Why is my evil lecturer forcing me to learn statistics? 1

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

1.1. What will this chapter tell me? ①

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 did go through a phase of driving my parents mad by asking 'Why?' every five seconds. 'Dad, why is the sky blue?', 'Dad, why doesn't mummy have a willy?', etc. Children are naturally curious about the world. I remember at the age of 3 being at a party of my friend Obe (this was 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. As I said, children are natural scientists and my little scientific 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.[1] My point is this: my curiosity to explain the world never went away, and that's why

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

I'm a scientist, and that's also why your evil lecturer is forcing you to learn statistics. It's because you have a curious mind too and you want to answer new and exciting questions. To answer these questions we need statistics. Statistics is a bit like sticking your finger into a revolving fan blade: sometimes it's very painful, but it does give you the power to answer interesting questions. This chapter is going to attempt to explain why 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.

1.2. What the hell am I doing here? I don't belong here ①

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 need 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 can scan a 700-page textbook)? 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 this much about statistics but now find yourself inexplicably wading, neck high, through the Victorian sewer that is data analysis. The reason that you're in the mess that you find yourself in is because you have a curious mind. You might have asked yourself questions like why do people behave the way they do (psychology) or why do behaviours differ across cultures (anthropology), how do businesses maximize their profit (business), how did the dinosaurs die? (palaeontology), does eating tomatoes protect you against cancer (medicine, biology), is it possible to build a quantum computer (physics, chemistry), is the planet hotter than it used to be and in what regions (geography, environmental studies)? Whatever it is you're studying or researching, the reason you're studying it is probably because you're interested in answering questions. Scientists are curious people, and you probably are too. However, you might not have bargained on the fact that to answer interesting questions, you need two things: data and an explanation of those data.

The answer to 'what the hell are you doing here?' is, therefore, simple: to answer interesting questions you need data. Therefore, one of the reasons why your evil statistics lecturer is forcing you to learn about numbers is because 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, media broadcasts and so on). 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.[2]

[2] This is a joke. I thought long and hard about whether to include it because, like many of my jokes, there are people who won't find it remotely funny. Its inclusion is also making me fear being hunted down and forced to eat my own entrails by a hoard of rabid qualitative researchers. However, it made me laugh, a lot, and despite being vegetarian I'm sure my entrails will taste lovely.

1.2.1. The research process ①

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[3]) or could be based on some data (you've got several cat owners to keep diaries of their cat's TV habits and have noticed that lots of them watch birds on TV). From your initial observation you generate explanations, or theories, of those observations, from which you can make predictions (hypotheses). Here's where the data come into the process because to test your predictions you need data. First you collect some relevant data (and to do that you need to identify things that can be measured) and then you analyse those data. The analysis of the data may support your theory or give you cause to modify 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.

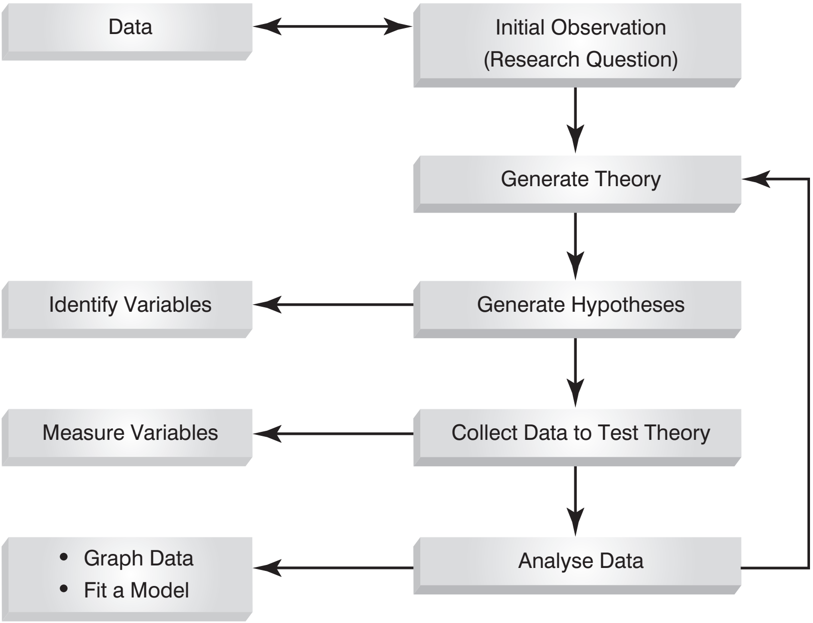


FIGURE 1.2 The research process

1.3. Initial observation: finding something that needs explaining ①

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. Every year I swear that I won't get hooked on *Big Brother*, and yet every year I find myself glued to the TV screen waiting for

[3] My cat does actually climb up and stare at the TV when it's showing birds flying about.

the next contestant's meltdown (I am a psychologist, so really this is just research – honestly). One question I am constantly perplexed by is why every year there are so many contestants with really unpleasant personalities (my money is on narcissistic personality disorder[4]) on the show. A lot of scientific endeavour starts this way: not by watching *Big Brother*, but by observing something in the world and wondering why it happens.

Having made a casual observation about the world (*Big Brother* contestants on the whole have profound personality defects), I need to collect some data to see whether this observation is true (and not just a biased observation). To do this, I need to define one or more **variables** that I would like 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 *Big Brother* contestants have extreme personalities.

1.4. Generating theories and testing them ①

The next logical thing to do is to explain these data (Figure 1.2). One explanation could be that people with narcissistic personality disorder are more likely to audition for *Big Brother* than those without. This is a **theory**. Another possibility is that the producers of *Big Brother* are more likely to select people who have narcissistic personality disorder to be contestants than those with less extreme personalities. This is another theory. We verified our original observation by collecting data, and we can collect more data to test our theories. We can make two predictions from these two theories. The first is that the number of people turning up for an audition that have narcissistic personality disorder will be higher than the general level in the population (which is about 1%). A prediction from a theory, like this one, is known as a **hypothesis** (see Jane Superbrain Box 1.1). We could test this hypothesis by getting a team of clinical psychologists to interview each person at the *Big Brother* audition and diagnose them as having narcissistic personality disorder or not. The prediction from our second theory is that if the *Big Brother* selection panel are more likely to choose people with narcissistic personality disorder then the rate of this disorder in the final contestants will be even higher than the rate in the group of people going for auditions. This is another hypothesis. Imagine we collected these data; they are in Table 1.1.

In total, 7662 people turned up for the audition. Our first hypothesis is 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,

TABLE 1.1 A table of the number of people at the *Big Brother* 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

[4] 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.

854 were diagnosed with the disorder; this is about 11% (854/7662 × 100) which is much higher than the 1% we'd expect. Therefore, the first hypothesis is supported by the data. The second hypothesis was that the *Big Brother* selection panel have a bias to chose 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 data again support our hypothesis. Therefore, my initial observation that contestants have personality disorders was verified by data, then my theory was tested using specific hypotheses that were also verified using data. Data are *very* important!

JANE SUPERBRAIN 1.1

When is a hypothesis not a hypothesis? ①

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 are ones that 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',[5] '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 record 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 should, therefore, produce hypotheses that are scientific statements.

I would now be smugly sitting in my office with a contented grin on my face about how my theories and observations 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. After another few hours (well, days probably) locked up at home watching *Big Brother* I would emerge triumphant with another profound observation, which is that these personality-disordered contestants, despite their obvious character flaws, enter the house convinced that the public will love them and that they will win.[6] My hypothesis would, therefore, be that if I asked the contestants if they thought that they would win, the people with a personality disorder would say yes.

[5] It was pretty awesome actually.

[6] One of the things I like about *Big Brother* in the UK is that year upon year the winner tends to be a nice person, which does give me faith that humanity favours the nice.

Let's imagine I tested my hypothesis by measuring their expectations of success in the show, by just asking them, 'Do you think you will win *Big Brother*?'. Let's say that 7 of 9 contestants with personality disorders said that they thought that they will win, which confirms my observation. Next, I would come up with another theory: these contestants think that they will win because they don't realize that they have a personality disorder. My hypothesis would be that if I asked these people about whether their personalities were different from other people they would say 'no'. As before, I would collect some more data and perhaps ask those who thought that they would win whether they thought that their personalities were different from the norm. All 7 contestants said that they thought their personalities were different from the norm. These data seem to contradict my theory. 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 *Big Brother* have extreme personalities and think that they will win. Imagine these 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 they 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 my theory sucks: it cannot explain all of the data, predictions from the theory are not supported by subsequent data, and it cannot explain other research findings. At this point I would start to feel intellectually inadequate and people would find me curled up on my desk in floods of tears wailing and moaning about my failing career (no change there then).

At this point, a rival scientist, Fester Ingpant-Stain, appears on the scene with a rival theory to mine. In his new theory, he suggests 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 (put another way, they believe that their personality makes them likeable, not dislikeable). One hypothesis 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. To test this hypothesis, Fester Ingpant-Stain collected yet more data. When each contestant came to the diary room they had to fill out a questionnaire evaluating all of the other contestants' personalities, and also answer each question as if they were each of the contestants responding about them. (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' view of them; in comparison 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 the rival theory that the contestants with personality disorders know 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: '*Big Brother* contestants who have personality disorders overestimate how much other people like their personality characteristics'.

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

1.5. Data collection 1: what to measure ①

We have seen already that data collection is vital for testing theories. When we collect data we need to decide on two things: (1) what to measure, (2) how to measure it. This section looks at the first of these issues.

1.5.1. Variables ①

1.5.1.1. Independent and dependent variables ①

To test hypotheses we need to measure variables. Variables are just 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'[7] then proposed cause is 'Coca-Cola' and the proposed effect is dead sperm. Both the cause and the outcome are variables: for the cause we could vary the type of drink, and for the outcome, these drinks will kill different amounts of sperm. The key to testing such 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 actually manipulated by the experimenter (as we will see in section 1.6.2). In cross-sectional research we don't manipulate any variables and we cannot make causal statements about the relationships between variables, so it doesn't make sense to talk of dependent and independent variables because all variables are dependent variables in a sense. One possibility is to abandon the terms dependent and independent variable and use the terms **predictor variable** and **outcome variable**. In experimental work the cause, or independent variable, is a predictor, and the effect, or dependent variable, is simply an outcome. This terminology also suits cross-sectional work where, statistically at least, we can use one or more variables to make predictions about the other(s) without needing to imply causality.

CRAMMING SAM'S TIPS **Some Important Terms**

When doing research there are some important generic terms for variables that you will encounter:

- **Independent variable:** A variable thought to be the cause of some effect. This term is usually used in experimental research to denote 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 is basically another term for 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. This term could be synonymous with 'dependent variable' for the sake of an easy life.

[7] 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, and different types of Coke are more or less effective – Diet Coke is best apparently (Umpierre, Hill, & Anderson, 1985). Nevertheless, a Coke douche is ineffective at preventing pregnancy.

1.5.1.2. Levels of measurement ①

As we have seen in the examples so far, 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 (not even one in a nice PVC suit) exist. 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 rugby or football (soccer) team. In rugby, the numbers of shirts denote specific field positions, so the number 10 is always worn by the fly-half (e.g. England's Jonny Wilkinson),[8] and the number 1 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 1. A number 10 player is not necessarily better than a number 1 (most managers would not want their fly-half stuck in the front of the scrum!). It is equally as 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 England coach found that Jonny Wilkinson (his number 10) was injured he would not get his number 4 to give number 6 a piggy-back and then take the kick. The only way that nominal data can be used is to consider frequencies. For example, we could look at how frequently number 10s score tries compared to number 4s.

JANE SUPERBRAIN 1.2

Self-report data ①

A lot of self-report data are ordinal. Imagine if two judges at our beauty pageant were asked to rate Billie's beauty on a 10-point scale. We might be confident that a judge who gives a rating of 10 found Billie more beautiful than one who gave a rating of 2, but can we be certain that the first judge found her five times more beautiful than the second? What about if both judges gave a rating of 8, could we be sure they found her equally beautiful? Probably not: their ratings will depend on their subjective feelings about what constitutes beauty. For these reasons, in any situation in which we ask people to rate something subjective (e.g. rate 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.

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

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. Imagine we went to a beauty pageant in which the three winners 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 much higher ratings from the judges than Freema and Elizabeth, or it might have been a very close contest that she won by only a point. 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. 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 (i.e. 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 also 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 422 and 472 ms, but also it is 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 twice as short as a reaction time of 400 ms.

Continuous variables can be, well, continuous (obviously) but also discrete. This is quite a tricky distinction (Jane Superbrain Box 1.3). A truly continuous variable can be measured to any level of precision, whereas a **discrete variable** can take on only certain values (usually whole numbers) on the scale. What does this actually mean? Well, our example above 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).

JANE SUPERBRAIN 1.3

Continuous and discrete variables ①

The distinction between discrete and continuous variables can be very 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 weird 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).

1.5.2. Measurement error ①

We have seen that to test hypotheses we need to measure variables. Obviously, it’s also important that we measure these variables 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 directly measured (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).

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 and another scientist, Dr Jack Q. Late, replicated the study but found that sperm motility 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 to 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 100 ml of Coke and 5 ml 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: 100 ml of a liquid is known to be twice as much as only 50 ml.

The second reason for the difference in results between the studies could have been to do with how sperm motility was measured. Perhaps in our original study we measured motility using absorption spectrophotometry, whereas Dr Late used laser light-scattering techniques.[9] Perhaps his measure is more sensitive than ours.

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

1.5.3. Validity and reliability ①

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 actually 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; 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 the instrument is measuring what it claims to measure (does your lecturer's helpfulness rating scale actually measure lecturers' helpfulness?). In an ideal world, you could assess this by relating scores on your measure to real-world observations.

[9] In the course of writing this chapter I have discovered more than I think is healthy about the measurement of sperm.

For example, we could take an objective measure of how helpful lecturers were and compare these observations to student's ratings on ratemyprofessor.com. This is often impractical and, of course, with attitudes you might not be interested in the reality so much as the person's perception of reality (you might not care whether they are 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 17).

SELF-TEST What is the difference between reliability and validity?

1.6. Data collection 2: how to measure ①

1.6.1. Correlational research methods ①

So far we've learnt that scientists want to answer questions, and that to do this they have to generate data (be they numbers or words), and to generate good data they need to use accurate measures. We move on now to look briefly at how the 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.

The main distinction between what we could call **correlational** or **cross-sectional research** (where we observe what naturally goes on in the world without directly interfering with it) and **experimental research** (where we manipulate one variable to see its effect on another) is that experimentation involves the direct manipulation of variables. In correlational research we do things like observe natural events or we take a snapshot of many variables at a single point in time. As some examples, 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 are 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 predict 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.

1.6.2. Experimental research methods ①

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, such as '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 anxious on dates and others won't (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; is it the case that the people who have a low opinion of themselves are the same people that get anxious on dates?

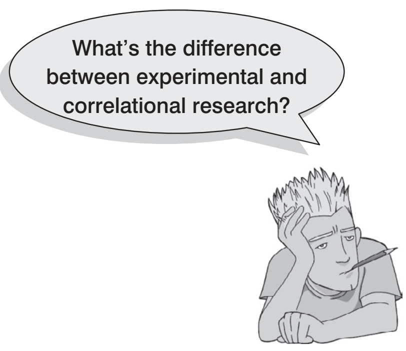
David Hume (see Hume, 1739–40; 1748 for more detail),[10] an influential philosopher, said that to infer cause and effect: (1) cause and effect must occur close together in time (contiguity); (2) the cause must occur before an effect does; and (3) the effect should never occur without the presence of the cause. These conditions imply that causality can be inferred through corroborating evidence: cause 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 whenever someone had low self-esteem they would feel anxious when on a date, that the low self-esteem emerged before the dating anxiety did, and that the person should never have dating anxiety if they haven't been suffering from low self-esteem.

In the previous section on correlational research, we saw that 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 the low self-esteem or the dating anxiety came first.

Let's imagine that we find that there are people who have low self-esteem but do not get dating anxiety. This finding doesn't violate Hume's rules: he doesn't say anything about the cause happening without the effect. 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 put pressure on you in dating situations). This illustrates a second problem

[10] Both of these can be read online at http://www.utilitarian.net/hume/ or by doing a Google search for David Hume.

with correlational evidence: 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.


What's the difference between experimental and correlational research?

The shortcomings of Hume's criteria led John Stuart Mill (1865) to add a further criterion: that all other explanations of the cause–effect relationship be ruled out. Put simply, Mill proposed that, to rule out confounding variables, an effect should be present when the cause is present and that when the cause is absent the effect should be absent also. Mill's ideas can be summed up by saying that the only way to infer causality is through comparison of 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.

As a simple case, we might want to see what the effect of positive encouragement has on learning about statistics. I might, therefore, randomly split some students into three different groups in which I change my style of teaching in the seminars on the course:

- **Group 1 (positive reinforcement)**: 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 (negative reinforcement)**: 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.
- **Group 3 (no reinforcement)**: This group receives normal university style seminars (some might argue that this is the same as group 2!). Students are not praised or punished and instead I give them no feedback at all.

The thing that I have manipulated is the teaching method (positive reinforcement, negative reinforcement or no reinforcement). As we have seen earlier in this chapter, 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. reinforcement has been split into three types: positive, negative and none). Once I have carried out this manipulation I must have some kind of outcome that I am interested in measuring. In this case it is statistical ability, and I could measure this variable using a statistics exam after the last seminar. We have also already discovered that this outcome variable is known as 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 reinforcement' group because this is a group where our proposed cause (reinforcement) is absent, and we can compare the outcome in this group against the two situations where the proposed cause is present. If the statistics scores are different in each of the reinforcement groups (cause is present) compared to the group for which no reinforcement was given (cause is absent) then this difference can be attributed to the style of reinforcement. In other words, the type of reinforcement caused a difference in statistics scores (Jane Superbrain Box 1.4).

JANE SUPERBRAIN 1.4

Causality and statistics ①

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 because 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.

1.6.2.1. Two methods of data collection ①

When we collect data in an experiment, we can choose between two methods of data collection. The first is to manipulate the independent variable using different participants. This method is the one described above, in which different groups of people take part in each experimental condition (a **between-groups**, **between-subjects**, or **independent design**). The second method is to manipulate the independent variable using the same participants. Simplistically, this method means that we give a group of students positive reinforcement for a few weeks and test their statistical abilities and then begin to give this same group negative reinforcement for a few weeks before testing them again, and then finally giving them no reinforcement and testing them for a third time (a **within-subject** or **repeated-measures design**). As you will discover, the way in which the data are collected determines the type of test that is used to analyse the data.

1.6.2.2. Two types of variation ①

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 which change various parameters of the economy; once these parameters have been changed a figure appears on the screen indicating the economic growth resulting from those parameters. Now, chimps can't read (I don't think) so this feedback is meaningless. A second training phase is the same except that if the economic growth is good, they get a banana (if growth is bad they do not) – this feedback is valuable to the average chimp. This is a repeated-measures design with two conditions: the same chimps participate in condition 1 *and* in condition 2.

Let's take a step back and think what would happen if we did *not* introduce an experimental manipulation (i.e. there were no bananas in the second training phase so condition 1 and condition 2 were identical). If there is no experimental manipulation then we expect a chimp's behaviour to be similar in both conditions. We expect this because external factors such as age, gender, IQ, motivation and arousal will be the same for both conditions (a chimp's gender etc. will not change from when they are tested in condition 1 to when they are tested in condition 2). If the performance measure is reliable (i.e. our test of how well they run the economy), 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 to what we do to them 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). Therefore, any difference between the means of the two conditions is probably due to the experimental manipulation. So, if the chimps perform better in one training phase than the other then this *has* to be due to the fact that bananas were used to provide feedback in one training phase but not the other. Differences in performance created by a specific experimental manipulation are known as **systematic variation**.

Now let's think about what happens when we use different participants – an independent design. In this design we still have two conditions, but this time different participants participate in each condition. Going back to our example, one group of chimps receives training without feedback, whereas a second group of different chimps does receive feedback on their performance via bananas.[11] 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, IQ and other factors. In short, the type of factors that were held constant in the repeated-measures design are free to vary in the independent measures 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-measures design there are always two sources of variation:

- **Systematic variation**: This variation is due to the experimenter doing something to all of the participants 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.).

The role of statistics is to discover how much variation there is in performance, and then to work out how much of this is systematic and 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 a person 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 people 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. Therefore, the effect of our experimental manipulation is likely to be more apparent in a repeated-measures design than in a between-groups design,

[11] When I say 'via' I don't mean that the bananas developed little banana mouths that opened up and said 'well done old chap, the economy grew that time' in chimp language. I mean that when they got something right they received a banana as a reward for their correct response.

because in the former unsystematic variation can be caused only by differences in the way in which someone behaves at different times. In independent designs we have differences in innate ability contributing to the unsystematic variation. Therefore, this error variation will almost always be much larger than if the same participants had been used. When we look at the effect of our experimental manipulation, it is always against a background of 'noise' caused 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 have more power to detect effects than independent designs.

1.6.3. Randomization ①

In both repeated measures and independent measures designs it is important to try to keep the unsystematic variation to a minimum. By keeping the unsystematic variation as small as possible we get a more sensitive measure of the experimental manipulation. Generally, scientists use the **randomization** of participants 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. When the same people 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 reinforcement and negative reinforcement. If the same participants were used in all conditions, then we might find that statistical ability was higher after the negative reinforcement condition. However, if every student experienced the negative reinforcement after the no reinforcement then they would enter the negative reinforcement condition already having a better knowledge of statistics than when they began the no reinforcement condition. So, the apparent improvement after negative reinforcement would not be due to the experimental manipulation (i.e. it's not because negative reinforcement works), but because participants had attended more statistics seminars by the end of the negative reinforcement condition compared to the no reinforcement 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 negative reinforcement seminars first or the no reinforcement seminars first.

If we turn our attention to independent designs, a similar argument can be applied. We know that different participants participate in different experimental conditions and that these participants 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* the systematic variation. The way to ensure that confounding variables are unlikely to contribute systematically to the variation between experimental conditions is to randomly allocate participants to a particular experimental condition. This should ensure that these confounding variables are evenly distributed across conditions.

A good example is the effects of alcohol on personality. You might give one group of people 5 pints of beer, and keep a second group sober, and then count how many fights each person gets into. The effect that alcohol has on people can be very variable because of different tolerance levels: teetotal people can become very drunk on a small amount, while alcoholics need to consume vast quantities before the alcohol affects them. Now, if you allocated a bunch of teetotal participants to the condition that consumed alcohol, then you might find no difference between them and the sober group (because the teetotal participants are all unconscious after the first glass and so can't become involved in any fights). As such, the person's prior experiences with alcohol will create systematic variation that cannot be dissociated from the effect of the experimental manipulation. The best way to reduce this eventuality is to randomly allocate participants to conditions.

SELF-TEST Why is randomization important?

1.7. Analysing data ①

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 to see what the general trends in the data are, and also fitting statistical models to the data.

1.7.1. Frequency distributions ①

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 4.

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 basically implies that the majority of scores lie around the centre of the distribution (so the largest bars on the histogram are all around the central value).

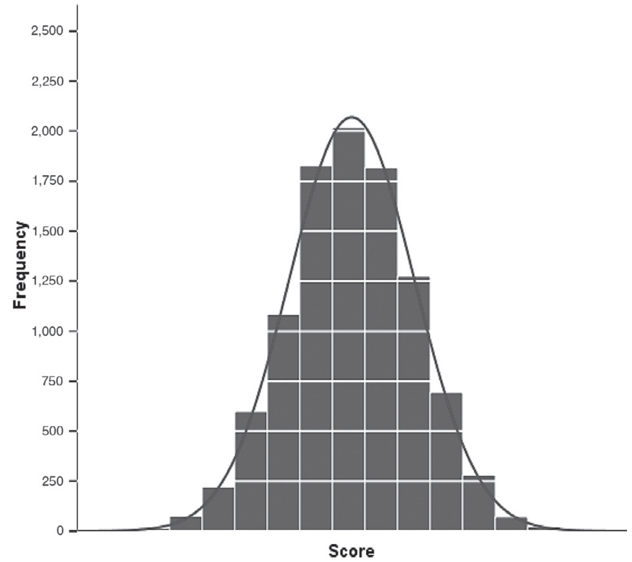
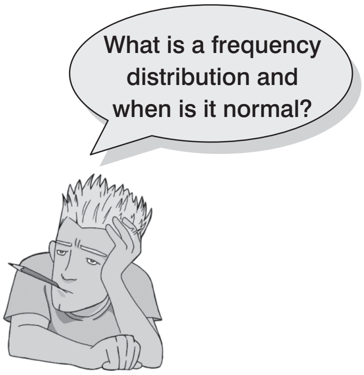


FIGURE 1.3 A 'normal' distribution (the curve shows the idealized shape)



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 about 175 cm tall,[12] 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.

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*) and how pointy a distribution is (but there are other factors that can affect how pointy the distribution looks – see Jane Superbrain Box 2.3). 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). 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 +1 (left panel) and –4 (right panel).

[12] I am exactly 180 cm tall. In my home country this makes me smugly above average. However, I'm writing this in The Netherlands where the average male height is 185 cm (a massive 10 cm higher than the UK), and where I feel like a bit of a dwarf.

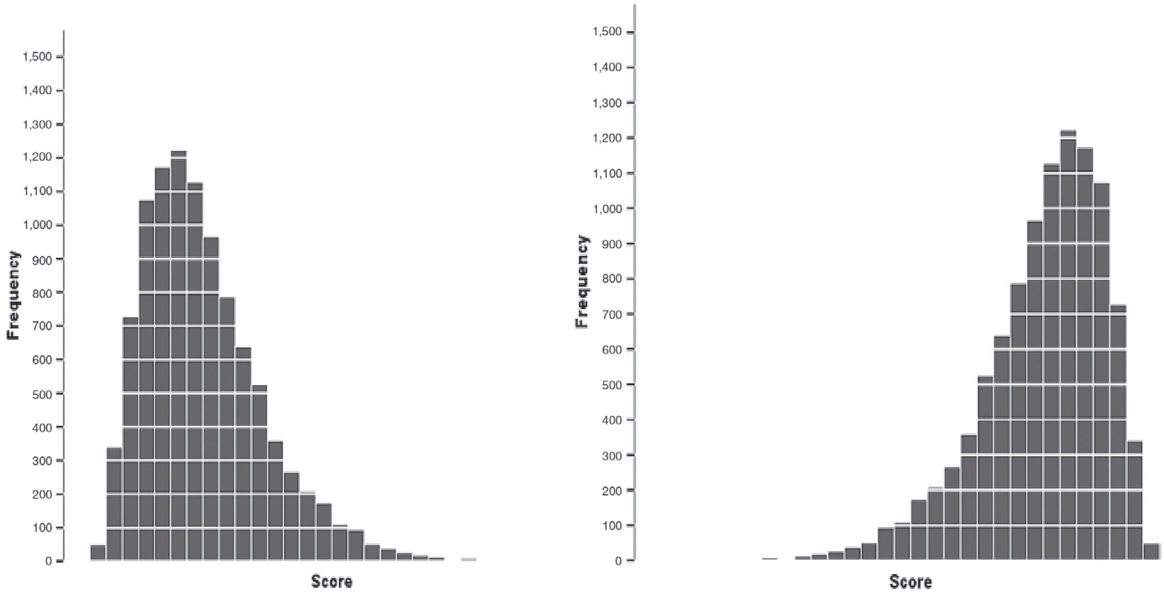


FIGURE 1.4 A positively (left figure) and negatively (right figure) skewed distribution

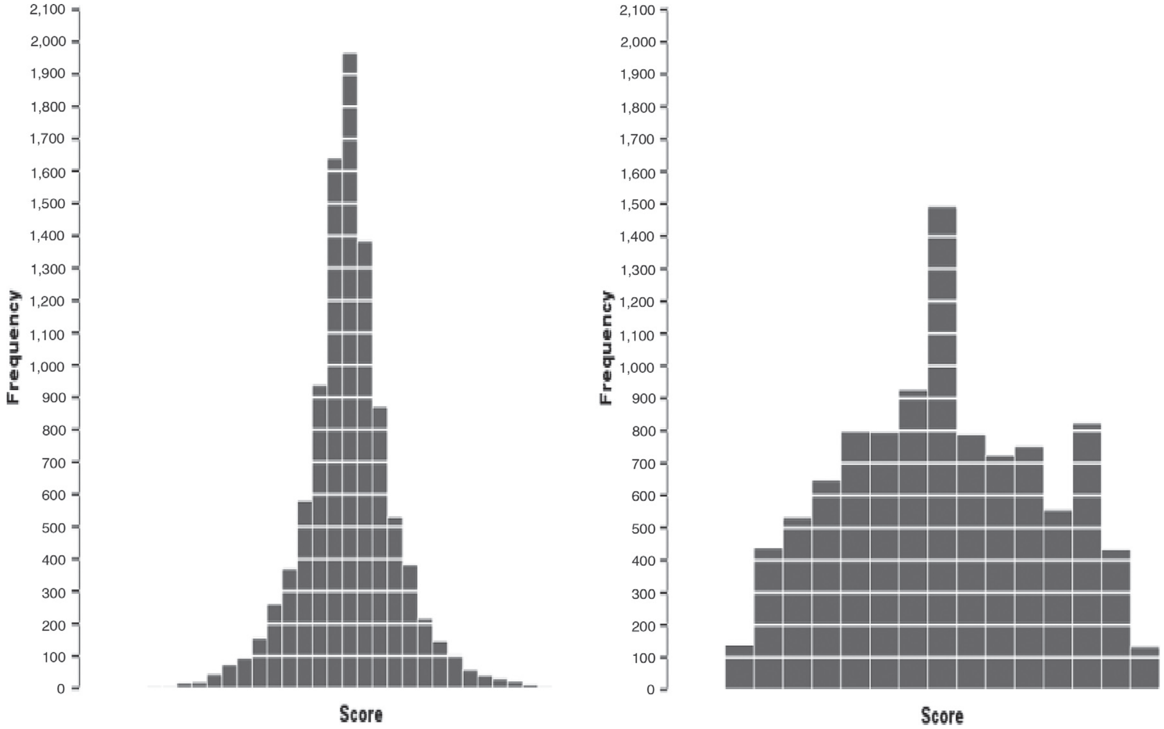


FIGURE 1.5 Distributions with positive kurtosis (leptokurtic, left figure) and negative kurtosis (platykurtic, right figure)

1.7.2. The centre of a distribution ①

We can also calculate where the centre of a frequency distribution lies (known as the **central tendency**). There are three measures commonly used: the mean, the mode and the median.

1.7.2.1. The mode ①

The **mode** is simply 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, simply 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 often 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**. It's also possible to find data sets with more than two modes (**multimodal**). Also, if the frequencies of certain scores are very similar, then the mode can be influenced by only a small number of cases.

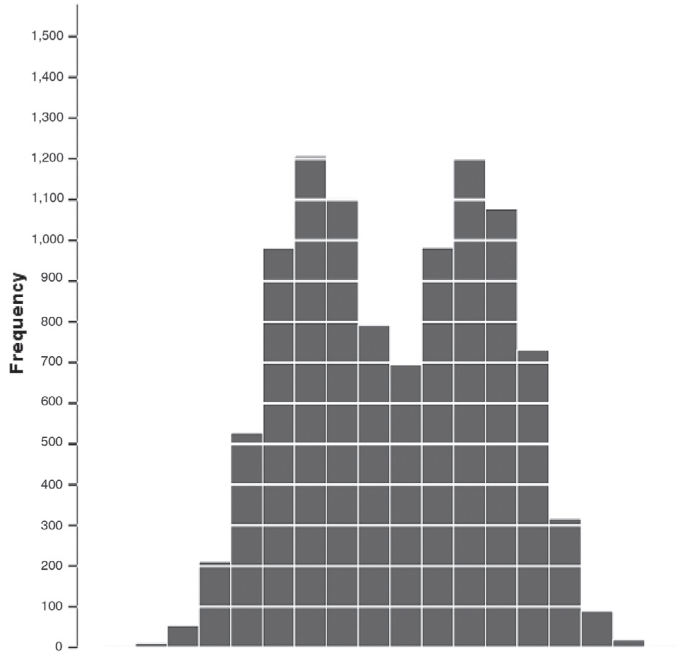


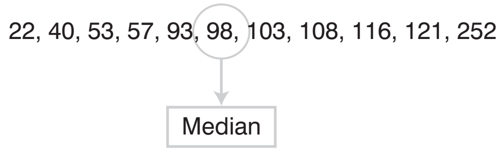
FIGURE 1.6 A bimodal distribution

1.7.2.2. The median ①

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**. For example, Facebook is a popular social networking website, in which users can sign up to be 'friends' of other users. Imagine we looked at the number of friends that a selection (actually, some of my friends) of 11 Facebook users had. Number of friends: 108, 103, 252, 121, 93, 57, 40, 53, 22, 116, 98.

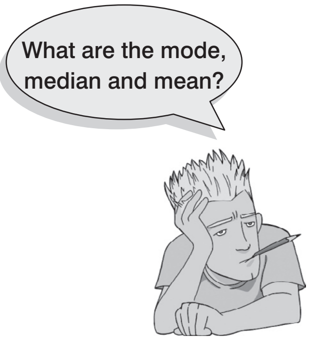
To calculate the median, we first arrange these scores into ascending order: 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252.

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:



22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252

Median



This 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 (more than twice as large as the next biggest number), we would ignore it. (For one thing, this person is far too popular and we hate them.) We have only 10 scores now. As before, we should 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 $(n + 1)/2 = 11/2 = 5.5$. This 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 252. 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).

1.7.2.3. The mean ①

The **mean** is the measure of central tendency that you are most likely to have heard of because it is simply the average score and the media are full of average scores.[13] To calculate the mean we simply 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:

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

This may look complicated, but the top half of the equation simply means 'add up all of the scores' (the x_i just 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 first add up all of the scores:

$$\begin{aligned}\sum_{i=1}^{n} x_i &= 22 + 40 + 53 + 57 + 93 + 98 + 103 + 108 + 116 + 121 + 252 \\ &= 1063\end{aligned}$$

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

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1063}{11} = 96.64$$

The mean is 96.64 friends, which is not a value we observed in our actual data (it would be ridiculous to talk of having 0.64 of a friend). In this sense the mean is a statistical model – more on this in the next chapter.

[13] I'm writing this on 15 February 2008, and to prove my point the BBC website is running a headline about how PayPal estimates that Britons will spend an average of £71.25 each on Valentine's Day gifts, but uSwitch.com said that the average spend would be £22.69!

SELF-TEST Compute the mean but excluding the score of 252.

If you calculate the mean without our extremely popular person (i.e. excluding the value 252), the mean drops to 81.1 friends. One disadvantage of the mean is that it can be influenced by extreme scores. In this case, the person with 252 friends on Facebook increased the mean by about 15 friends! Compare this difference with that of the median. Remember that the median hardly changed if we included or excluded 252, which illustrates how the median is 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 all of the time? 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.

1.7.3. The dispersion in a distribution ①

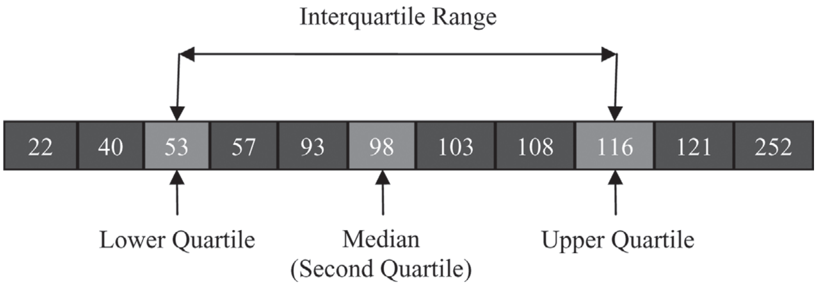
It can also be interesting to try to quantify the spread, or dispersion, of scores in the data. The easiest way to look at dispersion is to take the largest score and subtract from it the smallest score. This is known as the **range** of scores. For our Facebook friends data, if we order these scores we get 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252. The highest score is 252 and the lowest is 22; therefore, the range is 252 – 22 = 230. One problem with the range is that because it uses only the highest and lowest score it is affected dramatically by extreme scores.

SELF-TEST Compute the range but excluding the score of 252.

If you have done the self-test task you'll see that without the extreme score the range drops dramatically from 230 to 99 – less than half the size!

One way around this problem is to calculate the range when we exclude 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. One rule of thumb is that 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 1.7 shows how we would calculate these values for the Facebook data. Like the median, the upper and lower quartile need not be values that actually appear in the 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

FIGURE 1.7 Calculating quartiles and the interquartile range



of two values in the data set). 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!).

SELF-TEST 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

1.7.4. Using a frequency distribution to go beyond the data ①

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' induces suicidal ideation in most people (myself included) so it seems fitting that we use an example about throwing ourselves off a cliff. Beachy Head is a large, windy cliff on the Sussex coast (not far from where I live) that has something of a reputation for attracting suicidal people, who seem to like throwing themselves off it (and after several months of rewriting this book I find my thoughts drawn towards that peaceful chalky cliff top more and more often). Figure 1.8 shows a frequency distribution of some completely made up data of the number of suicides at Beachy Head in a year by people of different ages (although I made these data up, they are roughly based on general suicide statistics such as those in Williams, 2001). There were 172 suicides in total and you can see that the suicides were most frequently aged between about 30 and 35 (the highest bar). The graph also tells us that, for example, very few people aged above 70 committed suicide at Beachy Head.

I said earlier that we could think of frequency distributions in terms of probability. To explain this, imagine that someone asked you 'how likely is it that a 70 year old committed suicide at Beach Head?' What would your answer be? The chances are that if you looked at the frequency distribution you might respond 'not very likely' because you can see that only 3 people out of the 172 suicides were aged around 70. What about if someone asked you 'how likely is it that a 30 year old committed suicide?' Again, by looking at the graph, you might say 'it's actually quite likely' because 33 out of the 172 suicides were by people aged around 30 (that's more than 1 in every 5 people who committed suicide). So based

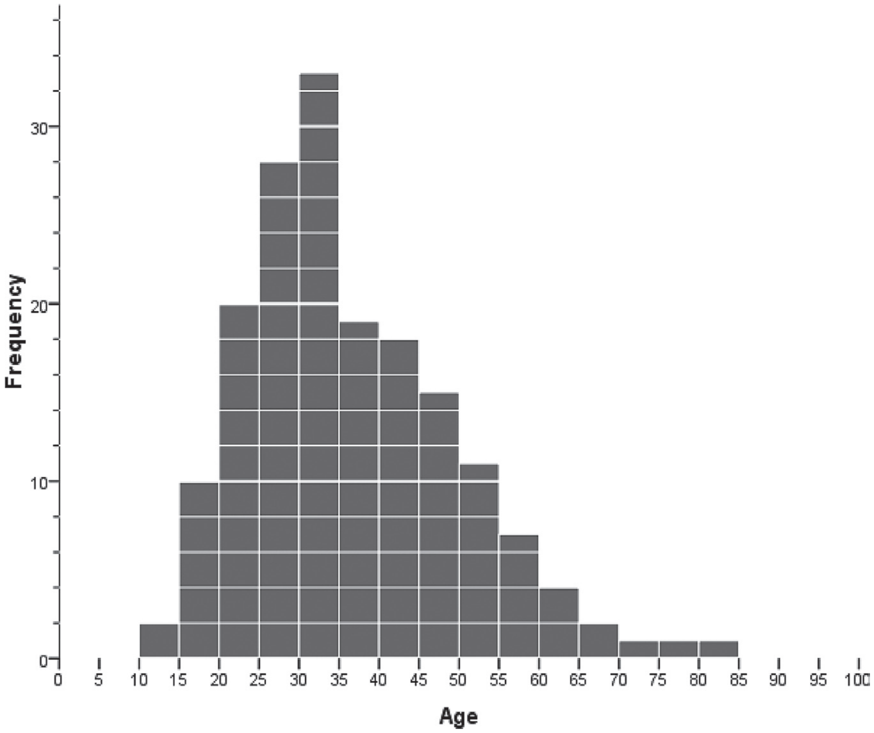
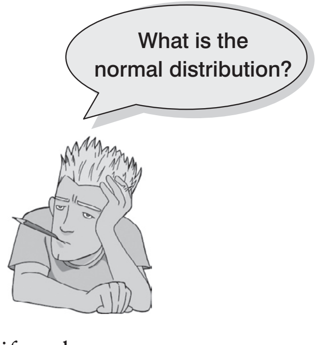


FIGURE 1.8 Frequency distribution showing the number of suicides at Beachy Head in a year by age

on the frequencies of different scores it should start to become clear that we could use this information to estimate the probability that a particular score will occur. We could ask, based on our data, 'what's the probability of a suicide victim being aged 16–20?' 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 April 2008. However, when I talk to anyone else, I might, more realistically, tell them that there's a .10 probability of me finishing the revisions on time (or put another way, a 10% chance, or 1 in 10 chance that I'll complete the book in time). In reality, the probability of my meeting the deadline is 0 (not a chance in hell) because I never manage to meet publisher's deadlines! If probabilities don't make sense to you then just ignore the decimal point and think of them as percentages instead (i.e. .10 probability that something will happen = 10% chance that something will happen).

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 that specify idealized versions of these distributions (they are specified in terms of a curved line). These idealized distributions are known as **probability distributions** and from these distributions it is possible to calculate the probability of getting particular scores based on the frequencies with which a particular score occurs in a distribution with these common shapes. One of these 'common' distributions is the normal distribution, which I've already mentioned in section 1.7.1. Statisticians have calculated the probability of certain scores occurring in a normal distribution with a mean of 0 and a standard deviation of 1. Therefore, if we have any data that are shaped like a normal distribution, then if the mean and standard deviation



are 0 and 1 respectively we can use the tables of probabilities for the normal distribution to see how likely it is that a particular score will occur in the data (I've produced such a table in the Appendix to this book).

The obvious problem is that not all of the data we collect will have a mean of 0 and standard deviation of 1. For example, we might have a data set that has a mean of 567 and a standard deviation of 52.98. Luckily 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 and subtract from it the mean of all. Then, we divide the resulting score by the standard deviation to ensure the data have a standard deviation of 1. The resulting scores are known as **z-scores** and in equation form, the conversion that I've just described is:

$$z = \frac{X - \overline{X}}{s} \tag{1.2}$$

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 suicide data, we can answer the question 'What's the probability that someone who threw themselves off of Beachy Head was 70 or older?' First we convert 70 into a *z*-score. Say, the mean of the suicide scores was 36, and the standard deviation 13; then 70 will become (70 – 36)/13 = 2.62. We then look up this value in the column labelled 'Smaller Portion' (i.e. the area above the value 2.62). You should find that the probability is .0044, or put another way, only a 0.44% chance that a suicide victim would be 70 years old or more. By looking at the column labelled 'Bigger Portion' we can also see the probability that a suicide victim was aged 70 or less. This probability is .9956, or put another way, there's a 99.56% chance that a suicide victim was less than 70 years old.

Hopefully you can see from these examples that the normal distribution and *z*-scores allow us to go a first step beyond our data in that from a set of scores we can calculate the probability that a particular score will occur. So, we can see whether scores of a certain size are likely or unlikely to occur in a distribution of a particular kind. You'll see just how useful this is in due course, but it is worth mentioning at this stage that certain *z*-scores are particularly important. This is because their value cuts off certain important percentages of the distribution. The first important value of *z* is 1.96 because this cuts off the top 2.5% of the distribution, and its counterpart at the opposite end (–1.96) cuts off the bottom 2.5% of the distribution. As such, taken together, this value cuts of 5% of scores, or put another way, 95% of *z*-scores lie between –1.96 and 1.96. The other two important benchmarks are ±2.58 and ±3.29, which cut off 1% and 0.1% of scores respectively. Put another way, 99% of *z*-scores lie between –2.58 and 2.58, and 99.9% of them lie between –3.29 and 3.29. Remember these values because they'll crop up time and time again.

SELF-TEST Assuming the same mean and standard deviation for the Beachy Head example above, what's the probability that someone who threw themselves off Beachy Head was 30 or younger?

1.7.5. Fitting statistical models to the data ①

Having looked at your data (and there is a lot more information on different ways to do this in Chapter 4), the next step is to fit a statistical model to the data. I should really just write 'insert the rest of the book here', because most of the remaining chapters discuss the various models that you can fit to the data. However, I do want to talk here briefly about two very important types of hypotheses that are used when analysing the data. Scientific statements, as we have seen, can be split into testable hypotheses. The hypothesis or prediction that comes from your theory is usually saying that an effect will be present. This hypothesis is called the **alternative hypothesis** and is denoted by H_1. (It is sometimes also called the *experimental hypothesis* but because this term relates to a specific type of methodology it's probably best to use 'alternative hypothesis'.) There is another type of hypothesis, though, and this is called the **null hypothesis** and is denoted by H_0. This hypothesis is the opposite of the alternative hypothesis and so would usually state that an effect is absent. Taking our *Big Brother* example from earlier in the chapter we might generate the following hypotheses:

- **Alternative hypothesis:** *Big Brother* contestants will score higher on personality disorder questionnaires than members of the public.
- **Null hypothesis**: *Big Brother* contestants and members of the public will not differ in their scores on personality disorder questionnaires.

The reason that we need the null hypothesis is because we cannot prove the experimental hypothesis using statistics, but we can reject the null hypothesis. If our data give us confidence to reject the null hypothesis then this provides support for our experimental hypothesis. However, be aware that even if we can reject the null hypothesis, this doesn't prove the experimental hypothesis – it merely supports it. So, rather than talking about accepting or rejecting a hypothesis (which some textbooks tell you to do) we should be talking about 'the chances of obtaining the data we've collected assuming that the null hypothesis is true'.

Using our *Big Brother* example, when we collected data from the auditions about the contestants' personalities we found that 75% of them had a disorder. When we analyse our data, we are really asking, 'Assuming that contestants are no more likely to have personality disorders than members of the public, is it likely that 75% or more of the contestants would have personality disorders?' Intuitively the answer is that the chances are very low: if the null hypothesis is true, then most contestants would not have personality disorders because they are relatively rare. Therefore, we are very unlikely to have got the data that we did if the null hypothesis were true.

What if we found that only 1 contestant reported having a personality disorder (about 8%)? If the null hypothesis is true, and contestants are no different in personality to the general population, then only a small number of contestants would be expected to have a personality disorder. The chances of getting these data if the null hypothesis is true are, therefore, higher than before.

When we collect data to test theories we have to work in these terms: we cannot talk about the null hypothesis being true or the experimental hypothesis being true, we can only talk in terms of the probability of obtaining a particular set of data if, hypothetically speaking, the null hypothesis was true. We will elaborate on this idea in the next chapter.

Finally, hypotheses can also be directional or non-directional. A directional hypothesis states that an effect will occur, but it also states the direction of the effect. For example, 'readers will know more about research methods after reading this chapter' is a one-tailed hypothesis because it states the direction of the effect (readers will know more). A non-directional hypothesis states that an effect will occur, but it doesn't state the direction of the effect. For example, 'readers' knowledge of research methods will change after they have read this chapter' does not tell us whether their knowledge will improve or get worse.

What have I discovered about statistics? ①

Actually, not a lot because we haven't really got to the statistics bit yet. However, we have discovered some stuff about the process of doing research. We began by looking at how research questions are formulated through observing phenomena or collecting data about a 'hunch'. Once the observation has been confirmed, theories can be generated about why something happens. From these theories we formulate hypotheses that we can test. To test hypotheses we need to measure things and this leads us to think about the variables that we need to measure and how to measure them. Then we can collect some data. The final stage is to analyse these data. In this chapter we saw that we can begin by just looking at the shape of the data but that ultimately we should end up fitting some kind of statistical model to the data (more on that in the rest of the book). In short, the reason that your evil statistics lecturer is forcing you to learn statistics is because it is an intrinsic part of the research process and it gives you enormous power to answer questions that are interesting; or it could be that they are a sadist who spends their spare time spanking politicians while wearing knee-high PVC boots, a diamond-encrusted leather thong and a gimp mask (that'll be a nice mental image to keep with you throughout your course). We also discovered that I was a curious child (you can interpret that either way). As I got older I became more curious, but you will have to read on to discover what I was curious about.

Key terms that I've discovered

Alternative hypothesis
Between-group design
Between-subject design
Bimodal
Binary variable
Boredom effect
Categorical variable
Central tendency
Confounding variable
Content validity
Continuous variable
Correlational research
Counterbalancing
Criterion validity
Cross-sectional research
Dependent variable
Discrete variable
Ecological validity
Experimental hypothesis
Experimental research
Falsification
Frequency distribution
Histogram
Hypothesis
Independent design
Independent variable
Interquartile range
Interval variable
Kurtosis
Leptokurtic
Level of measurement
Lower quartile
Mean
Measurement error
Median
Mode
Multimodal
Negative skew
Nominal variable
Normal distribution
Null hypothesis
Ordinal variable
Outcome variable
Platykurtic
Positive skew
Practice effect

Predictor variable
Probability distribution
Qualitative methods
Quantitative methods
Quartile
Randomization
Range
Ratio variable
Reliability
Repeated-measures design
Second quartile
Skew
Systematic variation
Tertium quid
Test–retest reliability
Theory
Unsystematic variance
Upper quartile
Validity
Variables
Within-subject design
z-scores

Smart Alex's tasks

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

- **Task 1:** What are (broadly speaking) the five stages of the research process? ①
- **Task 2:** What is the fundamental difference between experimental and correlational research? ①
- **Task 3:** What is the level of measurement of the following variables? ①
  a. The number of downloads of different bands' songs on iTunes.
  b. The names of the bands that were downloaded.
  c. The position in the iTunes download chart.
  d. The money earned by the bands from the downloads.
  e. The weight of drugs bought by the bands with their royalties.
  f. The type of drugs bought by the bands with their royalties.
  g. The phone numbers that the bands obtained because of their fame.
  h. The gender of the people giving the bands their phone numbers.
  i. The instruments played by the band members.
  j. 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 friends CD-counting device? ①
- **Task 5:** Sketch the shape of a normal distribution, a positively skewed distribution and a negatively skewed distribution. ①

Answers can be found on the companion website.

Further reading

Field, A. P., & Hole, G. J. (2003). *How to design and report experiments*. London: Sage. (I am rather biased, but I think this is a good overview of basic statistical theory and research methods.)

Miles, J. N. V., & Banyard, P. (2007). *Understanding and using statistics in psychology: a practical introduction*. London: Sage. (A fantastic and amusing introduction to statistical theory.)

Wright, D. B., & London, K. (2009). *First steps in statistics* (2nd ed.). London: Sage. (This book is a very gentle introduction to statistical theory.)

Interesting real research

Umpierre, S. A., Hill, J. A., & Anderson, D. J. (1985). Effect of Coke on sperm motility. *New England Journal of Medicine*, *313*(21), 1351.