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

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

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

EXTENDED CONTENTS

PREFACE

Introduction

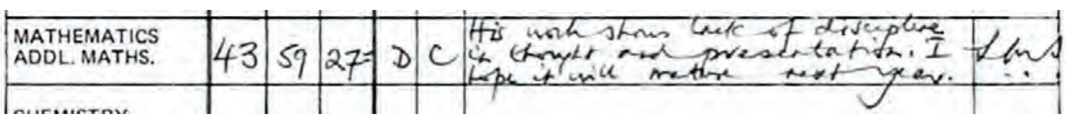
Many behavioural and social science students (and researchers for that matter) despise statistics. Most of us have a non-mathematical background, which makes understanding complex statistical equations very difficult. Nevertheless, the Evil Horde (a.k.a. professors) force our non-mathematical brains 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.

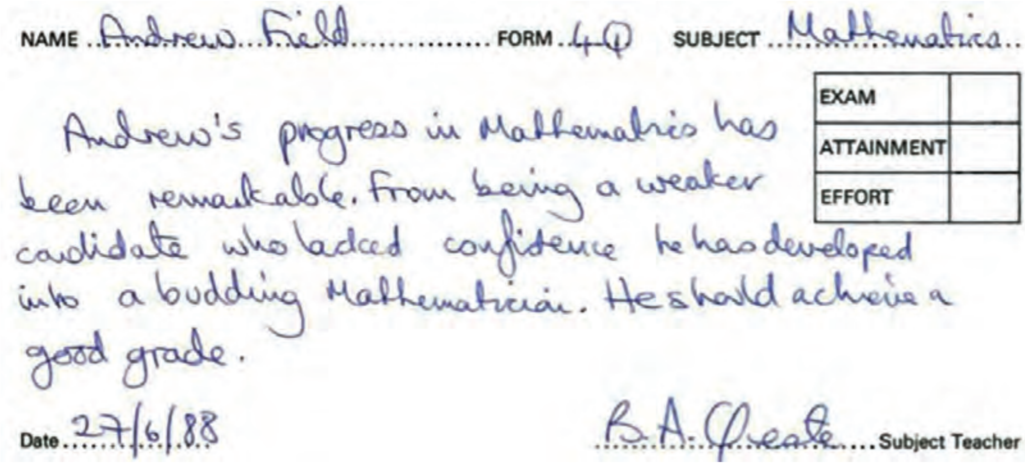
A second (ridiculously ambitious) aim is to make this the only statistics book that you'll ever need to buy (sort of). It's a book that I hope will become your friend from your first year at university right through to your professorship. The start of the book is aimed at first-year undergraduates (Chapters 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.	[illegible]

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 4Q 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

B.A. Cheate 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

$\sum$	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 *i*th 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,

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

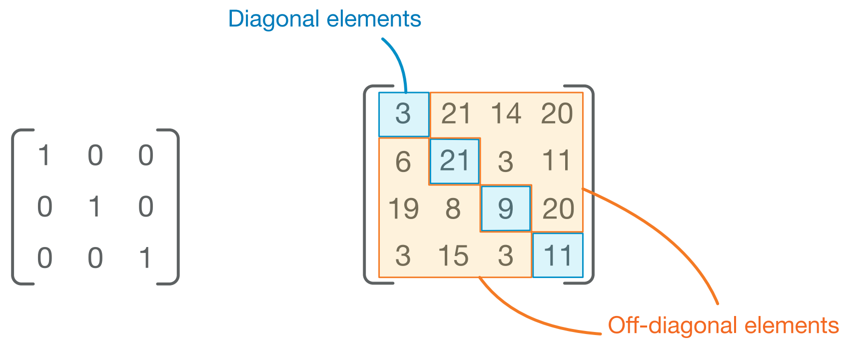
Matrices

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.

$$\begin{bmatrix} 2 & 5 & 6 \\ 3 & 5 & 8 \end{bmatrix} \qquad \begin{bmatrix} 3 & 21 & 14 & 20 \\ 6 & 21 & 3 & 11 \\ 19 & 8 & 9 & 20 \\ 3 & 15 & 3 & 11 \\ 23 & 1 & 14 & 11 \end{bmatrix}$$

2×3 matrix $\qquad$ 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.



Identity matrix $\qquad$ Square matrix

1

WHY IS MY EVIL LECTURER FORCING ME TO LEARN STATISTICS?

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.

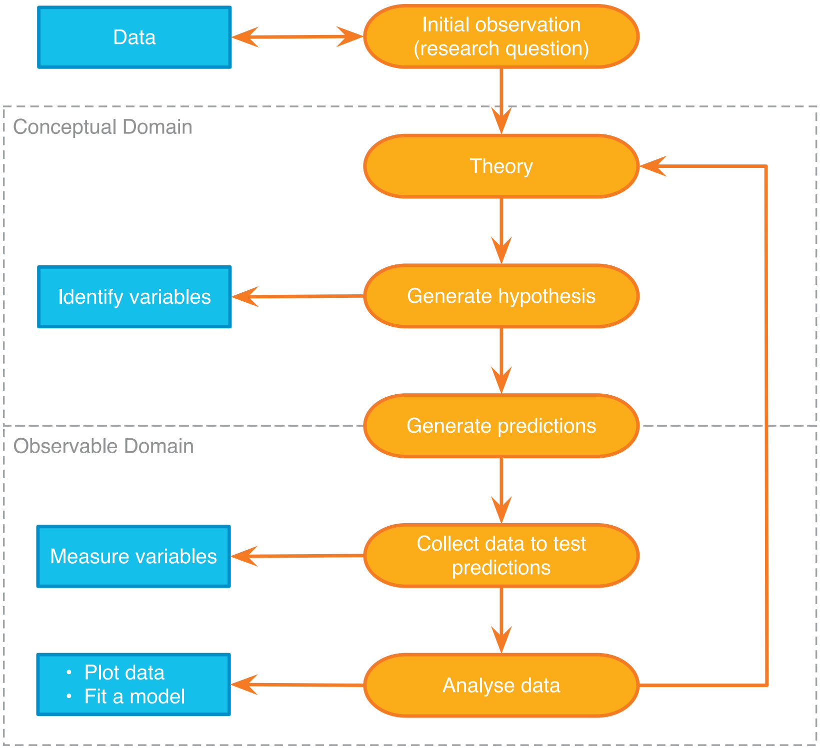


Figure 1.2 The research process

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'.

SELF TEST

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: