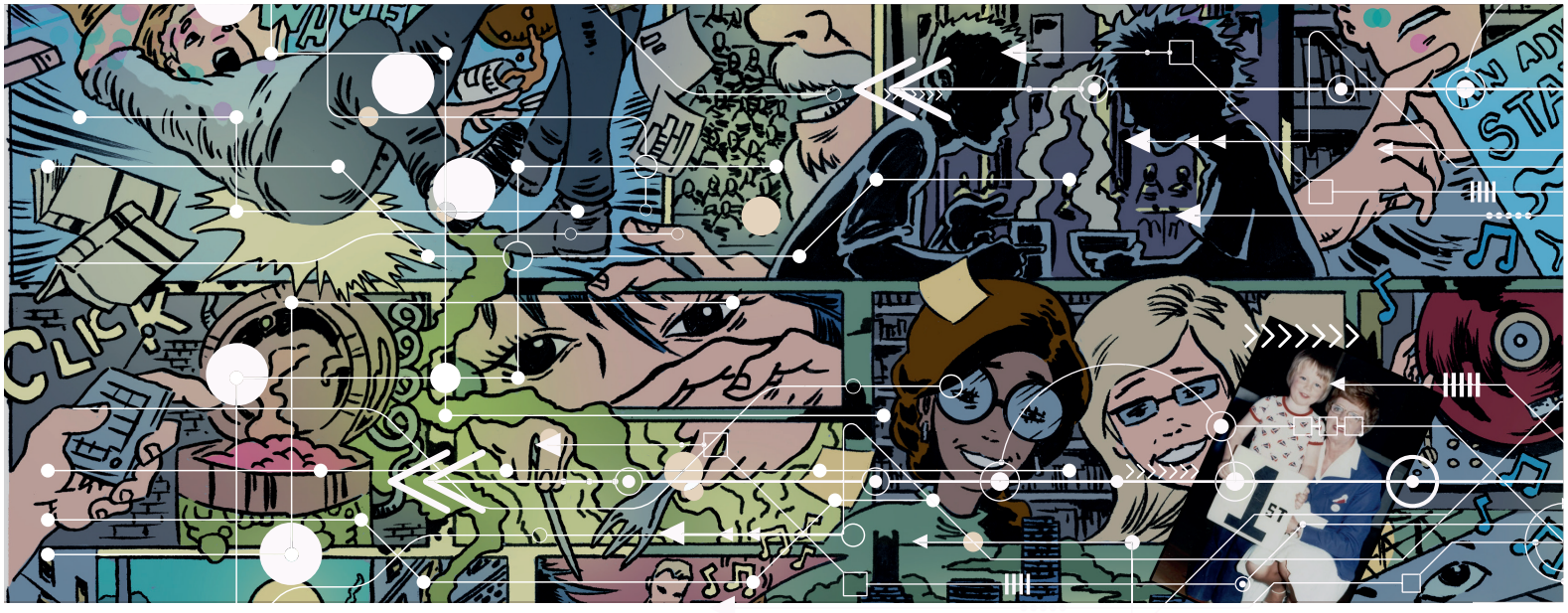
DISCOVERING STATISTICS USING IBM SPSS STATISTICS

CATISFIED CUSTOMERS

5TH EDITION

DISCOVERING STATISTICS USING IBM SPSS STATISTICS



ANDY FIELD

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

First edition published 2000
Second edition published 2005
Third editon published 2009. Reprinted 2009, 1010, 2011 (twice), 2012
Fourth edition published 2013. Reprinted 2014 (twice), 2015, 2016, 2017

Library of Congress Control Number: 2017954636

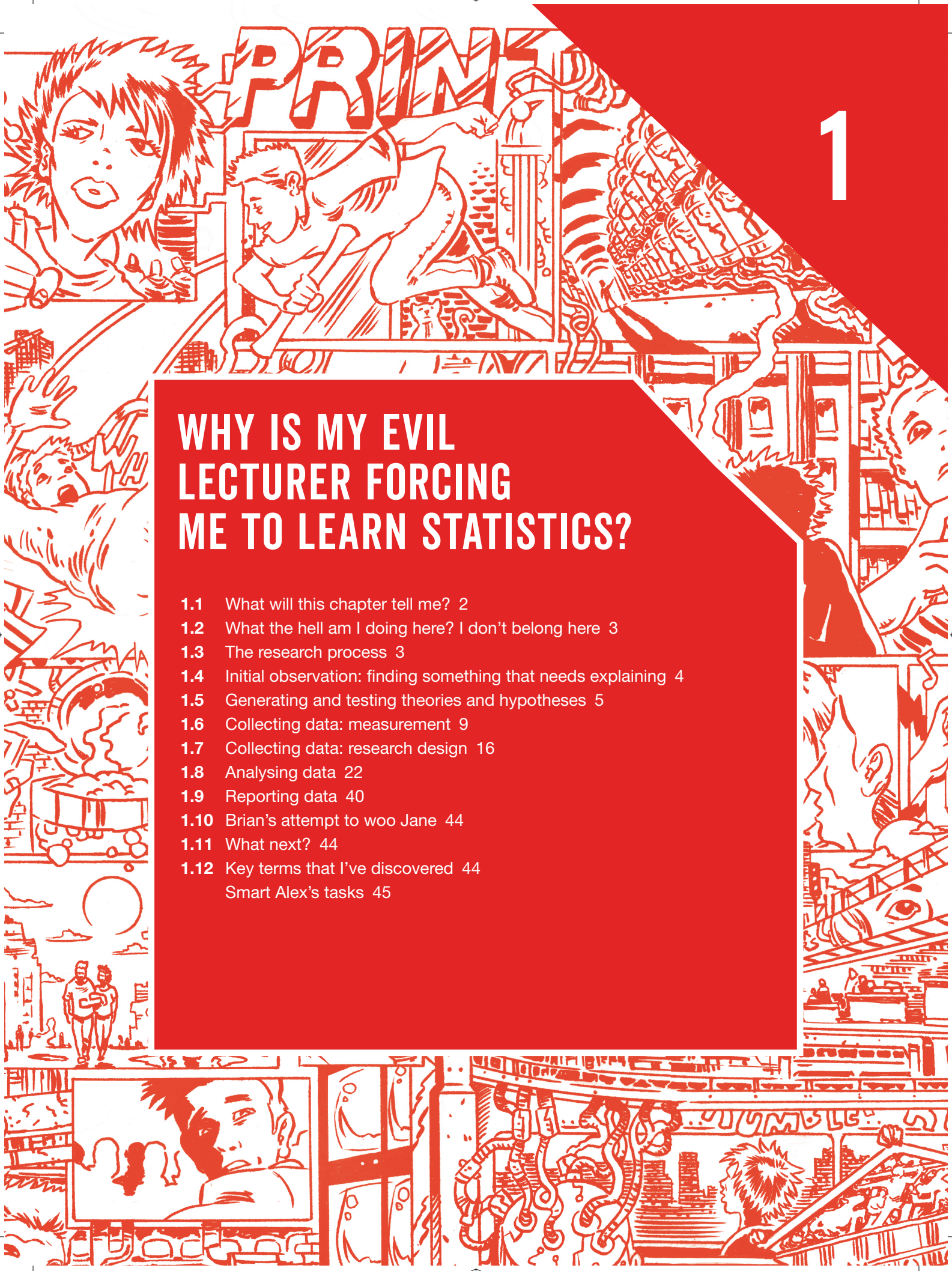
British Library Cataloguing in Publication data

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

Editor: Jai Seaman
Development editors: Sarah Turpie & Nina Smith
Assistant editors, digital: Chloe Statham
Production editor: Ian Antcliff
Copyeditor: Richard Leigh
Indexer: David Rudeforth
Marketing manager: Ben Griffin-Sherwood
Cover design: Wendy Scott
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in Germany by: Mohn Media Mohndruck GmbH

Illustrated by: James Iles

ISBN 978-1-5264-1951-4
ISBN 978-1-5264-1952-1 (pbk)



1

WHY IS MY EVIL LECTURER FORCING ME TO LEARN STATISTICS?

1.1 What will this chapter tell me?

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 driving my dad mad 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 are 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 He was practising 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. Had he hit me with the club end rather than the shaft he probably would have. Fortunately (for me, but not for you), I survived, although some might argue that this incident explains the way my brain functions.

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

1.2 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), 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 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]

1.3 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 explanations (hypotheses) for those observations, from which you can make predictions. To

3 This is a joke. Like many of my jokes, there are people who won't find it remotely funny. Passions run high between qualitative and quantitative researchers, so its inclusion will likely result in me being hunted down, locked in a room and forced to do discourse analysis by a horde of rabid qualitative researchers.

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

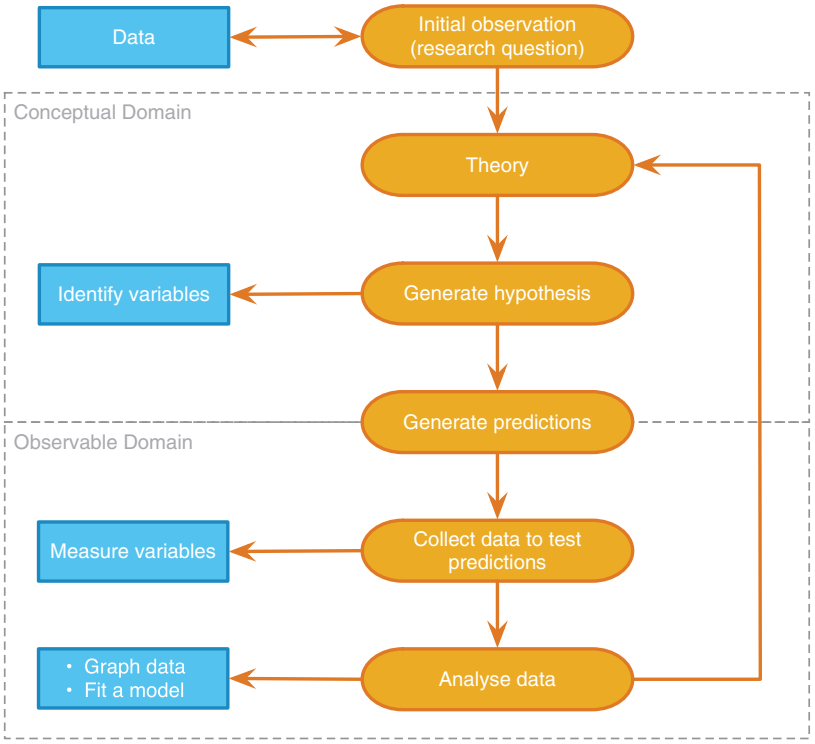


Figure 1.2 The research process

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.

1.4 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.

Having made a casual observation about the world (reality TV contestants on the whole 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 define one or more **variables** to measure that quantify the thing

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, and exploitation of others.

I'm trying to measure. There's one variable in this example: the personality of the contestant. I could measure this variable by giving them 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 did have narcissistic personality disorder. These data support my observation: a lot of reality TV contestants have extreme personalities.

1.5 Generating and testing theories and hypotheses

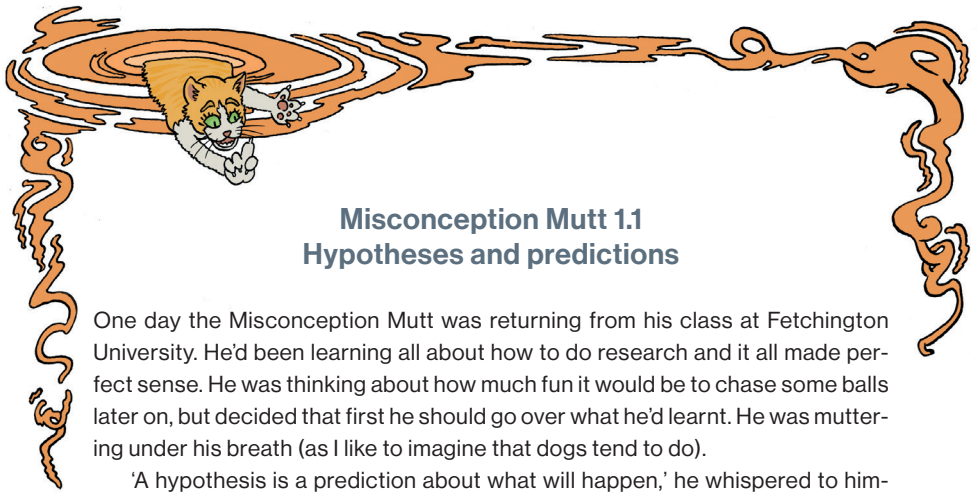
The next logical thing to do is to explain these data (Figure 1.2). The first step is to look for relevant theories. A **theory** is an explanation or set of principles that is well substantiated by repeated testing and explains a broad phenomenon. We might begin by looking at theories of narcissistic 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, Johnson, Clouthier, Scala, & Temes, 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. 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. 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



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.

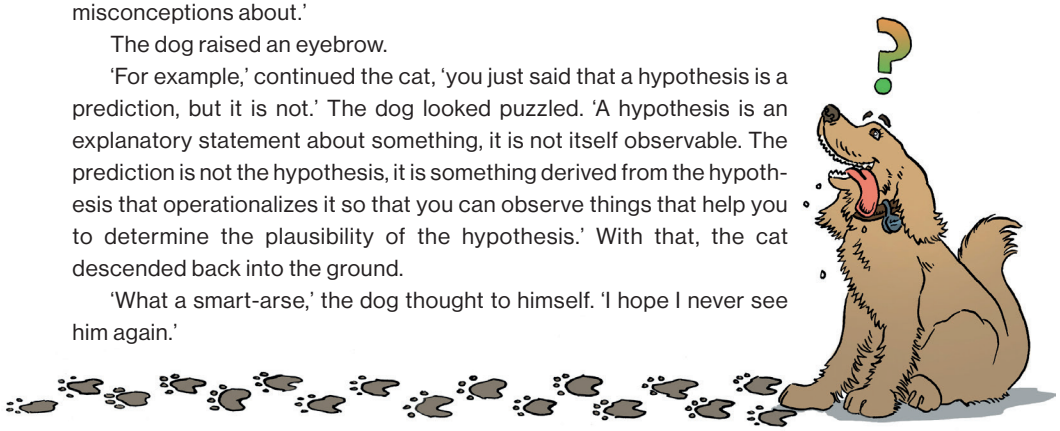
The mutt twitched as he inhibited the urge to chase the cat. 'Who are you?' he asked.

'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.

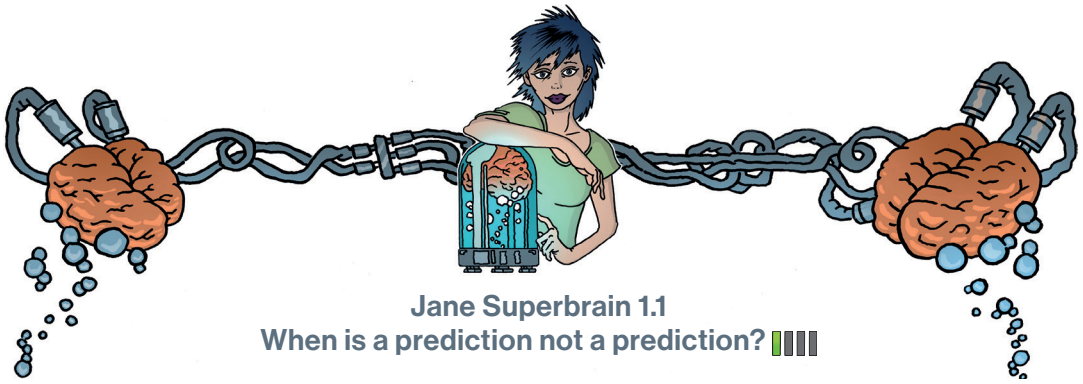
'For example,' continued the cat, 'you just said that a hypothesis is a prediction, but it is not.' The dog looked puzzled. 'A hypothesis is an explanatory statement about something, it is not itself observable. The prediction is not the hypothesis, it is something derived from the hypothesis that operationalizes 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.'



about 1%, so we'd be able to see whether the ratio of narcissistic personality disorder to not 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.



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 global warming. 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 were nonsense and 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

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

6 It was pretty awesome actually.

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 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 prediction 2 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.

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 that 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 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 all of the data,

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

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 to complete the questionnaire about themselves but answering from the perspective of each of their housemates. (So, for every contestant there is a measure of what they thought of every other contestant, and also a measure of 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 'Reality TV contestants who have personality disorders overestimate how much other people like their personality characteristics'.

1.6 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.

1.6.1 Independent and dependent variables

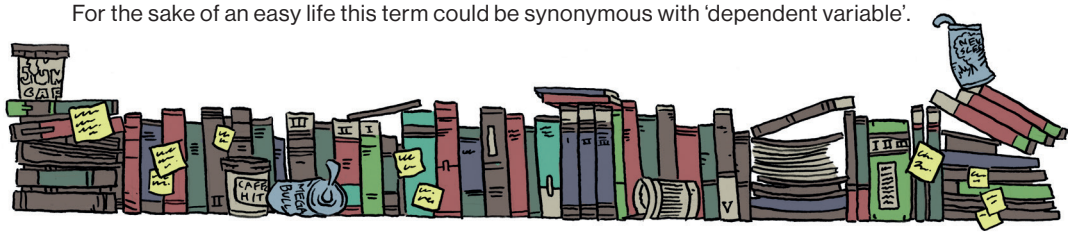
To test hypotheses we need to measure variables. Variables are things that can change (or vary); they might vary between people (e.g., IQ, behaviour) or locations (e.g., unemployment) or even time (e.g., mood, profit, number of cancerous cells). Most hypotheses can be expressed in terms of two variables: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

8 Actually, 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 (movement), and some types of Coke are more effective than others – Diet Coke is best, apparently (Umpierre, Hill & Anderson, 1985). In case you decide to try this method out, I feel it worth mentioning that despite the effects on sperm motility a Coke douche is ineffective at preventing pregnancy.

Cramming Sam's Tips
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'.



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.

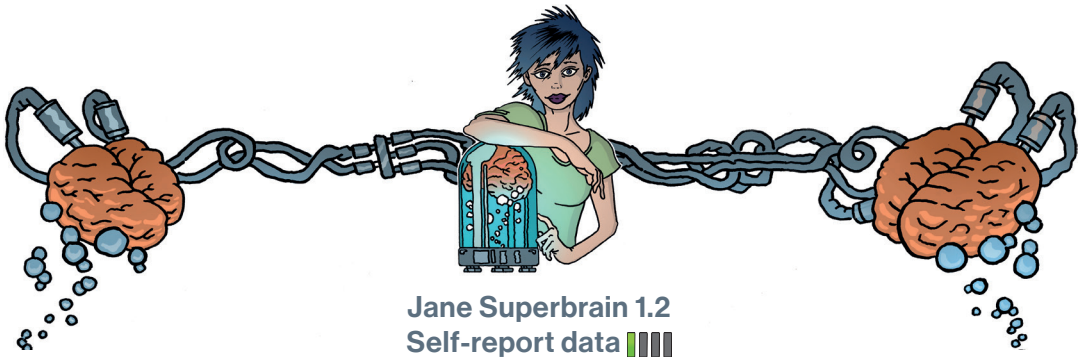
A variable that we think is a cause is known as an **independent variable** (because its value does not depend on any other variables). A variable that we think is an effect is called a **dependent variable** because the value of this variable depends on the cause (independent variable). These terms are very closely tied to experimental methods in which the cause is manipulated by the experimenter (as we will see in Section 1.7.2). 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 1.7), for which it doesn't make sense to talk of dependent and independent variables because all variables 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.

1.6.2 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, etc.). 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 (despite many fantasies to the contrary) a catwoman exist (not even one in a PVC suit). A categorical variable is one that names distinct entities. In its simplest form it names just two distinct types of things, for example male or female. 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**.

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 ugly-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



Jane Superbrain 1.2
Self-report data

A lot of self-report data are ordinal. Imagine two judges on *The X Factor* 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 that they found her equally talented? Probably not: their ratings will depend on their subjective feelings about what constitutes talent (the quality of singing? showmanship? 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.

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

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.

So far, the categorical variables we have considered 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 *The X Factor*, *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 'luck' more than hard work.[10] Imagine that the three winners of a particular *X Factor* series were Billie, Freema and Elizabeth. The names of the winners don't provide any information about where they came in the contest; however, labelling them according to their performance does – first, second and third. These categories are ordered. In using ordered categories we now know that the woman who won was better than the women who came second and third. We still know nothing about the differences between categories, though. 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 Elizabeth, 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 rely on having data measured at this level at least. 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.

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 with a 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 does mean a complete absence of time.

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

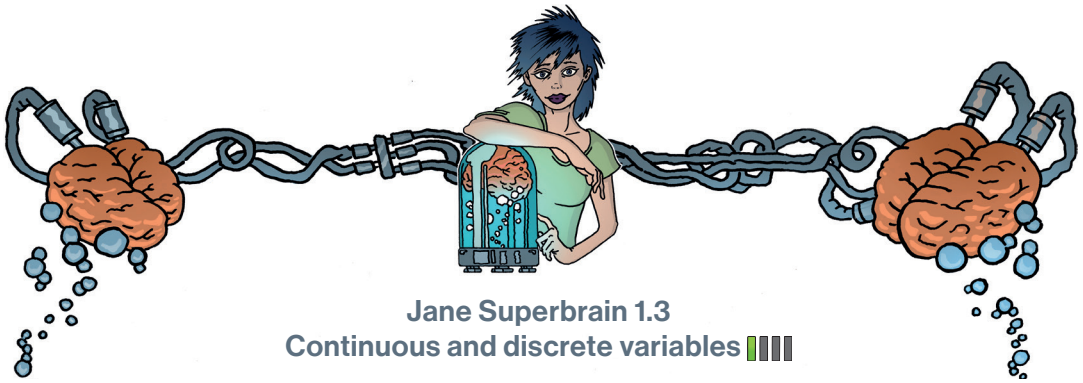
10 I am in no way bitter about spending years learning musical instruments and trying to create original music, only to be beaten to musical fame and fortune by 15-year-olds who can sing, sort of.

precision, whereas a **discrete variable** can take on only certain values (usually whole numbers) on the scale. What does this actually mean? Well, our example of rating lecturers on a 5-point scale is 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 exists underneath the scale (i.e., a rating of 3.24 makes sense), the actual values that the variable takes on are limited. A 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).

1.6.3 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. A few years passed, as you might expect given that Coke and sperm rarely top scientists' research lists, before another scientist, Dr Jack Q. Late, replicated the study. Dr Late found that sperm motility

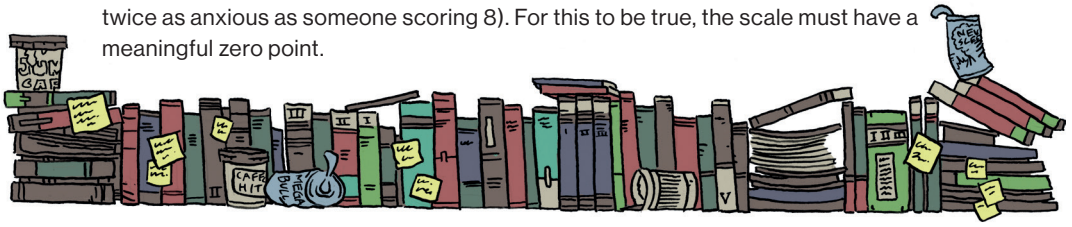


Jane Superbrain 1.3
Continuous and discrete variables

The distinction between continuous and discrete 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 years). Also, we often treat discrete variables as if they were continuous. For example, the number of boyfriends/girlfriends 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 boyfriends that women in their 20s 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 boyfriends.

Cramming Sam's Tips
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.



was 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 you had reported that you'd used 100ml of Coke and 5ml of sperm, then Dr Late could have ensured that he had used the same amount – because millilitres are a standard unit of measurement – we would know that Dr Late used exactly the same amount of Coke that we used. Direct measurements such as the millilitre provide an objective standard: 100ml of a liquid is known to be twice as much as only 50ml.

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 original study we measured motility using absorption spectrophotometry, whereas Dr Late used laser light-scattering techniques.[11] Perhaps his measure is more sensitive than ours.

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

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 83kg. One day you step on the bathroom scales and they read 80kg. There is a difference of 3kg between your actual weight and the weight given by your measurement tool (the scales): this is a measurement error of 3kg. 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 3kg), 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?

1.6.4 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 other factors other than the one we're interested in that can influence them.

Criterion validity is whether you can establish that an instrument measures what it claims to measure through comparison to objective criteria. In an ideal world, you assess this 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 ratemyprofessor.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 attitudes, you might be interested in the person's perception of reality and not reality itself (you might not care whether a person *is* a psychopath but whether they *think* they are a psychopath). With self-report measures/questionnaires we can also assess the degree to which individual items represent the construct being measured, and cover the full range of the construct (**content validity**).

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).

SELF TEST

What is the difference between reliability and validity?

1.7 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.

1.7.1 Correlational research methods

In correlational research we observe natural events; we can do this by either 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 (smoking, 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 like I'm absolutely obsessed with using Coke as a contraceptive (I'm not, but my discovery that people in the 1950s and 1960s actually tried this has, I admit, 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, etc.). 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. Even if we could, it would be totally unethical to insist on some people using Coke as a contraceptive (or indeed to do anything that would make a person likely to produce a child that they didn't intend to produce). However, there is a price to pay, which relates to causality: correlational research tells us nothing about the causal influence of variables.

1.7.2 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 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, some people will perceive themselves in a negative way (so it is something that varies); and, for the outcome, some people will get more anxious on dates than others (again, this is something that 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 who have a low opinion of themselves 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' (1739–40/1965).[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 low self-esteem and feeling anxious when on a date co-occur, and that the low self-esteem emerged before the dating anxiety did.

In correlational research variables are often measured simultaneously. The first problem with doing this is that it provides no information about the contiguity between different variables: we might find from a questionnaire study that people with low self-esteem also have dating anxiety but we wouldn't know whether it was the low self-esteem or the dating anxiety that came first. Longitudinal research addresses this issue to some extent, but there is 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 (e.g., poor social skills which might make you feel generally worthless but also puts pressure on you in dating situations). Therefore, low self-esteem and dating anxiety do always co-occur (meeting Hume's definition of cause) but only because poor 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 having breast implants and suicide (Koot, Peeters, Granath, Grobbee, & Nyren, 2003). However, it is unlikely that having breast implants causes you to commit suicide – presumably, there is an external factor (or factors) that causes both; for example, low self-esteem might lead you to have breast implants and also attempt suicide. These extraneous factors are sometimes 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 provide a comparison of situations (usually called *treatments* or *conditions*) in which the proposed cause is present or absent.

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

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 nice 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 are not praised or punished, instead I give them 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: 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).

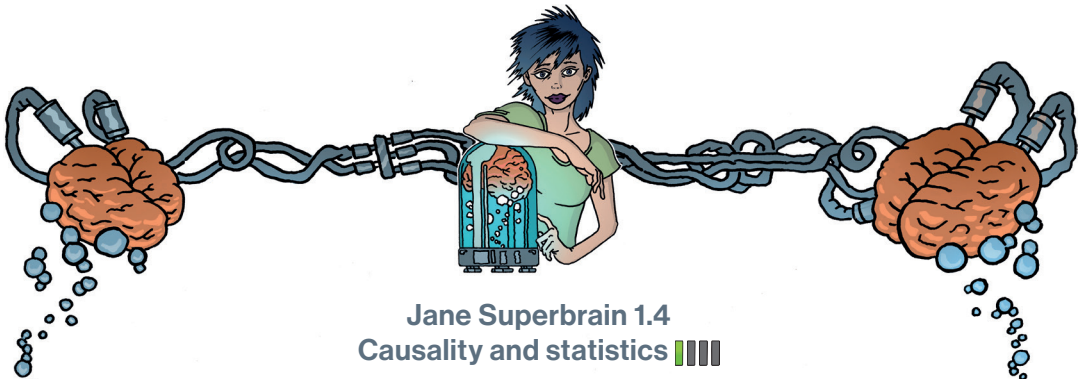
1.7.3 Two methods of data collection

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.

1.7.4 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

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



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, etc.) 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.

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, this *has* to be due to the fact that bananas were used to provide feedback in one training phase but not in 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 (such as 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.

1.7.5 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

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 eradicated.

15 Obviously I mean that they receive 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.

sensitive measure of the experimental manipulation. Generally, scientists use the **randomization** of 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 randomly determine 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 use randomization to ensure that the number of statistics seminars does not introduce a systematic bias by randomly assigning students to have the harsh feedback seminars first or the no feedback seminars first.

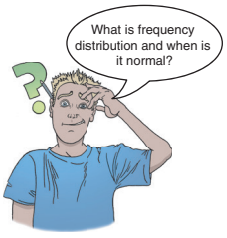
If we turn our attention to independent designs, a similar argument can be applied. We know that participants in different experimental conditions will differ in many respects (their IQ, attention span, etc.). Although we know that these confounding variables contribute to the variation between conditions, we need to make sure that these variables contribute to the unsystematic variation and *not* to the systematic variation. A good example is the effects of alcohol on behaviour. You might give one group of people 5 pints of beer, and keep a second group sober, and then count how many times you can persuade them to do a fish impersonation. The effect that alcohol has varies because people differ in their tolerance: teetotal people can become drunk on a small amount, while alcoholics need to consume vast quantities before the alcohol affects them. If you allocated a bunch of hardened drinkers to the condition that consumed alcohol, and teetotal people to the no alcohol condition, then you might find that alcohol doesn't increase the number of fish impersonations you get. However, this finding could be because (1) alcohol does not make people engage in frivolous activities, or (2) the hardened drinkers were unaffected by the dose of alcohol. You have no way to dissociate these explanations

because the groups varied not just on dose of alcohol but also on their tolerance of alcohol (the systematic variation created by their past experience with alcohol cannot be separated from the effect of the experimental manipulation). The best way to reduce this eventuality is to randomly allocate participants to conditions: by doing so you minimize the risk that groups differ on variables other than the one you want to manipulate.

1.8 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 also fitting statistical models to the data (all other chapters). Given that the rest of the book is dedicated to this process, we'll begin here by looking at a few fairly basic ways to look at and summarize the data you have collected.

1.8.1 Frequency distributions



Once you've collected some data a very useful thing to do is to plot a graph of how many times each score occurs. This is known as a **frequency distribution**, or **histogram**, which is a graph plotting values of observations on the horizontal axis, with a bar showing how many times each value occurred in the data set. Frequency distributions can be very useful for assessing properties of the distribution of scores. We will find out how to create these types of charts in Chapter 5.

Frequency distributions come in many different shapes and sizes. It is quite important, therefore, to have some general descriptions for common types of distributions. In an ideal world our data would be distributed symmetrically around the centre of all scores. As such, if we drew a vertical line through the centre of the distribution then it should look the same on both sides. This is known as a **normal distribution** and is characterized by the bell-shaped curve with which you might already be familiar. 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). Also, as we get further away from the centre, the bars get smaller, implying that as scores start to deviate from the centre their frequency is decreasing. As we move still further away from the centre our 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). An example of a normal distribution is shown in Figure 1.3.

16 I am exactly 180 cm tall. In my home country this makes me smugly above average. However, I often visit the Netherlands, where the average male height is 185 cm (a little over 6A, and a massive 10 cm higher than the UK), and where I feel like a bit of a dwarf.

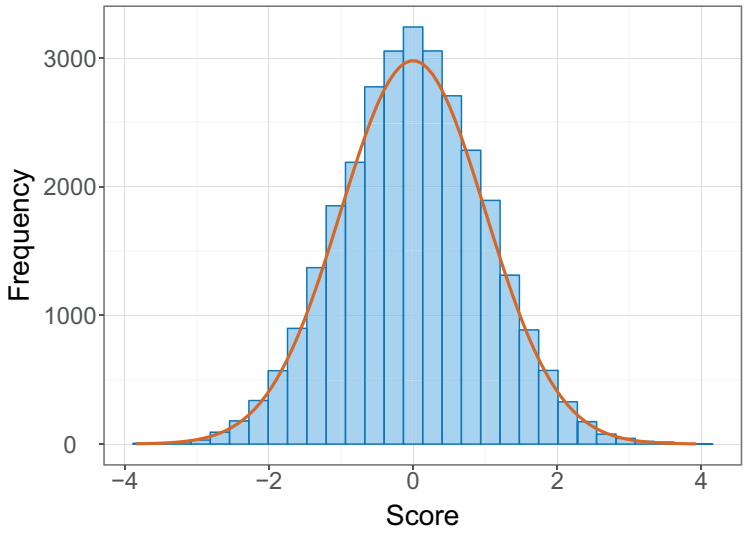


Figure 1.3 A 'normal' distribution (the curve shows the idealized shape)

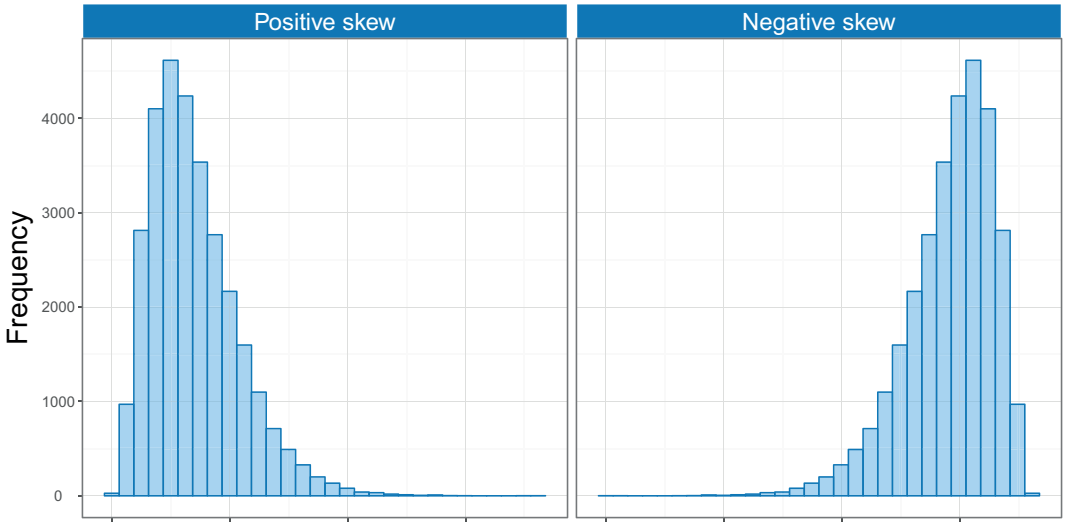


Figure 1.4 A positively (left) and negatively (right) skewed distribution

There are two main ways in which a distribution can deviate from normal: (1) lack of symmetry (called **skew**) and (2) pointyness (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* (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.

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*)

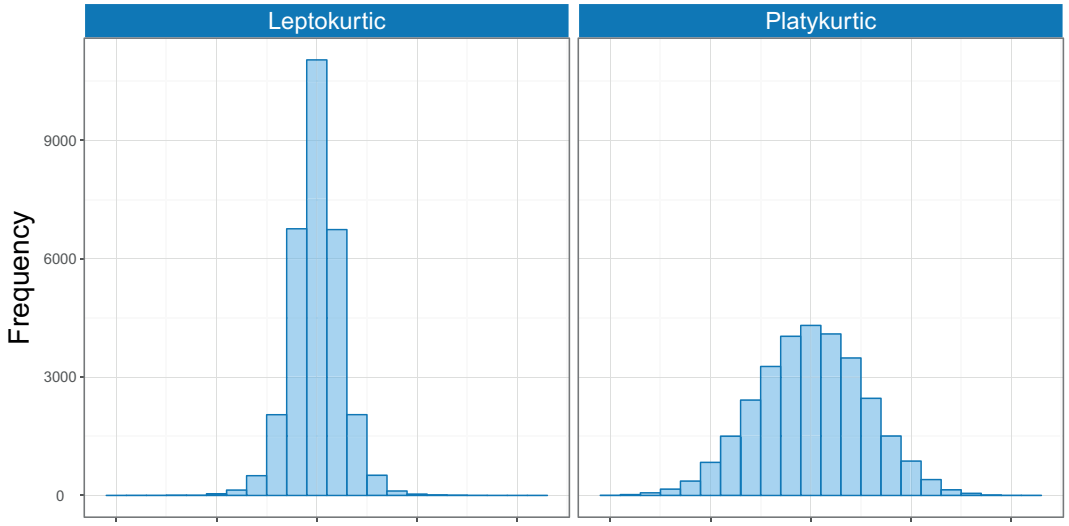


Figure 1.5 Distributions with positive kurtosis (leptokurtic, left) and negative kurtosis (platykurtic, right)

and this tends to express itself in how pointy a distribution is (but there are other factors that can affect how pointy the distribution looks – see Jane Superbrain Box 1.5). A distribution with *positive kurtosis* has many scores in the tails (a so-called heavy-tailed distribution) and is pointy. This is known as a **leptokurtic** distribution. In contrast, a distribution with *negative kurtosis* is relatively thin in the tails (has light tails) and tends to be flatter than normal. This distribution is called **platykurtic**. Ideally, we want our data to be normally distributed (i.e., not too skewed, and not too many or too few scores at the extremes). For everything there is to know about kurtosis, read DeCarlo (1997).

In a normal distribution the values of skew and kurtosis are 0 (i.e., the tails of the distribution are as they should be).[17] If a distribution has values of skew or kurtosis above or below 0 then this indicates a deviation from normal: Figure 1.5 shows distributions with kurtosis values of +2.6 (left panel) and –0.09 (right panel).

1.8.2 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 be 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.

17 Sometimes no kurtosis is expressed as 3 rather than 0, but SPSS uses 0 to denote no excess kurtosis.

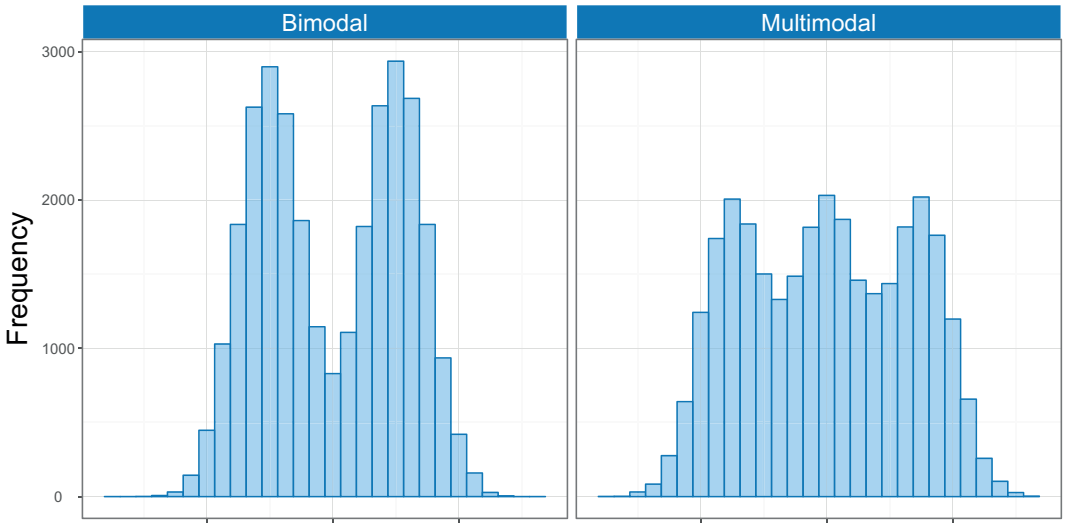
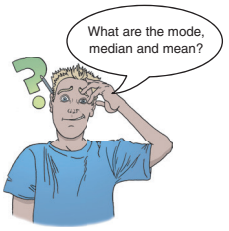


Figure 1.6 Examples of bimodal (left) and multimodal (right) distributions

1.8.3 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 friends that 11 users of the social networking website Facebook had. Figure 1.7 shows the number of friends 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 $(n + 1)/2 = (11 + 1)/2 = 12/2 = 6$. 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).

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

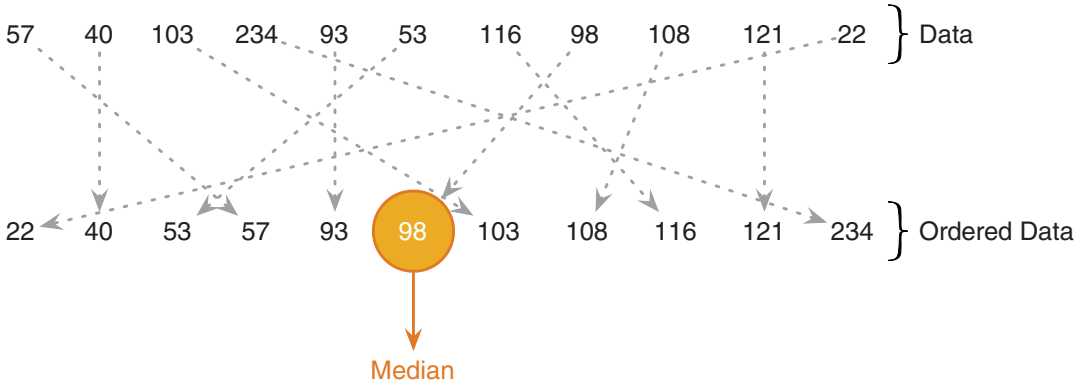


Figure 1.7 The median is simply the middle score when you order the data

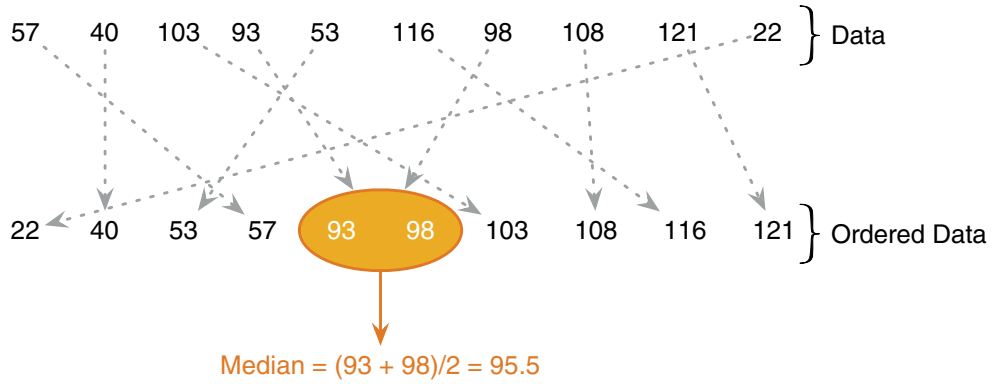


Figure 1.8 When the data contain an even number of scores, the median is the average of the middle two values

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).

1.8.4 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.[18] To calculate the mean we add up all of 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 simply 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 instead), and the bottom bit means, 'divide this total by the number of scores you have got (n)'. Let's calculate the mean for the Facebook data. First, we add up all the scores:

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

18 I wrote this on 15 February, and to prove my point, the BBC website ran a headline today 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 is full of lies and contradictions.

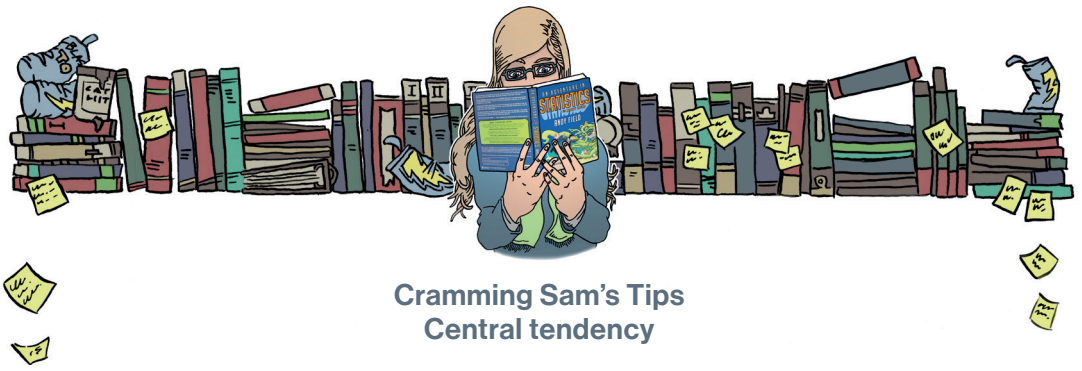
We then divide by the number of scores (in this case 11) as in equation (1.3):

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

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 friends on Facebook increased the mean by about 14 friends; 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 do we use it so often? 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
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.

1.8.5 The dispersion in a distribution

It can also be interesting 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 Facebook 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. One problem with the range is that because it uses only the highest and lowest score, it is affected dramatically by extreme scores.

If you have done the self-test task you'll see that without the extreme score the range drops from 212 to 99 – less than half the size.

One way around this problem is to calculate the range but 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 Facebook data. First, we need to calculate what are 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. 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 Facebook 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 Facebook 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 a lot of data (half of it, in fact).

It's worth noting here 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

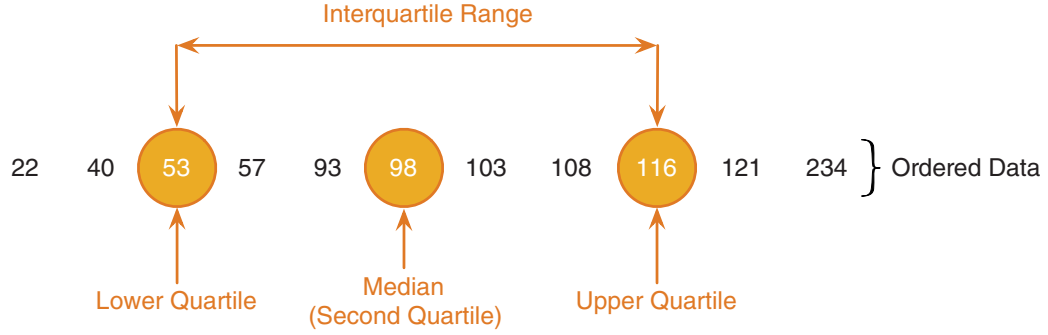


Figure 1.9 Calculating quartiles and the interquartile range

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.

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.

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** (Eq. 1.4):

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

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:

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

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

Let's try this with the Facebook data. Table 1.2 shows the number of friends for each person in the Facebook 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 Queen of England if they don a bowler hat and pretend to be human. Everyone knows that the Queen 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 exactly the same, the resulting value will be bigger than zero, indicating that there is some deviance from the mean. As an equation, we would write: equation (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):

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

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

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

Number of Friends (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
		$\sum_{i=1}^{n} x_i - \bar{x} = 0$	$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 32246$

the data. The sum of squares for the Facebook 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 also 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 simply the sum of squares divided by the number of observations (N). Actually, we normally divide the SS by the number of observations minus 1 as in equation (1.7) (the reason why is explained in the next chapter and Jane Superbrain Box 2.2):

$$\text{variance}\left(s^2\right) = \frac{\text{SS}}{N-1} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{N-1} = \frac{32{,}246}{10} = 3224.6 \tag{1.7}$$

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 friends squared. It makes very little sense to talk about friends squared, so we often take the square root of the variance (which ensures that the measure of average error is in the same units as the original measure). This measure is known as the **standard deviation** and is the square root of the variance (Eq. 1.8).

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

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)

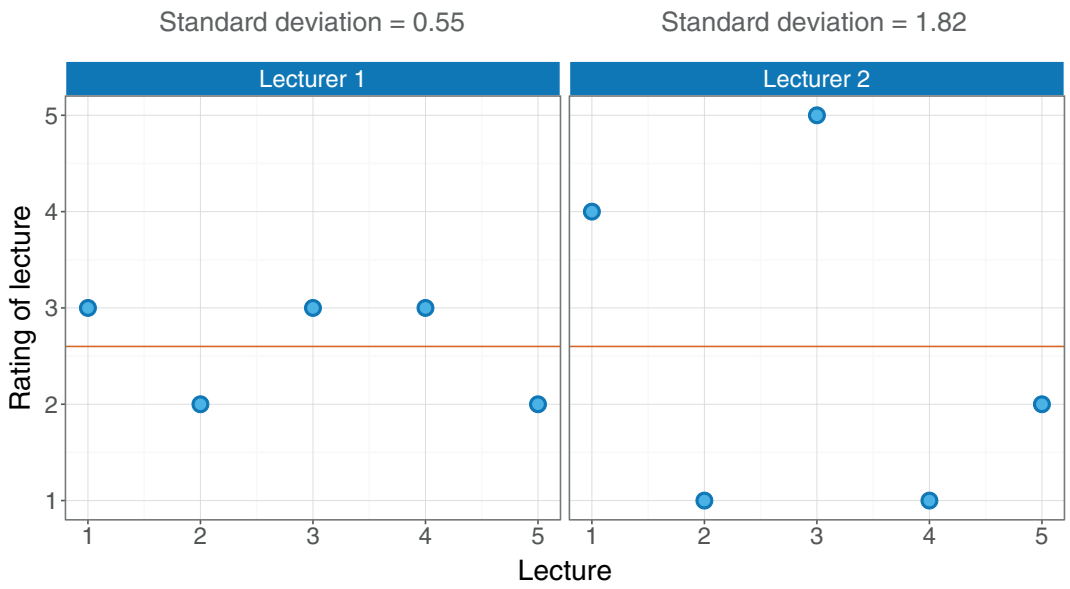


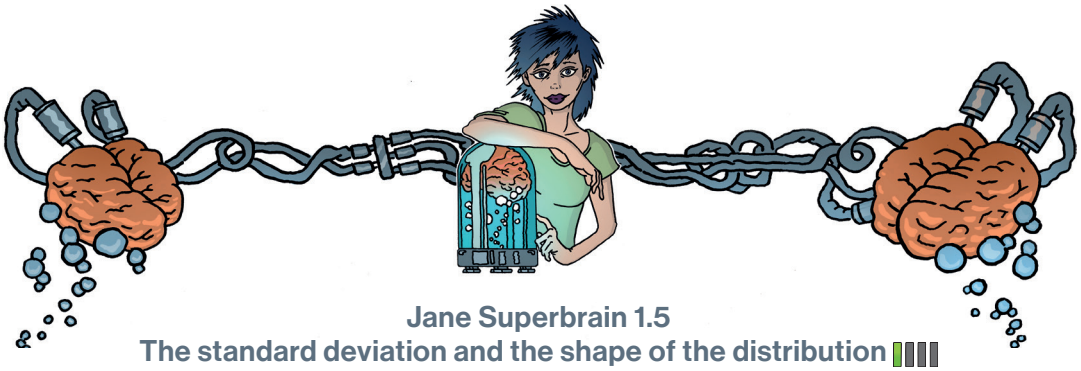
Figure 1.10 Graphs illustrating data that have the same mean but different standard deviations

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 first lecturer had a standard deviation of 0.55 (relatively small compared to the mean). It should be clear from the left-hand graph 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. Put another way, the scores are not spread too widely around the mean. The second lecturer, however, had a standard deviation of 1.82 (relatively high compared to the mean). The ratings for this second lecturer are more spread from the mean than the first: for some lectures she received very high ratings, and for others her ratings were appalling.

1.8.6 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 tend to 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; in

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.



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 represents the data well then most of the scores will cluster close to the mean and the resulting standard deviation is small relative to the mean. When the mean is a worse representation 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.

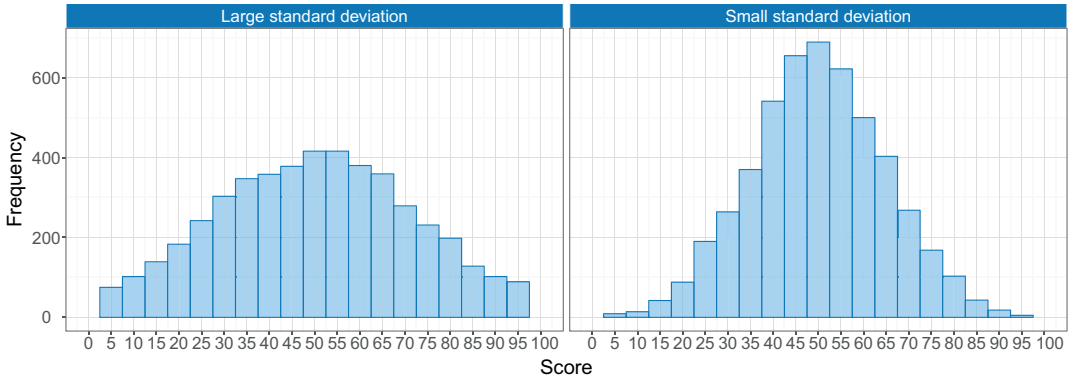
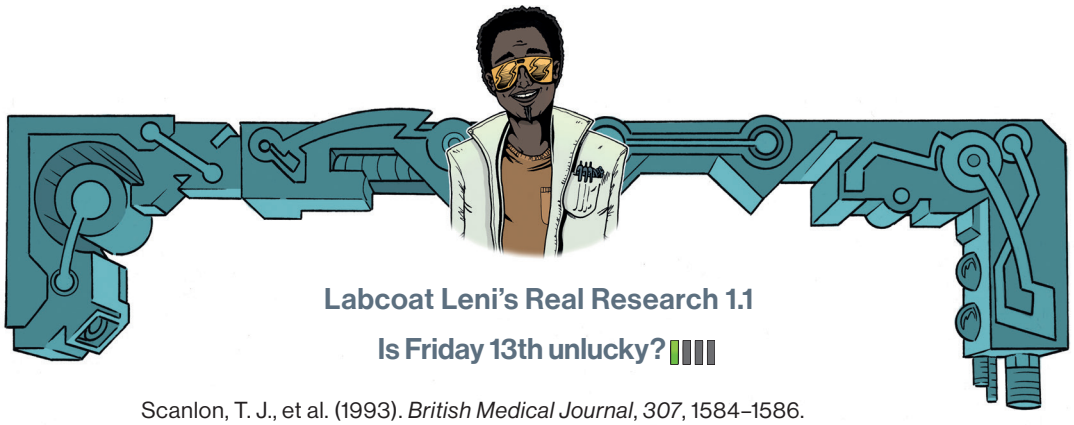


Figure 1.11 Two distributions with the same mean, but large and small standard deviations

this video you 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 Facebook and 2.3 million on YouTube. I mentioned that memes often follow a normal distribution, and Figure 1.12 shows this: the insert shows the 'interest' score from Google Trends



Labcoat Leni's Real Research 1.1

Is Friday 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, Luben, Scanlon, & Singleton, 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

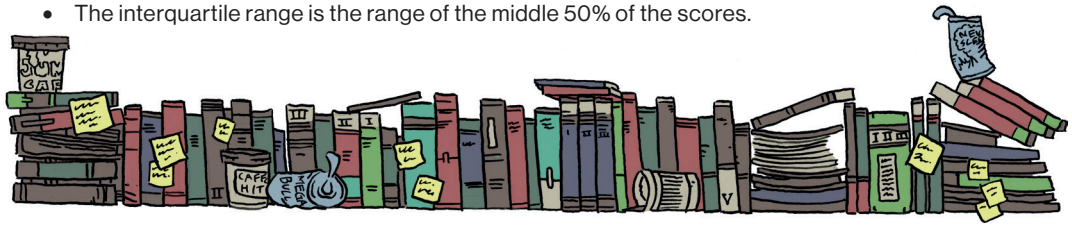
Calculate the mean, median, standard deviation and interquartile range for each type of accident and on each date. Answers are on the companion website.

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 fairly similar, don't

20 You can generate the insert 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.

Cramming Sam's Tips
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.



they? Once it got going (about 2–3 weeks after the first video) it went viral, and popularity increased rapidly, reaching a peak at around 21 August (about 36 days after Chris Kennedy got the ball rolling). After this peak, popularity rapidly declines as people tire of the meme.

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 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 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 the type of people like me, and statistics lectures more generally, who don't check Facebook 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, hipster scum!

21 Very very approximately indeed. I have converted the Google interest data into videos posted on YouTube by using the fact that I know that 2.33 million videos were posted during this period and by 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.

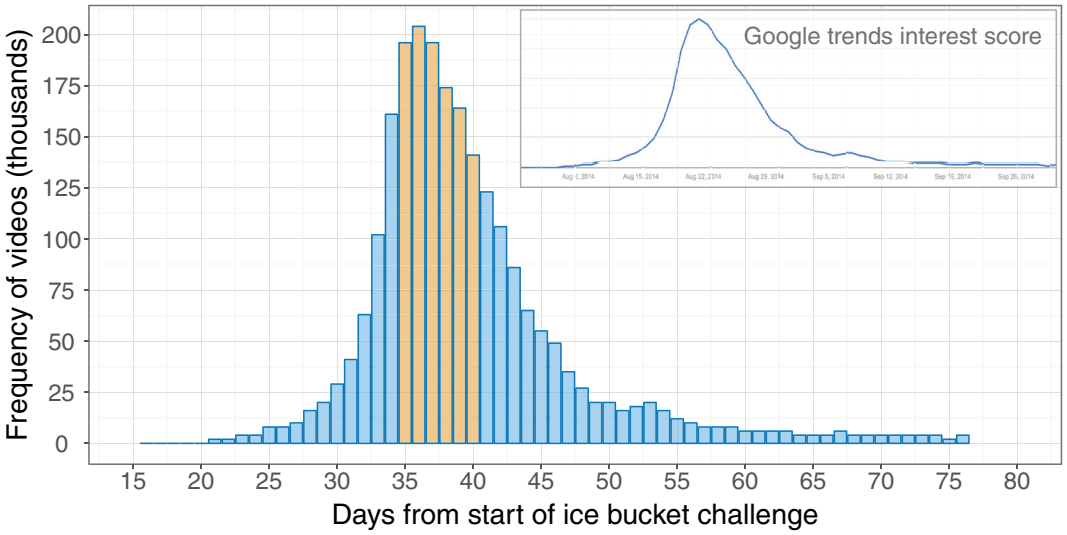


Figure 1.12 Frequency distribution showing the number of ice bucket challenge videos on YouTube by day since the first video (the insert shows the actual Google Trends data on which this example is based)

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 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 size 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

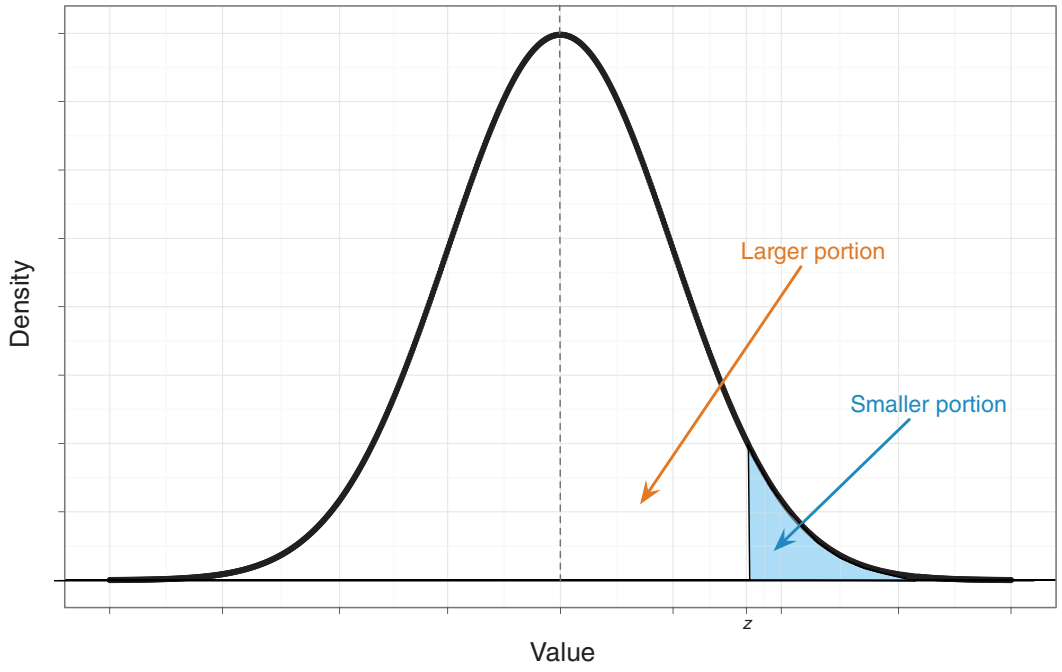


Figure 1.13 The normal probability distribution

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, PDF**) 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 1.8.1) it would look like Figure 1.13, which has the characteristic bell shape that we saw already in Figure 1.3.

A probability distribution is just 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 probability density function. 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 all of

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

these distributions have something in common: they are all defined by an equation that enables us to calculate precisely the probability of obtaining a given score.

As we have seen, distributions can have different means and standard deviations. This isn't a problem for the probability density function – it will still give us the probability of a given value occurring – but it is a problem for us because probability density functions are difficult enough to spell, let alone use to compute probabilities. Therefore, to avoid a brain meltdown we often use a normal distribution with a mean of 0 and a standard deviation of 1 as a standard. This has the advantage that we can pretend that the probability density function 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 a 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 data set can be converted into a data set that has 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:

$$z = \frac{X - \bar{X}}{s} \tag{1.9}$$

The table of probability values that have been calculated for the standard normal distribution is shown 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 2.63 (Eq. 1.10):

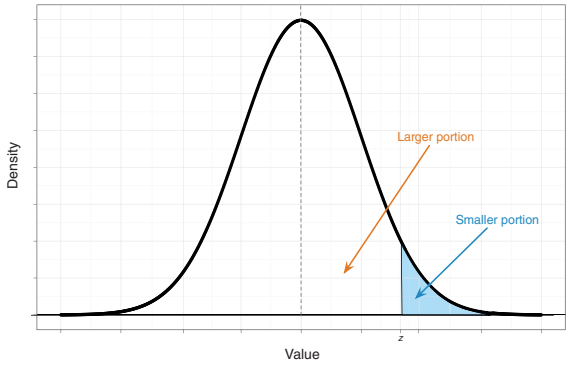
$$z = \frac{60 - 39.68}{7.74} = 2.63 \tag{1.10}$$

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 portions. 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?' To answer this question we need to use the table the opposite way around. We know that the

A.1. Table of the standard normal distribution



z	Larger Portion	Smaller Portion	y	z	Larger Portion	Smaller Portion	y
.00	.50000	.50000	.3989	.12	.54776	.45224	.3961
.01	.50399	.49601	.3989	.13	.55172	.44828	.3956
.02	.50798	.49202	.3989	.14	.55567	.44433	.3951
.03	.51197	.48803	.3988	.15	.55962	.44038	.3945
.04	.51595	.48405	.3986	.16	.56356	.43644	.3939
1.56	.94062	.05938	.1182	1.86	.96856	.03144	.0707
1.57	.94179	.05821	.1163	1.87	.96926	.03074	.0694
1.58	.94295	.05705	.1145	1.88	.96995	.03005	.0681
1.59	.94408	.05592	.1127	1.89	.97062	.02938	.0669
1.60	.94520	.05480	.1109	1.90	.97128	.02872	.0656
1.61	.94630	.05370	.1092	1.91	.97193	.02807	.0644
1.62	.94738	.05262	.1074	1.92	.97257	.02743	.0632
1.63	.94845	.05155	.1057	1.93	.97320	.02680	.0620
1.64	.94950	.05050	.1040	1.94	.97381	.02619	.0608
1.65	.95053	.04947	.1023	1.95	.97441	.02559	.0596
1.66	.95154	.04846	.1006	1.96	.97500	.02500	.0584
1.67	.95254	.04746	.0989	1.97	.97558	.02442	.0573
1.68	.95352	.04648	.0973	1.98	.97615	.02385	.0562
2.27	.98840	.01160	.0303	2.57	.99492	.00508	.0147
2.28	.98870	.01130	.0297	2.58	.99506	.00494	.0143
2.29	.98899	.01101	.0290	2.59	.99520	.00480	.0139
2.30	.98928	.01072	.0283	2.60	.99534	.00466	.0136
2.31	.98956	.01044	.0277	2.61	.99547	.00453	.0132
2.32	.98983	.01017	.0270	2.62	.99560	.00440	.0129
2.33	.99010	.00990	.0264	2.63	.99573	.00427	.0126

Figure 1.14 Using tabulated values of the standard normal distribution

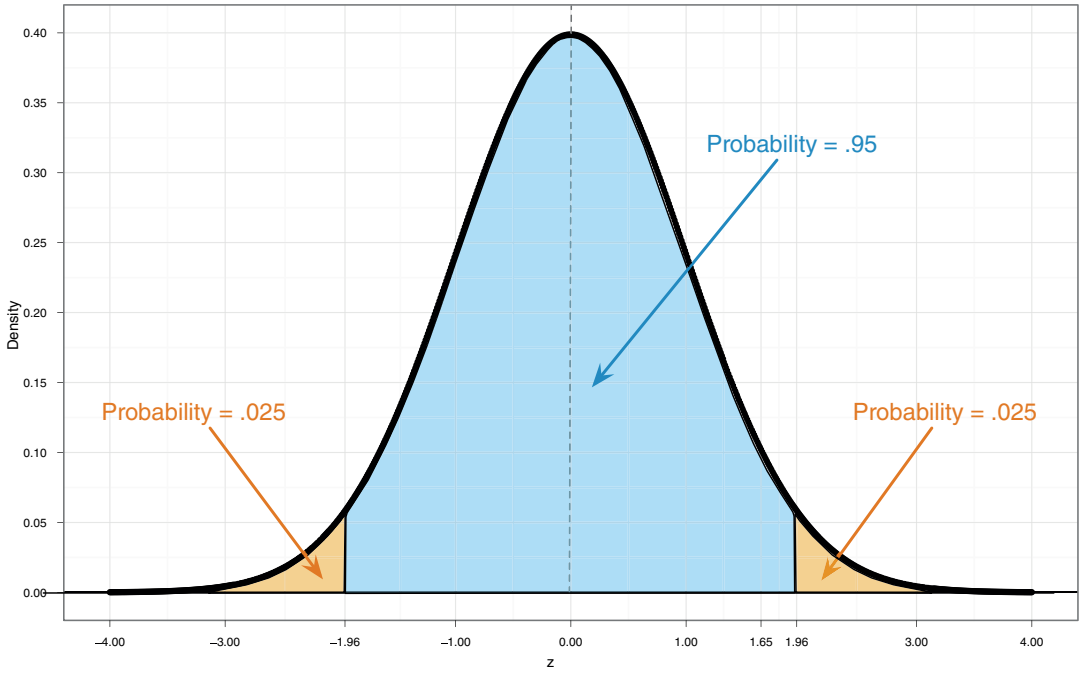
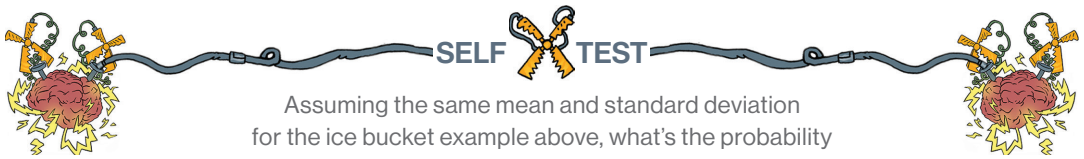


Figure 1.15 The probability density function of a normal distribution

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?

Cramming Sam's Tips
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.

1.8.7 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 5), 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.

1.9 Reporting data

1.9.1 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 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 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 then 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 when they doubt themselves.

1.9.2 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 graphs (see Chapter 5) 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 Pyschological Association, 2010), 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 sensible guiding principles that you can extrapolate to most situations you'll encounter.

1.9.3 Some initial guiding principles

When reporting data, your first decision is whether to use text, a graph or a table. You want to be succinct, so you shouldn't present the same values in multiple ways: if you have a graph 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 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 graph rather than a table to present 4–20 values because a graph will show up the pattern of data most clearly. Similarly, I can imagine some graphs 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 present graphs 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 the mathematical notation ($\bar{X}$) too. However, you should be consistent: if you use M to represent the mean you should do so throughout your article. 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 friends on Facebook. 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) friends on Facebook 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 suggests *Mdn*) and we should report the range or interquartile range as well (the APA does not have an abbreviation for either of these terms, but IQR is commonly used for the interquartile range). Therefore, we could report:

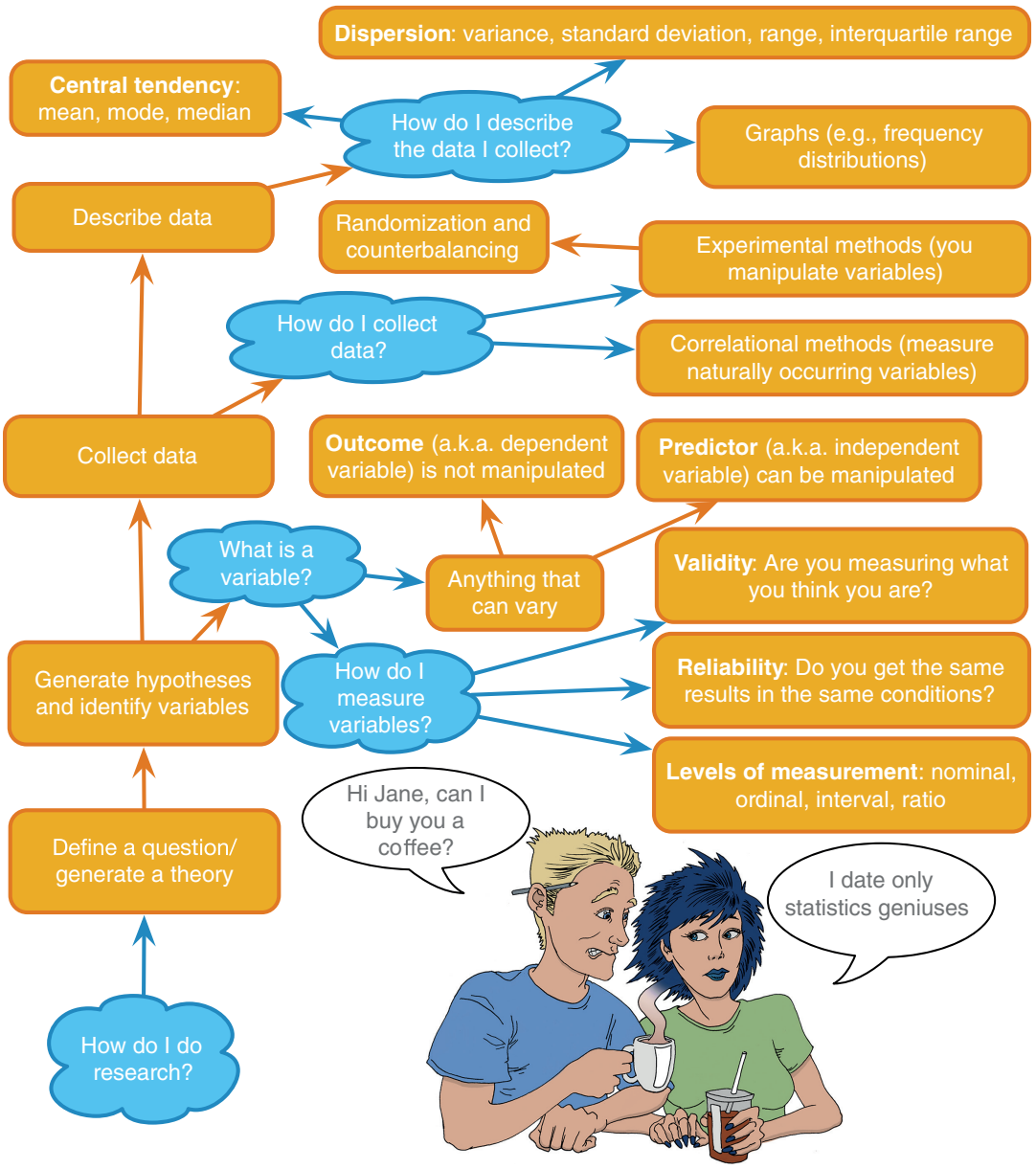


Figure 1.16 What Brian learnt from this chapter

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

1.10 Brian's attempt to woo Jane

Brian had a crush on Jane. He'd seen her around campus a lot, always rushing with a big bag and looking sheepish. People called her a weirdo, but her reputation for genius was well earned. She was mysterious, no one had ever spoken to her or knew why she scuttled around the campus with such purpose. Brian found her quirkiness sexy. He probably needed to reflect on that someday.

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.

Jane looked Brian up and down. He was handsome, but he looked like he might be an idiot ... and Jane didn't trust people, especially guys. To her surprise, Brian tried to woo her with what he'd learnt in his statistics lecture that morning. Maybe she was wrong about his idiocy, 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.

1.11 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.

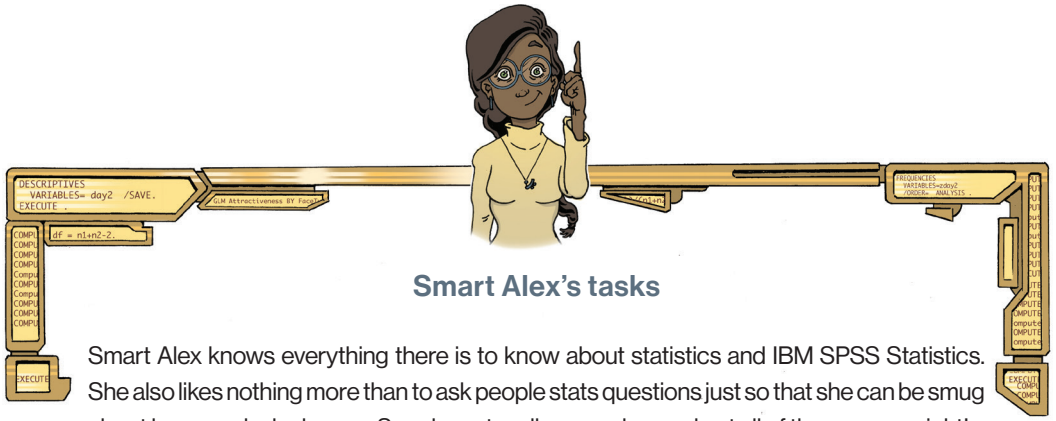
My early childhood wasn't all full of pain, on the contrary it was filled with a lot of fun: the nightly 'from how far away can I jump into bed' competition (which sometimes involved a bit of pain) and being carried by my brother and dad to bed as they hummed Chopin's *Marche Funèbre* before lowering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

1.12 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

WHY IS MY EVIL LECTURER FORCING ME TO LEARN STATISTICS?

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
Sum of squared errors
Systematic variation
Tertium quid
Test–retest reliability
Theory
Unsystematic variance
Upper quartile
Validity
Variables
Variance
Within-subject design
z-scores



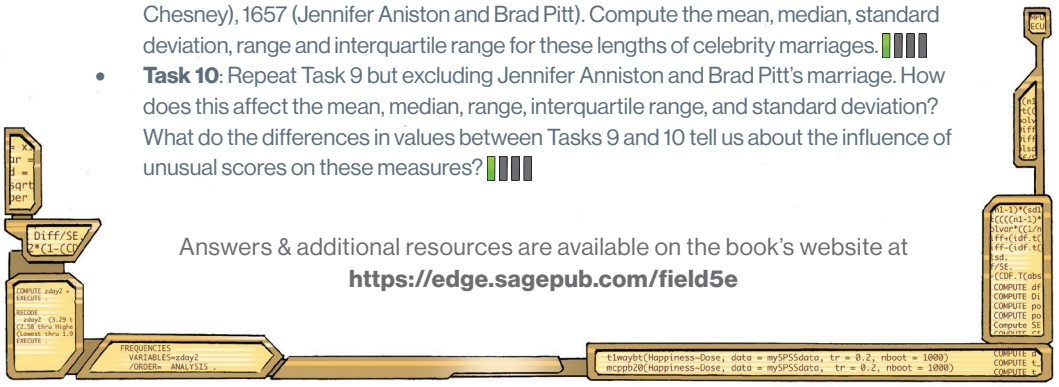
Smart Alex's tasks

Smart Alex knows everything there is to know about statistics and IBM SPSS Statistics. She also likes nothing more than to ask people stats questions just so that she can be smug about how much she knows. So, why not really annoy her and get all of the answers right!

- **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?



Answers & additional resources are available on the book's website at
https://edge.sagepub.com/field5e