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

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)

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.

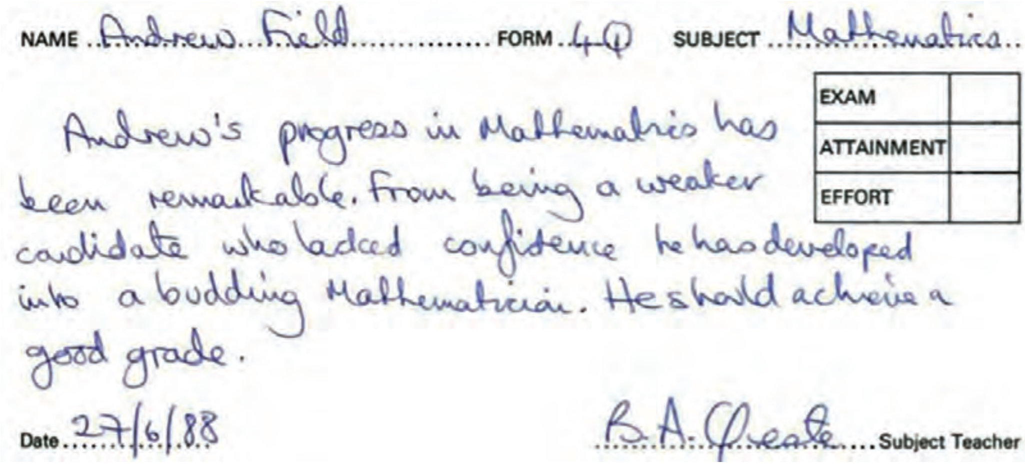
Too many books create the impression that there is a 'right' and 'wrong' way to do statistics. Data analysis is more subjective than you might think. Therefore, although I make recommendations, within the limits imposed by the senseless destruction of rainforests, I hope to give you enough background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

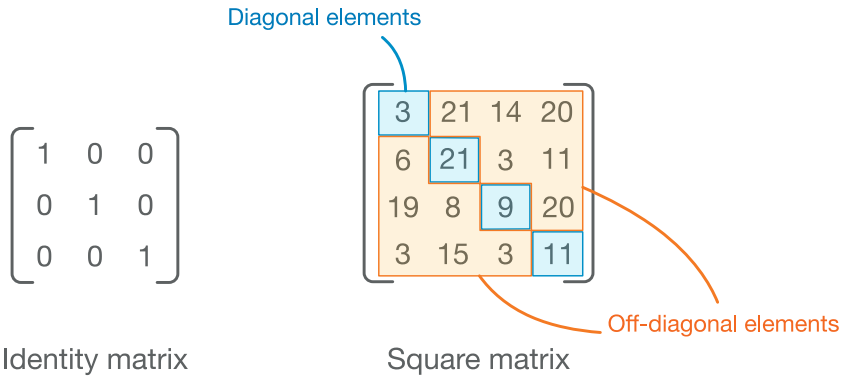
$$\begin{bmatrix} 2 & 5 & 6 \\ 3 & 5 & 8 \end{bmatrix}$$

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?

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.

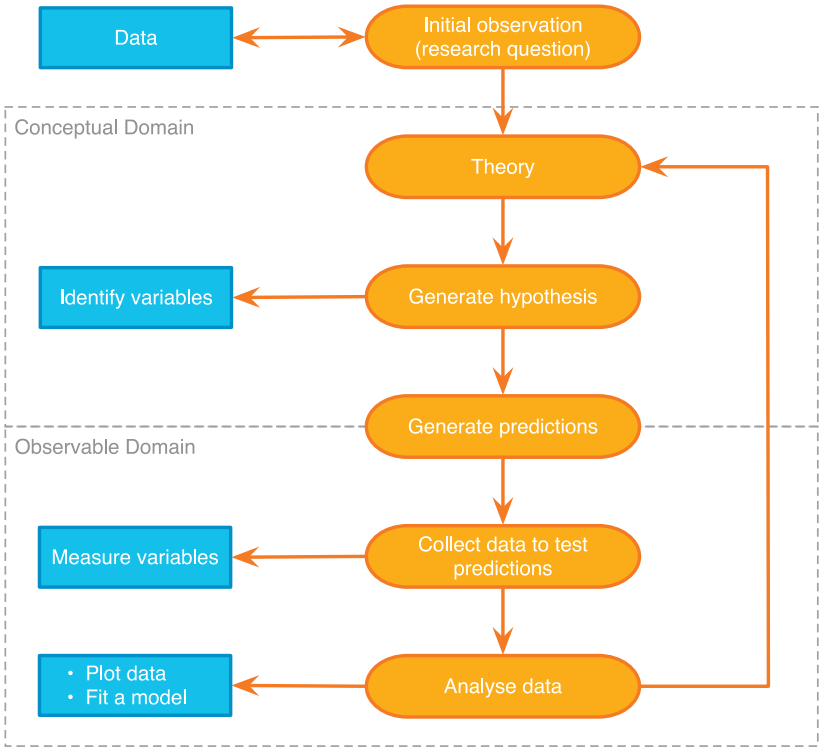


Figure 1.2 The research process

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.

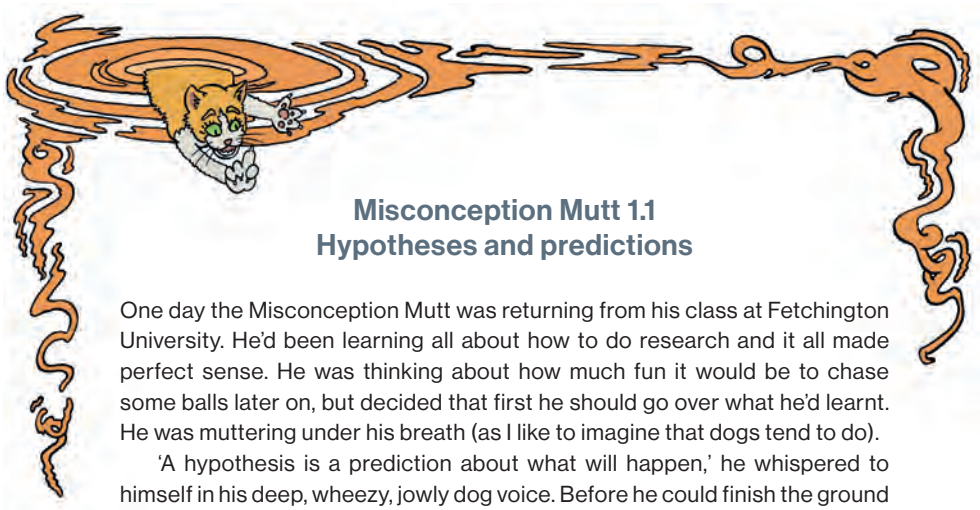
To continue the example, having studied the attachment theory of personality disorders, we might decide that this theory implies that people with personality disorders seek out the attention that a TV appearance provides because they lack close interpersonal relationships. From this we can generate a 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.

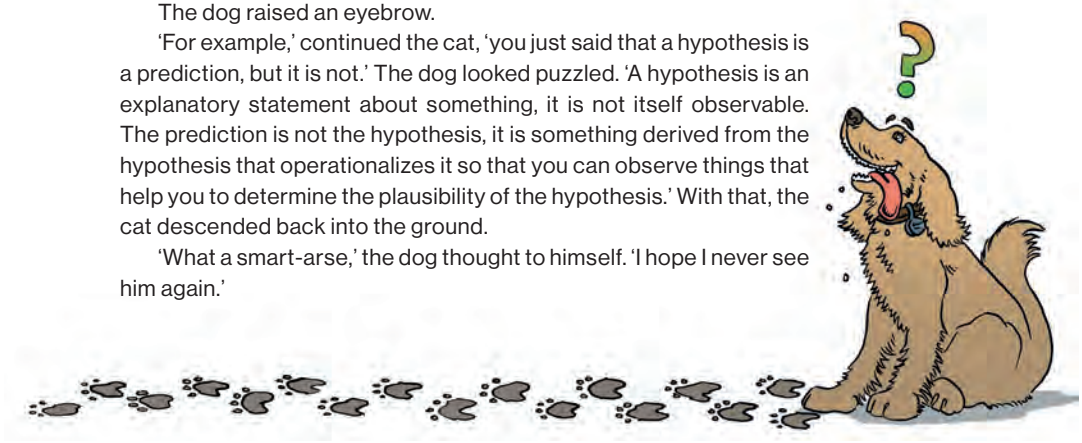
The mutt twitched as he inhibited the urge to chase the cat. 'Who are you?' he asked.

'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.

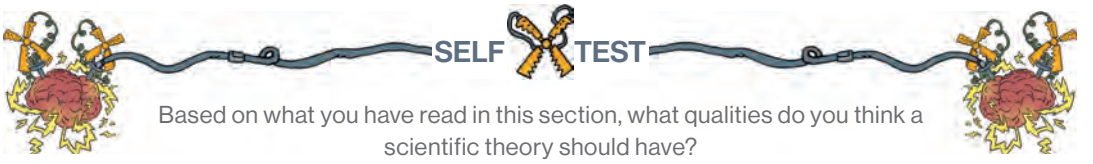
It's unlikely that we would be the only people interested in why individuals who go on reality TV have extreme personalities. Imagine that these other researchers discovered that: (1) people with narcissistic personality disorder think that they are more interesting than others; (2) they also think that

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'.

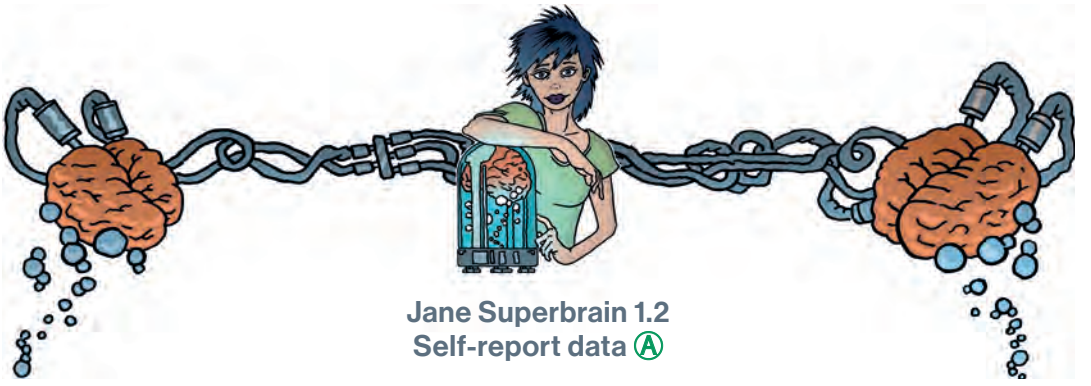
8 There is a long-standing urban myth that a post-coital douche with the contents of a bottle of Coke is an effective contraceptive. Unbelievably, this hypothesis has been tested and Coke does affect sperm motility, and some types of Coke are more effective than others – Diet Coke is best apparently (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.

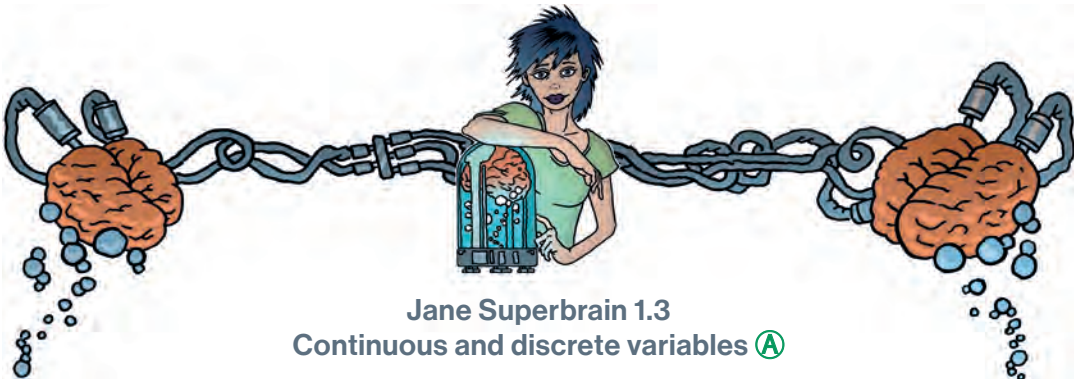
The categorical variables we have considered up to now have been unordered (e.g., different brands of Coke with which you're trying to kill sperm), but they can be ordered too (e.g., increasing concentrations of Coke with which you're trying to skill sperm). When categories are ordered, the variable 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

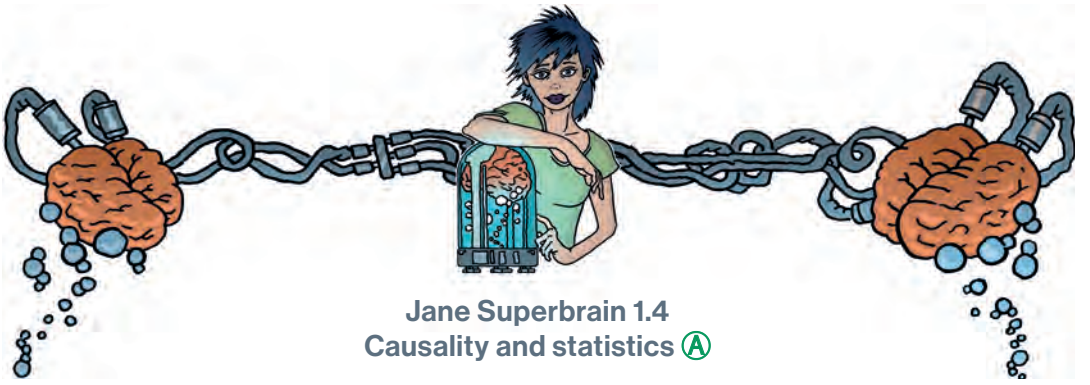
12 As you might imagine, his view was a lot more complicated than this definition, but let's not get sucked down that particular wormhole.

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.

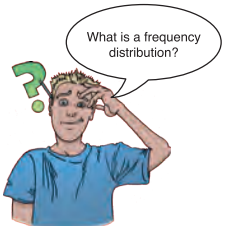
If we turn our attention to independent designs, a similar argument can be applied. We know that participants in different experimental conditions will differ in many respects (their IQ, attention span, etc.). We know that these confounding variables contribute to the variation between conditions, but 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.

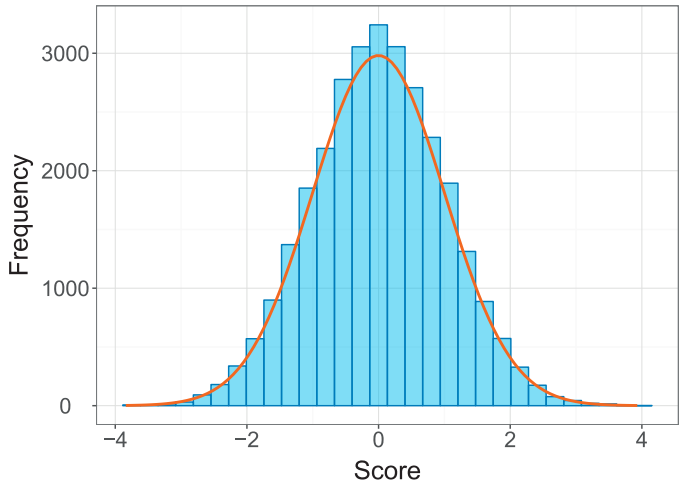


Figure 1.3 A 'normal' distribution (the curve shows the idealized shape)

(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.

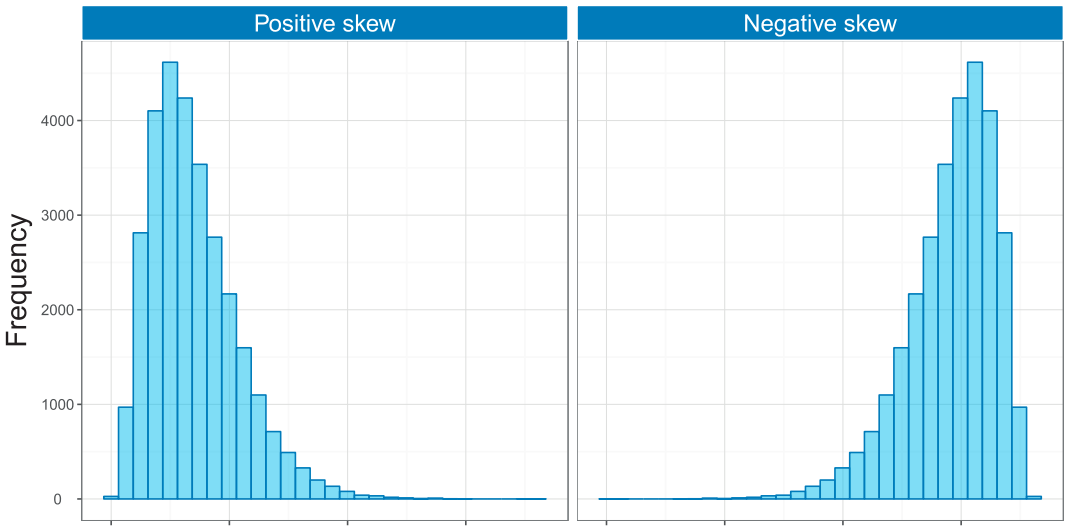


Figure 1.4 A positively (left) and negatively (right) skewed distribution

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).

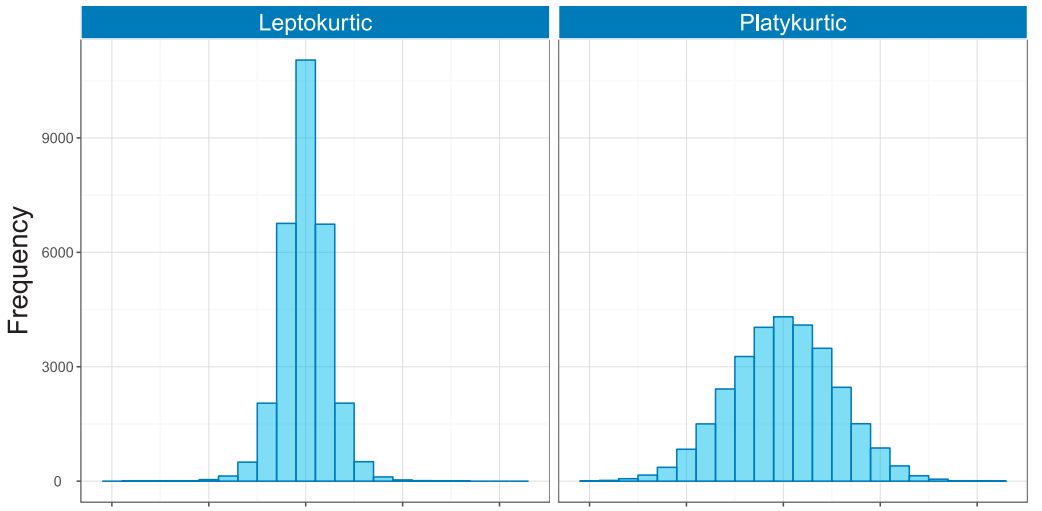


Figure 1.5 Distributions with positive kurtosis (leptokurtic, left) and negative kurtosis (platykurtic, right)

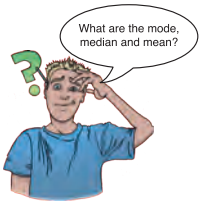
1.7.2 The mode Ⓐ

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

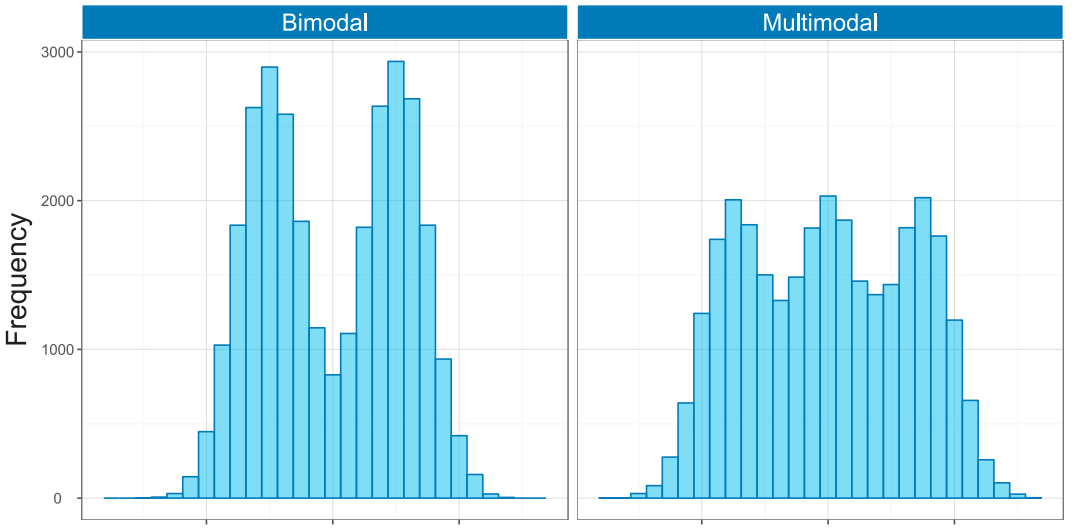


Figure 1.6 Examples of bimodal (left) and multimodal (right) distributions

$(n + 1)/2 = (11 + 1)/2 = 12/2 = 6$. 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).

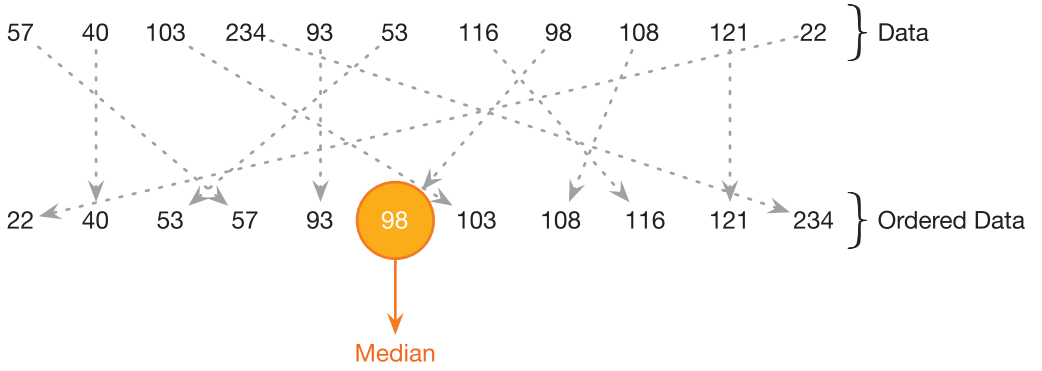


Figure 1.7 The median is simply the middle score when you order the data

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).

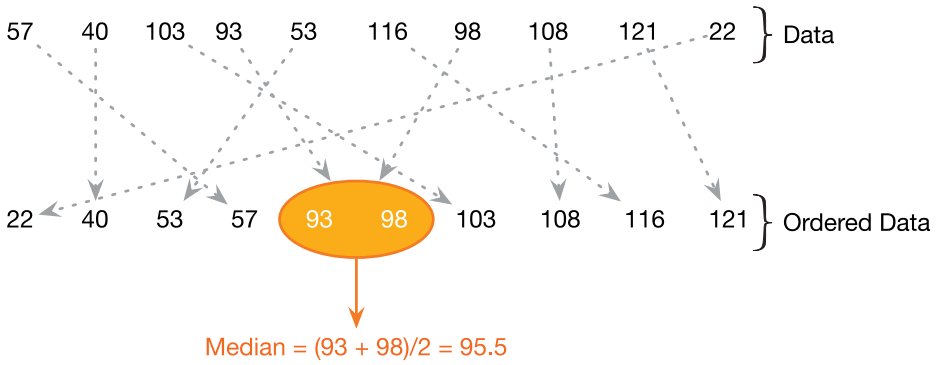


Figure 1.8 When the data contain an even number of scores, the median is the average of the middle two values

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.

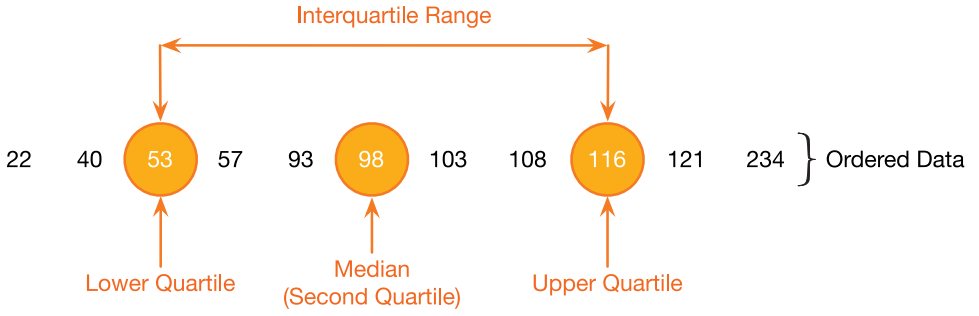


Figure 1.9 Calculating quartiles and the interquartile range

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.

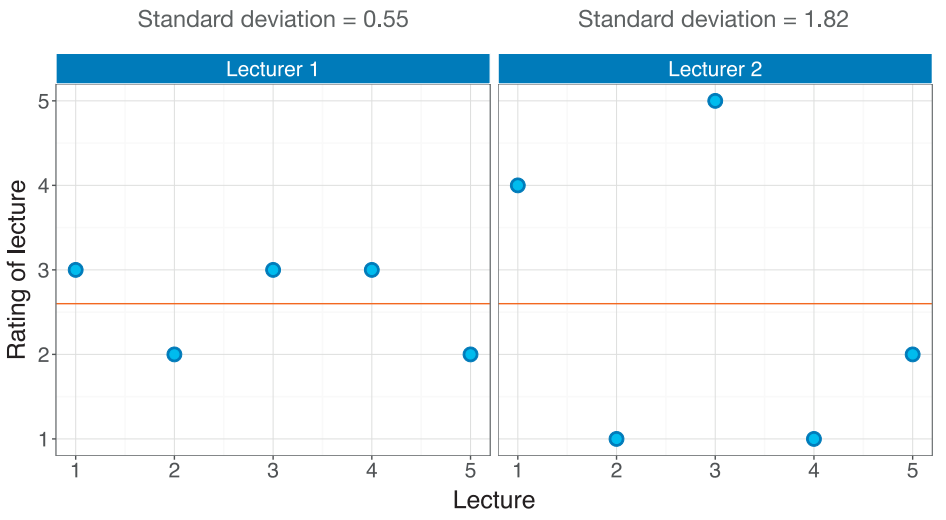


Figure 1.10 Graphs illustrating data that have the same mean but different standard deviations

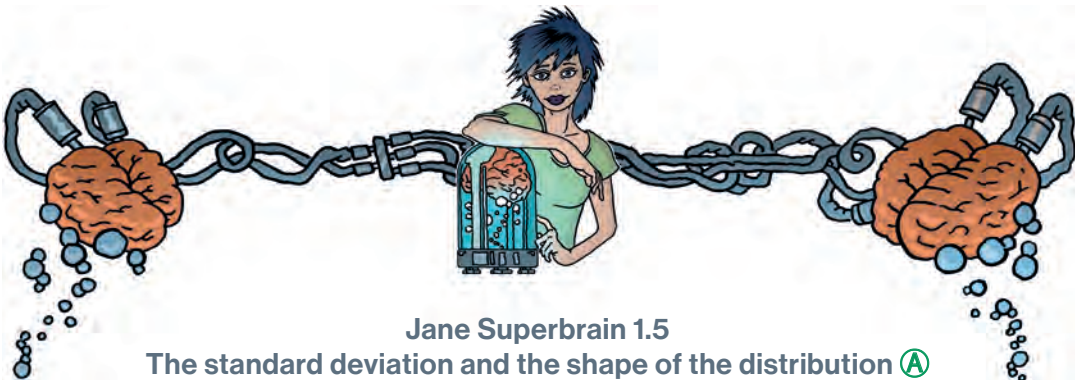
1.7.6 Using a frequency distribution to go beyond the data Ⓐ

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.

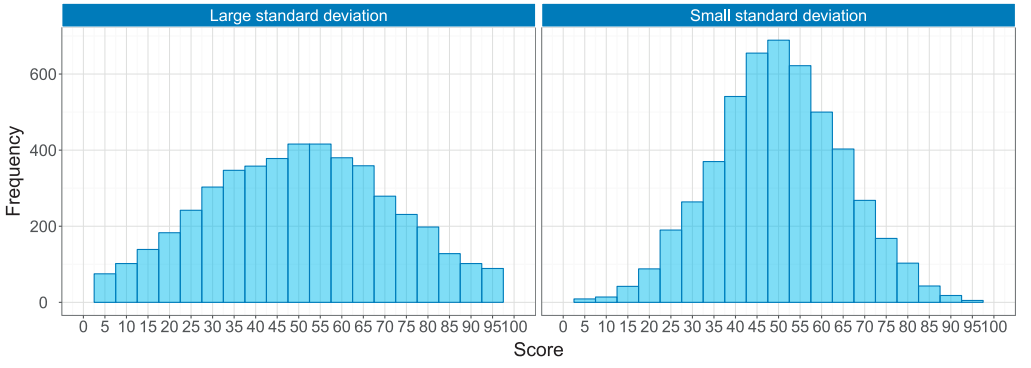


Figure 1.11 Two distributions with the same mean, but large and small standard deviations

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

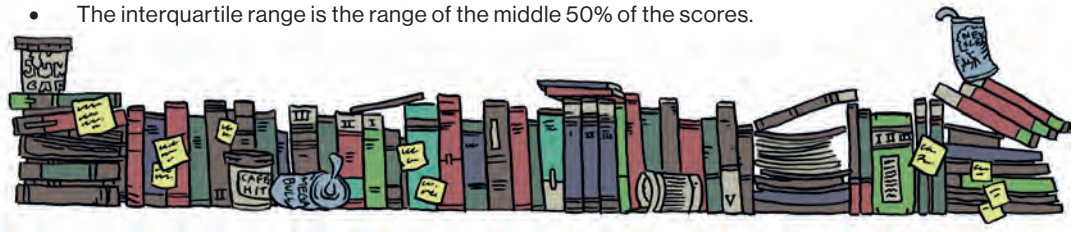
Calculate the mean, median, standard deviation and interquartile range for each type of accident and on each date. You can view the solutions at www.discoverspss.com/solutions/

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

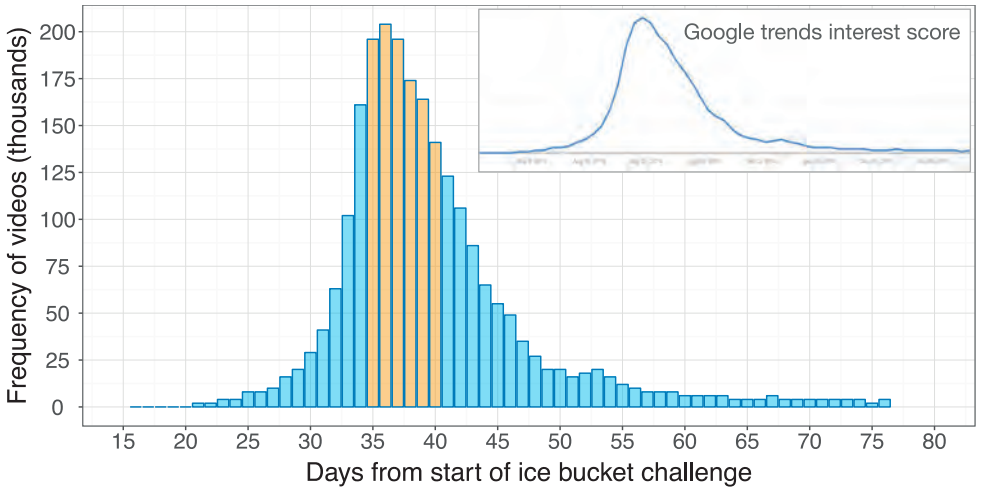


Figure 1.12 Frequency distribution showing the number of ice bucket challenge videos on YouTube by day since the first video (the inset shows the actual Google Trends data on which this example is based)

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.

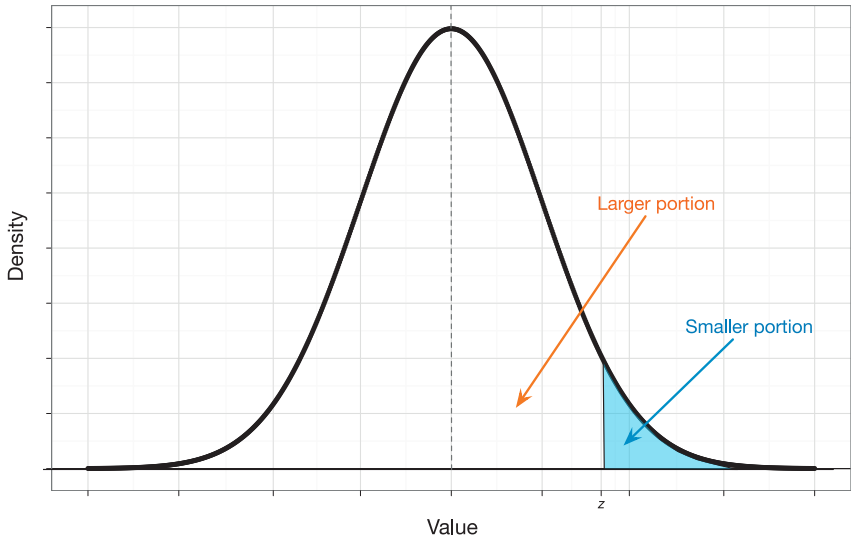
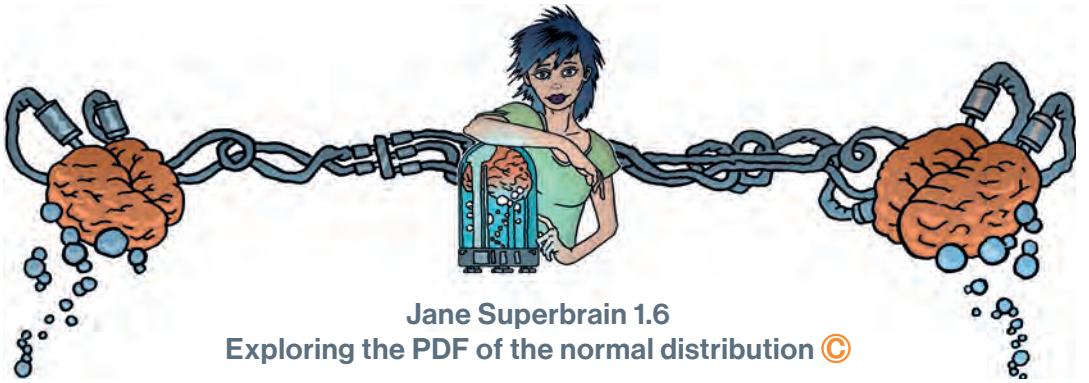


Figure 1.13 The normal probability distribution

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:

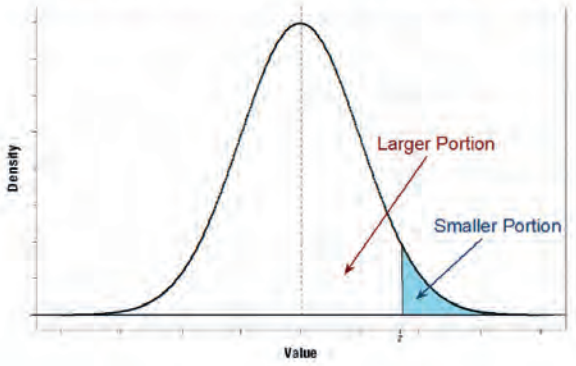
$$f(x=2;\mu=0,\sigma=1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{2-0}{1}\right)^2}=\frac{1}{\sqrt{2\pi}}e^{-2}=0.054$$

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?'

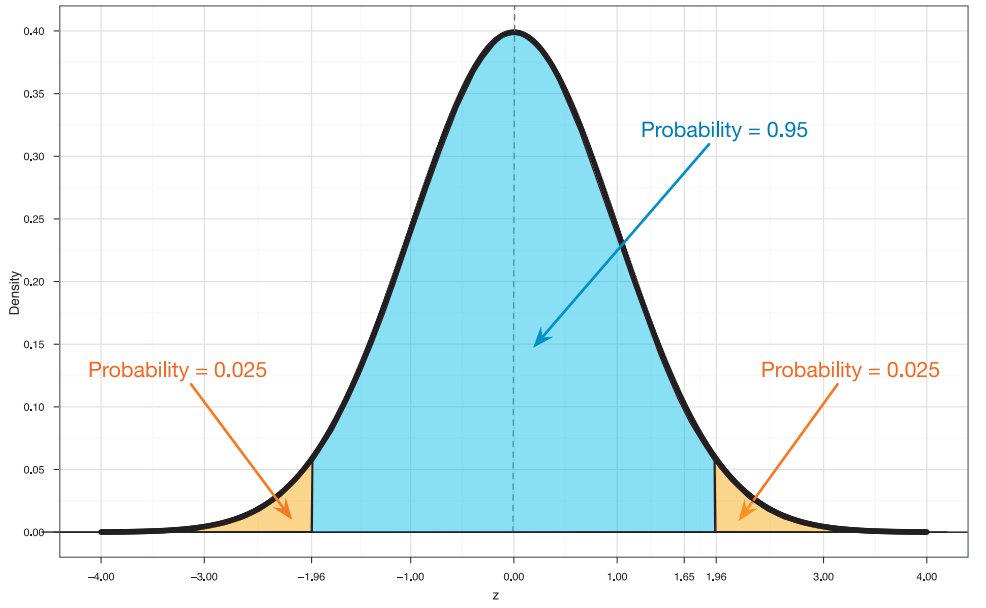
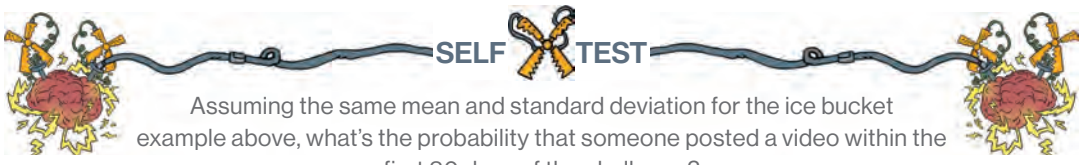


Figure 1.15 The probability density function of a normal distribution

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.

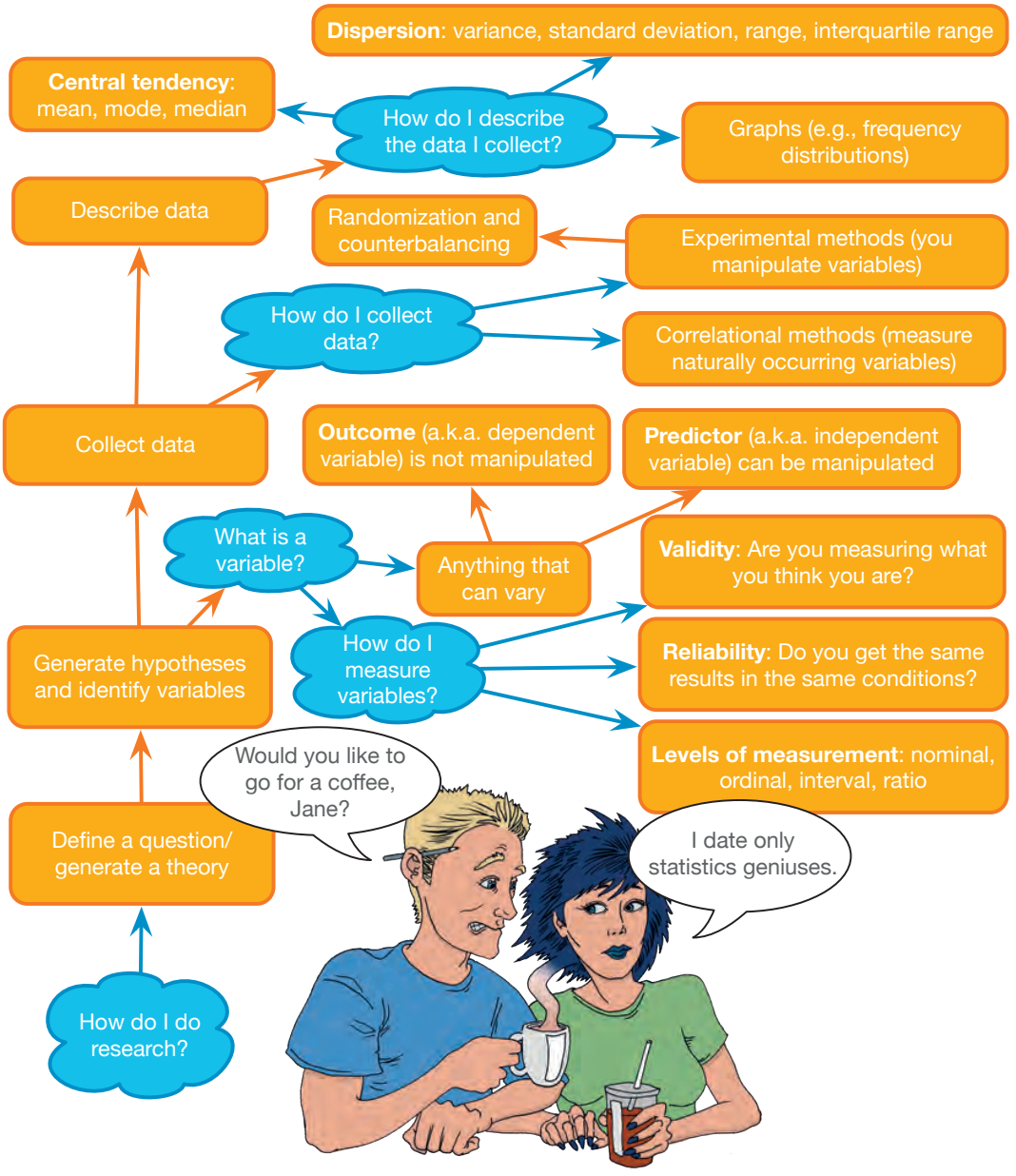


Figure 1.16 What Brian learnt from this chapter

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.

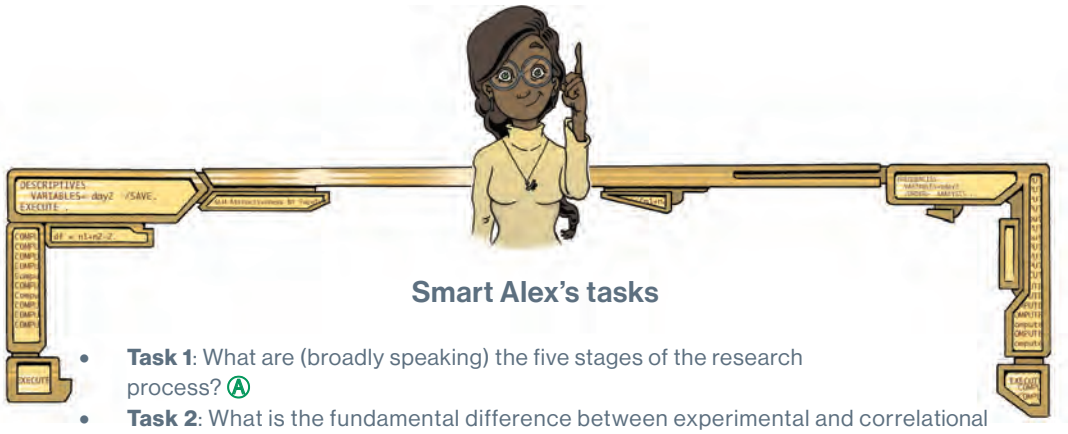
1.10 What next? Ⓐ

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



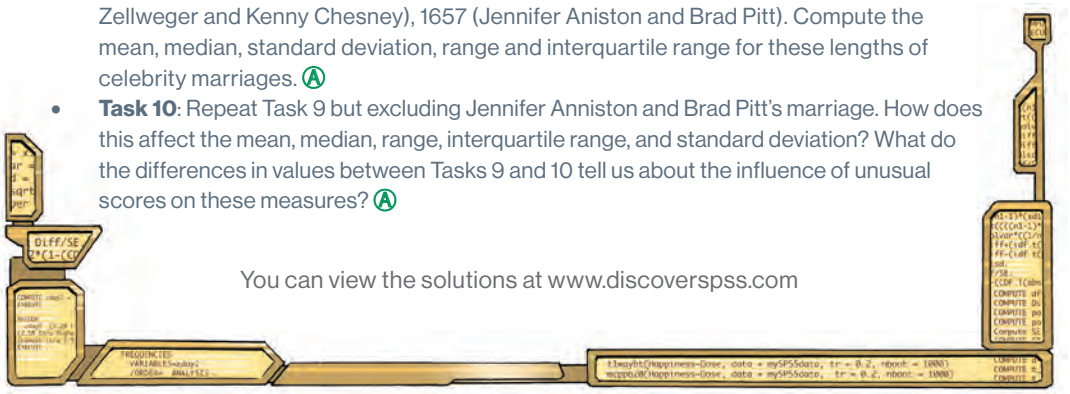
Smart Alex's tasks

- **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

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.

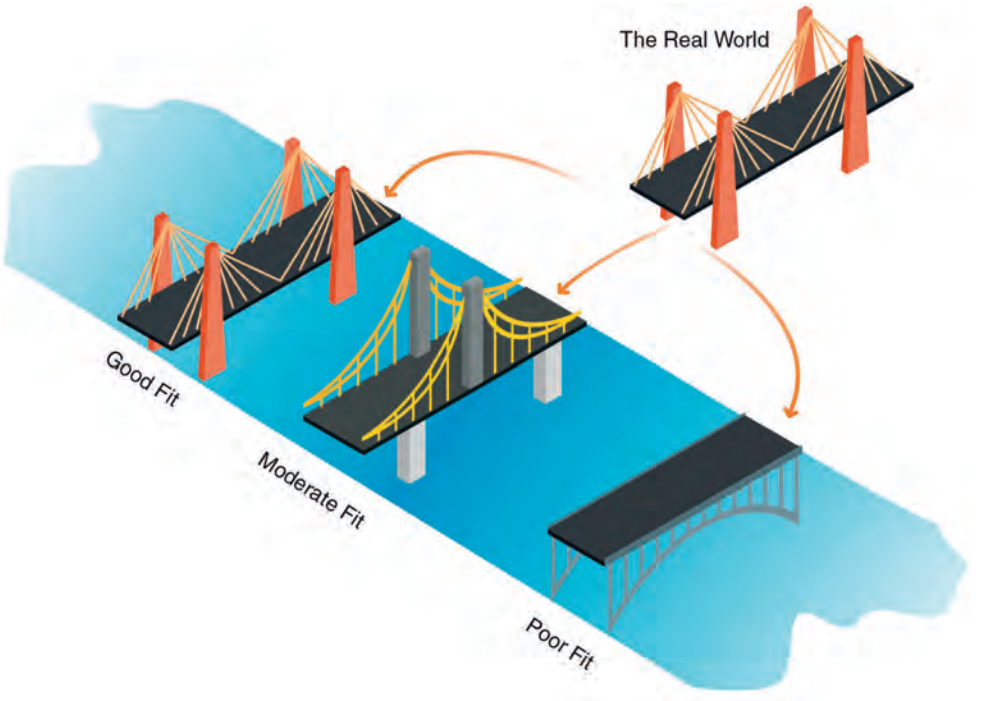
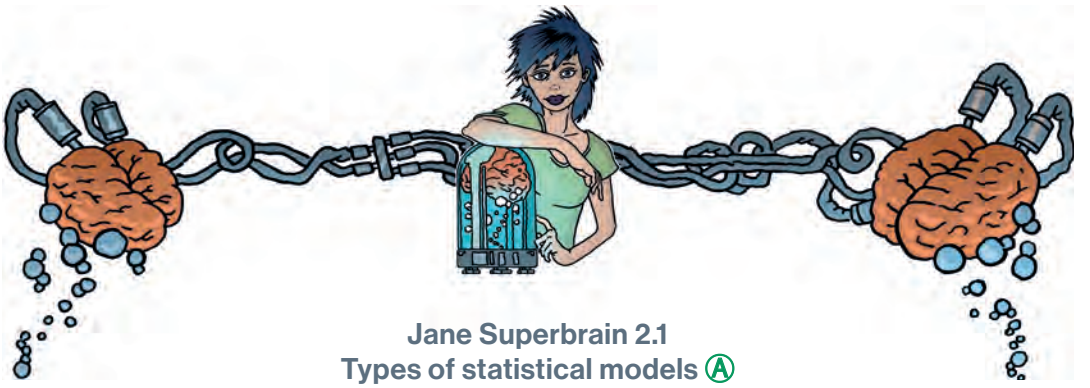


Figure 2.2 Fitting models to real-world data (see text for details)

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).

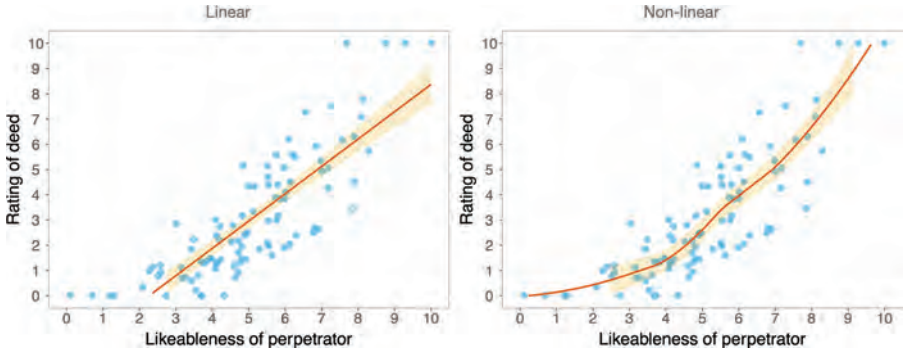


Figure 2.3 A scatterplot of the same data with a linear model fitted (left), and with a non-linear model fitted (right)

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.

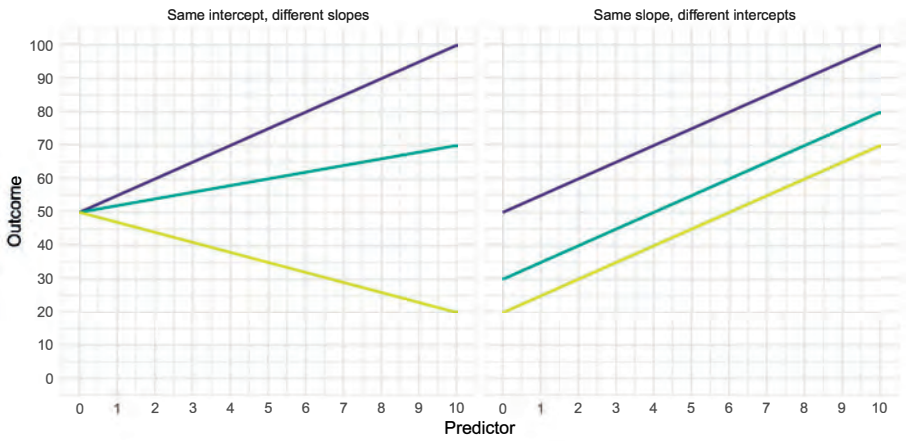


Figure 2.5 Lines that share the same intercept but have different gradients, and lines with the same gradients but different intercepts

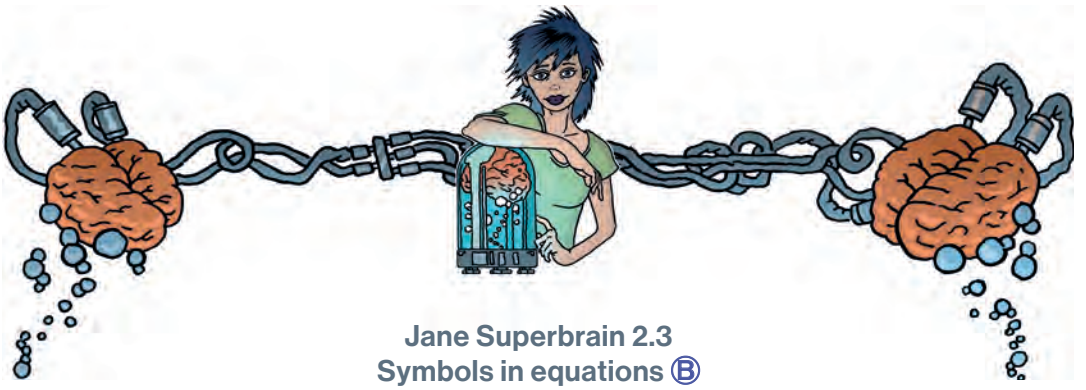
2.6 P is for parameters Ⓐ

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.

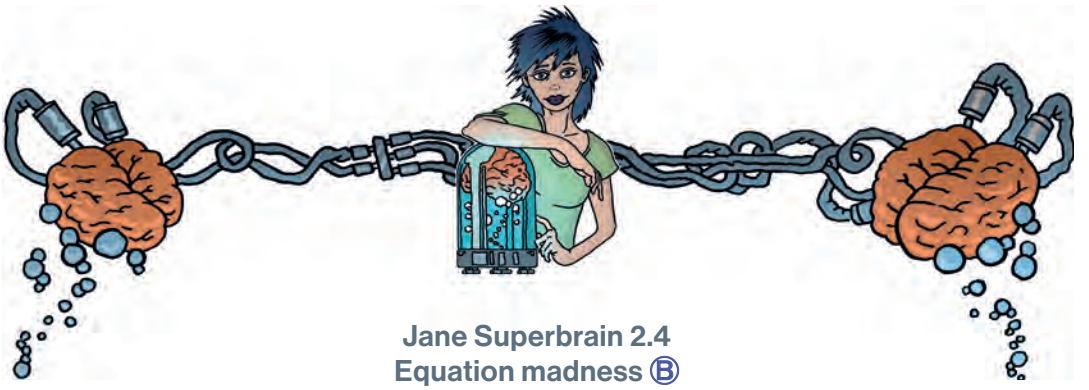
Hopefully what you can take from this section is that this book boils down to the idea that we can predict values of an outcome variable based on a model. The form of the model changes, there will always be some error in prediction, and there will always be parameters that tell us about the shape or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

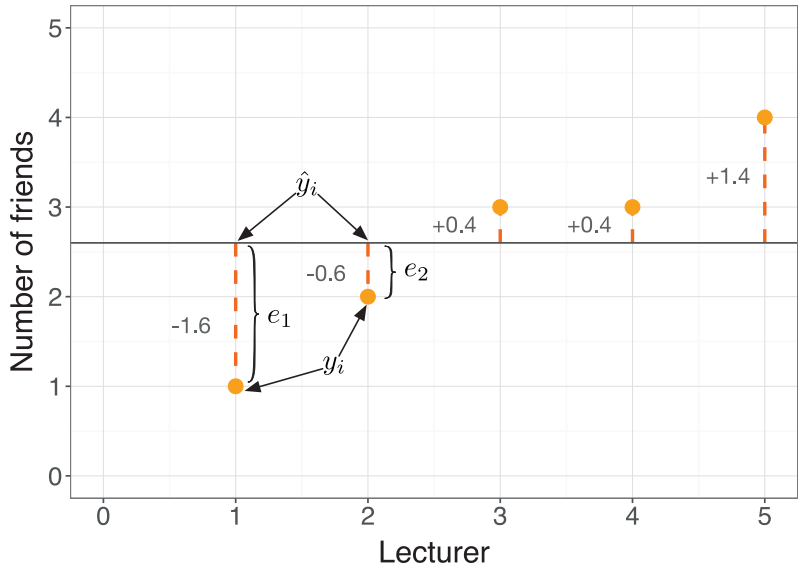


Figure 2.7 Plot showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends

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

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:

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

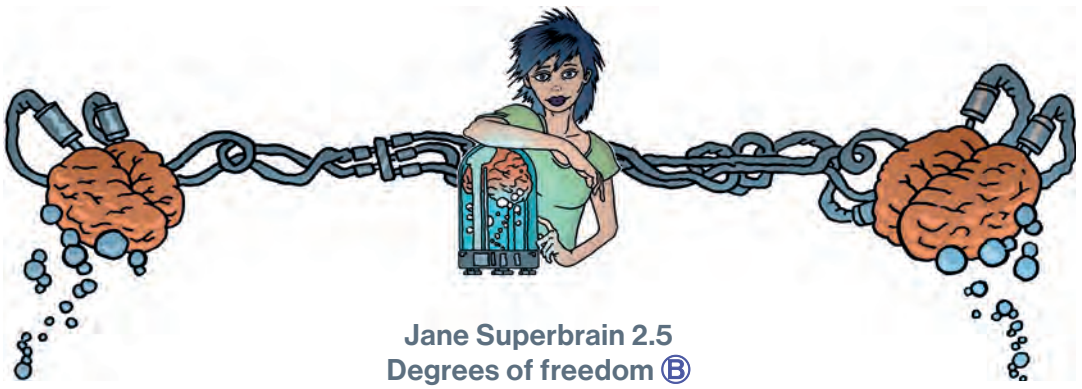
This equation shows how something we have used before (the sum of squares) can be used to assess the total error in any model (not just the mean).

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

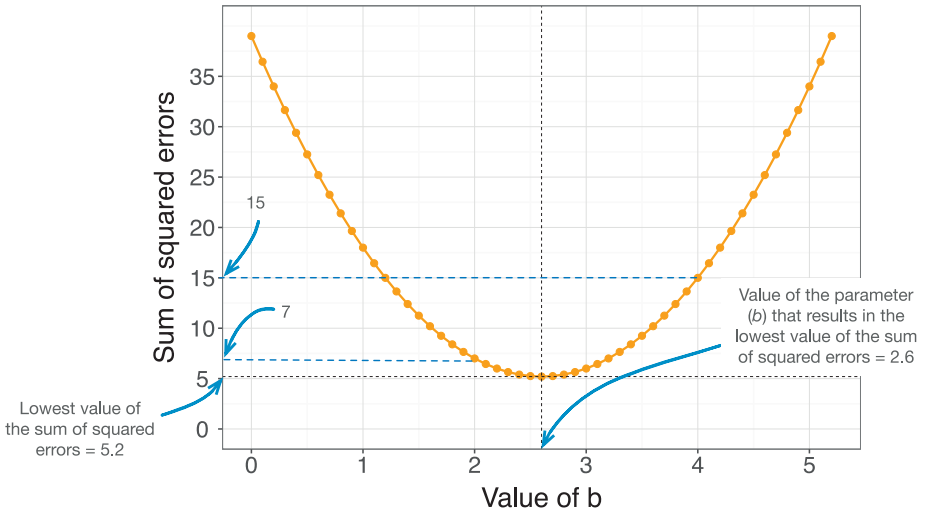


Figure 2.8 Plot showing the sum of squared errors for different 'guesses' of the mean

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.

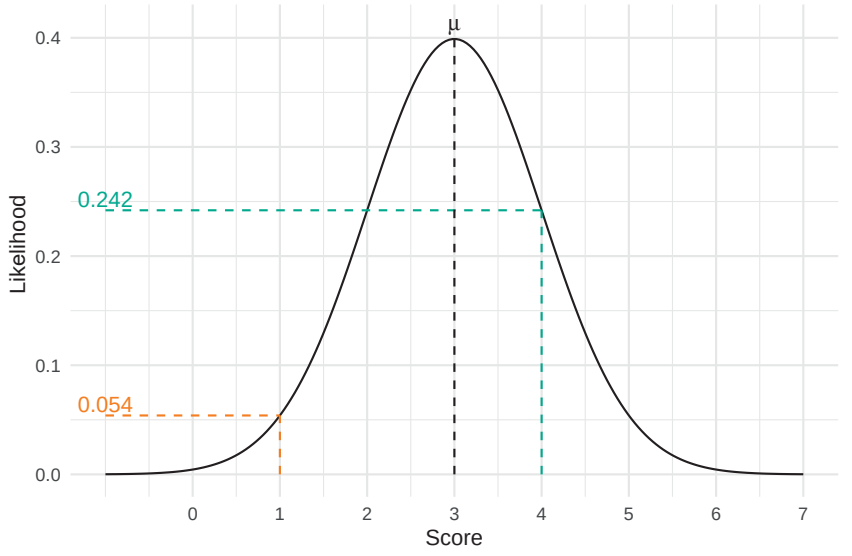


Figure 2.9 Illustration of the likelihood

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
		$\prod_{i=1}^{n} L_1 = 3.54 \times 10^{-4}$		$\prod_{i=1}^{n} L_2 = 1.64 \times 10^{-5}$

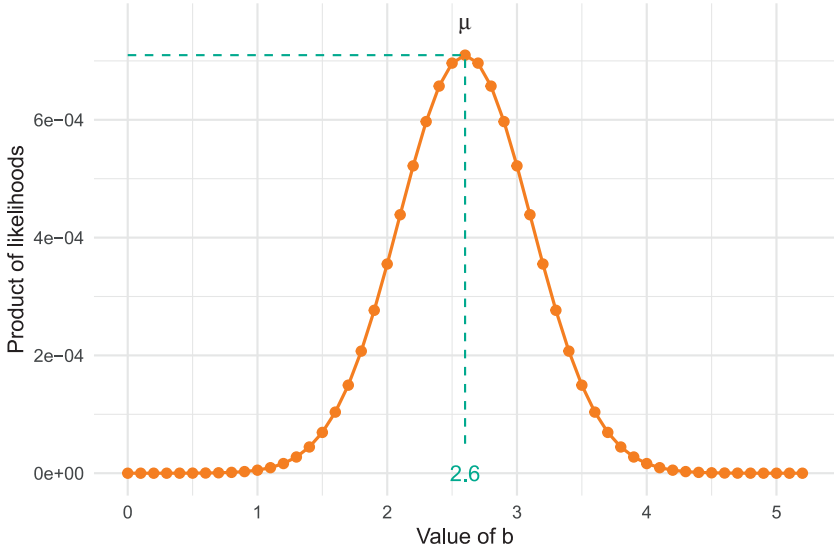


Figure 2.10 Plot showing the likelihood for different 'guesses' of the mean

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

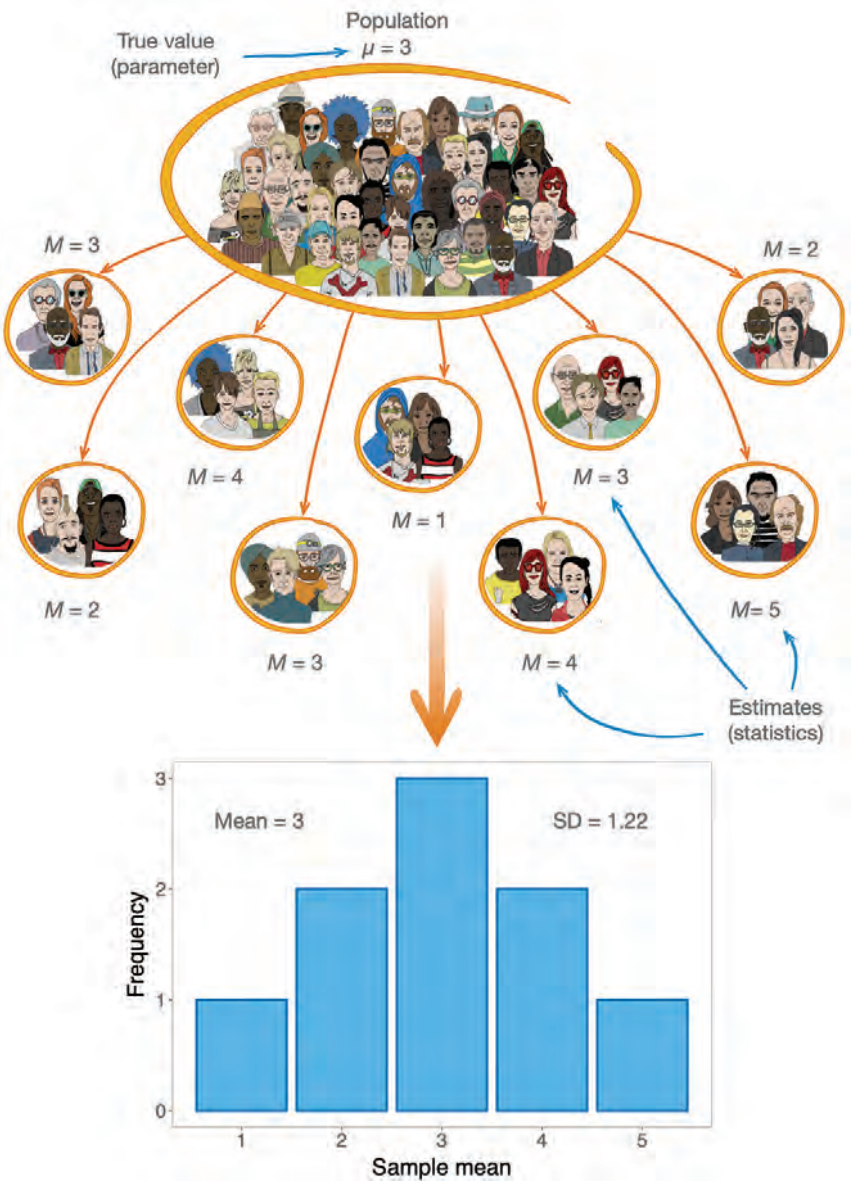
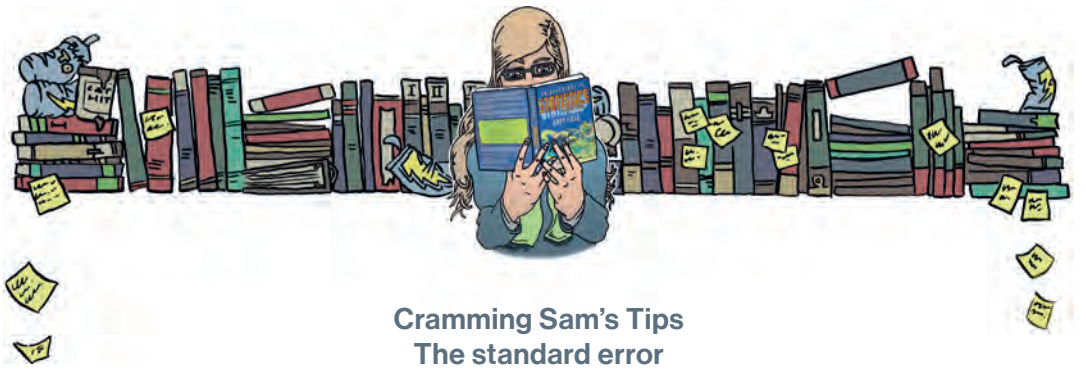


Figure 2.11 Illustration of the standard error (see text for details)

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

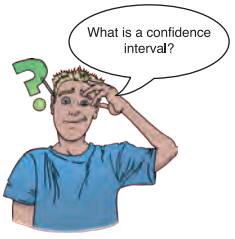
- The standard error of the mean is the standard deviation of sample means. As such, it is a measure of how representative of the population a sample mean is likely to be. A large standard error (relative to the sample mean) means that there is a lot of variability between the means of different samples 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.

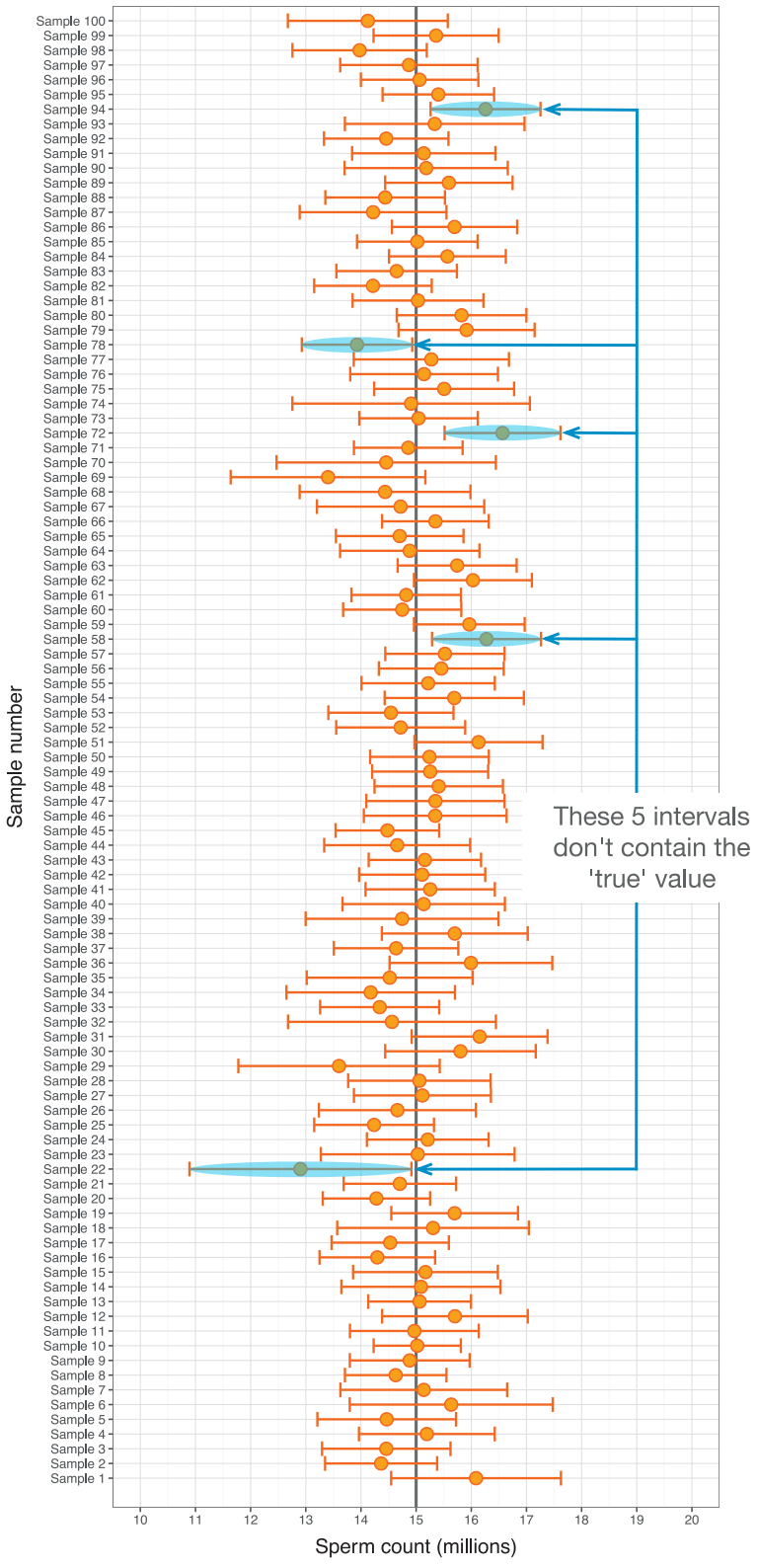


Figure 2.12 The confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 100 different samples (vertical axis)

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:

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

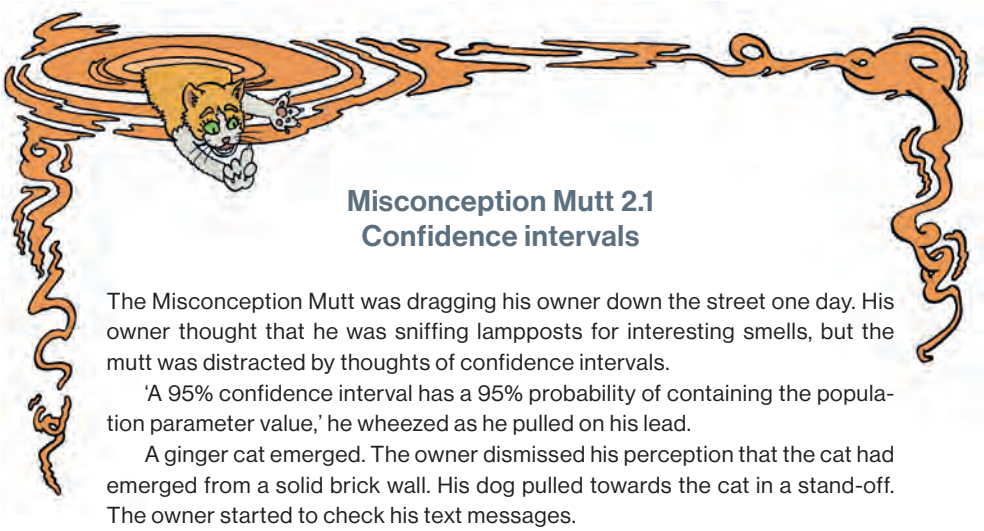
$$X = \bar{X} + (1.96 \times s) \qquad X = \bar{X} - (1.96 \times s)$$

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:

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

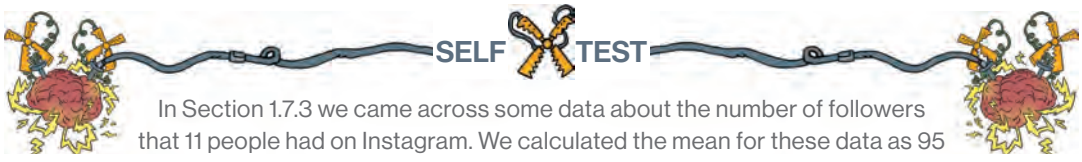
in which p is the probability value for the confidence interval. So, if you want a 95% confidence interval, then you want the value of *z* for (1–0.95)/2 = 0.025. Look this up in the 'smaller portion' column of the table of the standard normal distribution (look back at Figure 1.14) and you'll find that *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.

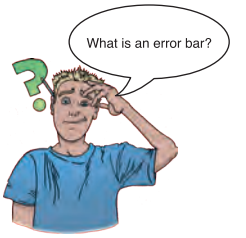


In Section 1.7.3 we came across some data about the number of followers that 11 people had on Instagram. We calculated the mean for these data as 95 and standard deviation as 56.79.

- 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.

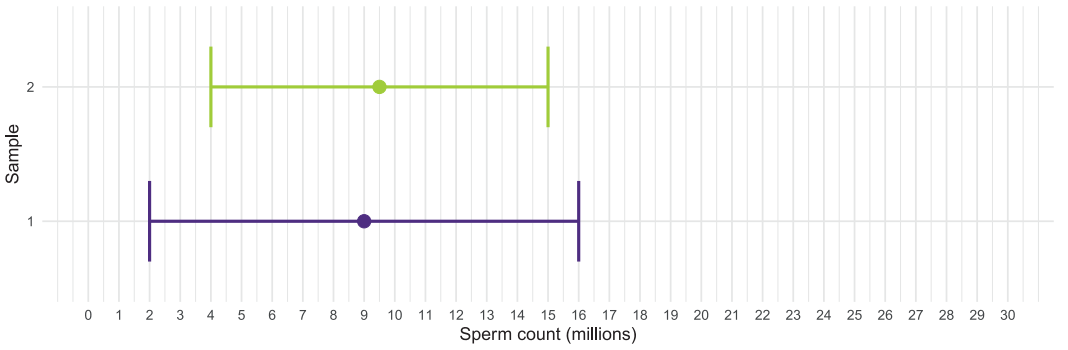


Figure 2.13 Two overlapping 95% confidence intervals

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

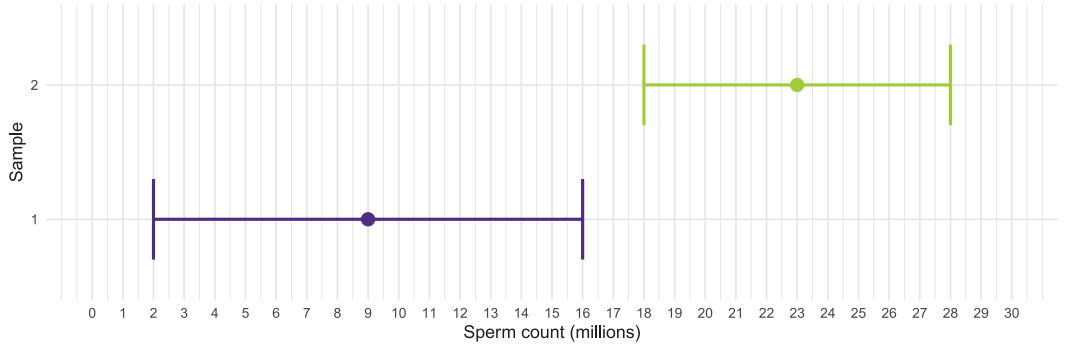


Figure 2.14 Two 95% confidence intervals that don't overlap

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.

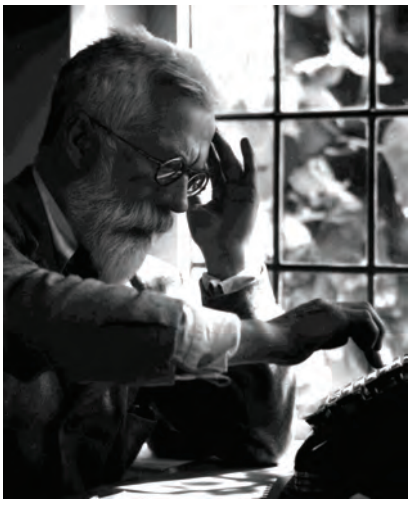


Figure 2.15 Sir Ronald A. Fisher

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.