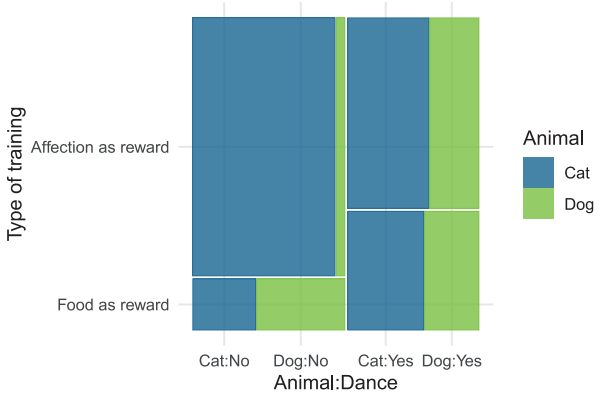


Figure 19.8 Percentage of different animals who danced or not after being trained with affection or food

19.10.3 Effect sizes in loglinear analysis Ⓑ

As with Pearson's chi-square, let's look at odds ratios again. Odds ratios are easiest to understand for 2 × 2 contingency tables, and so if you have significant higher-order interactions, or your variables have more than two categories, it is worth trying to break these effects down into logical 2 × 2 tables and calculating odds ratios that reflect the nature of the interaction. In this example we could calculate odds ratios for dogs and cats separately. We have the odds ratio for cats already (Section 19.8.3), and for dogs we would get 0.35:

$$\begin{aligned}\text{odds}_{\text{dancing after food}} &= \frac{\text{number that had food and danced}}{\text{number that had food but didn't dance}} \\ &= \frac{20}{14} \\ &= 1.43\end{aligned} \tag{19.43}$$

$$\begin{aligned}\text{odds}_{\text{dancing after affection}} &= \frac{\text{number that had affection and danced}}{\text{number that had affection but didn't dance}} \\ &= \frac{29}{7} \\ &= 4.14\end{aligned} \tag{19.44}$$

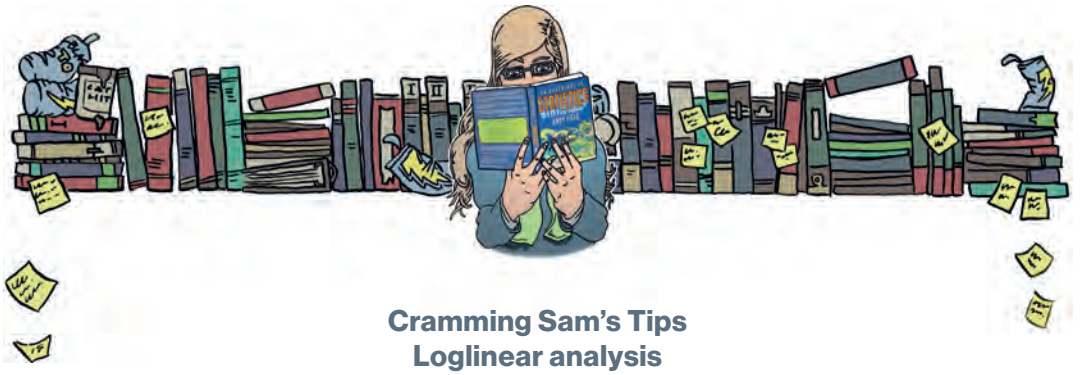
$$\begin{aligned}\text{odds Ratio} &= \frac{\text{odds}_{\text{dancing after food}}}{\text{odds}_{\text{dancing after affection}}} \\ &= \frac{1.43}{4.14} \\ &= 0.35\end{aligned} \tag{19.45}$$

This tells us is that if a dog was trained with food the odds of their dancing were 0.35 times the odds if they were rewarded with affection (i.e., they were less likely to dance). Another way to say this is that the odds of their dancing were 1/0.35 = 2.90 times lower if they were trained with food instead of affection. Compare this to cats where the odds of dancing were 6.65 higher if they were trained with food rather than affection. As you can see, comparing the odds ratios for dogs and cats is an elegant way to present the three-way interaction term in the model.

19.11 Reporting the results of loglinear analysis Ⓑ

For loglinear analysis report the likelihood ratio statistic for the final model, usually denoted just by χ^2. For any terms that are significant you should report the chi-square change, or you could consider reporting the *z*-score for the effect and its associated confidence interval. If you break down any higher-order interactions in subsequent analyses then you need to report the relevant chi-square statistics (and odds ratios). For this example we could report:

- ✓ The three-way loglinear analysis produced a final model that retained all effect. This indicated that the highest-order interaction (the animal × training × dance interaction) was significant, $\chi^2(1) = 20.31$, $p < 0.001$. To break down this effect, separate chi-square tests on the training and dance variables were performed separately for dogs and cats. For cats, there was a significant association between the type of training and whether they would dance, χ^2 (1) = 25.36, $p < 0.001$; this was true in dogs also, $\chi^2(1) = 3.93$, $p = 0.047$. Odds ratios indicated that the odds of dancing were 6.65 higher after food than affection in cats, but only 0.35 in dogs (i.e., in dogs, the odds of dancing were 2.90 times lower if trained with food compared to affection). The analysis seems to reveal a fundamental difference between dogs and cats: cats are more likely to dance for food than affection, whereas dogs are more likely to dance for affection than food.



Cramming Sam's Tips
Loglinear analysis

- Test the relationship between more than two categorical variables with *loglinear analysis*.
- Loglinear analysis is hierarchical: the initial model contains all main effects and interactions. Starting with the highest-order interaction, terms are removed to see whether their removal significantly affects the fit of the model. If it does then this term is not removed and all lower-order effects are ignored.
- Look at the table labelled *K-Way and Higher-Order Effects* to see which effects have been retained in the final model. Then look at the table labelled *Partial Associations* to see the individual significance of the retained effects (look at the column labelled *Sig*. values less than or equal to 0.05 indicate significance).
- Look at the *Goodness-of-Fit Tests* for the final model: if this model is a good fit to the data then this statistic should be non-significant (*Sig* should be bigger than 0.05).
- Look at the contingency table to interpret any significant effects (percentage of total for cells is the best thing to look at).



19.12 Jane and Brian's story Ⓐ

Jane looked at the discoloured brass of the jar in front of her. It had a timeless class. She brushed dust from the glass and stared at the brain within. What to do? She'd decided to stop her crazy experiment weeks ago. It had been easy at first: the time she'd spent with Brian made her feel more human than she ever had; she hadn't given the corridors of the Pleiades building a thought. But like an addict, she returned there as soon as life got tough. Maybe she'd been working too hard. Weeks of little sleep and long stretches of work were taking their toll on her resolve. She'd hit a wall with what she was working on and no matter how much she chiselled away at it with her mind, it was defeating her. She didn't like defeat. Like a ghost she had been floating in and out of Brian's life. When she ran out of energy to work she would go to him and he would take care of her before she disappeared again. He rolled with it. He didn't ask questions, but he listened and supported. It was so important to her to find a way, any way, to break the impasse and finish what she'd started. 'Just one more,' she thought, 'one more to give my mind the kick it needs.'

She looked deeper into the jar, trying to find a way into the soul of whoever used to possess this mind. She felt a tingle of excitement at her own depravity, and then disgust as she thought of Brian, waiting for her somewhere, oblivious to her current dilemma, and her past.

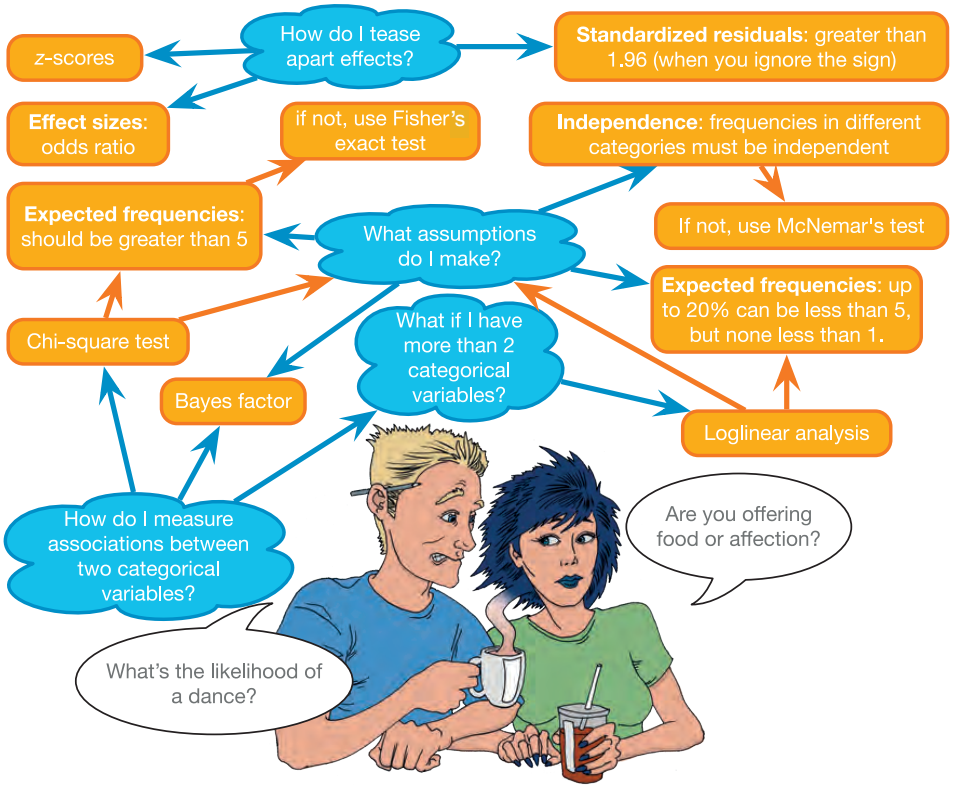


Figure 19.9 What Brian learnt from this chapter

'Just one more,' she thought.

She had no choice, did she? She had to finish her work. This was the only way.

'Just one more.'

She thought of Brian. Her hands trembled. Her chest felt tight as she tried to suppress her emotions.

'Just one more,' she whispered, 'one more … one more … one more … NO MORE!'

Jane punched the lever that retuned the jar to its cubby hole and ran. She didn't stop until she reached Brian's.

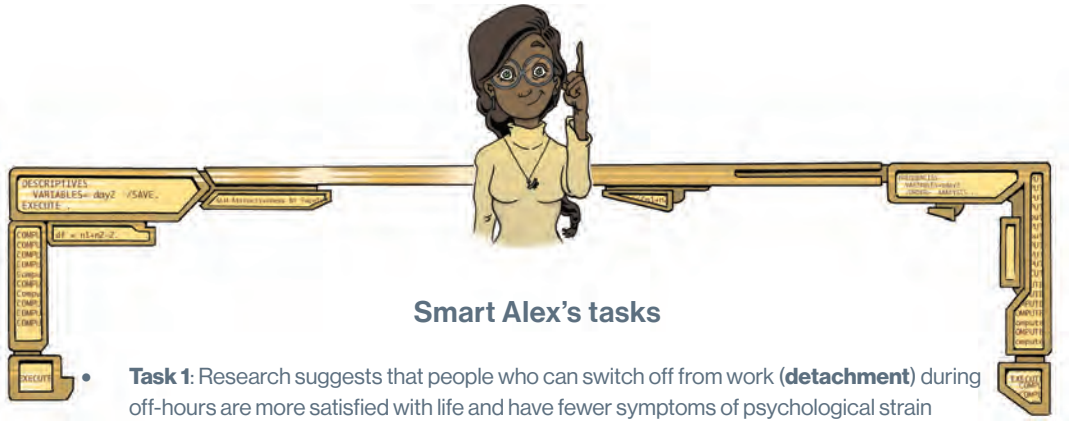
19.13 What next? Ⓐ

When I wrote the first edition of this book I had no idea what journey I was starting. My main ambition was to write a statistics book that I would enjoy reading. I didn't expect anyone else to enjoy reading it, but it turns out some people do, which is a nice feeling. One of the weird side effects of writing statistics books is that everyone assumes that I'm a statistician. I'm not, which means that I constantly disappoint people by not being able to answer their statistics questions. This book is pretty much my total knowledge about statistics. It's probably more than the sum. The loglinear section in this chapter, for example, was the result of reading a lot of stuff about loglinear models, which I have now forgotten (because I never use it). Should I ever need to do a loglinear analysis I will read the chapter in this book and be amazed at how good I am at pretending to know what I'm talking about. If you knew how often I look stuff up in my own book, you probably wouldn't have bought it.

Over the editions of this book I've had 'fans' set up appreciation societies, had videos on YouTube dedicated to me, had hundreds of people send me pictures of their dogs, cats, lizards, children, horses, birds posing with my book, been befriended by the manager of a black metal band, been invited to an autopsy (seriously …), had people randomly turn up at my office from all over the world to say 'hi', had strangers ask for selfies with me at conferences, and all because I did the most uncool thing you can think of: wrote stats books. It bemuses me, but it is heart-warming to have strangers from all walks of life want to spend their valuable time saying hi to you. In its own tiny, unimportant (to most people) microcosm, it's a bit like being the rock star that the younger me always wanted to be, except writing about stats is nowhere near as much fun as playing music. It was only a matter of time before the musical itch needed to be scratched once again.

19.14 Key terms that I've discovered

Chi-square distribution
Chi-square test
Cramér's V
Fisher's exact test
Goodman and Kruskal's λ
Loglinear analysis
Phi
Saturated model
Yates's continuity correction

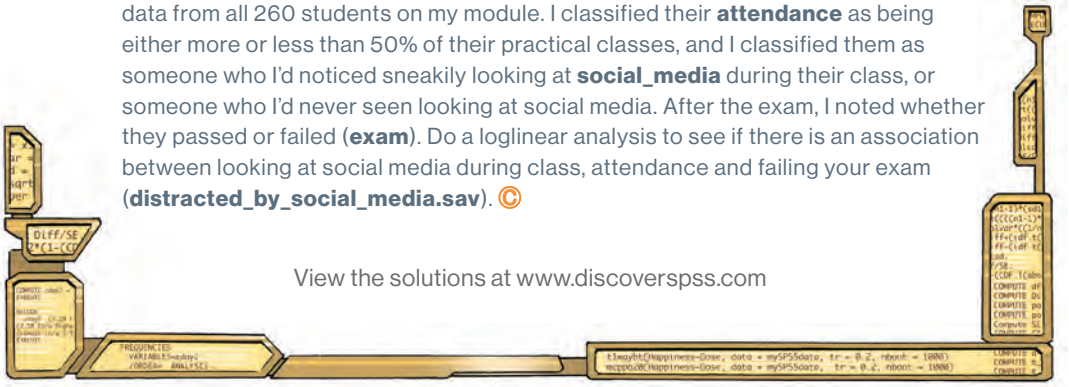


Smart Alex's tasks

- **Task 1**: Research suggests that people who can switch off from work (**detachment**) during off-hours are more satisfied with life and have fewer symptoms of psychological strain (Sonnentag, 2012). Factors at work, such as time pressure, affect your ability to detach when away from work. A study of 1709 employees measured their time pressure (**time_pressure**) at work (no time pressure, low, medium, high and very high time pressure). Data generated to approximate Figure 1 in Sonnentag (2012) are in the file **sonnentag_2012.sav**. Carry out a chi-square test to see if time pressure is associated with the ability to detach from work. Ⓐ
- **Task 2**: Labcoat Leni's Real Research 19.1 describes a study (Daniels, 2012) that looked at the impact of sexualized images of athletes compared to performance pictures on women's perceptions of the athletes and of themselves. Women looked at different types of pictures (**picture**) and then did a writing task. Daniels identified whether certain themes were present or absent in each written piece (**theme_present**). We looked at the self-evaluation theme, but Daniels identified others: commenting on the athlete's body/appearance (**athletes_body**), indicating admiration or jealousy for the athlete (**admiration**), indicating that the athlete was a role model or motivating (**role_model**), and their own physical activity (**self_physical_activity**). Test whether the type of picture viewed was associated with commenting on the athlete's body/appearance (**daniels_2012.sav**). Ⓐ
- **Task 3**: Using the data in Task 2, see whether the type of picture viewed was associated with indicating admiration or jealousy for the athlete. Ⓐ
- **Task 4**: Using the data in Task 2, see whether the type of picture viewed was associated with indicating that the athlete was a role model or motivating. Ⓐ

- **Task 5**: Using the data in Task 2, see whether the type of picture viewed was associated with the participant commenting on their own physical activity. Ⓐ
- **Task 6**: I wrote much of the third edition of this book in the Netherlands (I have a soft spot for it). The Dutch travel by bike much more than the English. I noticed that many more Dutch people cycle while steering with only one hand. I pointed this out to one of my friends, Birgit Mayer, and she said that I was a crazy English fool and that Dutch people did not cycle one-handed. Several weeks of me pointing at one-handed cyclists and her pointing at two-handed cyclists ensued. To put it to the test I counted the number of Dutch and English cyclists who ride with one or two hands on the handlebars (**handlebars.sav**). Can you work out which one of us is correct? Ⓐ
- **Task 7**: Compute and interpret the odds ratio for Task 6. Ⓑ
- **Task 8**: Certain editors at Sage like to think they're great at football (soccer). To see whether they are better than Sussex lecturers and postgraduates we invited employees of Sage to join in our football matches. Every person played in one match. Over many matches, we counted the number of players that scored goals. Is there a significant relationship between scoring goals and whether you work for Sage or Sussex (**sage_editors_can't_play_football.sav**)? Ⓐ
- **Task 9**: Compute and interpret the odds ratio for Task 8. Ⓑ
- **Task 10**: I was interested in whether horoscopes are tosh. I recruited 2201 people, made a note of their star sign (this variable, obviously, has 12 categories: Capricorn, Aquarius, Pisces, Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpio and Sagittarius) and whether they believed in horoscopes (this variable has two categories: believer or unbeliever). I sent them an identical horoscope about events in the next month, which read: 'August is an exciting month for you. You will make friends with a tramp in the first week and cook him a cheese omelette. Curiosity is your greatest virtue, and in the second week, you'll discover knowledge of a subject that you previously thought was boring. Statistics perhaps. You might purchase a book around this time that guides you towards this knowledge. Your new wisdom leads to a change in career around the third week, when you ditch your current job and become an accountant. By the final week you find yourself free from the constraints of having friends, your boy/girlfriend has left you for a Russian ballet dancer with a glass eye, and you now spend your weekends doing loglinear analysis by hand with a pigeon called Hephzibah for company.' At the end of August I interviewed these people and I classified the horoscope as having come true, or not, based on how closely their lives had matched the fictitious horoscope. Conduct a loglinear analysis to see whether there is a relationship between the person's star sign, whether they believe in horoscopes and whether the horoscope came true (**horoscope.sav**). Ⓒ
- **Task 11**: On my statistics module students have weekly practical statistics classes in a computer laboratory. I've noticed that many students are studying social media more than the very interesting statistics assignments that I have set them. I wanted to see the impact that this behaviour had on their exam performance. I collected data from all 260 students on my module. I classified their **attendance** as being either more or less than 50% of their practical classes, and I classified them as someone who I'd noticed sneakily looking at **social_media** during their class, or someone who I'd never seen looking at social media. After the exam, I noted whether they passed or failed (**exam**). Do a loglinear analysis to see if there is an association between looking at social media during class, attendance and failing your exam (**distracted_by_social_media.sav**). Ⓒ

View the solutions at www.discoverspss.com





CATEGORICAL OUTCOMES: LOGISTIC REGRESSION

20.1 What will this chapter tell me? 926
20.2 What is logistic regression? 927
20.3 Theory of logistic regression 927
20.4 Sources of bias and common problems 939
20.5 Binary logistic regression 943
20.6 Interpreting logistic regression 952
20.7 Interactions in logistic regression: a sporty example 962
20.8 Reporting logistic regression 971
20.9 Jane and Brian's story 973
20.10 What next? 975
20.11 Key terms that I've discovered 975
Smart Alex's tasks 976

Letters A to D indicate increasing difficulty

20.1 What will this chapter tell me? Ⓐ

Over the last few chapters we've seen how my childhood dreams of rock stardom crumbled as I became a statistical model in human form. I can scarcely imagine a more dramatic failure to achieve one's ambitions. I felt like I had become numbers without a soul, and I needed salvation before the transformation became complete. It was time to unlock the latent rock star once more, and get some therapy in the process. So, the hollow husk of a 29-year-old Andy decided to learn the drums (feel free to insert your own joke about it being the perfect instrument for a failed musician, but they're much harder to play than people think). A couple of years later I had a call from an old friend of mine, Doug. Back in my Scansion days, Doug had played in another local band, so we went back a long way. The conversation went a bit like this:

'Remember the last time I saw you we talked about you coming and having a jam with us?' asked Doug.

I had absolutely no recollection whatsoever of him saying this, so I responded 'Yes'.

'How about it then?' he said.

'OK,' I said, 'you arrange it and I'll bring my guitar'.

'No, you whelk,' he said, 'we need a drummer. Learn some of the songs on the CD I gave you last year.'

I'd played his band's CD and I liked it, but their songs were ridiculously fast and there was no way on earth that I could play them. 'Sure, no problem,' I lied.

I spent the next two weeks trying to become a much better drummer than I was. It'd be nice to report that at the rehearsal I astounded them with my brilliance, but I didn't. I did, however, nearly have a heart attack and herniate everything in my body. Still, we had another rehearsal, and then another, and we're still having them.[1] It's curious that I started off playing guitar (which I can still 'play', incidentally), and then took up drums,



Figure 20.1 Having a therapy session

1 Although not with Doug in the band, as it turns out, and not very frequently these days because of, you know, having children, jobs, being old and boring. That sort of thing.

because there are assumptions about the personalities of musicians in rock bands: the singers are egocentric, guitarists are the cool ones, bassists are laid back and introverted, and drummers are alleged to be wild hedonists. I am *not* a hedonist, by the way. These assumptions are speculative. If we wanted to test what personality characteristics predict the instrument you choose to play, we'd have a categorical outcome (type of instrument) with several categories (drums, guitar, bass, singing, keyboard, tuba, etc.) and continuous predictors (personality traits such as neuroticism and extroversion). We've looked at how we can quantify associations between purely categorical variables, but if we have continuous predictors too then surely there's no model on earth that can handle that kind of complexity? Actually, we can do **logistic regression**, but there's a compelling case that we should go out and do something fun instead.

20.2 What is logistic regression? Ⓐ

In the previous chapter we looked models of the relationships between categorical variables, and we'll now extend this discussion to logistic regression – a model for predicting categorical outcomes from categorical and continuous predictors. In its simplest form, this means predicting which of two categories a person is likely to belong to, given their scores on various predictor variables. For example, in medical research logistic regression is used to generate models from which predictions can be made about the likelihood that a tumour is cancerous or benign (for example). Based on existing data, the logistic regression model is used to establish variables that predict the malignancy of a tumour. These variables can then be measured for a new patient and their values placed in the logistic regression model to obtain a probability of malignancy. If the probability of the tumour being malignant is low then the doctor may decide not to carry out expensive and painful surgery that is likely to be unnecessary. When we are trying to predict membership of only two categories the model is known as **binary logistic regression**, but when we want to predict membership of more than two categories we use **multinomial** (or **polychotomous**) **logistic regression**.

20.3 Theory of logistic regression Ⓒ

20.3.1 Modelling categorical outcomes Ⓒ

I won't dwell on the maths behind logistic regression (I am living proof that you don't need to know it). Instead, I'll draw a few parallels to the linear model to give you the conceptual gist of what's going on. To keep things as straightforward as possible I'll explain binary logistic regression, but most of the principles extend to predicting membership of more than two categories. We've seen many times that the linear model can be expressed as:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i \tag{20.1}$$

when there is one predictor variable and as:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni} + \varepsilon_i \tag{20.2}$$

when there are two or more. Remember that b_0 is the value of the outcome when the predictors are zero (the intercept), the bs quantify the relationship between each predictor and outcome, X is the value of each predictor variable and ε is the error in prediction.

Imagine that we are sports scientists wanting to predict whether a soccer player scores or misses a penalty kick.[2] We have a binary outcome: success (the player scores) or failure (the player misses). We'll return to this example later in the chapter, but for now we decide that past success is likely to be a strong predictor of success: if the player has scored penalties in the past then they are likely to score in the future. At the start of a season, we obtain the percentage of previous penalties scored by the 75 penalty takers across various leagues, and we then record whether they score the first penalty that they take in the subsequent season. We could dummy-code the outcome variable as 0 (missed the penalty) and 1 (scored). These values equate to the probability of scoring because when someone misses the penalty the probability of scoring is 0 (it definitely didn't happen), and when someone scores then the probability of scoring is 1 (it definitely did happen). For each value of previous success Figure 20.2 shows whether the player scored their first penalty of the season (triangles) or not (circles). Notice that there are only two possible values of the outcome: 0 (did not score/the probability of scoring is 0) and 1 (scored/the probability of scoring is 1).

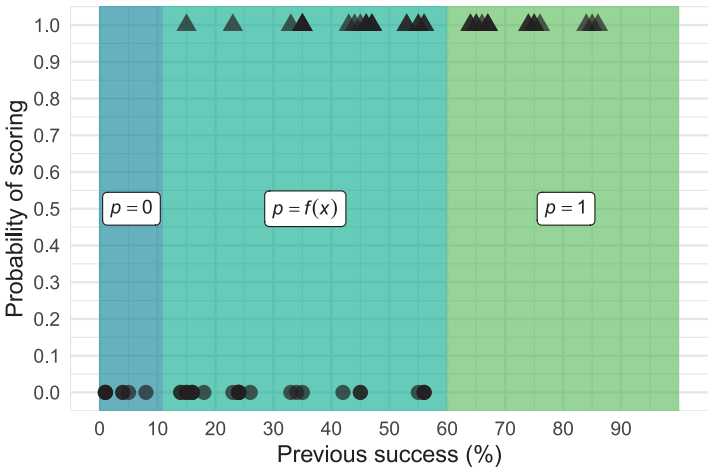
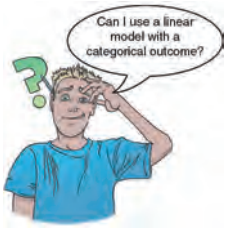


Figure 20.2 Success or failure of penalty takers as a function of the previous success



One of the assumptions of the linear model is that the relationship between the predictors and outcome is linear (Section 9.4.1). When the outcome variable is categorical, this assumption is violated (Berry, 1993). For example, Figure 20.3 (left) shows the result of fitting a linear model to the data. Do you think this model is a good fit? It is not, because most of the observed data are really far from the line (the exceptions being the circles when previous success is less than 10%, and the triangles when previous success falls between 65% and 75%). The residuals of this model (Figure 20.3, right) show that linearity has been violated. If linearity were a reasonable assumption then the observed (red) trend line for the residuals would be flat at 0 (the grey horizontal line), but instead it shows curvature. Basically, if we model binary outcomes with the type of model we have learnt about so far, we will violate a major assumption of that model, and it will not be very representative of the data.

2 If you prefer some other sport that involves penalty kicks or throws then imagine I'm writing about that sport instead of soccer.

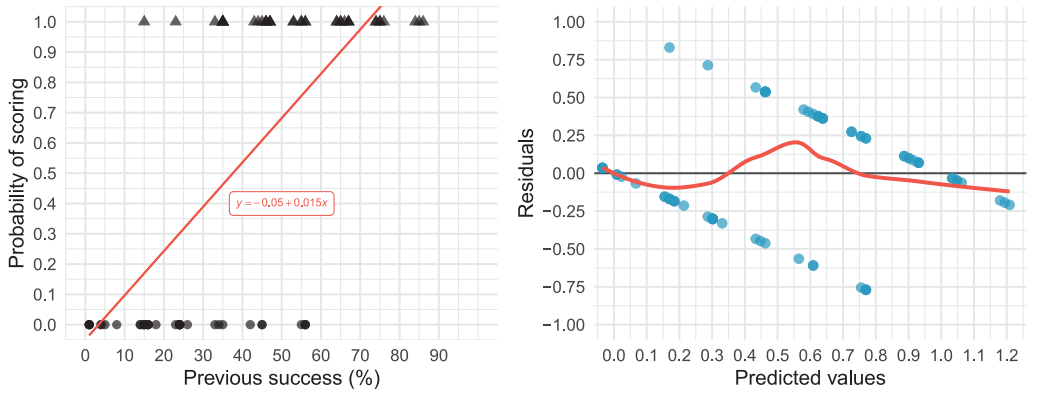


Figure 20.3 A linear model for a binary outcome (left) and its residuals (right)

If we can't fit a linear model, what can we do instead? Look back at Figure 20.2 and note that I have split it into three sections. The left of the figure shows the data for players whose previous success rate is below 11%. None of these players score, so the probability of scoring in this part of the data is 0 (when previous success is below 11% no one scores). Conversely, the right of the figure shows the data for players whose previous success rate is above 60%. All of these players scored, so the probability of scoring in this portion of data is 1 (when previous success is above 60% everyone scores). The middle section of Figure 20.2 (between 11% and 60%) shows a region of the data where some players score and others miss, so the probability of scoring will be somewhere between 0 (it definitely won't happen) and 1 (it definitely will happen). In this middle section the probability of scoring will be some kind of function of previous experience. We can see, for example, that when previous success is between 10% and 30% most players miss but a few score (so the probability of scoring is still quite low), whereas when previous success is between 40% and 60% few players miss and most score (so the probability of scoring is getting larger). A crude way to get at the probabilities in the middle section of the data would be to split the previous success variable into bins (e.g., 0–9%, 10–29%, 30–39% and so on), count the number of scorers in each bin, and then convert that to a proportion. That's what I've done in Table 20.1. Previous success (%) has been split into 10 equally sized categories each of which spans a 10% range. For example, the first category would contain any players whose previous success rate was 0% or greater and less than 10%, the second contains any player whose previous success rate was 10% or greater and less than 20% and so on.[3] Having allocated players to these categories, we count how many scored. For example, no players whose previous success rate was 0% or greater and less than 10% scored, and 9 players whose previous success rate was 40% or greater and less than 50% scored. Out of the 75 players, 40 of them scored, so we can convert these counts into proportions. For example, 9 players whose previous success rate was 40% or greater and less than 50% scored, which is 9/40 = 0.225 (nearly a quarter) of the players that scored. The final column shows the cumulative proportion, which are the proportions added down the rows. For example, the proportions in the first two rows are 0 and 0.025, so the cumulative proportion for the second row is 0 + 0.025 = 0.025, and the proportions in the first four rows are 0, 0.025, 0.025 and 0.175 so the cumulative proportion for the fourth row is 0 + 0.025 + 0.025 + 0.175 = 0.225, and so on. Eventually the cumulative proportion equals 1 (the maximum possible value for a proportion) meaning that all the data are accounted for. These proportions equate to empirical probabilities. If you ask 'what's the probability that a player whose previous success is 50% will score?', these proportions tell us that 0.45 (45%) of players who score had previous success rates under 50%, so a reasonable probability is 0.45. Similarly, the probability that a player scores who has a previous success rate of 80% might be approximated by 0.925 (i.e. extremely likely) because we know that 0.925 of scorers had previous success rates less than 80%.

3 The notation (0, 10] means 'all values from 0 (not including 0) up to, and including, 10'.

Table 20.1 The number and proportion of scorers for categories of previous success

Previous success (bins)	Number of players who scored	Proportion of players who scored	Cumulative proportion
(0, 10]	0	0.000	0.000
(10, 20]	1	0.025	0.025
(20, 30]	1	0.025	0.050
(30, 40]	7	0.175	0.225
(40, 50]	9	0.225	0.450
(50, 60]	6	0.150	0.600
(60, 70]	8	0.200	0.800
(70, 80]	5	0.125	0.925
(80, 90]	3	0.075	1.000
(90, 100]	0	0.000	1.000

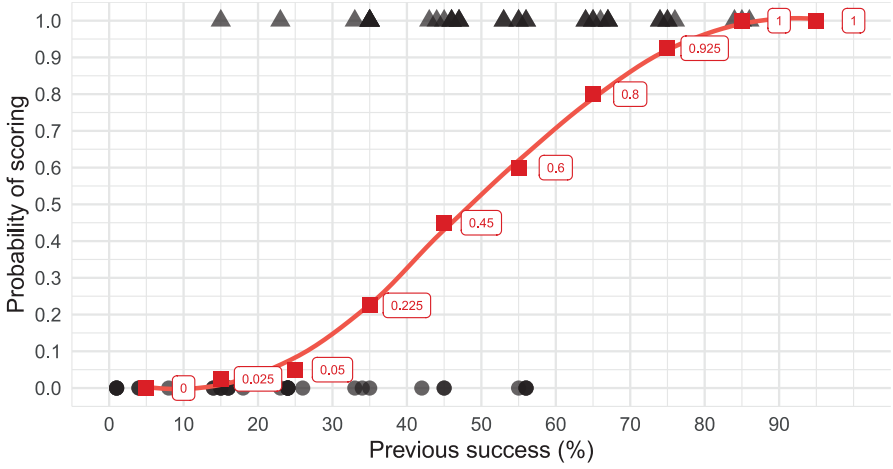


Figure 20.4 The estimated probability of scoring based on the observed proportions of scorers

When discussing Figure 20.2 I noted that for the middle section when previous success ranges from 11% to 60% the probability of scoring will be some kind of function of previous success. We can use the probability estimates from Table 20.1 (last column) to get an idea of what this function might look like. Figure 20.4 plots the probability estimates as squares, and by connecting these estimates we can see that the function that links the probability of scoring with previous success looks S-shaped. Following this curve, we can see that when previous success is low, the probability of scoring is also low, but as previous success increases so does the probability of scoring until the point where, at about an 80–90% previous success rate, you are almost certain to score (the estimated probability is around 1). Throughout this book we have modelled the relationship between predictors and an outcome using a mathematical function that describes a straight line, but to model the probability of scoring we now need a mathematical function that describes an S-shaped curve. One such function is the logistic function, from which logistic regression gets its name:

$$P(Y_i) = \frac{1}{1 + e^{-(b_0 + b_1 X_i)}} \qquad (20.3)$$

This equation says that we can predict the probability of an outcome, $P(Y_i)$, from a function of the variable, X. A couple of points to make. First, e is the base of natural logarithms, and, second, you might have noticed that the linear model that you've seen countless times already is cosily tucked up inside the parentheses. As with all the other models we've met, we estimate the values of the bs from the data. In the models we have encountered so far parameters are (typically) estimated using the method of least squares (Section 2.7) whereas in logistic regression maximum likelihood estimation is used (Section 2.7).

Figure 20.5 shows the logistic function applied to the penalty-taker data. Compare the line with the one in Figure 20.4, where we approximated the probabilities from the data. Hopefully you can see that the logistic function does a very good job of modelling the S-shape relationship that we observed between previous success and the probability of scoring.

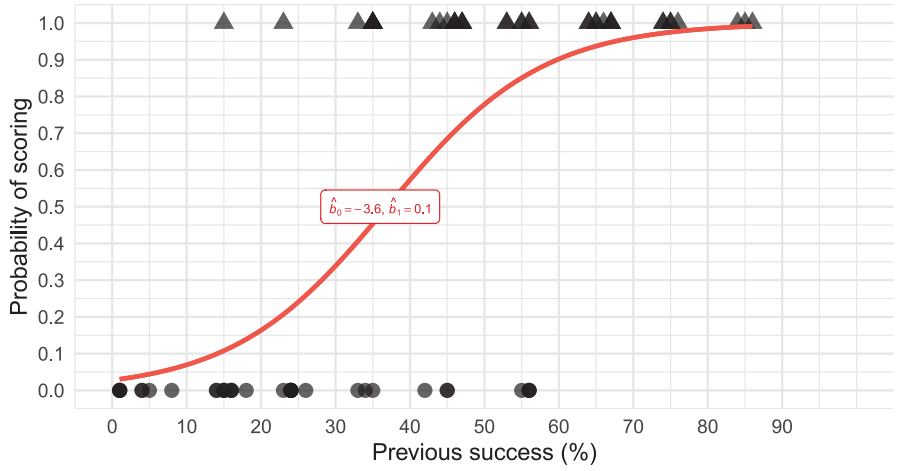


Figure 20.5 The logistic function applied to the penalty-taker data

Note that Figure 20.5 tells us the parameter estimates and that they are $\hat{b}_0 = -3.6$ and $\hat{b}_1 = 0.1$. We can substitute these values into Eq. 20.3 to get the model for the penalty data:

$$P(\text{Scored}_i) = \frac{1}{1 + e^{-(-3.6 + 0.1X_i)}} \tag{20.4}$$

With other models in this book we have seen that the parameters of the model determine its shape and location. The same is true with the logistic function. Figure 20.6 (top) shows that the value of b_0 determines the horizontal location of the curve (as it increases the curve moves to the right) and b_1 determines the slope of the curve. If b_1 is positive then as the predictor increases so does the probability of the outcome, whereas a negative value flips the curve such that as the predictor increases the probability of the outcome *decreases* (Figure 20.6, middle). Therefore, as in other models we have considered, the direction of b_1 indicates the direction of the relationship between the predictor and the probability of the outcome. Also as with other models we've considered, the magnitude of b_1 indicates the rate of change of (the probability of) the outcome as the predictor changes. A value of 0 indicates that as the predictor changes the probability of the outcome doesn't (i.e. zero relationship) and as b_1 increases in relationship the S-shape gets less flat (Figure 20.6, bottom).

Just as the linear model expands to accommodate multiple predictors, so does the logistic function. With several predictors the model becomes:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni})}} \tag{20.5}$$

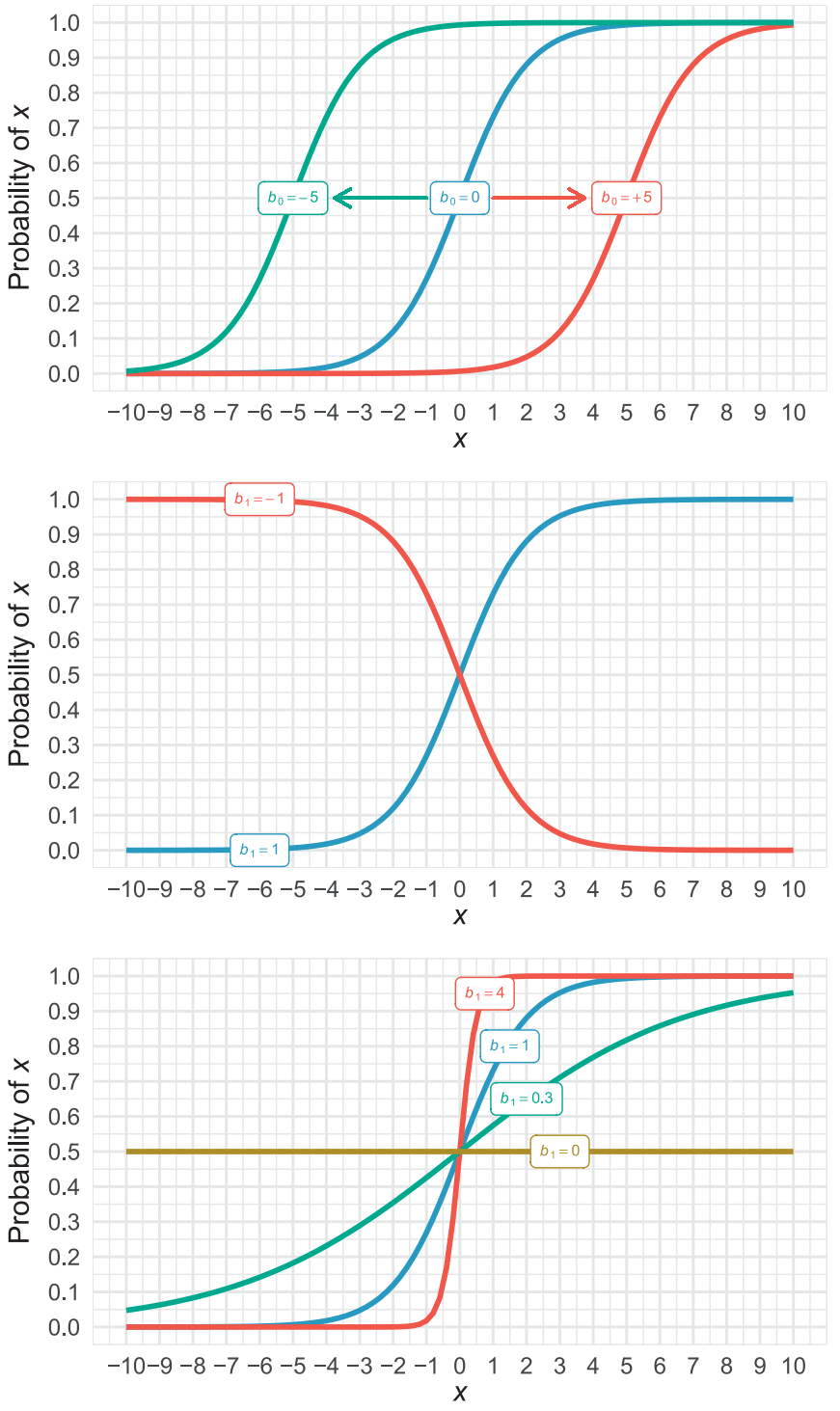


Figure 20.6 The effects of the parameters on the shape and location of the logistic curve

Note that instead of the parentheses containing the linear model for one predictor, some friends have been invited round for tea so the parentheses party contains the full linear model (Eq. 20.2). Note that because we're predicting the probability of the outcome occurring, the predicted values from the model will lie between 0 and 1. A value close to 0 means that the outcome is very unlikely to occur, and a value close to 1 means that the outcome is very likely to occur.

20.3.2 An alternative way to express the model Ⓓ

The logistic regression model can be expressed in a different, but mathematically identical, form. First let's substitute the linear part of the model with a symbol. Any symbol will do, but I'll use η. We can express the model as:

$$\begin{aligned} P(Y) &= \frac{1}{1+e^{-\eta}} \\ \eta &= b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni} \end{aligned} \tag{20.6}$$

Now let's rearrange the equation to make the linear part (η) the subject:

$$\begin{aligned} P(Y) &= \frac{1}{1+e^{-\eta}} \\ P(Y)\left(1+e^{-\eta}\right) &= 1 \\ P(Y) + P(Y)e^{-\eta} &= 1 \\ e^{-\eta} &= \frac{1-P(Y)}{P(Y)} \end{aligned} \tag{20.7}$$

Remembering that a value to a negative power is the same as the reciprocal of that value to the corresponding positive power means that $e^{-\eta} = 1/e^{\eta}$, which gives us:

$$\frac{1}{e^{\eta}} = \frac{1-P(Y)}{P(Y)} \tag{20.8}$$

Flipping the fractions on both sides of the equation, and taking the natural log of both side (remembering that $\log(e^x) = x$) results in:

$$\begin{aligned} e^{\eta} &= \frac{P(Y)}{1-P(Y)} \\ \log(e^{\eta}) &= \log\left(\frac{P(Y)}{1-P(Y)}\right) \\ \log\left(\frac{P(Y)}{1-P(Y)}\right) &= \eta \end{aligned} \tag{20.9}$$

Substituting the linear part back in for η gives us this alternative, but equivalent form of the model:

$$\log\left(\frac{P(Y)}{1-P(Y)}\right) = b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni} \tag{20.10}$$

This form of the model makes it much more obvious that we are fitting a linear model because the right-hand side of the equation is identical to the other models we have fitted in this book (which were introduced in Chapter 6). The left-hand side is different. Instead of predicting the raw value of

the outcome variable, instead we're predicting the natural log of the probability of the event occurring divided by the probability of it *not* occurring.[4] In Section 3.7.5 we discovered that the probability of the event occurring divided by the probability of it not occurring is called the odds:

$$\text{odds} = \frac{P(\text{event})}{P(\text{no event})} \tag{20.11}$$

So, in the logistic regression model we are predicting the log of the odds of the outcome from a linear combination of predictors:

$$\log(\text{odds of } Y) = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} \tag{20.12}$$

As with other models, we would estimate the model parameters from the data and when we do we stick hats on them to remind us that they are estimates:

$$\log(\text{odds of } Y) = \hat{b}_0 + \hat{b}_1X_{1i} + \hat{b}_2X_{2i} + \ldots + \hat{b}_nX_{ni} \tag{20.13}$$

Therefore, the model parameters can be interpreted exactly as we have throughout this book, except that where we've talked about 'the outcome' we now need to talk about 'the log odds of the outcome':

- $\hat{b}_0$ is the log odds of the outcome when all predictors have a value of 0.
- $\hat{b}_1$ is the change in the log odds of the outcome when predictor 1 increases by one unit (and all other predictors are fixed).
- $\hat{b}_n$ is the change in the log odds of the outcome when the nth predictor increases by one unit (and all other predictors are fixed).

20.3.3 The odds ratio: exp(*b*) ©

In the previous section we discovered that the logistic regression model can be expressed in a way that has a cosy familiarity with all the other linear models in this book. It's as cosy as snuggling under a soft blanket with a cup of tea. Unfortunately, under that blanket you're pressed painfully against the bony elbow of 'log odds'. Never date someone called log odds.

So, although it's nice to know that the model parameters can basically be interpreted in the way you're used to, it's annoying that they are expressed in terms of the log odds of the outcome. Odds are difficult enough to understand and logs make no sense to anyone (well, not to me), so log odds are the dysfunctional relationship of two units that met in a cesspit. No one understands what a change in the log odds means. No one.

It's common, therefore, to exponentiate the parameter estimates, which means that instead of interpreting $\hat{b}$ we interpret the exponential of it ($e^{\hat{b}}$ or $\mathbf{exp(\hat{b})}$). Taking the exponential of a log value reverses the log. Therefore, the exponential of a parameter estimate expresses the relationship between the predictor and odds of the outcome, rather than then log odds. Avoiding thinking about logarithms is always a good thing. For example, we can interpret the exponential of the parameter estimates as follows:

4 If an outcome can only occur or not, it has two probabilities associated with it. The probability that it happens, denoted $P(Y)$, and the probability that it does not, denoted $P(\overline{Y})$. These probabilities sum to 1, so $P(Y) + P(\overline{Y}) = 1$, and rearranging this equation gives us $P(\overline{Y}) = 1 - P(Y)$.

- $\exp(\hat{b}_0)$ is the odds of the outcome when all predictors have a value of 0.
- $\exp(\hat{b}_1)$ is the proportional change in the odds of the outcome when predictor 1 increases by one unit (and all other predictors are fixed).
- $\exp(\hat{b}_n)$ is the proportional change in the odds of the outcome when the *n*th predictor increases by one unit (and all other predictors are fixed).

Notice that I've used the phrase 'proportional change' rather than 'change'. This is because of how logarithms work. In the linear models we've covered up to now, $\hat{b}$ is the literal change in the outcome ($\hat{Y}$) as a predictor increases by one unit:

$$\hat{b} = \hat{Y}_{\text{after a unit change in the predictor}} - \hat{Y}_{\text{original value of predictor}} \tag{20.14}$$

In the logistic model the outcome is the log odds, so the equivalent is:

$$\hat{b} = \log(\text{odds}_{\text{after a unit change in the predictor}}) - \log(\text{odds}_{\text{original value of predictor}}) \tag{20.15}$$

Because of the rules of logarithms, which I summarize in the preface, this is mathematically equivalent to:

$$\hat{b} = \log\left(\frac{\text{odds}_{\text{after a unit change in the predictor}}}{\text{odds}_{\text{original value of predictor}}}\right) \tag{20.16}$$

When we take the exponent of each side of the equation we get:

$$\exp\left(\hat{b}\right) = \frac{\text{odds}_{\text{after a unit change in the predictor}}}{\text{odds}_{\text{original value of predictor}}} \tag{20.17}$$

Therefore $\exp(\hat{b})$ is commonly referred to the *odds ratio* because it is the ratio of the odds after a unit change in the predictor to the odds before the unit change (also see Section 3.7.5). It expresses the proportionate change in the odds. This is interpreted as follows:

- $\exp(\hat{b}) = 1$. If the odds ratio is 1 it means that the top and bottom half of Eq. 20.17 are the same value and the odds have not changed. In other words, the predictor has no association with the outcome.
- $\exp(\hat{b}) > 1$. If the odds ratio is greater than 1 it means that the top half of Eq. 20.17 is greater than the bottom half. In other words, the odds of the outcome increase as the predictor increases. There is a positive association between the predictor and the odds of the outcome.
- $\exp(\hat{b}) < 1$. If the odds ratio is less than 1 it means that the top half of Eq. 20.17 is smaller than the bottom half. In other words, the odds of the outcome decrease as the predictor increases. There is a negative association between the predictor and the odds of the outcome.

20.3.4 The Wald statistic Ⓑ

In the linear model, we used the parameter estimates (*b*) and their standard errors to compute a *t*-statistic. In logistic regression there is an analogous statistic – the *z*-statistic – which follows the normal distribution. Like the *t*-statistic in the linear model, the *z*-statistic tells us whether the *b*-value for that predictor is significantly different from zero. If the coefficient is significantly different from zero then we assume that the predictor is making a significant contribution to the prediction of the outcome.

The following equation shows how the *z*-statistic is calculated, and it's basically identical to the *t*-statistic in the linear model (see Eq. 9.22) – it is the *b*-value divided by its associated standard error:

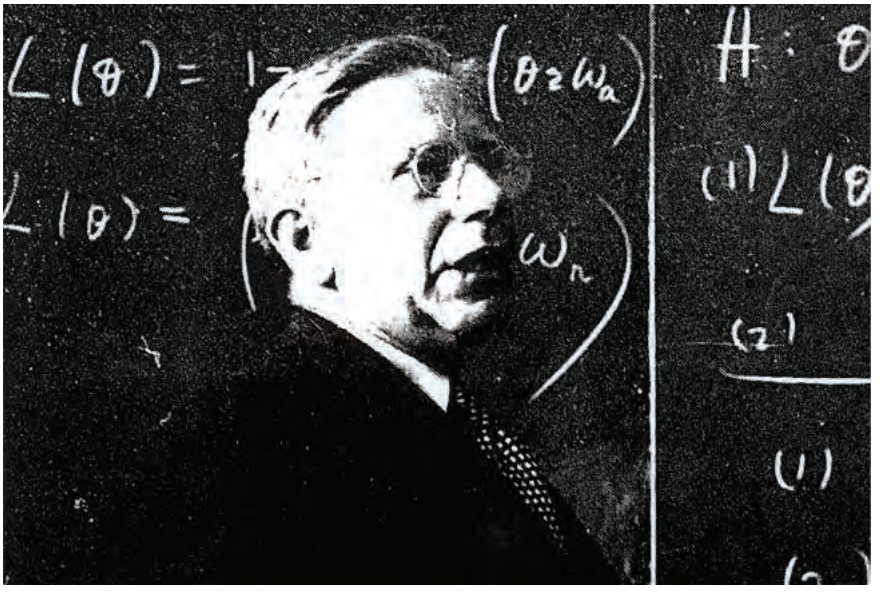


Figure 20.7 Abraham Wald writing 'I must not devise test statistics prone to having inflated standard errors' on the blackboard 100 times

$$z = \frac{\hat{b}}{SE_{\hat{b}}} \tag{20.18}$$

The z-statistic was developed by Abraham Wald (Figure 20.7) and is known as the **Wald statistic**. IBM SPSS Statistics software reports the Wald statistic as z^2, which transforms it so that it has a chi-square distribution. The z-statistic should be used a little cautiously because, when the b-value is large, the standard error tends to become inflated, resulting in the z-statistic being underestimated (see Menard, 2002). The inflation of the standard error increases the probability of rejecting a predictor as being significant when, in reality, it is making a significant contribution to the model (i.e., a Type II error). When assessing whether predictors significantly predict the outcome, it is probably more accurate to enter predictors hierarchically and examine the change in likelihood ratio statistics.

20.3.5 Assessing the model: the log-likelihood statistic ©

As we have seen, the logistic regression models the probabilities of an event occurring, $P(Y)$, based on observations of whether the event *did* occur, Y. For a given person, i, the *observed* outcome, Y_i, will be either 0 (the outcome didn't occur) or 1 (the outcome did occur), but the *predicted* value, $P(\hat{Y}_i)$, will be a value between 0 (there is no chance that the outcome will occur) and 1 (the outcome will certainly occur). When we assessed the fit of the linear model we compared the observed and predicted values of the outcome (if you remember, we use R^2, which is the squared Pearson correlation between the observed values of the outcome and the values predicted by the model). We do much the same in logistic regression using the **log-likelihood**:

$$\text{log-likelihood} = \sum_{i=1}^{N}\left[Y_i \ln\left(P\left(\hat{Y}_i\right)\right) + \left(1 - Y_i\right)\ln\left(1 - P\left(\hat{Y}_i\right)\right)\right] \tag{20.19}$$

The log-likelihood is based on summing the probabilities associated with the predicted, $P(\hat{Y}_i)$, and actual, Y_i, outcomes. It is analogous to the residual sum of squares in the sense that it is an indicator

of how much unexplained information there is after the model has been fitted. It follows, therefore, that large values of the log-likelihood statistic indicate poorly fitting statistical models, because the larger the value of the log-likelihood, the more unexplained observations there are.

20.3.6 Assessing the model: the deviance statistic ©

The **deviance** is closely related to the log-likelihood described in the previous section. It's given by:

$$\text{deviance} = -2 \times \text{log-likelihood} \tag{20.20}$$

The deviance is often referred to as ***–2LL*** because of the way it is calculated. It's rather convenient to (almost) always use the deviance rather than the log-likelihood because it has a chi-square distribution (see Chapter 19 and the Appendix), which makes it easy to calculate its significance.

One important use of the log-likelihood and deviance is to compare models. For example, it's useful to compare a logistic regression model against a baseline state – usually the model when only the intercept is included (i.e., no predictors). In the standard linear model, the baseline model we use is the mean of all scores (i.e., we predict the outcome from the intercept). With a categorical outcome it doesn't make sense to use the overall mean (because is the outcome is whether an event happened or not), so instead we use the frequency with which the outcome occurred. If the outcome occurs 107 times, and doesn't occur 72 times, then our best guess of the outcome will be that it occurs (because it occurs more times than it doesn't). As such, like the linear model, our baseline model is the model that gives us the best prediction when we know nothing other than the values of the outcome: in logistic regression this will be the outcome category that occurs most often, which is the same as predicting the outcome from the intercept. If we add one or more predictors to the model, we can compute the improvement of the model as:

$$\begin{aligned}\chi^2 &= (-2LL_{\text{baseline}}) - (-2LL_{new}) \\ &= 2LL_{\text{new}} - 2LL_{\text{baseline}} \\ df &= k_{\text{new}} - k_{\text{baseline}}\end{aligned} \tag{20.21}$$

We take the new model deviance and subtract from it the deviance for the baseline model (the model when only the constant is included). This difference is known as a likelihood ratio[5] and has a chi-square distribution with degrees of freedom equal to the number of parameters, k, in the new model minus the number of parameters in the baseline model. The number of parameters in the baseline model will always be 1 (the constant is the only parameter); any subsequent model will have degrees of freedom equal to the number of predictors plus 1 (i.e., the number of predictors plus one parameter representing the constant).

If we build up models hierarchically (i.e., adding one predictor at a time) we can also use Eq. 20.21 to compare these models. For example, if you have a model (we'll call it the 'old' model) and you add a predictor (the 'new' model), you can see whether the new model has improved the fit using Eq. 20.21 in which the baseline model is the 'old' model. The degrees of freedom will be the difference between the degrees of freedom of the two models.

5 You might wonder why it is called a 'ratio' when a 'ratio' means something is divided by something else, and we're subtracting not dividing. As we saw for $\exp(\hat{b})$, subtracting two logs gives the same result as the log of their ratio. For example, 10/5 = 2 and (try it on your calculator) log(10) – log(5) = log(2).

20.3.7 Assessing the model: *R* and R^2 ©



When we discussed the linear model we saw that the multiple correlation coefficient *R* and the corresponding R^2 are useful measures of how well the model fits the data. The likelihood ratio is similar in that it is based on the level of correspondence between predicted and observed values of the outcome. However, there is a more literal analogue of the multiple correlation in logistic regression known as the *R*-statistic. It is the partial correlation between the outcome variable and each of the predictor variables and it varies between –1 and 1. A positive value indicates that as the predictor variable increases, so does the likelihood of the event occurring. A negative value implies that as the predictor value increases, the likelihood of the outcome occurring decreases. If a predictor variable has a small value of *R* then it contributes only a small amount to the model.

To compute *R* use the following equation:

$$R = \sqrt{\frac{z^2 - 2df}{-2LL_{\text{basline}}}} \tag{20.22}$$

in which the $-2LL$ is the deviance for the original model, the Wald statistic (z) is calculated as described in Section 20.3.4. and the degrees of freedom can be read from the SPSS output for the variables in the equation. We'll see in the next section that the Wald statistic can be inaccurate under certain circumstances and, because of this, *R* is by no means an accurate measure. For this reason treat the value of *R* with caution, and it is invalid to square it and interpret it as you would in a linear model with a continuous outcome.

There is some controversy over what *would* make a good analogue to R^2 in logistic regression. **Hosmer and Lemeshow's** (1989) measure, R_L^2 is calculated by dividing the model chi-square, which represents the change from the baseline (based on the log-likelihood), by the baseline $-2LL$ (the deviance of the model before any predictors were entered):

$$R_{\text{L}}^2 = \frac{\chi^2_{\text{model}}}{-2LL_{\text{basline}}} \tag{20.23}$$

Given what the model chi-square represents (see Eq. 20.21), another way to express this statistic is:

$$R_{\text{L}}^2 = \frac{(-2LL_{\text{basline}}) - (-2LL_{\text{new}})}{-2LL_{\text{basline}}} \tag{20.24}$$

R_L^2 is the proportional reduction in the absolute value of the log-likelihood measure. It is a measure of how much the badness of fit improves as a result of the inclusion of the predictor variables. It can vary between 0 (indicating that the predictors are useless at predicting the outcome variable) and 1 (indicating that the model predicts the outcome variable perfectly).

SPSS uses **Cox and Snell's** (1989) measure, R_{CS}^2 which is based on the deviance of the model ($-2L$–$2LL_{\text{new}}$), the deviance of the original model ($-2L$–$2LL_{\text{baseline}}$), and the sample size, n:

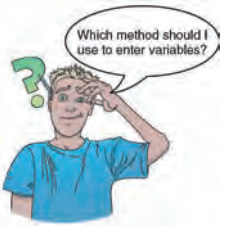
$$R_{\text{CS}}^2 = 1 - \exp\left(\frac{-2LL_{\text{new}} - (-2LL_{\text{baseline}})}{n}\right) \tag{20.25}$$

This statistic never reaches its theoretical maximum of 1, therefore Nagelkerke (1991) suggested the following amendment (**Nagelkerke's** R_N^2):

$$R_{\text{N}}^2 = \frac{R_{\text{CS}}^2}{1 - \exp\left(-\frac{-2LL_{\text{basline}}}{n}\right)} \tag{20.26}$$

Although all these measures differ in their computation (and the answers you get), conceptually they are somewhat the same. They are somewhat similar to the R^2 in linear models with continuous outcomes in that they provide a gauge of the substantive significance of the model.

20.3.8 Model building and parsimony Ⓑ



When you have more than one predictor, you can choose between the same methods to build your model as described for the linear model (Section 9.9.1). Forced entry and hierarchical methods are preferred, but if you are undeterred by the criticisms of stepwise methods then you can choose a forward or backward stepwise method. Stepwise methods work in the same way as we've seen for the linear model, except that different statistics are used to determine whether predictors are entered or removed from the model. The forward method enters predictors based on their score statistic, then assesses removal based on one of three statistics: the likelihood ratio statistic described in Section 19.3.4 (*Forward: LR*), an arithmetically less intense version of the likelihood ratio statistic called the conditional statistic (*Forward: Conditional*), or significance values of the Wald statistic (*Forward: Wald*) above a removal criterion (0.1 by default). The likelihood ratio method is the best removal criterion because the Wald statistic can be unreliable (see Section 20.3.4). Backward stepwise methods begin with all predictors in the model and remove them if their removal is not detrimental to the fit of the model (assessed using the same three methods as the forward approach).

As we have seen before, stepwise methods are best avoided for theory testing but can be used when no previous research exists on which to base hypotheses, and in situations where causality is not of interest and you wish only to find a model to fit your data (see Agresti, 2013; Menard, 2002). If you use a stepwise method then the backward method is preferable because forward methods are more likely to exclude predictors involved in **suppressor effects**.

As with the linear model, it is best to use hierarchical methods and to build models in a systematic and theory-driven way. Although we haven't yet discussed this for linear models, when building a model we should strive for **parsimony**. In a scientific context, parsimony refers to the idea that simpler explanations of a phenomenon are preferable to complex ones. The statistical implication of using a parsimony heuristic is that models be kept as simple as possible. In other words, do not include predictors unless they have explanatory benefit. To implement this strategy we first need to fit the model that includes all potential predictors, and then systematically remove any that don't seem to contribute to the model. This is a bit like a backward stepwise method, except that the decision-making process is in the researcher's hands: they make informed decisions about what predictors should be removed. It's also worth bearing in mind that if you have interaction terms in your model then for an interaction term to be valid *you must retain the main effects involved in the interaction term as well* (even if they don't appear to contribute much).

20.4 Sources of bias and common problems Ⓓ

20.4.1 Assumptions Ⓑ

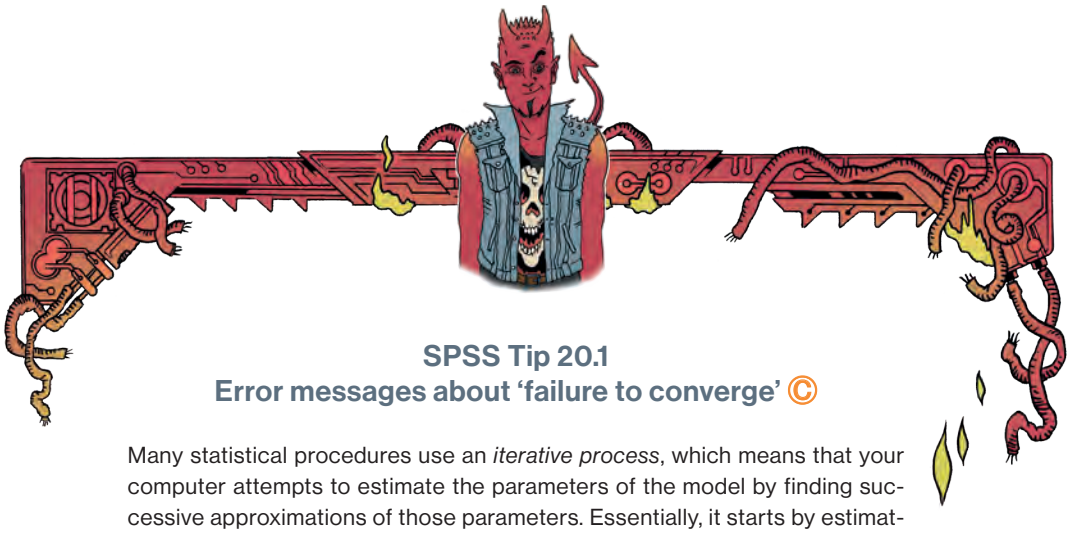
Logistic regression, like any linear model, is open to the sources of bias discussed in Chapter 6 and Section 9.3. In the context of logistic regression, it's worth noting a couple of points about the assumptions of linearity and independence:

- *Linearity*: In the linear model we assume that the outcome has linear relationships with the predictors. As we have seen, logistic regression deals with this by predicting the log odds of

the outcome (Eq. 20.13). The assumption of linearity in logistic regression, therefore, assumes that there is a linear relationship between any continuous predictors and the log odds of the outcome (also known as the *logit* of the probability of the outcome variable). This assumption can be tested by looking at whether the interaction term between the predictor and its log transformation is significant (Hosmer et al., 2013). We will go through an example in Section 20.7.3.

- *Independence of errors*: As for other linear models, we assume independent errors.

Logistic regression also has some unique problems. These are not sources of bias so much as things that can go wrong. Logistic regression parameters are estimated by an iterative process (SPSS Tip 20.1). Sometimes, instead of pouncing on the correct solution quickly, you'll notice nothing happening: the computer begins to move infinitely slowly, or appears to have just got fed up with you asking it to do stuff and has gone on strike. If it can't find a correct solution, then sometimes it does give up, quietly offering you a result which is completely incorrect. Usually this is revealed by implausibly large standard errors. Two situations can provoke this situation, both of which are related to the ratio of cases to variables: incomplete information and complete separation.



SPSS Tip 20.1
Error messages about 'failure to converge' ©

Many statistical procedures use an *iterative process*, which means that your computer attempts to estimate the parameters of the model by finding successive approximations of those parameters. Essentially, it starts by estimating the parameters with a 'best guess'. It then attempts to approximate them more accurately (known as an *iteration*). It then tries again, and then again, and so on through many iterations. It stops either when the approximations of parameters converge (i.e., at each new attempt the 'approximations' of parameters are the same or very similar to the previous attempt), or it reaches the maximum number of attempts (iterations).

Sometimes you will get an error message in the output that says something like 'Maximum number of iterations were exceeded, and the log-likelihood value and/or the parameter estimates cannot converge'. What this means is that SPSS has attempted to estimate the parameters the maximum number of times (as specified in the options) but they are not converging (i.e., at each iteration estimates are quite different). If this happens you should ignore any output, and it might mean that your data are beyond help. You can try increasing the number of iterations that are attempted or make the criteria to assess 'convergence' less strict.

20.4.2 Incomplete information from the predictors Ⓓ

Imagine you're trying to predict lung cancer from smoking and whether you eat tomatoes (which are believed to reduce the risk of cancer). You collect data from people who do and don't smoke, and from people who do and don't eat tomatoes; however, this isn't sufficient unless you collect data from all combinations of smoking and tomato eating. Imagine you ended up with the data in Table 20.2. Observing only the first three possibilities does not prepare you for the outcome of the fourth. You have no way of knowing whether this last person will have cancer or not based on the other data you've collected. Therefore, unless you've collected data from all combinations of your variables it'll be tricky to estimate the model. You can check for incomplete information using a contingency table before you fit the model (I describe how to do this in Chapter 19). While you're checking these tables, look at the expected frequencies in each cell of the table to make sure that they are greater than 1 and no more than 20% are less than 5 (see Section 19.5). This is because the goodness-of-fit tests in logistic regression make this assumption.

Table 20.2 Incomplete contingency table

Do you smoke?	Do you eat tomatoes?	Do you have cancer?
Yes	No	Yes
Yes	Yes	Yes
No	No	Yes
No	Yes	??????

This point applies not only to categorical predictors, but also to continuous ones. Suppose that you wanted to investigate factors related to human happiness. These might include age, gender identity, sexual orientation, religious beliefs, levels of anxiety and even whether a person is right-handed. You interview 1000 people, record their characteristics, and whether they are happy ('yes' or 'no'). Although a sample of 1000 seems quite large, is it likely to include many 80-year-old, highly anxious, Buddhist, left-handed lesbians? Probably not. If you found one or two such people and they were happy, should you conclude that everyone else in the same category is happy? It would, obviously, be better to have several more people in this category to confirm that this combination of characteristics is associated with happiness. One solution is to collect more data.

As a general point, whenever samples are broken down into categories and one or more combinations are empty it creates problems. These will probably be signalled by coefficients that have unreasonably large standard errors. Conscientious researchers produce and check multi-way crosstabulations of all categorical independent variables. Lazy but cautious ones don't bother with crosstabulations but look carefully at the standard errors. Those who don't do either should expect trouble.

20.4.3 Complete separation Ⓓ

A second situation in which logistic regression collapses might surprise you: it's when the outcome variable can be perfectly predicted by one variable or a combination of variables. This situation is known as **complete separation**. Imagine you placed a pressure pad under your door mat and connected it to your security system so that you could detect burglars when they creep in at night. However, because your teenage children (which you would have if you're old enough and rich enough to have security systems and pressure pads) and their friends are often coming home in the middle of the night, and when they tread on the pad you want it to work out the probability that the person is a burglar and not one of your teenagers. Therefore, you could measure the weight of some burglars and some teenagers and use logistic regression to predict the outcome (teenager or burglar) from the

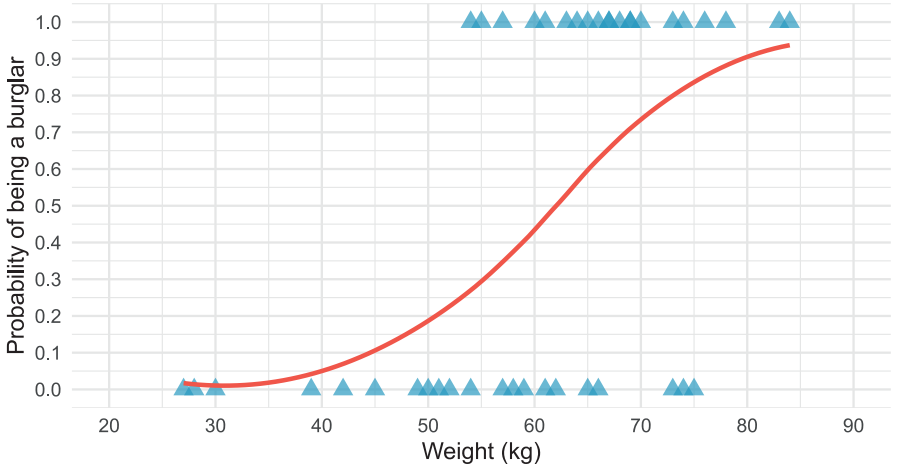


Figure 20.8 An example of the relationship between weight and a dichotomous outcome variable (being a burglar or not). Note that the weights of burglars ($y = 1$) and non-burglars ($y = 0$) overlap.

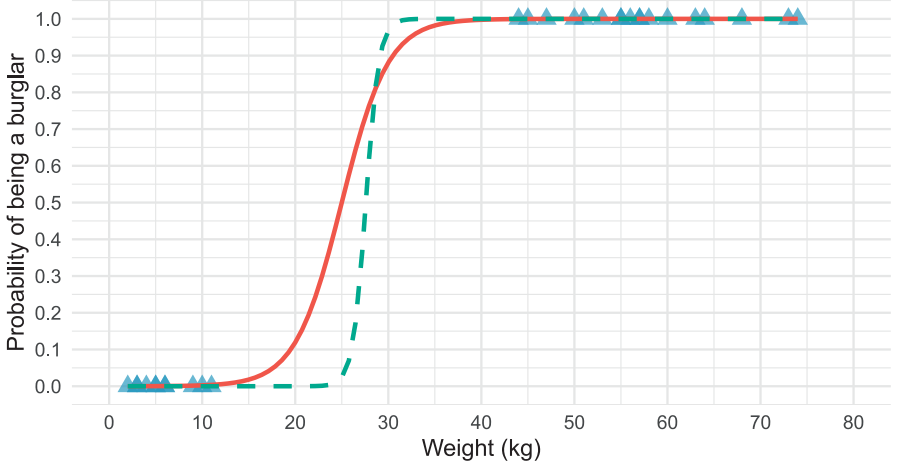


Figure 20.9 An example of *complete separation*. Note that the weights of the burglars ($y = 1$) and non-burglars ($y = 0$) do not overlap.

weight. The plot (Figure 20.8) would show a line of triangles at 0 (the data points for the teenagers you weighed) and a line of triangles at 1 (the data points for burglars you weighed). Note that these lines of triangles overlap (some teenagers are as heavy as burglars). As we've discovered, logistic regression predicts the probability of the outcome using a characteristic S-shaped curve. In this case, at low weights the fitted probability follows the bottom line of the plot, at high weights it starts to follow the top line, and at intermediate values it tries to follow the probability as it changes.

Imagine that we had the same pressure pad, but our teenage children had left home to go to university. We're now interested in distinguishing burglars from our pet cat based on weight. Again, we can weigh some cats and weigh some burglars. This time the plot (Figure 20.9) still has a row of triangles at zero (the cats) and a row at 1 (the burglars) but the rows of triangles do not overlap: there is no burglar who weighs the same as a cat – obviously there were no cat burglars in the sample (groan). This situation is known as complete (or sometimes perfect) separation: the outcome (cats and burglars) can be perfectly predicted from weight (anything less than 15 kg is a cat, anything more than 40 kg



Cramming Sam's Tips
Issues in logistic regression

- In logistic regression, we assume some of the same things as the linear model.
- The linearity assumption is that each predictor has a linear relationship with the log odds of the outcome variable.
- If we created a table that combined all possible values of all variables then we should ideally have some data in every cell of this table. If you don't then watch out for big standard errors.
- If the outcome variable can be predicted perfectly from one predictor variable (or a combination of predictor variables) then we have *complete separation*. This problem creates large standard errors too.



is a burglar). If we try to calculate the probabilities of the outcome given a certain weight then we run into trouble. When the weight is low, the probability is 0, and when the weight is high, the probability is 1, but what happens in between? We have no data between 15 and 40 kg on which to base these probabilities. The figure shows two possible probability curves that we could fit to these data, one steeper than the other. Either one of these curves is valid based on the data we have available. The lack of data means that the computer will be uncertain about how steep it should make the intermediate slope and it will try to bring the centre as close to vertical as possible, but its estimates veer unsteadily towards infinity (hence large standard errors). Complete separation often arises when too many variables are fitted to too few cases. Often the only satisfactory solution is to collect more data, but sometimes a neat answer is found by using a simpler model.

20.5 Binary logistic regression Ⓑ

20.5.1 An example that will make you feel eel Ⓑ

A hobby of mine is unearthing bizarre academic papers (really, if you find any, please email them to me) – it's amazing what you find. A research paper by Lo et al., (2004) describes the case of a 50-year-old man who reported to the emergency department of a hospital with abdominal pain. A physical examination revealed peritonitis, so they X-rayed the man's abdomen. The X-ray revealed the shadow of an eel. The authors don't directly quote the man's response to this news, but I like to imagine it was something to the effect of 'Oh, that! Erm, yes, well, I didn't think it was terribly relevant to my abdominal pain so I didn't mention it, but I did insert an eel into my

anus earlier today. Do you think *that's* the problem?' He probably didn't say that, but he did admit that he inserted the eel to 'relieve constipation'.



I have a lively imagination, and my thoughts turn to the poor eel. There it was, minding its own business swimming about, thinking to itself 'Today seems like a nice day, there are no eel-eating sharks about, the sun is out, the water is nice, what could possibly go wrong?' The next thing it knows, it's being shoved up a man's anus. 'I didn't see that coming,' thinks the eel. It finds itself in a tight, dark tunnel, there's no light, there's a distinct lack of water compared to its usual habitat, and it's scared. Its day has gone *very* wrong. It considers its fate and, noticing that the walls of the prison cell are fairly soft, it does what any self-respecting eel would do: it decides, 'OK, I'll *eat* my way out of here'. Unfortunately, the eel didn't make it, but it went out with a fight (there's a fairly unpleasant photograph in the article of the eel biting the splenic flexure). Lo et al. conclude that 'Insertion of a live animal into the rectum causing rectal perforation has never been reported. This may be related to a bizarre healthcare belief, inadvertent sexual behaviour, or criminal assault. However, the true reason may never be known.' Quite.

This tale is grim[6] and bizarre.[7] I'm no medic, but if constipation is a failure to empty the bowel, inserting more stuff up there seems, at best, a counterintuitive remedy. But perhaps I am showing my ignorance, and of course it is an empirical question whether an eel up the anus can cure constipation. To answer the question, we could do a randomized controlled trial of eel therapy. Our outcome variable would be 'constipated' vs. 'not constipated', which is a dichotomous variable that we're trying to predict. The main predictor variable would be the intervention condition (eel up the anus vs. waiting list/no treatment), but we might also factor in how many days the patient had been constipated before treatment. This scenario is perfect for logistic regression (but not for eels).

Some statistics lecturers don't share my unbridled joy at discussing eel-created rectal perforations with students, so in the data file (**eel.sav**) I have used general variables names and descriptions:

- *Outcome* (dependent variable): **cured** (cured or not cured).
- *Predictor* (independent variable): **intervention** (intervention or no treatment).
- *Predictor* (independent variable): **duration** (the number of days before treatment that the patient had the problem).

In doing so, your tutor can adapt the example to something more palatable if they wish, but you will secretly know that it's all about having eels up your bum.

6 As it happens, it isn't an isolated grim tale. Through this article I found myself hurtling down a wormhole of rectal insertion that involved a very large stone (Sachdev, 1967), a test tube (Hughes et al., 1976), a baseball (McDonald & Rosenthal, 1977), an aerosol deodorant can, hose pipe, iron bar, broomstick, penknife, marijuana, banknotes, blue plastic tumbler, vibrator and primus stove (Clarke et al., 2005), and a toy pirate ship, with or without pirates I'm not sure (Bemelman & Hammacher, 2005). I encourage you to send me bizarre research, but if it involves objects in the rectum then probably don't, unless someone has managed to put Buckingham Palace up there.

7 Possibly not as bizarre as the study I subsequently found of a 14-year-old boy who reported to hospital unable to urinate (Vezhaventhan & Jeyaraman, 2007). A small fish was discovered in his bladder that had swum up his penis while he was having a wee while cleaning a fish tank. Yes, of course it did.

20.5.2 Building a model Ⓐ

In Section 20.3.8 we discussed the idea of building models based on the principle of parsimony. In this example, we have three potential predictors: **intervention**, **duration** and the **intervention** × **duration** interaction. The most complex model we can fit would be to include all these variables, but the parsimony principle would suggest building up to this model in steps, looking at what predictors didn't improve the model and going back to a simpler model that doesn't include them. The key effect of interest is whether the intervention has an effect, so the first model would have only **intervention** as a predictor. Figure 20.10 shows how we then build this model up by adding in the other main effect of **duration** (model 2) and then the interaction term (model 3). Our job is to determine which of these models best fits the data, while adhering to the general idea of parsimony. If adding the interaction term doesn't improve the model then we roll back to model 2 as our final model, and if **duration** doesn't add anything we roll back to the first model. Remember, though, that if you want to look at an interaction *you must include any main effects involved in that interaction in the model even if the main effects weren't significant*. In this example, if we want to assess the contribution of the **duration** × **intervention** interaction, we must also include **intervention** and **duration** in the model involving the interaction.

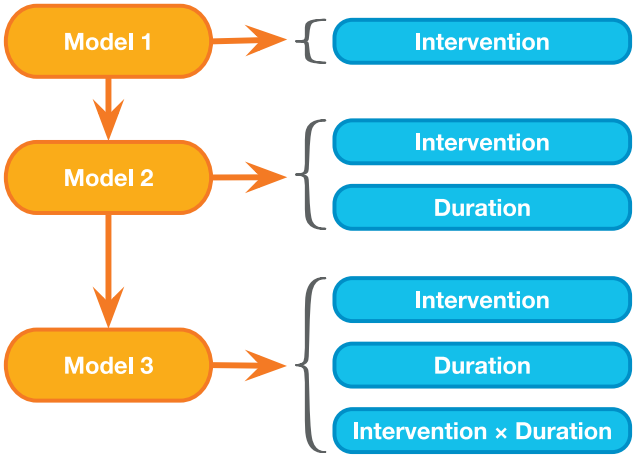


Figure 20.10 Building models based on the principle of parsimony

20.5.3 Logistic regression: the general procedure Ⓐ

Figure 20.11 shows the general process of conducting logistic regression. First, we run an initial hierarchical analysis to fit competing models and decide which one is best (i.e., the three models identified in Section 20.5.2). Having done this, we refit the model that we chose but save diagnostic statistics and inspect them to look for signs of bias (outliers and influential cases). We then check for linearity of the logit (in fact, it's a good idea to do this first, but it's a little complicated so I want to deal with it later in the chapter), and check for multicollinearity.

20.5.4 Data entry Ⓐ

The data should be entered as for the linear model: for this example, the data are arranged in three columns (one for each variable). Note in **eel.sav** that both categorical variables are coding variables (Section 4.6.5) in which numbers specify categories. In general, code the event occurring as 1 and not occurring as 0; in this example, 1 represents being cured and 0 represents not being cured. For the no treatment control is the baseline and so is coded with 0, whereas the intervention is coded as 1.

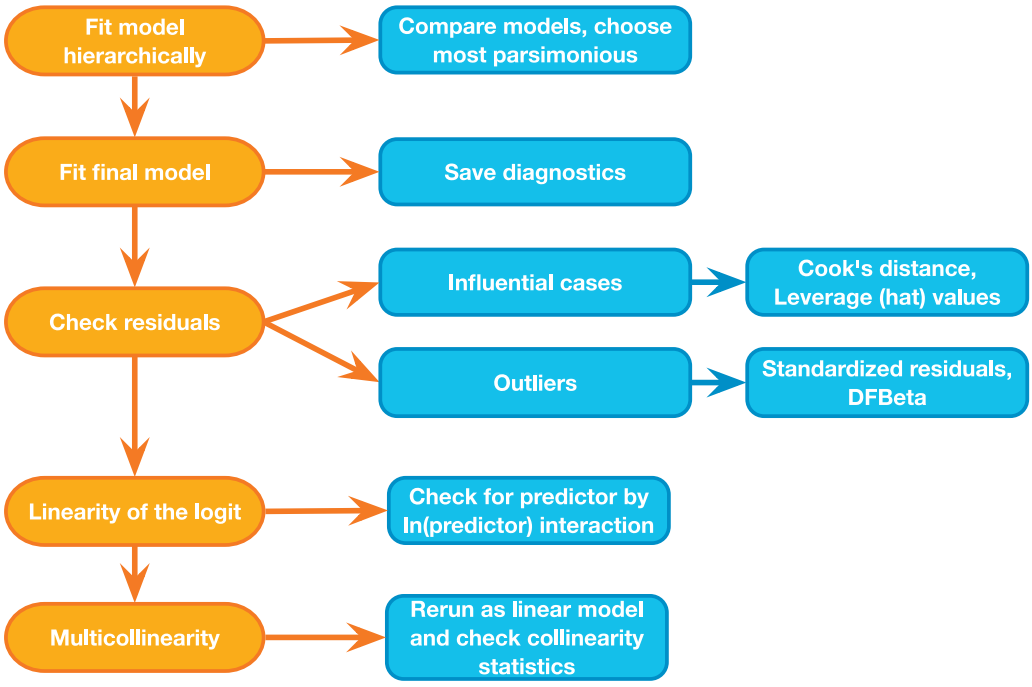


Figure 20.11 The process of fitting a logistic regression model

20.5.5 Building the models using SPSS Ⓑ

To build the three models in Figure 20.10 select *Analyze* > *Regression* > Binary Logistic... to access the dialog box in Figure 20.12. The models need to be specified in blocks, with each block adding a new variable to the model. To specify the first model, drag the outcome variable (**cured**) to the *Dependent* box (or select it and click ▸). There is a box labelled *Covariates* for specifying the predictor variables. Model 1 has one predictor variable (the main effect of **intervention**), so drag this variable to the *Covariates* box (or click ▸). Make sure that the *Method* is set to Enter. That's model 1 sorted.

Click Next, which clears the *Covariates* box and relabels it to indicate which block of the model you are in. To add a new predictor in a second block. Model 2 adds the main effect of **duration**, so drag this variable to the *Covariates* box (or click ▸). Because **intervention** is already forced into the model in block 1, this second block creates a model that includes both **intervention** and **duration** (i.e., model 2). Again, make sure that the method is set to Enter. That's model 2 done.

To specify the final model we need to add in the **duration × intervention** interaction. Click Next to create a new block (block 3), then select **duration** and **intervention** simultaneously by holding down the *Ctrl* (Windows) or *Cmd* (Mac OS) key when selecting them and click >a*b>. This should add the interaction term. Model 3 is complete. To move between models use the Previous and Next buttons.

20.5.6 Method of regression Ⓑ

For each of the models that we specified you can select a method (see Section 20.5.6) of variable entry by clicking Enter and selecting a method in the drop-down list. We are doing this analysis hierarchically so we want to use the Enter method in each block (i.e., you don't need to change anything) – just bear in mind that other methods exist.

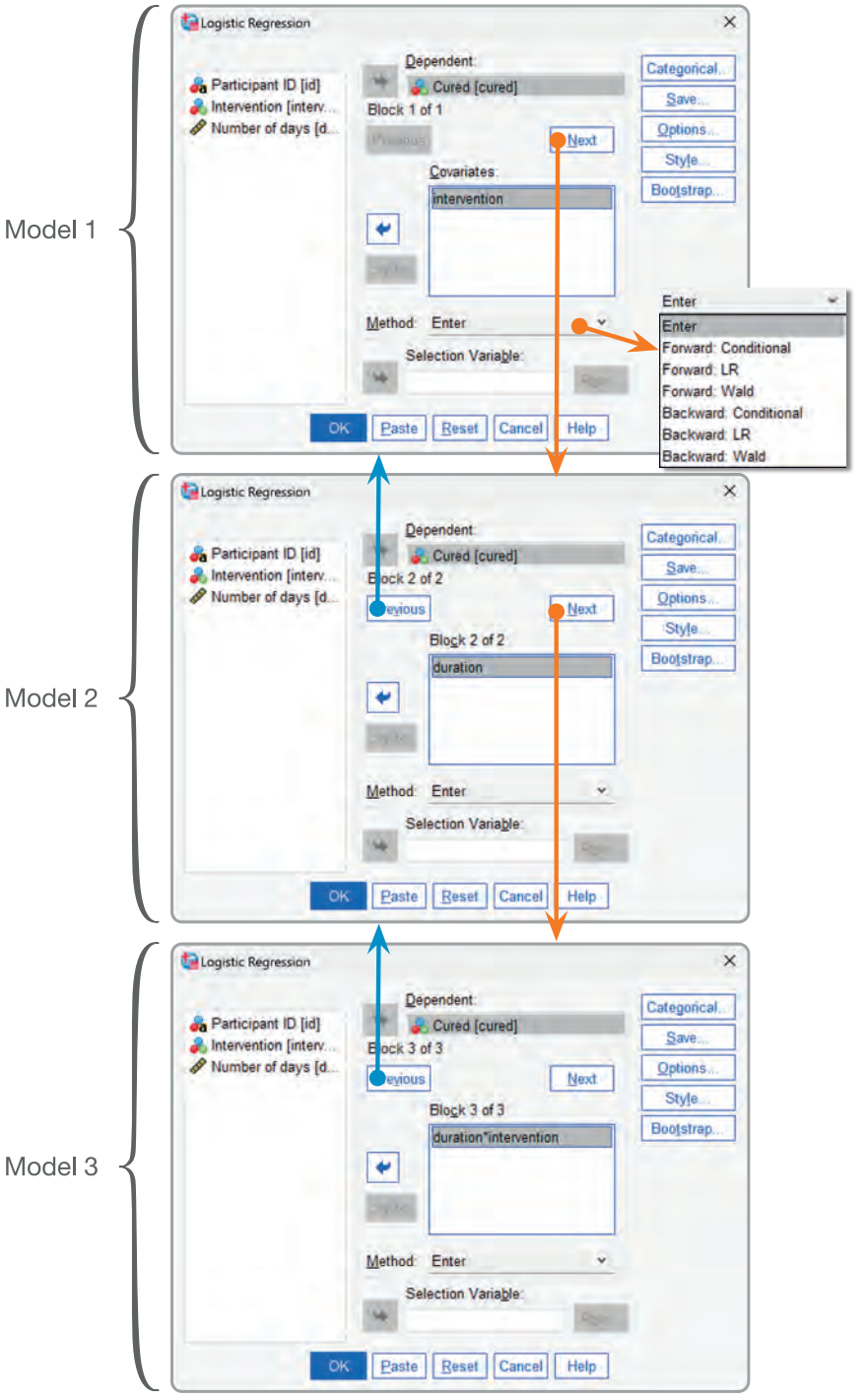


Figure 20.12 Specifying models using the *Logistic Regression* dialog box

20.5.7 Categorical predictors Ⓑ

Click Categorical... in the main dialog box to activate the dialog box in Figure 20.13, which you use to specify categorical predictors. Drag any categorical predictor variables in your model (in this example we have only one, **intervention**) to the *Categorical Covariates* box (or click ➔). There are different

ways to code categorical predictors, and we have discussed some of them (e.g., Section 12.5.4). By default, *Indicator* coding is used, which is standard dummy variable coding where you choose either the first or last category as the baseline category. To use a different type of coding click Indicator and select from simple contrasts, difference contrasts, Helmert contrasts, repeated contrasts, polynomial contrasts and deviation contrasts in the drop-down list (see Table 12.8). Let's use standard dummy coding (indicator) for this example. We need to decide whether to use the First or Last category as a baseline. If you had a categorical predictor with more than two categories then you should either use the highest number to code your control category in the data editor and select Last, or use the lowest number to code your control category and select First. In our data, I coded 'cured' as 1 and 'not cured' (our control category) as 0; therefore, select First and then Change so that the completed dialog box looks like Figure 20.13.

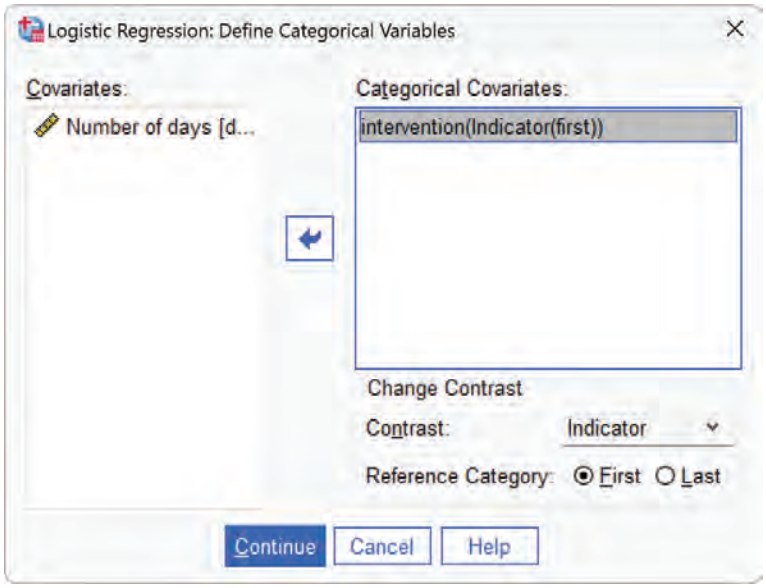


Figure 20.13 Defining categorical variables in logistic regression

20.5.8 Comparing the models Ⓑ

Before we look at some of the other options, let's compare our models with just the basic settings (which is all we need to assess their fit). Having selected the options that I have already described, click OK in the main dialog box. Output 20.1 tells us both how we coded our outcome variable (it reminds us that 0 = not cured, and 1 = cured)[8] and how it has coded the categorical predictors (the parameter codings for **intervention**). We chose indicator coding and so the coding is the same as the values in the data editor (0 = no treatment, 1 = treatment). If *deviation* coding had been chosen then the coding would have been –1 (treatment) and 1 (no treatment). With a *simple* contrast, if First was selected as the reference category the codes would have been –0.5 (**intervention** = no treatment) and 0.5 (**intervention** = treatment) and if Last was selected the signs of the codes would be reversed. The parameter codes are important for calculating the probability of the outcome variable ($P(Y)$), but we will come to that later.

8 These values are the same as the data editor, so this table might seem pointless; however, had we used codes other than 0 and 1 (e.g., 1 = not cured, 2 = cured) then these codes change to zeros and ones and this table informs you of which category is represented by 0 and which by 1, which is important for interpretation.



Oditi's Lantern
Logistic regression

'I, Oditi, believe that my loyal brethren will find it difficult to master the secrets within the data if their bowels are creaking at the seams because the curse of constipation has afflicted them. You could do the magic dance of the turtle head and hope that it brings you relief, but it is my belief that to remove an intestinal log we need log-istic regression. Stare into my lantern and feel immediate relief.'



Dependent Variable Encoding

Original Value	Internal Value
Not cured	0
Cured	1

Categorical Variables Codings

		Frequency	Parameter coding (1)
Intervention	No treatment	56	.000
	Intervention	57	1.000

Output 20.1

Output 20.2 shows the overall model summary statistics for each of the three models. The table labelled *Omnibus Tests of Model Coefficients* includes the chi-square statistic (which is related to *–2LL*) for the model overall (*Model*) and the change since the previous model (*Block*). Model 1 yields a chi-square of 9.926, which is highly significant, $p = 0.002$. We can compare the models using Eq. 20.21. For model 1, the *–2LL* would have been compared to that obtained from a model that included only the intercept, so we're comparing a model that includes **intervention** against a model that has no predictors. The chi-square tells us that the model has improved significantly by adding **intervention** as a predictor.

In model 2, we added the effect of **duration**, and this model is a significant fit of the data because the *Model* chi-square in the table labelled *Omnibus Tests of Model Coefficients* is significant, $\chi^2(2) = 9.928$,

p = 0.007. However, we're not interested in the model overall because the previous model was also a significant fit; we're interested in the *improvement* of model 2 over model 1, and this information is given by the chi-square for *Block*. The *Block* chi-square tells us about the change in chi-square in this block: it is the *change* in the chi-square resulting from adding **duration** to the model. The value is obtained by taking the difference between the model chi-square for the two models (in this case 9.928 – 9.926 = 0.002). This change is very non-significant, $\chi^2(1) = 0.002$, $p = 0.964$, indicating that adding **duration** to the model had virtually no effect on the fit (the chi-square has hardly changed).

In model 3, we added the **intervention × duration** interaction. Again, this model is a significant fit of the data because the *Model* chi-square is significant, $\chi^2(3) = 9.989$, $p = 0.019$. However, as with model 2, we're interested in the *improvement* of model 3 over the previous one (the chi-square for *Block*). As before, the value is obtained by taking the difference between the model chi-square for the two models (in this case 9.989 – 9.928 = 0.061). This change is very non-significant, $\chi^2(1) = 0.061$, p = 0.805, indicating that adding the interaction term has had virtually no effect on the fit.

Model 1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	9.926	1	.002
	Block	9.926	1	.002
	Model	9.926	1	.002

Model 2

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	.002	1	.964
	Block	.002	1	.964
	Model	9.928	2	.007

Model 3

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	.061	1	.805
	Block	.061	1	.805
	Model	9.989	3	.019

Output 20.2

We could, if we wanted, look at the difference between models 1 and 3 as well using Eq. 20.21. The result is:

$$\chi^2 = \chi^2_{\text{model 3}} - \chi^2_{\text{model 1}} = 9.989 - 9.926 = 0.063$$
$$df = df_{\text{model 3}} - df_{\text{model 1}} = 3 - 1 = 2 \qquad (20.27)$$

We could compare this against the critical values for the chi-square distribution with 2 degrees of freedom, but we don't need to because 0.063 is so small it barely warrants turning the pages to confirm what we already know: **duration** and the **duration × intervention** interaction add nothing to the model. Based on this comparison, we would choose model 1.

20.5.9 Refitting the model Ⓐ

Comparing the models shows that **duration** and the **duration × intervention** interaction add little and that we should proceed with model 1. We set up the main dialog box as we did before for model 1, with only **intervention** as a predictor (the top dialog box in Figure 20.12). Set the same options as before, but we'll now also get some more detailed information about the model.

20.5.10 Obtaining residuals Ⓑ

To save residuals (see Section 9.3) click Save... in the main dialog box to access the dialog box in Figure 20.14. Most of the options are similar to what we encountered for the linear model (Section 9.10.4). The *predicted probabilities* and *predicted group memberships* are unique to logistic regression. The predicted probabilities are the probabilities of Y occurring (derived from Eq. 20.3), given the values of each predictor for a given case. The predicted group membership tells us to which of the two outcome categories a participant is most likely to belong, based on the model. The group memberships are based on the predicted probabilities, and I will explain these values in due course. The deviance residuals are the residuals from the model. As a bare minimum select the options in Figure 20.14. Note that these variables won't save if you activate bootstrapping (see below).

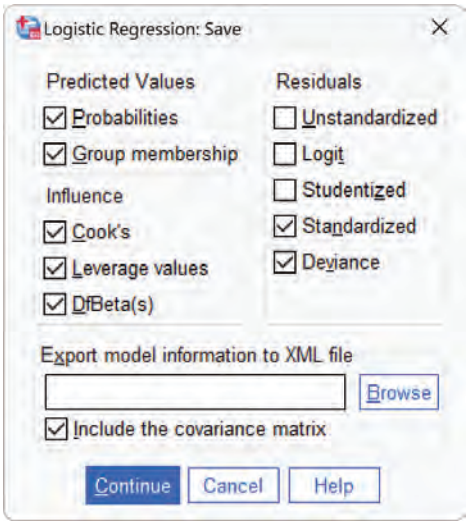


Figure 20.14 Dialog box for obtaining residuals for logistic regression

20.5.11 Further options Ⓑ

Finally, click Options... in the main dialog box to obtain the dialog box in Figure 20.15. For the most part, the default settings are fine. I mentioned in Section 20.5.6 that when a stepwise method is used there are default criteria for selecting and removing predictors from the model. These default settings are displayed under *Probability for Stepwise*. The probability thresholds can be changed, but there is really no need. Another default is to arrive at a model after a maximum of 20 iterations (SPSS Tip 20.1). Unless you have a very complex model, 20 iterations will be enough. Linear models (like logistic regression) usually contain a constant (the b_0 in Eq. 20.5) and by default it is included, but you can deselect this option and force the model to pass through the origin (i.e., Y is 0 when X is 0). Normally we don't want to do this.

A useful way to assess the fit of the model to the observed data is a histogram of the actual and predicted values of the outcome variable (select Classification plots). It is possible to do a Casewise listing of residuals for any cases for which the standardized residual is greater than 2 standard deviations (this default value is sensible, but change it if you like), or for all cases. I recommend a more thorough examination of residuals, but this option can be useful for a quick inspection. Select CI for exp(B) to produce a 95% confidence interval for the odds ratio (you can change it, but 95% is what is conventionally reported). The Hosmer-Lemeshow goodness-of-fit statistic is used to assess how well the chosen model fits the data. The remaining options are fairly unimportant: you can choose to display all statistics and plots at each stage of an analysis (the default), or only after the final model

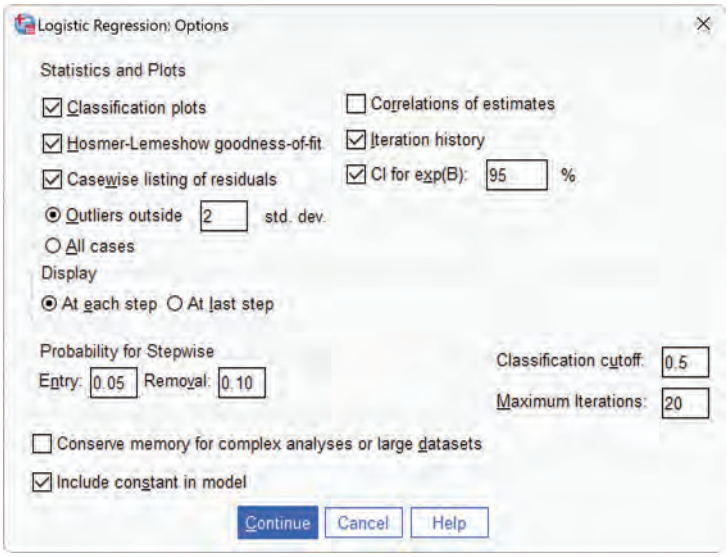


Figure 20.15 Dialog box for logistic regression options

has been fitted. Finally, you can display a correlation matrix of parameter estimates for the terms in the model (*Correlations of estimates*) – the practical function of doing this is lost on most of us. You can display coefficients and log-likelihood values at each iteration of the parameter estimation process (☑ Iteration history), which is useful because it's the only way you can display the initial *–2LL*, and we need this value if we want to compute *R*. When you have selected the options I've just described, click OK and watch the output spew forth.

20.5.12 Bootstrapping Ⓑ

If you use forced entry then you can bootstrap your model by clicking on Bootstrap... in the main dialog box and selecting appropriate options. This function doesn't work with stepwise methods so the button will be inactive unless you choose Enter. It's also worth remembering that if you activate bootstrapping then any residuals that you have asked to be saved won't be saved. This is annoying because it means that to bootstrap the model parameters we have to run the analysis again, but deselecting the options in Figure 20.14 (I would also deselect the options in Figure 20.15 just to keep things simple). Let's do this: select the usual bootstrapping options (Section 6.10.8) and rerun the model by clicking OK.

20.5.13 Listing residuals Ⓑ

As we saw for the linear model, residuals are saved in columns in the data editor. We can list them in the output viewer using the *Analyze* > *Reports* > Case Summaries... dialog box (see Section 9.11.6).

20.6 Interpreting logistic regression Ⓑ

20.6.1 Block 0 Ⓑ

The output is split into two blocks: the model before (block 0) and after (block 1) **intervention** is included. As such, block 1 is the main bit in which we're interested, but Output 20.3 from block 0 is useful because it tells us the initial *–2LL* value (154.084), which we'll use later, so don't forget it.[9]

9 If you can't see this output it'll be because you didn't select ☑ Iteration history in Figure 20.10.

Iteration History[a,b,c]

Iteration		−2 Log likelihood	Coefficients Constant
Step 0	1	154.084	.301
	2	154.084	.303
	3	154.084	.303

a. Constant is included in the model.

b. Initial −2 Log Likelihood: 154.084

c. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Output 20.3

20.6.2 Model summary Ⓑ

With **intervention** included in the model a patient is classified as being cured or not based on whether they had an intervention or not. To illustrate this principle, look at the crosstabulation for the variables **intervention** and **cured** in Table 20.3.[10] The model applies this classification table to decide whether a patient was cured or not based on what intervention they had. For example, there were 57 patients who had the intervention, so the model predicts that these 57 patients were cured. From Table 20.3 we can see that the model is correct for 41 patients, but misclassifies 16 patients as 'cured' who were not. Similarly, the model predicts that all the 56 patients who received no treatment were not cured and in doing so it correctly classifies 32 patients but misclassifies 24 patients as 'not cured' who were. The classification table in Output 20.4 shows this pattern of predicted values and observed values, with the diagonal values being the correctly classified cases and the off-diagonal elements being misclassified cases. This output also tells us that the model correctly classifies 66.7% of not cured cases and 63.1% of cured cases and the overall accuracy is 64.6% (the weighted average of these two values).[11]

Table 20.3 Crosstabulation of intervention with outcome status (cured or not)

		Intervention	
		No Treatment	Intervention
Cured	Not Cured	32	16
	Cured	24	41
	Total	56	57

10 The dialog box to produce this table can be obtained by selecting *Analyze* > *Descriptive Statistics* > Crosstabs....

11 Notice that the classification tables in the outputs for models 2 and 3 that we fitted earlier are identical to the one reported for this model, which means that adding **duration** and the interaction term did not lead to even a single person being more accurately classified than when we include only **intervention** as a predictor.

Output 20.4 also shows summary statistics for the model,[12] but we should also look at the table that we already inspected in Output 20.2 that showed the chi-square statistic for the model (remember it was 9.926 and highly significant, $p = 0.002$). This chi-square statistic is derived from Eq. 20.21 and is the difference between the current $-2LL$ (which for this model is 144.158) and the baseline $-2LL$ (i.e., the value before intervention was added, which is reported in as 154.084 in Output 20.3): 154.084 – 144.158 = 9.926. Output 20.4 also tells us the values of Cox and Snell's and Nagelkerke's R^2, but we will discuss these a little later.

Model Summary

Step	−2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	144.158[a]	.084	.113

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Cured		
	Observed		Not cured	Cured	Percentage Correct
Step 1	Cured	Not cured	32	16	66.7
		Cured	24	41	63.1
	Overall Percentage				64.6

a. The cut value is .500

Output 20.4

Output 20.5 tells us the estimates for the b-values, confidence intervals, p-values and odds ratio for the predictors in the model (namely **intervention** and the constant). We can replace the b-values in Eq. 20.5 to establish the probability that a case falls into a certain category, or into Eq. 20.13 to estimate the predicted log odds of being cured. As we saw earlier, these $\hat{b}$s also quantify the relationship between the predictor and outcome. As intervention increases by one unit, the log odds of being cured increase by 1.229 units. We'll pick this effect apart in more detail in the next section.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Intervention(1)	1.229	.400	9.447	1	.002	3.417	1.561	7.480
	Constant	−.288	.270	1.135	1	.287	.750		

a. Variable(s) entered on step 1: Intervention.

Bootstrap for Variables in the Equation

			Bootstrap[a]				
						BCa 95% Confidence Interval	
		B	Bias	Std. Error	Sig. (2-tailed)	Lower	Upper
Step 1	Intervention(1)	1.229	.019	.412	.002	.419	2.123
	Constant	−.288	.001	.272	.286	−.836	.251

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Output 20.5

12 If you use bootstrapping you'll notice a load of guff below this table. Ignore it.

The output also tells us the Wald statistic (Eq. 20.18),[13] from which a *p*-value can be computed. If the coefficient is significantly different from zero then we can assume that the predictor is making a significant contribution to the prediction of the outcome (*Y*). For these data, the Wald statistic indicates that having the intervention (or not) is a significant predictor of whether the patient was cured because the *p*-value is 0.002, which is less than the conventional threshold of 0.05 and also the less conventional one of 0.005.

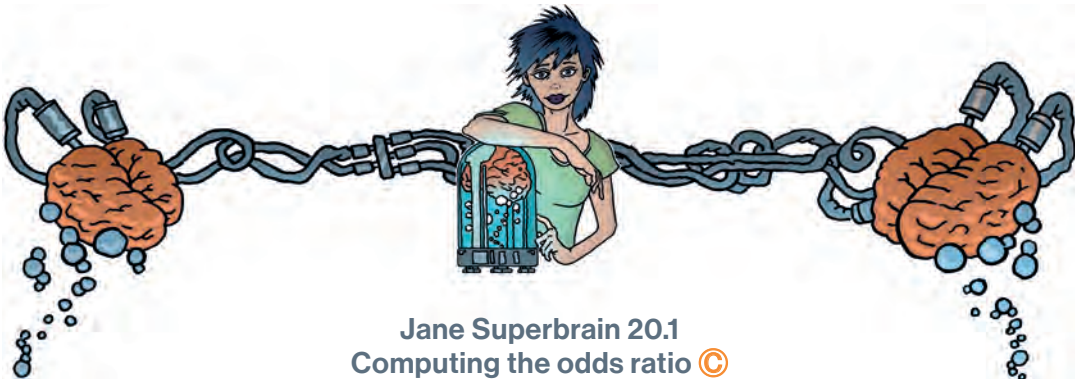
If you followed my advice and reran the model with bootstrapping activated you'll get the table labelled *Bootstrap for Variables in the Equation* in Output 20.5. This table reports the *b* values again but estimates the standard error using bootstrap resampling (Section 6.10.8). The change in the standard error results in a *p*-value that doesn't affect the interpretation. The bootstrap confidence interval for the *b*-values tells that the population value of *b* falls between 0.42 and 2.12 (assuming this sample is one of the 95% for which the confidence interval contains the population value). This interval doesn't include zero, so we might conclude that there is a genuine positive relationship between having the intervention (or not) and being cured (or not). The bootstrap confidence intervals and *p*-value will differ slightly every time you run the analysis, but they are nevertheless robust to violations of the underlying assumptions of the test.

20.6.3 The odds ratio, $\exp(\hat{b})$

We've seen that as intervention increases by one unit, the log odds of being cured increase by 1.229 units. I've also mentioned that 'log odds' make no sense to most of us and that instead we can work with $\exp(\hat{b})$, which is $\hat{b}$ expressed as odds rather than *log* odds (Section 20.3.3). The value can be found as *Exp(B)* in Output 20.5, and it is literally the exponential of the $\hat{b}$ for the predictor, in this case $e^{1.229} = 3.42$. We noted earlier that $\exp(\hat{b})$ is often called the odds ratio and it represents the ratio of the odds after a unit change in the predictor to the odds before that change (Eq. 20.17). To get an intuition about what the odds ratio represents see Jane Superbrain Box 20.1.

In the options (see Section 20.5.11), we requested a confidence interval for *Exp(B)*, and it can also be found in the output. Assuming the current sample is one of the 95% for which the confidence interval contains the true value, then the population value of $\exp(\hat{b})$ lies between 1.56 and 7.48. However, our sample could be one of the 5% that produces a confidence interval that 'misses' the population value. The important thing is that the interval doesn't contain 1 (both values are greater than 1). The value of 1 is important because it is the threshold at which the direction of the effect changes (Section 20.3.3): values greater than 1 mean that as the predictor variable increases, so do the odds of (in this case) being cured, but values less than 1 mean that as the predictor variable increases, the odds of being cured *decrease*. If the confidence interval contains 1 then the population value might be one that suggests that the intervention improves the probability of cure, or it might be a value that suggests that the intervention decreases the probability of being cured. The fact that both limits of our interval are above 1 suggests (with the usual assumptions) that the direction of the relationship that we have observed is true in the population (i.e., it's likely that having an intervention compared to not having it increases the odds of being cured). If the lower limit had been below 1 then it would tell us that there is a chance that in the population the direction of the relationship is the opposite to what we have observed. This would mean that it's ambiguous whether the intervention increases the odds of being cured.

13 As we have seen, the Wald statistic is *b* divided by its standard error (1.229/0.40 = 3.0725); however, SPSS quotes the Wald statistic squared, $3.0725^2 = 9.44$.



Jane Superbrain 20.1
Computing the odds ratio ©

To calculate the odds ratio, we need to calculate the odds of a patient being cured dependent on which intervention they received. Adapting Eq. 20.11 we know that:

$$\text{odds} = \frac{P(\text{cured})}{P(\text{not cured})}$$

To calculate the odds, we therefore need the probabilities of being cured and not cured. We can obtain these by plugging values of $\tilde{b}_0$ – 0.288, $\tilde{b}_1$ = 1.229 and X_i from the output into Eq. 20.3. To get X_i look at the parameter coding at the beginning of the output, which tells us that patients who did not have the intervention were coded with a 0 and those that did were coded with 1. So we have X_i = 0 (no treatment) and X_i = 1 (intervention).

Using this information the odds of being cured *if you are in the no treatment* group are:

$$\begin{aligned} P(\text{cured}) &= \frac{1}{1+e^{-(b_0 + b_1 X_i)}} \\ &= \frac{1}{1+e^{-[-0.288 + (1.229 \times 0)]}} \\ &= 0.428 \\ P(\text{not cured}) &= 1 - P(\text{cured}) \\ &= 1 - 0.428 \\ &= 0.572 \\ \text{Odds} &= \frac{0.428}{0.572} \\ &= 0.748 \end{aligned} \tag{20.28}$$

Now, we calculate the odds *after the predictor variable has changed by one unit*. In this case, because the predictor variable is dichotomous it means calculating the odds of a patient being cured given that they have had the intervention. The value of the intervention variable changes to X_i = 1 (rather than 0), but everything else remains the same. The resulting odds are:

$$\begin{aligned} P(\text{cured}) &= \frac{1}{1+e^{-[-0.288 + (1.229 \times 1)]}} \\ &= 0.719 \\ P(\text{not cured}) &= 1 - P(\text{cured}) \\ &= 1 - 0.719 \\ &= 0.281 \\ \text{Odds} &= \frac{0.719}{0.281} \\ &= 2.559 \end{aligned} \tag{20.29}$$

We now have the odds of being cured before and after a unit change in the predictor variable. The odds ratio is the ratio of these values (Eq. 20.17). The result is 3.42:

$$\begin{aligned}\text{odds ratio} &= \frac{\text{odds}_{\text{after a unit change in the predictor}}}{\text{odds}_{\text{original value of predictor}}} \\ &= \frac{2.559}{0.748} \\ &= 3.42\end{aligned} \tag{20.30}$$



20.6.4 Analogues of R^2 ©

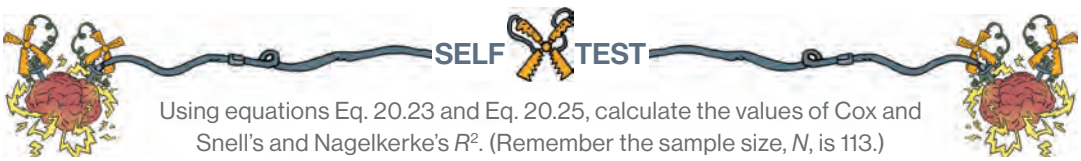
In Section 20.3.7 we saw that we could calculate an analogue of R using Eq. 20.22. For these data, z^2 (the Wald statistic in the output) and its *df* are in Output 20.5 (9.447 and 1, respectively) and the baseline $-2LL$ was 154.084 (Output 20.3). Therefore, R is 0.22:

$$R = \sqrt{\frac{9.447 - (2 \times 1)}{154.084}} = 0.22 \tag{20.31}$$

The Hosmer and Lemeshow's measure (R^2_L) from Section 20.3.7 is 0.06:

$$\begin{aligned}R^2_L &= \frac{(-2LL_{\text{baseline}}) - (-2LL_{\text{new}})}{-2LL_{\text{baseline}}} \\ &= \frac{154.084 - 144.158}{154.084} \\ &= 0.06\end{aligned} \tag{20.32}$$

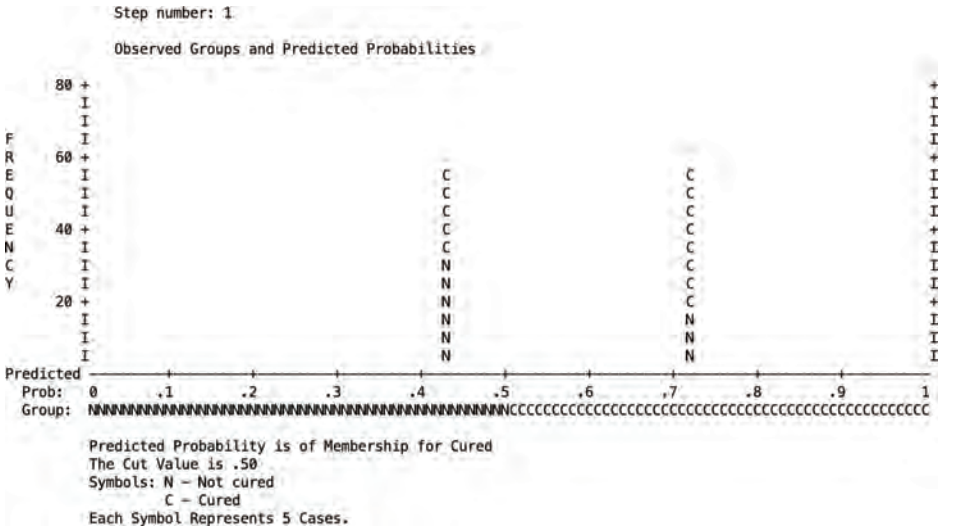
This is the same as dividing the model chi-square after **intervention** has been entered into the model (9.93) by the baseline $-2LL$ (before any variables were entered). The resulting value of 0.06 is different than the one we would get by squaring R above ($R^2 = 0.22^2 = 0.05$). Two other measures of R^2, which were described in Section 20.3.7, are in Output 20.4: Cox and Snell's measure (0.084) and Nagelkerke's adjusted value (0.113). All of these R^2 values differ, but they give us an approximate effect size measure for the model.



Using equations Eq. 20.23 and Eq. 20.25, calculate the values of Cox and Snell's and Nagelkerke's R^2. (Remember the sample size, N, is 113.)

20.6.5 Classification plots Ⓑ

Output 20.6 displays the classification plot, which is a histogram of the predicted probabilities of a patient being cured. If the model perfectly fits the data, then this histogram should show all the cases for which the event has occurred on the right-hand side, and all the cases for which the event hasn't occurred on the left-hand side. In this example, all the patients who were cured should appear on the right and all those who were not cured should appear on the left. Because the only predictor is dichotomous there are just two columns of cases on the plot. If the predictor is a continuous variable, the cases will be spread across many columns. As a rule of thumb, the more the cases cluster at each end of the graph, the better; such a plot would show that when the outcome did occur (i.e., the patient was cured) the predicted probability of the event occurring is also high (i.e., close to 1). Likewise, at the other end of the plot it would show that when the event didn't occur (i.e., the patient still had a problem) the predicted probability of the event occurring is also low (i.e., close to 0). This situation represents a model that correctly predicts the observed outcome data. If lots of points cluster in the centre of the plot then it shows that for many cases the model is predicting a probability of 0.5; in other words, there is close to a 50:50 chance that these cases are predicted correctly by the model – you could predict these cases as accurately as the model by tossing a coin. In Output 20.6 cured cases are predicted relatively well by the model (the probabilities are not that close to 0.5), but for not cured cases the model is less good (the probability of classification is only slightly lower than 0.5). Also, a good model will ensure that few cases are misclassified; for these data there are a few Ns (not cured) appearing on the cured side, but more worryingly there are quite a few Cs (cured) appearing on the N side.

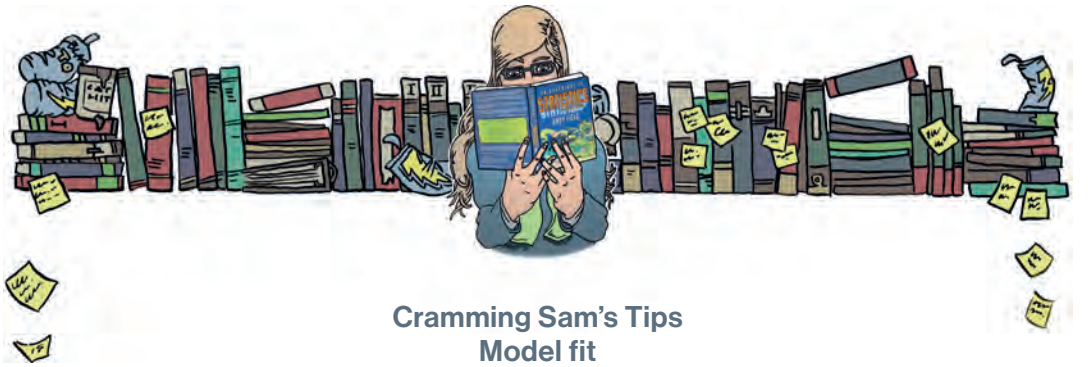


Output 20.6

20.6.6 Listing predicted probabilities Ⓑ

In Section 20.5.10 we saved residuals and predicted probabilities. The predicted probabilities and predicted group memberships are saved as variables in the data editor with the names **PRE_1** and **PGR_1**, respectively. These probabilities can be listed using the *Analyze* > *Reports* > Case Summaries... dialog box (see Section 9.11.6).

Output 20.7 shows a selection of the predicted probabilities (because the only predictor in the model was a dichotomous variable, there will be only two different probability values). I have also listed the predictor variables to clarify from where the predicted probabilities come. The only predictor



Cramming Sam's Tips
Model fit

- Build your model systematically and choose the most parsimonious model as the final one.
- The overall fit of the model is shown by the −2*LL* statistic and its associated chi-square statistic. If the significance of the chi-square statistic is less than or equal to 0.05, then the model is a significant fit to the data.
- Check the table labelled *Variables in the Equation* to see the regression parameters for any predictors you have in the model.
- For each variable in the model, look at the Wald statistic and its significance (which again should be less than or equal to 0.05). Use the odds ratio, *Exp*(*B*), for interpretation. If the value is greater than 1 then as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease. The confidence interval of *Exp*(*B*) should not cross 1.





Use the *case summaries* function to create a table for the first 15 cases in the file **eel.sav** showing the values of **cured**, **intervention**, **duration**, the predicted probability (**PRE_1**) and the predicted group membership (**PGR_1**) for each case.

in the final model was having the intervention, which could have a value of either 1 (had the intervention) or 0 (no intervention). If these two values are placed into Eq. 20.3 with the respective regression coefficients, then the two probability values in Output 20.7 are derived. In fact, we calculated these values in Jane Superbrain Box 20.1: the calculated probabilities (*P*(cured) in these equations) correspond to the values in **PRE_1**. These values tell us that when a patient is not treated (**intervention** = 0, no treatment group), there is a probability of 0.428 that they will be cured – basically, about 43% of people get better without any treatment. However, if the patient does have the intervention (**intervention** = 1, intervention group), there is a probability of 0.719 that they will get better – about 72% of people treated get better. When you consider that a probability of 0 indicates no chance of getting better, and a probability of 1 indicates that the patient will definitely get better, the values

obtained suggest that having the intervention increases your chances of getting better substantially (although the probability of recovery without the intervention is still not bad).

Assuming we are content that the model is accurate and that the intervention has some substantive significance, then we could conclude that our intervention (which, to remind you, was putting an eel up the anus) is the best predictor of being cured (of constipated). Furthermore, including the duration of the constipation pre-intervention and its interaction with the intervention did not improve how well we could predict whether a person got better.

Case Summaries[a]

	Case Number	Cured	Intervention	Number of days	Predicted probability	Predicted group
1	1	Not cured	No treatment	7	.42857	Not cured
2	2	Not cured	No treatment	7	.42857	Not cured
3	3	Not cured	No treatment	6	.42857	Not cured
4	4	Cured	No treatment	8	.42857	Not cured
5	5	Cured	Intervention	7	.71930	Cured
6	6	Cured	No treatment	6	.42857	Not cured
7	7	Not cured	Intervention	7	.71930	Cured
8	8	Cured	Intervention	7	.71930	Cured
9	9	Cured	No treatment	8	.42857	Not cured
10	10	Not cured	No treatment	7	.42857	Not cured
11	11	Cured	Intervention	7	.71930	Cured
12	12	Cured	No treatment	7	.42857	Not cured
13	13	Cured	No treatment	5	.42857	Not cured
14	14	Not cured	Intervention	9	.71930	Cured
15	15	Not cured	No treatment	6	.42857	Not cured
Total N	15		15	15	15	15

a. Limited to first 15 cases.

Output 20.7

20.6.7 Interpreting residuals Ⓑ

Fitting a model without checking how well it fits the data is like buying a new pair of trousers without trying them on: they might look fine on the hanger but get them home and you find they're too tight. The trousers do their job (they cover your legs and keep you warm) but they have little real-life value (because they cut off the blood circulation to your legs). Likewise, a model does its job regardless of the data, but the real-life value of the model may be limited (and may cut off the blood circulation to your career). So, our conclusions so far are fine in themselves, but to be sure that the model is a good one, it is important to examine the residuals.

We saw in Chapter 9 that the main purpose of examining residuals is to (1) isolate points for which the model fits poorly, and (2) isolate points that exert an undue influence on the model. To assess the former we examine the residuals, especially the Studentized residuals, standardized residuals and deviance statistics. To assess the latter we use influence statistics such as Cook's distance, DFBeta and leverage statistics. We'll look at deviance statistics in Section 20.7.3. These statistics were explained in Section 9.3 and their interpretation in logistic regression is the same; therefore, Table 20.4 summarizes the main statistics that you should look at and what to look for, but for more detail consult Chapter 9.

Remember that these residuals are saved in the data editor. If you inspect them you'll see that Cook's distance, leverage, standardized residuals and DFBeta values are pretty good: all cases have DFBetas less than 1, and leverage statistics (**LEV_1**) are very close to the calculated expected value of 0.018. There are also no unusually high values of Cook's distance (**COO_1**) which, all in all, means that there are no influential cases. The standardized residuals all have values of less than ±2 and so there seems to be very little here to concern us.



Oliver Twisted
Please, Sir, can I have some more ... diagnostics?

'What about the trees?' protests eco-Oliver. 'These SPSS outputs take up so much room, why don't you put them on the website instead?' It's a valid point, so I have produced a table of the diagnostic statistics for this example, but it's at www.discoverspss.com.

Table 20.4 Summary of residual statistics saved by SPSS

Label	*Name*	*Comment*
PRE_1	Predicted value	
PGR_1	Predicted group	
COO_1	Cook's distance	Should be less than 1
LEV_1	Leverage	Lies between 0 (no influence) and 1 (complete influence). The expected leverage is $(k+1)/N$, where k is the number of predictors and N is the sample size. In this example it would be 2/113 = 0.018
SRE_1	Studentized residual	About 5% should lie outside ±1.96, and about 1% should lie outside ±2.58. Cases above 3 are cause for concern, and cases close to 3 warrant inspection
ZRE_1	Standardized residual	
DEV_1	Deviance	
DFB0_1	DFBeta for the constant	Should be less than 1
DFB1_1	DFBeta for the first predictor **(intervention)**	

The residuals in this model are slightly unusual because they are based on a single predictor that is categorical. Consequently there isn't a lot of variability in the values of the residuals. Also, remember that if substantial outliers or influential cases had been isolated, you would not be justified in eliminating these cases to make the model fit better. Instead these cases should be inspected closely to try to isolate a good reason why they were unusual. It might simply be an error in inputting data, or it could be that the case was one which had a special reason for being unusual: for example, there were other medical complications that might contribute to the constipation and that were noted during the patient's assessment. In such a case, you may have good reason to exclude the case and duly note the reasons why.



Cramming Sam's Tips
Diagnostic statistics

- Look for cases that might be influencing the logistic regression model by checking residuals.
- Look at standardized residuals and check that no more than 5% of cases have absolute values above 2, and that no more than about 1% have absolute values above 2.5. Any case with a value above about 3 could be an outlier.
- Look in the data editor for the values of Cook's distance: any value above 1 indicates a case that might be influencing the model.
- Calculate the average leverage (the number of predictors plus 1, divided by the sample size) and then look for values greater than twice or three times this average value.
- Look for absolute values of DFBeta greater than 1.



20.6.8 Calculating the effect size Ⓑ

The best effect size to use in the context of logistic regression is the odds ratio, which we looked at in Section 20.6.3 (but see Jane Superbrain Box 20.2).

20.7 Interactions in logistic regression: a sporty example Ⓑ



I am English, and a stereotype of being English is believing that we can win sports events despite the weight of historical evidence to the contrary.[14] As each new tournament arrives, we are led by the media to believe that victory is assured. With every defeat, we feel crushed. My writing of each edition of this book has coincided with a soccer-related national failure. In 1998 when I wrote the first edition, the men's team were knocked out of the World Cup by losing a penalty shootout. In 2004 (second edition), they were knocked out of the European Championship in another penalty shootout. The men

14 The Rugby World Cup winning side of 2003 was the exception that proves the rule. Oh, and I think we won the Ashes in 2005, but we lost them the previous 18 years in a row, so that victory is hardly anything to feel smug about. Also, I don't like cricket (and not in the 10cc sense).



Jane Superbrain 20.2
Odds ratios and unobserved heterogeneity ©

Mood (2010) argues that odds ratios in logistic regression reflect not just the effect of a predictor but unobserved heterogeneity in the model. This observation is demonstrated by model parameters in logistic regression changing when an uncorrelated predictor is added. Imagine a linear model in which Y is predicted from X_1. Remember from our discussion of linear models that b-values represent the effect of a predictor adjusted for their relationship to other predictors in the model. If we add the predictor X_2, then the b-value for X_1 will be adjusted for its relationship with X_2: its value will change as a function of whether X_2 is in the model. If X_1 and X_2 are uncorrelated (i.e., independent), then the 'adjustment' will be zero, and so the b-value for X_1 will be unaffected by whether X_2 is also in the model. This doesn't happen in logistic regression: adding a predictor that is uncorrelated with an existing predictor still changes the b-value for the existing predictor. This is because estimates of residual variance are affected by what predictors are in the model. Consequently, Mood argues that: (1) it is problematic to use odds ratios (or log odds ratios) as effect size measures because they reflect unobserved heterogeneity as well as the size of effect; (2) odds ratios (and log odds ratios) should not be compared across models with different predictors because unobserved heterogeneity will vary across these models; and (3) where models use the same predictors, it is still problematic to compare odds ratios (and log odds ratios) across different samples, across groups within the same sample, and over time because unobserved heterogeneity will vary across samples, groups and time points.



didn't even manage to qualify for the 2008 European Championship (third edition); not a penalty shootout this time, just playing terribly. In 2012 (fourth edition), we lost yet another penalty shootout that sent us home from the European Championship. In 2016 (fifth edition), we lost to a semi-professional Icelandic team.[15] Then 2020 (as I was gearing up for this edition) we lost the final of the European Championship. Guess how we lost? That's right, a penalty shootout. What is *wrong* here?[16]

If I were the England men's soccer team coach, I'd take each one of the overpaid prima donnas and give them a kick in the testicles. A really hard kick. The kind of kick that lets them know that I'm the kind of guy that'll give you a good hard kick in your testicles if you miss a penalty. It wouldn't be as much fun, but I suppose, if push came to shove, I could also use science to find out which factors

15 Which, honestly, I found hilarious.

16 It should be noted that the women's national team (European champions 2021) and men's under-21 national team (European champions 2023) do not suffer the same affliction.

predict whether a player will score a penalty. Then I'd kick them in the testicles. One way or another, their testicles are getting kicked.

Imagine we wanted to research the factors that affect penalty-taking success. In soccer, games in knockout competitions are sometimes decided by a penalty shootout. The format is that each team takes five penalties using five different players. The winning team is the one that scores the most goals from the five penalties. If after this initial stage both teams have the same number of goals it goes to an excitingly named 'sudden death'. Fortunately no one suddenly dies, but instead each team takes a penalty and if both score or both miss another penalty is taken, but if one team scores and the other misses, the scoring team wins the game. As you might imagine, the pressure ramps up considerably with each successive penalty. If the first player misses his or her kick, then it's not necessarily a disaster because there are four more kicks to be taken, and (most important) opportunities for your opponent to miss or to make up for your mistake. However, if you are the player taking the fifth penalty you are under a lot more pressure because a miss is highly likely to result in your team losing. The pressure ramps up, the more penalties are taken. A sports scientist wanted to look at whether the position in the penalty shootout did predict success in a shootout. She collated data from penalty shootouts across various worldwide men's and women's soccer leagues as well as getting coaches' ratings of players.

As discussed at the start of this chapter, it seems reasonable that previous performance will predict penalty success, and let's imagine that past research had shown this to be true (Eriksson et al., 2004; Hoddle et al., 1998; Hodgson et al., 2012; Southgate et al., 2020). However, teams typically have one or two main penalty takers, so penalty shootouts invariably involve players not used to taking penalties in front of a stadium full of overly emotional and noisy fans, and for which there aren't any data on their previous success. Instead, the scientist got the coaching team to rate players' **ability** based on their penalty taking in training. Each player was rated against all other players the coaches had coached using a scale from 1 to 10, where 1 is 'their penalties are in the bottom 10% of players I've coached', 2 is 'their penalties are in the bottom 10–20% of players I've coached' and 10 is 'their penalties are in the top 10% of players I've coached'. In each shootout, the scientist noted the **position** order of the penalty (1 = first penalty, 5 = fifth penalty) and whether the goal was **scored**. The data are in **penalty_shootout.sav**. Those of you who hate soccer can read this example as being factors that predict success in a free throw in basketball or netball, a penalty in hockey, a penalty kick in rugby or a field goal in American football (although you might need to pretend that these sports involve the penalty shootout format!).

This research question is perfect for logistic regression because our outcome variable is a dichotomy: a penalty can be either scored or missed. We have two predictor variables (ability and position) but the scientist was also interested in their interaction (that is, is the effect of position moderated by ability?). The (fictional) data contain 868 different players taking kicks.

20.7.1 Fitting the model ©

Select *Analyze* > *Regression* > Binary Logistic... The sports scientist has a specific hypothesis about ability and position in the shootout interacting, so we want to force all predictors into the model. Drag the outcome variable (**scored**) to the *Dependent* box (or select it and click ▸). Select **ability** and **position** and drag them to the *Covariates* box (or click ▸), then select them again and click >a*b> to add the interaction term. The completed dialog box is in Figure 20.16. This is a big data set so don't select bootstrapping otherwise your computer will grind to a halt, but otherwise select the same options that we did in the previous example.

20.7.2 Interpreting the model ©

The main parts of the output show a few interesting things. First, Cox and Snell's measure (0.104) and Nagelkerke's adjusted value (0.234) suggest that the model is explaining around 10–20% of the

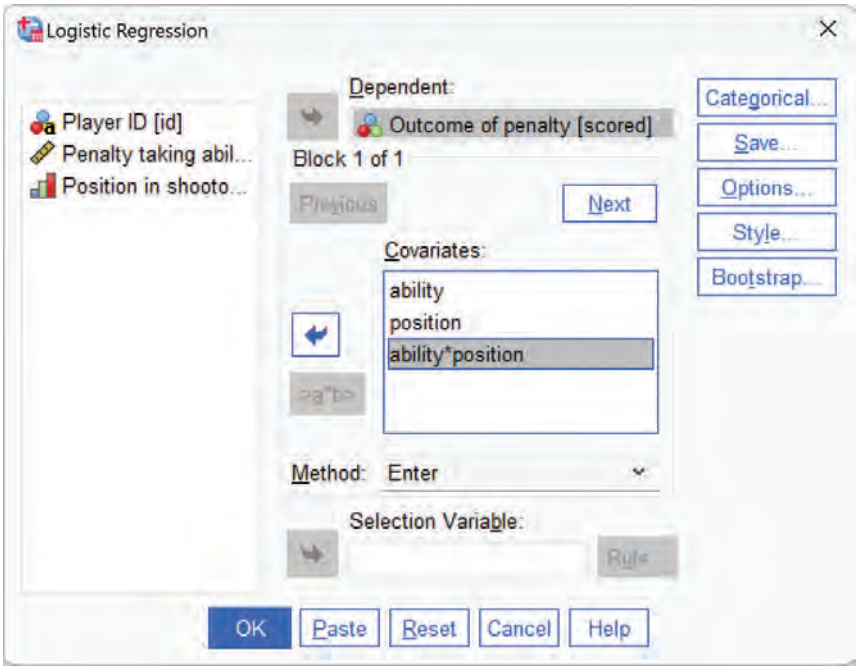


Figure 20.16 *Logistic Regression* dialog box for the penalty data

variance (Output 20.8). The classification table (Output 20.8) and plot (Output 20.9) show that the model is much better at classifying penalties that were scored. We can see this on the classification plot because there are a lot of S letters (for scored) to the far right of the plot and relatively few Ms (for miss). However, the left-hand side of the plot is basically empty, suggesting no misclassified scored kicks but also very few correctly classified misses. Similarly, in the table we can see that 99.4% of scored penalties are correctly classified compared to only 13.3% of missed penalties.



The classification table gives us a clue as to why scored penalties are better predicted than missed ones. The vast majority of kicks are scored because people tend not to end up as professional soccer players unless they're extremely good at kicking footballs. Of the 868 penalties in the data 793 are scored and only 75 missed! We can interpret the parameter estimates as follows:

- $\hat{b}_0 = 1.79$, which means that when all predictors are zero the log odds of scoring were 1.79. Looking at the odds ratio, we could say that a penalty was 5.97 times more likely to be scored than not if a player was rated as zero on ability and took kick number zero.[17]

17 Note that this conclusion isn't useful because these are impossible values in the data: no player was rated as zero on ability (because it was rated from 1 to 10) and no player took a penalty in position 0 (because the positions were labelled 1 to 5).

- $\hat{b}_{\text{ability}} = 0.84$, which means that when position is fixed, as the ability rating increases by 1, the log odds of scoring increase by 0.84. The corresponding odds ratio is 1.088, which tells us that when position is fixed, as the ability rating increases by 1, the odds of scoring are 1.088 times what they were before that change. Because this value is more than 1 it tells us that the odds of scoring increase very slightly as ability increases. However, the value is very close to 1, suggesting this increase is negligible. In fact, the effect is not significant, $z = 0.19$, $p = 0.661$.
- $\hat{b}_{\text{position}} = -1.18$, which means that when ability is fixed, as the position in the shootout increases by 1, the log odds of scoring decrease by 1.18. In other words, the further down the queue of penalty takers you are the less likely you are to score. The corresponding odds ratio is 0.309, which tells us that when ability is fixed, with each penalty kick that is taken, the odds of scoring are 0.309 times what they were for the previous player on the team. Because this value is less than 1 it tells us that the odds of scoring decrease, the further into the shootout we get. This is a substantial change in the odds, and the corresponding significance test is highly significant, $z = 10.06$, $p = 0.002$.
- $\hat{b}_{\text{interaction}} = 0.191$, which means that the increase in the log odds of scoring that occurs with every unit increase in ability, further increases by 0.19 each time position increases by one unit. An equally valid interpretation is that the change in the log odds of scoring that occurs with every unit increase in position gets 0.19 added to it each time ability increases by one unit. The corresponding odds ratio is 1.21, which tells us that the increase in the odds of scoring associated with increasing ability gets larger as position increases. Alternatively, the negative change in the odds of scoring associated with increasing position is becoming more positive as ability increases. The corresponding significance test is highly significant, $z = 10.00$, $p = 0.002$.

To help get a sense of what the parameter estimate for the interaction term means, see Jane Superbrain Box 20.3. Also, Figure 20.17 plots the interaction term. Unfortunately these plots were created in R, not SPSS, so you can't rely on these with your own data without using the R extension. In any case, both plots show the probability of scoring on the y-axis. The left plot shows the probability curves for position in the shootout separately for five different abilities (scores of 1, 3, 5, 7 and 9). Note that

Model Summary

Step	−2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	415.338[a]	.104	.234

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted Outcome of penalty		
	Observed		Missed penalty	Scored penalty	Percentage Correct
Step 1	Outcome of penalty	Missed penalty	10	65	13.3
		Scored penalty	5	788	99.4
	Overall Percentage				91.9

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Penalty taking ability	.084	.192	.192	1	.661	1.088	.746	1.585
	Position in shootout	−1.176	.371	10.057	1	.002	.309	.149	.638
	Penalty taking ability by Position in shootout	.191	.060	10.003	1	.002	1.210	1.075	1.363
	Constant	1.786	1.245	2.058	1	.151	5.967		

a. Variable(s) entered on step 1: Penalty taking ability, Position in shootout, Penalty taking ability * Position in shootout .

Output 20.8

when ability is high (7 and 9) the probability curves are flat: the position in the shootout doesn't affect the probability of scoring. However, at lower ability levels (1 and 3) the probability curves slope down, showing that as the position in the shootout increases, the probability of scoring decreases. In other words, the position in the shootout only matters for players rated low on penalty ability, and for those players the further from the first penalty, the less likely you are to score. The right-hand plot flips the variables and shows the probability curves for penalty-taking ability separately for the five different positions in the shootout (penalties 1 to 5). Note that when ability is high (6 and above) the probability curves are identical: all players have a very high probability of scoring regardless of which penalty they take in the shootout. However, when ability is low (below 6) the curves separate. In particular, at very low abilities, for example, the probability of scoring if you take the first penalty is very high at about 0.75, but if you take the last one (position = 5) it is only 0.1. Again, this shows that the interaction tells us that the penalty you take (first through to fifth) only matters for players rated low on their ability to take penalties. For these players they are progressively less likely to score the further their penalty is from the first one.

```
Step number: 1

Observed Groups and Predicted Probabilities

    160 +                                                                                                                +
        I                                                                                                                I
        I                                                                                                                I
F       I                                                                                                                I
R   120 +                                                                                                                +
E       I                                                                                                          S     I
Q       I                                                                                                        SS S  I
U       I                                                                                                        SS SS I
E    80 +                                                                                                        SS SSS+
N       I                                                                                                    S   SS SSSI
C       I                                                                                                   SS   SS SSSI
Y       I                                                                                                   SS   SS SSSI
    40 +                                                                                                    SS SSSSSSS+
       I                                                                                                    SS SSSSSSSSI
       I                                                                                          S         SS SSSSSSSSI
       I                                                                                    S  S      S SS SSSSSSSSI
       I                                                                 M        S       M      S   M  M   S SS S MM MSMMSSSSI
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9        1
 Group:   MMMM
```



Jane Superbrain 20.3
Interpreting interactions ©

To give you a flavour of what the interaction term means, I ran a logistic regression model for the effect of ability separately for those who took the first penalty (position = 1) and those that took the second (position = 2). Output 20.10 shows the parameter estimates from SPSS. Note that:

- For those who took the first penalty, $\hat{b}_{\text{ability}} = 0.22$, which means that as the ability rating increases by 1, the log odds of scoring increase by 0.22.
- For those who took the second penalty, $\hat{b}_{\text{ability}} = 0.42$, which means that as the ability rating increases by 1, the log odds of scoring increase by 0.42.

Note further that:

$$\hat{b}_{\text{ability, position 2}} - \hat{b}_{\text{ability, position 1}} = 0.42 - 0.22 = 0.2$$

which is approximately the value of the $\hat{b}_{\text{interaction}} = 0.19$ in our model (it's not identical because I've used a subset of the data). So, as position increases by one unit, the effect of ability gets stronger (it goes up from 0.22 to 0.42); in other words, the later your kick in the shootout (i.e. as position increases), the more important ability becomes to scoring.

Variables in the Equation[a]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[b]	Penalty taking ability	.220	.210	1.098	1	.295	1.246
	Constant	1.298	1.449	.802	1	.370	3.661

a. Position in shootout = 1

b. Variable(s) entered on step 1: Penalty taking ability.

Variables in the Equation[a]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[b]	Penalty taking ability	.416	.156	7.119	1	.008	1.516
	Constant	-.697	1.009	.477	1	.490	.498

a. Position in shootout = 2

b. Variable(s) entered on step 1: Penalty taking ability.

Output 20.10



20.7.3 Testing for linearity of the logit ©

In this example we have two continuous variables, therefore we have to check that each one is linearly related to the log of the outcome variable (**scored**). I mentioned earlier in this chapter that to test this assumption we run the logistic regression but include predictors that are the interaction between each predictor and the log of itself (Hosmer et al., 2013). To create these interaction terms, we use *Transform* > Compute Variable... (Section 6.10.4). For each variable create a new variable that is the log of the original variable. For example, for **ability**, create a new variable called **ln_ability** by entering this name into the box labelled *Target Variable*, click Type & Label... and give the variable a label such as *ln(ability)*. In the list labelled *Function group* click *Arithmetic* and then in the box labelled *Functions and Special Variables* click *Ln* (this is the natural log transformation) and transfer it to the command area by clicking ↑. The command will appear in the command area as 'LN(?)' and the question mark should be replaced with a variable name (which can be typed manually or transferred from the variables list). Replace the question mark with the variable **ability** by either selecting the variable in the list and clicking → or typing 'ability' where the question mark is. Click OK to create the variable.



To test the assumption we redo the analysis, but instead of the interaction term put in the interaction terms of each predictor and their logs. Select *Analyze* > *Regression* > Binary Logistic..., then in the main dialog box drag **scored** to the *Dependent* box (or click →). Specify the main effects by clicking on **ability**, and **position** while holding down the *Ctrl* key (or *Cmd* on a Mac) and dragging them to the *Covariates* box (or click →). To input the interactions, click the two variables in the interaction while holding down the *Ctrl* key (or *Cmd* on a Mac): for example, click **ability** then, while holding down *Ctrl*, click **ln_ability** and click >a*b> to move them to the *Covariates* box. This action specifies the **ability** × **ln_ability** interaction. Specify the **position** × **ln_position** interaction in the same way. The completed dialog box is in Figure 20.18.

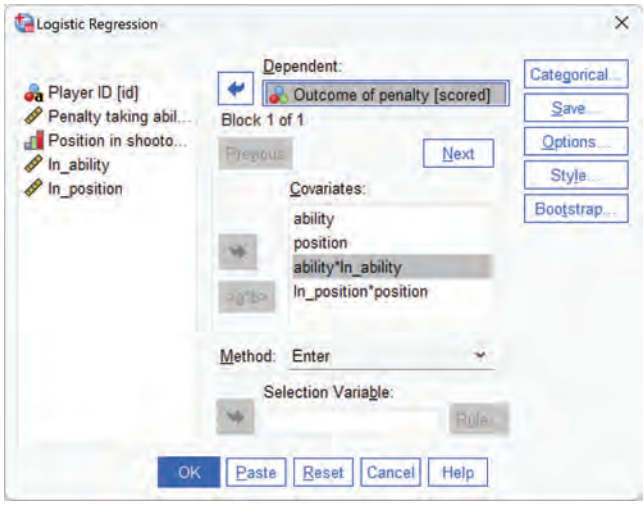


Figure 20.18 Dialog box for testing the assumption of linearity in logistic regression

Output 20.11 shows the part of the output that tests the assumption. We're interested only in whether the interaction terms are significant. Any interaction that is significant indicates that the main effect has violated the assumption of linearity of the logit. All interactions have significance values greater than 0.05, indicating that the assumption of linearity of the logit has been met for **ability** and **position**.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Penalty taking ability	.065	1.183	.003	1	.956	1.067	.105	10.844
	Position in shootout	-.461	.895	.266	1	.606	.630	.109	3.641
	Penalty taking ability by ln_ability	.221	.437	.256	1	.613	1.248	.530	2.939
	ln_position by Position in shootout	.198	.435	.208	1	.649	1.219	.520	2.859
	Constant	-.018	2.650	.000	1	.995	.982		

a. Variable(s) entered on step 1: Penalty taking ability, Position in shootout, Penalty taking ability * ln_ability , ln_position * Position in shootout .

Output 20.11

20.7.4 Testing for multicollinearity ©

In Section 9.9.3 we saw how multicollinearity can affect the parameters of a linear model. Logistic regression is just as prone to the biasing effect of collinearity, so we need to test for it. Unfortunately, SPSS does not produce collinearity diagnostics in logistic regression (which creates the illusion that multicollinearity doesn't matter). However, you can obtain statistics such as the tolerance and VIF by running a linear regression using the same outcome and predictors.

For the penalty example, select *Analyze* > *Regression* > Linear Regression. The completed dialog box is shown in Figure 20.19. It is unnecessary to specify any options other than clicking Statistics... and selecting ☑ Collinearity diagnostics (I'd also switch off all the default options). Click Continue to return to the main dialog box, and click OK to run the analysis.

The first table in Output 20.12 shows tolerance values of 1 for both **position** and **ability**. To recap, tolerance values less than 0.1 and VIF values greater than 10 (Menard, 2002) indicate a problem

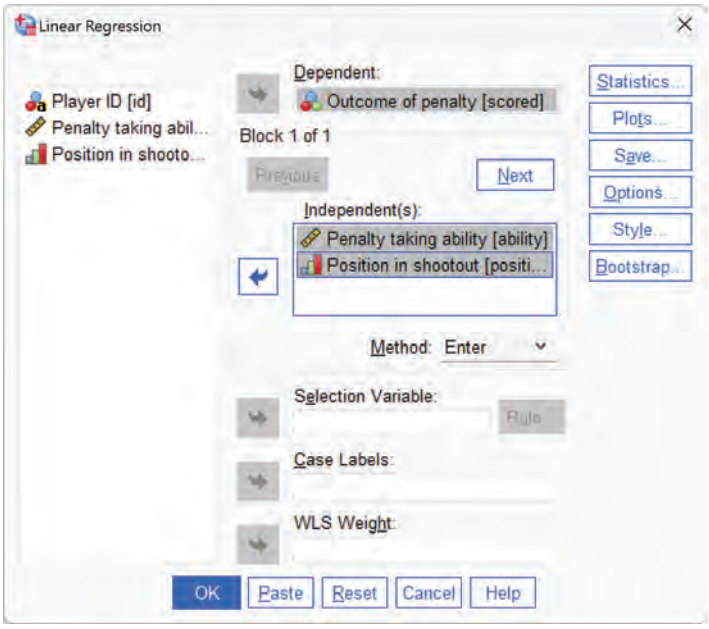


Figure 20.19 *Linear Regression* dialog box for the penalty data

(Chapter 9). In these data the VIF values are also 1 (because 1/1 = 1). Basically, no problems here. We can investigate collinearity further by examining the table labelled *Collinearity Diagnostics* (Section 9.9.3). For these data the final dimension has a condition index of 11.07, which is a little larger than the other dimensions. Although there are no hard-and-fast rules about how much larger a condition index needs to be to indicate collinearity problems, our condition indexes are not dramatically different. The variance proportions tell us the proportion of the variance of each predictor's *b*-value that is attributed to each eigenvalue. These proportions can be converted to percentages by multiplying them by 100. In terms of collinearity, we are looking for predictors that have high proportions on the same *small* eigenvalue, because this would indicate that the variances of their *b*-value are dependent. So we are interested mainly in the bottom few rows of the table (which represent small eigenvalues). In this example, for **ability** 90% of the variance of the *b*-value is associated with eigenvalue number 3, 9% is associated with eigenvalue number 2 and 1% is associated with eigenvalue number 1; whereas for **position** 9% of the variance of the *b*-value is associated with eigenvalue number 3, 89% is associated with eigenvalue number 2 and 2% is associated with eigenvalue number 1. Basically, **position** and **ability** have the bulk of their variance associated with *different* eigenvalues, suggesting that they are not collinear.

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Penalty taking ability	1.000	1.000
	Position in shootout	1.000	1.000

a. Dependent Variable: Outcome of penalty

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Penalty taking ability	Position in shootout
1	1	2.848	1.000	.00	.01	.02
	2	.129	4.698	.03	.09	.89
	3	.023	11.073	.97	.90	.09

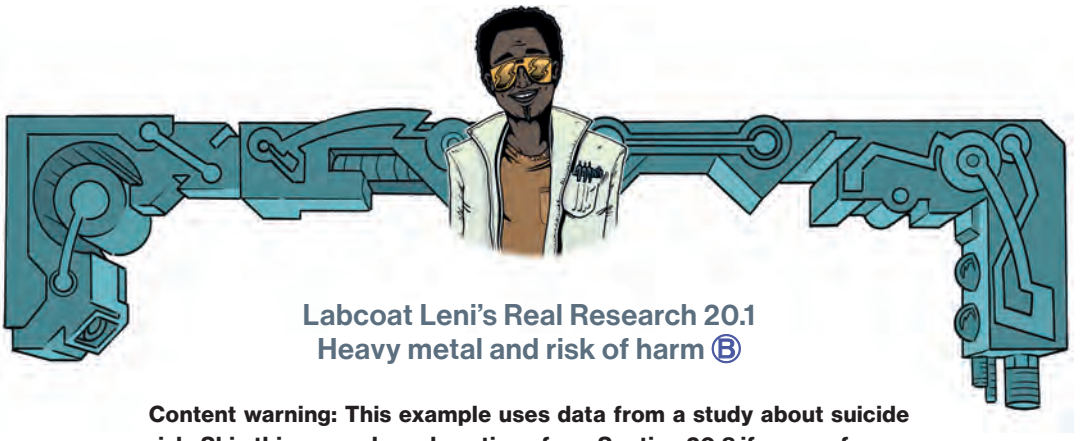
a. Dependent Variable: Outcome of penalty

Output 20.12 Collinearity diagnostics for the penalty data

If you identify collinearity then, unfortunately, there's not much that you can do about it. One obvious solution is to omit one of the variables. The problem with this strategy might be obvious: there is no way of knowing which variable to omit. The resulting theoretical conclusions are meaningless because, statistically speaking, any of the collinear variables could be omitted. There are no statistical grounds for omitting one variable over another. Even if a predictor is removed, Bowerman and O'Connell (2000) recommend that another equally important predictor that does not have such strong multicollinearity replace it. They also suggest collecting more data to see whether the multicollinearity can be lessened. Another possibility when there are several predictors involved in the multicollinearity is to run a principal components analysis on these predictors and to use the resulting component scores as a predictor (see Chapter 18). The safest (although unsatisfactory) remedy is to acknowledge the unreliability of the model.

20.8 Reporting logistic regression Ⓑ

I'd report logistic regression much the same as any linear model (see Section 9.13). I'd be inclined to tabulate the results, unless it's a very simple model. As a bare minimum, report the *b*-values (and their standard errors and significance value), the odds ratio (and its confidence interval) and some general statistics about the model (such as the R^2 and goodness-of-fit statistics). I'd also include the constant



Labcoat Leni's Real Research 20.1
Heavy metal and risk of harm Ⓑ

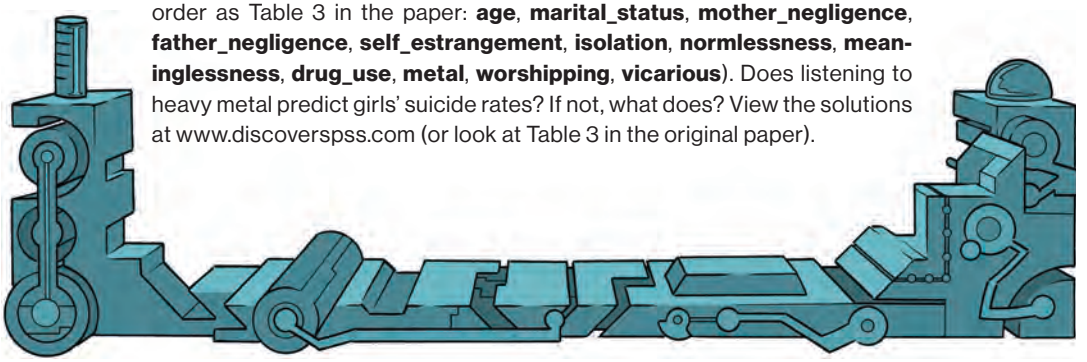
Content warning: This example uses data from a study about suicide risk. Skip this example and continue from Section 20.8 if you prefer.

Lacourse, E., et al. (2001). *Journal of Youth and Adolescence*, *30*, 321–332.

My favourite kind of music is heavy metal. When not listening (and often while listening) to heavy metal, I sometimes research clinical psychology in youths. I stumbled on a paper that combined these two interests by testing whether a love of heavy metal could predict suicide risk (Lacourse et al., 2001).

Eric Lacourse and his colleagues used questionnaires to measure suicide risk (yes or no), marital status of parents (together or divorced/separated), the extent to which the person experienced neglect, self-estrangement/powerlessness (adolescents who have negative self-perceptions, are bored with life, etc.), social isolation (feelings of a lack of support), normlessness (beliefs that socially disapproved behaviours can be used to achieve certain goals), meaninglessness (doubting that school is relevant to gaining employment) and drug use. In addition, the authors measured liking of heavy metal; they included the sub-genres of classic (Black Sabbath, Iron Maiden), thrash metal (Slayer, Metallica), death/black metal (Obituary, Burzum) and gothic (Marilyn Manson). As well as liking they measured behavioural manifestations of worshipping these bands (e.g., hanging posters, hanging out with other metal fans) and what the authors termed 'vicarious music listening' (whether music was used when angry or to bring out aggressive moods). They used logistic regression to predict suicide risk from these variables for those identifying as male and female separately.

The data for the female sample are in the file **lacourse_2001_females.sav**. Labcoat Leni wants you to carry out a logistic regression predicting **suicide_risk** from all the predictors (forced entry). (To make your results easier to compare to the published results, enter the predictors in the same order as Table 3 in the paper: **age**, **marital_status**, **mother_negligence**, **father_negligence**, **self_estrangement**, **isolation**, **normlessness**, **meaninglessness**, **drug_use**, **metal**, **worshipping**, **vicarious**). Does listening to heavy metal predict girls' suicide rates? If not, what does? View the solutions at www.discoverspss.com (or look at Table 3 in the original paper).



so that readers of your work can construct the full regression model if they need to. You might also consider reporting the variables that were not significant predictors, because this can be as valuable as knowing about which predictors were significant.

For the first example in this chapter we might produce something like Table 20.5. Hopefully you can work out from where the values came by looking back through the chapter so far. I've rounded off to 2 decimal places throughout. If you use APA style then for the R^2 and p-values there should be no zero before the decimal point (because these values cannot exceed 1). I have reported the bootstrap confidence intervals for $\hat{b}$.

Table 20.5 Coefficients of the model predicting whether a patient was cured (95% BCa bootstrap confidence intervals based on 1000 samples in brackets)

		95% CI for Odds Ratio		
	$\hat{b}$	Lower	Odds Ratio	Upper
Included				
Constant	−0.29 [−0.77, 0.20]			
Intervention	1.23* [0.42, 2.06]	1.56	3.42	7.48

Note. R^2 = 0.06 (Hosmer–Lemeshow), 0.08 (Cox–Snell), 0.11 (Nagelkerke). Model $\chi^2(1) = 9.93$, $p = 0.002$. *$p < 0.01$.

For the second example we might produce a table like Table 20.6.

Table 20.6 Coefficients of the model predicting whether a penalty was scored

			95% CI for Odds Ratio		
	$\hat{b}$	p	Lower	Odds Ratio	Upper
Included					
Constant	1.79	0.151		5.97	
Ability	0.08	0.661	0.75	1.09	1.56
Position in shootout	−1.18	0.002	0.15	0.31	0.64
Ability × Position	0.19	0.002	1.08	1.21	1.36

Note. R^2 = 0.10 (Cox–Snell), 0.23 (Nagelkerke).

20.9 Jane and Brian's story Ⓐ

'Do you and your dad talk about your mum?' Jane asked. She usually avoided emotional conversations, but since he'd opened up to her Jane had felt guilty that she hadn't asked Brian anything more about it. She wondered how someone could grow up so balanced with such a shadow lurking in their past.

Brian looked surprised by the question but he was open to it. 'When I was growing up … but not so much now,' he replied. 'I think it upsets him to talk about her.'

'But he used to …,' Jane said, challenging Brian's assumption.

'He had this thing about me not remembering her. It really hurt him, so he'd talk about her. He'd pull out photos every few weeks from when she was alive and talk about what we'd done, when the photo was taken … he talk about their routines, about things we did together, anything to keep her alive in the house and to jog my memory. But nothing. He hated the fact I couldn't remember her, it made him so desolate. I wanted to bring him some peace, so eventually I lied.'

'Lied?'

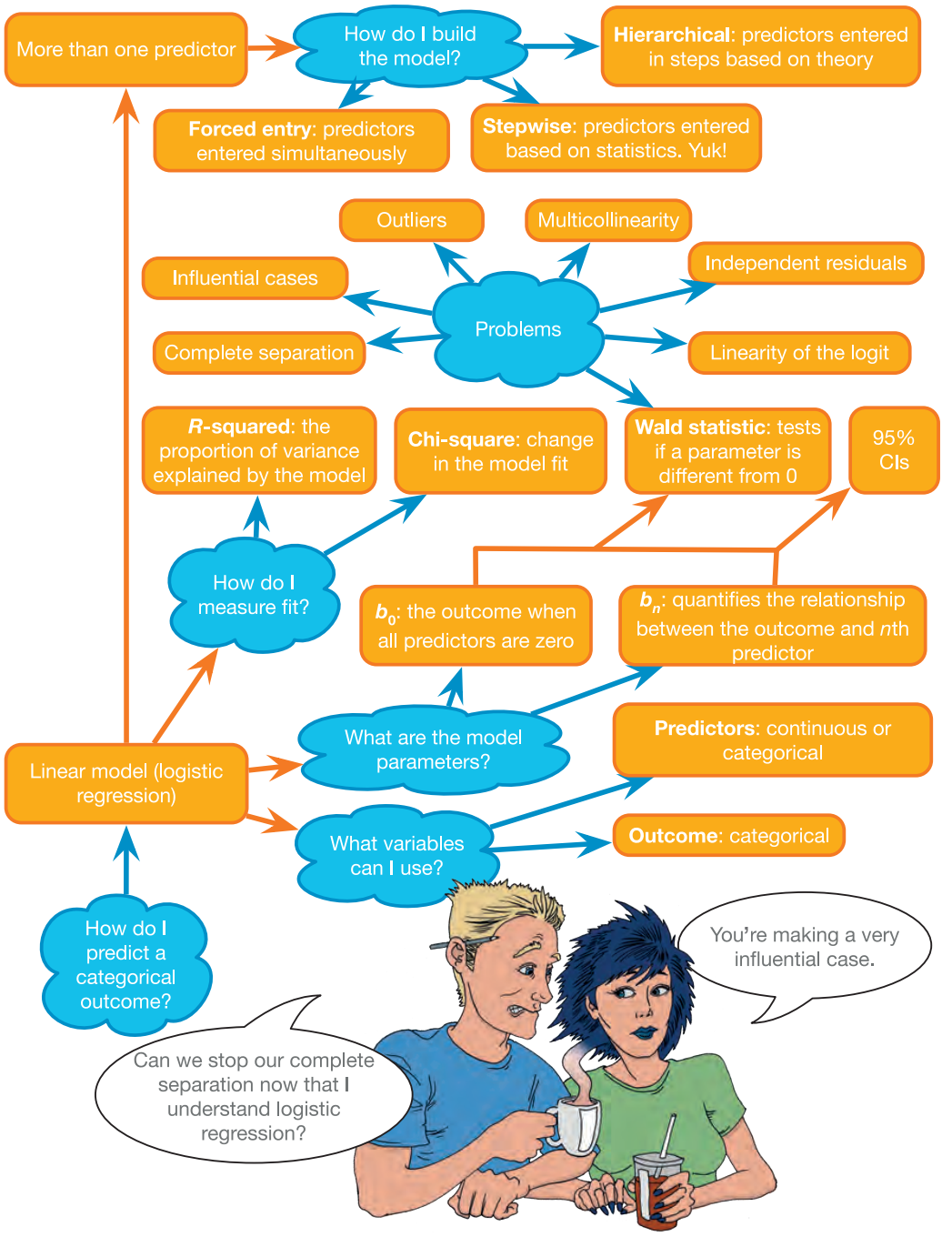


Figure 20.20 What Brian learnt from this chapter

'Yeah, you know … just little lies, like, I picked up on their routine: mum took me to school, he picked me up, so I'd tell him "I remember this one time on the way to school and mum and I saw this cat …" – just invent some cute little incident that he couldn't verify. His eyes sparkled every time. That was all I needed to keep going. I kept lying and eventually we stopped talking about her.

'Do you miss her?'

'How can you miss what you don't remember? That's what you'd think, wouldn't you? But, yeah, I think about her every day. It's like, I know my dad so well, I look at myself and I see him in what I look like and what I do, but then there's this bunch of stuff about me, that I guess comes from her, but I just don't know. I know who I am and yet half of me is missing … But I'm grateful for what I have. My dad is awesome. Some people have horrible, destructive parents, that mess them up. All I've ever had is love.'

Jane wondered what that would feel like.

20.10 What next? Ⓐ

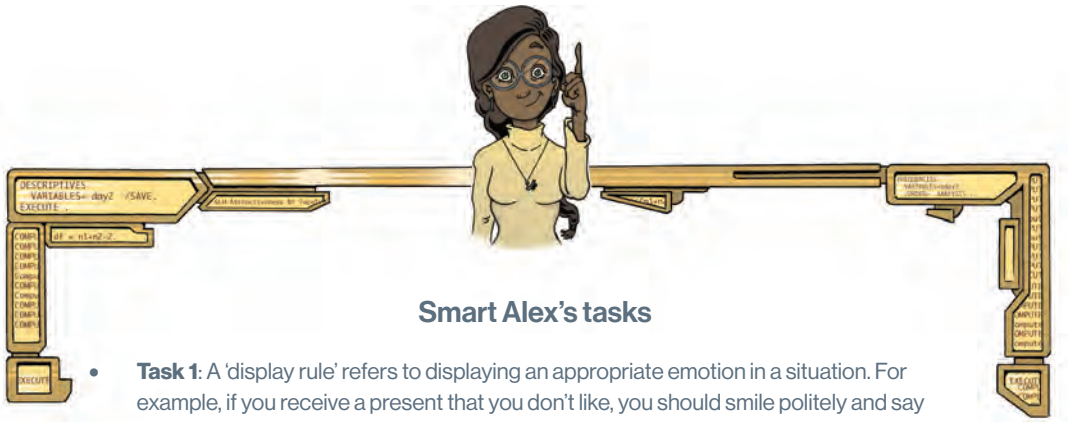
At the age of 10 I thought I was going to be a rock star. Such was my conviction about this that even today (many years on) I'm still not entirely sure how I ended up *not* being a rock star (possible explanations are lack of talent, being so uncool that I still use the word 'cool', inability to write good songs – the list is depressingly long). Instead of the glitzy and fun life that I anticipated I am instead reduced to writing textbook chapters about things that I don't even remotely understand.

The other thing that I thought at the age of 10 was that I would marry Clair Sparks. Such was my conviction that even today I'm still not entirely sure …. Nah, I'm just kidding. However, as a boy I was (for some inexplicable reason) convinced that I would get married at the age of 28. It was something of a shock to me when I reached 28 and discovered that I didn't have a wife and probably wasn't in any fit state to have one. I'd spent far too much time dedicated to music, then to working ridiculously hard to try to convince everyone I was clever enough to be a scientist, and the two years dating someone who destroyed my very soul didn't help either. Time ticked on until in my mid-thirties I found myself alone, without a wife. 'I'd better get one of those,' I thought to myself.[18]

20.11 Key terms that I've discovered

$-2LL$
Binary logistic regression
Complete separation
Cox and Snell's R^2_{CS}
Deviance
$\exp(\hat{b})$, or Exp(B)
Hosmer and Lemeshow's R^2_L
Logistic regression
Log-likelihood
Multinomial logistic regression
Nagelkerke's R^2_N
Parsimony
Polychotomous logistic regression
Suppressor effects
Wald statistic

18 Needless to say, I'm skipping over a fair few romantic events between the ages of 10 and my mid-thirties, some of them more pleasant than others.



Smart Alex's tasks

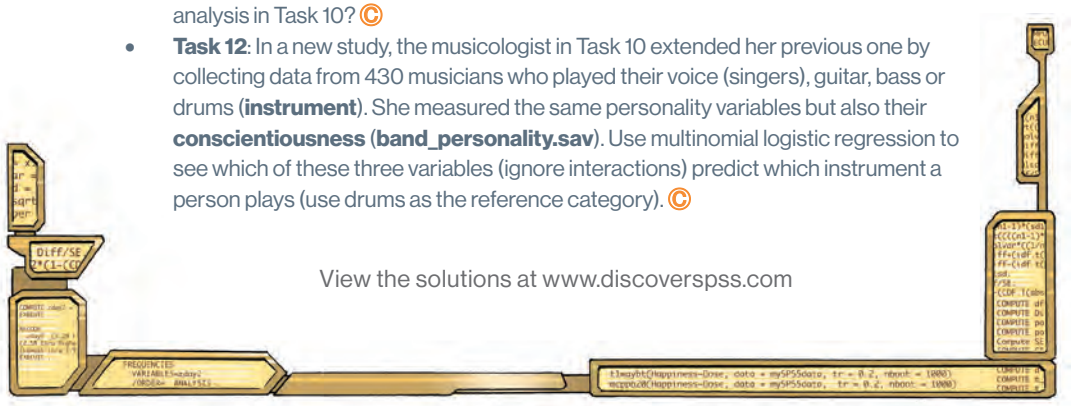
- **Task 1**: A 'display rule' refers to displaying an appropriate emotion in a situation. For example, if you receive a present that you don't like, you should smile politely and say 'Thank you, Auntie Kate, I've always wanted a rotting cabbage'; you do not start crying and scream 'Why did you buy me a rotting cabbage?!' A psychologist measured children's understanding of display rules (with a task that they could pass or fail), their age (months) and their ability to understand others' mental states ('theory of mind', measured with a false belief task that they could pass or fail). Can **display** rule understanding (did the child pass the test: yes/no?) be predicted from theory of mind (did the child pass the false belief, **fb**, task: yes/no?), **age** and their interaction? (**display.sav**.) Ⓒ
- **Task 2**: Are there any influential cases or outliers in the model for Task 1? Ⓒ
- **Task 3**: The behaviour of drivers has been used to claim that people of a higher social class are more unpleasant (Piff et al., 2012). Piff and colleagues classified social class by the type of car (**vehicle**) on a five-point scale and observed whether the drivers cut in front of other cars at a busy intersection (**vehicle_cut**). Do a logistic regression to see whether social class predicts whether a driver cut in front of other vehicles (**piff_2012_vehicle.sav**).[19] Ⓑ
- **Task 4**: In a second study, Piff et al. (2012) observed the behaviour of drivers and classified social class by the type of car (**vehicle**), but the outcome was whether the drivers cut off a pedestrian at a crossing (**pedestrian_cut**). Do a logistic regression to see whether social class predicts whether or not a driver prevents a pedestrian from crossing (**piff_2012_pedestrian.sav**). Ⓑ
- **Task 5**: Four hundred and sixty-seven lecturers completed questionnaire measures of **burnout** (burnt out or not), perceived **control** (high score = low perceived control), **coping** ability (high score = high ability to cope with stress), stress from **teaching** (high score = teaching creates a lot of stress for the person), stress from **research** (high score = research creates a lot of stress for the person) and stress from providing **pastoral** care (high score = providing pastoral care creates a lot of stress for the person). Cooper et al.'s (1988) model of stress indicates that perceived control and coping style are important predictors of burnout. The remaining predictors were measured to see the unique contribution of different aspects of a lecturer's work to their burnout. Conduct a logistic regression to see which factors predict burnout (**burnout.sav**). Ⓒ
- **Task 6**: An HIV researcher explored the factors that influenced condom use with a new partner (relationship less than 1 month old). The outcome measure was whether a condom was used (**use**: condom used = 1, not used = 0). The predictor variables were mainly scales from the Condom Attitude Scale (Sacco et al., 1991): **gender**; the degree to which the person views their relationship as 'safe' from sexually transmitted disease (**safety**); the degree to which previous experience influences attitudes towards condom use (**experience**); whether or not the couple used a condom in their previous encounter

19 I reconstructed the raw data from Figure 1 of the paper, so you will get basically the same values as reported there, but not the exact ones because they also controlled for the age and gender of drivers (and we don't have those variables).

(**previous**: 1 = condom used, 0 = not used, 2 = no previous encounter with this partner); the degree of self-control that a person has when it comes to condom use (**self_control**); the degree to which the person perceives a risk from unprotected sex (**risk_perception**). Previous research has shown that gender, relationship safety and perceived risk predict condom use (Sacco et al., 1991). Verify these previous findings and test whether self-control, previous usage and sexual experience predict condom use (**condom.sav**). Ⓒ

- **Task 7**: How reliable is the model in Task 6? Ⓑ
- **Task 8**: Using the final model from Task 6, what are the probabilities that participants 12, 53 and 75 will use a condom? Ⓒ
- **Task 9**: A female who used a condom in her previous encounter scores 2 on all variables except perceived risk (for which she scores 6). Use the model in Task 6 to estimate the probability that she will use a condom in her next encounter. Ⓒ
- **Task 10**: At the start of the chapter we looked at whether the type of instrument a person plays is connected to their personality. A musicologist measured **extroversion** and **agreeableness** in 200 singers and guitarists (**instrument**). Use logistic regression to see which personality variables (ignore their interaction) predict which instrument a person plays (**sing_or_guitar.sav**). Ⓑ
- **Task 11**: Which problem associated with logistic regression might we have in the analysis in Task 10? Ⓒ
- **Task 12**: In a new study, the musicologist in Task 10 extended her previous one by collecting data from 430 musicians who played their voice (singers), guitar, bass or drums (**instrument**). She measured the same personality variables but also their **conscientiousness** (**band_personality.sav**). Use multinomial logistic regression to see which of these three variables (ignore interactions) predict which instrument a person plays (use drums as the reference category). Ⓒ

View the solutions at www.discoverspss.com





MULTILEVEL LINEAR MODELS

21.1 What will this chapter tell me? 980
21.2 Hierarchical data 981
21.3 Multilevel linear models 984
21.4 Practical issues 998
21.5 Multilevel modelling using SPSS 1008
21.6 How to report a multilevel model 1025
21.7 A message from the octopus of inescapable despair 1026
21.8 Jane and Brian's story 1026
21.9 What next? 1026
21.10 Key terms that I've discovered 1027
Smart Alex's tasks 1028

Letters A to D indicate increasing difficulty

21.1 What will this chapter tell me? Ⓐ

Years at an all-boys' school carefully nurtured a fear of interacting with women and, along with a love of heavy metal, had made the world of relationships a tricky place for me to inhabit. I'd always dreamt that by my mid-thirties I would have a wife and a cute little child or two to stop me spending my life staring at a screen writing about statistics. But as I took my first furtive and depressing steps into middle age I found myself single and the closest thing to a child was a ginger cat and this book. However, something remarkable had happened since my teens: rock music had become popular. I needed to capitalize before the ephemeral guillotine of fashion spliced this opportunity from me. I met Zoë, who was not only happy to discuss Iron Maiden, but owned my favourite of their albums (*Piece of Mind*). She had no aversion to statistics, or soccer, and happened to be the loveliest woman ever placed on the face of the earth. Result. 'I'd better marry her before she realizes I'm a balding geek with slight hoarding tendencies,' I thought. So, that's what we did. A little later than anticipated, my dreams had come true: I started my late thirties with a wife and a cute little … book about a statistics package called R to which my wife contributed. I made a mental note to next time create a little human, not a big book.[1]



Figure 21.1 On the road to happiness

Marriage is a leap of faith into the unknown, a shared adventure full of challenges. A bit like this chapter really, because, when I started writing it, multilevel linear models were an 'unknown': I knew nothing about them. If you're reading this section then you probably don't know much about them either. So, we'll learn together – a shared adventure, and it sure will include some challenges …

1 A mental note that we heeded (see Chapter 3).

21.2 Hierarchical data Ⓑ



So far we have treated data as though they are organized at a single level. However, data are often structured in contexts that form a hierarchy. For example, part of my academic life has been spent researching anxiety in children. If we conduct research in a school, the children who participate have been assigned to different classes which are taught by different teachers. A child's class is a context that could affect my results. Imagine I collect data from two different classes. Mr. Nervous, who teaches the first class, is very anxious. He tells children to be careful, that things that they do are dangerous, and that they might hurt themselves. Little Miss Daredevil,[2] who teaches the second, is carefree, encourages children to feel their fear, and gives them freedom and support to explore new situations.

Imagine my experiment involves giving children information about an animal in a big wooden box. For some children I focus on the positive aspects of the animal (e.g., soft fur) whereas for others I focus on the negative (e.g., big teeth). I then invite the children to put their hand into the box and stroke the animal (which they can't see). I measure whether they will. The boxes actually contain cuddly toys. Children taught by Mr. Nervous have been taught in an environment that reinforces caution, whereas children taught by Miss Daredevil have been encouraged to embrace new experiences. Therefore, irrespective of the information I have given the children, we might expect Mr. Nervous's children to be more reluctant to put their hand in the box than Miss Daredevil's children because of the classroom experiences that they have had. Similarly, the information itself might have a different effect for the two classes (e.g., Mr. Nervous' children might be more sensitive to the negative information). Figure 21.2 illustrates this scenario: the child (or case) is at the bottom of the hierarchy (known as a *level 1* unit) and is nested (or grouped) within the class that they are in (the class is a level up from the child in the hierarchy and is said to be a *level 2* unit).

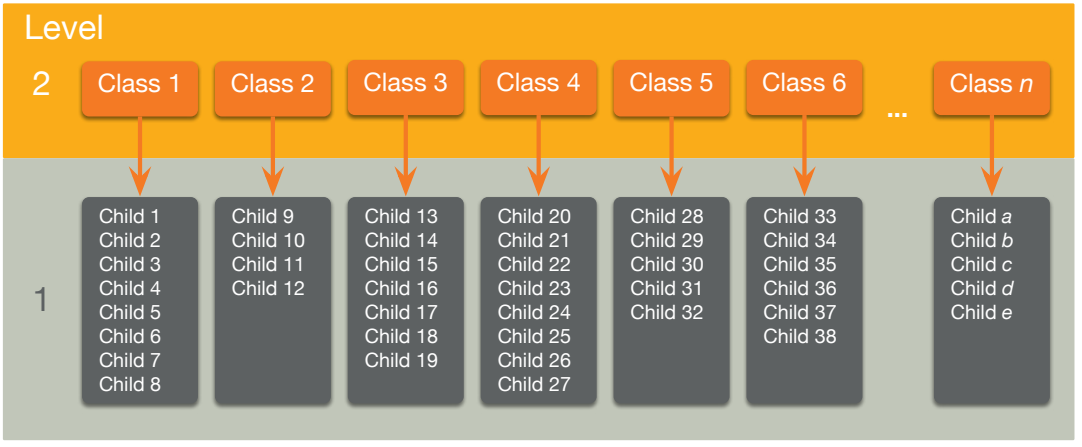


Figure 21.2 An example of a two-level hierarchical data structure. Children (level 1) are organized within classrooms (level 2)

2 I loved the Mr. Men/Little Miss (http://www.mrmen.com) stories as a kid, but as an adult reading them to my own kids I have been disappointed to discover that theirs is a world full of punishment-obsessed, name-calling sociopaths.

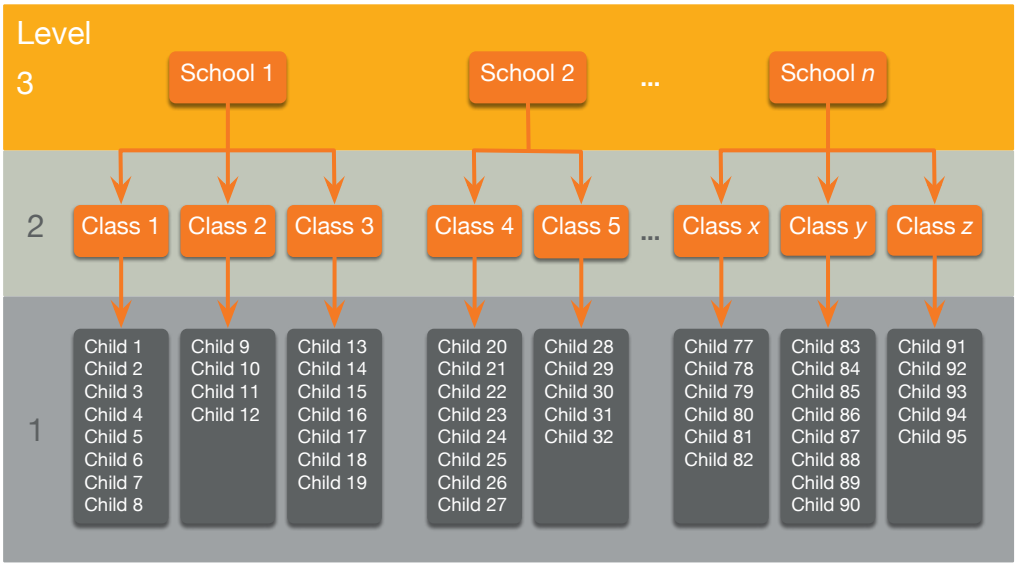


Figure 21.3 An example of a three-level hierarchical data structure

You can have more complex hierarchies than a two-level one. Sticking with our example, an obvious third level is that classrooms are nested within schools. If I collected data not just in different classrooms but at different schools, then I'd have another level in the hierarchy. The logic is the same as before: children in the same school will be more like each other than like children in different schools because schools have their own teaching environments and reflect their social demographic (which differs from school to school). Figure 21.3 shows this three-level hierarchy: the child (level 1) is nested within the class (level 2) to which he or she belongs which is nested within the school (level 3) to which the class belongs. There are two contexts: school and classroom.

Repeated-measures designs, in which participants contribute multiple data points over time or across different conditions are also an example of a hierarchical data structure. In these situations, observations (level 1) are nested within entities (level 2). Imagine that I was interested in creating a culture of kindness within primary/elementary schools. Throughout the school year participating schools embed a programme of structured activities that teach, discuss and practice kindness. I want to see whether it works. To this end I design a measure of kindness that taps each child's perceptions of their own and others' kind behaviours as well as their beliefs around kindness. Each school administers this measure to the children twice during each term so we have a measure of kindness taken at the start of the school year (in September) and then in October, January, February, April and June. For a given child there are several measures of kindness and these scores are not independent: if a child scores highly on kindness at one point in time, they are likely to score highly at another people in time. Therefore, kindness scores are not independent units. As such, the person acts as a context within which kindness is measured. Figure 21.4 illustrates this scenario: the child is the level 2 variable, and within each child there are kindness scores taken at the various points during the school year (the level 1 variable). Notice that some children are missing scores for certain months (perhaps they were absent on the day that kindness was measured), and child *y* moved school after the first term.

Logically, you'd expect kindness scores within a classroom or school to be more similar than scores in different classrooms or schools. We can, therefore, also include levels of the hierarchy above the child; the class to which the child belongs could be a level 3 variable, and their school a level 4 variable. This is an example of a growth model, which is where observations (at different times) are a level 1 variable nested within some other level 2 variable (such as people or organizations).

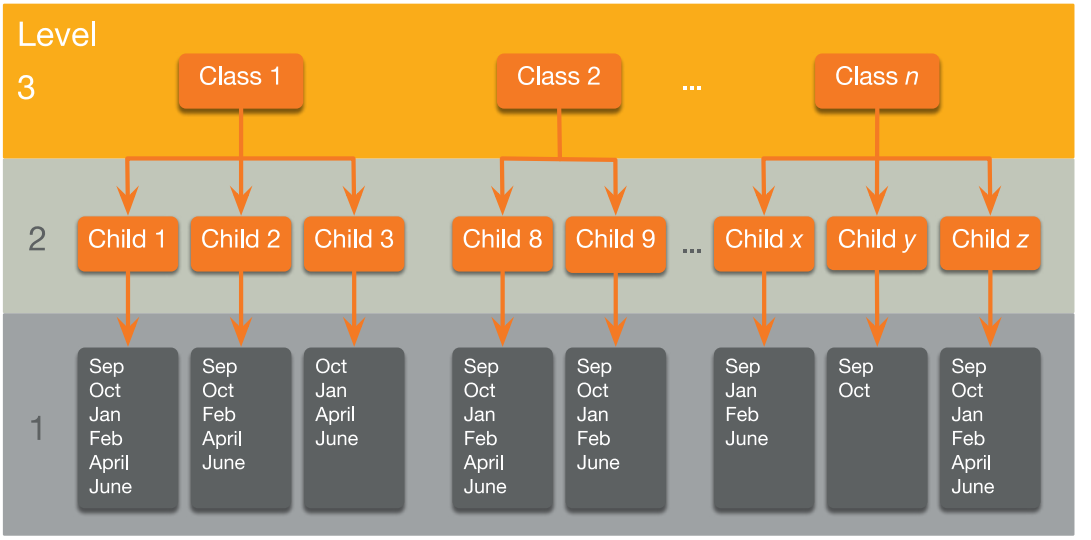


Figure 21.4 An example of a three-level hierarchical data structure, where the level 1 variable is a repeated measure (kindness scores measured at different points in time)

The reason we care that data are hierarchical (or not) is that the **contextual variables** in the hierarchy introduce dependency in observations. In a nutshell, residuals in the model will be correlated. To understand why, imagine that Charlotte and Emily are two children taught by Mr. Nervous, and Kiki and Jip are taught by Miss Daredevil. Charlotte and Emily's responses to the animal in the carrier have both been influenced by Mr. Nervous's cautious manner, so their behaviour may be similar. Their scores are likely to be correlated or dependent (because of the shared influence of Mr. Nervous). Likewise, Kiki and Jip's responses to the animal in the box are likely to be correlated because they have both been influenced by Miss Daredevil's carefree manner. However, Charlotte and Emily's responses shouldn't correlate with Kiki and Jip's, because the former pair have not been influenced by Mr. Nervous and the latter have not been influenced by Miss Daredevil.

This similarity is a problem because the linear model we've used until now assumes that errors are independent (Chapter 6), and when entities are sampled from similar contexts this independence is unlikely to be realistic. By thinking about contextual variables and factoring them into the model we can overcome this problem of non-independent observations.

We can use the intraclass correlation (which we came across as a measure of inter-rater reliability in Section 18.10) to estimate the dependency between scores. We'll skip the formalities of calculating the ICC (see Oliver Twisted if you're keen to know), and I'll try to give you a conceptual grasp of what it represents. In our two-level example of children within classes, the ICC represents the proportion of the total variability in the outcome that is attributable to the classes. It follows that if a class has had a big effect on the children within it then the variability within the class will be small (the children will behave similarly). As such, variability in the outcome within classes is minimized, and variability in the outcome between classes is maximized; therefore, the ICC is large. Conversely, if the class has little effect on the children then the outcome will vary a lot within classes, which will make differences between classes relatively small. Therefore, the ICC is small too. Thus, the ICC tells us that variability within levels of a contextual variable (in this case the class to which a child belongs) is small, but between levels of a contextual variable (comparing classes) is large. As such the ICC is a good gauge of the extent to which a contextual variable has an effect on the outcome.



Oliver Twisted
Please, Sir, can I have some more ... ICC?

'I have a dependency on gruel,' whines Oliver. 'Maybe I could measure this dependency if I knew more about the ICC.' We'll you're so high on gruel, Oliver, that you have rather missed the point. Still, I did write an article on the ICC once upon a time (Field, 2005a), and it's reproduced at www.discoverspss.com for your delight and amusement.

21.3 Multilevel linear models ©

The class of models we discuss in this chapter are called **multilevel linear models** (Figure 21.5) and they are able to include information about hierarchies in the data. The underlying theory of multilevel models is far too complicated for my peanut of a brain to comprehend. Fortunately, the advent of computers and software makes it possible for feeble-minded individuals such as me to take advantage of this tool without needing to know the maths. Better still, this means I can get away with not explaining the maths (I'm not kidding, I don't understand any of it). Instead, I'll try to give you a flavour of what multilevel models are and what they do by describing the key concepts within the framework of linear models that has permeated this whole book.



Figure 21.5 Thanks to the Confusion machine, there are many ways to refer to a multilevel model

21.3.1 A surgical example Ⓑ

In 2020 over 900,000 people had aesthetic plastic surgery procedures in the USA (Aesthetic Society, 2020). There are two main reasons to have cosmetic surgery: (1) to help a physical problem (i.e., breast reduction surgery to relieve backache, skin grafts after accidents); and (2) to change your external appearance when there is no underlying physical pathology. Related to this second reason, some have proposed cosmetic surgery as a psychological intervention to improve self-esteem (S. A. Cook et al., 2006; Kellett et al., 2008), an idea that I find abhorrent, but not abhorrent enough to stop me from creating an example around the effects of cosmetic surgery on quality of life. The variables in the data file (**cosmetic.sav**) are:

- **id**: The participant's identifier code.
- **post_qol**: This is the outcome variable and it measures quality of life after cosmetic surgery.
- **base_qol**: The quality of life before the surgery. We need to adjust our outcome for this variable.
- **days**: The number of days after surgery that post-surgery quality of life was measured.
- **clinic**: This variable specifies which of 21 clinics the person attended to have their surgery.
- **bdi**: Some research suggests that people volunteering for cosmetic surgery (especially when the surgery is purely for vanity) might, on average, have lower self-esteem or be more depressed than the general population (S. A. Cook et al., 2006). This variable measures levels of depression using the Beck Depression Inventory (BDI).
- **reason**: This variable specifies whether the person had surgery purely to change their appearance or because of a physical reason.

Figure 21.6 shows the hierarchical structure of the data. Essentially, people being treated in the same surgeries are not independent of each other because they will have had surgery from the same surgeon

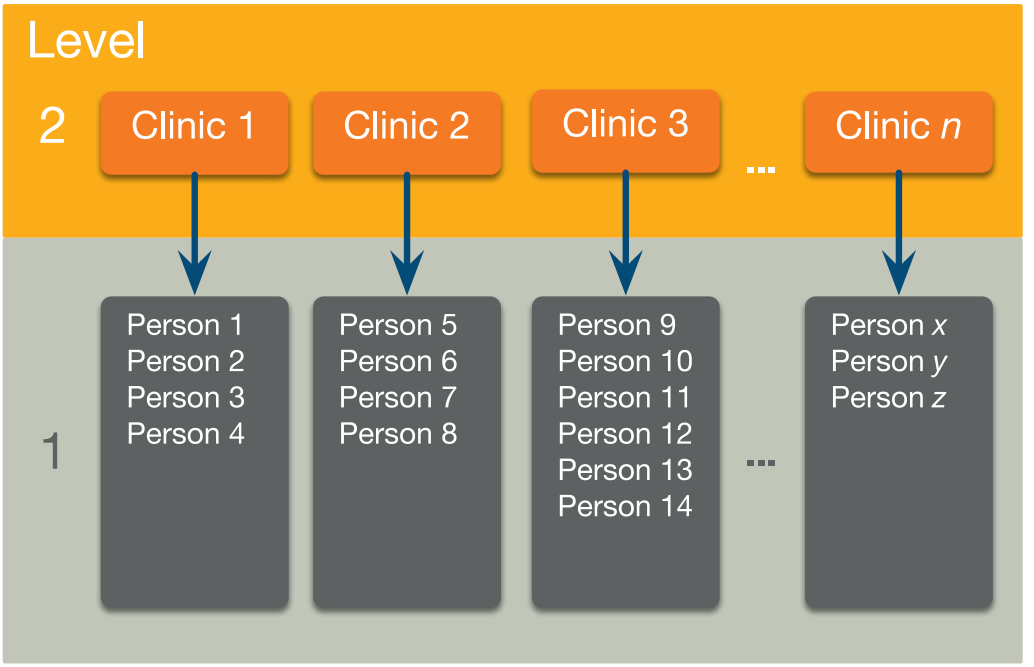


Figure 21.6 Diagram to show the hierarchical structure of the cosmetic surgery data set. People are clustered within clinics. For each person several variables were measured: post-surgery quality of life, BDI, the number of days post surgery that quality of life was measured, the reason for the surgery and pre-surgery quality of life

(or team). Surgeons will vary in how good they are, and quality of life will to some extent depend on how well the surgery went (if they did a neat job then quality of life should be higher than if they left unpleasant scars or there were complications). Therefore, people within clinics will be more similar to each other than people in different clinics. As such, the person undergoing surgery is the level 1 variable, and the clinic attended is a level 2 variable.

21.3.2 Fixed and random coefficients ©

The concepts of effects and variables should be very familiar to you by now. Throughout the book we have viewed these concepts a bit simplistically by treating all effects as 'fixed'. However, effects can also be conceptualized as 'random'. The terms 'fixed' and 'random' are confusing because they are used differently in a variety of contexts. For example, an effect in an experiment is said to be a **fixed effect** if all possible treatment conditions in which a researcher is interested are present in the experiment, but it is a **random effect** if the experiment contains only a random sample of possible treatment conditions. This distinction is important because fixed effects can be generalized only to the situations in your experiment, whereas random effects can be generalized beyond the treatment conditions in the experiment (provided that the treatment conditions are representative). For example, in our puppy therapy example from Chapter 12, the effect is fixed if we say that we are interested only in the three conditions that we had (no puppies, 15 minutes and 30 minutes) and we can generalize our findings only to the situation of treating people with puppies. However, perhaps kittens would be equally as effective, or rats? Ponies? Piglets? Lambs? Ferrets? Tarantulas? OK, not tarantulas. Anyway, imagine we believe that therapy with any kind of furry mammal would suffice. We set up a study comparing a waiting list (no therapy) to therapy groups involving puppies, kittens and rats. If we treat these groups as a fixed effect then we can't generalize beyond treating people with puppies, kittens or rats. However, if we were to conceptualize the three animals as a sample of possible mammals (we could have used piglet therapy), then it is a random effect and we can generalize beyond puppy, kitten and rat therapy to therapies with other mammals. The predictor variables in the examples up to now have been treated as fixed effects and the vast majority of academic research that you read will treat predictor variables as fixed effects.

People also talk about **fixed variables** and **random variables**. A fixed variable is one that is not supposed to change over time (e.g., eye colour is a fixed variable – it doesn't change), whereas a random one varies over time (e.g., your weight is likely to fluctuate over time).

In the context of multilevel models we make a distinction between **fixed coefficients** and **random coefficients**. In the linear models we have used we have assumed that the *b*-values (parameters) are fixed. For example, we have seen numerous times (e.g., Chapter 9) that the general linear model with one predictor is expressed as:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i \tag{21.1}$$

We estimate the various parameters (and put hats on them to remind us that they are estimates) from data collected in a sample and express the model as:

$$Y_i = \hat{b}_0 + \hat{b}_1 X_i + e_i \tag{21.2}$$

The outcome (Y_i), predictor (X_i) and residual (e_i) all vary as a function of i, which normally represents a particular case of data. In other words, it represents the level 1 variable. If, for example, we wanted to model Sam's score, we could replace the is with her name:

$$Y_{\text{Sam}} = \hat{b}_0 + \hat{b}_1 X_{\text{Sam}} + e_{\text{Sam}} \tag{21.3}$$

This is revision. When we fit this linear model we assume that the bs are fixed values that we estimate from the data. In doing so, we assume that the model holds true across the entire sample and that for

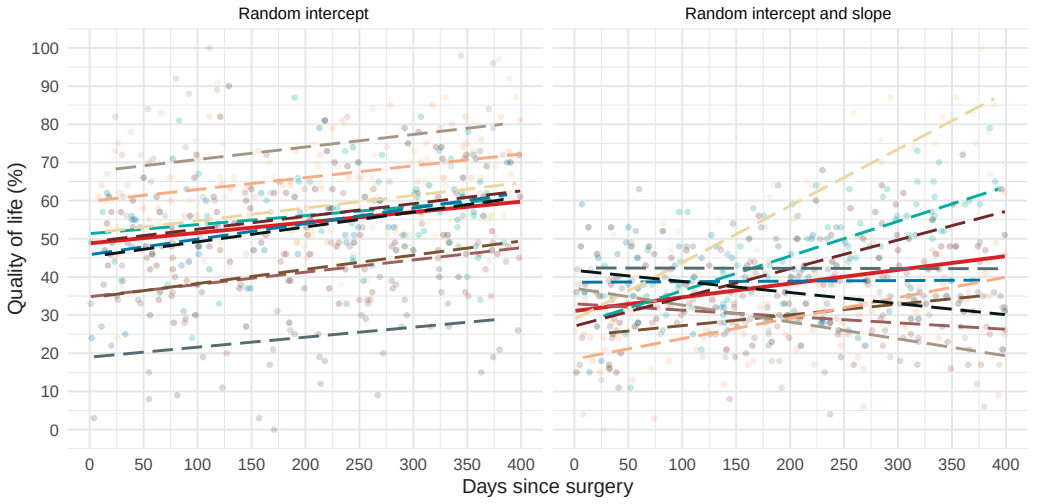


Figure 21.7 Data sets showing an overall model (red line) and the models for separate contexts within the data (i.e., individual clinics are represented by different colours)

every case of data we can predict a score using the same *b*-values for the predictor and intercept. However, these parameters can also conceptualized as 'random' variables,[3] meaning that they can vary from context to context. In other words, their values are not fixed.

Up until now we have thought of linear models as having **fixed intercepts** and **fixed slopes**, but the idea that parameters can vary across contexts creates the two possibilities shown in Figure 21.7. This figure shows both the overall relationship between the time since surgery and quality of life (as a solid red line), but also the same relationship in 10 different clinics (shown in different colours and as dashed lines). The first possibility is to model intercepts that vary across contexts (or groups) – a **random intercepts** model. This model is like assuming that the relationship between the time since surgery and quality of life is the same across the 10 clinics (i.e., the slope is constant across contexts), but that the level of quality of life when surgery is zero differs across the contexts (i.e., the intercepts are different in different clinics). In other words, the model allows for the possibility that when the post-surgery measure is taken on the same day as the surgery took place (so the variable **day** = 0) the level of quality of life varies across the 10 clinics, but assumes that the change in quality of life over the subsequent days is constant across the 10 clinics (i.e., the change in quality of life after surgery is the same in the 10 clinics). This model is represented in the left-hand panel of Figure 21.7, in which the same relationship exists between the days since surgery and quality of life in the 10 clinics (the slopes are the same) but level of quality of life for those measured immediately post surgery varies across clinics (which can be seen by the lines starting at different points along the quality of life scale).

A second possibility is that slopes, as well as intercepts, vary across contexts – that is, we model **random slopes**. This scenario is shown in the right-hand panel of Figure 21.7. In this situation, the models within the different contexts (colours) have different intercepts as they do in the left-hand panel, but they also have different slopes.[4] For our surgery data, this equates to modelling the possibility

3 'Random' isn't an intuitive term for us non-statisticians because it implies that values are plucked out of thin air (randomly selected). However, this is not the case; they are carefully estimated just as fixed parameters are.

4 Mathematically there's no reason why you couldn't model a situation where the slopes are free to vary but intercepts are fixed. In this example, that would force the quality of life on the day of surgery (i.e., **day** = 0) to be identical in each clinic but would allow its change over time to be different in each clinic. This situation is rarely plausible because variability in the relationship (slopes) would normally be associated with variability in the level of the outcome variable (intercepts).

that the relationship between the number of days post surgery and quality of life differs across the 10 clinics, and that levels of quality of life for those measured immediately post surgery (i.e. when **days** = 0) also differ across the clinics.

Let's build up our model step-by-step. Let's begin with a model containing fixed effects and then introduce parameters to the model that reflect the variability in these effects across contexts (e.g., allow the fixed effects to become random effects). Let's use our cosmetic surgery example and imagine that we first wanted to predict someone's quality of life (QoL) based on the number of days since cosmetic surgery. We can represent this as the following linear model:

$$\begin{aligned} Y_i &= b_0 + b_1 X_i + \varepsilon_i \\ \text{QoL}_i &= b_0 + b_1 \text{Days}_i + \varepsilon_i \end{aligned} \tag{21.4}$$

People typically use the Greek letter beta (β) to represent fixed effect parameters in multilevel models. To get us used to the idea of using β, let's rewrite this model as

$$\text{QoL}_i = \beta_0 + \beta_1 \text{Days}_i + \varepsilon_i \tag{21.5}$$

Notice that the two equations are the same except that one uses the letter b to represent parameters and the other uses β. As I've mentioned before, the choice of symbols is arbitrary (Jane Superbrain Box 2.3), so the interpretations of parameters that you have already learnt are unchanged by whether we represent them with b or β. Finally, let's remind ourselves that the errors in this model are normally distributed (Jane Superbrain Box 6.3) around zero, by noting:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{21.6}$$

21.3.3 Random intercepts ©

In this example, cosmetic surgery is conducted in different clinics. We might expect the effect of the time since surgery on QoL to vary as a function of the clinic in which the surgery was conducted because clinics use different surgeons, may have different care plans, and so on. This is a level 2 unit. We could allow the model that represents the effect of surgery on QoL to vary across the different contexts (clinics) by allowing the intercepts, slopes or both to vary across clinics.

To begin with, let's change the intercept for QoL from a fixed to a random effect. To model random intercepts we add a term, u_{0j}, that estimates the difference between the intercept of a particular clinic and the overall intercept (the intercept when we ignore clinics). In doing so, we expand β_0 to become $\beta_0 + u_{0j}$. We estimate both the intercept of the overall model ignoring the clinic attended, β_0, and the variability from that overall intercept of the intercepts for each individual clinic, u_{0j}. The overall model becomes:

$$\begin{aligned} Y_{ij} &= (\beta_0 + u_{0j}) + \beta_1 X_{ij} + \varepsilon_{ij} \\ \text{QoL}_{ij} &= (\beta_0 + u_{0j}) + \beta_1 \text{Days}_{ij} + \varepsilon_{ij} \end{aligned} \tag{21.7}$$

The *j*s in the equation reflect values of the level 2 variable over which the intercept varies (in this case the clinic). It is common to organize the terms so that the fixed effects are collected separate from the random parts:

$$\text{QoL}_{ij} = [\beta_0 + \beta_1 \text{Days}_{ij}] + [u_{0j} + \varepsilon_{ij}] \tag{21.8}$$

The first set of brackets contains the fixed effects (the overall intercept and slope) and is referred to as the *structural* part of the model. The second set of square brackets contains the random effects (the deviations of the intercepts from the overall intercept and the individual error term) and is referred to as the *stochastic* part of the model.

Let's visualize what u_{oj} represents. Figure 21.8 (left) shows a simplified situation with only five of the 20 clinics displayed. I've also plotted **months** since surgery rather than days (for reasons that will become apparent later in the chapter). The estimated overall relationship between the days since surgery and quality of life for the five clinics is shown as a dashed red line. The estimated individual relationships in each clinic are shown as solid coloured lines. The intercept, as with any intercept, is the value of the outcome when the predictor (in this case **months**) is zero. The overall estimated intercept ($\hat{\beta}_0$) is the red square when months is 0 (it has a value of 30.96) and the estimated intercept for each clinic is the corresponding-coloured circle. The vertical distance of each circle from the overall intercept (the red square) is the deviation of a clinic's intercept from the overall intercept. This distance from the estimated overall intercept for clinic *j* is estimated as $\hat{u}_{0j}$ and is shown for clinics 1, 5, 8 and 12 in Figure 21.8 (left).

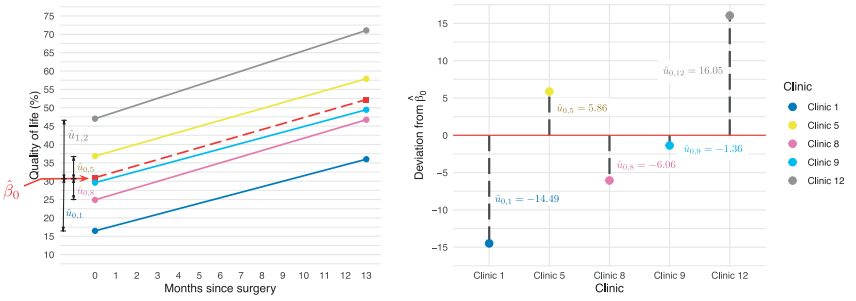


Figure 21.8 Delving into random intercepts

Figure 21.8 (right) makes these distances from $\hat{\beta}_0$ more explicit. The red horizontal line represents the overall intercept and therefore has a deviation of 0; around this line we see each clinic's deviation from the overall intercept. We can see, for example, that the intercept for clinic 1 is 14.49 units *below* the overall intercept, $\hat{u}_{01} = -14.49$. For clinic 2 the intercept is 5.86 units *above* the overall intercept, $\hat{u}_{02} = 5.86$. In general, the intercept of clinic *j* (β_{0j}) is made up of the overall intercept plus its deviation from it. For example, for clinic 1 the estimated intercept is:

$$\begin{aligned} \beta_{0j} &= \beta_0 + u_{0j} \\ \hat{\beta}_{01} &= \hat{\beta}_0 + \hat{u}_{01} \\ &= 30.96 + (-14.49) \\ &= 16.47 \end{aligned} \tag{21.9}$$

Notice that this value of $\hat{\beta}_{01}$= 16.41 corresponds to the *y*-axis value for clinic 1 in the left-hand panel of Figure 21.8. The fact that each clinic's intercept has a deviation from the overall intercept highlights that each intercept is, in fact, an observation from a variable that has an associated distribution. More formally, there is a variable, u_0, from which the specific deviation from the overall intercept for a clinic *j* (u_{0j}) is sampled. We assume that these deviations are modelled by a normally distributed with mean 0 (on average clinic intercepts do not deviate from the overall intercept) and a variance, $\sigma^2_{u_0}$, that quantifies the degree to which intercepts from different clinics vary around the overall intercept. To conceptualize this variance, imagine that, having estimated the deviance of each intercept from the overall intercept in the 10 clinics (i.e., the 10 values of u_{0j} in Figure 21.8, right), we calculate the variance in the usual way. First, for each deviance we subtract the mean (in this case zero, leaving the deviations as they were), square them, sum the squared deviations and finally divide by a function of the number of scores. Although the calculation is more complicated than that, *conceptually*, that's what $\hat{\sigma}^2_{u_0}$ represents: it is an estimate (so, we give it a hat) of the variance of the intercepts. More formally, we can state:

$$u_0 \sim \mathcal{N}(0, \sigma^2_{u_0}) \tag{21.10}$$

which says that the intercepts are normally distributed with mean 0 and variance $\sigma^2_{u_0}$. Therefore, by making the intercepts a random effect we have added a new parameter to the model – the variance of intercepts ($\hat{\sigma}^2_{u_0}$) – that needs to be estimated.

21.3.4 Random slopes ©

If we want to make the slope for the effect of surgery on QoL a random effect, we do something similar. We add a component to the slope of the overall model, β_1, that quantifies the extent to which slopes in each context deviate from it, u_{1j}. Remember that we currently have the model in Eq. 21.7:

$$Y_{ij} = (\beta_0 + u_{0j}) + \beta_1 X_{ij} + \varepsilon_{ij}$$

By adding the additional term for the deviation in slopes, the model becomes:

$$\begin{aligned} Y_{ij} &= (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})\, X_{ij} + \varepsilon_{ij} \\ \text{QoL}_{ij} &= (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})\text{Days}_{ij} + \varepsilon_{ij} \end{aligned} \tag{21.11}$$

If we expand the brackets, we get:

$$\text{QoL}_{ij} = \beta_0 + u_{0j} + \beta_1 \text{Days}_{ij} + u_{1j}\text{Days}_{ij} + \varepsilon_{ij}$$

If we express this model in the composite form by collecting the fixed and random parts in brackets we get

$$\text{QoL}_{ij} = [\beta_0 + \beta_1 \text{Days}_{ij}] + [u_{0j} + u_{1j}\text{Days}_{ij} + \varepsilon_{ij}] \tag{21.12}$$

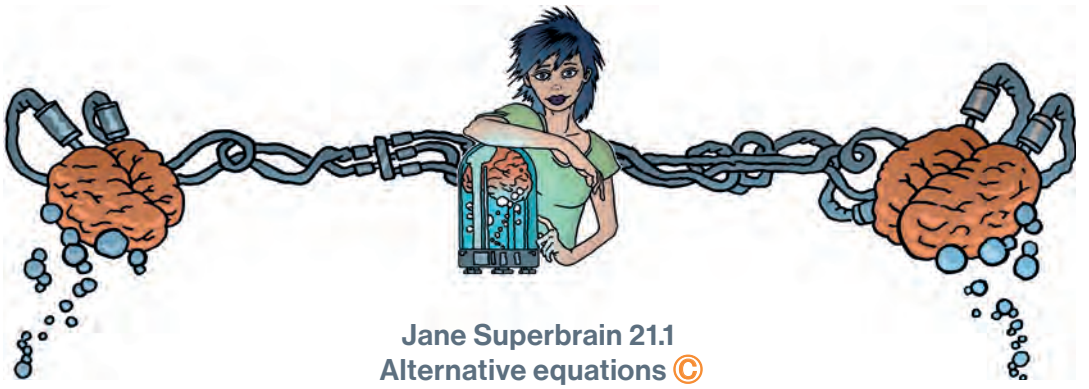
As with the random intercept model, the first set of brackets contains the fixed effects (the overall intercept and slope) or *structural* part of the model. The second set of square brackets contains the random effects (the deviations of the intercepts and slopes from the overall values and the individual error term) or *stochastic* part of the model. These models are sometimes expressed in a different format in which the models at different levels are separated (Jane Superbrain Box 21.1).

Let's delve into what u_{1j} represents. Figure 21.9 shows another five of the 20 clinics and, as before, I've plotted the months since surgery rather than days. The estimated overall model is the dashed red line. To avoid clutter, in the left hand panel I've included the individual models within only clinics 1 and 3. We're familiar with the idea that the slope is the change in the outcome as the predictor (in this case the months since surgery) increases by one unit. Therefore, the slope in this example represents the change in quality of life for every month since surgery. The estimated overall slope ($\hat{\beta}_1$) is the gradient of the red line, which has a value of 0.75, meaning that quality of life increases by 0.75 units for every month that passes since surgery.

For clinic 1 the equivalent estimated change in the outcome ($\hat{\beta}_{1,\,\text{clinic1}}$) is 1.50. We can calculate the difference between the estimated of clinic 1's slope ($\hat{\beta}_{11}$) and the overall slope ($\hat{\beta}_1$) as:

$$\begin{aligned} u_{1j} &= \beta_{1j} - \beta_1 \\ \hat{u}_{11} &= \hat{\beta}_{11} - \hat{\beta}_1 \\ &= 1.50 - 0.75 \\ &= 0.75 \end{aligned} \tag{21.13}$$

This value is the estimated deviation of clinic 1's slope from the overall slope as shown in Figure 21.9 (left). The positive value tells us that the slope in clinic 1 was larger than the overall effect – it is twice as big – in other words, in clinic 1 the rate of change of people's quality of life ratings was greater than the overall rate of change.



Jane Superbrain 21.1
Alternative equations ©

In this chapter the author tends to use the composite form of the equation to describe the multilevel model. However, it is common to find the models written in a (mathematically equivalent) way in which the level 1 part of the model is separated from the level 2 part (and indeed the level 3 part if there is one). For a random intercept model, the level 1 model looks like an ordinary linear model equation, except that the intercept has changed from fixed, β_0, to random, β_{0j}, with the j subscript telling us that the intercept varies by levels of j (in this case the clinics attended). This random intercept is defined in a separate equation underneath, which tells us that it is made up of the overall intercept (β_0) plus a deviation (u_{0j}) from that overall intercept:

$$\begin{aligned} \text{QoL}_{ij} &= \beta_{0j} + \beta_1 \text{Days}_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \end{aligned} \tag{21.14}$$

In this format, the first equation represents the part of the model at level 1 (the participant) and the second equation represents the part of the model at level 2 (the clinic).

For the random slopes model, we again start with the level 1 model, which is a standard linear model that replaces the fixed intercept and slope (β_0 and β_1) with their random counterparts (β_{0j} and β_{1j}). These random coefficients are then defined as separate equations underneath. Compare Eq. 21.14 with Eq. 21.15. In the first equation the only difference is that the fixed slope (β_1) has become random (β_{1j}), the second equation is the same (it defines the random intercept) and a third equation has been added that defines β_{1j} (from the first equation) in terms of the overall slope (β_1) and the deviation from it (u_{1j}) for each level of j (i.e.. each clinic):

$$\begin{aligned} \text{QoL}_{ij} &= \beta_{0j} + \beta_{1j} \text{Days}_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j} \end{aligned} \tag{21.15}$$

Again the first equation represents the part of the model at level 1 (the participant), but now there are two subsequent equations that represent the part of the model at level 2 (the clinic).



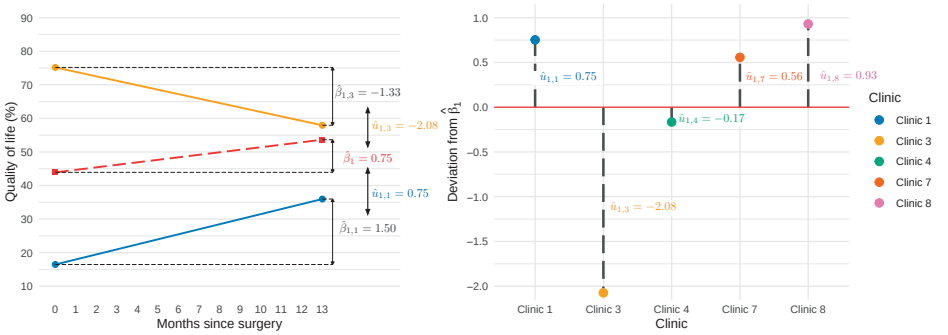


Figure 21.9 Delving into random slopes

Similarly, the difference between clinic 3's slope ($\hat{\beta}_{13}$) and the overall slope ($\hat{\beta}_1$) is:

$$\begin{aligned}\hat{u}_{13} &= \hat{\beta}_{13} - \hat{\beta}_1 \\ &= -1.33 - 0.75 \\ &= -2.08\end{aligned} \tag{21.16}$$

This value is the deviation of clinic 3's slope from the overall slope as shown in Figure 21.9 (left). The fact that the value is negative tells us that the overall slope was larger than the slope in clinic 3. In fact, the slope in clinic 3 was in the opposite direction to the overall slope, so we get a fairly large negative deviation. Think carefully about what these deviations mean: a value of zero means that the slope in the clinic is identical to the overall slope, a positive deviation means that the slope in the clinic is more positively sloped than the overall slope, and a negative value means that the slope in the clinic is more negatively sloped. Note, for example, that the magnitude of the slopes in clinics 1 and 3 are similar (1.50 and 1.33), but the directions are opposite (clinic 1 has a positive slope but clinic 3 has a negative slope).

As with the intercepts, each clinic has a value quantifying the deviation of its slope from the overall slope, and I've plotted these deviations in Figure 21.9 (right) as well as the deviations for another three clinics (which are not in the left-hand figure). The red horizontal line represents the estimated overall slope ($\hat{\beta}_1$) and has a deviation of 0, and each $\hat{u}_{1j}$ is the estimated deviation of clinic j from this value. We have just calculated, for example, that the slope for clinic 1 is 0.75 units *more than* the overall slope, $\hat{u}_{11} = 0.75$, and the slope for clinic 3 is 2.08 units *less than* the overall slope, $\hat{u}_{13} = -2.08$. By rearranging Eq. 21.15 we can see that the estimated slope of clinic j ($\hat{\beta}_{1j}$) is made up of the overall slope plus its deviation from it. For example, for clinic 1 the slope is:

$$\begin{aligned}\beta_{1j} &= \beta_1 + u_{1j} \\ \hat{\beta}_{11} &= \hat{\beta}_1 + \hat{u}_{11} \\ &= 0.75 + 0.75 \\ &= 1.50\end{aligned} \tag{21.17}$$

Each slope is an observation from a variable which has an associated distribution. More formally, there is a variable, u_1, from which the specific deviation from the overall slope for a clinic j (u_{1j}) is sampled. Again, we assume that these deviations are modelled by a normally distributed with mean 0 (on average slopes within clinics do not deviate from the overall slope) and a variance, $\sigma^2_{u_1}$, that quantifies the degree to which slopes from different clinics vary around the overall slope. To visualize this variance, imagine that we calculate the variance of the five values of u_{1j} in Figure 21.9, right. Conceptually, the resulting value is what the estimate of the variance of the slopes ($\hat{\sigma}^2_{u_1}$) represents. More formally, we can state:

$$u_1 \sim N(0, \sigma^2_{u_1}) \tag{21.18}$$

which says that the deviations from the overall slope are normally distributed with mean 0 and variance $\sigma^2_{u_1}$. By modelling slopes as random effects we add a parameter to the model that needs to be estimated, the variance of the deviations of slopes, $\hat{\sigma}^2_{u_1}$.

In fact, we add *two* new parameters to the model because the slopes and intercepts will correlate. To explain why, let's simplify things and imagine we're predicting quality of life from time since surgery in years, and we've taken measures for exactly 1 year. This means that the gradient for quality of life is the same as the difference in quality of life between someone measured at baseline and someone measured at 1 year.[5] Now, imagine a clinic in which the intercept is 95% on the quality of life scale. Let's assume that quality of life *increases* as the time from surgery increases. Quality of life is measured as a percentage so, at the very most, predicted quality of life in this clinic can change over time by a maximum of five units from 95% (at the intercept) to 100% (at the final time point). In short, the slope in this clinic is constrained to a maximum value of 5 by the upper limit of the quality of life scale. Now imagine the same clinic but that quality of life *decreases* with time. In this scenario, the predicted quality of life can change over time by a maximum of 95 units from 95% (at the intercept) to 0% (at the final time point). Therefore, the slope has the potential to be very steep (at most, –95). This extreme example illustrates that slopes are not independent of intercepts.

Given that slopes and intercepts are co-dependent, we need a parameter to estimate the covariance between them, which we'll denote by:

$$\text{covariance between slopes and intercepts} = \sigma_{u_0, u_1} \tag{21.19}$$

To sum up, by modelling the intercept and the slope for a predictor as random variables we introduce three new parameters: the variability in intercepts, the variability in slopes (for the predictor) and the covariance between slopes and intercepts.

21.3.5 The covariance structure of random effects Ⓓ

We can bring the random effects together more formally by combining Eq. 21.10 and Eq. 21.18 into a single expression:

$$\begin{bmatrix} u_0 \\ u_1 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0, u_1} \\ \sigma_{u_0, u_1} & \sigma^2_{u_1} \end{bmatrix} \right) \tag{21.20}$$

which looks terrifying. It's basically a mathsy way of saying that the deviations of the slopes and intercepts (the stuff on the left-hand side of the tilde) are assumed to follow a multivariate normal distribution (that's what the $\mathcal{N}()$ represents) with means of 0 (the first set of square brackets within $\mathcal{N}()$) and variances defined by a variance–covariance matrix (the second set of square brackets within $\mathcal{N}()$). We have briefly encountered variance–covariance matrices earlier in the book. A variance–covariance matrix contains variances along the diagonal elements and the off-diagonal elements are covariances. Matrices can seem scary, but imagine them as a table where the row names and column headers have been removed. For example, the variance–covariance matrix for the random effects in our model could be viewed as a table with intercepts represented in the first row and column and slopes in the second row and column. The matrix itself is the part in parentheses:

5 The gradient is the change in the outcome divided by the change in the predictor. If we measure the predictor in years then the change in the predictor is 1 – 0 years = 1, and the gradient becomes the change in the outcome (divided by 1).

$$\begin{array}{c} \\ \text{Intercepts} \\ \text{Slopes} \end{array} \begin{array}{c} \begin{array}{cc} \text{Intercepts} & \text{Slopes} \end{array} \\ \begin{pmatrix} \text{Variance}_{\text{Intercepts}} & \text{Covariance}_{\text{Intercepts,Slopes}} \\ \text{Covariance}_{\text{Intercepts,Slopes}} & \text{Variance}_{\text{Slopes}} \end{pmatrix} \end{array}$$

Replacing the parts in the matrix with their symbols gives us

$$\begin{array}{c} \\ \text{Intercepts} \\ \text{Slopes} \end{array} \begin{array}{c} \begin{array}{cc} \text{Intercepts} & \text{Slopes} \end{array} \\ \begin{pmatrix} \sigma^2_{u_0} & \sigma_{u_0,u_1} \\ \sigma_{u_0,u_1} & \sigma^2_{u_1} \end{pmatrix} \end{array}$$

This variance–covariance matrix defines the *covariance structure* for the random effects. The key thing is that this matrix contains information about the degree to which intercepts and slopes vary across contexts (along the diagonal) and the degree to which the value of the intercept is associated with the value of the slope (on the off-diagonal). By default, this matrix has an *unstructured* form, meaning that all variances and covariances are estimated, but other structures can be used (see Section 21.4.7):

$$\begin{pmatrix} \sigma^2_{u_0} & \sigma_{u_0,u_1} & \cdots & \sigma_{u_0,u_n} \\ \sigma_{u_0,u_1} & \sigma^2_{u_1} & \cdots & \sigma_{u_1,u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{u_0,u_n} & \sigma_{u_1,u_n} & \cdots & \sigma^2_{u_n} \end{pmatrix} \tag{21.21}$$

21.3.6 Level 1 errors ©

We noted in Eq. 21.6 that the linear model contains an error term (ε_i) that is assumed to come from a normal distribution with mean 0 and constant variance. The multilevel model in Eq. 21.8 also contains this error term, but its subscripts reflect the fact that it now varies not just by individual i but also within context j (ε_{ij}). Figure 21.10 shows what these errors represent. To simplify the diagram, I have included only the data from clinic 3. As in Figure 21.8 and Figure 21.9 the red line represents the overall fixed effect, with the overall intercept β_0 represented by quality of life when the time since surgery is 0 and the overall slope (β_1) being the rate of change of quality of life as the time since surgery increases. Three participants have been highlighted on the plot in different colours. How do the observed values of these individuals differ from those predicted by the overall model (the red line)? The extent to which a participant's quality of life differs from that predicted by the overall model is affected by two things. First, it is affected by the clinic to which they belong. These participants all belong to clinic 3, so we know that the intercept for the clinic to which all of these individuals belong differs from the overall intercept by the distance u_{03} (the 0 tells us this is an intercept, the 3 comes from clinic 3), and we also know that the slope for their clinic differs from the overall slope by the distance u_{13} (the 1 tells us this is the slope for the first predictor, the 3 comes from clinic 3). Second, it is affected by the error in prediction for the model within their clinic. That is, for each participant there is a difference between their observed quality of life and that predicted by the clinic 3 model. These are the distances ε_{13}, ε_{23} and ε_{33} in the figure (the first subscript relates to their participant number, i = 1, 2, or 3, and the second their clinic, j = 3). Therefore, the extent to which a participant's quality of life differs from that predicted by the overall model is made up of the deviation of the intercept and slope for their clinic from the overall intercept and slope (which is at level 2 of the model, the clinic level) and the deviation of their outcome from the value predicted by the model within their clinic (which is at level 1 of the model, the participant level). Therefore, the level 1 errors

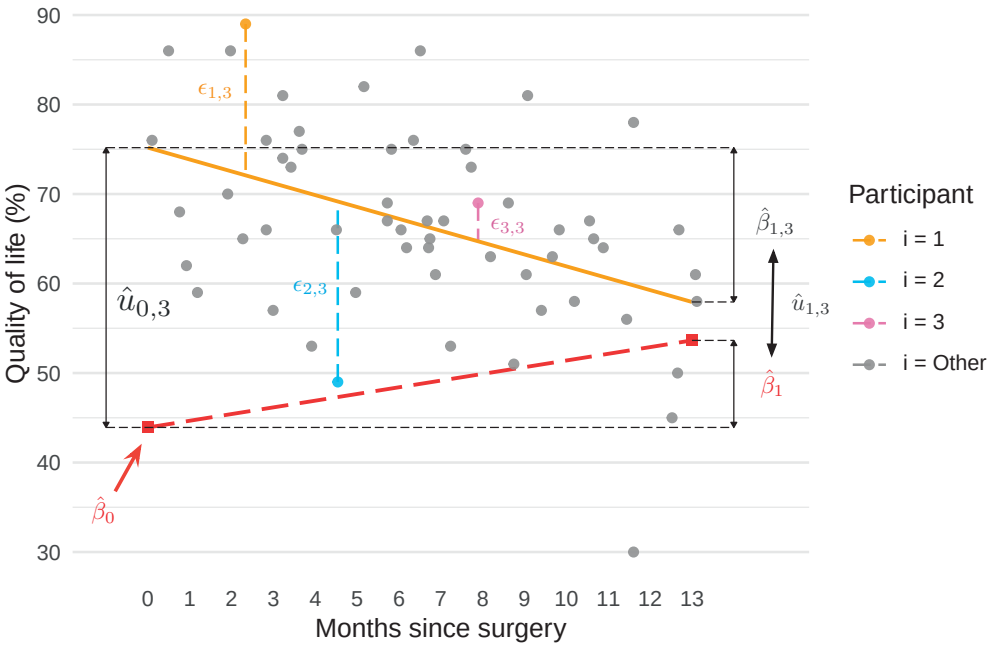


Figure 21.10 Level 1 errors

are the deviations of each participant's observed score from the model within their clinic. These level 1 errors are assumed to have a normal distribution around 0 and constant variance (just as was assumed in the standard linear model). We can express this assumption as:

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{21.22}$$

21.3.7 The covariance structure of the level 1 errors ©

We have talked about the variance–covariance matrix of the random effects, but we should also consider the variance–covariance matrix of the level 1 errors. These are the ε_{ij} that we have just discussed. Back in Section 6.2.8 we talked about the covariance structure of errors in a standard linear model and found that the assumption of spherical errors implied that the variance–covariance matrix looks like this (Eq. 6.30):

$$\Phi = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

To recap, this variance–covariance matrix assumes that errors are independent (represented by the off-diagonal elements – the covariances – being zero) and have constant variance (represented by the diagonal elements all being σ^2). By default, this covariance structure is used in a multilevel model, but other structures can be used (see Section 21.4.7).

21.3.8 Adding predictors ©

Imagine we wanted to add another predictor, for example the quality of life before surgery, which we'll call **Pre QoL**. Knowing what we do about linear models, we shouldn't be invading the personal space of the idea that we can extend Eq. 21.12 to include this variable along with an associated parameter (represented by a β). This predictor acts at the participant level (level 1) so we'd give it subscripts of i and j:

$$\text{QoL}_{ij} = [\beta_0 + \beta_1\text{Days}_{ij} + \beta_2\text{Pre QoL}_{ij}] + [\text{u}_{0j} + \text{u}_{1j}\text{Days}_{ij} + \varepsilon_{ij}] \quad (21.23)$$

Notice that we've included the parameter as a fixed effect (in the structural part of the model). We know this because there is no parameter that estimates the deviations in slopes for this predictor across levels of j (i.e., clinics). However, we *could* model this predictor as a random effect by including this parameter u_{2j} to the stochastic part of the model for the predictor **Pre QoL**:

$$\text{QoL}_{ij} = [\beta_0 + \beta_1\text{Days}_{ij} + \beta_2\text{Pre QoL}_{ij}] + [u_{0j} + u_{1j}\text{Days}_{ij} + u_{2j}\text{ Pre QoL}_{ij} + \varepsilon_{ij}] \quad (21.24)$$

In short, to add new predictors to a model we decide whether they are treated as fixed or random effects then expand the model accordingly. At the surface there isn't anything terribly complicated about this idea, but lurking below is a world of pain. By adding u_{2j} we expand the diagonal of the variance–covariance matrix in Eq. 21.20 to include a new parameter ($\sigma^2_{u_2}$) that represents the variance in slopes of baseline QoL:

$$\begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0,\,u_1} & ? \\ \sigma_{u_0,\,u_1} & \sigma^2_{u_1} & ? \\ ? & ? & \sigma^2_{u_2} \end{bmatrix}$$

By expanding the diagonal of the matrix, gaps are created that alert us to the fact that we also need to consider the relationship (covariance) between the slope of baseline QoL and the intercept ($\sigma_{u_0,\,u_2}$) and the relationship between the slope of baseline QoL and the slope for the other predictor, **days** (σ_{u_1,u_2}). The complete covariance structure for the random effects becomes:

$$\begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0,\,u_1} & \sigma_{u_0,\,u_2} \\ \sigma_{u_0,\,u_1} & \sigma^2_{u_1} & \sigma_{u_1,\,u_2} \\ \sigma_{u_0,\,u_2} & \sigma_{u_1,\,u_2} & \sigma^2_{u_2} \end{bmatrix}$$

Therefore, by modelling the slope of a new predictor as a random effect we have added *three* parameters to be estimated from the data: $\sigma^2_{u_2}$, σ_{u_0,u_2} and σ_{u_1,u_2}. If we added another predictor with a random slope that would create a further four parameters (a variance and three covariances) to be estimated and so on. In short, the number of parameters that need estimating from the data quickly escalates as you model more random effects because for every random effect (other things being equal) you need to estimate the covariances between that effect and all other random effects. You can quickly reach the point where you don't have enough data to estimate the parameters and the model will fail to converge.

21.3.9 The model we're going to fit ©

The final model that we will fit in this chapter takes Eq. 21.12 as its starting point. We're going to include baseline QoL as a fixed effect, making the model

$$\text{QoL}_{ij} = [\beta_0 + \beta_1\text{Days}_{ij} + \beta_2\text{Pre QoL}_{ij}] + [u_{0j} + u_{1j}\text{Days}_{ij} + \varepsilon_{ij}] \quad (21.25)$$

We're also going to include the reason for having cosmetic surgery as a fixed effect predictor, which would expand the model:

$$\text{QoL}_{ij} = [\beta_0 + \beta_1\text{Days}_{ij} + \beta_2\text{Pre QoL}_{ij} + \beta_3\text{Reason}_{ij}] + [u_{0j} + u_{1j}\text{Days}_{ij} + \varepsilon_{ij}] \quad (21.26)$$

Finally, we'll include the interaction between surgery and the reason for having it as a fixed effect predictor. This interaction term will establish whether the effect of surgery is moderated by the reason. The final model is expressed as:

$$\begin{aligned}\text{QoL}_{ij} = {} & [\beta_0 + \beta_1\text{Days}_{ij} + \beta_2\text{Pre QoL}_{ij} + \beta_3\text{Reason}_{ij} + \beta_4\text{Days} \times \text{Reason}_{ij}] \\ & + [u_{0j} + u_{1j}\text{Days}_{ij} + \varepsilon_{ij}]\end{aligned} \quad (21.27)$$

Notice that because the predictors we've added are treated as fixed effects it is only the structural part of the model that has changed. Jane Superbrain Box 21.2 shows these models expressed in the alternative form where the level 1 and level 2 models are separated.



Jane Superbrain 21.2
More alternative equations ©

A mathematically equivalent form of Eq. 21.25 is:

$$\begin{aligned}\text{QoL}_{ij} &= \beta_{0j} + \beta_{1j}\text{Days}_{ij} + \beta_2\text{Pre QoL}_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j}\end{aligned} \quad (21.28)$$

Similarly, Eq. 21.26 can be expressed as:

$$\begin{aligned}\text{QoL}_{ij} &= \beta_{0j} + \beta_{1j}\,\text{Days}_{ij} + \beta_2\text{Pre QoL}_{ij} + \beta_3\text{Reason}_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j}\end{aligned} \quad (21.29)$$

Finally, Eq. 21.27 can be expressed as:

$$\begin{aligned}\text{QoL}_{ij} &= \beta_{0j} + \beta_{1j}\text{Days}_{ij} + \beta_2\text{Pre QoL}_{ij} + \beta_3\text{Reason}_{ij} + \beta_4\text{Days} \times \text{Reason}_{ij} + \varepsilon_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j}\end{aligned} \quad (21.30)$$

Notice that because the predictors we've added are treated as fixed effects it is only the level 1 model that changes each time.



21.3.10 Benefits of multilevel models Ⓑ

To convince you that trawling through this chapter is going to reward you with statistical possibilities beyond your wildest dreams, here are just a few (slightly overstated) benefits of multilevel models:

- **Cast aside the assumption of homogeneity of regression slopes**: When we use analysis of covariance we assume that the relationship between the covariate and outcome is the same across the different groups that make up our predictor variable (Chapter 13). In multilevel models we can model this variability in slopes explicitly and forget about this assumption (provided we have enough groups).
- **Say 'bye bye' to the assumption of independence**: Linear models assume independent errors (Chapter 6). If errors are dependent, little lizards climb out of your mattress while you're asleep and eat you. Multilevel models allow you to model dependencies between errors (Section 21.4.7) and sleep safely.
- **Laugh in the face of missing data**: Linear models based on ordinary least squares (OLS) can do strange things when data are missing or when (experimental) designs are not balanced. Missing data are a particular problem within clinical trials and other longitudinal designs in which you might want to collect follow-up data months or years after an intervention (or other baseline) took place and patients/participants might be difficult to track down. There are ways to correct for and impute missing data, but these techniques are complicated. Multilevel models are estimated using maximum likelihood estimation (Section 2.7) and do not require complete data in repeated measures designs. Instead parameters can often be estimated successfully with the available data, which offers a relatively convenient solution to dealing with missing data, although the discussion of whether data are missing at random still applies. It is important to stress that no statistical procedure can overcome data that are missing. Good methods, designs and research execution should be used to minimize missing values and reasons for missing values should always be explored.

'Is there anything multilevel models can't do?' I hear you ask. Well, I've never had one make me tea.

21.4 Practical issues Ⓒ

Having raised your hopes by cheerleading multilevel models, it seems only right and proper that in this section I crush your hopes and dreams. Sorry. Multilevel models are incredibly complicated beasts and there are lots of difficult issues to consider, lots of differing advice and lots of things we simply don't yet know enough about.[6] This section delves into this abyss of pain.

6 Ben Bolker has written a lot of very useful things and maintains a lot of very helpful resources about multilevel modelling. He frequently pops up online offering really useful and detailed answers to people's questions and, to me at least, seems like an unsung hero. Much of 'my' advice is in fact a collation of Ben's advice across various forums/websites with the added risk of my having misunderstood him.



Cramming Sam's Tips
Multilevel models

- Multilevel models should be used to analyse data that have a hierarchical structure. For example, you might measure depression after psychotherapy. In your sample, patients will see different therapists within different clinics. This is a three-level hierarchy with depression scores from patients (level 1), nested within therapists (level 2), who are themselves nested within clinics (level 3).
- Hierarchical models are linear models in which you can allow parameters to vary (this is called a random effect) across contexts. In the standard linear model, parameters generally are a fixed value estimated from the sample (a fixed effect).
- If we estimate a linear model within each context (the therapist or clinic, to use the example above) rather than the sample as whole, then we can estimate the variation in the intercepts around the overall intercept (so-called 'random intercepts'), and also model the variation in the slopes around the overall slope (so-called 'random slopes').
- When we model variation in slopes and intercepts we add a parameter to the model that estimates this variation, but we also add parameters estimating the covariance (relationship) between all random effects in the model.



21.4.1 Computing *p*-values Ⓓ

To compute a *p*-value you need the cumulative distribution function of the test statistic of interest when the null hypothesis is true. In the linear models we've covered up to now this is possible – provided some assumptions are met – because we know how the test statistics are distributed. However, multilevel models can be fitted to a wide range of data structures with differing numbers of random effects and sometimes the cumulative distribution function of the test statistic is not known, therefore you can't always compute a *p*-value. To put the scale of this problem into perspective, across the 350 or so pages of Goldstein's seminal multilevel model text, he barely mentions *p*-values (Goldstein, 2011). Much as I have tried to emphasize estimation and interpreting parameters, we have also talked a lot about *p*-values. If you're a student, then your lecturer/professor/supervisor might like *p*-values. If you're a researcher, you'll find that a lot of scientific journal editors and reviewers work themselves into quite the frenzy if they see a results section devoid of *p*-values. What to do?

It turns out that is a difficult question to answer (for a review see Bolker, 2015). Out there is a gateway to hell labelled '*p*-values in multilevel models'. Open the door find a room full of books and articles that offer mental anguish in abundance but very little in terms of concrete advice. I have sat in this room for many hours so that you don't have to. I left the room with the general impression

that the advice that will never lead you astray is not to use *p*-values for multilevel models. Interpret the parameter estimates (be they Bayesian or frequentist) and move on with life a happier person. The more nuanced advice is … here goes …

You can get *p*-values using a likelihood test (Section 19.3.4) and in balanced designs with large samples this approach is reasonable. For models based on normally distributed errors (which is the topic of this chapter), when you have a balanced design that is nested, you can use an *F*-distribution to get a *p*-value. However, in models of unbalanced (the example in this chapter) or crossed designs the degrees of freedom for the *F*-distribution are unknown. There are numerous ways to approximate them, for example the Satterthwaite and Kenward–Roger methods (Kenward & Roger, 2009; Satterthwaite, 1946), but there is no way of knowing whether these approximations are correct (Baayen et al., 2008). The Satterthwaite and Kenward–Roger methods seem to perform well in the limited model designs in which they have been studied; for example, in a single categorical predictor model, with a random slope and intercept the Satterthwaite and Kenward–Roger methods had the expected Type I error rates (Luke, 2017). However, it's unclear whether this would be true in more complex designs or with a more complex covariance structure. These approximations aren't implemented for models not based on normally distributed errors (e.g. modelling binary outcomes such as in Chapter 20).

Another possibility is the bootstrap (Section 6.10.8). The idea here is to fit a baseline or 'null' model containing no predictors to the data. Then simulate new data from this null model. The simulated data act as a bootstrap sample within which you fit the baseline model and the full model with all fixed and random effects of interest and estimate parameters and calculate the relevant tests statistics. Repeat this process across lots of bootstrap samples simulated from the null model. You now have some idea of the distribution of test statistics/parameter estimates generated from the null model, from which you can obtain an empirical probability of observing a test statistic at least as large as the one observed under the null model. This seems like a promising option (for confidence intervals too) but we don't currently know very much about how this approach performs under a range of models.

To sum up, there is no unifying method to compute *p*-values in multilevel models. The approximations we've discussed are probably (but not certainly) fine for the sorts of models in this chapter, but we should focus on interpreting the parameter estimates of our effects and treat significance cautiously.

21.4.2 Should an effect be fixed or random? ©

There are three broad approaches to deciding whether an effect should be fixed or random: (1) theoretical; (2) always fit the maximal model; and (3) data-driven. The theoretical approach is to take the view that if, theoretically, you'd expect something to vary across contexts then you model that variance. In our specific example, if we're expecting the relationship between the days since surgery and quality of life to vary across clinics we include a random intercept and slope for **days** and estimate that variability. Conversely, if we are not expecting the relationship between the days since surgery and quality of life to vary across clinics then we leave it as a fixed effect. Basically, your decision is dictated entirely by theoretical concerns.

Alternatively, some argue that you should always fit the model including all random effects (the maximal model) because (in some situations) overfitting the model has little consequence for Type I error rates and power (Barr et al., 2013). The data-driven approach is to build up the model from a minimal baseline model, add random effects and assess them (perhaps with a significance test). Random effects with parameters close to zero or that are not significant are replaced with fixed effects. A reverse approach is to start with the maximal model (if it will converge) and systematically remove random effects that are not significant (Matuschek et al., 2017).

In previous editions of this book (and its related books) I did a hybrid of theoretical and data-driven; that is, build the model up from a baseline but don't throw out a random effect if it is

theoretically meaningful. However, the older and wiser (ha, ha, ha!) me prefers a theory-driven approach. Here's why. The data-driven approach is problematic because, as we have seen, significance tests in multilevel models are not without complications and the decisions that you make are at the mercy of both the power and accuracy of these tests. In short, your model is too beholden to the whims of significance testing. The maximal model approach is problematic because, as we've seen, adding random effects proliferates the number of parameters to be estimated; you can quickly find yourself with a model with more parameters than can be estimated from the data. Also, although overfitting models may (in some situations) not affect Type I error rates, it can (in different situations) affect power (Matuschek et al., 2017) and lead to a 'singular fit' (Section 21.4.3), which is a bad thing.

In practice, the theory-driven and maximal model approaches might not be very different. In experimental designs a theory-driven approach would typically equate to fitting the maximal model. Therefore, the theory-driven and maximal model approaches align in the situation (a properly designed and powered experiment) where you stand the most chance of the maximal model converging (but note the previous caveats about overfitting models). The theory-driven and maximal model approaches might not align for models based on non-experimental designs (e.g., data from natural observation) where there are many predictors, and a theory-driven model is likely to be simpler than the maximal model. In this situation, there is more chance that the maximal model would not converge and would need to be simplified anyway to fit the model (more on this in Section 21.4.3).

The other consideration is the number of levels of the contextual variable. In our example we have 20 clinics and each clinic is a level. Are 20 levels enough to estimate the parameters associated with the random effects? Thinking about what the parameters represent (variance from an overall effect), would you be confident about a variance based on 20 scores? You probably would, but what about three scores? Fifty scores? This is the simplistic answer to the question. You can *try* to estimate a random effect based on three levels, and maybe the model will converge, but there would be a lot of uncertainty around that estimate. With 20 levels you'd have more certainty around the estimate. Some have suggested that with fewer than five levels you may as well treat an effect as fixed (Bolker, 2015; Gelman & Hill, 2007).

There are other considerations, nicely summarized by Bolker (2015), who suggests treating a predictor as a random effect if:

- You're *not* interested in differences between the levels. In our example, we treat clinic as a random effect because we're not interested in comparing the differences in the effect of surgery between different clinics. If we wanted to compare the effect of surgery across clinics then we'd treat clinic as a fixed effect.
- You're interested in quantifying the variability across levels of the variable. In our example, we're interested in the variability in the effect of surgery across clinics so we treat clinic as a random effect.
- You're interested in generalizing beyond the observed levels of the contextual variable. If we want to draw conclusions only about the 20 clinics that we observed then it makes sense to treat clinic as a fixed effect. However, if we'd like our statements to generalize beyond the 20 clinics we observed we would treat it as a random effect. Related to this point, to treat a variable (like clinic) in this way the levels ought to be randomly sampled from and representative of a larger population. In the case of our clinics, they are probably not randomly sampled, but we might argue that they are representative of 'all' cosmetic surgery clinics in the same country.
- You have an unbalanced design. If each level of the variable contains different numbers of observations then treat it as a random effect. In our case, we have different numbers of participants within each clinic so, by this argument, we treat it as a random effect.
- You have a categorical predictor that is not directly relevant to the hypothesis but for which you need to adjust (a nuisance variable). In our example, if we don't care about variability in effects across clinics, we don't want to make general statements about clinics but we feel that we should adjust for the context in which surgery takes place, then we'd treat clinic as a random effect.

21.4.3 'Don't give me your problems': convergence and singular fits ©

I've mentioned 'model convergence' a few times. Models are fitted using iterative methods, which crudely means that the models are fitted using a series of successive approximations (iterations) with the hope that these approximations converge towards stable estimates of the model parameters. A model 'converges' at the point where the estimates from new iteration are very close to those from previous recent iterations. Exactly how close estimates need to be is determined by what's known as a tolerance value. To give an example, if your tolerance is 10 and you have a value of 48, then any value that is within 48 ± 10 (i.e., between 38 and 58) is within the tolerance; so, 57 is within the tolerance but 37 is not. Needless to say, how multilevel models use tolerances to decide on whether a model has converged is more complicated than this, but you get the general idea: if running more iterations is not meaningfully changing the parameter estimates then these estimates can be assumed to be 'stable' for the model. Some models don't converge. Despite tens of thousands of attempts (iterations) to find stable estimates the tolerance is always exceeded. Non-convergence is common, largely because the maths behind multilevel models is extremely complicated. It's best to be prepared for the fact that at some point you *will* get an error message about your model not converging. If that happens, the more straightforward things to try are:

- Check and simplify the model: check you have specified the model correctly and, if it is correct, simplify it (perhaps by removing some random effects).
- Centre and/or standardize continuous predictors (Section 11.3.3).

There are other, more technical, things to try but I'm not going to get into them.

Another problem is a singular fit, which is where the estimated variance–covariance matrix for the random effects is singular (i.e., has a determinant of 0). Put less mathematically, estimates of the variances of – and covariances between – random effects need to be constrained so that variances cannot be negative (because it makes no sense to have negative variance) and correlations cannot exceed ±1 (because values outside of these limits are impossible). Therefore, your estimated model might include a matrix for the random effects containing variances of 0 and correlations of 1 (Bates et al., 2018), not because these are reasonable estimates based on the data, but because of the constraints placed upon the matrix. A variance of 0 might well lead you to conclude that the associated random effect is zero, when in fact this estimate has arisen from trying to estimate more parameters than the data can support.

Common causes of singular fits are (1) having small numbers of random effect levels (i.e., three clinics, not 20); and (2) complex random effect models. To combat the first cause, we can change the random effect to be fixed (Section 21.4.2). In doing so we address a different question about the variable in question, but this might be a price worth paying. The second cause can arise from having lots of random effects (remember that as we add random effects the numbers of parameters in the model proliferate) or from categorical predictors with many levels that are treated as random (because n categories will be represented by $n - 1$ dummy variables and each will have a random intercept, slope and covariances with all the others). A solution is, therefore, to find ways to simplify the random effects within the model so that there are fewer parameters to estimate.

21.4.4 Centring predictors Ⓓ

Centring is the process of transforming a variable into deviations around a fixed point (Section 11.3.3). As we have seen, it can help with model convergence. One such fixed point is the mean of a variable (**grand mean centring**). This form of centring is used in multilevel models too, but sometimes **group mean centring** is used instead. Group mean centring occurs when for a given variable



we take each score and subtract from it the mean of the scores (for that variable) *within a given group*. For multilevel models, it is usually only level 1 predictors that are centred (in our cosmetic surgery example this would be predictors such as BDI and pre-surgery quality of life). If group mean centring is used then a level 1 variable is typically centred on means of a level 2 variable (in our cosmetic surgery data this would mean that, for example, the BDI of a person would be centred around the mean BDI for the clinic at which the person had their surgery).

In multilevel models centring can be a useful way to combat multicollinearity between predictor variables. It's also helpful when predictors do not have a meaningful zero point. Multilevel models with centred predictors tend to be more stable and estimates from these models can be treated as more or less independent of each other, which might be desirable. However, as with standard linear models (Section 11.3.3), centring affects the model. I'll try to sum up some excellent reviews (Enders & Tofighi, 2007; Kreft et al., 1995; Kreft & de Leeuw, 1998). Essentially, if you fit a multilevel model using the raw score predictors and then fit the same model but with grand mean centred predictors then the resulting models are equivalent. They will fit the data equally well, have the same predicted values, and the residuals will be the same. The parameters themselves (the *b*s) will, of course, be different, but there will be a direct relationship between the parameters from the two models. Therefore, grand mean centring doesn't change the multilevel model, but it would change your interpretation of the parameters (you can't interpret them as though they are raw scores). It can also help with models that won't converge.

Group mean centring is more complicated because the raw score model is not equivalent to the centred model in either the fixed part or the random part. This is because when you group mean centre a predictor you convert the raw scores to deviations from a group mean, therefore you're asking a different question to the one being asked by the raw score model. An effect in a raw score model, the effect of **days** in our example, combines a within and between effect of a predictor. The within effect operates at level 1; for our example it quantifies the difference in quality of life between individuals measured earlier or later than the average number of days at which people in the same clinic were measured. The between effect quantifies the difference in quality of life overall between clinics (which differ in the average number of days at which measures were taken). Grand mean centring eliminates the between effect (because the mean of the predictor becomes zero for all clinics), leaving only the within effect. There are many valid situations where you might want to isolate the within effect, so grand mean centring is useful, but just be aware that it changes what the effects represent (Bell et al., 2018).

People learning statistics often worry about there being a 'best' way to do things, but the 'best' method usually depends on what you're trying to do. Centring is a good example. Although some people make a decision about whether to use group or grand mean centring based on some statistical criterion, there is no statistically correct choice between not centring, grand mean centring and group mean centring (Kreft et al., 1995). Enders and Tofighi (2007) make four recommendations when analysing data with a two-level hierarchy: (1) group mean centring should be used if the primary interest is in an association between variables measured at level 1 (i.e., the aforementioned relationship between days and quality of life after surgery); (2) grand mean centring is appropriate when the primary interest is in the level 2 variable but you want to adjust for the level 1 covariate (i.e., you want to look at the effect of clinic on quality of life after surgery while adjusting for the time at which quality of life was measured); (3) both types of centring can be used to look at the differential influence of a variable at level 1 and 2 (i.e., is the effect of time on quality of life different at the clinic level than at the client level?); and (4) group mean centring is preferable for examining cross-level interactions (e.g., the interactive effect of clinic and time on quality of life). If group mean centring is used then the group means should be reintroduced as a level 2 variable unless you want to look at the effect of your 'group' or level 2 variable uncorrected for the mean effect of the centred level 1 predictor, such as when fitting a model when time is your main explanatory variable (Kreft & de Leeuw, 1998).



Oliver Twisted
Please, Sir, can I have some more ... group mean centring?

'Centring was so much fun when we did it in Chapter 11. It was like being gently tickled to sleep in a warm bath of octopuses. I want some more!' gurgles Oliver as he splashes around in his bath. Fair enough, Oliver, if you want to know how to do group mean centring using SPSS, then the material at www.discoverspss.com will tell you. Oh, and be careful with that toy pirate ship ...

21.4.5 Assessing fit and comparing multilevel models Ⓓ

The overall fit of a multilevel model is tested using a chi-square likelihood ratio test (see Section 19.3.4); we can obtain the deviance, which is minus twice the log-likelihood, $-2LL$ (see Section 20.3.2). Essentially, the smaller the value of the log-likelihood, the better. We can also obtain adjusted versions of the log-likelihood value, which can be interpreted in the same way as the log-likelihood:

- **Akaike's information criterion (AIC)**: This statistic is a goodness-of-fit measure that is corrected for model complexity. That means that it takes into account how many parameters have been estimated.
- *Schwarz's* **Bayesian information criterion (BIC)**: This statistic is comparable to the AIC, but it is slightly more conservative (it corrects more harshly for the number of parameters being estimated). It should be used when sample sizes are large and the number of parameters is small.

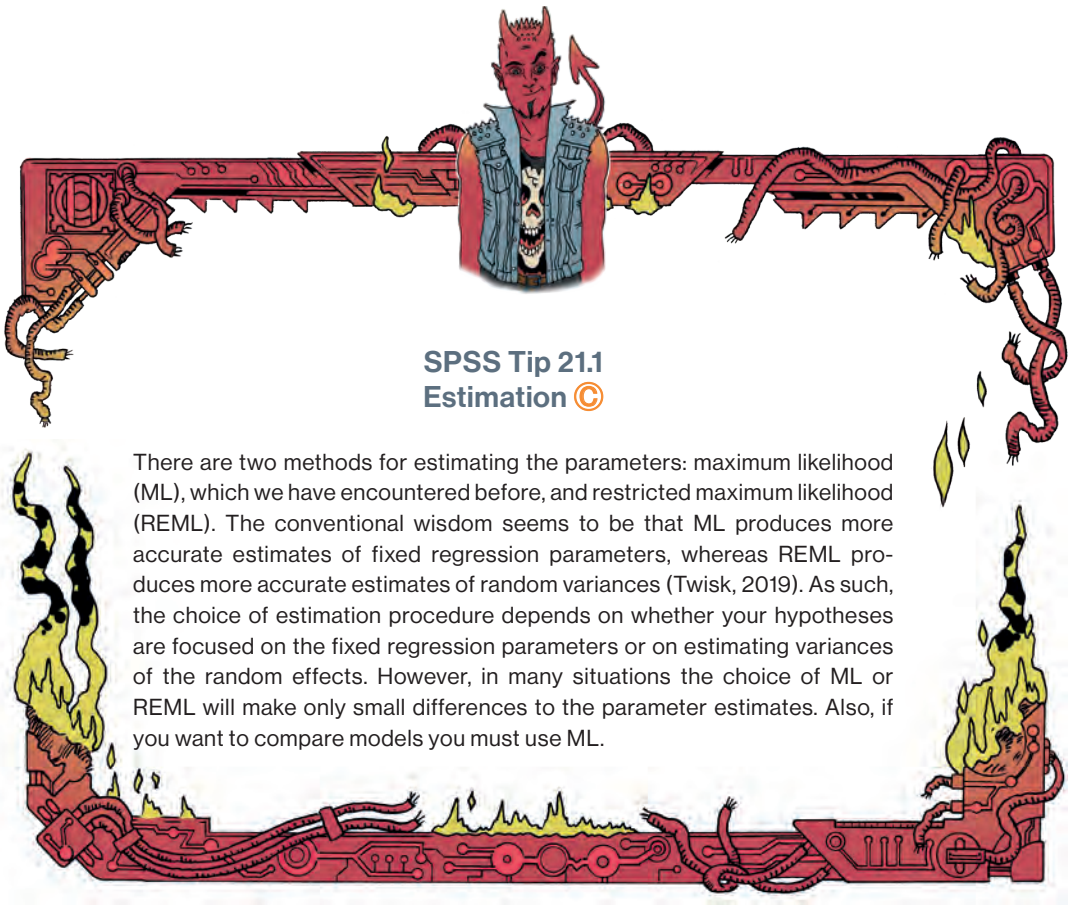
It's not meaningful to talk about values of the AIC and BIC as being large or small *per se*, but they can be compared across models that are built up hierarchically. Smaller values mean better-fitting models. If using data-driven approaches to deciding on whether effects are fixed or random, an option is to start with the maximal model and look at the extent to which removing a random effect worsens the fit. Alternatively (but do this cautiously), build up models starting with a 'basic' model in which all parameters are fixed and then add random coefficients as appropriate and explore confounding variables (Raudenbush & Bryk, 2002; Twisk, 2019). The higher-order model (the one with more parameters) can be compared to the lower-order model at each step using the AIC and BIC values, or using the change in the log-likelihood:

$$\chi^2_{change} = (-2LL_{higher}) - (-2LL_{lower})$$
$$df_{change} = K_{higher} - k_{lower} \quad (21.31)$$

in which k is the number of parameters in the respective models. This equation is basically the same as Eq. 19.37 and Eq. 20.21. There are two caveats to this approach: (1) it works only if full maximum likelihood estimation is used (and not restricted maximal likelihood – see SPSS Tip 21.1); and (2) the higher-order model must contain all the effects of the lower-order model. If you're going to use this method you should try to adjust the model one parameter at a time so that each test of the model focuses on the value (to the model) of a single parameter/effect. Likelihood ratio tests are OK in models based on normal errors in large samples, but for models based on normal errors (like in this chapter) tests based on Satterthwaite and Kenward–Roger degrees of freedom or bootstrapping are preferable.

21.4.6 Assumptions ©

Multilevel linear models are estimated using maximum likelihood rather than OLS, but the usual assumptions (see Chapter 6) apply. For example, we're still fitting a model that assumes linearity and additivity. The assumption of normality of errors in the standard linear model still applies in that we assume (in Eq. 21.6) that level 1 errors are normally distributed with mean zero and constant variance (i.e. homoscedasticity). When using maximum likelihood estimation, this is a formal assumption (remember in Chapter 6 we said this assumption wasn't too important for OLS estimation). Relatedly, we saw in Eq. 21.20 that any random effects (slopes and intercepts) are also assumed to be normally distributed with mean zero and constant variance (i.e. homoscedasticity).



SPSS Tip 21.1
Estimation ©

There are two methods for estimating the parameters: maximum likelihood (ML), which we have encountered before, and restricted maximum likelihood (REML). The conventional wisdom seems to be that ML produces more accurate estimates of fixed regression parameters, whereas REML produces more accurate estimates of random variances (Twisk, 2019). As such, the choice of estimation procedure depends on whether your hypotheses are focused on the fixed regression parameters or on estimating variances of the random effects. However, in many situations the choice of ML or REML will make only small differences to the parameter estimates. Also, if you want to compare models you must use ML.

In terms of the assumptions of independence and independent errors, we can factor in the correlations between cases caused by higher-level variables. As such, if a lack of independence is being caused by a level 2 or level 3 variable then a multilevel model should make this problem go away (although not always). Also in repeated-measures designs we need not restrict the covariance structure of the level 1 errors to being spherical (i.e., we don't need to assume sphericity) because we can model less restrictive structures. So multilevel models give us versatility to model dependence between scores, and, therefore, errors.

Finally, it's worth mentioning that multicollinearity can be a particular problem in multilevel models if you have interactions that cross levels in the data hierarchy (cross-level interactions). However, centring predictors can help matters (Kreft & de Leeuw, 1998).

21.4.7 Covariance structures Ⓓ

We have already discussed the structure of the variance–covariance matrices of the random effects (Section 21.3.5) and level 1 errors (Section 21.3.7). We saw that, by default, the random effects are modelled using an **unstructured** form. Expressed generally (Eq. 21.32), this structure means that each variable (random effect) has a unique variance and pairs of variables have unique covariances to be estimated from the data.

$$\begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,n} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,n} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 & \cdots & \sigma_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n} & \sigma_{2,n} & \sigma_{3,n} & \cdots & \sigma_n^2 \end{pmatrix} \tag{21.32}$$

We also saw that the level 1 residuals are modelled, by default, with what's known as a **variance components** structure in which variables are assumed to have the same constant variance (σ^2) and are independent and so have covariances of zero:

$$\begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{pmatrix} \tag{21.33}$$

We can model different structures to these defaults. For example, a **diagonal** structure, as in:

$$\begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \tag{21.34}$$

is like variance components in that random effects or errors are assumed to be independent (hence covariances of zero) but the variances are not constant. Instead, each random effect/error has a unique variance to be estimated from the data (σ_1^2, σ_2^2 … σ_n^2). We could use this structure in the level 1 errors to model heteroscedasticity. A different way to think of the diagonal structure is as the

unstructured variance–covariance matrix but with the constraint that variables are independent. Therefore, we might use this structure for the random effects if we do not want to estimate the covariances between them.

A final common covariance structure is a first-order autoregressive structure, or **AR(1)**:

$$\begin{pmatrix} \sigma^2 & \rho & \rho^2 & \rho^3 & \dots & \rho^{n-1} \\ \rho & \sigma^2 & \rho & \rho^2 & \dots & \rho^{n-2} \\ \rho^2 & \rho & \sigma^2 & \rho & \dots & \rho^{n-3} \\ \rho^3 & \rho^2 & \rho & \sigma^2 & \dots & \rho^{n-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \dots & \sigma^2 \end{pmatrix} \tag{21.35}$$

This structure is often used for the level 1 errors in repeated-measures data (especially when measurements are taken over time such as in growth models). Like variance components, this structure models constant variance (σ^2) but the relationships follow a distinctive pattern. Imagine the rows and columns of the matrix as points in time. What the matrix is saying is that the covariance between repeated measurements is highest at adjacent time points. So, in the first column, the correlation between time points 1 and 2 is ρ. For argument's sake, let's assume that $\rho = 0.3$. As we move to time point 3, the correlation between time points 1 and 3 is $\rho^2 = 0.09$. In other words, the covariance has decreased: scores at time point 1 are more related to scores at time 2 than they are to scores at time 3. At time 4, the covariance goes down again to $\rho^3 = 0.027$.

To sum up, the covariances between adjacent time points are assumed to be ρ, scores two intervals apart are assumed to have a weaker covariance of ρ^2, and scores three intervals apart are assumed to have a still weaker covariance of ρ^3 In short, the dependency between scores gets weaker over time. When this covariance structure is used the parameter ρ has to be estimated from the data.

$$\begin{pmatrix} \sigma^2 & \rho & \rho^2 & \rho^3 & \dots & \rho^{n-1} \\ \rho & \sigma^2 & \rho & \rho^2 & \dots & \rho^{n-2} \\ \rho^2 & \rho & \sigma^2 & \rho & \dots & \rho^{n-3} \\ \rho^3 & \rho^2 & \rho & \sigma^2 & \dots & \rho^{n-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \dots & \sigma^2 \end{pmatrix}$$

These are some of the more commonly used covariance structures, not an exhaustive list. For example, it's possible to have an AR(1) covariance structure in which variances are not constrained to be equal.

21.4.8 Sample size and power ©

As you might well imagine, the situation with power and sample size is very complex indeed. One complexity is that we are trying to make decisions about our power to detect both fixed and random effects coefficients. Kreft and de Leeuw (1998) do a tremendous job of making sense of things for us. Essentially, the message is the more data, the better. As more levels are introduced into the model, more parameters need to be estimated and the larger the sample sizes need to be. Kreft and de Leeuw conclude that if you are looking for cross-level interactions then you should aim to have more than 20 contexts (groups) in the higher-level variable, and that group sizes 'should not be too small'. They conclude that there are so many factors involved in multilevel analysis that it is impossible to produce any meaningful rules of thumb.

Twisk (2019) agrees that the number of contexts relative to individuals within those contexts is important. He also points out that standard sample size and power calculations can be used but then 'corrected' for the multilevel component of the analysis (by factoring, among other things, the intraclass correlation). However, he discusses two corrections that yield very different sample sizes. He recommends using sample size calculations with caution.

Having said that, a typical approach to power calculations in multilevel models is to simulate samples of data based on the expected parameters of your model and then, using these, to determine the proportion of samples in which the effect(s) of interest are detected (P. C. D. Johnson et al., 2015; Kain et al., 2015). By adjusting the sample size of the simulated samples it's possible to estimate the sample size necessary for the desired level of power. Good luck doing that in IBM SPSS Statistics software. If you use Windows investigate MLPowSIm, Optimal Design and, for two-level hierarchies, PinT. Also, various R packages exist such as *simr* (P. Green & MacLeod, 2016) and *simglm* (LeBeau, 2022).

21.5 Multilevel modelling using SPSS Ⓓ

Bolker et al. (2009) outline a general process for fitting multilevel models that I have tried to summarize in Figure 21.11. You begin by specifying the model either based on theory or using a maximal model (Section 21.4.2), which in turn affects whether you define effects as fixed or random and what effects you include in the model. Having defined the model, you'd perform some visual checks that linearity is reasonable, try to identify outlying cases or groups, and look for potential heteroscedasticity. For these initial checks you can work with the raw data rather than residuals. At this stage you might choose to address any potential problems; although robust methods exist to estimate multilevel models, they can fail, or take a prohibitively long time, to converge, so transformations might be the most reasonable solution. You'd then fit a fixed effect model to the data overall and to the individual levels of the contextual variable. In our case, this would mean fitting the model ignoring clinic and then fitting it to the individual clinics. You wouldn't necessarily report these models, but it ought to allow you spot contexts (clinics) where the model might be a poor fit, or just different. Essentially, it's a largely descriptive exercise to get a feel for the data.

Having done this, you'd fit the multilevel model whilst crossing your fingers and gently chanting incantations to the Statistical Spirits in the hope that they will grant your model convergence. It can help to make an offering of lemon drizzle cake, but best to avoid agreeing to them having your soul because no matter how much your soul seems a fair exchange for the joy of a converging model, in time you will regret it. Depending on the quality of your incantations/gifts, your model might converge. If so, interpret it and get out while the going is good. If not, try reducing the model complexity, centre predictors if you haven't already, get a new incantation or sob uncontrollably. All are appropriate responses. Then try refitting your new model. Rinse and repeat.

21.5.1 Entering the data Ⓑ

Data entry depends a bit on the study design: when the same variables are measured at several points in time, or you have several scores nested within an entity, you need to use the long format, but otherwise you can use the wide format (Section 4.6.1). The surgery example contains a single outcome score from each person (not several outcomes scores nested within each person) so we use the wide format shown in Figure 21.12. Each row represents a case of data (in this case a person who had surgery). Their scores on the various variables are entered in different columns. So, for example, the first person has identifier qx069, had a baseline quality of life score of 56, a post-surgery quality of life score of 71 measured 342 days after surgery, and had surgery for a physical reason at clinic 5.

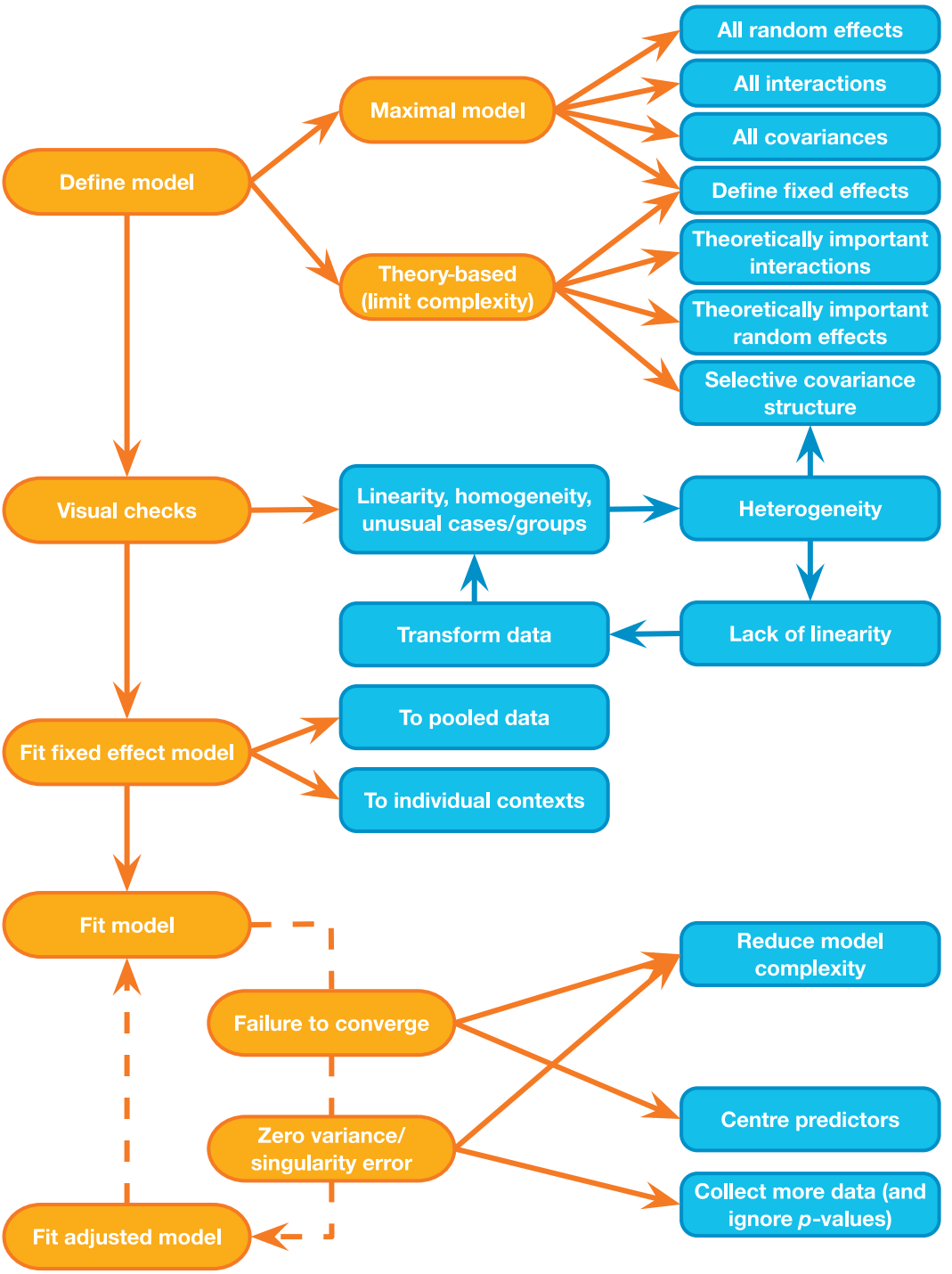


Figure 21.11 The process of fitting a multilevel model (based on Bolker et al., 2009)

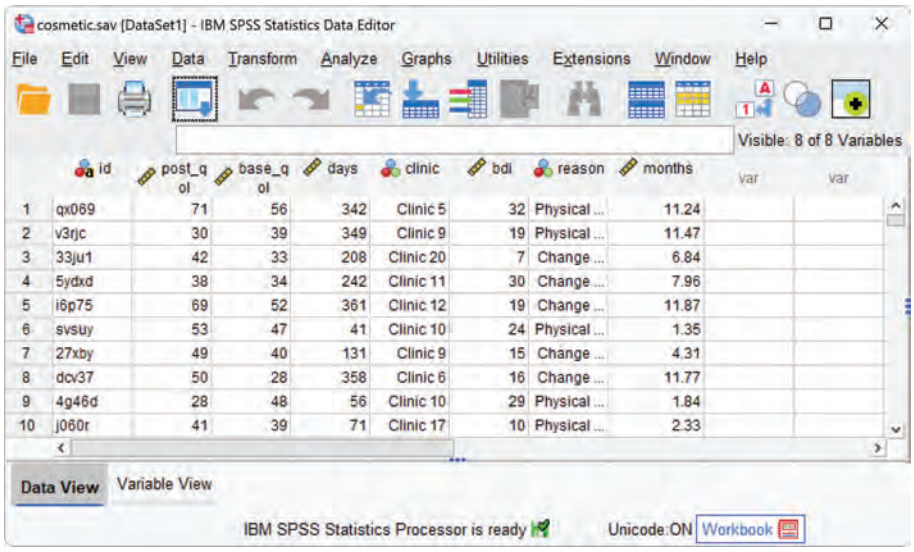


Figure 21.12 Data layout for multilevel modelling when outcome scores are not nested within cases

21.5.2 Defining the model Ⓒ

We have already discussed the model we plan to fit in Section 21.3.9 and defined it in Eq. 21.27. We are fitting a theory-based model that includes the random intercept and slope of **surgery** across clinics. We have fixed effects of the **reason** for surgery, baseline quality of life (**base_qol**) and the interaction between surgery and the reason for it.

21.5.3 Visualizing the data Ⓑ

The next phase is visual checks of the model. Our main example looks at **days** and baseline quality of life (**base_qol**) as predictors of quality of life after surgery (**post_qol**). Remember that surgery was conducted at one of 20 clinics. Ideally, we'd like to see the relationship between surgery and quality of life separately for the 20 clinics.



Produce a scatterplot of **days** (*x*-axis) against **post_qol** (*y*-axis), with each clinic plotted as a different colour. Include a line of best fit for each clinic.

The resulting plot is in Figure 21.13, and it looks ghastly. Nevertheless, the plot suggests variability in intercepts because overall levels of post-surgery quality of life are not the same across the 20 clinics. You can see this by looking at where the days since surgery is 0 and noting that at this point the lines for each clinic are at different heights. The plot also suggests variability in the relationship between time since surgery and quality of life; for example, quality of life increases over time in clinics such as 1, 5, 8 and 11, decreases over time in clinics such as 2, 3, 6 and 10 and stays constant in clinic 17. Based on the plot is seems reasonable to model days since surgery using a random slope and intercept.

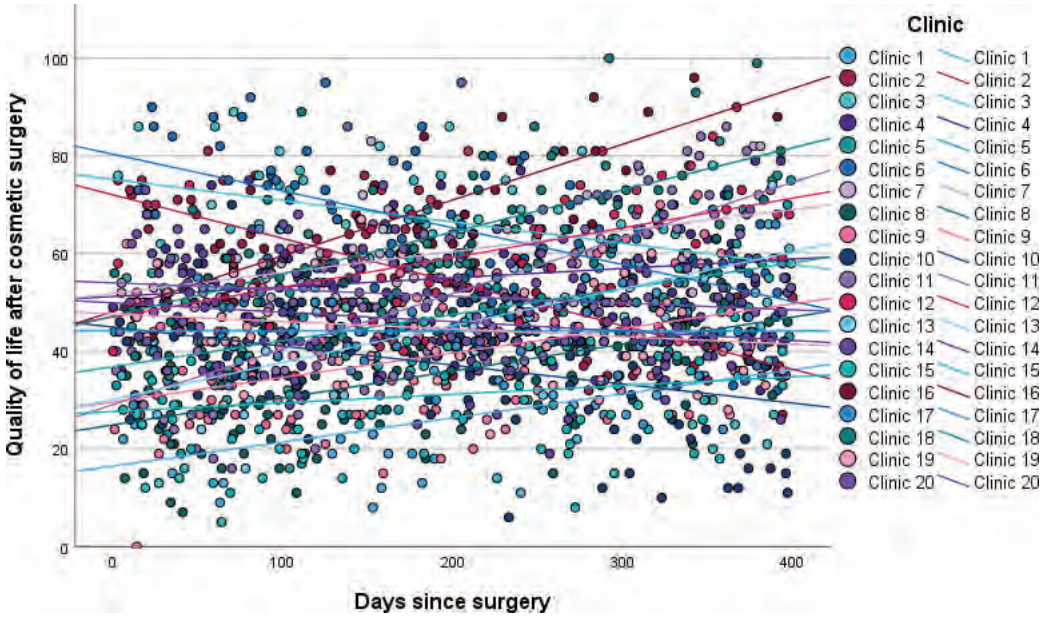


Figure 21.13 Plot of quality of life measured days post surgery at 20 clinics.

21.5.4 Fitting fixed effect models Ⓑ

The next phase is to fit the model without the random effects to get a feel for the fit across contexts. This model is expressed in Eq. 21.36. I have left the brackets for the structural and stochastic parts of the model to make the connection to the subsequent model that contains the random effects.

$$\text{QoL}_i = [\beta_0 + \beta_1\text{Surgery}_i + \beta_2\text{Pre QoL}_i + \beta_3\text{Reason}_i + \beta_4\text{Surgery} \times \text{Reason}_i] + [\varepsilon_i] \qquad (21.36)$$

Bolker et al. (2009) suggest fitting the model to the pooled data. Partly this is because of the complexity of multilevel models and the potential for unexpected consequences of your modelling decisions. Fitting a fully fixed effect model allows you to get a sense of what to expect from the multilevel model.

To access the multilevel model dialog boxes select *Analyze* > *Mixed Models* > Linear... to activate the dialog box in Figure 21.14. This dialog box is for specifying the hierarchical structure of the data, and we will use it to specify our level 2 variable (**clinic**). In general, we specify contextual variables that group participants (or subjects) by dragging them to the box labelled *Subjects*. In this example, we need to drag (or select and click ▶) **clinic** from the list of variables into this box as in Figure 21.14.

Click Continue to access the main dialog box in Figure 21.15. First, specify the outcome variable, which is quality of life after surgery, by dragging **post_qol** to the box labelled *Dependent Variable* (or select it and click ▶). Next, specify the predictors by selecting **base_qol**, **days** and **reason** and dragging them to the box labelled *Covariate(s)* (or click ▶). If you are wondering why we have dragged reason to this box rather than the one labelled *Factor(s)* when it is a categorical variable then read SPSS Tip 21.2.

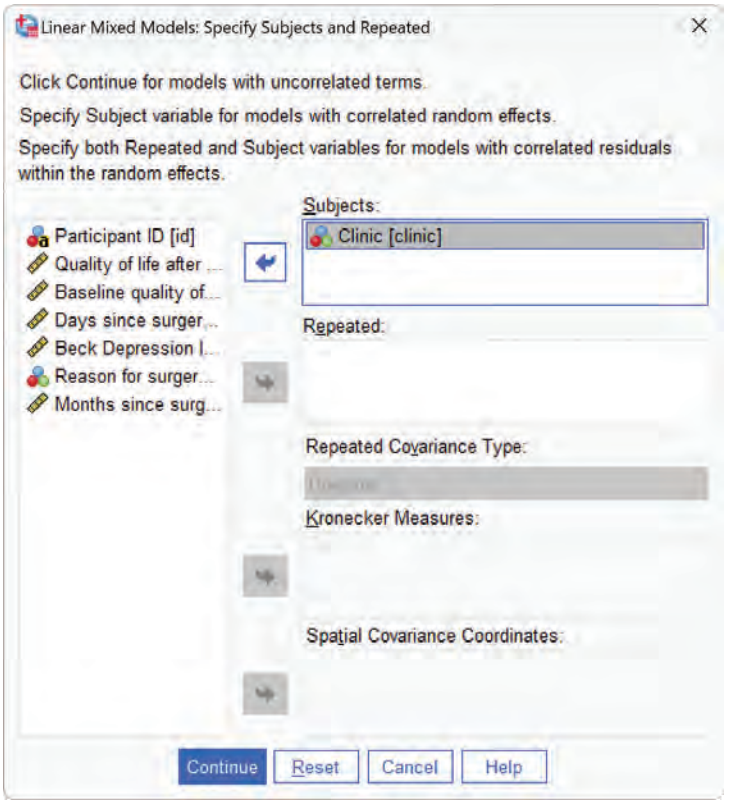


Figure 21.14 Specifying a level 2 variable in a hierarchical linear model

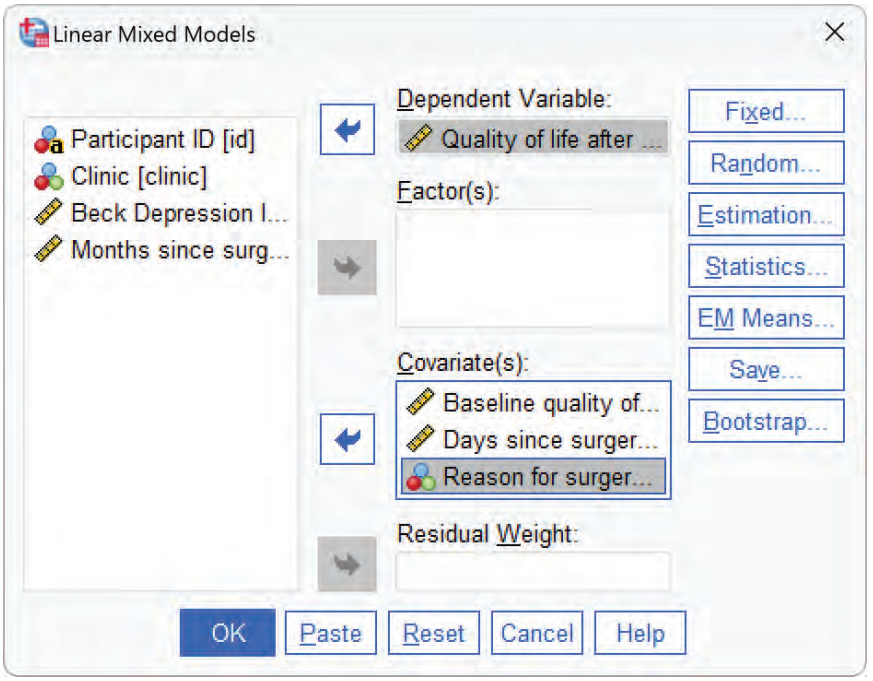
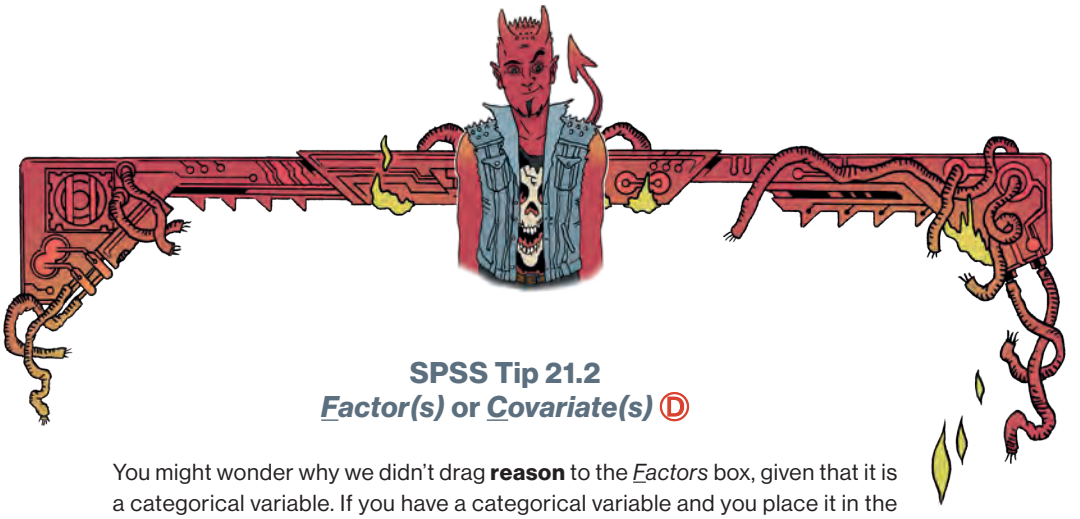
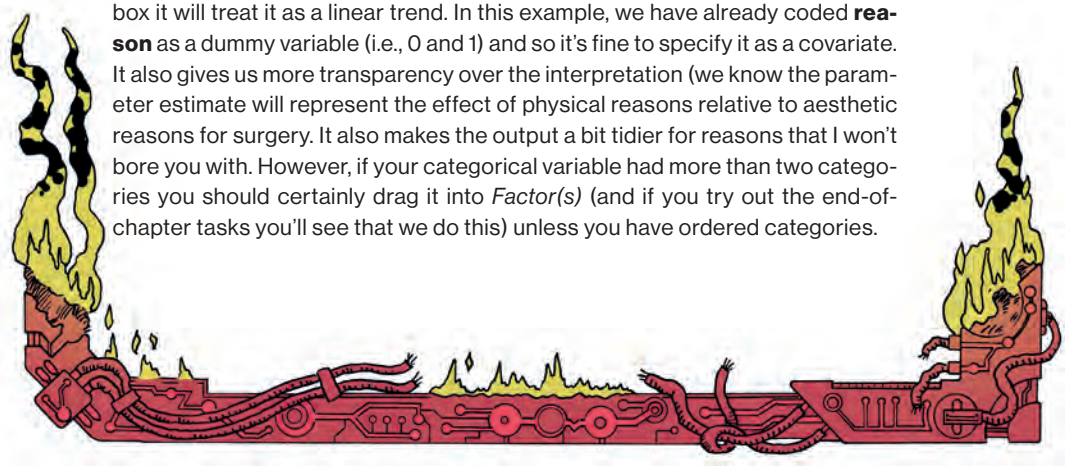


Figure 21.15 The main mixed models dialog box



SPSS Tip 21.2
Factor(s) or *Covariate(s)* Ⓓ

You might wonder why we didn't drag **reason** to the *Factors* box, given that it is a categorical variable. If you have a categorical variable and you place it in the *Factor(s)* box then SPSS will convert it into dummy variables for you and place these dummy variables into the model. It will also treat the final category as the baseline (in this case 'change appearance'). If you place it into the *Covariate(s)* box it will treat it as a linear trend. In this example, we have already coded **reason** as a dummy variable (i.e., 0 and 1) and so it's fine to specify it as a covariate. It also gives us more transparency over the interpretation (we know the parameter estimate will represent the effect of physical reasons relative to aesthetic reasons for surgery. It also makes the output a bit tidier for reasons that I won't bore you with. However, if your categorical variable had more than two categories you should certainly drag it into *Factor(s)* (and if you try out the end-of-chapter tasks you'll see that we do this) unless you have ordered categories.



Click on Fixed... in Figure 21.15 to access the dialog box for specifying fixed effects (Figure 21.16). We want to add fixed effects to the model, which means specifying **base_qol**, **days**, **reason** and **days × reason**. To do this, select **base_qol**, **days** and **reason** from the list labelled *Factors and Covariates*, then change Factorial to Main effects and click Add. The main effects of **base_qol**, **days** and **reason** should appear in the list labelled *Model*. To specify the interaction, change Factorial to Interaction. Next, select **days** from *Factors and Covariates* and, while holding down the *Ctrl* (*Cmd* on a Mac) key, select **reason**. With both variables selected, click Add to transfer them to *Model* as an interaction effect. The completed dialog box should look like Figure 21.16. Click Continue to return to the main dialog box.

Click Estimation... to open the dialog box in Figure 21.17 in which you can change the settings for the estimation process. For example, if you don't get a solution then you could increase the number of iterations (SPSS Tip 20.1). The defaults can be left alone, but you should decide whether to use the maximum likelihood, or restricted maximum likelihood estimation method. There are pros and cons

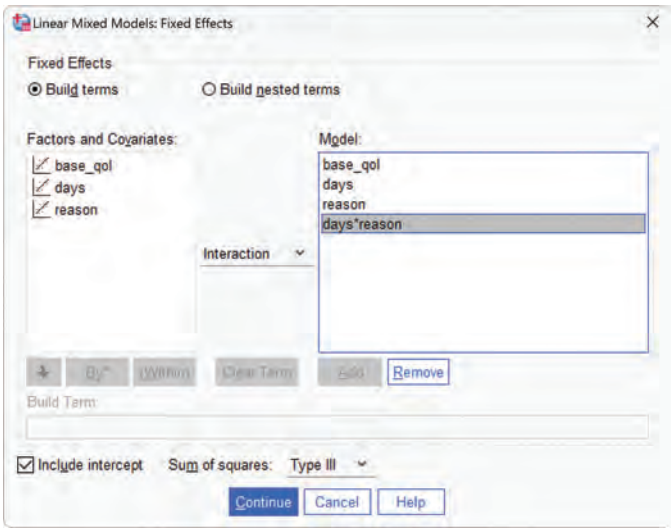


Figure 21.16 The dialog box for specifying fixed effects in mixed models

to both (see SPSS Tip 21.1). If we were intending to fit this model with only fixed effects and then compare it to a later model where we add one or more random effects then we would need to select ◉ Maximum Likelihood (ML). However, we are not going to do this, so we can leave the default of ◉ Restricted Maximum Likelihood (REML). You can also select a method for computing the degrees of freedom for significance tests (Section 21.4.1), but again the default of the Satterthwaite approximation is fine for our models. Click Continue to return to the main dialog box. To fit the fixed effect model we don't need to worry about any other options, so click OK to fit the model.

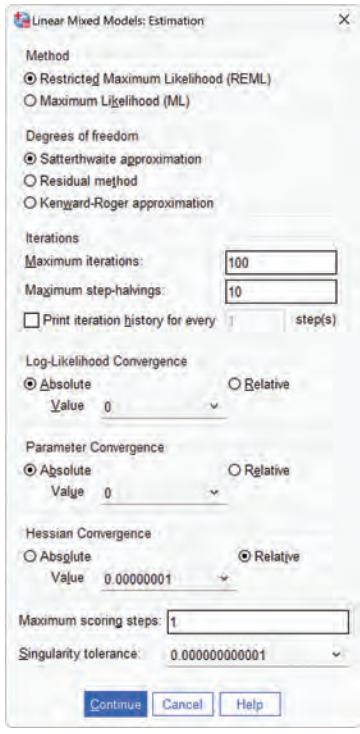


Figure 21.17 The *Estimation* options for mixed models

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	1571.000	144.732	<.001
base_qol	1	1571	132.695	<.001
days	1	1571	4.276	.039
reason	1	1571.000	2.353	.125
days * reason	1	1571.000	10.271	.001

a. Dependent Variable: Quality of life after cosmetic surgery.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	24.755	2.058	1571.000	12.030	<.001	20.719	28.791
base_qol	.459	.040	1571	11.519	<.001	.380	.537
days	.009	.004	1571	2.068	.039	.000	.017
reason	-2.479	1.616	1571.000	-1.534	.125	-5.649	.691
days * reason	.023	.007	1571.000	3.205	.001	.009	.037

a. Dependent Variable: Quality of life after cosmetic surgery.

Output 21.1

Output 21.1 shows the overall tests and parameter estimates for the pooled fixed effects model. As you'd expect with such a large sample size, the variables **days**, **base_qol** and the **days × reason** interaction are significant predictors of quality of life after surgery. We're interested in the parameter estimates to guide our expectations about what to expect in the more complicated model that includes the random effects. For example, the overall effect of **days** appears to be that for every extra day since surgery, quality of life increases by $\hat{\beta} = 0.009$ units. This effect seems tiny, but how much would you expect quality of life to change in only 1 day? It's helpful to think about what this effect equates to over, say, 4 months (~120 days) or a year (365 days). This effect equates to about a unit change in quality of life over 4 months and about 3 units over a year.[7] In the context of a 100-point scale, this is not a huge change. Similarly, the parameter estimate for the interaction is tiny, $\hat{\beta} = 0.023$, but this also reflects the fact that the variable **days** has a huge range so the change we can expect from a single day is tiny (but might add up over time).

21.5.5 Scaling variables Ⓑ

The discussion about **days** highlights a point about scaling variables. In this data set, our outcome variable (**post_qol**) and one of the predictors (**pre_qol**) can range from 0 to 100, the **reason** for surgery can range from 0 (change appearance) to 1 (physical reason), but the predictor **days** varies from 0 to 400. Our model has variables measured on very different scales. We have just seen how the wide scale of **days** makes it difficult to get an intuition about the size of the effects involving this variable. When we see small numbers we tend to assume small effects. Also, although our model has converged, sometimes they don't when variables are measured on very different scales. The solution in all cases is to scale variables so that they are broadly similar.

One way to rescale variables is to use standardized scores instead of raw ones. Standardizing scores transforms them into standard deviation units; it rescales the raw scores such that they have mean 0 and a maximum range likely to be within ±4. If we standardize all of the continuous variables in the model they'd be expressed in the same units (standard deviations) and have a similar range. A downside to standardization is (in my opinion) that parameter estimates are less intuitive to interpret because variables are expressed in standard deviation units. The people to whom you might need to

7 $120 \times 0.009 = 1.08$; $365 \times 0.009 = 3.285$

communicate your results probably don't have an intuition of what it means that a '1 standard deviation change in **days** results in an *x* standard deviation change in quality of life'. I certainly don't.

Another solution is to change the metric of particularly problematic variables. For example, expressing time since surgery in days results in a very wide range of values. We have expressed the time since surgery in days, but this choice is arbitrary. We could have chosen hours, minutes, seconds, months or years, all of which are re-expressions of the same information. If we were to express time in hours rather than days the range of 0 to 400 days becomes 0 to 9600 hours, so this would make things worse because the range increases. However, if we expressed the time since surgery in months rather than days the range becomes 0 to 13 (approx.), which might help because the range has decreased and is more in line with the other variables in the model. The data set contains a variable **months**, which is the variable **days** approximately transformed to months.[8]



Assuming you're running this model after the previous one, some of the dialog boxes will be partially set up already. Select *Analyze* > *Mixed Models* > Linear... and leave the first dialog box like we set it up for the previous model in Figure 21.14. Click Continue and select **days** in the box labelled *Covariate(s)* and drag it back to the variable list (or select it and click). Next, select **months** and drag it to the box labelled *Covariate(s)* (or click).The main dialog box should look like Figure 21.18.

Click on Fixed... to access the dialog box for specifying fixed effects. The effects involving **days** will have been removed but the main effects of **base_qol** and **reason** should still be listed. To add the effect of **months** and the **months** × **reason** interaction select Factorial, then select **months** and **reason** simultaneously from *Factors and Covariates* (by holding down the *Ctrl*/*Cmd* on a Windows/Mac) and click Add. By selecting select Factorial both the main effects and interaction will be added. The completed dialog box should look like Figure 21.19. Click Continue to return to the main dialog box and OK to fit the model.

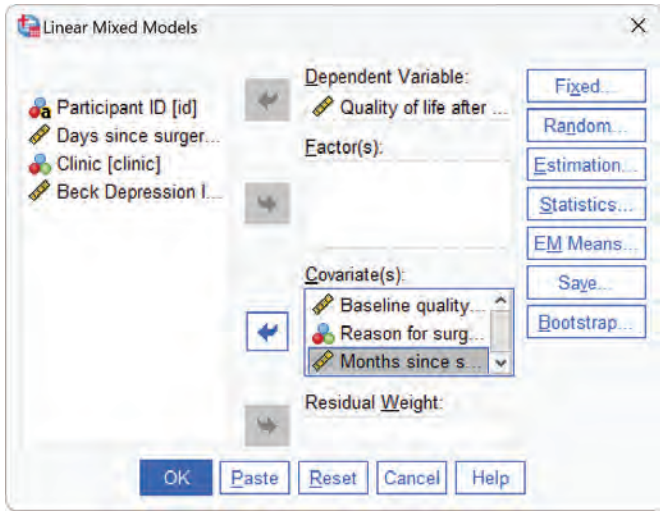


Figure 21.18 The main mixed models dialog box

8 The months of the year have different numbers of days, but there are 365 days in most years, and 12 months, so a good enough approximation is Months since surgery = Days since surgery × 12/365

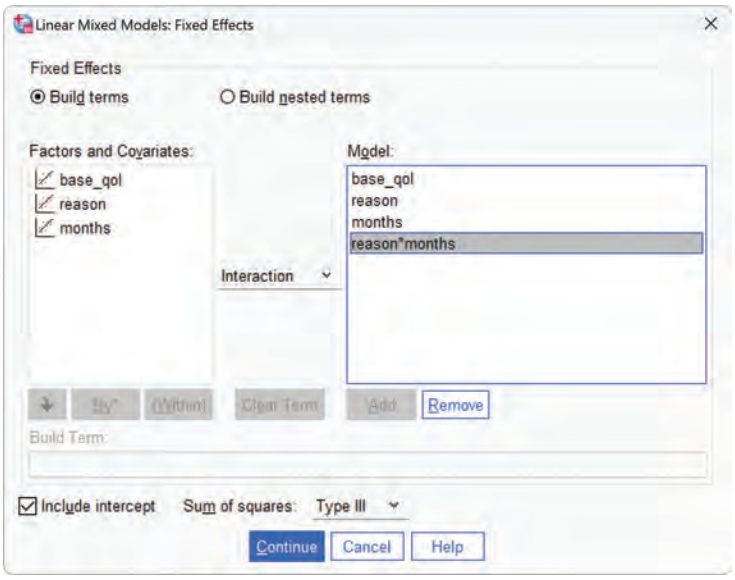


Figure 21.19 The dialog box for specifying fixed effects in mixed models

Output 21.2 shows the overall tests and parameter estimates when we use **months** instead of **days** in the model. Compare this output with Output 21.1 and note that everything is identical *except* for the parameter estimates for the effect of time since treatment. Of course, this is what you'd expect because the variable **months** is a redescription of **days** so significance tests will be identical using the two variables, but because the $\hat{\beta}$s are expressed in the units measured, they have changed. So instead of concluding that for every extra day since surgery, quality of life increases by $\hat{\beta} = 0.009$ units, we conclude that for every extra *month* since surgery, quality of life increases by $\hat{\beta} = 0.271$ units. To my mind, it is easier to grasp how small the effect of time is using months: a 0.271 change on a 100-point scale over a period of a month is tiny. There's no correct thing to do, but because model convergence *can* be affected by the scales of measurement of the variables, it's useful to explore this in the chapter.

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	1571	144.732	<.001
base_qol	1	1571	132.695	<.001
months	1	1571	4.276	.039
reason	1	1571.000	2.353	.125
months * reason	1	1571.000	10.271	.001

a. Dependent Variable: Quality of life after cosmetic surgery.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	24.755	2.058	1571	12.030	<.001	20.719	28.791
base_qol	.459	.040	1571	11.519	<.001	.380	.537
months	.271	.131	1571	2.068	.039	.014	.528
reason	-2.479	1.616	1571.000	-1.534	.125	-5.649	.691
months * reason	.690	.215	1571.000	3.205	.001	.268	1.112

a. Dependent Variable: Quality of life after cosmetic surgery.

Output 21.2

21.5.6 Adding random effects to the model Ⓑ

Let's stick with **months** as our unit of time but add in the random intercept and slope. Again select *Analyze* > *Mixed Models* > Linear..., leave the first dialog box set up as in Figure 21.14, leave the main dialog box as it is in Figure 21.18, and make sure the fixed effects are set up as in Figure 21.19. To add random effects click Random... to access the dialog box in Figure 21.20. First, we specify our contextual variable by selecting it (**clinic**) from the section labelled *Subjects* (which will contain any variables that we specified in Figure 21.14) and dragging it (or clicking ➔) to the area labelled *Combinations*. To specify that the intercept is random select ☑ Include intercept (Figure 21.20). To specify a random slope for a variable, select the variable (in this case **months**) from the list of *Factors and Covariates* and add it to the *Model* by clicking Add (Figure 21.20). There is a drop-down list to specify the type of covariance structure (Variance Components ×). A decent default option is to use an unstructured covariance structure. SPSS offers two variants: the standard one will express the relationship between slopes and intercepts as the raw covariance, whereas Unstructured: Correlation Metric will express it as a correlation. Given that most people find correlations easier to interpret than covariance, I suggest choosing this option. When everything looks like Figure 21.20 click Continue to get back to the main dialog box.

Click Statistics... to open the dialog box in Figure 21.21. There are two useful options in this dialog box. The first is ☑ Parameter estimates for fixed effects, which will give us *b*-values for each effect and their significance. The second is ☑ Tests for covariance parameters, which will give us a significance test of each of the covariance estimates in the model (i.e., the estimates of $\sigma^2_{u_0}$, $\sigma^2_{u_1}$ and $\sigma^2_{u_0,u_1}$ in Eq. 21.20). These estimates tell us about the variability of intercepts or slopes across our contextual variable and so significance-testing them can be useful (we can then say that there was significant, or non-significant, variability in intercepts or slopes). Select these two options and click Continue to return to the main dialog box and OK to fit the model.

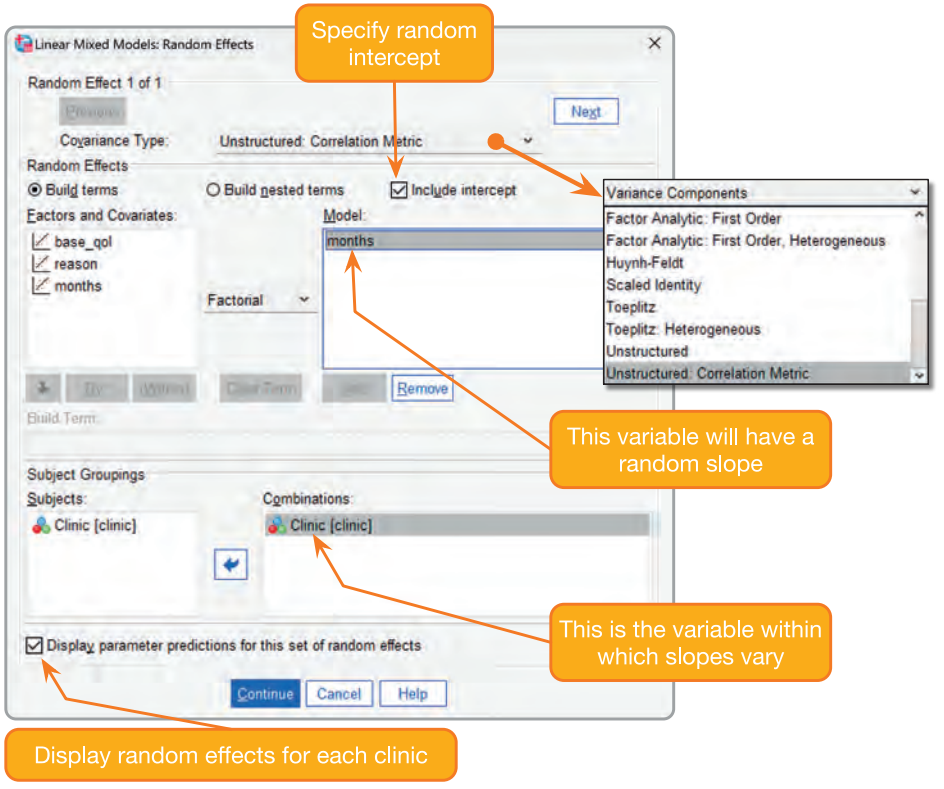


Figure 21.20 The dialog box for specifying random effects in mixed models

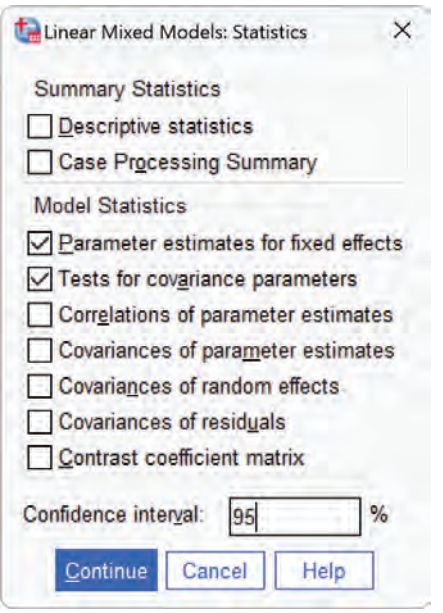


Figure 21.21 The *Statistics* dialog box

21.5.7 Interpreting the model ©

Output 21.3 shows the resulting table of fixed effects. By default Type III sums of squares are used and *F*-tests are based on Satterthwaite's approximation (see Section 21.4.1). You can change this in Figure 21.17 should you need to. To sum up, there are significant effects of baseline quality of life, $F(1, 1534.92) = 382.66$, $p < 0.001$, and the months × reason interaction, $F(1, 1535.12) = 11.84$, $p = 0.001$, but not the overall effect of months, $F(1, 19.62) = 1.48$, $p = 0.239$, or the main effect of reason, $F(1, 1535.47) = 3.36$, $p = 0.067$. We'll return to these effects, but note that the interaction is significant.

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	22.492	38.986	<.001
base_qol	1	1534.926	382.665	<.001
reason	1	1535.470	3.362	.067
months	1	19.627	1.477	.239
reason * months	1	1535.124	11.841	<.001

a. Dependent Variable: Quality of life after cosmetic surgery.

Output 21.3

Output 21.4 shows the resulting table of parameter estimates. The overall effect of time on quality of life was small and non-significant, $\hat{\beta} = 0.48$ [–0.35, 1.31], $t(19.63) = 1.22$, $p = 0.239$. The effect of the reason for surgery was also small and non-significant, $\hat{\beta} = -1.79$ [–3.71, 0.13], $t(1535.47) = -1.83$, $p = 0.067$. To put this effect in context, with other variables held constant the difference in quality of life in those seeking treatment for cosmetic reasons rather than physical treasons was 1.79 lower (on the 100-point scale). The effect of baseline quality of life on post-surgery quality of life was more substantial and significant, $\hat{\beta} = 0.47$ [0.42, 0.52], $t(1534.93) = 19.56$, $p < 0.001$. This effect suggests that with other variables held constant, for every unit increase in baseline quality of life there is a half unit

increase in post-surgery quality of life, which is a fairly strong relationship (because both are measured on the same scale). The combined effect of months and reason (their interaction) on post-surgery quality of life was significant, $\hat{\beta} = 0.45$ [0.19, 0.70], $t(1535.12) = 3.44$, $p = 0.001$. The parameter estimate suggests that the gradient (the change over time) of quality of life is 0.45 bigger in those having surgery for physical reasons than in those having it for cosmetic reasons.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	25.069	4.015	22.492	6.244	<.001	16.753	33.385
base_qol	.470	.024	1534.926	19.562	<.001	.423	.517
reason	−1.792	.977	1535.470	−1.834	.067	−3.709	.125
months	.482	.397	19.627	1.215	.239	−.346	1.310
reason * months	.447	.130	1535.124	3.441	<.001	.192	.702

a. Dependent Variable: Quality of life after cosmetic surgery.

Output 21.4

Output 21.5 shows the estimates of $\sigma^2_{u_0}$, $\sigma^2_{u_1}$ and σ_{u_0,u_1} in Eq. 21.20. The estimated variance of the intercepts is $\hat{\sigma}^2_{u_0} = 290.49$, which is significant, $z = 3.02$, $p = 0.003$, suggesting significant variation in intercepts across clinics. If you find standard deviations easier to think about then take the square root of the variance (in this case the standard deviation of intercepts is $\hat{\sigma}_{u_0} = \sqrt{290.49} \cong 17.04$). The estimated variation in slopes is $\hat{\sigma}^2_{u_{\text{months}}} = 2.99$ (or as a standard deviation $\hat{\sigma}_{u_{\text{months}}} \sqrt{2.99} \cong 1.73$), which is significant, $z = 2.98$, $p = 0.003$, suggesting significant variation in the effect of time on post-surgery quality of life across clinics. The correlation between slopes and intercepts is estimated to be $\hat{r}_{u_0, u_{\text{months}}} = -0.71$. To sum up, there was non-zero variability in intercepts and slopes, and they were very strongly and negatively related; that is, clinics with large intercepts tended to have smaller slopes. You can see this relationship in Figure 21.13 where clinics with small intercepts (where the line starts lower on the *y*-axis) tend to have positive slopes and clinics with large intercepts (where the line starts higher on the *y*-axis) tend to have negative (i.e. smaller) slopes.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Residual		85.677	3.095	27.684	<.001	79.821	91.962
Intercept + months [subject = clinic]	Var(1)	290.488	96.120	3.022	.003	151.870	555.627
	Var(2)	2.990	1.004	2.979	.003	1.549	5.773
	Corr(2,1)	−.705	.118	−5.955	<.001	−.871	−.393

a. Dependent Variable: Quality of life after cosmetic surgery.

Output 21.5

The variability in slopes and intercepts for **months** is shown in Output 21.6, which breaks down the estimated difference within each clinic from the overall slope and intercept. These are the values of $\hat{u}_{0,j}$ and $\hat{u}_{1,j}$ illustrated in Figure 21.8 and Figure 21.9 (the figures used the variable **days** and so won't match the values in Output 21.6). For example, in clinic 5, $\hat{u}_{0,j} = -7.68$ suggesting that the intercept in clinic 5 was seven units lower than the overall intercept, and $\hat{u}_{1,j} = 0.85$ suggesting that the slope of months was 0.85 greater than the overall slope. Scanning across the clinics you can see that some intercepts and slopes are greater than the overall values, and others smaller, and in some clinics the effect of months is significant, whereas in others it is not. All of which consolidates our conclusion that the effect of time since surgery on quality of life (slope), and overall post-surgery quality of life (intercept) vary across clinics.

We can tease this interaction apart like we did when we looked at moderation effects by looking at simple slopes; that is, estimate the slope of months separately for physical and cosmetic reasons for surgery. In SPSS this entails splitting the file by **reason** and the fitting the model but without **reason** (or the interaction term) included.

Empirical Best Linear Unbiased Predictions[a]

Clinic	Parameter	Prediction	Std. Error	df	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Clinic 5	Intercept	-7.678	4.117	24.792	.074	-16.161	.805
	months	.845	.439	29.277	.064	-.053	1.743
Clinic 9	Intercept	-15.935	4.181	26.336	.001	-24.525	-7.346
	months	.832	.452	32.731	.075	-.089	1.753
Clinic 20	Intercept	7.729	4.316	29.749	.083	-1.087	16.546
	months	-.997	.465	36.360	.039	-1.940	-.053
Clinic 11	Intercept	-17.788	4.236	27.687	.000	-26.469	-9.107
	months	3.006	.465	36.407	1.621E-7	2.063	3.950
Clinic 12	Intercept	.203	4.555	36.480	.965	-9.031	9.436
	months	1.396	.501	47.610	.008	.387	2.404
Clinic 10	Intercept	-.687	4.199	26.775	.871	-9.307	7.932
	months	-1.617	.449	31.772	.001	-2.531	-.702
Clinic 6	Intercept	34.220	4.292	29.126	8.306E-9	25.444	42.996
	months	-2.873	.460	34.892	3.742E-7	-3.807	-1.939
Clinic 17	Intercept	-.691	4.527	35.658	.880	-9.875	8.494
	months	-.710	.507	49.532	.168	-1.729	.309
Clinic 8	Intercept	-22.059	4.117	24.795	1.520E-5	-30.542	-13.576
	months	1.249	.439	29.242	.008	.352	2.147
Clinic 3	Intercept	28.240	4.512	35.220	3.440E-7	19.082	37.398
	months	-1.777	.503	48.031	.001	-2.788	-.766
Clinic 4	Intercept	5.888	4.289	29.048	.180	-2.883	14.659
	months	-.295	.473	38.624	.536	-1.252	.662
Clinic 13	Intercept	-13.934	4.910	47.850	.007	-23.808	-4.060
	months	1.471	.565	70.312	.011	.344	2.599
Clinic 18	Intercept	2.540	4.395	31.899	.567	-6.414	11.494
	months	1.855	.494	45.040	.000	.861	2.849
Clinic 15	Intercept	-18.432	4.178	26.250	.000	-27.016	-9.849
	months	-.049	.447	31.392	.913	-.961	.863
Clinic 7	Intercept	8.355	4.663	39.781	.081	-1.072	17.781
	months	.113	.547	63.435	.837	-.979	1.205
Clinic 19	Intercept	1.223	4.336	30.287	.780	-7.628	10.074
	months	-.891	.474	38.865	.068	-1.850	.067
Clinic 1	Intercept	-28.666	5.055	52.862	6.065E-7	-38.806	-18.526
	months	.832	.614	87.933	.178	-.387	2.052
Clinic 14	Intercept	4.320	4.159	25.787	.309	-4.232	12.872
	months	-1.182	.441	29.726	.012	-2.083	-.280
Clinic 16	Intercept	5.727	5.152	56.336	.271	-4.593	16.047
	months	2.234	.625	91.956	.001	.993	3.476
Clinic 2	Intercept	27.425	4.365	31.065	5.442E-7	18.524	36.326
	months	-3.443	.482	41.330	9.837E-9	-4.416	-2.470

a. Dependent Variable: Quality of life after cosmetic surgery

Output 21.6



Split the file by **reason** and then run a multilevel model predicting **post_qol** with a random intercept, and random slopes for **months**, and including **base_qol** and **months** as predictors.

Output 21.7 shows the effect of months for each reason for surgery. For those who had surgery to change their appearance (top table), their quality of life increased over time but not significantly so, $\hat{\beta} = 0.51$ [–0.29, 1.31], $t(19.13) = 1.32$, $p = 0.200$. For every additional month since surgery, quality of life increased by 0.51 units (on the 100-point scale). Over a year, this amounts to a change of 12 × 0.51 = 6.12 units. In contrast, for those who had surgery to help with a physical problem, their quality of life increased significantly over time, $\hat{\beta} = 0.92$ [0.03, 1.81], $t(18.75) = 2.16$, $p = 0.044$. For every additional month since surgery, quality of life increased by 0.92 units (on the 100-point scale). Over a year, this amounts to a change of 12 × 0.92 = 11.04 units. The change in quality of life over the same time period for people who had surgery for a physical reason is about double that of people who had it for a cosmetic reason. Notice that if we subtract the two parameter estimates we get (approximately) the



Cramming Sam's Tips
Fitting multilevel models

- Plot the substantive variables (ones for which you expect effects to vary across groups) in the model, splitting by the grouping variable across which you expect effects to vary (in this case clinics). Use these plots to get a feel for whether the effects (intercepts and slopes) do vary across contexts.
- Fit the model to the complete data (ignoring the context variable). Note the model parameters for the overall model.
- Fit the full model including random effects. Note warnings and failures to converge. Try to address these by looking at rescaling variables or simplifying the model (removing some random effects).
- Interpret overall *F*-statistics for the fixed effects in the table of *Type III Tests of Fixed Effects*: look in the column labelled *Sig.* If the value is less than or equal to your criterion (typically 0.05) then the effect is significant. Interpret these *p*-values cautiously.
- Inspect the random effects in the table of *Estimates of Covariance Parameters*. Are these values meaningfully non-zero? What's the relationship between slopes and intercepts? The significance of these estimates should be treated cautiously.
- Interpret the parameter estimates ($\hat{\beta}$s) in the table of *Estimates of Fixed Effects*. Interpret the size and direction of the parameter estimates and their confidence intervals and cautiously interpret *p*-values. Try to communicate to your results in terms of the amount of change in the outcome that you get from a unit change in each predictor – is this change a meaningful change in the outcome?



parameter estimate for the interaction term (in Output 21.4), which reminds us that the interaction term represents the difference in the effect of months on quality of life in the two groups:

$$\begin{aligned}\hat{\beta}_{\text{months} \times \text{reason}} &\cong \hat{\beta}_{\text{months(physical)}} - \hat{\beta}_{\text{months(appearance)}} \\ &\cong 0.917 - 0.507 \\ &\cong 0.41\end{aligned} \tag{21.37}$$

Estimates of Fixed Effects[a,b]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	24.909	4.020	23.937	6.196	<.001	16.611	33.207
base_qol	.473	.031	936.504	15.347	<.001	.413	.534
months	.507	.382	19.126	1.328	.200	-.292	1.305

a. Reason for surgery = Change appearance

b. Dependent Variable: Quality of life after cosmetic surgery.

Estimates of Fixed Effects[a,b]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	23.380	4.473	26.880	5.228	<.001	14.202	32.559
base_qol	.468	.039	562.660	11.931	<.001	.391	.545
months	.917	.424	18.745	2.160	.044	.028	1.806

a. Reason for surgery = Physical reason

b. Dependent Variable: Quality of life after cosmetic surgery.

Output 21.7



Oditi's Lantern
Multilevel models

'I, Oditi, believe that you know that experimental manipulations happen within cults, erm, I mean contexts; and people within cults, erm, contexts, become more similar to each other than to people outside of that cult, erm, context. To eliminate this dependency we must make everyone join our cult, erm, no, seriously, I mean we must factor the dependency by using a multilevel model. Stare into my lantern and you will become worthy to call yourself one of my cult of undiscovered numerical truths.'





Labcoat Leni's Real Research 21.1
Please release me ©

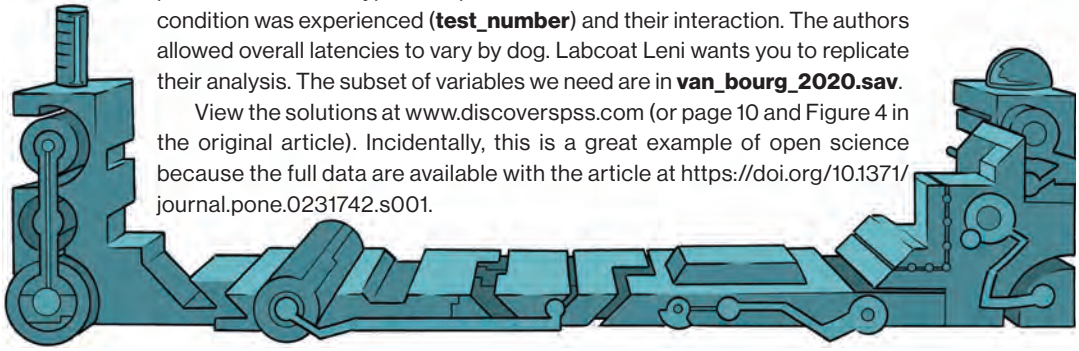
Van Bourg, J., Patterson, J. E., & Wynne, C. D. L. (2020). *PLoS One*, *15*(4), e0231742. https://doi.org/10.1371/journal.pone.0231742

Pet dogs often engage in behaviours helpful to their owners (mine likes to cuddle me when I've had a bad day, and in fact when I've had a good day, and now I think of it, pretty much any day regardless of how good or bad it has been). It's unclear whether these behaviours are truly prosocial. Can a dog engage in prosocial behaviours that haven't been explicitly trained? Van Bourg et al. (2020) addressed this question by trapping some dog's owners in boxes! In the study 60 dogs were tested in three conditions all of which involved being in a room with large restrainer box (a large acrylic box with holes in the side that could be closed by resting a foam board door across its opening). Each dog had three experiences in the room and each time the experimenters were interested in whether the dog would open the restrainer box within 120 seconds. The order of the three experiences was counterbalanced so different dogs completed the experiences in different orders

- The **food** condition: food was dropped into the restrainer. This condition was to test whether the dog was capable of moving the foam board door to open the box (to get the food).
- The **distress** condition: the dogs' owner was placed in the restrainer and was instructed to call for help in a distressed tone.
- The **reading** condition: the dogs' owner was placed in the restrainer and was instructed to read from a magazine at the same pace and in the same tone as in the distress condition.

The researchers did several analyses, but we will focus on the time for the dog to open the restrainer up to a maximum of 120 s (**latency**). Create a variable **ln_latency** that is the log of the latencies, filter out dogs that did not open the compartment in the food condition, then filter out data from the food condition. Fit a model in which the natural log of time taken to open the box is predicted from the type of experience (**condition**), the order in which the condition was experienced (**test_number**) and their interaction. The authors allowed overall latencies to vary by dog. Labcoat Leni wants you to replicate their analysis. The subset of variables we need are in **van_bourg_2020.sav**.

View the solutions at www.discoverspss.com (or page 10 and Figure 4 in the original article). Incidentally, this is a great example of open science because the full data are available with the article at https://doi.org/10.1371/journal.pone.0231742.s001.



21.6 How to report a multilevel model ©

Multilevel models take on so many forms that giving standard advice is not straightforward. For any model you need to say something about the random effects. For the final model of the cosmetic surgery example you could write something like:

✓ There was non-zero variability in intercepts and slopes. The estimated variance of the intercepts is $\hat{\sigma}^2_{u_0} = 290.49$, which is significant, $z = 3.02$, $p = 0.003$, suggesting significant variation in intercepts across clinics. The estimated variation in slopes is $\hat{\sigma}^2_{u_{\text{months}}} = 2.99$, which is significant, $z = 2.98$, $p = 0.003$, suggesting significant variation in the effect of time on post-surgery quality of life across clinics. The correlation between slopes and intercepts is estimated to be $\hat{r}_{u_0, u_{\text{months}}} = -0.71$, suggesting that clinics with large intercepts tended to have smaller slopes.

For the model itself, you have two choices. The first is to report the results with the *F*s and degrees of freedom for the fixed effects, and then report the parameters for the random effects in the text. The second is to produce a table of parameters as you would for a linear model. For example, you might report the cosmetic surgery example as follows:

✓ The overall effect of time on quality of life was small and non-significant, $\hat{\beta} = 0.48$ [–0.35, 1.31], $t(19.63) = 1.22$, $p = 0.239$. The effect of the reason for surgery was also small and non-significant, $\hat{\beta} = -1.79$ [–3.71, 0.13], $t(1535.47) = -1.83$, $p = 0.067$. To put this effect in context, with other variables held constant the difference in quality of life in those seeking treatment for cosmetic reasons rather than physical treasons was 1.79 lower (on the 100-point scale). The effect of baseline quality of life on post-surgery quality of life was more substantial and significant, $\hat{\beta} = 0.47$ [0.42, 0.52], $t(1534.93) = 19.56$, $p < 0.001$. This effect suggests that, with other variables held constant, for every unit increase in baseline quality of life there is a half unit increase in post-surgery quality of life, which is a fairly strong relationship (because both are measured on the same scale). The combined effect of months and reason (their interaction) on post-surgery quality of life was significant, $\hat{\beta} = 0.45$ [0.19, 0.70], $t(1535.12) = 3.44$, $p = 0.001$. For those who had surgery to change their appearance, their quality of life increased over time but not significantly so, $\hat{\beta} = 0.51$ [–0.29, 1.31], $t(19.13) = 1.32$, $p = 0.200$. In contrast, for those who had surgery to help with a physical problem, their quality of life significantly increased over time, $\hat{\beta} = 0.92$ [0.03, 1.81], $t(18.75) = 2.16$, $p = 0.044$. The change in quality of life over the same time period for people who had surgery for a physical reason is about double that of people who had it for a cosmetic reason.

Alternatively, you could present parameter information as in Table 21.1.

Table 21.1 Summary of the multilevel model

	$\hat{\beta}$	95% CI	*t*	*df*	*p*
Baseline QoL	0.47	[0.42, 0.52]	19.56	1534.93	< 0.001
Months	0.48	[−0.35, 1.31]	1.22	19.63	0.239
Reason	−1.79	[−3.71, 0.13]	−1.83	1535.47	0.067
Months × Reason	0.45	[0.19, 0.70]	3.44	1535.12	0.001

21.7 A message from the octopus of inescapable despair Ⓐ

When I started writing this chapter I didn't know anything about multilevel models, but by its completion I felt a tiny bit smug that I had them nailed. However, I don't, and if you now feel like you understand multilevel models too, then you're wrong. You're not wrong because you're daft, but because multilevel modelling is very complicated and this chapter barely scratches the surface of what there is to know. Multilevel models often fail to converge, with no apology or explanation, and trying to fathom out what's happening can feel like hammering nails into your head.

21.8 Jane and Brian's story Ⓐ

Jane was giddy with excitement. Theoretically the device would work, but she wouldn't know for sure until she tried it. It had been easy enough to steal a portable TMS device, but getting it to focus its magnetic pulse into a precise location had beaten her at first. Then last week it came to her, in the shower of all places, and she'd spent another three weeks away from the world, beating away at the problem. Occasional visits to Brian, but mainly in lockdown in the lab.

Looking at the machine, Jane felt proud, perhaps for the first time. She'd solved this problem on her own, with no visits to the Pleiades building. She'd taken her intellect for granted, played it down even, but this was a real achievement. She was desperate to test the device, but she needed to sleep, and to double-check everything with a fresh head. The field test would have to wait. She went home, showered, ate and messaged Brian. 'Sorry, been busy, I'll come over soon.' She curled up under her duvet and felt utterly content.

21.9 What next? Ⓑ

This brings my life story somewhat up to date. I left out some of the more colourful bits, but only because I couldn't find even an extremely tenuous way to link them to statistics. We saw that over my life I managed to completely fail to achieve any of my childhood dreams. It's OK, I have other ambitions now (on a slightly smaller scale than 'rock star') and I'm looking forward to failing to achieve them too. I did at least manage to marry my lovely wife. Writing this chapter, like marriage, was a leap into the unknown. Marriage, however, has proved to be infinitely more enjoyable than writing about multilevel models. I think marriage is a useful metaphor for learning about statistics: if you think about both things logically you might never do them because they are full of uncertainty and potential scariness. However, you have to go with your heart, knowing that jumping in will enrich you. Admittedly, the kind of enrichment that marriage bestows is more obviously pleasant than knowing about autoregressive covariance structures, but statistics does give you enormous power to negotiate the scientific world (not just as practising scientists, but as normal people evaluating the, often misleading, media reports of scientific findings).

My wife and I think a lot about what makes a marriage work, and we think it comes down to reciprocal effort to enrich the other person's life. There is a parallel to this book: you and I have entered into a statistical relationship of sorts. For my part, I've put as much effort as I can into trying to pass on what I know about statistics, and if you have reciprocated that effort in reading the book and working through the examples, then hopefully our time together has enriched you. In return, your reactions to this book, more often than not, enrich me.

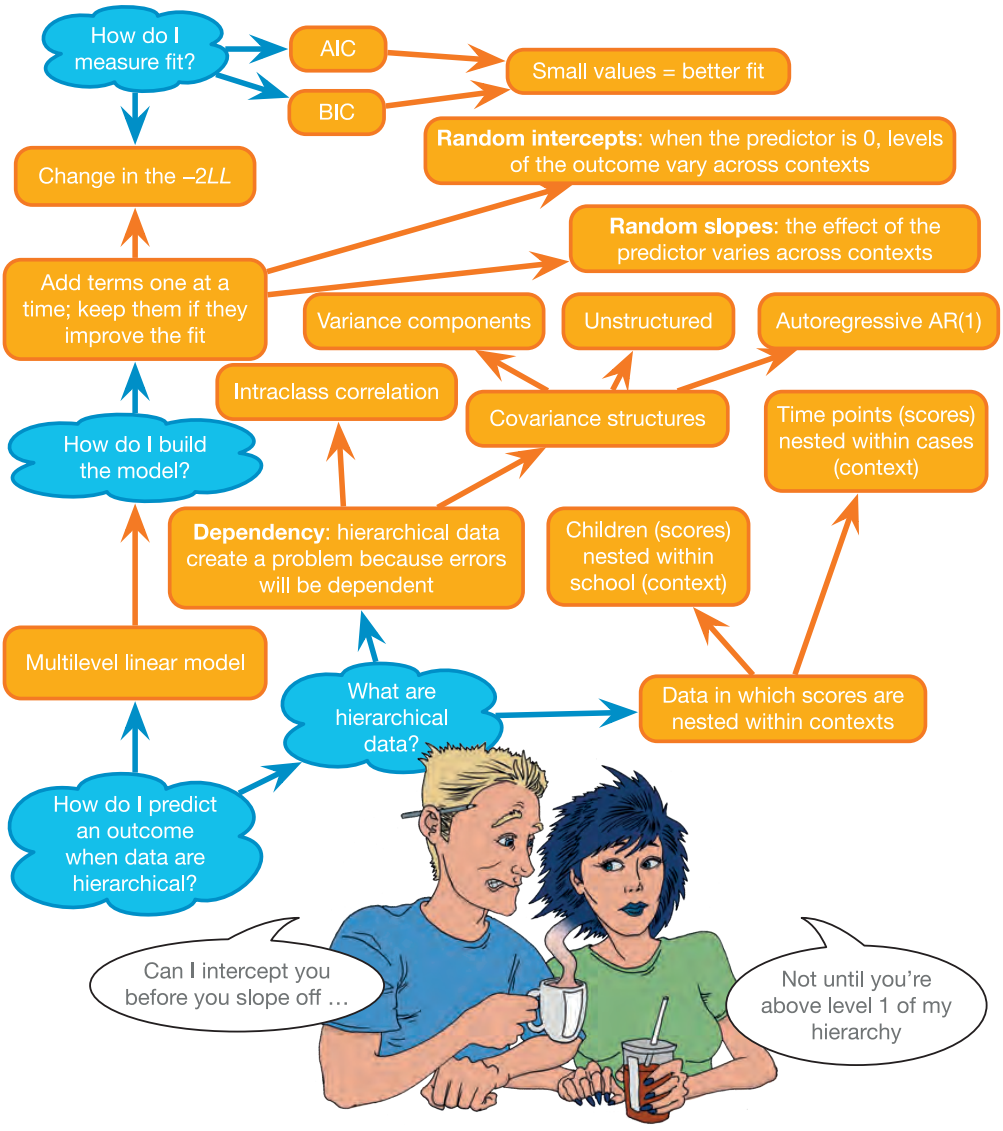
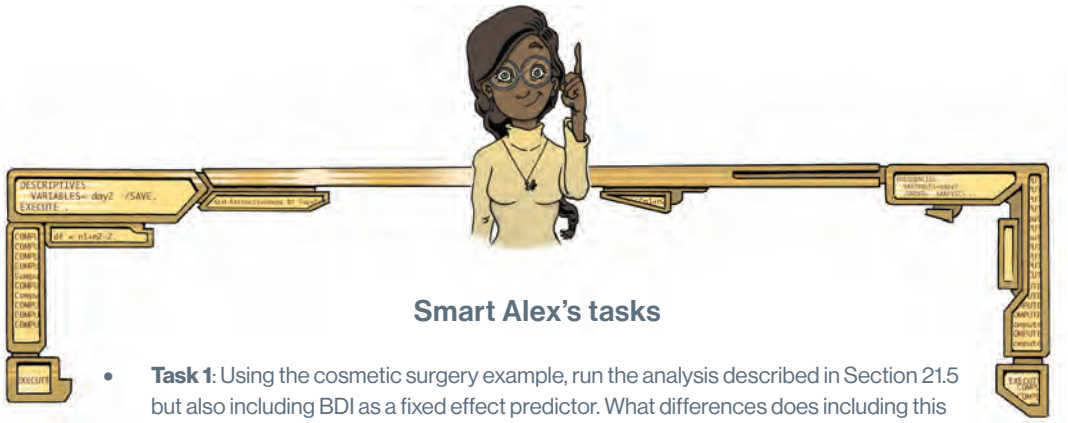


Figure 21.22 What Brian learnt from this chapter

21.10 Key terms that I've discovered

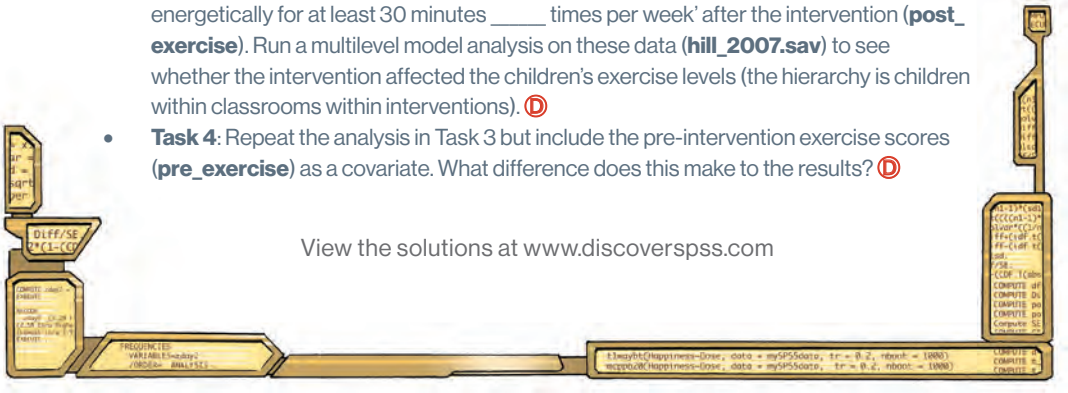
Akaike's information criterion (AIC)
AR(1)
Bayesian information criterion (BIC)
Centring
Contextual variable
Diagonal
Fixed coefficient
Fixed effect
Fixed intercept
Fixed slope
Fixed variable
Grand mean centring
Group mean centring
Multilevel linear model
Random coefficient
Random effect
Random intercept
Random slope
Random variable
Unstructured
Variance components



Smart Alex's tasks

- **Task 1**: Using the cosmetic surgery example, run the analysis described in Section 21.5 but also including BDI as a fixed effect predictor. What differences does including this predictor make? Ⓓ
- **Task 2**: Miller et al. (2007) tested the 'hidden-estrus' theory, which suggests that unlike other female mammals, humans do not experience an 'estrus' phase during which they are more sexually receptive, proceptive, selective and attractive. If this theory is wrong then human heterosexual men should find women most attractive during the fertile phase of their menstrual cycle compared to the pre-fertile (menstrual) and post-fertile (luteal) phase. Miller used the **tips** obtained by dancers (**id**) at a lap dancing club as a proxy for their sexual attractiveness and also recorded the phase of the dancer's menstrual cycle during a given shift (**cyclephase**), and whether they were using hormonal contraceptives (**contraceptive**). Dancers provided data from between 9 and 29 of their shifts. Fit a multilevel model using these data (**miller_2007.sav**) to see whether **tips** can be predicted from **cyclephase**, **contraceptive** and their interaction, allowing the overall level of tips to vary across dancers. Is the 'hidden-estrus' hypothesis supported? Ⓓ
- **Task 3:** Hill et al. (2007) examined whether providing children with a leaflet based on the 'theory of planned behaviour' increased their exercise. There were four different interventions (**intervention**): a control group, a leaflet, a leaflet and quiz, and a leaflet and a plan. A total of 503 children from 22 different classrooms were sampled (**classroom**). The 22 classrooms were randomly assigned to the four different conditions. Children were asked 'On average over the last three weeks, I have exercised energetically for at least 30 minutes _____ times per week' after the intervention (**post_exercise**). Run a multilevel model analysis on these data (**hill_2007.sav**) to see whether the intervention affected the children's exercise levels (the hierarchy is children within classrooms within interventions). Ⓓ
- **Task 4**: Repeat the analysis in Task 3 but include the pre-intervention exercise scores (**pre_exercise**) as a covariate. What difference does this make to the results? Ⓓ

View the solutions at www.discoverspss.com



EPILOGUE

Here's some questions that the writer sent
Can an observer be a participant?
Have I seen too much?
Does it count if it doesn't touch?
If the view is all I can ascertain,
Pure understanding is out of range

(Fugazi, 2001)

Brian woke with a spasm. A sharp pain had woken him up, but he couldn't move. His arms and legs were tied to the bed. His head felt heavy, like it was clamped. As his eyes focused he saw Jane walking away with a syringe in her hand. He realized that the top of his head was in a metal device of some description. He sensed it humming slightly, vibrating his skin. He followed the wires from his head to a metal box by the side of the bed. It looked like a failed DIY electronics project. He panicked. There had been rumours about Jane around campus, that she was a psychopath, but Brian put it down to her quirkiness, and other girls envying her genius. As he considered his predicament, he felt like a fool, and terrified. She placed the syringe on the breakfast bar and turned to face him.

His throat congealed. He was sweating. He remembered reading that hostages should humanize themselves to their captors. 'Please,' he rasped, 'think what this will do to my dad.'

Jane looked confused. She walked calmly over, stroked his hand and said, 'Let yourself go, it won't hurt.'

Brian was sure that brain extraction would hurt a great deal. Whatever Jane had injected was taking effect, though, and he felt drowsy and weak. She turned to the machine, flicked some switches, and the thing on Brian's head hummed more intensely, pulsing periodically. Jane faded from view.

Brian's eyes opened. He couldn't speak, and he was cocooned. His eyes darted around, everything was blurred. He could make out contrast, but nothing tangible. His gaze fixed on a familiar configuration of blurred shapes. They made him feel content and safe. A voice was quietly singing near him. 'Lavender's blue, dilly dilly, lavender's green. When you are king, dilly dilly, I shall be queen. A penny for your thoughts my dear, a penny for your thoughts my dear, IOU for your love, IOU for your love, for your love,' she sang, and Brian fell asleep.

Brian's eyes opened. He was standing, but his legs felt like he was too heavy for them. Two big hands grabbed his own from behind. In front of him was a woman in her early thirties crouching down, smiling, her arms open. She had long brown hair, big brown eyes. Seeing her made him smile, but he didn't know why. She was talking to him, beckoning with her hands. He couldn't understand her, but her face and voice excited him. He bounced a little, trying out his legs. A deeper voice spoke as the

giant hands that propped him up let go. Brian had one intense urge: to go to the woman. He put one foot forward and his weight shifted awkwardly onto it. He felt himself falling and pulled his other leg forward to regain his balance. Moving felt exciting. He moved the first leg again, then the other, then again, and again, and faster and faster until he tumbled into the arms of the woman. The two voices cheered and the woman looked him in the eyes. It felt as though she was pouring her soul into him. He'd never felt so happy.

Brian's eyes opened. He was in a children's playground. A woman embraced him. His head was nestled in her brown hair. It smelt of coconut. He loved the smell. His wrist hurt and he was crying. The woman kissed his wrist. 'It'll be better soon,' she said. It was his mum. Her words made him feel safe. A man came over. He looked concerned. Brian recognized him. It was his dad. He looked young.

'Is he OK?' he asked the woman.

'Yeah, he just had a little fall, nothing major,' she replied.

Brian's mum locked eyes with his dad for perhaps just a second and they exchanged reassuring smiles. Brian studied them. Each of them had given him the same look a thousand times. 'Pouring their souls into each other,' he thought, but he wasn't sure why.

Brian had dismissed the idea of your life flashing before your eyes before you die, but maybe he'd been wrong. His head was pulsing from the flood of thoughts and images. Christmas mornings opening presents, sitting on his bed listening to storybooks, helping him to build Lego, walking through woods, sitting in cafés, scooting in the park, eating beans on toast, splashing in the bath, and the sound of that lavender song at bedtime when he couldn't sleep. From the excitement of him scoring a goal for his team at sports day to the mundane putting on of shoes. Brian remembered a thousand things and more, all with one thing in common: the brown-haired woman. His mum. He saw her tears as she left him at nursery, her worry as he climbed in the playground, her patience as he asked questions and pushed boundaries, her empathy when he was upset, her rage when he was wronged, and her joy when he was happy. But these were not the photographic images that his dad showed him, these were like bite-sized chunks of reality sprouting in his consciousness. He could hear her, touch her, smell her, feel her. As his memories engulfed him, he found it harder to breathe. Was this the end? He tensed.

Brian's eyes opened. He saw school gates. His mum knelt to face him like she did every morning. Brian couldn't look her in the eye. He felt guilty. At breakfast he had told her that it was embarrassing having his mum take him to school. He was 10, he could go on his own. When he'd seen how hurt she'd looked, he wanted to take the words back. He'd be at high school next term, and he'd get the bus. He didn't need to say anything, he just needed to be patient.

'About this morning,' she began. 'Maybe you're right, maybe you are old enough to walk here on your own. It's hard, though, … as a parent. … Children can't help but take their parents for granted because they've never known a world without them, but parents had a life before their children, so they know what they have to lose. It can make you … overprotective. I'm proud of you for having the courage to tell me. If you don't want me to walk you to school, that's fine.'

Brian hugged her tightly. He didn't want to let go. He secretly hoped that she'd change her mind because he didn't want this to be the last time that she walked him to school, but it was, although not because of what he'd said.

Brian's eyes opened. He was in his apartment. Jane smiled at him. He convulsed violently as reality sank in, his juddering body eventually settling into a rhythm of uninhibited sobbing. Simultaneously he felt as euphoric and as lonely as he could ever remember. He wished that another human being wasn't here to witness his vulnerability.

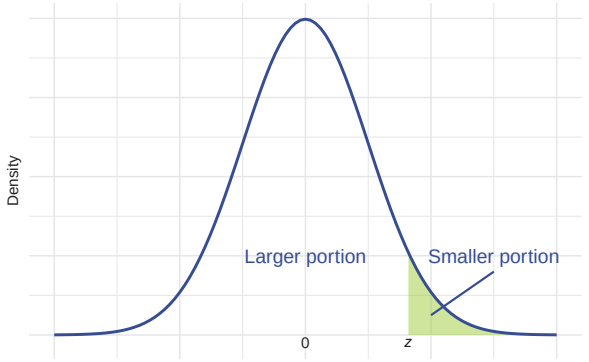
Jane switched off the machine, removed the cap from Brian's twitching head, and slowly untied the binding on each of his limbs. Brian curled up into a ball as if trying to hide from Jane. She got on the bed and lay next to him, put her arm around him and pulled herself into his back. It didn't feel awkward at all. She held him until his breathing returned to normal and he was still. Brian turned to face Jane. His eyes were red. He looked spent.

'I remember her,' he said.

'You're welcome,' Jane replied.

APPENDIX

A.1 Table of the standard normal distribution



z	Larger Portion	Smaller Portion	y
.00	.50000	.50000	.3989
.01	.50399	.49601	.3989
.02	.50798	.49202	.3989
.03	.51197	.48803	.3988
.04	.51595	.48405	.3986
.05	.51994	.48006	.3984
.06	.52392	.47608	.3982
.07	.52790	.47210	.3980
.08	.53188	.46812	.3977
.09	.53586	.46414	.3973
.10	.53983	.46017	.3970
.11	.54380	.45620	.3965
.12	.54776	.45224	.3961
.13	.55172	.44828	.3956
.14	.55567	.44433	.3951
.15	.55962	.44038	.3945
.16	.56356	.43644	.3939
.17	.56749	.43251	.3932
.18	.57142	.42858	.3925
.19	.57535	.42465	.3918
.20	.57926	.42074	.3910
.21	.58317	.41683	.3902
.22	.58706	.41294	.3894
.23	.59095	.40905	.3885
.24	.59483	.40517	.3876

z	Larger Portion	Smaller Portion	y
.25	.59871	.40129	.3867
.26	.60257	.39743	.3857
.27	.60642	.39358	.3847
.28	.61026	.38974	.3836
.29	.61409	.38591	.3825
.30	.61791	.38209	.3814
.31	.62172	.37828	.3802
.32	.62552	.37448	.3790
.33	.62930	.37070	.3778
.34	.63307	.36693	.3765
.35	.63683	.36317	.3752
.36	.64058	.35942	.3739
.37	.64431	.35569	.3725
.38	.64803	.35197	.3712
.39	.65173	.34827	.3697
.40	.65542	.34458	.3683
.41	.65910	.34090	.3668
.42	.66276	.33724	.3653
.43	.66640	.33360	.3637
.44	.67003	.32997	.3621
.45	.67364	.32636	.3605
.46	.67724	.32276	.3589
.47	.68082	.31918	.3572
.48	.68439	.31561	.3555
.49	.68793	.31207	.3538

z	Larger Portion	Smaller Portion	*y*
.50	.69146	.30854	.3521
.51	.69497	.30503	.3503
.52	.69847	.30153	.3485
.53	.70194	.29806	.3467
.54	.70540	.29460	.3448
.55	.70884	.29116	.3429
.56	.71226	.28774	.3410
.57	.71566	.28434	.3391
.58	.71904	.28096	.3372
.59	.72240	.27760	.3352
.60	.72575	.27425	.3332
.61	.72907	.27093	.3312
.62	.73237	.26763	.3292
.63	.73565	.26435	.3271
.64	.73891	.26109	.3251
.65	.74215	.25785	.3230
.66	.74537	.25463	.3209
.67	.74857	.25143	.3187
.68	.75175	.24825	.3166
.69	.75490	.24510	.3144
.70	.75804	.24196	.3123
.71	.76115	.23885	.3101
.72	.76424	.23576	.3079
.73	.76730	.23270	.3056
.74	.77035	.22965	.3034
.75	.77337	.22663	.3011
.76	.77637	.22363	.2989
.77	.77935	.22065	.2966
.78	.78230	.21770	.2943
.79	.78524	.21476	.2920
.80	.78814	.21186	.2897
.81	.79103	.20897	.2874
.82	.79389	.20611	.2850
.83	.79673	.20327	.2827
.84	.79955	.20045	.2803
.85	.80234	.19766	.2780
.86	.80511	.19489	.2756
.87	.80785	.19215	.2732
.88	.81057	.18943	.2709
.89	.81327	.18673	.2685
.90	.81594	.18406	.2661
.91	.81859	.18141	.2637
.92	.82121	.17879	.2613
.93	.82381	.17619	.2589
.94	.82639	.17361	.2565

z	Larger Portion	Smaller Portion	*y*
.95	.82894	.17106	.2541
.96	.83147	.16853	.2516
.97	.83398	.16602	.2492
.98	.83646	.16354	.2468
.99	.83891	.16109	.2444
1.00	.84134	.15866	.2420
1.01	.84375	.15625	.2396
1.02	.84614	.15386	.2371
1.03	.84849	.15151	.2347
1.04	.85083	.14917	.2323
1.05	.85314	.14686	.2299
1.06	.85543	.14457	.2275
1.07	.85769	.14231	.2251
1.08	.85993	.14007	.2227
1.09	.86214	.13786	.2203
1.10	.86433	.13567	.2179
1.11	.86650	.13350	.2155
1.12	.86864	.13136	.2131
1.13	.87076	.12924	.2107
1.14	.87286	.12714	.2083
1.15	.87493	.12507	.2059
1.16	.87698	.12302	.2036
1.17	.87900	.12100	.2012
1.18	.88100	.11900	.1989
1.19	.88298	.11702	.1965
1.20	.88493	.11507	.1942
1.21	.88686	.11314	.1919
1.22	.88877	.11123	.1895
1.23	.89065	.10935	.1872
1.24	.89251	.10749	.1849
1.25	.89435	.10565	.1826
1.26	.89617	.10383	.1804
1.27	.89796	.10204	.1781
1.28	.89973	.10027	.1758
1.29	.90147	.09853	.1736
1.30	.90320	.09680	.1714
1.31	.90490	.09510	.1691
1.32	.90658	.09342	.1669
1.33	.90824	.09176	.1647
1.34	.90988	.09012	.1626
1.35	.91149	.08851	.1604
1.36	.91309	.08691	.1582
1.37	.91466	.08534	.1561
1.38	.91621	.08379	.1539
1.39	.91774	.08226	.1518

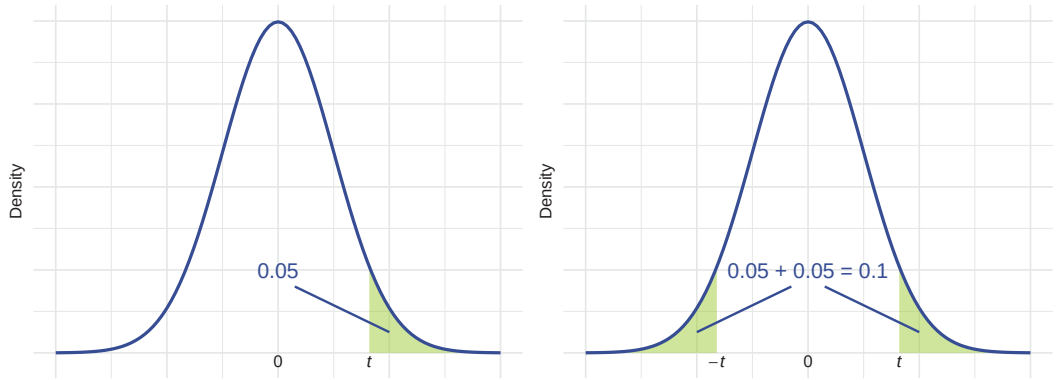
z	Larger Portion	Smaller Portion	y
1.40	.91924	.08076	.1497
1.41	.92073	.07927	.1476
1.42	.92220	.07780	.1456
1.43	.92364	.07636	.1435
1.44	.92507	.07493	.1415
1.45	.92647	.07353	.1394
1.46	.92785	.07215	.1374
1.47	.92922	.07078	.1354
1.48	.93056	.06944	.1334
1.49	.93189	.06811	.1315
1.50	.93319	.06681	.1295
1.51	.93448	.06552	.1276
1.52	.93574	.06426	.1257
1.53	.93699	.06301	.1238
1.54	.93822	.06178	.1219
1.55	.93943	.06057	.1200
1.56	.94062	.05938	.1182
1.57	.94179	.05821	.1163
1.58	.94295	.05705	.1145
1.59	.94408	.05592	.1127
1.60	.94520	.05480	.1109
1.61	.94630	.05370	.1092
1.62	.94738	.05262	.1074
1.63	.94845	.05155	.1057
1.64	.94950	.05050	.1040
1.65	.95053	.04947	.1023
1.66	.95154	.04846	.1006
1.67	.95254	.04746	.0989
1.68	.95352	.04648	.0973
1.69	.95449	.04551	.0957
1.70	.95543	.04457	.0940
1.71	.95637	.04363	.0925
1.72	.95728	.04272	.0909
1.73	.95818	.04182	.0893
1.74	.95907	.04093	.0878
1.75	.95994	.04006	.0863
1.76	.96080	.03920	.0848
1.77	.96164	.03836	.0833
1.78	.96246	.03754	.0818
1.79	.96327	.03673	.0804
1.80	.96407	.03593	.0790
1.81	.96485	.03515	.0775
1.82	.96562	.03438	.0761
1.83	.96638	.03362	.0748
1.84	.96712	.03288	.0734

z	Larger Portion	Smaller Portion	y
1.85	.96784	.03216	.0721
1.86	.96856	.03144	.0707
1.87	.96926	.03074	.0694
1.88	.96995	.03005	.0681
1.89	.97062	.02938	.0669
1.90	.97128	.02872	.0656
1.91	.97193	.02807	.0644
1.92	.97257	.02743	.0632
1.93	.97320	.02680	.0620
1.94	.97381	.02619	.0608
1.95	.97441	.02559	.0596
1.96	.97500	.02500	.0584
1.97	.97558	.02442	.0573
1.98	.97615	.02385	.0562
1.99	.97670	.02330	.0551
2.00	.97725	.02275	.0540
2.01	.97778	.02222	.0529
2.02	.97831	.02169	.0519
2.03	.97882	.02118	.0508
2.04	.97932	.02068	.0498
2.05	.97982	.02018	.0488
2.06	.98030	.01970	.0478
2.07	.98077	.01923	.0468
2.08	.98124	.01876	.0459
2.09	.98169	.01831	.0449
2.10	.98214	.01786	.0440
2.11	.98257	.01743	.0431
2.12	.98300	.01700	.0422
2.13	.98341	.01659	.0413
2.14	.98382	.01618	.0404
2.15	.98422	.01578	.0396
2.16	.98461	.01539	.0387
2.17	.98500	.01500	.0379
2.18	.98537	.01463	.0371
2.19	.98574	.01426	.0363
2.20	.98610	.01390	.0355
2.21	.98645	.01355	.0347
2.22	.98679	.01321	.0339
2.23	.98713	.01287	.0332
2.24	.98745	.01255	0325
2.25	.98778	.01222	.0317
2.26	.98809	.01191	.0310
2.27	.98840	.01160	.0303
2.28	.98870	.01130	.0297
2.29	.98899	.01101	.0290

z	Larger Portion	Smaller Portion	*y*
2.30	.98928	.01072	.0283
2.31	.98956	.01044	.0277
2.32	.98983	.01017	.0270
2.33	.99010	.00990	.0264
2.34	.99036	.00964	.0258
2.35	.99061	.00939	.0252
2.36	.99086	.00914	.0246
2.37	.99111	.00889	.0241
2.38	.99134	.00866	.0235
2.39	.99158	.00842	.0229
2.40	.99180	.00820	.0224
2.41	.99202	.00798	.0219
2.42	.99224	.00776	.0213
2.43	.99245	.00755	.0208
2.44	.99266	.00734	.0203
2.45	.99286	.00714	.0198
2.46	.99305	.00695	.0194
2.47	.99324	.00676	.0189
2.48	.99343	.00657	.0184
2.49	.99361	.00639	.0180
2.50	.99379	.00621	.0175
2.51	.99396	.00604	.0171
2.52	.99413	.00587	.0167
2.53	.99430	.00570	.0163
2.54	.99446	.00554	.0158
2.55	.99461	.00539	.0154
2.56	.99477	.00523	.0151
2.57	.99492	.00508	.0147
2.58	.99506	.00494	.0143
2.59	.99520	.00480	.0139
2.60	.99534	.00466	.0136
2.61	.99547	.00453	.0132
2.62	.99560	.00440	.0129
2.63	.99573	.00427	.0126
2.64	.99585	.00415	.0122
2.65	.99598	.00402	.0119
2.66	.99609	.00391	.0116
2.67	.99621	.00379	.0113
2.68	.99632	.00368	.0110
2.69	.99643	.00357	.0107
2.70	.99653	.00347	.0104
2.71	.99664	.00336	.0101
2.72	.99674	.00326	.0099
2.73	.99683	.00317	.0096
2.74	.99693	.00307	.0093
2.75	.99702	.00298	.0091
2.76	.99711	.00289	.0088
2.77	.99720	.00280	.0086
2.78	.99728	.00272	.0084
2.79	.99736	.00264	.0081
2.80	.99744	.00256	.0079
2.81	.99752	.00248	.0077
2.82	.99760	.00240	.0075
2.83	.99767	.00233	.0073
2.84	.99774	.00226	.0071
2.85	.99781	.00219	.0069
2.86	.99788	.00212	.0067
2.87	.99795	.00205	.0065
2.88	.99801	.00199	.0063
2.89	.99807	.00193	.0061
2.90	.99813	.00187	.0060
2.91	.99819	.00181	.0058
2.92	.99825	.00175	.0056
2.93	.99831	.00169	.0055
2.94	.99836	.00164	.0053
2.95	.99841	.00159	.0051
2.96	.99846	.00154	.0050
2.97	.99851	.00149	.0048
2.98	.99856	.00144	.0047
2.99	.99861	.00139	.0046
3.00	.99865	.00135	.0044
...	...	...	...
3.25	.99942	.00058	.0020
...	...	...	...
3.50	.99977	.00023	.0009
...	...	...	...
4.00	.99997	.00003	.0001

Values computed by the author using IBM SPSS Statistics

A.2 Critical values of the *t*-distribution

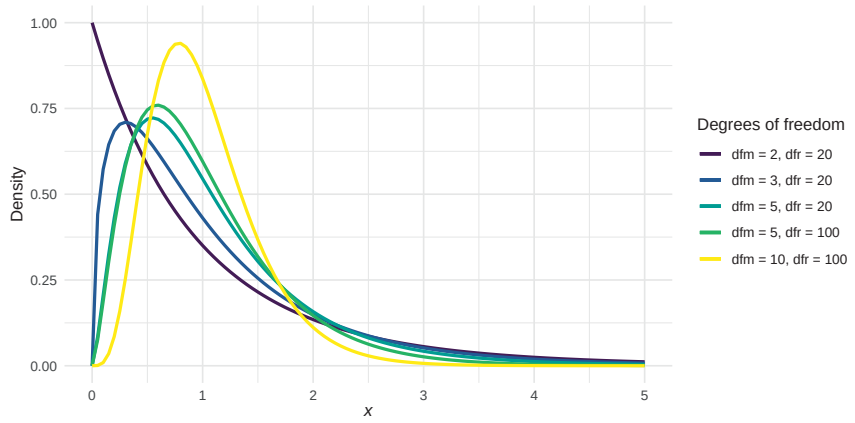


	0.10	0.05	0.025	0.01	0.005
			Proportion across Two Tails		
df	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845

	Proportion across Two Tails				
df	0.20	0.10	0.05	0.02	0.01
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
55	1.297	1.673	2.004	2.396	2.668
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
∞	1.282	1.645	1.960	2.326	2.576

Values computed by the author using R

A.3 Critical values of the *F*-distribution



df_R	df_M													
	1	2	3	4	5	6	7	8	9	10	15	25	50	100
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	249.26	251.77	253.04
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.46	19.48	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.63	8.58	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.77	5.70	5.66
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.52	4.44	4.41
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.83	3.75	3.71
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.40	3.32	3.27
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.11	3.02	2.97
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.89	2.80	2.76
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.73	2.64	2.59
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.50	2.40	2.35
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.34	2.24	2.19
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.23	2.12	2.07
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.14	2.04	1.98
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.07	1.97	1.91
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.02	1.91	1.85
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	1.97	1.86	1.80
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.94	1.82	1.76
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.91	1.79	1.73
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.88	1.76	1.70
33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13	1.98	1.84	1.72	1.66
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	1.96	1.82	1.70	1.63

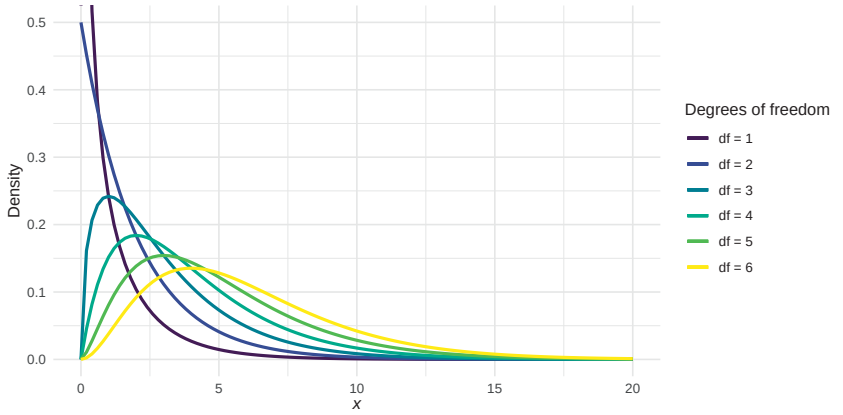
df_R	df_M													
	1	2	3	4	5	6	7	8	9	10	15	25	50	100
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.78	1.66	1.59
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.89	1.75	1.63	1.55
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.73	1.60	1.52
55	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01	1.85	1.71	1.58	1.50
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.69	1.56	1.48
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.03	1.98	1.82	1.68	1.54	1.46
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.81	1.66	1.53	1.45
75	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96	1.80	1.65	1.52	1.44
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.79	1.64	1.51	1.43
85	3.95	3.10	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.94	1.79	1.64	1.50	1.42
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.78	1.63	1.49	1.41
95	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93	1.77	1.62	1.48	1.40
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.62	1.48	1.39

df_R	df_M													
	1	2	3	4	5	6	7	8	9	10	15	25	50	100
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6157.28	6239.83	6302.52	6334.11
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43	99.46	99.48	99.49
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87	26.58	26.35	26.24
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	13.91	13.69	13.58
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.45	9.24	9.13
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.30	7.09	6.99
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.06	5.86	5.75
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.26	5.07	4.96
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.71	4.52	4.41
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.31	4.12	4.01
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.76	3.57	3.47
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.41	3.22	3.11
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.16	2.97	2.86
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	2.98	2.78	2.68
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.84	2.64	2.54
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.73	2.53	2.42
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89	2.64	2.44	2.33
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.81	2.57	2.36	2.25
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.75	2.51	2.30	2.19
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.45	2.25	2.13
33	7.47	5.31	4.44	3.95	3.63	3.41	3.24	3.11	3.00	2.91	2.63	2.39	2.18	2.06
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.60	2.35	2.14	2.02
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.27	2.06	1.94
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.46	2.21	2.00	1.88

df_R	df_M													
	1	2	3	4	5	6	7	8	9	10	15	25	50	100
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42	2.17	1.95	1.82
55	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.38	2.13	1.91	1.78
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.10	1.88	1.75
65	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.80	2.69	2.61	2.33	2.07	1.85	1.72
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.31	2.05	1.83	1.70
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57	2.29	2.03	1.81	1.67
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.27	2.01	1.79	1.65
85	6.94	4.86	4.02	3.55	3.24	3.02	2.86	2.73	2.62	2.54	2.26	2.00	1.77	1.64
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.24	1.99	1.76	1.62
95	6.91	4.84	3.99	3.52	3.22	3.00	2.83	2.70	2.60	2.51	2.23	1.98	1.75	1.61
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	1.97	1.74	1.60

Values computed by the author using R

A.4 Critical values of the chi-square distribution



df	$p = 0.05$	$p = 0.01$
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.92	40.29
23	35.17	41.64
24	36.42	42.98

df	$p = 0.05$	$p = 0.01$
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89
35	49.80	57.34
40	55.76	63.69
45	61.66	69.96
50	67.50	76.15
60	79.08	88.38
70	90.53	100.43
80	101.88	112.33
90	113.15	124.12
100	124.34	135.81
200	233.99	249.45
300	341.40	359.91
400	447.63	468.72
500	553.13	576.49
600	658.09	683.52
700	762.66	789.97
800	866.91	895.98
900	970.90	1001.63
1000	1074.68	1106.97

Values computed by the author using R

GLOSSARY

0: How much clue Sage have about the amount of effort I put into writing this book.

–2*LL*: The *log-likelihood* multiplied by minus 2. This version of the likelihood is used in *logistic regression*.

α-level: The probability of making a *Type I error* (usually this value is 0.05).

A life: What you don't have when writing statistics textbooks.

Adjusted mean: In the context of *analysis of covariance* this is the value of the group mean adjusted for the effect of the *covariate*.

Adjusted predicted value: A measure of the influence of a particular case of data. It is the predicted value of a case from a model estimated without that case included in the data. The value is calculated by re-estimating the model without the case in question, then using this new model to predict the value of the excluded case. If a case does not exert a large influence over the model then its predicted value should be similar regardless of whether the model was estimated including or excluding that case. The difference between the predicted value of a case from the model when that case was included and the predicted value from the model when it was excluded is the *DFFit*.

Adjusted R^2: A measure of proportion of variance in an *outcome variable* explained by a model adjusted for the number of predictors in the model. It measures the *shrinkage*. By adjusting for the number of predictors, it gives us a better estimate (than the unadjusted value) of the R^2 in the population.

AIC (Akaike's information criterion): A *goodness-of-fit* measure that is corrected for model complexity. That just means that it takes account of how many *parameters* have been estimated. It is not intrinsically interpretable, but can be compared in different models to see how changing the model affects the fit. A small value represents a better fit to the data.

Alpha factoring: A method of *factor analysis*.

Alternative hypothesis: The prediction that there will be an effect (i.e., that your experimental manipulation will have some effect or that certain variables will relate to each other).

Analysis of covariance: A statistical procedure that uses the *F-statistic* to test the overall fit of a linear model, adjusting for the effect that one or more *covariates* have on the *outcome variable*. In experimental research this linear model tends to be defined in terms of group means, and the resulting ANOVA is therefore an overall test of whether group means differ after the variance in the outcome variable explained by any *covariates* has been removed.

Analysis of variance: A statistical procedure that uses the *F-statistic* to test the overall fit of a linear model. In experimental research this linear model tends to be defined in terms of group means, and the resulting ANOVA is therefore an overall test of whether group means differ.

ANCOVA: Acronym for *analysis of covariance*.

Anderson–Rubin method: A way of calculating *factor scores* which produces scores that are uncorrelated and *standardized* with a mean of 0 and a standard deviation of 1.

ANOVA: Acronym for *analysis of variance*.

Assumption of normality: A misunderstood assumption of linear models. A common misunderstanding is that this assumption relates to the observed data (i.e., 'my data need to be normally distributed') but when using ordinary least squares the assumption relates to the *errors* of the model (which will be reflected in the model *residuals*), not the observed data. When errors cannot be assumed to have a normal distribution, model *parameter* estimates are unbiased (i.e., the assumption is largely irrelevant). However, for the *standard errors* of model *parameters* to be unbiased we need normally distributed errors or a large sample size otherwise confidence intervals and significance tests or the model *parameter* estimates will be biased (because they both use the standard error in their calculation).

Autocorrelation: When the *residuals* of two observations in a regression model are correlated.

b_i: Unstandardized *parameter* (aka *regression coefficient*). It indicates the strength of relationship between a given predictor, i, and an outcome in the units of measurement of the predictor. It is the change in the outcome associated with a unit change in the predictor.

β_i: Standardized *parameter* (aka *regression coefficient*). It indicates the strength of relationship between a given predictor, i, and an outcome in a *standardized* form. It is the change in the outcome (in standard deviations) associated with a one standard deviation change in the predictor.

β-level: The probability of making a *Type II error* (Cohen, 1992, suggests a maximum value of 0.2).

Bar chart: A graph in which a summary statistic (usually the mean) is plotted on the y-axis against a categorical variable on the x-axis (this categorical variable could represent, for example, groups of people, different times or different experimental conditions). The value of the mean for each category is shown by a bar. Different-coloured bars may be used to represent levels of a second categorical variable.

Bartlett's test of sphericity: Unsurprisingly, this is a test of the assumption of *sphericity*. This test examines whether a *variance–covariance matrix* is proportional to an *identity matrix*. Therefore, it effectively tests whether the diagonal elements of the variance–covariance matrix are equal (i.e., group variances are the same), and whether the off-diagonal elements are approximately zero (i.e., the *dependent variables* are not *correlated*). Jeremy Miles, who does a lot of multivariate stuff, claims he's never ever seen a matrix that reached non-significance using this test and, come to think of it, I've never seen one either (although I do less multivariate stuff), so you've got to wonder about its practical utility.

Bayes factor: The ratio of the probability of the observed data given the *alternative hypothesis* to the probability of the observed data given the *null hypothesis*. Put another way, it is the *likelihood* of the alternative hypothesis relative to the null. A Bayes factor of 3, for example, means that the observed data are 3 times more likely under the alternative hypothesis than under the null hypothesis. A Bayes factor less than 1 supports the null hypothesis by suggesting that the probability of the data given the null is higher than the probability of the data given the alternative hypothesis. Conversely, a Bayes factor greater than 1 suggests that the observed data are more likely given the alternative hypothesis than the null. Values between 1 and 3 are considered evidence for the alternative hypothesis that is 'barely worth mentioning', values between 3 and 10 are considered 'substantial evidence' ('having substance' rather than being 'very strong') for the alternative hypothesis, and values greater than 10 are strong evidence for the alternative hypothesis.

Bayesian statistics: A branch of statistics in which hypotheses are tested or model *parameters* are estimated using methods based on *Bayes' theorem*.

Bayes' theorem: A mathematical description of the relationship between the *conditional probability* of events A and B, $p(A|B)$, their reverse conditional probability, $p(B|A)$, and individual probabilities of the events, $p(A)$ and $p(B)$. The theorem states that

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

Beer-goggles effect: The phenomenon that people appear much more attractive after a few alcoholic drinks.

Between-groups design: Another name for *independent design*.

Between-subjects design: Another name for *independent design*.

BIC (Schwarz's Bayesian information criterion): A *goodness-of-fit* statistic comparable to the *AIC*, although it is slightly more conservative (it corrects more harshly for the number of *parameters* being estimated). It should be used when sample sizes are large and the number of *parameters* is small. It is not intrinsically interpretable, but can be compared in different models to see how changing the model affects the fit. A small value represents a better fit to the data.

Bimodal: A description of a distribution of observations that has two *modes*.

Binary logistic regression: *Logistic regression* in which the outcome variable has exactly two categories.

Binary variable: A *categorical variable* that has only two mutually exclusive categories (e.g., being dead or alive).

Biserial correlation: A standardized measure of the strength of relationship between two variables when one of the two variables is *dichotomous*. The biserial correlation coefficient is used when one variable is a continuous dichotomy (e.g., has an underlying continuum between the categories).

Bivariate correlation: A correlation between two variables.

Bonferroni correction: A correction applied to the α-*level* to control the overall *Type I error rate* when multiple significance tests are carried out. Each test conducted should use a criterion of significance of the α-level (normally 0.05) divided by the number of tests conducted. This is a simple but effective correction, but tends to be too strict when lots of tests are performed.

Bootstrap: A technique for estimating the sampling distribution of a statistic by taking repeated samples (with replacement) from the data set (in effect, treating the data as a population from which smaller samples are taken). The statistic of interest (e.g., the *mean*, or *b* coefficient) is calculated for each sample, from which the sampling distribution of the statistic is estimated. The standard error of the statistic is estimated as the standard deviation of the sampling distribution created from the bootstrap samples. From this, confidence intervals and significance tests can be computed.

Boredom effect: Refers to the possibility that performance in tasks may be influenced (the assumption is a negative influence) by boredom or lack of concentration if there are many tasks, or the task goes on for a long period of time. In short, what you are experiencing reading this glossary is a boredom effect.

Boxplot (a.k.a. box-whisker diagram)**:** A graphical representation of some important characteristics of a set of observations. At the centre of the plot is the *median*, which is surrounded by a box the top and bottom of which are the limits within which the middle 50% of observations fall (the *interquartile range*). Sticking out of the top and bottom of the box are two whiskers which extend to the highest and lowest extreme scores, respectively.

Box's test: A test of the assumption of *homogeneity of covariance matrices*. This test should be non-significant if the matrices are roughly the same. Box's test is very susceptible to deviations from *multivariate normality* and so may be non-significant not because the *variance–covariance matrices* are similar across groups, but because the assumption of multivariate normality is not tenable. Hence, it is vital to have some idea of whether the data meet the multivariate normality assumption (which is extremely difficult) before interpreting the result of Box's test.

Box–whisker plot: See *boxplot*.

Brown–Forsythe *F*: A version of the *F-statistic* designed to be accurate when the assumption of *homogeneity of variance* has been violated.

Categorical variable: Any variable made up of categories of objects/entities. The university you attend is a good example of a categorical variable: students who attend the University of Sussex are not also enrolled at Harvard or UV Amsterdam, therefore students fall into distinct categories.

Catterplot: When your data have fried your mind so much that every graph you look at transforms before your eyes into cute kittens that beckon you to 'come play with a ball of string'. Under no circumstances follow them.

Central limit theorem: This theorem states that when samples are large (above about 30) the *sampling distribution* will take the shape of a *normal distribution* regardless of the shape of the population from which the sample was drawn. For small samples the *t*-distribution better approximates the shape of the sampling distribution. We also know from this theorem that the *standard deviation* of the sampling distribution (i.e., the *standard error* of the sample *mean*) will be equal to the standard deviation of the sample (s) divided by the square root of the sample size (N).

Central tendency: A generic term describing the centre of a *frequency distribution* of observations as measured by the *mean*, *mode* and *median*.

Centring: The process of transforming a variable into deviations around a fixed point. This fixed point can be any value that is chosen, but typically the mean is used. To centre a variable the mean is subtracted from each score. See *grand mean centring*, *group mean centring*.

Chartjunk: Superfluous material that distracts from the data being displayed on a graph.

Chi-square distribution: A *probability distribution* of the sum of squares of several normally distributed variables. It tends to be used to test hypotheses about categorical data, and to test the fit of models to the observed data.

Chi-square test: Although this term can apply to any *test statistic* having a *chi-square distribution*, it generally refers to Pearson's chi-square test of the independence of two categorical variables. Essentially it tests whether two categorical variables forming a *contingency table* are associated.

Cochran's *Q*: This test is an extension of *McNemar's test* and is basically a *Friedman's ANOVA* for *dichotomous* data. So, imagine you asked 10 people whether they'd like to shoot Justin Timberlake, David Beckham and Simon Cowell and they could answer only 'yes' or 'no'. If we coded responses as 0 (no) and 1 (yes) we could do Cochran's test on these data.

Coefficient of determination: The proportion of variance in one variable explained by a second variable. It is *Pearson's correlation coefficient* squared.

Cohen's $\hat{d}$: An *effect size* that expresses the difference between two means in standard deviation units. In general, it can be estimated using:

$$\hat{d} = \frac{\bar{X}_1 - \bar{X}_2}{s}$$

Common factor: A *factor* that affects all measured *variables* and, therefore, explains the *correlations* between those variables.

Common variance: Variance shared by two or more variables.

Communality: The proportion of a variable's variance that is *common variance*. This term is used primarily in *factor analysis*. A variable that has no *unique variance* (or *random variance*) would have a communality of 1, whereas a variable that shares none of its variance with any other variable would have a communality of 0.

Complete separation: A situation in *logistic regression* when the outcome variable can be perfectly predicted by one predictor or a combination of predictors! Suffice it to say this situation makes your computer have the equivalent of a nervous breakdown: it'll start weeping and saying it doesn't know what to do.

Component matrix: General term for the *structure matrix* in *principal components analysis*.

Compound symmetry: A condition that holds true when both the variances across conditions are equal (this is the same as the *homogeneity of variance* assumption) and the *covariances* between pairs of conditions are also equal.

Concurrent validity: A form of *criterion validity* where there is evidence that scores from an instrument correspond to concurrently recorded external measures conceptually related to the measured construct.

Confidence interval: For a given statistic calculated for a sample of observations (e.g., the mean), the confidence interval is a range of values around that statistic that is believed to contain, in a certain proportion of samples (e.g., 95%), the true value of that statistic (i.e., the population *parameter*). What that also means is that for the other proportion of samples (e.g., 5%), the confidence interval won't contain that true value. The trouble is, you don't know which category your particular sample falls into.

Confirmatory factor analysis (CFA): A version of *factor analysis* in which specific hypotheses about structure and relations between the *latent variables* that underlie the data are tested.

Confounding variable: A variable (that we may or may not have measured) other than the *predictor variables* in which we're interested that potentially affects an *outcome variable*.

Contaminated normal distribution: See *mixed normal distribution*.

Content validity: Evidence that the content of a test corresponds to the content of the construct it was designed to cover.

Contingency table: A table representing the cross-classification of two or more *categorical variables*. The levels of each variable are arranged in a grid, and the number of observations falling into each category is noted in the cells of the table. For example, if we took the categorical variables of **glossary** (with two categories: whether an author was made to write a glossary or not) and **mental state** (with three categories: normal, sobbing uncontrollably and utterly psychotic), we could construct a table as below. This instantly tells us that 127 authors who were made to write a glossary ended up as utterly psychotic, compared to only 2 who did not write a glossary.

		Glossary		
		Author made to write glossary	No glossary	Total
Mental state	Normal	5	423	428
	Sobbing uncontrollably	23	46	69
	Utterly psychotic	127	2	129
	Total	155	471	626

Continuous variable: A variable that can be measured to any level of precision. (Time is a continuous variable, because there is in principle no limit on how finely it could be measured.)

Cook's distance: A measure of the overall influence of a case on a model. Cook and Weisberg (1982) have suggested that values greater than 1 may be cause for concern.

Correlation coefficient: A measure of the strength of association or relationship between two variables. See *Pearson's correlation coefficient*, *Spearman's correlation coefficient*, *Kendall's tau*.

Correlational research: A form of research in which you observe what naturally goes on in the world without directly interfering with it. This term implies that data will be analysed so as to look at

relationships between naturally occurring variables rather than making statements about cause and effect. Compare with *cross-sectional research*, *longitudinal research* and *experimental research*.

Counterbalancing: A process of systematically varying the order in which experimental conditions are conducted. In the simplest case of there being two conditions (A and B), counterbalancing simply implies that half of the participants complete condition A followed by condition B, whereas the remainder do condition B followed by condition A. The aim is to remove systematic bias caused by *practice effects* or *boredom effects*.

Covariance: A measure of the 'average' relationship between two variables. It is the average *cross-product deviation* (i.e., the cross-product divided by one less than the number of observations).

Covariance ratio (CVR): A measure of whether a case influences the variance of the *parameters* in a *regression model*. When this ratio is close to 1 the case has very little influence on the variances of the model *parameters*. Belsey et al. (1980) recommend the following: if the CVR of a case is greater than $1 + [3(k + 1)/n]$ (where k is the number of predictors and n is the sample size) then deleting that case will damage the precision of some of the model's *parameters*, but if it is less than $1 - [3(k + 1)/n]$ then deleting the case will improve the precision of some of the model's *parameters*.

Covariate: A variable that has a relationship with (in terms of *covariance*), or has the potential to be related to, the *outcome variable* we've measured.

Cox and Snell's R^2_{CS}: A version of the *coefficient of determination* for logistic regression. It is based on the log-likelihood of a model (LL_{new}), the log-likelihood of the original model ($LL_{baseline}$) and the sample size, n. However, it is notorious for not reaching its maximum value of 1 (see *Nagelkerke's* R^2_N).

Cramér's *V*: A measure of the strength of association between two *categorical variables* used when one of these variables has more than two categories. It is a variant of *phi* used because when one or both of the categorical variables contain more than two categories, phi fails to reach its minimum value of 0 (indicating no association).

Credible interval: In Bayesian statistics, a credible interval is an interval within which a certain percentage of the posterior distribution falls (usually 95%). It can be used to express the limits within which a *parameter* falls with a fixed probability. For example, if we estimated the average length of a romantic relationship to be 6 years with a 95% credible interval of 1 to 11 years, then this would mean that 95% of the posterior distribution for the length of romantic relationships falls between 1 and 11 years. A plausible estimate of the length of romantic relationships would, therefore, be 1 to 11 years.

Criterion validity: Evidence that scores from an instrument correspond with (*concurrent validity*) or predict (*predictive validity*) external measures conceptually related to the measured construct.

Cronbach's α: A measure of the reliability of a scale defined by

$$\alpha = \frac{N^2\overline{\text{cov}}}{\sum s^2_{\text{item}} + \sum \text{cov}_{\text{item}}}$$

in which the top half of the equation is the number of items (N) squared multiplied by the average covariance between items (the average of the off-diagonal elements in the *variance–covariance matrix*). The bottom half is the sum of all the elements in the *variance–covariance matrix*.

Cross-product deviations: A measure of the 'total' relationship between two variables. It is the deviation of one variable from its mean multiplied by the other variable's deviation from its mean.

Cross-sectional research: A form of research in which you observe what naturally goes on in the world without directly interfering with it by measuring several variables at a single time point. In psychology, this term usually implies that data come from people at different age points, with different people representing each age point. See also *correlational research*, *longitudinal research*.

Cross-validation: Assessing the accuracy of a model across different samples. This is an important step in *generalization*. In a *regression model* there are two main methods of cross-validation: *adjusted* R^2 or data splitting, in which the data are split randomly into two halves, and a regression model is estimated for each half and then compared.

Crying: What you feel like doing after writing statistics textbooks.

Cubic trend: If you connected the means in ordered conditions with a line then a cubic trend is shown by two changes in the direction of this line. You must have at least four ordered conditions.

Currency variable: A variable containing values of money.

Data view: There are two ways to view the contents of the *data editor* window. The data view shows you a spreadsheet and can be used for entering raw data. See also *variable view*.

Date variable: A variable made up of dates. The data can take forms such as dd-mmm-yyyy (e.g., 21-Jun-1973), dd-mmm-yy (e.g., 21-Jun-73), mm/dd/yy (e.g., 06/21/73) and dd.mm.yyyy (e.g., 21.06.1973).

Degrees of freedom: An impossible thing to define in a few pages, let alone a few lines. Essentially it is the number of 'entities' that are free to vary when estimating some kind of statistical *parameter*. In a more practical sense, it has a bearing on significance tests for many commonly used *test statistics* (such as the *F-statistic*, *t-test*, *chi-square statistic*) and determines the exact form of the *probability distribution* for these *test statistics*. The explanation involving soccer players in Chapter 2 is far more interesting…

Deleted residual: A measure of the influence of a particular case of data. It is the difference between the *adjusted predicted value* for a case and the original observed value for that case.

Density plot: Similar to a *histogram* except that rather than having a summary bar representing the frequency of scores, it shows each individual score as a dot. They can be useful for looking at the shape of a distribution of scores.

Dependent *t*-test: See *paired-samples t-test*.

Dependent variable: Another name for *outcome variable*. This name is usually associated with experimental methodology (which is the only time it really makes sense) and is used because it is the variable that is not manipulated by the experimenter and so its value depends on the variables that have been manipulated. To be honest, I just use the term *outcome variable* all the time – it makes more sense (to me) and is less confusing.

Deviance: The difference between the observed value of a variable and the value of that variable predicted by a statistical model.

Deviation contrast: A non-orthogonal *planned contrast* that compares the mean of each group (except for the first or last, depending on how the contrast is specified) to the overall mean.

DFBeta: A measure of the influence of a case on the values of b_i in a *regression model*. If we estimated a *parameter* b_i and then deleted a particular case and re-estimated the same *parameter* b_i, then the difference between these two estimates would be the DFBeta for the case that was deleted. By looking at the values of the DFBetas, it is possible to identify cases that have a large influence on the *parameters* of the model; however, the size of DFBeta will depend on the units of measurement of the *parameter*.

DFFit: A measure of the influence of a case. It is the difference between the *adjusted predicted value* and the original predicted value of a particular case. If a case is not influential then its DFFit should be zero – hence, we expect non-influential cases to have small DFFit values. However, we have the problem that this statistic depends on the units of measurement of the outcome, and so a DFFit of 0.5 will be very small if the outcome ranges from 1 to 100, but very large if the outcome varies from 0 to 1.

Diagonal: A covariance structure used in *multilevel linear models*. In this variance structure variances are assumed to be heterogeneous and all of the covariances are 0.

Dichotomous: Description of a variable that consists of only two categories (e.g., assuming you have a good definition of that point at which someone is asleep, it is a dichotomous variable because it consists of only two categories: asleep or awake).

Difference contrast: A non-orthogonal *planned contrast* that compares the mean of each condition (except the first) to the overall mean of all previous conditions combined.

Direct effect: The effect of a *predictor variable* on an *outcome variable* when a *mediator* is present in the model (cf. *indirect effect*).

Direct oblimin: A method of *oblique rotation*.

Discrete variable: A variable that can only take on certain values (usually whole numbers) on the scale.

Discriminant analysis: See discriminant function analysis.

Discriminant function analysis: This analysis identifies and describes the *discriminant function variates* of a set of variables and is useful as a follow-up test to *MANOVA* as a means of seeing how these variates allow groups of cases to be discriminated.

Discriminant function variate: A linear combination of variables created such that the differences between group means on the transformed variable are maximized. It takes the general form: $\text{Variate}_{1i} = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni}$.

Discriminant score: A score for an individual case on a particular *discriminant function variate* obtained by substituting that case's scores on the measured variables into the equation that defines the variate in question.

Dummy variables: A way of recoding a categorical variable with more than two categories into a series of variables all of which are *dichotomous* and can take on values of only 0 or 1. There are seven basic steps to create such variables: (1) count the number of groups you want to recode and subtract 1; (2) create as many new variables as the value you calculated in step 1 (these are your dummy variables); (3) choose one of your groups as a baseline (i.e., a group against which all other groups should be compared, such as a control group); (4) assign that baseline group values of 0 for all of your dummy variables; (5) for your first dummy variable, assign the value 1 to the first group that you want to compare against the baseline group (assign all other groups 0 for this variable); (6) for the second dummy variable assign the value 1 to the second group that you want to compare against the baseline group (assign all other groups 0 for this variable); (7) repeat this process until you run out of dummy variables.

Durbin–Watson test: A test for serial correlations between errors in *regression models*. Specifically, it tests whether adjacent residuals are correlated, which is useful in assessing the assumption of *independent errors*. The test statistic can vary between 0 and 4, with a value of 2 meaning that the residuals are uncorrelated. A value greater than 2 indicates a negative correlation between adjacent residuals, whereas a value below 2 indicates a positive correlation. The size of the Durbin–Watson statistic depends upon the number of predictors in the model and the number of observations. For accuracy, look up the exact acceptable values in Durbin and Watson's (1951) original paper. As a very conservative rule of thumb, values less than 1 or greater than 3 are definitely cause for concern; however, values closer to 2 may still be problematic, depending on the sample and model.

Ecological validity: Evidence that the results of a study, experiment or test can be applied, and allow inferences, to real-world conditions.

Eel: Long, snakelike, scaleless fishes that lacks pelvic fins. From the order Anguilliformes or Apodes, eels should probably not be inserted into your anus to cure constipation (or for any other reason).

Effect size: An objective and (usually) standardized measure of the magnitude of an observed effect. Measures include Cohen's $\hat{d}$, Glass's g and Pearson's correlations coefficient, r.

Empirical probability: The empirical probability is the *probability* of an *event* based on the observation of many trials. For example, if you define the collective as all men, then the empirical probability of infidelity in men will be the proportion of men who have been unfaithful while in a relationship. The probability applies to the collective and not to the individual events. You can talk about there being a 0.1 probability of men being unfaithful, but the individual men were either faithful or not, so their *individual* probability of infidelity was either 0 (they were faithful) or 1 (they were unfaithful).

Equamax: A method of *orthogonal rotation* that is a hybrid of *quartimax* and *varimax*. It is reported to behave fairly erratically (see Tabachnick & Fidell, 2012) and so is probably best avoided.

Error bar chart: A graphical representation of the mean of a set of observations that includes the 95% confidence interval of the mean. The mean is usually represented as a circle, square or rectangle at the value of the mean (or a bar extending to the value of the mean). The confidence interval is represented by a line protruding from the mean (upwards, downwards or both) to a short horizontal line representing the limits of the confidence interval. Error bars can be drawn using the standard error or standard deviation instead of the 95% confidence interval.

Error SSCP (E): The error sum of squares and cross-products matrix. This is a *sum of squares and cross-products matrix* for the error in a predictive *linear model* fitted to *multivariate* data. It represents the *unsystematic variance* and is the multivariate equivalent of the *residual sum of squares*.

Eta squared (η^2): An *effect size* measure that is the ratio of the *model sum of squares* to the *total sum of squares*. So, in essence, *the coefficient of determination* by another name. It doesn't have an awful lot going for it: not only is it biased, but also it typically measures the overall effect of an ANOVA, and effect sizes are more easily interpreted when they reflect specific comparisons (e.g., the difference between two means).

Exp(*B*): The label that SPSS applies to the *odds ratio*. It is an indicator of the change in *odds* resulting from a unit change in the predictor in *logistic regression*. If the value is greater than 1 then it indicates that as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease.

Experimental hypothesis: Synonym for *alternative hypothesis*.

Experimental research: A form of research in which one or more variables are systematically manipulated to see their effect (alone or in combination) on an *outcome variable*. This term implies that data will be able to be used to make statements about cause and effect. Compare with *cross-sectional research* and *correlational research*.

Experimentwise error rate: The probability of making a *Type I error* in an experiment involving one or more statistical comparisons when the null hypothesis is true in each case.

Extraction: A term used for the process of deciding whether a *factor* in *factor analysis* is statistically important enough to 'extract' from the data and interpret. The decision is based on the magnitude of the eigenvalue associated with the factor. See *Kaiser's criterion, scree plot*.

F_{max}: See *Hartley's F_{max}*.

***F*-statistic:** A test statistic with a known *probability distribution* (the *F*-distribution). It is the ratio of the average variability in the data that a given model can explain to the average variability unexplained by

that same model. It is used to test the overall fit of the model in *simple regression* and *multiple regression*, and to test for overall differences between group means in experiments.

Factor: Another name for an *independent variable* or *predictor* that's typically used when describing experimental designs. However, to add to the confusion, it is also used synonymously with *latent variable* in *factor analysis*.

Factor analysis: A *multivariate* technique for identifying whether the correlations between a set of observed variables stem from their relationship to one or more *latent variables* in the data, each of which takes the form of a *linear model*.

Factor matrix: General term for the *structure matrix* in *factor analysis*.

Factor loading: The *regression coefficient* of a variable for the *linear model* that describes a *latent variable* or *factor* in *factor analysis*.

Factor score: A single score from an individual entity representing their performance on some *latent variable*. The score can be crudely conceptualized as follows: take an entity's score on each of the variables that make up the factor and multiply it by the corresponding *factor loading* for the variable, then add these values up (or average them).

Factor transformation matrix, Λ: A matrix used in *factor analysis*. It can be thought of as containing the angles through which factors are rotated in factor *rotation*.

Factorial ANOVA: An analysis of variance involving two or more *independent variables* or *predictors*.

Falsification: The act of disproving a hypothesis or theory.

Familywise error rate: The probability of making a *Type I error* in any family of tests when the null hypothesis is true in each case. The 'family of tests' can be loosely defined as a set of tests conducted on the same data set and addressing the same empirical question.

Fisher's exact test: Fisher's exact test (Fisher, 1922) is not so much a test as a way of computing the exact probability of a statistic. It was designed originally to overcome the problem that with small samples the sampling distribution of the chi-square statistic deviates substantially from a chi-square distribution. It should be used with small samples.

Fit: How sexually attractive you find a statistical test. Alternatively, it's the degree to which a statistical model is an accurate representation of some observed data. (Incidentally, it's just plain *wrong* to find statistical tests sexually attractive.)

Fixed coefficient: A coefficient or model *parameter* that cannot vary over situations or contexts (cf. *random coefficient*).

Fixed effect: An effect in an experiment is said to be a fixed effect if all possible treatment conditions that a researcher is interested in are present in the experiment. Fixed effects can be generalized only to the situations in the experiment. For example, the effect is fixed if we say that we are interested only in the conditions that we had in our experiment (e.g., placebo, low dose and high dose) and we can generalize our findings only to the situation of a placebo, low dose and high dose.

Fixed intercept: A term used in *multilevel linear modelling* to denote when the intercept in the model is not free to vary across different groups or contexts (cf. *random intercept*).

Fixed slope: A term used in *multilevel linear modelling* to denote when the slope of the model is not free to vary across different groups or contexts (cf. *random slope*).

Fixed variable: A fixed variable is one that is not supposed to change over time (e.g., for most people their eye colour is a fixed variable – it never changes).

Frequency distribution: A graph plotting values of observations on the horizontal axis, and the frequency with which each value occurs in the data set on the vertical axis (a.k.a. *histogram*).

Friedman's ANOVA: A non-parametric test of whether more than two related groups differ. It is the non-parametric version of one-way *repeated-measures ANOVA*.

General linear model: A term to represent the fact that the *linear model* can encompass a range of different research designs such as multiple outcome variables (a.k.a. *MANOVA*), comparing means of categorical predictors (a.k.a. *t*-test, ANOVA), and including both categorical and continuous predictors (a.k.a. ANCOVA).

Generalization: The ability of a statistical model to say something beyond the set of observations that spawned it. If a model generalizes it is assumed that predictions from that model can be applied not just to the sample on which it is based, but to a wider population from which the sample came.

Glossary: A collection of grossly inaccurate definitions (written late at night when you really ought to be asleep) of things that you thought you understood until some evil book publisher forced you to try to define them.

Goodman and Kruskal's λ: This measures the proportional reduction in error that is achieved when membership of a category of one variable is used to predict category membership of the other variable. A value of 1 means that one variable perfectly predicts the other, whereas a value of 0 indicates that one variable in no way predicts the other.

Goodness of fit: An index of how well a model fits the data from which it was generated. It's usually based on how well the data predicted by the model correspond to the data that were actually collected.

Grand mean: The *mean* of an entire set of observations.

Grand mean centring: The transformation of a variable by taking each score and subtracting the mean of all scores (for that variable) from it (cf. *group mean centring*).

Grand variance: The *variance* within an entire set of observations.

Greenhouse–Geisser estimate: An estimate of the departure from *sphericity*. The maximum value is 1 (the data completely meet the assumption of sphericity) and the minimum is the *lower bound*. Values below 1 indicate departures from sphericity and are used to correct the *degrees of freedom* associated with the corresponding *F-statistics* by multiplying them by the value of the estimate. Some say the Greenhouse–Geisser correction is too conservative (strict) and recommend the *Huynh–Feldt correction* instead.

Group mean centring: The transformation of a variable by taking each score and subtracting from it the mean of the scores (for that variable) for the group to which that score belongs (cf. *grand mean centring*).

Growth curve: A curve that summarizes the change in some outcome over time. See *polynomial*.

HARKing: Hypothesizing after the results are known is the practice in research articles of presenting a hypothesis that was made *after* data were collected as though it were made *before* data collection.

Harmonic mean: A weighted version of the *mean* that takes account of the relationship between variance and sample size. It is calculated by summing the reciprocal of all observations, then dividing by the number of observations, n. The reciprocal of the end product is the harmonic mean:

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Hartley's F_{max}: Also known as the *variance ratio*, this is the ratio of the variances between the group with the biggest variance and the group with the smallest variance. This ratio is compared to critical values in a table published by Hartley as a test of *homogeneity of variance*. Some general rules are that with sample sizes (n) of 10 per group, an F_{max} less than 10 is more or less always going to be non-significant, with 15–20 per group the ratio needs to be less than about 5, and with samples of 30–60 the ratio should be below about 2 or 3.

Hat values: Another name for *leverage*.

HE^{-1}: This is a matrix that is functionally equivalent to the *hypothesis SSCP* divided by the *error SSCP* in *MANOVA*. Conceptually it represents the ratio of *systematic* to *unsystematic variance*, so is a *multivariate* analogue of the *F-statistic*.

Helmert contrast: A non-orthogonal *planned contrast* that compares the mean of each condition (except the last) to the overall mean of all subsequent conditions combined.

Heterogeneity of variance: The opposite of *homogeneity of variance*. This term means that the variance of one variable varies (i.e., is different) across levels of another variable.

Heteroscedasticity: The opposite of *homoscedasticity*. This occurs when the residuals at each level of the predictor variables(s) have unequal variances. Put another way, at each point along any predictor variable, the spread of residuals is different.

Hierarchical regression: A method of *multiple regression* in which the order in which predictors are entered into the regression model is determined by the researcher based on previous research: variables already known to be predictors are entered first, new variables are entered subsequently.

Histogram: A frequency distribution.

Homogeneity of covariance matrices: An assumption of some *multivariate* tests such as *MANOVA*. It is an extension of the *homogeneity of variance* assumption in *univariate* analyses. However, as well as assuming that *variances* for each *dependent variable* are the same across groups, it also assumes that relationships (*covariances*) between these dependent variables are roughly equal. It is tested by comparing the population *variance–covariance matrices* of the different groups in the analysis.

Homogeneity of regression slopes: An assumption of *analysis of covariance*. This is the assumption that the relationship between the *covariate* and *outcome variable* is constant across different treatment levels. So, if we had three treatment conditions, and if there's a positive relationship between the covariate and the outcome in one group, we assume that there is a similar-sized positive relationship between the covariate and outcome in the other two groups too.

Homogeneity of variance: The assumption that the variance of one variable is stable (i.e., relatively similar) at all levels of another variable.

Homoscedasticity: An assumption in regression analysis that the residuals at each level of the predictor variable(s) have similar variances. Put another way, at each point along any predictor variable, the spread of residuals should be fairly constant.

Hosmer and Lemeshow's R_L^2: A version of the *coefficient of determination* for logistic regression. It is a fairly literal translation in that it is the $-2LL$ for the model divided by the original $-2LL$, in other words, it's the ratio of what the model can explain compared to what there was to explain in the first place.

Hotelling–Lawley trace (T^2): A *test statistic* in *MANOVA*. It is the sum of the eigenvalues for each *discriminant function variate* of the data and so is conceptually the same as the *F-statistic* in *ANOVA*: it is the sum of the ratio of *systematic* and *unsystematic variance* (SS_M/SS_R) for each of the variates.

Huynh–Feldt estimate: An estimate of the departure from *sphericity*. The maximum value is 1 (the data completely meet the assumption of sphericity). Values below this indicate departures from

sphericity and are used to correct the *degrees of freedom* associated with the corresponding *F-statistics* by multiplying them by the value of the estimate. It is less conservative than the *Greenhouse–Geisser estimate*, but some say it is too liberal.

Hypothesis: A proposed explanation for a fairly narrow phenomenon or set of observations. It is not a guess, but an informed, theory-driven attempt to explain what has been observed. A hypothesis cannot be tested directly but must first be operationalized as predictions about variables that can be measured (see *experimental hypothesis* and *null hypothesis*).

Hypothesis SSCP (H): The hypothesis sum of squares and cross-products matrix. This is a *sum of squares and cross-products matrix* for a predictive *linear model* fitted to *multivariate* data. It represents the *systematic variance* and is the multivariate equivalent of the *model sum of squares*.

Identity matrix: A square matrix (i.e., having the same number of rows and columns) in which the diagonal elements are equal to 1, and the off-diagonal elements are equal to 0. The following are all examples:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Independence: The assumption that one data point does not influence another. When data come from people, it basically means that the behaviour of one person does not influence the behaviour of another.

Independent ANOVA: *Analysis of variance* conducted on any design in which all *independent variables* or *predictors* have been manipulated using different participants (i.e., all data come from different entities).

Independent design: An experimental design in which different treatment conditions utilize different organisms (e.g., in psychology, this would mean using different people in different treatment conditions) and so the resulting data are independent (a.k.a. between-groups or between-subjects designs).

Independent errors: For any two observations in regression the *residuals* should be uncorrelated (or independent).

Independent factorial design: An experimental design incorporating two or more *predictors* (or *independent variables*) all of which have been manipulated using different participants (or whatever entities are being tested).

Independent *t*-test: A test using the *t-statistic* that establishes whether two means collected from independent samples differ significantly.

Independent variable: Another name for a *predictor variable*. This name is usually associated with experimental methodology (which is the only time it makes sense) and is used because it is the variable that is manipulated by the experimenter and so its value does not depend on any other variables (just on the experimenter). I just use the term *predictor variable* all the time because the meaning of the term is not constrained to a particular methodology.

Index of mediation: A standardized measure of an *indirect effect*. In a mediation model, it is the *indirect effect* multiplied by the ratio of the standard deviation of the *predictor variable* to the standard deviation of the *outcome variable*.

Indirect effect: The effect of a *predictor variable* on an *outcome variable* through a *mediator* (cf. *direct effect*).

Informative prior distribution: In Bayesian statistics an informative *prior distribution* is a distribution representing your beliefs in a model *parameter* where the distribution narrows those beliefs to some degree. For example, a prior distribution that is normal with a peak at 5 and range from 2 to 8

would narrow your beliefs in a *parameter* such that you most strongly believe that its value will be 5, and you think it is impossible for the *parameter* to be less than 2 or greater than 8. As such, this distribution constrains your prior beliefs. Informative priors can vary from weakly informative (you are prepared to believe a wide range of values) to strongly informative (your beliefs are very constrained) (cf. *uninformative prior*).

Interaction effect: The combined effect of two or more *predictor variables* on an *outcome variable*. It can used to gauge *moderation*.

Interaction graph: A graph showing the means of two or more *independent variables* in which means of one variable are shown at different levels of the other variable. Usually the means are connected with lines, or are displayed as bars. These graphs are used to help understand *interaction effects*.

Interquartile range: The limits within which the middle 50% of an ordered set of observations fall. It is the difference between the value of the *upper quartile* and *lower quartile*.

Interval variable: Data measured on a scale along the whole of which intervals are equal. For example, people's ratings of this book on Amazon.com can range from 1 to 5; for these data to be interval it should be true that the increase in appreciation for this book represented by a change from 3 to 4 along the scale should be the same as the change in appreciation represented by a change from 1 to 2, or 4 to 5. Compare to *ratio variable*.

Intraclass correlation (ICC): A *correlation coefficient* that assesses the consistency between measures of the same class, that is, measures of the same thing (cf. *Pearson's correlation coefficient*, which measures the relationship between variables of a different class). Two common uses are in comparing paired data (such as twins) on the same measure, and assessing the consistency between judges' ratings of a set of objects. The calculation of these correlations depends on whether there is a measure of consistency (in which the order of scores from a source is considered but not the actual value around which the scores are anchored) or absolute agreement (in which both the order of scores and the relative values are considered), and whether the scores represent averages of many measures or just a single measure is required. This measure is also used in *multilevel linear models* to measure the dependency in data within the same context.

Jonckheere–Terpstra test: This statistic tests for an ordered pattern of medians across independent groups. Essentially it does the same thing as the *Kruskal–Wallis test* (i.e., test for a difference between the medians of the groups) but it incorporates information about whether the order of the groups is meaningful. As such, you should use this test when you expect the groups you're comparing to produce a meaningful order of medians.

Journal: In the context of academia a journal is a collection of articles on a broadly related theme, written by scientists, that report new data, new theoretical ideas or reviews/critiques of existing theories and data. Their main function is to induce learned helplessness in scientists through a complex process of self-esteem regulation using excessively harsh or complimentary peer feedback that has seemingly no obvious correlation with the actual quality of the work submitted.

Kaiser–Meyer–Olkin measure of sampling adequacy (KMO): The KMO can be calculated for individual and multiple variables and represents the ratio of the squared correlation between variables to the squared *partial correlation* between variables. It varies between 0 and 1: a value of 0 means that the sum of partial correlations is large relative to the sum of correlations, indicating diffusion in the pattern of correlations (hence, *factor analysis* is likely to be inappropriate); a value close to 1 indicates that patterns of correlations are relatively compact and so factor analysis should yield distinct and reliable factors. Values between 0.5 and 0.7 are mediocre, values between 0.7 and 0.8 are good, values between 0.8 and 0.9 are great and values above 0.9 are superb (see Kaiser & Rice, 1974).

Kaiser's criterion: A method of *extraction* in *factor analysis* based on the idea of retaining factors with associated eigenvalues greater than 1. This method appears to be accurate when the number of variables

in the analysis is less than 30 and the resulting *communalities* (after *extraction*) are all greater than 0.7, or when the sample size exceeds 250 and the average communality is greater than or equal to 0.6.

Kendall's tau: A non-parametric correlation coefficient similar to *Spearman's correlation coefficient*, but should be used in preference for a small data set with a large number of tied ranks.

Kendall's *W*: This is much the same as *Friedman's ANOVA* but is used specifically for looking at the agreement between raters. So, if, for example, we asked 10 different women to rate the attractiveness of Justin Timberlake, David Beckham and Brad Pitt we could use this test to look at the extent to which they agree. Kendall's *W* ranges from 0 (no agreement between judges) to 1 (complete agreement between judges).

Kolmogorov–Smirnov test: A test of whether a distribution of scores is significantly different from a *normal distribution*. A significant value indicates a deviation from normality, but this test is notoriously affected by large samples in which small deviations from normality yield significant results.

Kolmogorov–Smirnov *Z*: Not to be confused with the *Kolmogorov–Smirnov test* that tests whether a sample comes from a normally distributed population. This tests whether two groups have been drawn from the same population (regardless of what that population may be). It does much the same as the *Mann–Whitney test* and *Wilcoxon rank-sum test*! It tends to have better power than the Mann–Whitney test when sample sizes are less than about 25 per group.

Kruskal–Wallis test: A non-parametric test of whether more than two independent groups differ. It is the non-parametric version of one-way *independent ANOVA*.

Kurtosis: This measures the degree to which scores cluster in the tails of a frequency distribution. Kurtosis is calculated such that kurtosis in a *normal distribution* yields a value of 3. To make the measure more intuitive, SPSS (and some other packages) subtract 3 from the value so that no kurtosis is expressed as 0 and positive and negative kurtosis take on positive and negative values, respectively. A distribution with positive kurtosis (*leptokurtic*, kurtosis > 0) has too many scores in the tails and is too peaked, whereas a distribution with negative kurtosis (*platykurtic*, kurtosis < 0) has too few scores in the tails and is quite flat.

Latent variable: A variable that cannot be directly measured, but is assumed to be related to several variables that can be measured.

Leptokurtic: See *kurtosis*.

Levels of measurement: The relationship between what is being measured and the numbers obtained on a scale.

Levene's test: This tests the hypothesis that the variances in different groups are equal (i.e., the difference between the variances is zero). It basically does a one-way ANOVA on the *deviations* (i.e., the absolute value of the difference between each score and the mean of its group). A significant result indicates that the variances are significantly different – therefore, the assumption of *homogeneity of variances* has been violated. When samples sizes are large, small differences in group variances can produce a significant Levene's test. I do not recommend using this test – instead interpret statistics that have been adjusted for the degree of heterogeneity in variances.

Leverage: Leverage statistics (or hat values) gauge the influence of the observed value of the outcome variable over the predicted values. The average leverage value is $(k+1)/n$, in which k is the number of predictors in the model and n is the number of participants. Leverage values can lie between 0 (the case has no influence whatsoever) and 1 (the case has complete influence over prediction). If no cases exert undue influence over the model then we would expect all of the leverage values to be close to the average value. Hoaglin and Welsch (1978) recommend investigating cases with values greater than twice the average ($2(k + 1)/n$) and Stevens (2002) recommends using three times the average ($3(k + 1)/n$) as a cut-off point for identifying cases having undue influence.

Likelihood: The probability of obtaining a set of observations given the *parameters* of a model fitted to those observations. When using *Bayes' theorem* to test a hypothesis, the likelihood is the probability that the observed data could be produced given the hypothesis or model being considered, *p*(data|-model). It is the inverse conditional probability of the *posterior probability*. See also *marginal likelihood*.

Linear model: A statistical model that is based upon an equation of the form $\mathbf{Y} = \mathbf{BX} + \mathbf{E}$, in which **Y** is a vector containing scores from an outcome variable, **B** represents the *b*-values, **X** the predictor variables and **E** the error terms associated with each predictor. The equation can represent a solitary predictor variable (**B**, **X** and **E** are vectors) as in *simple regression* or multiple predictors (**B**, **X** and **E** are matrices) as in *multiple regression*. The key is the form of the model, which is linear (e.g., with a single predictor the equation is that of a straight line).

Line chart: A graph in which a summary statistic (usually the mean) is plotted on the *y*-axis against a categorical variable on the *x*-axis (this categorical variable could represent, for example, groups of people, different times or different experimental conditions). The value of the mean for each category is shown by a symbol, and means across categories are connected by a line. Different-coloured lines may be used to represent levels of a second categorical variable.

Logistic regression: A version of *multiple regression* in which the outcome is a *categorical variable*. If the categorical variable has exactly two categories the analysis is called *binary logistic regression*, and if the outcome has more than two categories it is called *multinomial logistic regression*.

Log-likelihood: A measure of error, or unexplained variation, in categorical models. It is based on summing the probabilities associated with the predicted and actual outcomes and is analogous to the *residual sum of squares* in multiple regression in that it is an indicator of how much unexplained information there is after the model has been fitted. Large values of the log-likelihood statistic indicate poorly fitting statistical models, because the larger the value of the log-likelihood, the more unexplained observations there are. The log-likelihood is the logarithm of the *likelihood*.

Loglinear analysis: A procedure used as an extension of the *chi-square test* to analyse situations in which we have more than two *categorical variables* and we want to test for relationships between these variables. Essentially, a *linear model* is fitted to the data that predicts expected frequencies (i.e., the number of cases expected in a given category). In this respect it is much the same as *analysis of variance* but for entirely categorical data.

Long format data: Data that are arranged such that scores on an outcome variable appear in a single column and rows represent a combination of the attributes of those scores – the entity from which the scores came, when the score was recorded, etc. In long format data, scores from a single entity can appear over multiple rows where each row represents a combination of the attributes of the score – for example, levels of an independent variable or time points at which the score was recorded (cf. *wide format data*).

Longitudinal research: A form of research in which you observe what naturally goes on in the world without directly interfering with it, by measuring several variables at multiple time points. See also *correlational research*, *cross-sectional research*.

Lower-bound estimate: The name given to the lowest possible value of the *Greenhouse–Geisser estimate* of *sphericity*. Its value is $1/(k-1)$, in which k is the number of treatment conditions.

Lower quartile: The value that cuts off the lowest 25% of the data. If the data are ordered and then divided into two halves at the median, then the lower quartile is the median of the lower half of the scores.

M-estimator: A robust measure of location. One example is the median. In some cases it is a measure of location computed after outliers have been removed; unlike a *trimmed mean*, the amount of trimming used to remove outliers is determined empirically.

Mahalanobis distances: These measure the influence of a case by examining the distance of cases from the mean(s) of the predictor variable(s). One needs to look for the cases with the highest values. It is not easy to establish a cut-off point at which to worry, although Barnett and Lewis (1978) have produced a table of critical values dependent on the number of predictors and the sample size. From their work it is clear that even with large samples ($N = 500$) and five predictors, values above 25 are cause for concern. In smaller samples ($N = 100$) and with fewer predictors (namely three), values greater than 15 are problematic, and in very small samples ($N = 30$) with only two predictors, values greater than 11 should be examined. However, for more specific advice, refer to Barnett and Lewis's (1978) table.

Main effect: The unique effect of a *predictor variable* (or *independent variable*) on an *outcome variable*. The term is usually used in the context of *ANOVA*.

Mann–Whitney test: A *non-parametric test* that looks for differences between two independent samples. That is, it tests whether the populations from which two samples are drawn have the same location. It is functionally the same as *Wilcoxon's rank-sum test*, and both tests are non-parametric equivalents of the *independent t-test*.

MANOVA: Acronym for multivariate analysis of variance**.**

Marginal likelihood (evidence): When using *Bayes' theorem* to test a hypothesis, the marginal likelihood (sometimes called evidence) is the probability of the observed data, p(data). See also *likelihood*.

Matrix: A collection of numbers arranged in columns and rows. The values within a matrix are typically referred to as *components* or *elements*.

Mauchly's test: A test of the assumption of *sphericity*. If this test is significant then the assumption of *sphericity* has not been met and an appropriate correction must be applied to the *degrees of freedom* of the *F-statistic* in *repeated-measures ANOVA*. The test works by comparing the *variance–covariance matrix* of the data to an *identity matrix*; if the variance–covariance matrix is a scalar multiple of an *identity matrix* then sphericity is met.

Maximum likelihood estimation: A way of estimating statistical *parameters* by choosing the *parameters* that make the data most likely to have happened. Imagine for a set of *parameters* that we calculated the probability (or likelihood) of getting the observed data; if this probability was high then these particular *parameters* yield a good fit to the data, but conversely if the probability was low, these *parameters* are a bad fit to our data. Maximum likelihood estimation chooses the *parameters* that maximize the probability.

McNemar's test: This tests differences between two related groups (see *Wilcoxon signed-rank test* and *sign test*), when *nominal data* have been used. It's typically used when we're looking for changes in people's scores and it compares the proportion of people who changed their response in one direction (i.e., scores increased) to those who changed in the opposite direction (scores decreased). So, this test needs to be used when we've got two related dichotomous variables.

Mean: A simple statistical model of the centre of a distribution of scores. A hypothetical estimate of the 'typical' score.

Mean squares: A measure of average variability. For every *sum of squares* (which measure the total variability) it is possible to create mean squares by dividing by the number of things used to calculate the sum of squares (or some function of it).

Measurement error: The discrepancy between the numbers used to represent the thing that we're measuring and the actual value of the thing we're measuring (i.e., the value we would get if we could measure it directly).

Median: The middle score of a set of ordered observations. When there is an even number of observations the median is the average of the two scores that fall either side of what would be the middle value.

Median test: A non-parametric test of whether samples are drawn from a population with the same median. So, in effect it does the same thing as the *Kruskal–Wallis test*. It works on the basis of producing a contingency table that is split for each group into the number of scores that fall above and below the observed median of the entire data set. If the groups are from the same population then these frequencies would be expected to be the same in all conditions (about 50% above and about 50% below).

Mediation: Perfect mediation occurs when the relationship between a *predictor variable* and an *outcome variable* can be completely explained by their relationships with a third variable. For example, taking a dog to work reduces work stress. This relationship is mediated by positive mood if (1) having a dog at work increases positive mood; (2) positive mood reduces work stress; and (3) the relationship between having a dog at work and work stress is reduced to zero (or at least weakened) when positive mood is included in the model.

Mediator: A variable that reduces the size and/or direction of the relationship between a *predictor variable* and an *outcome variable* (ideally to zero) and is associated statistically with both.

Meta-analysis: This is a statistical procedure for assimilating research findings. It is based on the simple idea that we can take effect sizes from individual studies that research the same question, quantify the observed effect in a standard way, and then combine these effects to get a more accurate idea of the true effect in the population.

Method of least squares: A method of estimating *parameters* (such as the *mean*, or a regression coefficient) that is based on minimizing the *sum of squared errors*. The *parameter* estimate will be the value, out of all of those possible, that has the smallest *sum of squared errors*.

Mixed ANOVA: Analysis of variance used for a mixed design.

Mixed design: An experimental design incorporating two or more *predictors* (or *independent variables*) at least one of which has been manipulated using different participants (or whatever entities are being tested) and at least one of which has been manipulated using the same participants (or entities). Also known as a split-plot design because Fisher developed ANOVA for analysing agricultural data involving 'plots' of land containing crops.

Mixed normal distribution: A normal-looking distribution that is contaminated by a small proportion of scores from a different distribution. These distributions are not normal and have too many scores in the tails (i.e., at the extremes). The effect of these heavy tails is to inflate the estimate of the population variance. This, in turn, makes significance tests lack power.

Mode: The most frequently occurring score in a set of data.

Model sum of squares: A measure of the total amount of variability for which a model can account. It is the difference between the *total sum of squares* and the *residual sum of squares*.

Moderation: Moderation occurs when the relationship between two variables changes as a function of a third variable. For example, the relationship between watching horror films (predictor) and feeling scared at bedtime (outcome) might increase as a function of how vivid an imagination a person has (moderator).

Moderator: A variable that changes the size and/or direction of the relationship between two other variables.

Monte Carlo method: A term applied to the process of using data simulations to solve statistical problems. Its name comes from the use of Monte Carlo roulette tables to generate 'random' numbers in the pre-computer age. Karl Pearson, for example, purchased copies of *Le Monaco*, a weekly Paris periodical that published data from the Monte Carlo casinos' roulette wheels. He used these data as pseudo-random numbers in his statistical research.

Moses extreme reactions: A non-parametric test that compares the variability of scores in two groups, so it's a bit like a non-parametric *Levene's test*.

Multicollinearity: A situation in which two or more variables are very closely linearly related.

Multilevel linear model: A linear model (just like regression, ANCOVA, ANOVA, etc.) in which the hierarchical structure of the data is explicitly considered. In this analysis *parameters* can be fixed (as in regression and ANOVA) but also random (i.e., free to vary across different contexts at a higher level of the hierarchy). This means that for each *parameter* there is a fixed component but also an estimate of how much the *parameter* varies across contexts (see *fixed coefficient*, *random coefficient*).

Multimodal: Description of a distribution of observations that has more than two *modes*.

Multinomial logistic regression: *Logistic regression* in which the outcome variable has more than two categories.

Multiple *R*: The multiple correlation coefficient. It is the correlation between the observed values of an outcome and the values of the outcome predicted by a *multiple regression* model.

Multiple regression: An extension of *simple regression* in which an outcome is predicted by a linear combination of two or more predictor variables. The form of the model is:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni} + \varepsilon_i$$

in which the outcome is denoted by Y, and each predictor is denoted by X. Each predictor has a *parameter* coefficient b associated with it, and b_0 is the value of the outcome when all predictors are zero.

Multivariate: Means 'many variables' and is usually used when referring to analyses in which there is more than one *outcome variable* (*MANOVA*, *principal component analysis*, etc.).

Multivariate analysis of variance: A family of tests that extend the basic *analysis of variance* to situations in which more than one *outcome variable* has been measured.

Multivariate normality: An extension of a normal distribution to multiple variables. It is a *probability distribution* of a set of variables $\mathbf{v}' = [\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_n]$ given by:

$$f(\mathbf{v}') = 2\pi^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{v}-\boldsymbol{\mu})\right\}$$

in which $\boldsymbol{\mu}$ is the vector of means of the variables, and $\boldsymbol{\Sigma}$ is the *variance–covariance* matrix. If that made any sense to you then you're cleverer than I am.

Nagelkerke's R^2_N: A version of the *coefficient of determination* for logistic regression. It is a variation on *Cox and Snell's* R^2_{CS} which overcomes the problem that this statistic has of not being able to reach its maximum value.

Negative skew: See *skew*.

Nominal variable: Where numbers merely represent names. For example, the numbers on sports players' shirts: a player with the number 1 on her back is not necessarily worse than a player with a 2 on her back. The numbers have no meaning other than denoting the type of player (full back, centre forward, etc.).

Noniles: A type of *quantile*; they are values that split the data into nine equal parts. They are commonly used in educational research.

Non-parametric tests: A family of statistical procedures that do not rely on the restrictive assumptions of parametric tests. In particular, they do not assume that the sampling distribution is normally distributed.

Normal distribution: A *probability distribution* of a random variable that is known to have certain properties. It is perfectly symmetrical (has a *skew* of 0), and has a *kurtosis* of 3 or excess kurtosis of 0.

Null hypothesis: The reverse of the *experimental hypothesis*, it states that your prediction is wrong and the predicted effect doesn't exist.

Numeric variables: Variables involving numbers.

Oblique rotation: A method of *rotation* in *factor analysis* that allows the underlying factors to be correlated.

Odds: The probability of an event occurring divided by the probability of that event not occurring.

Odds ratio: The ratio of the *odds* of an event occurring in one group compared to another. So, for example, if the odds of dying after writing a glossary are 4, and the odds of dying after not writing a glossary are 0.25, then the odds ratio is 4/0.25 = 16. This means that the *odds* of dying if you write a glossary are 16 times higher than if you don't. An odds ratio of 1 would indicate that the *odds* of a particular outcome are equal in both groups.

Omega squared: An *effect size* measure associated with ANOVA that is less biased than *eta squared*. It is a (sometimes hideous) function of the *model sum of squares* and the *residual sum of squares* and isn't actually much used because it measures the overall effect of the ANOVA and so can't be interpreted in a meaningful way. In all other respects it's great, though.

One-tailed test: A test of a directional hypothesis. For example, the hypothesis 'the longer I write this glossary, the more I want to place my editor's genitals in a starved crocodile's mouth' requires a one-tailed test because I've stated the direction of the relationship. I would generally advise against using them because of the temptation to interpret interesting effects in the opposite direction to that predicted (see also *two-tailed test*).

Open science: A movement to make the process, data and outcomes of scientific research freely available to everyone.

Ordinal variable: Data that tell us not only that things have occurred, but also the order in which they occurred. These data tell us nothing about the differences between values. For example, gold, silver and bronze medals are ordinal: they tell us that the gold medallist was better than the silver medallist, but they don't tell us how much better (was gold a lot better than silver, or were gold and silver very closely competed?).

Ordinary least squares (OLS): A method of *regression* in which the *parameters* of the model are estimated using the *method of least squares*.

Orthogonal: This means perpendicular (at right angles) to something. It tends to be equated to *independence* in statistics because of the connotation that perpendicular *linear models* in geometric space are completely independent (one is not influenced by the other).

Orthogonal rotation: A method of *rotation* in *factor analysis* that keeps the underlying factors independent (i.e., not correlated).

Outcome variable: A variable whose values we are trying to predict from one or more *predictor variables*.

Outlier: An observation or observations very different from most others. Outliers bias statistics (e.g., the mean) and their standard errors and confidence intervals.

Overdispersion: When the observed variance is bigger than expected from the logistic regression model. Like leprosy, you don't want it.

***p*-curve:** A curve summarizing the frequency distribution of *p*-values you'd expect to see in published research. On a graph that shows the value of the *p*-value on the horizontal axis against the frequency

(or proportion) on the vertical axis, the *p*-curve is the line reflecting how frequently (or proportionately) each value of *p* should occur for a given *effect size*.

_p_-hacking: Research practices that lead to selective reporting of significant *p*-values. Some examples of *p*-hacking are: (1) trying multiple analyses and reporting only the one that yields significant results; (2) stopping collecting data at a point other than when the predetermined sample size is reached; (3) deciding whether to include data based on the effect they have on the *p*-value; (4) including (or excluding) variables in an analysis based on how they affect the *p*-value; (5) measuring multiple *outcome* or *predictor* variables but reporting only those for which the effects are significant; (6) merging groups of variables or scores to yield significant results; and (7) transforming, or otherwise manipulating, scores to yield significant *p*-values.

Paired-samples *t*-test: A test using the *t-statistic* that establishes whether two means collected from the same sample (or related observations) differ significantly.

Pairwise comparisons: Comparisons of pairs of means.

Parameter: A very difficult thing to describe. When you fit a statistical model to your data, that model will consist of *variables* and *parameters*: variables are measured constructs that vary across entities in the sample, whereas *parameters* describe the relations between those variables in the population. In other words, they are constants believed to represent some fundamental truth about the measured variables. We use sample data to estimate the likely value of *parameters* because we don't have direct access to the population. Of course, it's not quite as simple as that.

Parametric test: A test that requires data from one of the large catalogue of distributions that statisticians have described. Normally this term is used for parametric tests based on the *normal distribution*, which require four basic assumptions that must be met for the test to be accurate: a normally distributed sampling distribution (see *normal distribution*), *homogeneity of variance*, *interval* or *ratio data*, and *independence*.

Parsimony: In a scientific context, parsimony refers to the idea that simpler explanations of a phenomenon are preferable to complex ones. This idea relates to Ockham's (or Occam's if you prefer) razor, which is a phrase referring to the principle of 'shaving' away unnecessary assumptions or explanations to produce less complex theories. In statistical terms, parsimony tends to refer to a general heuristic that models be kept as simple as possible – in other words, not including variables that don't have real explanatory benefit.

Part correlation: Another name for a *semi-partial correlation*.

Partial correlation: A measure of the relationship between two variables while 'adjusting' the effect of one or more additional variables has on both.

Partial eta squared (partial η^2): A version of *eta squared* that is the proportion of variance that a variable explains when excluding other variables in the analysis. Eta squared is the proportion of total variance explained by a variable, whereas partial eta squared is the proportion of variance that a variable explains that is not explained by other variables.

Pattern matrix: A matrix in *factor analysis* containing the *regression coefficients* for each variable on each *factor* in the data. See also *structure matrix*.

Pearson's correlation coefficient: Pearson's product-moment correlation coefficient, to give it its full name, is a *standardized* measure of the strength of relationship between two variables. It can take any value from –1 (as one variable changes, the other changes in the opposite direction by the same amount), through 0 (as one variable changes the other doesn't change at all), to +1 (as one variable changes, the other changes in the same direction by the same amount).

Peer Reviewers' Openness Initiative: An initiative to get scientists to commit to the principles of *open science* when they act as expert reviewers for journals. Signing up is a pledge to review submissions

only if the data, stimuli, materials, analysis scripts and so on are made publicly available (unless there is a good reason, such as a legal requirement, not to).

Percentiles: A type of *quantile*; they are values that split the data into 100 equal parts.

Perfect collinearity: This exists when at least one predictor in a *regression model* is a perfect linear combination of the others (the simplest example being two predictors that are perfectly correlated – they have a correlation coefficient of 1).

Phi: A measure of the strength of association between two *categorical variables*. Phi is used with 2 × 2 *contingency tables* (tables which have two categorical variables and each variable has only two categories). Phi is a variant of the *chi-square test,* χ^2:

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

in which N is the total number of observations.

Pillai–Bartlett trace (*V*): A *test statistic* in *MANOVA*. It is the sum of the proportion of explained variance on the *discriminant function variates* of the data. As such, it is similar to the ratio of SS_M/SS_T.

Pilot fish (*Naucrates ductor*)**:** A carnivorous fish in the Carangidae family known for congregating around larger more impressive beings (e.g., sharks) and feeding parasitically from their bodies. A bit like Courtney Love.

Planned comparisons: Another name for *planned contrasts*.

Planned contrasts: A set of comparisons between group means that are constructed before any data are collected. These are theory-led comparisons and are based on the idea of partitioning the variance created by the overall effect of group differences into gradually smaller portions of variance. These tests have more power than *post hoc tests*.

Platykurtic: See *kurtosis***.**

Point-biserial correlation: A standardized measure of the strength of relationship between two variables when one of the two variables is *dichotomous*. The point-biserial correlation coefficient is used when the dichotomy is a discrete, or true, dichotomy (i.e., one for which there is no underlying continuum between the categories). An example of this is pregnancy: you can be either pregnant or not, there is no in-between.

Polychotomous logistic regression: Another name for *multinomial logistic regression*.

Polynomial: A posh name for a *growth curve* or trend over time. If *time* is our predictor variable, then any polynomial is tested by including a variable that is the predictor to the power of the order of polynomial that we want to test: a linear trend is tested by *time* alone, a quadratic or second-order polynomial is tested by including a predictor that is $time^2$, for a fifth-order polynomial we need a predictor of $time^5$ and for an nth-order polynomial we would have to include $time^n$ as a predictor.

Polynomial contrast: A contrast that tests for trends in the data. In its most basic form it looks for a linear trend (i.e., that the group means increase proportionately).

Population: In statistical terms this usually refers to the collection of units (be they people, plankton, plants, cities, despondent authors, etc.) to which we want to generalize a set of findings or a statistical model.

Positive skew: See *skew*.

***Post hoc* tests:** A set of comparisons between group means that were not thought of before data were collected. Typically these tests involve comparing the means of all combinations of pairs of groups. To compensate for the number of tests conducted, each test uses a strict criterion for significance. As

such, they tend to have less power than *planned contrasts*. They are usually used for exploratory work for which no firm hypotheses were available on which to base planned contrasts.

Posterior distribution: A distribution of *posterior probabilities*. This distribution should contain our subjective beliefs about a *parameter* or hypothesis after considering the data. The posterior distribution can be used to ascertain a value of the posterior probability (usually by examining some measure of where the peak of the distribution lies or a *credible interval*).

Posterior odds: The ratio of *posterior probability* for one hypothesis to another. In Bayesian hypothesis testing the posterior odds are the ratio of the probability of the alternative hypothesis given the data, *p*(alternative|data), to the probability of the null hypothesis given the data, *p*(null|data).

Posterior probability: When using *Bayes' theorem* to test a hypothesis, the posterior probability is our belief in a hypothesis or model *after* we have considered the data, *p*(model|data). This is the value that we are usually interested in knowing. It is the inverse conditional probability of the *likelihood*.

Power: The ability of a test to detect an effect of a particular size (a value of 0.8 is a good level to aim for).

P-P plot: Short for 'probability–probability plot'. A graph plotting the cumulative probability of a variable against the cumulative probability of a particular distribution (often a normal distribution). Like a *Q-Q plot*, if values fall on the diagonal of the plot then the variable shares the same distribution as the one specified. Deviations from the diagonal show deviations from the distribution of interest.

Practice effect: This refers to the possibility that participants' performance in a task may be influenced (positively or negatively) if they repeat the task because of familiarity with the experimental situation and/or the measures being used.

Predictive validity: A form of *criterion validity* where there is evidence that scores from an instrument predict external measures (recorded at a different point in time) conceptually related to the measured construct.

Predicted value: The value of an outcome variable based on specific values of the predictor variable or variables being placed into a statistical model.

Predictor variable: A variable that is used to try to predict values of another variable known as an *outcome variable*.

Pre-registration: A term referring to the practice of making all aspects of your research process (rationale, hypotheses, design, data processing strategy, data analysis strategy) publicly available before data collection begins. This can be done in a *registered report* in an academic journal, or more informally (e.g., on a public website such as the Open Science Framework). The aim is to encourage adherence to an agreed research protocol, thus discouraging threats to the validity of scientific results such as *researcher degrees of freedom*.

Principal component analysis (PCA): A *multivariate* technique for identifying the linear components of a set of variables.

Prior distribution: A distribution of *prior probabilities*. This distribution should contain our subjective beliefs about a *parameter* or hypothesis before, or prior to, considering the data. The prior distribution can be an *informative prior* or an *uninformative prior*.

Prior odds: The ratio of the probability of one hypothesis/model to a second. In Bayesian hypothesis testing, the prior odds are the probability of the *alternative hypothesis*, *p*(alternative), divided by the probability of the *null hypothesis*, *p*(null). The prior odds should reflect your belief in the alternative hypothesis relative to the null *before* you look at the data.

Prior probability: When using *Bayes' theorem* to test a hypothesis, the prior probability is our belief in a hypothesis or model before, or prior to, considering the data, *p*(model). See also *posterior probability*, *likelihood*, *marginal likelihood*.

Probability density function (PDF): The function that describes the probability of a random variable taking a certain value. It is the mathematical function that describes the *probability distribution*.

Probability distribution: A curve describing an idealized *frequency distribution* of a particular variable from which it is possible to ascertain the probability with which specific values of that variable will occur. For categorical variables it is simply a formula yielding the probability with which each category occurs.

Promax: A method of *oblique rotation* that is computationally faster than *direct oblimin* and so useful for large data sets.

Publication bias: The fact that articles published in scientific journals tend to over-represent positive findings. This can be because (1) non-significant findings are less likely to be published; (2) scientists don't submit their non-significant results to journals; (3) scientists selectively report their results to focus on significant findings and exclude non-significant ones; and (4) scientists capitalize on *research degrees of freedom* to shed their results in the most favourable light possible.

Q-Q plot: Short for 'quantile–quantile plot'. A graph plotting the *quantiles* of a variable against the quantiles of a particular distribution (often a normal distribution). Like a *P-P plot*, if values fall on the diagonal of the plot then the variable shares the same distribution as the one specified. Deviations from the diagonal show deviations from the distribution of interest.

Quadratic trend: If the means in ordered conditions are connected with a line then a quadratic trend is shown by one change in the direction of this line (e.g., the line is curved in one place); the line is, therefore, U-shaped. There must be at least three ordered conditions.

Qualitative methods: Extrapolating evidence for a theory from what people say or write (contrast with *quantitative methods*).

Quantiles: Values that split a data set into equal portions. *Quartiles*, for example, are a special case of quantiles that split the data into four equal parts. Similarly, *percentiles* are points that split the data into 100 equal parts and *noniles* are points that split the data into nine equal parts (you get the general idea).

Quantitative methods: Inferring evidence for a theory through measurement of variables that produce numeric outcomes (cf. *qualitative methods*).

Quartic trend: If the means in ordered conditions are connected with a line then a quartic trend is shown by three changes in the direction of this line. There must be at least five ordered conditions.

Quartiles: A generic term for the three values that cut an ordered data set into four equal parts. The three quartiles are known as the first or *lower quartile*, the second quartile (or *median*) and the third or *upper quartile*.

Quartimax: A method of *orthogonal rotation*. It attempts to maximize the spread of factor loadings for a variable across all *factors*. This often results in lots of variables loading highly on a single *factor*.

Random coefficient: A coefficient or model *parameter* that is free to vary over situations or contexts (cf. *fixed coefficient*).

Random effect: An effect is said to be random if the experiment contains only a sample of possible treatment conditions. Random effects can be generalized beyond the treatment conditions in the experiment. For example, the effect is random if we say that the conditions in our experiment (e.g., placebo, low dose and high dose) are only a sample of possible conditions (perhaps we could have tried a very high dose). We can generalize this random effect beyond just placebos, low doses and high doses.

Random intercept: A term used in *multilevel linear modelling* to denote when the intercept in the model is free to vary across different groups or contexts (cf. *fixed intercept*).

Random slope: A term used in *multilevel linear modelling* to denote when the slope of the model is free to vary across different groups or contexts (cf. *fixed slope*).

Random variable: A random variable is one that varies over time (e.g., your weight is likely to fluctuate over time).

Random variance: Variance that is unique to a particular variable but not reliably so.

Randomization: The process of doing things in an unsystematic or random way. In the context of experimental research the word usually applies to the random assignment of participants to different treatment conditions.

Range: The range of scores is the value of the smallest score subtracted from the highest score. It is a measure of the dispersion of a set of scores. See also *variance*, *standard deviation* and *interquartile range*.

Rank-biserial correlation: A measure of *effect size* that quantifies the proportion of favourable to unfavourable evidence for a hypothesis. Used as an effect size measure for non-parametric statistical tests, it is the difference between the sum of positive and negative ranks divided by the sum of all ranks.

Ranking: The process of transforming raw scores into numbers that represent their position in an ordered list of those scores. The raw scores are ordered from lowest to highest and the lowest score is assigned a rank of 1, the next highest score is assigned a rank of 2, and so on.

Ratio variable: An *interval variable* but with the additional property that ratios are meaningful. For example, people's ratings of this book on Amazon.com can range from 1 to 5; for these data to be ratio not only must they have the properties of *interval variables*, but in addition a rating of 4 should genuinely represent someone who enjoyed this book twice as much as someone who rated it as 2. Likewise, someone who rated it as 1 should be half as impressed as someone who rated it as 2. These variables should also have a meaningful zero value.

Registered report: An article in a journal usually outlining an intended research process (rationale, hypotheses, design, data processing strategy, data analysis strategy). The report is reviewed by relevant expert scientists, ensuring that authors get useful feedback before data collection. If the protocol is accepted by the journal editor it typically comes with a guarantee to publish the findings no matter what they are, thus reducing *publication bias* and discouraging *researcher degrees of freedom* aimed at achieving significant results.

Regression coefficient: Another name for a *parameter* in a linear model, See b_i and β_i.

Regression model: See multiple regression and simple regression**.**

Regression line: A line on a scatterplot representing the *regression model* of the relationship between the two variables plotted.

Related design: Another name for a *repeated-measures design***.**

Related factorial design: An experimental design incorporating two or more *predictors* (or *independent variables*) all of which have been manipulated using the same participants (or whatever entities are being tested).

Reliability: The ability of a measure to produce consistent results when the same entities are measured under different conditions.

Repeated contrast: A non-orthogonal *planned contrast* that compares the mean in each condition (except the first) to the mean of the preceding condition.

Repeated-measures ANOVA: An *analysis of variance* conducted on any design in which the *independent variable* (*predictor*) or *variables* (*predictors*) have all been measured using the same participants in all conditions.

Repeated-measures design: An experimental design in which different treatment conditions utilize the same organisms (i.e., in psychology, this would mean the same people take part in all experimental conditions) and so the resulting data are related (a.k.a. related design or within-subject designs).

Researcher degrees of freedom: The analytic decisions a researcher makes that potentially influence the results of the analysis. Some examples are: when to stop data collection, which control variables to include in the statistical model, and whether to exclude cases from the analysis.

Residual: The difference between the value a model predicts and the value observed in the data on which the model is based. Basically, an error. When the residual is calculated for each observation in a data set the resulting collection is referred to as the *residuals*.

Residuals: See *residual*.

Residual sum of squares: A measure of the variability that cannot be explained by the model fitted to the data. It is the total squared *deviance* between the observations, and the value of those observations predicted by whatever model is fitted to the data.

Reverse Helmert contrast: Another name for a *difference contrast*.

Robust methods: A term applied to a family of procedures to estimate statistics that produce unbiased estimates, confidence intervals and significance tests even when the normal assumptions of the statistic are not met.

Rotation: A process in *factor analysis* for improving the interpretability of factors. In essence, an attempt is made to transform the *factors* that emerge from the analysis in such a way as to maximize *factor loadings* that are already large, and minimize factor loadings that are already small. There are two general approaches: *orthogonal rotation* and *oblique rotation*.

Roy's largest root: A *test statistic* in *MANOVA*. It is the eigenvalue for the first *discriminant function variate* of a set of observations. So, it is the same as the *Hotelling–Lawley trace* but for the first variate only. It represents the proportion of explained variance to unexplained variance (SS_M/SS_R) for the first discriminant function.

Sample: A smaller (but hopefully representative) collection of units from a *population* used to determine truths about that population (e.g., how a given population behaves in certain conditions).

Sampling distribution: The *probability distribution* of a statistic. We can think of this as follows: if we take a *sample* from a *population* and calculate some statistic (e.g., the *mean*), the value of this statistic will depend somewhat on the sample we took. As such the statistic will vary slightly from sample to sample. If, hypothetically, we took lots and lots of samples from the population and calculated the statistic of interest we could create a frequency distribution of the values we get. The resulting distribution is what the sampling distribution represents: the distribution of possible values of a given statistic that we could expect to get from a given population.

Sampling variation: The extent to which a statistic (the mean, median, *t*, *F*, etc.) varies in samples taken from the same population.

Saturated model: A model that perfectly fits the data and, therefore, has no error. It contains all possible *main effects* and *interactions* between variables.

Scatterplot: A graph that plots values of one variable against the corresponding value of another variable (and the corresponding values of a third variable can also be included on a 3-D scatterplot).

Scree plot: A graph plotting each *factor* in a *factor analysis* (*X*-axis) against its associated eigenvalue (*Y*-axis). It shows the relative importance of each factor. This graph has a very characteristic shape (there is a sharp descent in the curve followed by a tailing off), and the point of inflexion of this curve

is often used as a means of *extraction*. With a sample of more than 200 participants, this provides a fairly reliable criterion for *extraction* (Stevens, 2002)

Second quartile: Another name for the *median*.

Semi-partial correlation: A measure of the relationship between two variables while adjusting for the effect that one or more additional variables have on one of those variables. If we call our variables x and y, it gives us a measure of the variance in y that x alone shares.

Shapiro–Wilk test: A test of whether a distribution of scores is significantly different from a *normal distribution*. A significant value indicates a deviation from normality, but this test is notoriously affected by large samples in which small deviations from normality yield significant results.

Shrinkage: The loss of predictive power of a regression model if the model has been derived from the population from which the sample was taken, rather than the sample itself.

Šidák correction: A slightly less conservative variant of the *Bonferroni correction*.

Sign test: This tests whether two related samples are different. It does the same thing as the *Wilcoxon signed-rank test*. Differences between the conditions are calculated and the sign of this difference (positive or negative) is analysed because it indicates the direction of differences. The magnitude of change is completely ignored (unlike in Wilcoxon's test, where the rank tells us something about the relative magnitude of change), and for this reason it lacks *power*. However, its computational simplicity makes it a nice party trick if ever anyone drunkenly accosts you needing some data quickly analysed without the aid of a computer – doing a sign test in your head really impresses people. Actually it doesn't, they just think you're a sad gimboid.

Simple contrast: A non-orthogonal *planned contrast* that compares the mean in each condition to the mean of either the first or last condition, depending on how the contrast is specified.

Simple effects analysis: This analysis looks at the effect of one *independent variable* (categorical *predictor variable*) at individual levels of another *independent variable*.

Simple regression: A *linear model* in which one variable or outcome is predicted from a single predictor variable. The model takes the form:

$$Y_i = b_0 + b_1X_i + \varepsilon_i$$

in which Y is the outcome variable, X is the predictor, b_1 is the *parameter* associated with the predictor and b_0 is the value of the outcome when the predictor is zero.

Simple slopes analysis: An analysis that looks at the relationship (i.e., the *simple regression*) between a *predictor variable* and an *outcome variable* at low, mean and high levels of a third (*moderator*) variable.

Singularity: Mathematically, a singular matrix is one that has a *determinant* of 0. In statistics singularity describes variables that are perfectly correlated (i.e., the *correlation coefficient* is 1 or –1), because such variables would result in a correlation matrix with a determinant of 0.

Skew: A measure of the symmetry of a *frequency distribution*. Symmetrical distributions have a skew of 0. When the frequent scores are clustered at the lower end of the distribution and the tail points towards the higher or more positive scores, the value of skew is positive. Conversely, when the frequent scores are clustered at the higher end of the distribution and the tail points towards the lower more negative scores, the value of skew is negative.

Spearman's correlation coefficient: A standardized measure of the strength of relationship between two variables that does not rely on the assumptions of a *parametric test*. It is *Pearson's correlation coefficient* performed on data that have been converted into ranked scores.

Sphericity: A less restrictive form of *compound symmetry* which assumes that the variances of the differences between data taken from the same participant (or other entity being tested) are equal. This assumption is most commonly found in *repeated-measures ANOVA* but applies only where there are more than two points of data from the same participant. See also *Greenhouse–Geisser correction*, *Huynh–Feldt correction*.

Split-half reliability: A measure of *reliability* obtained by splitting items on a measure into two halves (in some random fashion) and obtaining a score from each half of the scale. The correlation between the two scores, corrected to take account of the fact the correlations are based on only half of the items, is used as a measure of reliability. There are two popular ways to do this. Spearman (1910) and Brown (1910) developed a formula that takes no account of the standard deviation of items:

$$r_{sh} = \frac{2r_{12}}{1 + r_{12}}$$

in which r_{12} is the correlation between the two halves of the scale. Flanagan (1937) and Rulon (1939), however, proposed a measure that does account for item variance:

$$r_{sh} = \frac{4r_{12} \times s_1 \times s_2}{s_T^2}$$

in which s_1 and s_2 are the standard deviations of each half of the scale, and s_T^2 is the variance of the whole test. See Cortina (1993) for more details.

Square matrix: A *matrix* that has an equal number of columns and rows.

Standard deviation: An estimate of the average variability (spread) of a set of data measured in the same units of measurement as the original data. It is the square root of the *variance*.

Standard error: The standard deviation of the *sampling distribution* of a statistic. For a given statistic (e.g., the *mean*) it tells us how much variability there is in this statistic across *samples* from the same *population*. Large values, therefore, indicate that a statistic from a given sample may not be an accurate reflection of the population from which the sample came.

Standard error of differences: If we were to take several pairs of samples from a population and calculate their means, then we could also calculate the difference between their means. If we plotted these differences between sample means as a *frequency distribution*, we would have the *sampling distribution* of differences. The standard deviation of this sampling distribution is the standard error of differences. As such it is a measure of the variability of differences between sample means.

Standard error of the mean (SE): The *standard error* associated with the mean. Did you really need a glossary entry to work that out?

Standardization: The process of converting a variable into a standard unit of measurement. The unit of measurement typically used is *standard deviation* units (see also *z-scores*). Standardization allows us to compare data when different units of measurement have been used (we could compare weight measured in kilograms to height measured in inches).

Standardized: See *standardization*.

Standardized DFBeta: A *standardized* version of *DFBeta*. These standardized values are easier to use than DFBeta because universal cut-off points can be applied. Stevens (2002) suggests looking at cases with absolute values greater than 2.

Standardized DFFit: A *standardized* version of *DFFit*.

Standardized residuals: The *residuals* of a model expressed in standard deviation units. Standardized residuals with an absolute value greater than 3.29 (actually, we usually just use 3) are cause for concern because in an average sample a value this high is unlikely to happen by chance; if more than 1% of our observations have standardized residuals with an absolute value greater than 2.58 (we usually just say 2.5) there is evidence that the level of error within our model is unacceptable (the model is a fairly poor fit to the sample data); and if more than 5% of observations have standardized residuals with an absolute value greater than 1.96 (or 2 for convenience) then there is also evidence that the model is a poor representation of the actual data.

Stepwise regression: A method of *multiple regression* in which variables are entered into the model based on a statistical criterion (the *semi-partial correlation* with the *outcome variable*). Once a new variable is entered into the model, all variables in the model are assessed to see whether they should be removed.

String variables: Variables involving words (i.e., letter strings). Such variables could include responses to open-ended questions such as 'How much do you like writing glossary entries?'; the response might be 'About as much as I like placing my ballbag on hot coals'.

Structure matrix: A matrix in *factor analysis* containing the *correlation coefficients* for each variable on each *factor* in the data. When *orthogonal rotation* is used this is the same as the *pattern matrix*, but when oblique rotation is used these matrices are different.

Studentized deleted residual: A measure of the influence of a particular case of data. This is a standardized version of the *deleted residual*.

Studentized residuals: A variation on *standardized residuals*. Studentized residuals are the *unstandardized residual* divided by an estimate of its standard deviation that varies point by point. These residuals have the same properties as the *standardized residuals* but usually provide a more precise estimate of the error variance of a specific case.

Sum of squared errors: Another name for the *sum of squares*.

Sum of squares (SS): An estimate of total variability (spread) of a set of observations around a *parameter* (such as the *mean*). First the *deviance* for each score is calculated, and then this value is squared. The SS is the sum of these squared deviances.

Sum of squares and cross-products matrix (SSCP matrix): A *square matrix* in which the diagonal elements represent the *sum of squares* for a particular variable, and the off-diagonal elements represent the *cross-products* between pairs of variables. The SSCP matrix is basically the same as the *variance–covariance matrix*, except that the SSCP matrix expresses variability and between-variable relationships as total values, whereas the variance–covariance matrix expresses them as average values.

Suppressor effect: A situation where a predictor has a significant effect but only when another variable is held constant.

Syntax: Predefined written commands that instruct SPSS what you would like it to do (writing 'bugger off and leave me alone' doesn't seem to work …).

Systematic variation: Variation due to some genuine effect (be that the effect of an experimenter doing something to all of the participants in one sample but not in other samples, or natural variation between sets of variables). We can think of this as variation that can be explained by the model that we've fitted to the data.

***t*-statistic:** A *test statistic* with a known *probability distribution* (the *t*-distribution). In the context of the *linear model* it is used to test whether a *b*-value is significantly different from zero; in the context of experimental work this *b*-value represents the difference between two means and so *t* is a test of whether the difference between those means is significantly different from zero. See also *paired-samples t-test* and *independent t-test*.

***Tertium quid*:** The possibility that an apparent relationship between two variables is actually caused by the effect of a third variable on them both (often called the 'third-variable problem').

Test of excess success (TES): A procedure designed for identifying sets of results within academic articles that are 'too good to be true'. For an article reporting multiple scientific studies examining the same effect, the test computes (based on the size of effect being measured and sample size of the studies) the probability that you would get significant results for all of the studies. If this probability is low it is highly unlikely that the researcher would get these results and the results appear 'too good to be true', implying *p-hacking* (Francis, 2013). It is noteworthy that the TES is not universally accepted as testing what it sets out to test (e.g., Morey, 2013).

Test–retest reliability: The ability of a measure to produce consistent results when the same entities are tested at two different points in time.

Test statistic: A statistic for which we know how frequently different values occur. The observed value of such a statistic is typically used to test *hypotheses*.

Theory: Although it can be defined more formally, a theory is a hypothesized general principle or set of principles that explain known findings about a topic and from which new hypotheses can be generated. Theories have typically been well substantiated by repeated testing

Tolerance: Tolerance statistics measure *multicollinearity* and are simply the reciprocal of the *variance inflation factor* (1/VIF). Values below 0.1 indicate serious problems, although Menard (1995) suggests that values below 0.2 are worthy of concern.

Total SSCP (T): The total sum of squares and cross-products matrix. This is a *sum of squares and cross-products matrix* for an entire set of observations. It is the *multivariate* equivalent of the *total sum of squares*.

Total sum of squares: A measure of the total variability within a set of observation. It is the total squared *deviance* between each observation and the overall mean of all observations.

Transformation: The process of applying a mathematical function to all observations in a data set, usually to correct some distributional abnormality such as *skew* or *kurtosis*.

Trimmed mean: A statistic used in many *robust tests*. It is a mean calculated using trimmed data. For example, a 20% trimmed mean is a mean calculated after the top and bottom 20% of ordered scores have been removed. Imagine we had 20 scores representing the annual income of students (in thousands), rounded to the nearest thousand: 0, 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 40. The mean income is 5 (£5000), which is biased by an outlier. A 10% trimmed mean will remove 10% of scores from the top and bottom of ordered scores before the mean is calculated. With 20 scores, removing 10% of scores involves removing the top and bottom two scores. This gives us: 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, the mean of which is 3.44. The mean depends on a symmetrical distribution to be accurate, but a trimmed mean produces accurate results even when the distribution is not symmetrical. There are more complex examples of robust methods such as the *bootstrap*.

Two-tailed test: A test of a non-directional hypothesis. For example, the hypothesis 'writing this glossary has some effect on what I want to do with my editor's genitals' requires a two-tailed test because it doesn't suggest the direction of the relationship. See also *one-tailed test*.

Type I error: Occurs when we believe that there is a genuine effect in our population, when in fact there isn't.

Type II error: Occurs when we believe that there is no effect in the population, when in fact there is.

Uninformative prior distribution: In Bayesian statistics an uninformative *prior distribution* is a distribution representing your beliefs in a model *parameter* where the distribution assigns equal

probability to all values of the model/*parameter*. For example, a prior distribution that is uniform across all potential values of a *parameter* suggests that you are prepared to believe that the *parameter* can take on any value with equal probability. As such, this distribution does not constrain your prior beliefs (cf. *informative prior*).

Unique factor: A *factor* that affects only one of many measured *variables* and, therefore, cannot explain the *correlations* between those variables.

Unique variance: Variance that is specific to a particular variable (i.e., is not shared with other variables). We tend to use the term 'unique variance' to refer to variance that can be reliably attributed to only one measure, otherwise it is called *random variance*.

Univariate: This means 'one variable' and is usually used to refer to situations in which only one *outcome variable* has been measured (*ANOVA*, *t-tests*, *Mann–Whitney tests*, etc.).

Unstructured: A covariance structure used in *multilevel linear modelling*. This covariance structure is completely general. Covariances are assumed to be completely unpredictable: they do not conform to a systematic pattern.

Unstandardized residuals: The *residuals* of a model expressed in the units in which the original outcome variable was measured.

Unsystematic variation: This is variation that isn't due to the effect in which we're interested (so could be due to natural differences between people in different samples such as differences in intelligence or motivation). We can think of this as variation that can't be explained by whatever model we've fitted to the data.

Upper quartile: The value that cuts off the highest 25% of ordered scores. If the scores are ordered and then divided into two halves at the median, then the upper quartile is the median of the top half of the scores.

Validity: Evidence that a study allows correct inferences about the question it was aimed to answer or that a test measures what it set out to measure conceptually. See also *content validity*, *criterion validity*.

Variables: Anything that can be measured and can differ across entities or across time.

Variable view: There are two ways to view the contents of the *data editor* window. The variable view allows you to define properties of the variables for which you wish to enter data. See also *data view*.

Variance: An estimate of average variability (spread) of a set of data. It is the sum of squares divided by the number of values on which the sum of squares is based minus 1.

Variance components: A covariance structure used in *multilevel linear modelling*. This covariance structure is very simple and assumes that all random effects are independent and that the variances of random effects are assumed to be the same and sum to the variance of the outcome variable.

Variance–covariance matrix: A square matrix (i.e., same number of columns and rows) representing the variables measured. The diagonals represent the *variances* within each variable, whereas the off-diagonals represent the *covariances* between pairs of variables.

Variance inflation factor (VIF): A measure of *multicollinearity*. The VIF indicates whether a predictor has a strong linear relationship with the other predictor(s). Myers (1990) suggests that a value of 10 is a good value at which to worry. Bowerman and O'Connell (1990) suggest that if the average VIF is greater than 1, then multicollinearity may be biasing the regression model.

Variance ratio: See *Hartley's* F_{max}.

Variance sum law: States that the variance of a sum of or difference between two independent random variables is equal to the sum of their variances.

Varimax: A method of *orthogonal rotation*. It attempts to maximize the dispersion of *factor loadings* within *factors*. Therefore, it tries to load a smaller number of variables highly onto each factor, resulting in more interpretable clusters of factors.

VIF: See variance inflation factor.

Wald statistic: A *test statistic* with a known *probability distribution* (a *normal distribution* or a *chi-square distribution* when squared) that is used to test whether the *b* coefficient for a predictor in a *logistic regression* model is significantly different from zero. It is analogous to the *t-statistic* in a *regression model* in that it is simply the *b* coefficient divided by its standard error. The Wald statistic is inaccurate when the *parameter* (*b*) is large, because the standard error tends to become inflated, resulting in the Wald statistic being underestimated.

Wald–Wolfowitz runs: Another variant on the *Mann–Whitney test*. Scores are rank-ordered as in the Mann–Whitney test, but rather than analysing the ranks, this test looks for 'runs' of scores from the same group within the ranked order. Now, if there's no difference between groups then obviously ranks from the two groups should be randomly interspersed. However, if the groups are different then one should see more ranks from one group at the lower end, and more ranks from the other group at the higher end. By looking for clusters of scores in this way the test can determine if the groups differ.

Weight: A number by which something (usually a variable in statistics) is multiplied. The weight assigned to a variable determines the influence that variable has within a mathematical equation: large weights give the variable a lot of influence.

Weighted least squares: A method of *regression* in which the *parameters* of the model are estimated using the *method of least squares* but observations are weighted by some other variable. Often they are weighted by the inverse of their *variance* to combat *heteroscedasticity*.

Welch's *F*: A version of the *F*-statistic designed to be accurate when the assumption of *homogeneity of variance* has been violated. Not to be confused with the squelch test which is where you shake your head around after writing statistics books to see if you still have a brain.

Wide format data: Data that are arranged such that scores from a single entity appear in a single row and levels of independent or predictor variables are arranged over different columns**.** As such, in designs with multiple measurements of an outcome variable within a case the *outcome variable* scores will be contained in multiple columns each representing a level of an independent variable, or a time point at which the score was observed. Columns can also represent attributes of the score or entity that are fixed over the duration of data collection, such as participant sex or employment status (cf. *long format data*).

Wilcoxon's rank-sum test: A *non-parametric test* that looks for differences between two independent samples. That is, it tests whether the populations from which two samples are drawn have the same location. It is functionally the same as the *Mann–Whitney test*, and both tests are non-parametric equivalents of the *independent t-test*.

Wilcoxon signed-rank test: A *non-parametric test* that looks for differences between two related samples. It is the non-parametric equivalent of the *related t-test*.

Wilks's lambda (Λ): A *test statistic* in *MANOVA*. It is the product of the unexplained variance on each of the *discriminant function variates*, so it represents the ratio of error variance to total variance (SS_R/SS_T) for each variate.

Winsorizing: A transformation of the data aimed at reducing the impact of extreme scores and *outliers*. When data are winsorized, scores outside of a certain threshold are replaced with values at the threshold. For example, if we set the threshold as 0.1 (or 10%) then we order the scores in ascending order and replace the bottom 10% of scores with the value of the score at the threshold (i.e., at 10%

along the distribution). Similarly, at the top end of the distribution of scores we replace the largest 10% of scores with the value of the score at the threshold (i.e., 90% along the distribution).

Within-subject design: Another name for a *repeated-measures design*.

Writer's block: Something I suffer from a lot. It's when you can't think of any decent examples and so end up talking about sperm the whole time. Seriously, look at this book, it's all sperm this, sperm that, quail sperm, human sperm. Frankly, I'm amazed donkey sperm didn't get in there somewhere. Oh, it just did.

Yates's continuity correction: An adjustment made to the *chi-square test* when the *contingency table* is 2 rows by 2 columns (i.e., there are two categorical variables both of which consist of only two categories). In large samples the adjustment makes little difference and is slightly dubious anyway (see Howell, 2012).

***z*-score:** The value of an observation expressed in standard deviation units. It is calculated by taking the observation, subtracting from it the mean of all observations, and dividing the result by the standard deviation of all observations. By converting a distribution of observations into *z*-scores a new distribution is created that has a mean of 0 and a standard deviation of 1.

REFERENCES

Aesthetic Society. (2020). *Aesthetic plastic surgery national database: Statistics 2020*. The Aesthetic Society. https://cdn.theaestheticsociety.org/media/statistics/aestheticplasticsurgerynationaldatabank-2020stats.pdf

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage.

Algina, J., & Olejnik, S. F. (1984). Implementing the Welch-James procedure with factorial designs. *Educational and Psychological Measurement*, *44*, 39–48.

American Psychological Association. (2019). *Publication manual of the American Psychological Association* (7th ed.). American Psychological Association.

Anderson, C. A., & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological Science*, *12*, 353–359.

Arrindell, W. A., & van der Ende, J. (1985). An empirical test of the utility of the observer-to-variables ratio in factor and components analysis. *Applied Psychological Measurement*, *9*, 165–178.

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. http://dx.doi.org/10.1037/met0000200

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, *35*, 73–80.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617.

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*(1), 158–175. https://doi.org/10.3758/s13428-011-0123-7

Barcikowski, R. S., & Robey, R. R. (1984). Decisions in single group repeated measures analysis: Statistical tests and three computer packages. *The American Statistician*, *38*, 148–150.

Bargman, R. E. (1970). *Interpretation and use of a generalized discriminant function*. University of North Carolina Press.

Barnard, G. A. (1963). Ronald Aylmer Fisher, 1890–1962: Fisher's contributions to mathematical statistics. *Journal of the Royal Statistical Society, Series A (General)*, *126*, 162–166.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data*. Wiley.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research – Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models.

arXiv:1506.04967 [stat]. http://arxiv.org/abs/1506.04967

Beckham, A. S. (1929). Is the Negro happy? A psychological analysis. *Journal of Abnormal and Social Psychology*, *24*, 186–190. https://doi.org/10.1037/h0072938

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389–396. https://doi.org/10.1037/1082-989X.10.4.389

Bell, A., Jones, K., & Fairbrother, M. (2018). Understanding and misunderstanding group mean centering: A commentary on Kelley et al.'s dangerous practice. *Quality & Quantity*, *52*(5), 2031–2036. https://doi.org/10.1007/s11135-017-0593-5

Belsey, D. A., Kuh, E., & Welsch, R. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley.

Bemelman, M., & Hammacher, E. R. (2005). Rectal impalement by pirate ship: A case report. *Injury Extra*, *36*, 508–510.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., Boeck, P. D., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*, 1–12.

Bernard, P., Gervais, S. J., Allen, J., Campomizzi, S., & Klein, O. (2012). Integrating sexual objectification with object versus person recognition: The sexualized-body-inversion hypothesis. *Psychological Science*, *23*, 469–471.

Berry, W. D. (1993). *Understanding regression assumptions*. Sage.

Bishop, D. V. M., & Thompson, P. A. (2016). Problems in using *p*-curve analysis and text-mining to detect rate of *p*-hacking and evidential value. *PeerJ*, *4*. https://doi.org/10.7717/peerj.1715

Board, B. J., & Fritzon, K. (2005). Disordered personalities at work. *Psychology, Crime & Law*, *11*, 17–32.

Bock, R. D. (1975). *Multivariate statistical methods in behavioural research*. McGraw-Hill.

Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, *46*, 241–255.

Bolker, B. M. (2015). Linear and generalized linear mixed models. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), *Ecological statistics: Contemporary theory and application* (pp. 309–333). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199672547.003.0014

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*, *65*(1), 23–28. https://doi.org/10.1007/BF02294183

Boot, W. R., Blakely, D. P., & Simons, D. J. (2011). Do action video games improve perception and cognition? *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00226

Bowerman, B. L., & O'Connell, R. T. (1990). *Linear statistical models: An applied approach* (2nd ed.). Duxbury.

Bowerman, B. L., & O'Connell, R. T. (2000). *Linear statistical models: An applied approach* (2nd ed.). Duxbury.

Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance*. Sage.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117. https://doi.org/10.1016/j.jarmac.2018.09.005

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behaviour of some statistics which test the equality of several means. *Technometrics*, *16*, 129–132.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. https://doi.org/10.1006/jmps.1999.1279

Bruns, S. B., & Ioannidis, J. P. A. (2016). *P*-curve and *p*-hacking in observational research. *PLoS*

One, *11*. https://doi.org/10.1371/journal.pone.0149144

Budescu, D. V. (1982). The power of the *F* test in normal populations with heterogeneous variances. *Educational and Psychological Measurement*, *42*, 609–616.

Budescu, D. V., & Appelbaum, M. I. (1981). Variance stabilizing transformations and the power of the F test. *Journal of Educational Statistics*, *6*, 55–74.

Caruso, J. C., & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for spearman's rho. *Educational and Psychological Measurement*, *57*(4), 637–654. https://doi.org/10.1177/0013164497057004009

Cattell, R. B. (1966a). *The scientific analysis of personality*. Aldine.

Cattell, R. B. (1966b). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Çetinkaya, H., & Domjan, M. (2006). Sexual fetishism in a quail (*Coturnix japonica*) model system: Test of reproductive success. *Journal of Comparative Psychology*, *120*, 427–432. https://doi.org/10.1037/0735-7036.120.4.427

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2. https://doi.org/10.1016/j.cortex.2015.03.022

Chamorro-Premuzic, T., Furnham, A., Christopher, A. N., Garwood, J., & Martin, N. (2008). Birds of a feather: Students' preferences for lecturers' personalities as predicted by their own personality and learning approaches. *Personality and Individual Differences*, *44*, 965–976.

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures*. Sage.

Chen, X., Wang, X. Y., Yang, D., & Chen, Y. G. (2014). The moderating effect of stimulus attractiveness on the effect of alcohol consumption on attractiveness ratings. *Alcohol and Alcoholism*, *49*, 515–519. https://doi.org/10.1093/alcalc/agu026

Chen, X. Z., Luo, Y., Zhang, J. J., Jiang, K., Pendry, J. B., & Zhang, S. A. (2011). Macroscopic invisibility cloaking of visible light. *Nature Communications*, *2*. https://doi.org/17610.1038/ncomms1176

Clarke, D. L., Buccimazza, I., Anderson, F. A., & Thomson, S. R. (2005). Colorectal foreign bodies. *Colorectal Disease*, *7*, 98–103.

Claxton, A., O'Rourke, N., Smith, J. Z., & DeLongis, A. (2012). Personality traits and marital satisfaction within enduring relationships: An intra-couple discrepancy approach. *Journal of Social and Personal Relationships*, *29*, 375–396. https://doi.org/10.1177/0265407511431183

Cliff, N. (1987). *Analyzing multivariate data*. Harcourt Brace Jovanovich.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*, 426–443.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Coldwell, J., Pike, A., & Dunn, J. (2006). Household chaos – links with parenting and child behaviour. *Journal of Child Psychology and Psychiatry*, *47*(11), 1116–1122. https://doi.org/10.1111/j.1469-7610.2006.01655.x

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin*, *115*, 465–474.

Collier, R. O., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. *Psychometrika*, *32*, 339–352.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Erlbaum.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman & Hall.

Cook, S. A., Rosser, R., & Salmon, P. (2006). Is cosmetic surgery an effective psychotherapeutic intervention? A systematic review of the evidence. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *59*, 1133–1151.

Cooper, C. L., Sloan, S. J., & Williams, S. (1988). *Occupational Stress Indicator management guide*. NFER-Nelson.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology*, *17*, 136–137.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Cox, D. R., & Snell, D. J. (1989). *The analysis of binary data (2nd edition)*. Chapman & Hall.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, *45*, 215–233. https://doi.org/10.1016/s0167-9473(02)00366-3

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G., & Calin-Jageman, R. J. (2017). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge. https://www.routledge.com/Introduction-to-the-New-Statistics-Estimation-Open-Science-and-Beyond/Cumming-Calin-Jageman/p/book/9781138825529

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–574. https://doi.org/10.1177/0013164401614002

Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika*, *21*(3), 287–290. https://doi.org/10.1007/BF02289138

Dai, X. C., Dong, P., & Jia, J. S. (2014). When does playing hard to get increase romantic attraction? *Journal of Experimental Psychology-General*, *143*, 521–526. https://doi.org/10.1037/a0032989

Daniels, E. (2012). Sexy versus strong: What girls and women think of female athletes. *Journal of Applied Developmental Psychology*, *33*, 79–90. https://doi.org/10.1016/j.appdev.2011.12.002

Davey, G. C. L., Startup, H. M., Zara, A., MacDonald, C. B., & Field, A. P. (2003). The perseveration of checking thoughts and mood–as–input hypothesis. *Journal of Behavior Therapy and Experimental Psychiatry*, *34*(2), 141–160. https://doi.org/10.1016/S0005-7916(03)00035-1

Davidson, M. L. (1972). Univariate versus multivariate tests in repeated-measures experiments. *Psychological Bulletin*, *77*, 446–452.

Davies, P., Surridge, J., Hole, L., & Munro-Davies, L. (2007). Superhero-related injuries in paediatrics: A case series. *Archives of Disease in Childhood*, *92*, 242–243. https://doi.org/10.1136/adc.2006.109793

De Groot, A. D. (1956). The meaning of 'significance' for different types of research [Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]. *Acta Psychologica*, *148*, 188–194.

DeCoster, J., Gallucci, M., & Iselin, A.-M. R. (2011). Best practices for using median splits, artificial categorization, and their continuous alternatives. *Journal of Experimental Psychopathology*, *2*, 197–209. https://doi.org/10.5127/jep.008310

DeCoster, J., Iselin, A.-M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods*, *14*, 349–366. https://doi.org/10.1037/a0016956

Di Falco, A., Ploschner, M., & Krauss, T. F. (2010). Flexible metamaterials at visible wavelengths. *New Journal of Physics*, *12*. https://doi.org/11300610.1088/1367-2630/12/11/113006

Dickersin, K., Min, Y.-I., & Meinert, C. L. (1992). Factors influencing publication of research results: Follow-up of applications submitted to two institutional review boards. *Journal of the American Medical Association*, *267*, 374–378.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290. https://doi.org/10.1177/1745691611406920

Docherty, M., Beardslee, J., Byrd, A. L., Yang, V. J. H., & Pardini, D. (2019). Developmental trajectories of interpersonal callousness from childhood to adolescence as predictors of antisocial behavior and psychopathic features in young adulthood. *Journal of*

Abnormal Psychology, *128*(7), 700–709. http://dx.doi.org/10.1037/abn0000449

Domjan, M., Blesbois, E., & Williams, J. (1998). The adaptive significance of sexual conditioning: Pavlovian control of sperm release. *Psychological Science*, *9*, 411–415.

Donaldson, T. S. (1968). Robustness of the *F*-test to errors of both kinds and the correlation between the numerator and denominator of the *F*-ratio. *Journal of the American Statistical Association*, *63*, 660–676.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Dunteman, G. E. (1989). *Principal components analysis*. Sage.

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression, II. *Biometrika*, *30*, 159–178.

Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap*. Chapman & Hall/CRC. https://doi.org/10.1201/9780429246593

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics*, *34*(2), 447–456. https://doi.org/10.1214/aoms/1177704156

Enders, C. K. (2022). *Applied missing data analysis* (2nd ed.). Guilford.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138. https://doi.org/10.1037/1082-989x.12.2.121

Eriksson, S.-G., Beckham, D., & Vassell, D. (2004). Why are the English so shit at penalties? A review. *Journal of Sporting Ineptitude*, *31*, 231–1072.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within- and between-subjects manipulations of the independent variable. *Psychological Bulletin*, *84*, 212–219.

Eysenck, H. J. (1953). *The structure of human personality*. Wiley.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One*, *4*. https://doi.org/10.1371/journal.pone.0005738

Fanelli, D. (2010a). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS One*, *5*. https://doi.org/10.1371/journal.pone.0010271

Fanelli, D. (2010b). 'Positive' results increase down the hierarchy of the sciences. *PLoS One*, *5*. https://doi.org/10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904. https://doi.org/10.1007/s11192-011-0494-7

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, *18*(10), 850–855. https://doi.org/10.1111/j.1467-9280.2007.01990.x

Feng, L., Gwee, X., Kua, E. H., & Ng, T. P. (2010). Cognitive function and tea consumption in community dwelling older Chinese in Singapore. *Journal of Nutrition Health & Aging*, *14*, 433–438.

Fesmire, F. M. (1988). Termination of intractable hiccups with digital rectal massage. *Annals of Emergency Medicine*, *17*, 872.

Field, A. P. (1998). A bluffer's guide to sphericity. *Newsletter of the Mathematical, Statistical and Computing Section of the British Psychological Society*, *6*, 13–22.

Field, A. P. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. Sage.

Field, A. P. (2005a). Intraclass correlation. In B. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 948–954). Wiley.

Field, A. P. (2005b). Sir Ronald Aylmer Fisher. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 658–659). Wiley.

Field, A. P. (2006). The behavioral inhibition system and the verbal information pathway to children's fears. *Journal of Abnormal Psychology*, *115*, 742–752. https://doi.org/10.1037/0021-843x.115.4.742

Field, A. P. (2010). Teaching statistics. In D. Upton & A. Trapp (Eds.), *Teaching Psychology in Higher Education* (pp. 134–163). Wiley-Blackwell.

Field, A. P. (2014). *Skills in mathematics and statistics in psychology and tackling transition* (p. 44) [Report]. The Higher Education Academy. https://www.advance-he.ac.uk/

knowledge-hub/skills-mathematics-and-statistics-psychology-and-tackling-transition

Field, A. P. (2016). *An adventure in statistics: The reality enigma*. Sage.

Field, A. P. (2024). *Discovering statistics using R and RStudio* (2nd ed.). Sage.

Field, A. P., & Davey, G. C. L. (1999). Reevaluating evaluative conditioning: A nonassociative explanation of conditioning effects in the visual evaluative conditioning paradigm. *Journal of Experimental Psychology-Animal Behavior Processes*, *25*, 211–224. https://doi.org/10.1037/0097-7403.25.2.211

Field, A. P., & Hole, G. J. (2003). *How to design and report experiments*. Sage.

Field, A. P., & Moore, A. C. (2005). Dissociating the effects of attention and contingency awareness on evaluative conditioning effects in the visual paradigm. *Cognition & Emotion*, *19*, 217–243. https://doi.org/10.1080/02699930441000292

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19–38. https://doi.org/10.1016/j.brat.2017.05.013

Fieller, E. C., Hartley, H. O., & Pearson, E. S. (1957). Tests for rank correlation coefficients. I. *Biometrika*, *44*(3/4), 470–481. JSTOR. https://doi.org/10.2307/2332878

Fienberg, S. E., Stigler, S. M., & Tanur, J. M. (2007). The William Kruskal Legacy: 1919–2005. *Statistical Science*, *22*, 255–261.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of *P*. *Journal of the Royal Statistical Society*, *85*, 87–94.

Fisher, R. A. (1925a). *Statistical methods, experimental design, and scientific inference*. Oxford University Press.

Fisher, R. A. (1925b). *Statistical methods for research workers*. Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Hafner.

Flanagan, J. C. (1937). A proposed procedure for increasing the efficiency of objective tests. *Journal of Educational Psychology*, *28*, 17–21.

Francis, G. (2014a). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, *21*, 1180–1187. https://doi.org/10.3758/s13423-014-0601-x

Francis, G. (2014b). Too much success for recent groundbreaking epigenetic experiments. *Genetics*, *198*, 449–451. https://doi.org/10.1534/genetics.114.163998

Francis, G., Tanzman, J., & Matthews, W. J. (2014). Excess success for psychology articles in the journal Science. *PLoS One*, *9*. https://doi.org/10.1371/journal.pone.0114255

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675–701.

Gallup, G. G. J., Burch, R. L., Zappieri, M. L., Parvez, R., Stockwell, M., & Davis, J. A. (2003). The human penis as a semen displacement device. *Evolution and Human Behavior*, *24*, 277–289. https://doi.org/10.1016/S1090-5138(03)00016-3

Games, P. A. (1983). Curvilinear transformations of the dependent variable. *Psychological Bulletin*, *93*, 382–387.

Games, P. A. (1984). Data transformations, power, and skew: A rebuttal to Levine and Dunlap. *Psychological Bulletin*, *95*, 345–347.

Games, P. A., & Lucas, P. A. (1966). Power of the analysis of variance of independent groups on non-normal and normally transformed data. *Educational and Psychological Measurement*, *26*, 311–327.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, *97*, 310–316.

Girden, E. R. (1992). *ANOVA: Repeated measures*. Sage.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*, 237–288.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). John Wiley & Sons.

Gönen, M., Johnson, W. O., Lu, Y. G., & Westfall, P. H. (2005). The Bayesian two-sample *t* test. *The American Statistician*, *59*, 252–257. https://doi.org/10.1198/000313005x55233

Graham, J. M., Guthrie, A. C., & Thompson, B. (2003). Consequences of not interpreting

structure coefficients in published CFA research: A reminder. *Structural Equation Modeling*, *10*, 142–153.

Grayson, D. (2004). Some myths and legends in quantitative psychology. *Understanding Statistics*, *3*, 101–134.

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science*, *18*(1), 88–94. https://doi.org/10.1111/j.1467-9280.2007.01853.x

Green, C. S., & Bavelier, D. (2012). Learning, attentional control, and action video games. *Current Biology*, *22*(6), R197–R206. https://doi.org/10.1016/j.cub.2012.02.012

Green, C. S., Pouget, A., & Bavelier, D. (2010). Improved probabilistic inference as a general learning mechanism with action video games. *Current Biology*, *20*(17), 1573–1579. https://doi.org/10.1016/j.cub.2010.07.040

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95–112.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337--350.

Greenwald, A. G. (1975). Consequences of prejudice against null hypothesis. *Psychological Bulletin*, *82*, 1–19.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450. https://doi.org/10.1037/1082-989X.6.4.430

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*, 265–275.

Ha, T., Overbeek, G., & Engels, R. C. M. E. (2010). Effects of attractiveness and social status on dating desire in heterosexual adolescents: An experimental study. *Archives of Sexual Behavior*, *39*, 1063–1071. https://doi.org/10.1007/s10508-009-9561-z

Hakstian, A. R., Roed, J. C., & Lind, J. C. (1979). Two-sample T2 procedure and the assumption of homogeneous covariance matrices. *Psychological Bulletin*, *86*, 1255–1263.

Hall, J., & Sammons, P. M. (2014). Mediation, moderation, and interaction: Definitions, discrimination and (some) means of testing. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 267–286). Sense.

Hamilton, B. L. (1977). An empirical-investigation of effects of heterogeneous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, *37*, 701–712.

Hardy, M. A. (1993). *Regression with dummy variables*. Sage.

Harman, B. H. (1976). *Modern factor analysis* (3rd ed.). University of Chicago Press.

Harris, R. J. (1975). *A primer of multivariate statistics*. Academic Press.

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. (2016). Distributions of *p*-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*. https://doi.org/10.7717/peerj.1935

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). The Guilford Press.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*, 709–722.

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, *41*, 924–936.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of P-hacking in science. *PloS Biology*, *13*. https://doi.org/10.1371/journal.pbio.1002106

Hedges, L. V. (1984). Estimation of effect size under non-random sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*, 61–85.

Hill, C., Abraham, C., & Wright, D. B. (2007). Can theory-based messages in combination with cognitive prompts promote exercise in classroom settings? *Social Science & Medicine*, *65*, 1049–1058.

Hoaglin, D., & Welsch, R. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, *32*, 17–22.

Hoagwood, K. E., Acri, M., Morrissey, M., & Peth-Pierce, R. (2017). Animal-assisted therapies for youth with or at risk for mental health problems: A systematic review. *Applied Developmental Science*, *21*, 1–13. https://doi.org/10.1080/10888691.2015.1134267

Hoddle, G., Batty, D., & Ince, P. (1998). How not to take penalties in important soccer matches. *Journal of Ineptitude*, *1*, 1–2.

Hodgson, R., Cole, A., & Young, A. (2012). The name of the game: Why can't people called Ashley score from a penalty kick? *Sporting Weakness Review*, *24*(6), 574–581.

Hoffmann, F., Musolf, K., & Penn, D. J. (2012). Spectrographic analyses reveal signals of individuality and kinship in the ultrasonic courtship vocalizations of wild house mice. *Physiology & Behavior*, *105*, 766–771. https://doi.org/10.1016/j.physbeh.2011.10.011

Hollingsworth, H. H. (1980). An analytical investigation of the effects of heterogeneous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, *40*, 611–618.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. Wiley.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Howell, D. C. (2012). *Statistical methods for psychology* (8th ed.). Wadsworth.

Huber, P. J. (1967). The behavior of maximum likelihood estimation under nonstandard conditions. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analysis. *Psychological Bulletin*, *105*, 302–308.

Hughes, J. P., Marice, H. P., & Gathright, J. B. (1976). Method of removing a hollow object from the rectum. *Diseases of the Colon & Rectum*, *19*, 44–45.

Hume, D. (1739). *A treatise of human nature (ed. L. A. Selby-Bigge)*. Clarendon Press, 1965.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomised block and split-plot designs. *Journal of Educational Statistics*, *1*, 69–82.

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, *50*(4), 361–365.

IBM Corp. (2022). *IBM SPSS for Windows* (Version 29). IBM Corp.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press,.

Joel, S., & Charlot, N. (2022). Dealbreakers, or dealbenders? Capturing the cumulative effects of partner information on mate choice. *Journal of Experimental Social Psychology*, *101*, 104328. https://doi.org/10.1016/j.jesp.2022.104328

Johns, S. E., Hargrave, L. A., & Newton-Fisher, N. E. (2012). Red is not a proxy signal for female genitalia in humans. *PLoS One*, *7*, e34669. https://doi.org/10.1371/journal.pone.0034669

Johnson, P. C. D., Barry, S. J. E., Ferguson, H. M., & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*, *6*(2), 133–142. https://doi.org/10.1111/2041-210X.12306

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their applications to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis, I: Artificial data. *Applied Statistics*, *21*, 160–173.

Jolliffe, I. T. (1986). *Principal component analysis*. Springer.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, *41*, 133–145.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in evaluation research. *Evaluation Research*, *5*, 602–619.

Julien, D., O'Connor, K. P., & Aardema, F. (2007). Intrusive thoughts, obsessions, and appraisals in obsessive-compulsive disorder: A critical review. *Clinical Psychology Review*, *27*, 366–383. https://doi.org/10.1016/j.cpr.2006.12.004

Kain, M. P., Bolker, B. M., & McCoy, M. W. (2015). A practical guide and power analysis for GLMMs: Detecting among treatment variation in random effects. *PeerJ*, *3*, e1226. https://doi.org/10.7717/peerj.1226

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.

Kaiser, H. F. (1970). A second-generation little jiffy. *Psychometrika*, *35*, 401–415.

Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika*, *30*, 1–14. https://doi.org/10.1007/bf02289743

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark 4. *Educational and Psychological Measurement*, *34*, 111–117. https://doi.org/10.1177/001316447403400115

Kanazawa, S. (2007). Beautiful parents have more daughters: A further implication of the generalized Trivers–Willard hypothesis. *Journal of Theoretical Biology*, *244*, 133–140.

Kass, R. A., & Tinsley, H. E. A. (1979). Factor analysis. *Journal of Leisure Research*, *11*, 120–138.

Kellett, S., Clarke, S., & McGill, P. (2008). Outcomes from psychological assessment regarding recommendations for cosmetic surgery. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *61*, 512–517.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. http://dx.doi.org/10.1037/a0028086

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis*, 13.

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, *3*, 11.IT.3.1. https://doi.org/10.2466/11.IT.3.1

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Keselman, H. J., & Keselman, J. C. (1988). Repeated measures multiple comparison procedures: Effects of violating multisample sphericity in unbalanced designs. *Journal of Educational Statistics*, *13*, 215–226.

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, *54*, 351–353. https://doi.org/10.1037/h0046737

King, B. M., Rosopa, P. J., & Minium, E. W. (2011). *Statistical reasoning in the behavioral sciences*. (6th ed.). John Wiley & Sons.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). Routledge.

Klockars, A. J., & Sax, G. (1986). *Multiple comparisons*. Sage.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Sage.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*, 1–21.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t*-test. *Journal of Experimental Psychology-General*, *142*, 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS and STAN* (2nd ed.). Academic Press.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583–621.

Lacourse, E., Claes, M., & Villeneuve, M. (2001). Heavy metal music and adolescent suicidal risk. *Journal of Youth and Adolescence*, *30*, 321–332.

Lakens, D. (2015). On the challenges of drawing conclusions from *p*-values just below 0.05. *PeerJ*, *3*. https://doi.org/10.7717/peerj.1142

Lambert, N. M., Negash, S., Stillman, T. F., Olmstead, S. B., & Fincham, F. D. (2012). A love that doesn't last: Pornography consumption and weakened commitment to one's romantic partner. *Journal of Social and Clinical Psychology*, *31*, 410–438. https://doi.org/10.1521/jscp.2012.31.4.410

LeBeau, B. (2022). *simglm: Simulate models based on the generalized linear model* (0.8.9). https://CRAN.R-project.org/package=simglm

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. R. (2013). The life of p: 'Just significant' results are on the rise. *Quarterly Journal of Experimental Psychology*, *66*, 2303–2309. https://doi.org/10.1080/17470218.2013.863371

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, *88*, 1242–1249.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*, 187–193.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and*

statistics: Essays in honor of Harold Hotelling (pp. 278–292). Stanford University Press.

Levine, D. W., & Dunlap, W. P. (1982). Power of the *F* test with skewed data: Should one transform or not? *Psychological Bulletin*, *92*, 272–280.

Levine, D. W., & Dunlap, W. P. (1983). Data transformation, power, and skew: A rejoinder to Games. *Psychological Bulletin*, *93*, 596–599.

Levy, K. N., Johnson, B. N., Clouthier, T. L., Scala, J. W., & Temes, C. M. (2015). An attachment theoretical framework for personality disorders. *Canadian Psychology–Psychologie Canadienne*, *56*, 197–207. https://doi.org/10.1037/cap0000025

Lo, S. F., Wong, S. H., Leung, L. S., Law, I. C., & Yip, A. W. C. (2004). Traumatic rectal perforation by an eel. *Surgery*, *135*, 110–111. https://doi.org/10.1016/S0039-6060(03)00076-X

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review*, *1*, 476–490.

Lombardi, C. M., & Hurlbert, S. H. (2009). Misprescription and misuse of one-tailed tests. *Austral Ecology*, *34*, 447–468. https://doi.org/10.1111/j.1442-9993.2009.01946.x

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*, 217–224. https://doi.org/10.2307/2685594

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305.

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, *72*, 336–337.

Lucke, J. F. (2005). The α and the ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement*, *29*(1), 65–81. https://doi.org/10.1177/0146621604270882

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, *7*, 263–269.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Erlbaum.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*(1), 49–55.

Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221080396. https://doi.org/10.1177/25152459221080396

Mair, P., & Wilcox, R. (2019). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01246-w

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50–60.

Marzillier, S. L., & Davey, G. C. L. (2005). Anxiety and disgust: Evidence for a unidirectional relationship. *Cognition & Emotion*, *19*, 729–750. https://doi.org/10.1080/02699930441000436

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology*, *65*, 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Massar, K., Buunk, A. P., & Rempt, S. (2012). Age differences in women's tendency to gossip are mediated by their mate value. *Personality and Individual Differences*, *52*, 106–109.

Mather, K. (1951). R. A. Fisher's *Statistical Methods for Research Workers*: An appreciation. *Journal of the American Statistical Association*, *46*, 51–54.

Matthews, R. C., Domjan, M., Ramsey, M., & Crews, D. (2007). Learning effects on sperm competition and reproductive fitness. *Psychological Science*, *18*, 758–762.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. *Journal of Educational Statistics*, *5*, 269–287.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Wadsworth.

McDonald, P. T., & Rosenthal, D. (1977). An unusual foreign body in the rectum – A baseball: Report of a case. *Diseases of the Colon & Rectum*, *20*, 56–57.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman & Hall/CRC Press. https://www.crcpress.com/Statistical-Rethinking-A-Bayesian-Course-with-Examples-in-R-and-STAN/McElreath/p/book/9780367139919

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods*, *11*, 386–401.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., & Yarkoni, T. (2016). How open science helps researchers succeed. *eLife*, *5*. https://doi.org/10.7554/eLife.16800

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. http://dx.doi.org/10.1037/met0000144

McNulty, J. K., Neff, L. A., & Karney, B. R. (2008). Beyond initial attraction: Physical attractiveness in newlywed marriage. *Journal of Family Psychology*, *22*, 135–143. https://doi.org/10.1037/0893-3200.22.1.135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Menard, S. (1995). *Applied logistic regression analysis*. Sage.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Sage. https://doi.org/10.4135/9781412983433

Mendoza, J. L., Toothaker, L. E., & Crain, B. R. (1976). Necessary and sufficient conditions for *F* ratios in the $L \times J \times K$ factorial design with two repeated factors. *Journal of the American Statistical Association*, *71*, 992–993.

Mendoza, J. L., Toothaker, L. E., & Nicewander, W. A. (1974). A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. *Multivariate Behavioural Research*, *9*, 165–177.

Meston, C. M., & Frohlich, P. F. (2003). Love at first fright: Partner salience moderates roller-coaster-induced excitation transfer. *Archives of Sexual Behavior*, *32*, 537–544. https://doi.org/10.1023/a:1026037527455

Miles, J. N. V., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. Sage.

Mill, J. S. (1865). *A system of logic: Ratiocinative and inductive*. Longmans, Green.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*, 40–48. https://doi.org/10.1037//0021-843x.110.1.40

Miller, G., Tybur, J. M., & Jordan, B. D. (2007). Ovulatory cycle effects on tip earnings by lap dancers: Economic evidence for human estrus? *Evolution and Human Behavior*, *28*, 375–381. https://doi.org/doi.org/10.1016/j.evolhumbehav.2007.06.002

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(3), 255–273. https://doi.org/10.1080/10705519509540013

Mishra, J., Zinni, M., Bavelier, D., & Hillyard, S. A. (2011). Neural basis of superior performance of action videogame players in an attention-demanding task. *Journal of Neuroscience*, *31*(3), 992–998. https://doi.org/10.1523/JNEUROSCI.4834-10.2011

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*, 67–82. https://doi.org/10.1093/esr/jcp006

Morewedge, C. K., Huh, Y. E., & Vosgerau, J. (2010). Thought for food: Imagined consumption reduces actual consumption. *Science*, *330*, 1530–1533. https://doi.org/10.1126/science.1195701

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Morey, R. D. (2013). The consistency test does not – and cannot – deliver what is advertised: A comment on Francis (2013). *Journal of Mathematical Psychology*, *57*,

180–183. https://doi.org/10.1016/j.jmp.2013.03.004

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., Lewandowsky, S., Morey, C. C., Newman, D. P., Schonbrodt, F. D., Vanpaemel, W., Wagenmakers, E. J., & Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*. https://doi.org/10.1098/rsos.150547

Mulaik, S. A. (2009). *Foundations of factor analysis* (2nd ed.). Chapman & Hall/CRC.

Muris, P., Huijding, J., Mayer, B., & Hameetman, M. (2008). A space odyssey: Experimental manipulation of threat perception and anxiety-related interpretation bias in children. *Child Psychiatry and Human Development*, *39*, 469–480. https://doi.org/10.1007/s10578-008-0103-z

Myers, R. (1990). *Classical and modern regression with applications* (2nd ed.). Duxbury.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691–692.

Namboodiri, K. (1984). *Matrix algebra: An introduction*. Sage.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289–337.

Ni, X., Wong, Z. J., Mrejen, M., Wang, Y., & Zhang, X. (2015). An ultrathin invisibility skin cloak for visible light. *Science*, *349*(6254), 1310–1314. https://doi.org/10.1126/science.aac9411

Nichols, L. A., & Nicki, R. (2004). Development of a psychometrically sound internet addiction scale: A preliminary step. *Psychology of Addictive Behaviors*, *18*, 381–384. https://doi.org/doi.org/10.1037/0893-164X.18.4.381

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141. https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. https://doi.org/10.1177/1745691612459058

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

O'Brien, M. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*, 316–333.

Ofcom. (2006). *Media literacy audit: Report on children's media literacy*. Media Literacy Audit: Report on Media Literacy amongst Children. https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens/children

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, *69*, 894–908.

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, *83*, 579–586.

Olson, C. L. (1979). Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens. *Psychological Bulletin*, *86*, 1350–1352.

Ong, E. Y. L., Ang, R. P., Ho, J. C. M., Lim, J. C. Y., Goh, D. H., Lee, C. S., & Chua, A. Y. K. (2011). Narcissism, extraversion and adolescents' self-presentation on Facebook. *Personality and Individual Differences*, *50*, 180–185. https://doi.org/10.1016/j.paid.2010.09.022

Oxoby, R. J. (2008). On the efficiency of AC/DC: Bon Scott versus Brian Johnson. *Economic Enquiry*, *47*, 598–602. https://doi.org/10.1111/j.1465-7295.2008.00138.x

Payne-Carter, S., Greenberg, K., & Walker, M. (2016). *The impact of computer usage on academic performance: Evidence from a randomized trial at the United States Military Academy*. MIT Department of Economics.

Pearson, E. S., & Hartley, H. O. (1954). *Biometrika tables for statisticians, volume I*. Cambridge University Press.

Pearson, K. (1894). Science and Monte Carlo. *The Fortnightly Review*, *55*, 183–193.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157–175.

Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design and analysis: An integrated approach*. Erlbaum.

Peirce, C. S. (1878). Deduction, induction, and hypothesis. *Popular Science Monthly*, *13*, 470–482.

Perham, N., & Sykora, M. (2012). Disliked music can be better for performance than liked music. *Applied Cognitive Psychology*, *26*, 550–555. https://doi.org/10.1002/acp.2826

Piff, P. K., Stancato, D. M., Côté, S., Mendoza-Dentona, R., & Keltner, D. (2012). Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences*, *109*, 4086–4091.

Plackett, R. L. (1983). Karl Pearson and the chi-squared test. *International Statistical Review*, *51*, 59–72.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods Instruments & Computers*, *36*, 717–731.

Preacher, K. J., & Hayes, A. F. (2008a). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Preacher, K. J., & Hayes, A. F. (2008b). Contemporary approaches to assessing mediation in communication research. In A. F. Hayes, M. D. Slater, & L. B. Snyder (Eds.), *The Sage sourcebook of advanced data analysis methods for communication research* (pp. 13–54). Sage.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*, 93–115. https://doi.org/10.1037/a0022658

R Core Team. (2023). *R: A language and environment for statistical computing* (4.3). R Foundation for Statistical Computing.

Rao, C. R. (1945). Generalisation of Markoff's theorem and tests of linear hypotheses. *Sankhyā: The Indian Journal of Statistics (1933–1960)*, *7*(1), 9–16. JSTOR.

Rao, C. R. (1948). Tests of significance in multivariate analysis. *Biometrika*, *35*(1/2), 58–79. JSTOR. https://doi.org/10.2307/2332629

Ratcliff, R. (1993). Methods for dealing with reaction-time outliers. *Psychological Bulletin*, *114*, 510–532.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. (2nd ed.). Sage.

Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1993). Music and spatial task-performance. *Nature*, *365*, 611–611. https://doi.org/10.1038/365611a0

Revelle, W. (2022). psych: Procedures for personality and pychological research. *R Package Version 2.2.5*. https://CRAN.R-project.org/package=psych

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. Business Premium Collection. https://doi.org/10.1007/s11336-008-9102-z

Rockwell, R. C. (1975). Assessment of multicollinearity: The Haitovsky test of the determinant. *Sociological Methods and Research*, *3*, 308–320.

Rogosa, D. (1981). On the relationship between the Johnson-Neyman region of significance and statistical tests of parallel within group regressions. *Educational and Psychological Measurement*, *41*.

Rouanet, H., & Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, *23*, 147–163.

Rowe, R., Costello, E. J., Angold, A., Copeland, W. E., & Maughan, B. (2010). Developmental pathways in oppositional defiant disorder and conduct disorder. *Journal of Abnormal Psychology*, *119*(4), 726–738. http://dx.doi.org/10.1037/a0020798

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416–428. https://doi.org/10.1037/h0042040

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, *9*, 99–103.

Ruxton, G. D., & Neuhaeuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution*, *1*, 114–117. https://doi.org/10.1111/j.2041-210X.2010.00014.x

Sacco, W. P., Levine, B., Reed, D., & Thompson, K. (1991). Attitudes about condom use as an AIDS-relevant behavior: Their factor structure and relation to condom use. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*, 265–272.

Sachdev, Y. V. (1967). An unusual foreign body in the rectum. *Diseases of the Colon & Rectum*, *10*, 220–221.

Salsburg, D. (2002). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Owl Books.

Sana, F., Weston, T., & Cepeda, N. J. (2013). Laptop multitasking hinders classroom learning for both users and nearby peers. *Computers & Education*, *62*, 24–31. https://doi.org/10.1016/j.compedu.2012.10.003

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114. https://doi.org/10.2307/3002019

Savage, L. J. (1976). On re-reading R. A. Fisher. *Annals of Statistics*, *4*, 441–500.

Scanlon, T. J., Luben, R. N., Scanlon, F. L., & Singleton, N. (1993). Is Friday the 13th bad for your health? *British Medical Journal*, *307*, 1584–1586.

Scariano, S. M., & Davenport, J. M. (1987). The effects of violations of independence in the one-way ANOVA. *The American Statistician*, *41*, 123–129.

Schützwohl, A. (2008). The disengagement of attentive resources from task-irrelevant cues to sexual and emotional infidelity. *Personality and Individual Differences*, *44*, 633–644.

Selfhout, M. H. W., Delsing, M. J. M. H., ter Bogt, T. F. M., & Meeus, W. H. J. (2008). Heavy metal and hip-hop style preferences and externalizing problem behavior: A two-wave longitudinal study. *Youth & Society*, *39*(4), 435–452. https://doi.org/10.1177/0044118X07308069

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, *25*, 4334–4344. https://doi.org/10.1002/sim.2682

Shee, J. C. (1964). Pargyline and the cheese reaction. *British Medical Journal*, *1*, 1441.

Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work. *Nature*, *526*, 189–191.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666–681. https://doi.org/10.1177/1745691614553988

Sobel, M. E. (1982). Asymptotic intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology 1982* (pp. 290–312). Jossey-Bass.

Sonnentag, S. (2012). Psychological detachment from work during leisure time: The benefits of mentally disengaging from work. *Current Directions in Psychological Science*, *21*, 114–118. https://doi.org/10.1177/0963721411434979

Southgate, G., Rashford, M., Sancho, J. M., & Saka, B. A. T. M. (2020). It's happening again, but we love you Saka. *Archives of Pain*, *07*(01), 616–666.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*, 271–295.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*(2), 164–182. https://doi.org/10.1037/1082-989X.9.2.164

Stephens, R., Atkins, J., & Kingston, A. (2009). Swearing as a response to pain. *Neuroreport*, *20*, 1056–1060. https://doi.org/10.1097/WNR.0b013e32832e64b1

Stevens, J. P. (1979). Comment on Olson: Choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, *86*, 355–360.

Stevens, J. P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin*, *88*, 728–737.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Erlbaum.

Strahan, R. F. (1982). Assessing magnitude of effect from rank-order correlation coefficients. *Educational and Psychological Measurement*, *42*, 763–765.

Studenmund, A. H., & Cassidy, H. J. (1987). *Using econometrics: A practical guide*. Little Brown.

Student. (1908). The probable error of a mean. *Biometrika*, *6*, 1–25. https://doi.org/doi:10.1093/biomet/6.1.1

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Allyn & Bacon.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, *14*, 327–333.

Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*, 414–424.

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*, 90–99.

Toothaker, L. E. (1993). *Multiple comparison procedures*. Sage.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.

Tuk, M. A., Trampe, D., & Warlop, L. (2011). Inhibitory spillover: Increased urination urgency facilitates impulse control in unrelated domains. *Psychological Science*, *22*, 627–633. https://doi.org/10.1177/0956797611404901

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling, Issue 2* (pp. 448–485). Stanford University Press.

Twenge, J. M. (2000). The age of anxiety? Birth cohort change in anxiety and neuroticism, 1952–1993. *Journal of Personality and Social Psychology*, *79*, 1007–1021.

Twisk, J. W. R. (2019). *Applied mixed model analysis: A practical guide*. Cambridge University Press.

Umpierre, S. A., Hill, J. A., & Anderson, D. J. (1985). Effect of Coke on sperm motility. *New England Journal of Medicine*, *313*, 1351.

Van Bourg, J., Patterson, J. E., & Wynne, C. D. L. (2020). Pet dogs (canis lupus familiaris) release their trapped and distressed owners: Individual variation and evidence of emotional contagion. *PLoS One*, *15*(4), e0231742. https://doi.org/10.1371/journal.pone.0231742

van Ravenzwaaij, D., Boekel, W., Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2014). Action video games do not improve the speed of information processing in simple perceptual tasks. *Journal of Experimental Psychology: General*, *143*(5), 1794–1805. https://doi.org/10.1037/a0036923

Vandekerckhove, J., Guan, M., & Styrcula, S. A. (2013). The consistency test may be too weak to be useful: Its systematic application would not improve effect size estimation in meta-analyses. *Journal of Mathematical Psychology*, *57*, 170–173. https://doi.org/10.1016/j.jmp.2013.03.007

Vezhaventhan, G., & Jeyaraman, R. (2007). Unusual foreign body in urinary bladder: A case report. *The Internet Journal of Urology*, *4*.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wainer, H. (1972). A practical note on one-tailed tests. *American Psychologist*, *27*, 775–776. https://doi.org/10.1037/h0020482

Wainer, H. (1984). How to display data badly. *The American Statistician*, *38*, 137–147.

Waller, N. G. (2008). Commingled samples: A neglected source of bias in reliability analysis. *Applied Psychological Measurement*, *32*(3), 211–223. https://doi.org/10.1177/0146621607300860

Wasserstein, R. L., & American Statistical Association. (2016). ASA statement on statistical significance and *p*-values. *The American Statistician*, *70*, 131–133.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*, 129–131. https://doi.org/10.1080/00031305.2016.1154108

Weaver, B., & Wuensch, K. L. (2013). SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behavior Research Methods*, *45*, 880–895. https://doi.org/10.3758/s13428-012-0289-7

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336.

Wendt, H. W. (1972). Dealing with a common problem in social science: A simplified rank-biserial coefficient of correlation based on the *U* statistic. *European Journal of Social Psychology*, *2*(4), 463–465. https://doi.org/10.1002/ejsp.2420020412

Wenger, M. R., & Frisco, M. L. (2021). Extradyadic sex and psychological distress among married and cohabiting young adults: An examination of internalized and externalized responses. *Journal of Family Issues*, *42*(4), 785–812. https://doi.org/10.1177/0192513X20927766

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. R.I.P. *The American Statistician*, *68*(3), 191–195. https://doi.org/10.1080/00031305.2014.917055

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (2nd ed.). Springer.

Wilcox, R. R. (2016). *Understanding and applying basic statistical methods using R*. John Wiley & Sons.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Elsevier.

Wilcox, R. R., Carlson, M., Azen, S., & Clark, F. (2013). Avoid lost discoveries, because of violations of standard assumptions, by using modem robust statistical methods. *Journal of Clinical Epidemiology*, *66*, 319–329. https://doi.org/10.1016/j.jclinepi.2012.09.003

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80–83.

Wildt, A. R., & Ahtola, O. (1978). *Analysis of covariance*. Sage.

Wilkinson, L. & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, *73*, 123–136.

Wright, D. B., London, K., & Field, A. P. (2011). Using bootstrap estimation and the plug-in principle for clinical psychology data. *Journal of Experimental Psychopathology*, *2*, 252–270. https://doi.org/doi:10.5127/jep.013611

Wu, Y. W. B. (1984). The effects of heterogeneous regression slopes on the robustness of 2 test statistics in the analysis of covariance. *Educational and Psychological Measurement*, *44*, 647–663.

Yates, F. (1951). The influence of *Statistical methods for research workers* on the development of the science of statistics. *Journal of the American Statistical Association*, *46*, 19–34.

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika*, *61*, 165–170. https://doi.org/10.1093/biomet/61.1.165

Zabell, S. L. (1992). R. A. Fisher and fiducial argument. *Statistical Science*, *7*, 369–387.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics: Proceedings of the first international meeting held in Valencia (Spain)* (pp. 585–603). University Press.

Zhang, S., Schmader, T., & Hall, W. M. (2013). L'eggo my ego: Reducing the gender gap in math by unlinking the self from performance. *Self and Identity*, *12*, 400–412. https://doi.org/10.1080/15298868.2012.687012

Zibarras, L. D., Port, R. L., & Woods, S. A. (2008). Innovation and the 'dark side' of personality: Dysfunctional traits and their relation to self-reported innovative characteristics. *Journal of Creative Behavior*, *42*, 201–215. https://doi.org/10.1002/j.2162-6057.2008.tb01295.x

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice and lives*. University of Michigan.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical & Statistical Psychology*, *57*, 173–181.

Zwick, R. (1985). Nonparametric one-way multivariate analysis of variance: A computational approach based on the Pillai-Bartlett trace. *Psychological Bulletin*, *97*, 148–152.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442.

INDEX

α-level 83, 93, 95
additivity and linearity 271, 427
adjusted means 621
adjusted predicted value 423
adjusted R_2 420
Akaike information criterion (AIC) 440, 1004
album sales and advertising (linear models) 414–25
all-or-nothing thinking 718
alpha factoring 833
alternative (experimental) hypotheses 82, 83
American Psychological Association (APA) 42, 124
American Statistical Association (ASA) 124
analysis of covariance (ANCOVA) 316, 617–48
  assumptions
    covariate and treatment effect 622–5
    heterogeneity of regression slopes 626
    homogeneity of regression slopes 625–7, 641–2
    violation 627
  Bayesian analysis with covariates 645–6
  effect size 622
  elimination of confounds 617
  and general linear model 618–22
  interpretation
    contrasts 638–40
    covariates 640, 643
    excluded covariate 634–5
    main analysis 635–8
    parameter estimates 637–8
    post hoc tests 638–40
  introduction 617
  planned contrasts 631–2
  reducing within-group error variance 617
  reporting results 647
  robust ANCOVA 644–5
  role of covariate 623
  in SPSS Statistics
    bootstrapping and plots 634
    contrasts 630
    dialog box for general linear model 629
    general procedure 627–8
    inputting data 628
    options 630–3
    testing independence of treatment variable and covariate 628–9
analysis of variance (ANOVA) 558, 575, 586
  approach to repeated-measures designs 700–4
  comparing several means 560
  factorial 653, 676, 685
  and linear model 654
  and MANOVA 705
  mixed 753
  one and two-tailed tests in 601
  one-way independent ANOVA 592, 608
    dialog box 594
    general procedure 593–4
    output from 599–606
    reporting results from 610
    robust comparisons 607
  partitioning variance for 578
  regression and 561
  robustness 576–7
analytic thinking 21–2
Anderson–Rubin method 831
Arrindell, W. A. 844
assumptions 269–70
  additivity and linearity 271
  confidence intervals 275–7
  homoscedasticity/homogeneity of variance 263, 271–5
  hypothesis tests 275–7
  independent errors 275
  normality 275–9
  sampling distributions 275–7
  and significance tests 295
  standard error 275–7
Auerswald, M. 837
autocorrelation 427
autoregressive structure/AR(1) 1007
average sum of squares 573

β-level 93, 95
bar charts 214–26
  Bar Chart Gallery 214–15
  clustered bar charts for independent means 216–18
    dialog boxes 219
    films: psychological arousal 218
  clustered bar charts for mixed designs 223–6
  clustered bar charts for related means 223
    Completed Chart Builder for repeated-measures graph 22
    hiccup sufferers: bar chart of number of hiccups after interventions 222
  simple bar chart for related means 219–22
    Element Properties for repeated-measures graph 221
    for repeated-measures data 220
    summary group dialog box 221
  simple bar charts for independent means 215–16
    dialog boxes 216

error bar 216
films: psychological arousal 217
Bargman, R. E. 815
Barnett, V. 424
Baron, R. M. 540, 542
Bartlett's test of sphericity 804, 845, 853, 855, 856
Basic and Applied Social Psychology 124
Bayes factor 143
Bayes' theorem 138
Bayesian analysis with covariates 645–6
Bayesian comparison, of several means 609
Bayesian models, of factorial designs 687–9
Bayesian statistics 135–44
alternative hypothesis 142–3
Bayes' theorem 138
benefits 143–4
priors on parameters 139–41
reporting Bayes factors 145–6
reporting effect sizes 145–6
updating beliefs 139
beautiful women and good-looking men 509
Beck Depression Inventory (BDI) 326
Beckham, A. S. 911
beer-goggles effect (factorial designs) 653
Belia, S. 96–7
Belsey, D. A. 425
Bernard, P. 764
Beta, β_i 454
between-groups designs 19
between-subjects designs 19
bias
checking for 279–302
linear models 422–5
central limit theorem 259–62
confidence intervals and significance tests 257–9
errors in models 247–52
errors vs. residuals 253–4
estimation 254–6
Gauss–Markov theorem 262–5
populations, samples and error in prediction 246–7
revision of 245–6
reducing
robust estimation method 303–16
transforming data 303–5
winsorizing 313–14
sources
outliers 266–9
violation of assumptions 266, 269–70
see also scientific research bias
bimodal distributions 24
binary logistic regression 927
binary variables 11, 14
biserial correlation 396–7
bivariate correlations 380, 384–6
Black Americans and happiness (categorical data analysis) 911
bladder control and inhibitory abilities 496
Board, B. J. 492
Boik, R. J. 704
Bolker, B. M. 1001, 1008, 1011
Bonferroni, C. 94
Bonferroni correction 94, 590, 591, 701, 704
bootstrapping 315–16, 597, 634, 675, 952
dialog box 316
percentile 315
boredom effects 21
Bowerman, B. L. 971
boxplots (box-whisker diagrams) 211–14
1-D 211
clustered 212
dialog box 212
simple 211–12
of success scores 214
Box's test 800
Bray, J. H. 801
Bronstein, M. V. 549–50
Brown–Forsythe *F* 575, 576
built-in contrasts 588

categorical data analysis 885–923
2 categorical variables 885–94
chi-square test as linear model 889–94
coefficients 894
contingency coefficient 889
contingency table 885
Cramer's *V* (effect size) 889
Fisher's exact test 887
furry example 885
likelihood ratio 887–8
other measures of association 888–9
Pearson's chi-square test 885–7
Phi statistic 889
Yates's correction 888
assumptions
expected frequencies 897–8
independence 897
row and column percentages 898
Bayesian test 907–8
multinomial model 908
Poisson model 908
crosstabulations 901, 903, 910, 912
entering frequencies and weighting cases 899–900
entering raw scores 899
general procedure 898–9
interpreting chi-square test 902–9
Bayesian test 907–8
calculating effect size 908–9
reporting the results 909
standardized residuals 904–5
summary 905–6
interpreting loglinear analysis 913–19
effect sizes 918–19
interaction 918
following up 917–18
loglinear analysis 895–7, 920
initial considerations 910–12
interpretation 913–19
main analysis 912–13
reporting the results 919
using SPSS 910–13
specifying chi-square test 900–1
SPSS procedure 910–13
symmetric measures 904

categorical variables 11, 12, 14, 132
Cattell, R. B. 825, 835
catterplots 231
cause
  and correlation 17, 380–1
  definitions 17
  and effect 9–10, 16, 17
  and statistics 18
central limit theorem 73, 259–62, 263, 306
  and sample size 277, 295
central tendency 26, 27, 35
centring 527–8, 1002
  predictors 1002–4
  variables 527–9
Çetinkaya, H. 359
Chamorro-Premuzic, T. 407
Chapman, J. P. 623
chartjunk 199
checking behaviour and mood 691–2
cheese causing nightmares (graphs) 201, 202
Chen, X. 479–80, 653
chi-square distribution 36, 90, 887
  critical values of 1043
children: interpretation of ambiguous information 641
chimps and bananas (causality) 19–20
Cliff, N. 834
cloak of invisibility and mischief (comparing two means) 480–5, 493–514
  coefficients 484
  descriptive 482
  tests of normality 482
Cobb, G. 123
Coca Cola as spermicide (correlations; variables) 10, 14–15, 16–17
Cochran's *Q* 363
coefficient of determination 393
Cohen, J. 93, 95, 108
Cohen's *d* 127–30, 134
Cole, D. A. 786
common factors 829
common variance 833
communality 833–4
comparing several means
  ANOVA 560, 561
    one and two-tailed tests in 601
    robustness 576–7
  assumptions 574–7
    ANOVA 576–7
    homoscedasticity 575–6
    violation 577
  Bayesian comparison 609
  contrast coding 577–89
  data in Puppies.sav 560
  dummy coding 562–8
  effect sizes 592
  F-statistic 560, 568–70, 573–4
    interpreting 574
  with linear model 560–74
  mean squares (MS) 573
  model sum of squares (SS_M) 571–2
  one and two-tailed tests 601
  one-way independent ANOVA
    dialog box 594
    general procedure 593–4
    options for 598
    output from 599–606
    reporting results from 610
    robust comparisons 607
  planned contrasts 577–89
    built-in contrasts 588
    defining contrasts using weights 582–7
    four-group experiment, using one/two control group 580, 581
    non-orthogonal contrasts 587
    orthogonal contrasts 583, 584
    output 602–3
    partitioning of model/experimental variance 579
    partitioning variance for ANOVA 578
    polynomial contrasts (trend analysis) 588–9
      cubic trends 589
      linear trends 589
      quadratic trend 589
      quartic trend 589
      in SPSS Statistics 594–5
    rules for assigning values to dummy variables 582
    standard contrasts in SPSS Statistics 588
  post hoc procedures 590–1
    bootstrapping 597
    output 603–6
    pairwise comparisons 590
    robust tests 590–1
    in SPSS Statistics 595–6
    summary 591
    Type I and Type II error rates 590
  puppy therapy data 566, 568, 571–3, 583, 587
  residual sum of squares (SS_R) 572–3
  robust comparisons 607–8
  SPSS Statistics 592–8
  total sum of squares (SS_T) 570–1
comparing two means
  Bayesian test 499–502, 512–14
    priors 499–500
    reporting 515–16
  differences 479
  effect sizes 498–9, 510–12
  entering data 502–3
  repeated-measures data 503–7
    adjusting values 506
    adjustment factor 504, 506
    descriptive statistics 504
    dialogue boxes 505
    error bar graphs 503–4
  t-tests 484–92
    assumptions 492
    dependent *t*-test 485
    independent *t*-test 485, 488–92
      output 495–6
    Leni's research 492
    paired-samples 485, 486–8, 507–8
      output 508–9
    procedure 492–3
    reporting 515–16

robust tests 497–8, 510
reporting 515
using *SD*s and *N*s 491
using SPSS 493–502
independent means 493–502
related means 502–14
complete separation 941–3
component loading 827
component matrices 829
compound symmetry 700
Comrey, A. L. 844
concurrent validity 15
confidence intervals 74–81, 94, 96–8, 275–7
calculating
95% intervals 74–7, 75
other than 95% intervals 78
in small samples 78
error bars 79, 80
moderate overlap 96
overlapping/not overlapping 79–80, 96
for Pearson's *r* 382
showing visually 79–81
and significance tests 257–9
and statistical significance 96–8
confirmatory factor analysis (CFA) 833
confounding variables 17
consistent estimator 266
contaminated normal distributions 278
content validity 15–16
contextual variables 983
contingency coefficient 889
continuous variables 12, 14
dichotomizing 480
and discrete variables 13
contrast coding 577–89
convergence and singular fits 1002
Cook, R. D. 425
Cook's distance 423
correlation analysis 375
bivariate correlation 384–97
biserial correlation 396–7
effect size 405
exam anxiety and performance 385
exam anxiety variables 386
general process 384–5
Kendall's tau 383, 395–6
point-biserial correlation 396–7
procedure 386–91
dialog boxes 387, 391
comparisons
dependent *r*s 404–5
independent *r*s 404
confidence intervals 382
correlation and causality 380–1
data entry for analysis 384
Kendall's tau 405
partial correlation 400, 401–3
Pearson's coefficient *r* 381–2, 384, 405
in SSPS 391–3
reporting 405–6
semi-partial (part) correlation 397–401
Spearman's coefficient 394–5, 405
z-scores 381
correlation coefficient 378–81
correlational research 16–17, 18
correlation/covariance matrix 849
Cortina, J. M. 871, 872
cosmetic surgery (multilevel linear models) 985–6
counterbalancing effects 21
covariance 376–8
correlation coefficient 378–81
observed data and means of two variables 377
positive and negative 377–8
standardization 378–81
structures 993–5, 1006–7
covariance ratio (CVR) 425
covariates 617, 640, 643
Bayesian analysis with 645–6
coverage 266
Cox and Snell's 938
Cramer's *V* (effect size) 889, 901, 908
credible intervals 141
criterion validity 15
critical values
of chi-square distribution 1043
of *F*-distribution 1039–42
of *t*-distribution 1037–8
Cronbach's alpha 870–3
cross-product deviations 376
cross-products 788
cross-sectional research 16
cross-validation 429
cubic trends 589
Cumming, G. 96, 97
cups of tea testing (Fisher's *p*-value) 81–2

Dai, X. 754
Daily Mirror 479
Daniels, E. A. 906
data
analysis 4, 22–40
collecting 4
measurement 9–16
research design 16–22
falsification 119, 120
need for 3, 8
reporting 40–4
dating anxiety and low self-esteem (causality) 17
Davenport, J. M. 577
Davey, G. C. L. 691, 810
David, Florence Nightingale 378, 379
de Leeuw, J. 1003, 1007
degrees of freedom 65–6, 789
deleted residual 423
dependent (outcome) variables 10, 17
depressant effects of two drugs 325–6, 331–4, 339
ranking the depression scores 326–7
determinant of matrices 846
deviance 29, 63
deviance statistic 937
deviation contrast 588
DFBeta 424
DFFit 425
Di Falco, A. 480–1
diagonal structure 1006–7
difference contrast (reverse Helmert contrast) 588

direct and indirect effect 539, 540
direct oblimin rotation 840
direct quartimin rotation 840
discrete variables 13
  and continuous variables 13
discriminant function analysis 785, 811–18
  classification options 813
  dialog box 812
  interpretations 814–17
    canonical discriminant function coefficient 816
    combined-groups plot 816–17
    covariance matrices 814
    eigenvalues 814–15
    Wilks' lambda 814–15
  reporting results 817–18
  save new variables dialog box 813
  statistic options 812
  variates 796–8
discriminant scores 814
Domjan, M. 345, 359
Download Music Festival and hygiene 280–302
drinking and mortality, in London 426
dummy coding 562–4
  dummy variables 562–4
  and linear model 565–8
dummy variables 483, 562–4
Dunlap, W. P. 306
Dunn and Scheffé test 590
Dunn, O. 94
Dunnett's T3 591, 604
Durbin–Watson test 427

ecological validity 16
eels for constipation (logistic regression) 943–4
effect sizes 95, 126–35
  Cohen's *d* 127–30
  compared with null hypothesis significance testing (NHST) 133–4
  meta-analysis 134–5
  odds ratio 132–3
  Pearson's *r* 130–1
efficient estimator 266
eigenvalues 835–9
eigenvectors and eigenvalues 458
empirical probability 116
Enders, C. K. 1003
English football (logistic regression) 962–71
equamax rotation 840
error bars 79
error in prediction, population model 246–7
error sum of squares and cross-products matrix (error SSCP) 788, 794
essential tau-equivalence assumption 870
eta squared 592, 601
  partial 622
exam anxiety and performance 228, 229
  and time spent revising 233–4, 397–401
experimental hypotheses 82, 83
experimental research 17–19
  manipulation 19, 479
experimentwise error rate 94
extraction 835
extreme views and shades of grey 117–18
Eysenck, H. J. 825

F-distribution 36, 90
  critical values of 1039–42
F-ratio *see* *F*-statistic
F-statistic 420, 560, 573–4
  in factorial designs 663–8
  interpreting 574
  overview of 568–70
  for repeated-measures designs 704–9
  and Type I errors 668, 669
F-test 306, 420, 568
Facebook: narcissism and behaviour 465–6
factor analysis 825–82
  adequacy of data 843–4
  common factors 829
  common variance 833
  communality 833–4, 835, 836, 844
  correlation/covariance matrix 849
  correlations between variables 845–7
  descriptives in 848
  determinant of matrices 846
  dialog box 847
  discovering factors 832–41
  distribution of data 847
  eigenvalues 835–9
  exploratory factor analysis results for SPSS anxiety questionnaire 867
  factor extraction 835–9, 848–9, 857–61
  factor loading 827
  factor rotation 839–41
  factor scores 830–2, 865–6
    coefficients 831
  factors and components 826
  graphical representation 827–8
    factor plots 827
  interpretation 853–66
    anti-image matrices 853–5
    case summaries 865, 866
    communalities 857, 858, 860
    correlation matrix 853
    factor correlation matrix 864
    factor extraction 857–61
    factor matrix 858, 861
    factor scores 865–6
    factor structure 841
    Kaiser's criterion 857–8, 859
    oblique rotation 863–5
    orthogonal rotation 861–2
    preliminary analysis 853–6
    reproduced correlations 860
    rotation 857
    scree plot 858, 859
    summary 866
    variance explained 857
  Kaiser's criterion 836, 837, 839
  mathematical representation 828–30
  methods of discovering factors 832–3
  multicollinearity 845, 847
  parallel analysis 836, 838
  and principal component analysis (PCA) 834–5
  procedure 843, 844

reporting 866–7
sample size 844–5
scree plots 835–9
in SPSS statistics 847–52
factor extraction 848–9
options 851–2
rotation 849–51
scores 851
unique factors 829
uses 825–6
factor loading 827
factor matrices 829
factor rotation 839–41
factor scores 830–2
factor transformation matrix 841
factorial ANOVA 653, 676, 685
factorial designs 653
ANOVAs and linear model 654
assumptions 668
Bayesian models 687–9
breaking down the variance 664
coding scheme 656, 688
contrasts 673–4
output 678–9
data for beer-goggles effect 655
effect sizes 689–90
F-statistics 663–8
model sum of squares 664–5
residual sum of squares 667–8
total sum of squares 663–4
and Type I errors 668, 669
goggly example 653–5
independent 655–9
interaction graphs 659
interactions effect 667
output 680–1
interpretation 659–62
main effect of alcohol 666–7
output 677–8
main effect of type of face 665–6
output 675–7
model sum of squares 664–5
post hoc tests 674, 679, 681–3
repeated-measures designs
interpretation 733–45
using SPSS 727–32
reporting results 691
residual sum of squares 667–8
robust models 683–7
simple effects analysis 662–3, 675, 681
in SPSS Statistics
bootstrapping 675
contrasts 673–4
dialog box for options 675
entering data 670
general procedure 669
interaction graph 671
plotting effects 670–2
post hoc tests 674
simple effects 675
total sums of squares 663–4
factors 653
falsification 119, 120
familywise error rate 94
Fanelli, D. 118, 119–20
feedback 126
styles 18–19, 21
Fidell, L. S. 800, 831, 897
Fienberg, S. E. 346
Finch, S. 96
Fisher, R. 81, 82, 84, 112, 379, 381
Fisher's exact test 887, 897
fit 52, 53, 62–6, 63
fixed coefficients 986
fixed effects 986
fixed intercepts 987
fixed slopes 987
fixed variables 986
frequency distributions 22–4
deviations from mean 30, 36
dispersion 27–31, 34
ice bucket interest score 31, 32
interquartile range 28, 34
kurtosis 22–4
mean 26–7, 32
median 24–6
mode 24
normal distributions 22–3, 36
probability 31, 34
quartiles 28
skewness 22
standard deviations 30, 32, 34
table of standard normal distribution 38
variance 29, 32, 34
Friday 13th experiment 33
Friedman's ANOVA 341, 359–68
data 359–60
effect size 366–7
follow up 364–5
inputting data 361–2
output 363–4, 367
procedure 362–3
results 367
theory 360–1
Fritzon, K. 492
Frohlich, P. F. 304
Fuzzy hiding up a fireplace 821

Gabriel's procedure 590
Gallup, G. G. J. 606
Games, P. A. 306
Games–Howell procedure 591, 604, 605
Garfield, H. 231
Gauss–Markov theorem 262–5
Gelman, A. 509
gender and recognition of inverted pictures (mixed designs) 764
gender, infidelity and jealousy (mixed designs) 778–9
general linear models 375
with covariates 618–22
generalization 422
Glass, G. V. 306
Glastonbury Music Festival and hygiene (dummy coding) 564
global warming (predictions) 7

Goodman and Kruskal's lambda 901
*G*Power* 96
Graham, J. M. 863
grand mean 504
grand mean centring 527, 1002, 1003
grand variance 570–1
graphs
  for factor analysis 827
  features of a good graph 199
    chartjunk 199
    example of bad graph 200
      properly drawn 201
  misleading graphs 201
  scale 201
  SPSS chart builder 201–4
Grayson, D. 872
green lizard aliens as humans (probability) 136–7
  contingency table 137
Greenhouse–Geisser estimate 701, 703, 715, 716, 717–18, 722, 726, 734, 762
Greenland, S. 108
group mean centring 1002–3, 1004
Guadagnoli, E. 834, 844

HARKing 120, 121, 123, 126
harmonic means 605
Hartgerink, C. H. J. 122
Hartley's F_{max} 302
hat values 423
HE^{-1} (analogue of *F*) 795–6
Helmert contrast 588
heterogeneity
  of regression slopes 626
  of variance 271–5, 298–302, 301
heteroscedasticity 314, 428
hiccup sufferers (bar charts) 219–22
hierarchical regression 437
histograms 22, 204, 205–10
  defining a histogram 207
  histogram gallery 206
  histogram options 207–8
  population pyramid 209–10
Hochberg's GT2 590–1
homogeneity
  of covariance matrices 800
  of regression 998
  of regression slopes 625–7
homogeneous groups 348
homoscedasticity/homogeneity of variance 263, 271–5, 298–302, 428, 575–6
  Hartley's F_{max} 302
  and heterogeneity of variance 271–5, 298–302
  when it matters 274–5
  Levene's test 301–2
  spotting 298–302
  standardized residuals 298
Hosmer and Lemeshow's 938
Hotelling, H. 799
Hotelling–Lawley trace (Hotelling's T^2) 799, 801
Howell, D. C. 94, 897
Hume, D. 17
Hurlbert, S. H. 92

Huynh–Feldt estimates 701, 703, 715, 717, 726, 762
hypotheses 4, 5
  changing 86, 120
  directional/non-directional 83, 91
  experimental 82
  falsification 8–9
  generation 4, 5, 85, 833
  and predictions 6
  sum of squares and cross-products matrix (SSCP) 788, 794–5
  testing 6, 9–10, 16, 265, 275–7
  and theories 5
  types 82–3
  see also null hypothesis significance testing (NHST)

ice bucket challenge (frequency distributions) 31, 32–4, 39–40
  interest score 31, 32
identity matrix 797
independent ANOVA *see* one-way independent analysis of variance
independent designs 19, 20–1
independent errors 263, 275, 427–8
independent factorial designs 653, 655–9
  defining contrasts for 674
  defining plots for 672
  dialog box for 671
  see also factorial designs
independent variables 10, 19, 653
index of mediation 543
indirect/direct effect 539, 540
informative prior distributions 139
Instagram data (frequency distributions) 24–9
interaction graph 671
interaction variables 529
internet addiction (factor analysis) 868
interquartile range 28, 34
interval estimates 74
interval variables 12, 14
intraclass correlation coefficient (ICC) 875
iterative processes 940

Jiminy Cricket.sav 204
Johns, S. E. 223
Johnson–Neyman interval 530
Joliffe, I. T. 836
Jonckheere–Terpstra test 351, 352, 354–6
journals 40–1, 118, 124

Kaiser, H. F. 843
Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy 843, 853, 856
Kaiser's criterion 836, 837, 839, 857–8, 859
Kanazawa, S. 509
Kelley, K. 542
Kendall's statistic 901
Kendall's tau 383, 385, 395–6
Kendall's *W* 363
Kennedy, C. 31
Kenny D. A. 540, 542
Keselman, H. J. 704

Keselman, J. C. 704
Kline, P. 872, 875
Kolmogorov, A. 296
Kolmogorov–Smirnov test 295, 296
  explore command 296–7
  Z test 332
Kreft, I. G. G. 1003, 1007
Kruskal, W. 346
Kruskal–Wallis test 346–59, 577
  data for soya example 347–8
  effect size 356–7
  follow-up analysis 348–9
  inputting data 349–50
  and median test 352
  output 351–4
  pairwise comparisons 348, 350, 353
    effect size 356
  procedure 350–1
  results 357–9
  stepwise procedure 350–1, 354
  theory 346–8
kurtosis 22–4, 295

Lacourse, E. 972
Lakens, D. 122
Lalande, D. R. 121, 122
Lambert, N. M. 540, 541, 543
latent variables 825, 833
least-significant difference (LSD) 590
Lee, H. B. 844
Lenth, R. V. 128
Lépine, D. 701
leptokurtic distributions 23
levels of measurement 11–14
Levene's test 295, 301–2, 493, 575, 763–4
  reporting 302
leverage (hat values) 423
Levine, D. W. 306
Lewis, T. 424
likelihood 138
likelihood ratio 887–8, 904
line charts 226
line-dancing cats
  categorical data analysis 885–94
  loglinear analysis 895–7
linear models 53, 54, 56–8, 245–6, 413–74
  assessing predictors: *t*-statistic 421
  assumptions 427–9
    multicollinearity 428
    non-zero variance 428
    predictors uncorrelated with external variables 428
    variable types 428
  Automatic Linear Modeling 437
  baseline model 417–18
  Bayesian regression 469–71
  bias 422–5
    assumptions 463–6
    casewise diagnostics 459–63
    covariance ratio 462
    influential cases 423–5
    outliers 422–3
    partial plots 463–4, 465
  categorical predictors in 482–4
  central limit theorem 259–62
  chi-square test as 889–94
  comparing models 439–40
  comparing several means with 560–74
  confidence intervals and significance tests 257–9
  cross-validation 429
    data splitting 429
  dummy coding and 565–8
  entering predictors
    hierarchical regression 437
    stepwise regression 437
  errors
    in models 247–52
    vs. residuals 253–4
  estimation 254–6
  expressing 252
  Gauss–Markov theorem 262–5
  generalizing 425–9
  goodness of fit 417–21
    F-test 420
    model sum of squares 418
    total sum of squares 418
  multicollinearity 440–1
  with one predictor 413–15
    fitting procedure 431–6
    interpreting 433–6, 434
      overall fit 433–4
      parameters 434–5
    using 435–6
  output
    collinearity 456–9
    descriptive statistics 448–9
    excluded variables 456
    parameters 452–6
  populations, samples and error in prediction 246–7
  reporting 471
  revision of 245–6
  robust regression 466–8
    bootstrap for coefficients 467
    dialog box 469
  sample size 429–30
  with several predictors 415–17, 436–41
  in SPSS 441–7
    dialog box 444
    main options 441–3
    matrix scatterplot 441–2
    Regression dialog box 443–4
    regression plots 444–5
    saving regression diagnostics 445–7
    statistics 443–4
  suppressor effects 439
linear trends 589
linearity 298–302, 384–5
Lo, S. F. 943
log-likelihood statistic 936–7
logistic regression 927–77
  alternative way to expressing the model 933–4
  assumptions
    independence of errors 940
    linearity 939–40
    testing 969–70

binary logistic regression 943–52
bootstrapping 952, 955
building models 945
using SPSS 946, 947
categorical predictors 947–8
comparing models 948–50
complete separation 941–3
data entry 945
deviance statistic 937
error messages: iterations 940
general procedure 945
incomplete information from predictors 941
interactions 962–71
fitting the model 964
interpreting the model 964–8
testing for linearity of logit 969–70
testing for multicollinearity 970–1
interpretation 952–62
analogues of *R* 957
blocks 952
classification plots 958
effect size 962
listing predicted probabilities 958–60
model summary 953–5
odds ratio 955–7
residuals 960–1
issues in 943
listing residuals 952
log-likelihood statistic 936–7
method of regression 946
model building 939, 945
modelling categorical outcomes 927–33
obtaining residuals 951
odds ratio 934–5, 955–7, 963, 971
options 951–2
process of fitting a model 946
R and R^2 938–9
refitting models 950
reporting 971–3
theory of 927–39
Wald statistic (*z*-statistic) 935–6, 955
loglinear analysis 895–7, 913–20
see also categorical data analysis
Lombardi, C. M. 92
longitudinal research 16
loud concerts and hearing (homoscedasticity) 272
lower-bound estimates 701
lower quartile 28

M-estimators 310, 311
MacCallum, R. C. 480, 844
MacKinnon, D. P. 543
Mahalanobis distances 423, 461, 463
Mahalanobis, Prasanta Chandra 424
Mann–Whitney test 326, 328, 331–4
dialog boxes 333
effect size 334–6
writing result 336–7
MANOVA *see* multivariate analysis of variance (MANOVA)
marginal homogeneity 341
marginal likelihood 138
Marzillier, S. L. 810
Masicampo, E. J. 121, 122
Massar, K. 548
mathematicians and Cthulhu 54
matrices 787–8
determinant 846
pattern and structure matrices 830
R-matrix 826, 827
variance–covariance 800, 871, 993–4, 995
Matthews, R. C. 345
Mauchly's test 701, 702, 715–16, 763
maximum likelihood (ML) 256, 833, 1005
Maxwell, S. E. 704, 801
McCloskey, D. N. 108
McNemar's test 340–1
mean 26–7, 32, 79, 306
as statistical model 61–2, 67, 71
mean squares (MS) 573, 708
mean sum of squares 420, 452
measurement 9–16
error 14–15
independent and dependent variables 9–11
levels of measurement 11–14
median 24–6
median in box-plots 211
median splits 480
median test 352
mediation 539–53
conceptual model 539–40
direct and indirect effect 539, 540
effect sizes 542–3
index of mediation 543
mediators 540
moderation and 544
reporting 548–9
Sobel test 542
in SPSS Statistics 543–4
dialog boxes 544
output 544–7
statistical model 540–2
with two mediators 549–53
dialog boxes 551
Meston, C. M. 304
meta-analysis 134–5
method of least squares 67, 274
Miles, J. 439
Mill, J. S. 17–18
Miller, G. A. 623
missing data 998
handling 281
mixed ANOVA 753
mixed designs 653, 751–82
assumptions 753–4
data 755
interactions 768–75
looks and charisma 769–71
looks, charisma and strategy 771–5
strategy and charisma 769
strategy and looks 768
Levene's test 763–4
main effect
of charisma 767–8
of looks 766–7
of strategy 763–5

output 762–75
reporting results 776–8
speed-dating example 754–5
in SPSS Statistics
contrasts 759–60
dialog box 757–61
entering data 756
fitting the model 757–9
general procedure 756
options 759–61
Plots dialog box 760
process 756
variable allocations 759
variable names and labels 757
mixed normal distributions 278
mode 24
bimodal distributions 24
multimodal distributions 24
model (hypothesis) sum of squares and cross-product matrix 788, 794–5
model sum of squares (SS_M) 418, 787, 788
comparing several means 571–2
factorial designs 664–5
repeated-measures designs 707
moderation 523, 526
centring variables 527–9
interaction effects
simple slopes analysis 529
linear model
categorical moderator 524
conceptual model 523–4, 523–5
continuous moderator 525
reporting 538
in SPSS Statistics 530–2
dialog boxes 531
output 532–7
PROCESS tool 530, 533
statistical model 525–7
Monte Carlo method 335
Mood, C. 963
mood induction 810–11
Morewedge, C. K. 82
Moses Extreme Reactions 332
Moshagen, M. 837
Mulaik, S. A. 834
multicollinearity 428, 440–1, 970–1
in factor analysis 845, 847
multilevel linear models
adding predictors 996
assessing fit and comparisons 1004–5
assumptions 1005–6
independence 1006
multicollinearity 1006
random effects 1005
benefits 998
centring predictors 1002–4
computing *p*-values 999–1000
convergence and singular fits 1002
cosmetic surgery data 985–6, 997–8
covariance structures
autoregressive structure/AR(1) 1007
diagonal structure 1006–7
of level 1 errors 995
of random effects 993–4
unstructured form 1006
variance components 1006
estimation 1005
Factors/*Covariates* 1013
fixed and random coefficients 986–8
fixed and random effects 1000–1
group mean centring 1002–3, 1004
growth models 982
hierarchical data 981–4
2-level structure 981–2
3-level structure 982, 983
intraclass correlation 983
level 1 errors 994–5
practical issues 998–1008
random intercepts 988–90
random slopes 990–3
reporting 1025
robust models 1008
sample size and power 1007–8
in SPSS Statistics 1008–23
adding random effects to model 1018–19
defining the model 1010
entering data 1008–10
fitting fixed effect models 1011–15
interpreting the model 1019–23
mixed models dialog box 1012, 1014, 1016–17
process 1009
scaling variables 1015–17
visualizing the data 1010–11
theory 984–98
variance components 1006
multilevel models 703
multimodal distributions 24
multinomial logistic regression *see under* logistic regression
multiple regression *see* linear models
multivariate analysis of variance (MANOVA) 594, 703, 705, 783–822
additional options 804
assumptions 800–1
Box's test 800
homogeneity of covariance matrices 800
independence 800
multivariate normality 800
random sampling 800
testing 804–5
violations 801
choosing outcomes 785
contrasts 803, 809
data 787
dialog box 803
final interpretation 818–19
means and confidence intervals between outcome variables in therapy group 818
outcome variables across each therapy group 818–19
follow-up analysis 801–2
see also discriminant function analysis
general procedure 802
interpretation 804–9
introducing 785
Levene's test of equality of variances 806

main analysis 802
multiple comparisons 803
power of 786
practical issues 800–2
and principal component analysis (PCA) 834–5
reporting results 810
residual SSCP matrix 804
in SPSS Statistics 802–4
SSCP matrices 804, 807–8
test statistics 801, 805–7
theory 787–99
discriminant function variates 796–8
error SSCP 788, 794
HE^{-1} (analogue of *F*) 795–6
Hotelling–Lawley trace (Hotelling's T^2) 799, 801
hypothesis SSCP 788, 794–5
matrices 787–8
Pillai–Bartlett trace (*V*) 798, 801
relationship between outcomes: cross-products 790–2
residual sum of squares 787
Roy's largest root 799, 801
sum of squares and cross-products (SSCP) matrices 787
total SSCP 788, 792–4
univariate *F* for outcome 1 (actions) 788–9
univariate *F* for outcome 2 (thoughts) 789–90
Wilks's lambda (Λ) 799, 801
univariate test statistics 806–7
see also discriminant function analysis
multivariate normality 800
Muris, P. 641

Nagelkerke's R_N^2 938
Namboodiri, K. 795
narcissistic personality disorder (hypotheses and theories) 4–9
and TV auditions 7–9
negative skew 23, 293
Neuhaeuser, M. 92
Neyman, J. 83, 84, 93, 112, 379
Nichols, L. A. 868
Nicki, R. 868
Nightingale, F. 200
nominal variables 11, 12, 14
non-linear models 54
non-orthogonal contrasts 587
non-parallel slopes model 641–2
non-parametric tests 303
Friedman's ANOVA 341, 359–68
Kolmogorov–Smirnov *Z* test 332
Kruskal–Wallis test 346–59
Mann–Whitney test 326, 328, 331–4
marginal homogeneity 341
McNemar's test 340–1
Moses Extreme Reactions 332
output 332–4
procedure 324–5
ranking scores 326–8
sign test 340
and statistical power 323
Wald-Wolfowitz runs 332
Wilcoxon rank-sum test 326, 328
Wilcoxon signed-rank test 337–45
non-positive definite matrix 853, 854
noniles 28
normality, assumption of
and central limit theorem 295
confidence intervals 275–7
when it matters 277–9
and parameter estimates 275
normality, spotting 291–8
normally distributed errors 428
Nosek, B. A. 118, 125
nQuery 96
null hypothesis significance testing (NHST) 81–101
confidence intervals 96–8, 97
and effect sizes 133–4
flow chart 85
inflated error rates 94–5
one-tailed and two-tailed tests 91–3
p-value 81–2, 90, 112, 123
principles for using NHST 124–5
and probability 111–17
problems
all-or-nothing thinking 109–11
statistical significance 108–9, 123
see also scientific research bias
process 83–9
researcher degrees of freedom 119, 126, 134
sample size and statistical significance 98–100
statistical power 95–6
test statistics 89–91
Type I and Type II errors 83, 93–4
Nunnally, J. C. 844

oblique rotation 830, 840, 841, 863–5
observation 3, 4–5
and hypotheses 4, 5
obsessive compulsive disorder (OCD) and cognitive behaviour therapy (CBT) (MANOVA) 785–819
O'Connell, R. T. 971
odds 132
odds ratio 132–3, 934–5, 955–7, 963, 971
contingency table 132
omega squared 592, 601–2, 724
one-tailed and two-tailed tests 91–3, 601
one-way independent analysis of variance 592
dialog box 594
general procedure 593–4
output from 599–606
reporting results from 610
robust comparisons 607
one-way repeated-measures designs *see* repeated-measures designs
Ong, E. Y. L. 465–6
open science 125, 126
Open Science Framework 40
order 528
ordinal variables 12, 14
ordinary least squares (OLS) 67, 256
orthogonal contrasts 583, 584
orthogonal rotation 830, 840, 841, 861–2
outcome (dependent) variables 10, 306–7
transforming 306

outliers 266–9
and parameters 267–9
spotting 280–7
stealth outliers 278
and sum of squared errors (SS) 268–9
transforming 303
using *z*-scores 286
Oxoby, R. J. 168, 211

p-curve 121
p-hacking 120, 121, 122, 123, 126
P-P plots (probability–probability plots) 288–91, 292–3
p-values 81–2, 85, 87, 98
computing 999–1000
influenced by intentions of scientists 111–17
pairwise comparisons 348, 350, 353, 590
parallel analysis 836, 838
parallel test 870
parameters 58–66, 265–6
estimates 60, 66–70, 73
and normality 275
and sample size 277
and outliers 267–9
testing 90
parsimony 939
partial correlation 400, 401–3
partial eta squared 622
PASS 96
pattern matrices, and structure matrices 830
Patterson, J. E. 1024
Pearson, E. 81, 83, 93, 112
Pearson, K. 379
Pearson's chi-square test 885–7, 897, 901, 904
as linear model 889–94
Pearson's correlation coefficient *r* 130–1, 378, 381–2, 382–4, 391–3, 827
Pedhazur, E. 872
Peer Reviewers' Openness Initiative 125
penis and sperm displacement 606
percentiles 28
perfect collinearity 440
Perham, N. 746
personality disorders and managers 492
Phi statistic 889, 901
Pillai–Bartlett trace (*V*) 798, 801
planned contrasts/comparisons 578
for ANCOVA 631–2
see also comparing several means, planned contrasts
platykurtic distributions 23
point-biserial correlation 396–7
point estimates 74
polychotomous logistic regression 927
polynomial contrasts (trend analysis) 588–9, 594–5
Popper, K. 7
popularity: aspects (factor analysis) 826–34
populations 55–6, 246–7
defined 248
errors in 247–52
population mean 71
and samples 55–6, 71–3
pornography use and marital fidelity (mediation) 540
positive skew 22–3, 293
post hoc procedures 590–1
output for 603–6
pairwise comparisons 590
robust tests 590–1
in SPSS Statistics 595–6
summary 591
Type I and Type II error rates 590
post hoc tests 578, 638–40, 674, 679, 681–3, 704, 714, 720–1
posterior distributions 139–41
posterior odds 142
posterior probability 138
Power and Precision 96
power of test: calculating 96
practice effects 21
pre-registration 93, 125
Preacher, K. J. 542
predicted value 415
predictions 4, 5–6, 7
and hypotheses 6–9, 7
and theories 7
predictive validity 15
predictor (independent) variables 10, 305–6, 479
transforming 305–6
principal component analysis (PCA) 825, 829, 830–1, 852
component scores 830
and factor analysis 834
and multivariate analysis of variance (MANOVA) 834–5
theory 834
prior distributions 139
prior odds 143
prior probability 138, 139
probability 31, 35
density function of normal distribution 39
distributions 35–6
empirical probability 116
normal distribution 36, 38
probability density function (PDF) 35, 36, 39
promax 840, 841
Psychological Science 121–2
publication bias 117–20
puppy therapy (comparing several means) 559–612
and love/dislike of puppies (ANCOVA) 617–48
pwr package 96

Q-Q plots 288–91, 292–3
quadratic trend 589
quails and fetishes 359
quails and terrycloth stimulus 359
quails' sperm and conditioning 345
quails' sperm release (confidence intervals) 74–81, 75
qualitative methods 3
quantiles 28
quantitative methods 3
quartic trend 589
quartiles 28
quartimax rotation 840

R and R^2 938–9
R-matrix 826, 827, 833
random coefficients 986, 1004
random effects 986
random intercepts 988–90
  models 699, 987
random slopes 987, 990–3, 1000
  model 699
random variables 250, 986
random variance 833
randomization 20–2
  counterbalancing effects 21
range 27–8
  see also interquartile range
ranking data 323
rating attractiveness of someone in photograph 304–6, 310
ratio variables 13, 14
reasons for learning statistics 3
registered reports 125
regression
  and ANOVA 561
  lines 229, 232, 235
  see also linear models
reliability 15–16
  meaning 868–70
reliability analysis 868–73
  Cronbach's alpha 870–3
  dialog box 874
  interpretation 875–8
  reporting 878–9
  split-half reliability 870
  in SPSS Statistics 873–5
  statistics for 874
repeated contrast 588
repeated-measures designs 19, 20, 21, 653, 695–749
  ANOVA 700–4
  assumptions 709
    Greenhouse–Geisser estimate 701, 703, 715, 716, 717–18, 722, 726, 734
    Huynh–Feldt estimates 701, 703, 715, 717, 726
    Mauchly's test 701, 702, 715–16
    sphericity 700–2, 704, 709, 714, 717
    violations 701, 702–3
  effect sizes 724–5
  F-statistic 704–9
    between-participant sum of squares 709
    mean squares 708
    model sum of squares 707
    output 716–18
    partitioning variance 705
    residual sum of squares 708
    total sum of squares 706
    within-participant sum of squares 706–7
  factorial
    contrasts 729–31
    dialog box 728, 729, 730, 731
    interpretation 733–45
      contrasts 741–3
      interaction plot 739
      main effect 736–44
      output 736, 738, 741
      pairwise comparisons 737, 738, 744, 745
      pivot trays 735
      post hoc tests 736, 742, 744
    plotting interactions 731
    simple effects 731–2
    using SPSS 727–32
  introduction 697
  and linear model 698–700
  nesting 982
  output 714–21
    contrasts 718–20
    correcting for sphericity 715
    descriptive statistics 714–15
    F-statistic 716–18
    pairwise comparisons 737
    post hoc tests 720–1, 722
  reporting 725–6, 747
  robust tests 721–3
  scented factorial 726
  in SPSS Statistics 709–14
    contrasts 711–12
    custom contrasts 712–13
    dialog boxes 711
    main analysis 710–11
    options dialog box 714
    post hoc tests 714
    process 710
replicating results 134
reporting
  bias in 118, 119
  data 40–4
  dissemination of research 40–1
  guiding principles 42–3
  numbers 42–3
  registered reports 125
  and reviews 42, 118, 125
reporting significance tests 100–1
research design 16–22
  correlational research 16–17
  data analysis 22–40
  data collection methods 19
  experimental research 17–19
  independent designs 19, 20–1
  randomization 20–2
  types of variation 19–20
research process 3–45
researcher degrees of freedom 119, 126, 134
residual mean sum of squares 708
residual sum of squares 256, 417–18, 787
  comparing several means 572–3
  and cross-product matrix 788, 794
  factorial designs 667–8
  repeated-measures designs 708
residuals 417, 422–3
  deleted 423
  and influence statistics 426
  standardized 422
  studentized 423
  studentized deleted residual 423
  unstandardized 422
restricted maximum likelihood (REML) 1005
Rice, J. 843
roaming cats 396, 397
robust methods 303

Rosenthal, R. 199
rotation 830
Rouanet, H. 701
Roy's largest root 799, 801
Ruxton, G. D. 92
Ryan, Einot, Gabriel and Welsch Q procedure (REGWQ) 590, 605, 679

Salsburg, David: *The Lady Tasting Tea* 81
samples, population model 246–7
sampling 60, 71–3
  central limit theorem 73
  confidence intervals 74–81
  from different populations 80–1
  distribution 71
  distributions 275–7
  Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy 843
  noise 95
  sample mean 71, 487
  sample size 43, 78, 95
    and central limit theorem 277
    factor analysis 844–5
    and level of power 96
    linear models 429–30
    and parameter estimates 277
    and significance tests 295
    and statistical significance 98–100
  standard error of differences 487
  standard error of mean 72–3
  variation 71
Satterthwaite and Kenward–Roger methods 1000
saturated model 892
Scanlon, T. J. 33
Scariano, S. M. 577
scatterplots 53, 227–36
  catterplots 231
  grouped scatterplots 230, 232
    dialog box 232
    exam anxiety and performance and biological sex 232
  matrix scatterplots 233–5
    dialog box 234
    exam anxiety, performance and revision time 235
  scatter/dot gallery 227–8
  simple scatterplot 228–30
    dialog box for scatterplot 228
    exam anxiety and performance 229
    grouped 3-D scatterplots 233
    Properties dialog box 230
Schmelkin, L. 872
Schützwohl, A. 778
Schwarz's Bayesian information criterion (BIC) 1004
scientific research bias
  falsifying data 120
  funding 119
  HARKing 120, 121, 123, 126
  incentives 117–19
  p-hacking 120, 121, 122, 123, 126
  publication bias 117–20
  researcher degrees of freedom 119, 126, 134
  test of excess success (TES) 122
scientific statements 7
scree plots 835
second quartile 28
seeing red and sexual attraction 223
self-report data 11
Senn, S. 624
sexual competitors and gossip 548
Shapiro–Wilk test 295, 296, 298, 349
Shee, J. C. 201
shrinkage 420
Šidák correction 630
sign test 340
significance tests
  skewness and kurtosis 297
simple contrast 588
simple effects analysis 662–3, 675, 681
simple regression *see* linear models
simple slopes analysis 529
singing to attract females 98–9, 126–7, 132
singularity 845
skew 22, 293
sniffer-dog data (repeated-measures designs) 698–749
Sobel test 542
Spearman, C. 383
Spearman's correlation coefficient (Spearman's rho) 382, 394–5
speed-dating example: charisma, looks and strategy (mixed designs) 754–78
sperm count and soya (Kruskal–Wallis test; Jonckheere–Terpstra test) 346–59
spherical errors
  homoscedasticity/homogeneity of variance 263, 271–5
sphericity 700–2, 704, 709, 717
SPINE acronym 51
split-half reliability 870
SPSS anxiety questionnaire (SAQ) 841–2, 853, 862, 866, 867, 869, 872
SPSS Statistics
  analysis of covariance (ANCOVA) 627–34
  Bayes factor package 908
  categorical data analysis 910–13
  Chart Builder 201–4
    The canvas 203
    colours 218–19, 238
    drop-line plot 236
    Drop zone 203
    Gallery 203
    initial dialog box 205
    simple dot lot or density plot 236
    Variable list 203
  Chart Editor 236–7, 238, 239
    Multiple comparisons 350
  comparing several means 588, 592–8
  comparing two means 493–514
  default options 185–7
  dialog box 152
  entering data 157–73
    coding variables 163–5
    copying and pasting 166–7
    currency variables 169–71
    data editor 151, 152–7
      data view 153

variable view 153, 160
data formats 157–9
long format 157
wide format 157–8
date variables 162–3
importing data 173
missing data 171, 172
naming variables 161, 162
numeric variables 163, 167
string variables 162, 163
value labels 165
exporting output 182, 183
extending SPSS Statistics
Integration plug-in for R 189–90
PROCESS tool 187–9
WRS2 package 190
factor analysis 847–52
factorial designs 669–75
Find and Replace 156
Go To 155
icons 155–7
importing data 173
linear models 441–7
on Linux 151
on Mac OS 151
mediation 543–7
menus
Analyze 154, 191
Compare Means 154
Correlate 154
Data 154
Descriptive Statistics 154
Dimension Reduction 155
Edit 154
Extensions 155
File 154
General Linear Model 154
Graphs 155
Help 155
Loglinear 155
Mixed Models 154
Nonparametric Tests 155
Regression 154
Scale 155
Transform 154
Utilities 155
View 154
Window 155
mixed designs 756–61
moderation 530–7
multilevel linear models 1008–23
multivariate analysis of variance (MANOVA) 802–4, 807–8
New Files 152
numbers with *E* 175
opening files 184–5
printing 178–9
Recent Files 152
reliability analysis 873–5
repeated-measures designs 709–14
saving files 182–4
SPSS syntax 174
SPSS viewer 151, 174, 177
classic mode 176–80
workbook mode 180–2
start up 151, 152
syntax editor 151, 179–80
Variables 156
square matrix 792
squared multiple correlation (SMC) 834
standard deviations 30, 32, 34, 378
standard error 275–7
of differences 487
of mean 72–3
standard normal distribution, table of 1033–6
standardization 378–81
standardized DFBeta 424
standardized DFFit 425
standardized residuals 298, 422
statistical models
fit 52, 53, 54, 62–6
lecturers and observed and mean number of friends 63
mean as 61–2
models and real-world data 51–2
parameters 58–67
populations 55–6
statistical power 95–6
statistical significance 98–100
misconceptions 108
statistical tests 18
stepwise regression 437
Stevens, J. P. 425, 704, 841
structure and pattern matrices 830
studentized deleted residual 423
Studentized Newman–Keuls procedure 590
studentized residuals 423
suicide risk (logistic regression) 972
sum of squared errors (SS) 29, 68
and outliers 268–9
sum of squared residuals (SS_R) 256
sum of squares and cross-products (SSCP) matrices 787, 804, 807–8
suppressor effects 439, 939
Sykora, M. 746
systematic/unsystematic variation 19–20, 89–90

t-distributions 36, 73, 85
critical values of 1037–8
t-statistic 90, 421
t-tests 18
t-values 86
Tabachnick, B. G. 800, 831, 897
table of standard normal distribution 1033–6
Tamhane's T2 591
tertium quid 17, 381
test statistic 89–91
test–retest reliability 16
tests of excess success (TES) 122
theories 4, 5
generation 4, 5
and hypotheses 5
toffees and adverts
data 376
observed data and means of two variables 377
Tofighi, D. 1003

tolerance statistic 441
TOSSE-R questionnaire 881
total sum of squares 418
  comparing several means 570–1
  and cross-products matrix (total SSCP) 788, 792–4
  factorial designs 663–4
  repeated-measures designs 706
Training Extra-terrestrial Reptile Detection (TERD) 698
transformations 303–5
  compute function 307–10
    dialog box 308
  square-root transformation 303–5
transforming data 303–5
  outcome variables 306
  predictor variables 305–6
Transparency and Openness Promotion (TOP) 125
trimming the data 303–5
  M-estimator 310, 311
  percentage based rule 310
  reaction time data 312
  standard deviation based 310, 312
  trimmed mean 310–11, 312
Tufte, E. R. 199, 200
Tuk, M. A. 496
Tukey, J. 92
Tukey's test 590, 604, 701, 704
Twisk, J. W. R. 1008
two-tailed tests *see* one-tailed and two-tailed tests
Type I errors
  and *F*-statistics 668, 669
  and multiple testing 684
  and Type II errors 93–4, 323–4, 353, 590
types of factorial designs *see* independent factorial designs

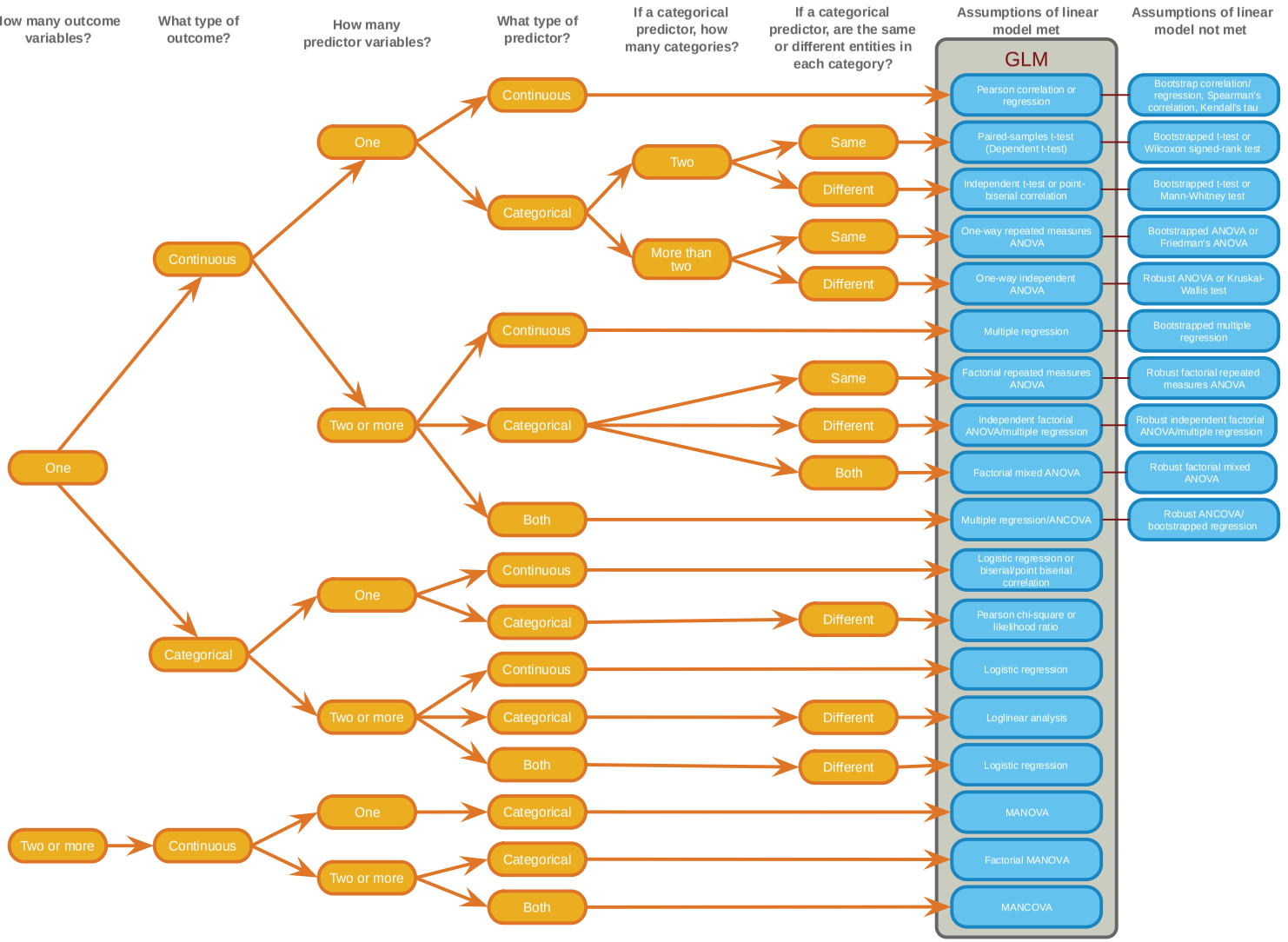
ultimatum game: offers to rock performers 168
unbiased estimator 265–6
unidimensionality 872
uninformative prior distributions 139
unique factors 829
unique variance 833
univariate models 785
unstandardized residuals 422
unsystematic variation 19, 20–1
upper quartiles 28

validity 15–16
Van Bourg, J. 1024
van der Ende, J. 844
variables
  confounding variables 17
  discrete 13
  identifying 5
  independent and dependent 9–11, 17, 18–19
  latent 825, 833
  measuring 5
  outcome 10
  predictor 10, 479
  ratio 13, 14
  relationships between 376
  categorical 11, 12, 14
  nominal 11, 12, 14
  binary 11, 14
  continuous 12, 14
  interval 12, 14
  ordinal variables 12, 14
variance 29, 32, 34, 376
  common 833
  components 1006
  and covariance 378
  homogeneity of 263, 271–5
  random 833
  unique 833
variance inflation factor (VIF) 441
variance ratio 302
variance-ratio method 560
variance sum law 489
variance–covariance matrices 800, 871, 993–4, 995
variation
  systematic/unsystematic 19–20, 89–90
  types 19–20
varimax 840, 861–2
Velicer, W. F. 834, 844
violent video games and aggression (moderation) 523–38

Wainer, H. 92
Wald, A. 936
Wald statistic (*z*-statistic) 935–6, 955
Wald-Wolfowitz runs 332
Weakliem, D. 509
Weaver, B. 405
weighted average 831
weighted least squares 274
weights 582
Weisberg, S. 425
Welch's *F* 575, 576, 577, 593, 601, 610
Why do you like your lecturers? 407, 472
Wilcox, R. R. 278, 316
Wilcoxon, F. 328
Wilcoxon rank-sum test 326, 328
Wilcoxon signed-rank test 337–45
  dialog boxes 341
  effect size 343–4
  output for alcohol group 342–3
  output for ecstasy group 342
  ranking data 337
  results 344
  split file command 339
  theory 337–9
Wilks's lambda (Λ) 799, 801, 814–15
winsorizing 313–14
within-participant sum of squares 706–7
within-subject design 19
women and idealized images (categorical data analysis) 906
World's Biggest Liar competition 394, 395, 406
Wuensch, K. L. 405
Wynne, C. D. L. 1024

Yates's correction 888, 904

z-scores 36, 40
Zabell, S. L. 84
Ziliak, S. T. 108
Zimmerman, D. W. 302


How many outcome variables?
What type of outcome?
How many predictor variables?
What type of predictor?
If a categorical predictor, how many categories?
If a categorical predictor, are the same or different entities in each category?
Assumptions of linear model met
Assumptions of linear model not met
GLM
One
Two or more
Continuous
Categorical
Continuous
Categorical
Two
More than two
Same
Different
Same
Different
Continuous
Categorical
Both
Continuous
Categorical
Both
Same
Different
Both
Different
Different
Different
Pearson correlation or regression
Paired-samples t-test (Dependent t-test)
Independent t-test or point-biserial correlation
One-way repeated measures ANOVA
One-way independent ANOVA
Multiple regression
Factorial repeated measures ANOVA
Independent factorial ANOVA/multiple regression
Factorial mixed ANOVA
Multiple regression/ANCOVA
Logistic regression or biserial/point biserial correlation
Pearson chi-square or likelihood ratio
Logistic regression
Loglinear analysis
Logistic regression
MANOVA
Factorial MANOVA
MANCOVA
Bootstrap correlation/ regression, Spearman's correlation, Kendall's tau
Bootstrapped t-test or Wilcoxon signed-rank test
Bootstrapped t-test or Mann-Whitney test
Bootstrapped ANOVA or Friedman's ANOVA
Robust ANOVA or Kruskal-Wallis test
Bootstrapped multiple regression
Robust factorial repeated measures ANOVA
Robust independent factorial ANOVA/multiple regression
Robust factorial mixed ANOVA
Robust ANCOVA/ bootstrapped regression