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.

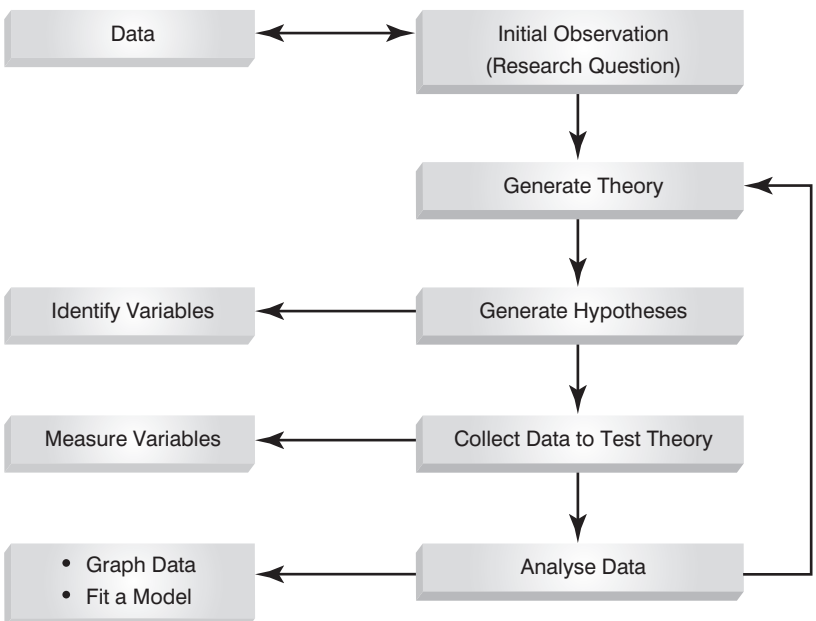


FIGURE 1.2 The research process

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).

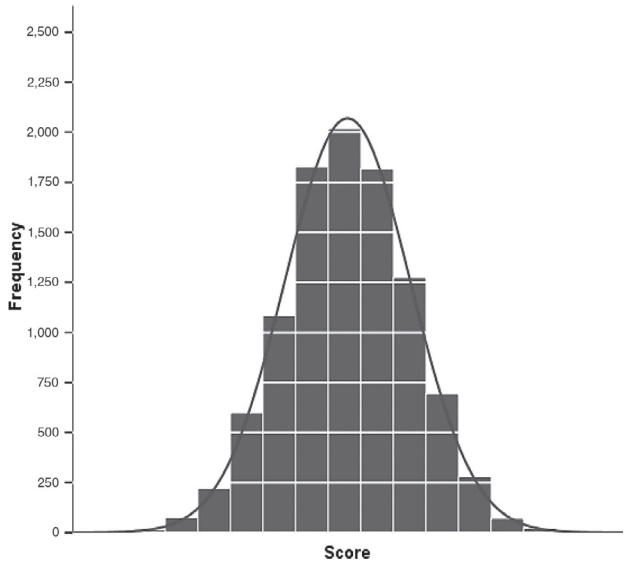
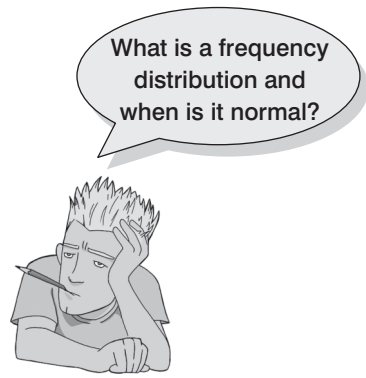


FIGURE 1.3
A 'normal' distribution (the curve shows the idealized shape)

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.

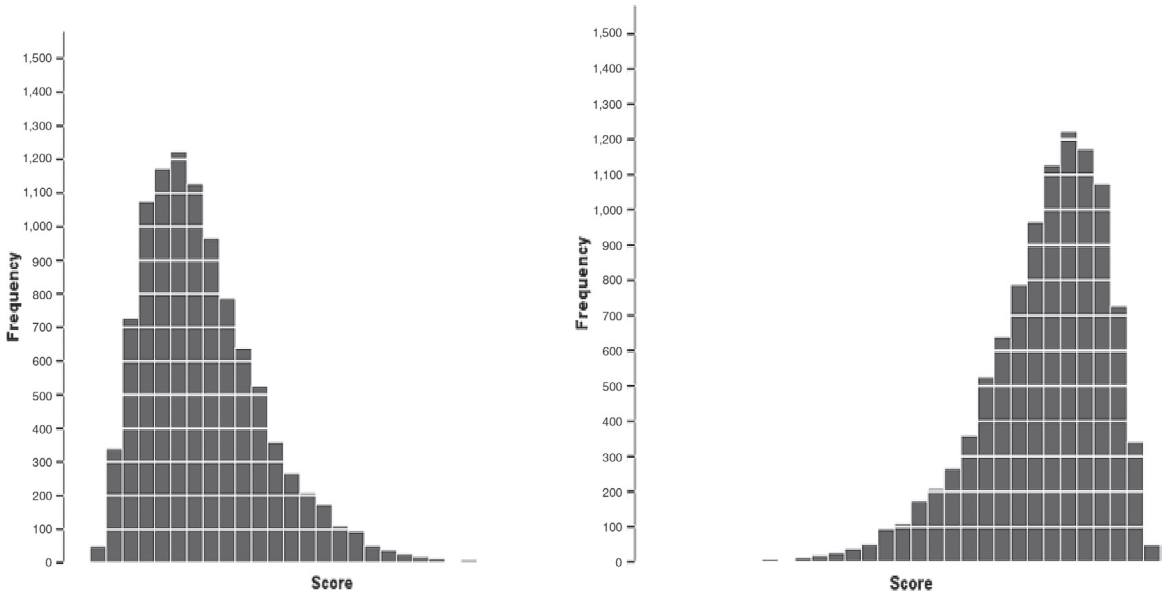


FIGURE 1.4 A positively (left figure) and negatively (right figure) skewed distribution

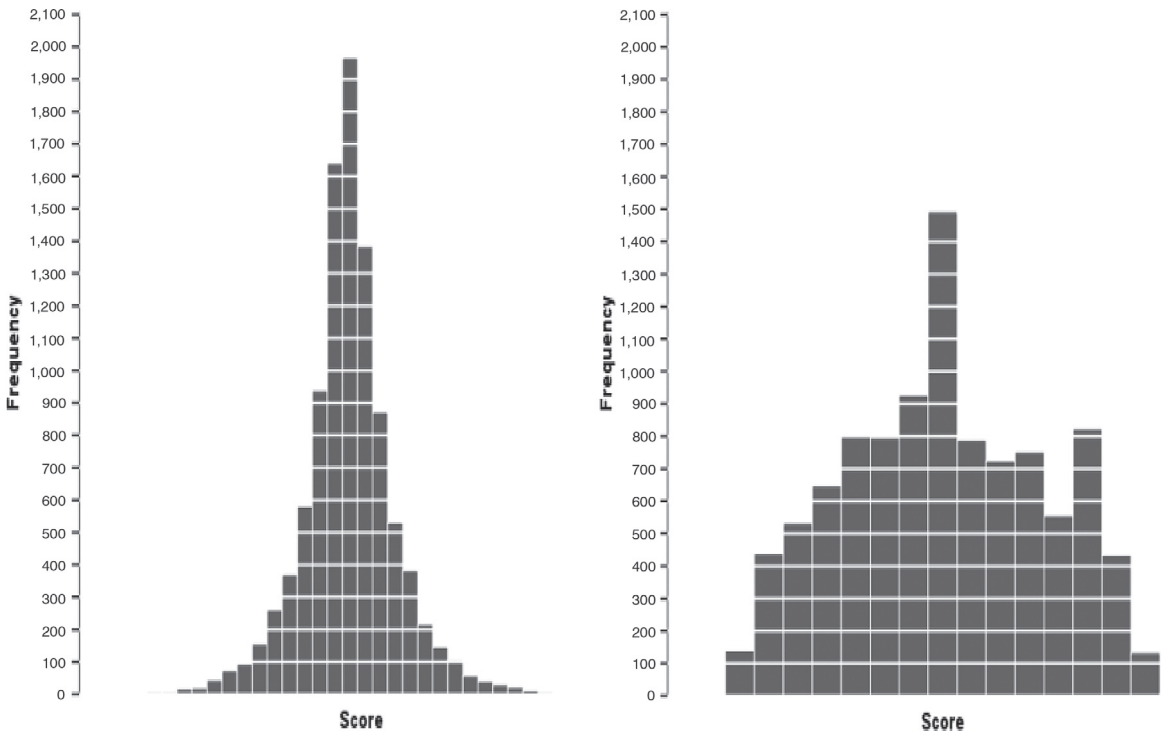


FIGURE 1.5 Distributions with positive kurtosis (leptokurtic, left figure) and negative kurtosis (platykurtic, right figure)

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.

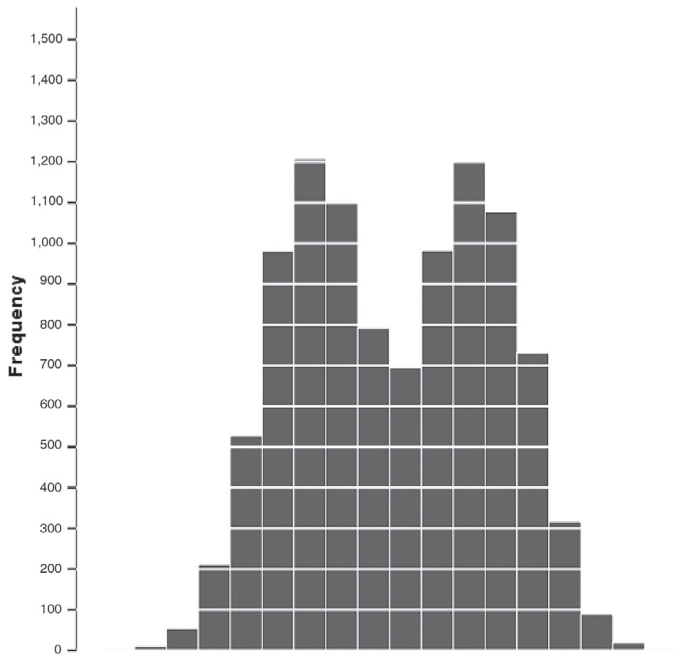


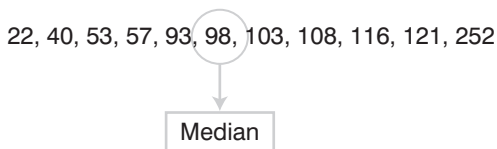
FIGURE 1.6
A bimodal distribution

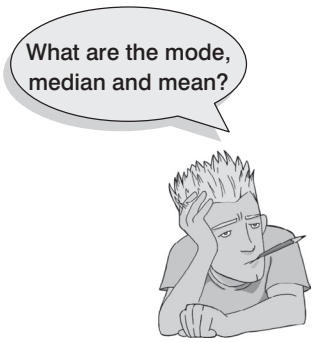
1.7.2.2. The median ①

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:





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

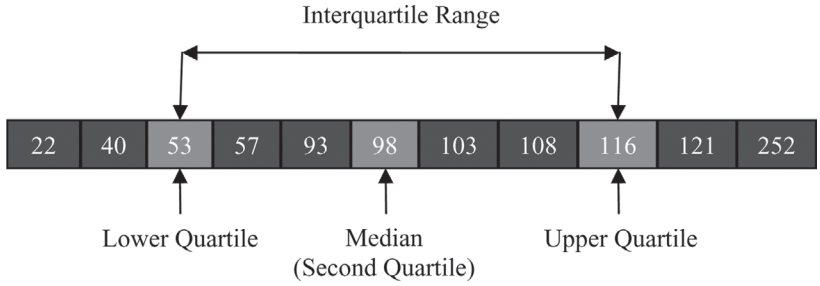


FIGURE 1.7 Calculating quartiles and the interquartile range

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

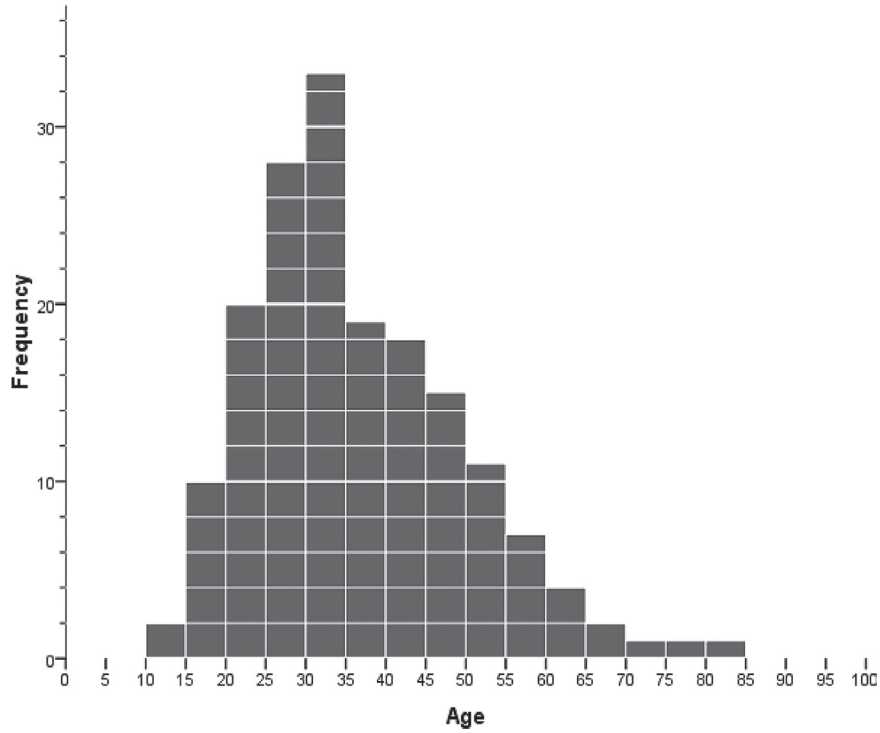
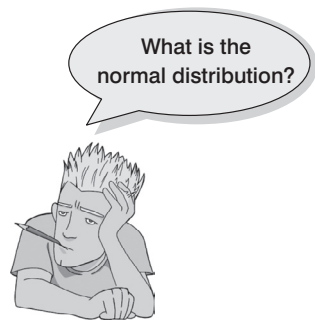


FIGURE 1.8 Frequency distribution showing the number of suicides at Beachy Head in a year by age

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.

Everything you ever wanted to know about statistics (well, sort of)

2

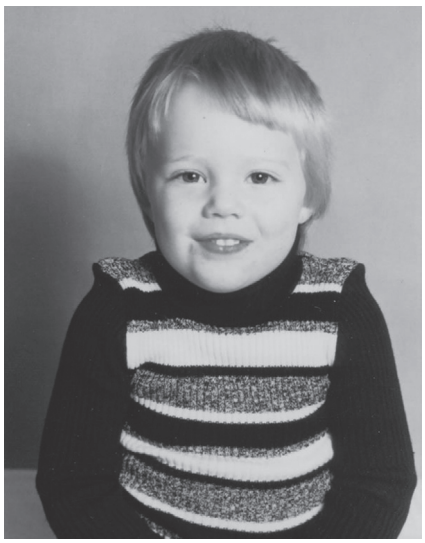


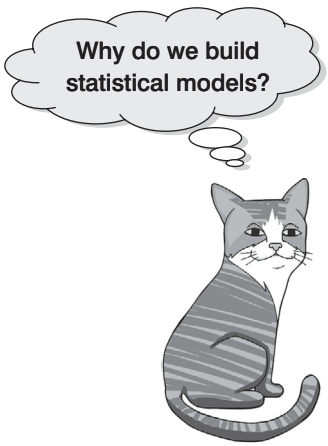
FIGURE 2.1
The face of innocence … but what are the hands doing?

2.1. What will this chapter tell me? ①

As a child grows, it becomes important for them to fit models to the world: to be able to reliably predict what will happen in certain situations. This need to build models that accurately reflect reality is an essential part of survival. According to my parents (conveniently I have no memory of this at all), while at nursery school one model of the world that I was particularly enthusiastic to try out was 'If I get my penis out, it will be really funny'. No doubt to my considerable disappointment, this model turned out to be a poor predictor of positive outcomes. Thankfully for all concerned, I soon learnt that the model 'If I get my penis out at nursery school the teachers and mummy and daddy are going to be quite annoyed' was

a better 'fit' of the observed data. Fitting models that accurately reflect the observed data is important to establish whether a theory is true. You'll be delighted to know that this chapter is all about fitting statistical models (and not about my penis). We edge sneakily away from the frying pan of research methods and trip accidentally into the fires of statistics hell. We begin by discovering what a statistical model is by using the mean as a straightforward example. We then see how we can use the properties of data to go beyond the data we have collected and to draw inferences about the world at large. In a nutshell then, this chapter lays the foundation for the whole of the rest of the book, so it's quite important that you read it or nothing that comes later will make any sense. Actually, a lot of what comes later probably won't make much sense anyway because *I've* written it, but there you go.

2.2. Building statistical models ①



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

The reason for building statistical models of real-world data is best explained by analogy. Imagine an engineer wishes to build a bridge across a river. That engineer would be pretty daft if she just built any old bridge, because the chances are that it would fall down. Instead, an engineer collects data from the real world: she looks at bridges in the real world and sees what materials they are made from, what structures they use and so on (she might even collect data about whether these bridges are damaged!). She then uses this information to construct a model. She builds a scaled-down version of the real-world bridge because it is impractical, not to mention expensive, to build the actual bridge itself. The model may differ from reality in several ways – it will be smaller for a start – but the engineer will try to build a model that best fits the situation of interest based on the data available. Once the model has been built, it can be used to predict things about the real world: for example, the engineer might test whether the bridge can withstand strong winds by placing the model in a wind tunnel. It seems obvious that it is important that the model is an accurate representation of the real world. Social scientists do much the same thing as engineers: they build models of real-world processes in an attempt to predict how these processes operate under certain conditions (see Jane Superbrain Box 2.1 below). We don't have direct access to the processes, so we collect data that represent the processes and then use these data to build statistical models (we reduce the process to a statistical model). We then use this statistical model to make predictions about the real-world phenomenon. Just like the engineer, we want our models to be as accurate as possible so that we can be confident that the predictions we make are also accurate. However, unlike engineers we don't have access to the real-world situation and so we can only ever *infer* things about psychological, societal, biological or economic processes based upon the models we build. If we want our inferences to be accurate then the statistical model we build must represent the data collected (the *observed data*) as closely as possible. The degree to which a statistical model represents the data collected is known as the **fit** of the model.

Figure 2.2 illustrates the kinds of models that an engineer might build to represent the real-world bridge that she wants to create. The first model (a) is an excellent representation of the real-world situation and is said to be a *good fit* (i.e. there are a few small differences but the model is basically a very good replica of reality). If this model is used to make predictions about the real world, then the engineer can be confident that these predictions will be very accurate, because the model so closely resembles reality. So, if the model collapses in a strong

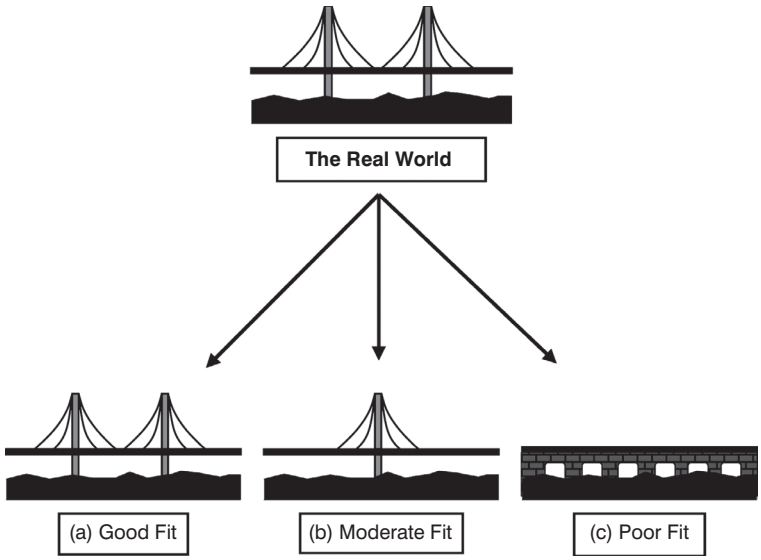


FIGURE 2.2 Fitting models to real-world data (see text for details)

wind, then there is a good chance that the real bridge would collapse also. The second model (b) has some similarities to the real world: the model includes some of the basic structural features, but there are some big differences from the real-world bridge (namely the absence of one of the supporting towers). This is what we might term a *moderate fit* (i.e. there are some differences between the model and the data but there are also some great similarities). If the engineer uses this model to make predictions about the real world then these predictions may be inaccurate and possibly catastrophic (e.g. the model predicts that the bridge will collapse in a strong wind, causing the real bridge to be closed down, creating 100-mile tailbacks with everyone stranded in the snow; all of which was unnecessary because the real bridge was perfectly safe – the model was a bad representation of reality). We can have some confidence, but not complete confidence, in predictions from this model. The final model (c) is completely different to the real-world situation; it bears no structural similarities to the real bridge and is a poor fit (in fact, it might more accurately be described as an abysmal fit!). As such, any predictions based on this model are likely to be completely inaccurate. Extending this analogy to the social sciences we can say that it is important when we fit a statistical model to a set of data that this model fits the data well. If our model is a poor fit of the observed data then the predictions we make from it will be equally poor.

JANE SUPERBRAIN 2.1

Types of statistical models ①

As behavioural and social scientists, most of the models that we use to describe data tend to be **linear models**. For example, analysis of variance (ANOVA) and regression are identical systems based on linear models (Cohen, 1968), yet they have different names and, in psychology at least, are used largely in different contexts due to historical divisions in methodology (Cronbach, 1957).

A linear model is simply a model that is based upon a straight line; this means that we are usually trying to summarize our observed data in terms of a straight line. Suppose we measured how many chapters of this book a person had read, and then measured their spiritual enrichment. We could represent these hypothetical data in the form of a scatterplot in which each dot represents an individual's score on both variables (see section 4.7). Figure 2.3 shows two versions of such a graph that summarize the pattern of these data with either a straight

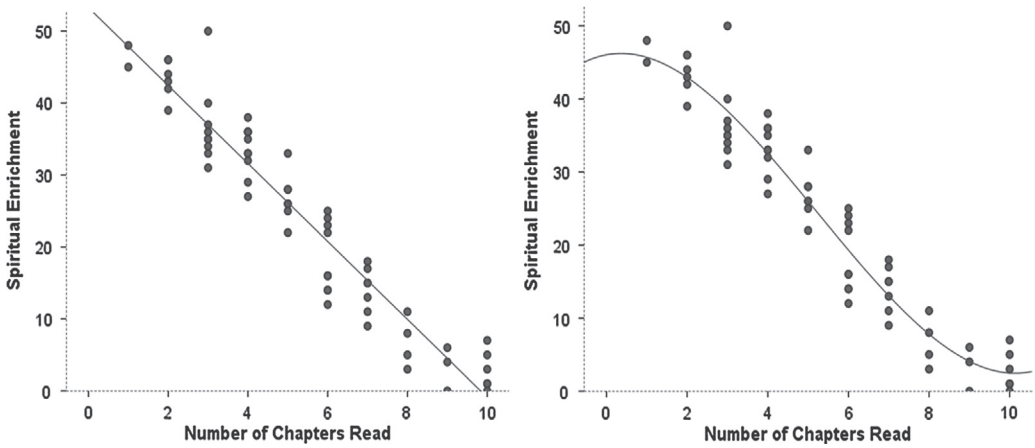


FIGURE 2.3 A scatterplot of the same data with a linear model fitted (left), and with a non-linear model fitted (right)

(left) or curved (right) line. These graphs illustrate how we can fit different types of models to the same data. In this case we can use a straight line to represent our data and it shows that the more chapters a person reads, the less their spiritual enrichment. However, we can also use a curved line to summarize the data and this shows that when most, or all, of the chapters have been read, spiritual enrichment seems to increase slightly (presumably because once the book is read everything suddenly makes sense – yeah, as if!). Neither of the two types of model is necessarily correct, but it will be the case that one model fits the data better than another and this is why when we use statistical models it is important for us to assess how well a given model fits the data.

It's possible that many scientific disciplines are progressing in a biased way because most of the models that we tend to fit are linear (mainly because books like this tend to ignore more complex curvilinear models). This could create a bias because most published scientific studies are ones with statistically significant results and there may be cases where a linear model has been a poor fit of the data (and hence the paper was not published), yet a non-linear model would have fitted the data well. This is why it is useful to plot your data first: plots tell you a great deal about what models should be applied to data. If your plot seems to suggest a non-linear model then investigate this possibility (which is easy for me to say when I don't include such techniques in this book!).

2.3. Populations and samples ①

As researchers, we are interested in finding results that apply to an entire population of people or things. For example, psychologists want to discover processes that occur in all humans, biologists might be interested in processes that occur in all cells, economists want to build models that apply to all salaries, and so on. A population can be very general (all human beings) or very narrow (all male ginger cats called Bob). Usually, scientists strive to infer things about general populations rather than narrow ones. For example, it's not very interesting to conclude that psychology students with brown hair who own a pet hamster named George recover more quickly from sports injuries if the injury is massaged (unless, like René Koning,[1] you happen to be a psychology student with brown hair who has a pet hamster named George). However, if we can conclude that *everyone's* sports injuries are aided by massage this finding has a much wider impact.

Scientists rarely, if ever, have access to every member of a population. Psychologists cannot collect data from every human being and ecologists cannot observe every male ginger cat called Bob. Therefore, we collect data from a small subset of the population (known as a **sample**) and use these data to infer things about the population as a whole. The bridge-building

[1] A brown-haired psychology student with a hamster called Sjors (Dutch for George, apparently), who, after reading one of my web resources, emailed me to weaken my foolish belief that this is an obscure combination of possibilities.

engineer cannot make a full-size model of the bridge she wants to build and so she builds a small-scale model and tests this model under various conditions. From the results obtained from the small-scale model the engineer infers things about how the full-sized bridge will respond. The small-scale model may respond differently to a full-sized version of the bridge, but the larger the model, the more likely it is to behave in the same way as the full-size bridge. This metaphor can be extended to scientists. We never have access to the entire population (the real-size bridge) and so we collect smaller samples (the scaled-down bridge) and use the behaviour within the sample to infer things about the behaviour in the population. The bigger the sample, the more likely it is to reflect the whole population. If we take several random samples from the population, each of these samples will give us slightly different results. However, on average, large samples should be fairly similar.

2.4. Simple statistical models ①

2.4.1. The mean: a very simple statistical model ①

One of the simplest models used in statistics is the mean, which we encountered in section 1.7.2.3. In Chapter 1 we briefly mentioned that the mean was a statistical model of the data because it is a hypothetical value that doesn't have to be a value that is actually observed in the data. For example, if we took five statistics lecturers and measured the number of friends that they had, we might find the following data: 1, 2, 3, 3 and 4. If we take the mean number of friends, this can be calculated by adding the values we obtained, and dividing by the number of values measured: $(1 + 2 + 3 + 3 + 4)/5 = 2.6$. Now, we know that it is impossible to have 2.6 friends (unless you chop someone up with a chainsaw and befriend their arm, which frankly is probably not beyond your average statistics lecturer) so the mean value is a *hypothetical* value. As such, the mean is a model created to summarize our data.

2.4.2. Assessing the fit of the mean: sums of squares, variance and standard deviations ①

With any statistical model we have to assess the fit (to return to our bridge analogy we need to know how closely our model bridge resembles the real bridge that we want to build). With most statistical models we can determine whether the model is accurate by looking at how different our real data are from the model that we have created. The easiest way to do this is to look at the difference between the data we observed and the model fitted. Figure 2.4 shows the number of friends that each statistics lecturer had, and also the mean number that we calculated earlier on. The line representing the mean can be thought of as our model, and the circles are the observed data. The diagram also has a series of vertical lines that connect each observed value to the mean value. These lines represent the **difference** between the observed data and our model and can be thought of as the error in the model. We can calculate the magnitude of these deviances by simply subtracting the mean value ($\bar{x}$) from each of the observed values (x_i).[2] For example, lecturer 1 had only 1 friend (a glove puppet of an ostrich called Kevin) and so the difference is $x_1 - \bar{x} = 1 - 2.6 = -1.6$. You might notice that the deviance is a negative number, and this represents the fact that our model *overestimates* this lecturer's popularity: it

[2] The x_i simply refers to the observed score for the ith person (so, the i can be replaced with a number that represents a particular individual). For these data: for lecturer 1, $x_i = x_1 = 1$; for lecturer 3, $x_i = x_3 = 3$; for lecturer 5, $x_i = x_5 = 4$.

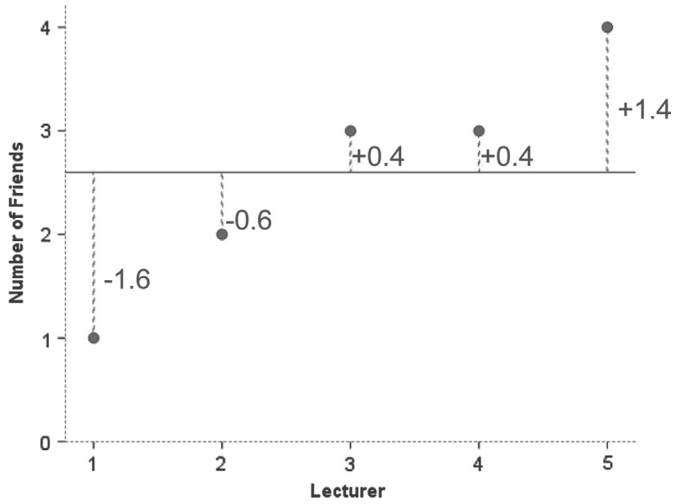


FIGURE 2.4 Graph showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends

predicts that he will have 2.6 friends yet in reality he has only 1 friend (bless him!). Now, how can we use these deviances to estimate the accuracy of the model? One possibility is to add up the deviances (this would give us an estimate of the total error). If we were to do this we would find that (don't be scared of the equations, we will work through them step by step – if you need reminding of what the symbols mean there is a guide at the beginning of the book):

$$\begin{aligned}\text{total error} &= \text{sum of deviances} \\ &= \sum(x_i - \overline{x}) = (-1.6) + (-0.6) + (0.4) + (0.4) + (1.4) = 0\end{aligned}$$

So, in effect the result tells us that there is no total error between our model and the observed data, so the mean is a perfect representation of the data. Now, this clearly isn't true: there were errors but some of them were positive, some were negative and they have simply cancelled each other out. It is clear that we need to avoid the problem of which direction the error is in and one mathematical way to do this is to square each error,[3] that is, multiply each error by itself. So, rather than calculating the sum of errors, we calculate the sum of squared errors. In this example:

$$\begin{aligned}\text{sum of squrared errors (SS)} &= \sum(x_i - \overline{x})(x_i - \overline{x}) \\ &= (-1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2 \\ &= 2.56 + 0.36 + 0.16 + 0.16 + 1.96 \\ &= 5.20\end{aligned}$$

The **sum of squared errors** (SS) is a good measure of the accuracy of our model. However, it is fairly obvious that the sum of squared errors is dependent upon the amount of data that has been collected – the more data points, the higher the SS. To overcome this problem we calculate the average error by dividing the SS by the number of observations (N). If we are interested only in the average error for the sample, then we can divide by N alone. However, we are generally interested in using the error in the sample to estimate the error in the population and so we divide the SS by the number of observations minus 1 (the reason why is explained in Jane Superbrain Box 2.2). This measure is known as the **variance** and is a measure that we will come across a great deal:

[3] When you multiply a negative number by itself it becomes positive.

JANE SUPERBRAIN 2.2

Degrees of freedom ②

Degrees of freedom (*df*) is a very difficult concept to explain. I'll begin with an analogy. Imagine you're the manager of a rugby team and you have a team sheet with 15 empty slots relating to the positions on the playing field. There is a standard formation in rugby and so each team has 15 specific positions that must be held constant for the game to be played. When the first player arrives, you have the choice of 15 positions in which to place this player. You place his name in one of the slots and allocate him to a position (e.g. scrum-half) and, therefore, one position on the pitch is now occupied. When the next player arrives, you have the choice of 14 positions but you still have the freedom to choose which position this player is allocated. However, as more players arrive, you will reach the point at which 14 positions have been filled and the final player arrives. With this player you have no freedom to choose where they play – there is only one position left. Therefore there are 14 degrees of freedom; that is, for 14 players you have some degree of choice over where they play, but for 1 player you have no choice. The degrees of freedom is one less than the number of players.

In statistical terms the degrees of freedom relate to the number of observations that are free to vary. If we take a sample of four observations from a population, then these four scores are free to vary in any way (they can be any value). However, if we then use this sample of four observations to calculate the standard deviation of the population, we have to use the mean of the sample as an estimate of the population's mean. Thus we hold one parameter constant. Say that the mean of the sample was 10; then we assume that the population mean is 10 also and we keep this value constant. With this parameter fixed, can all four scores from our sample vary? The answer is no, because to keep the mean constant only three values are free to vary. For example, if the values in the sample were 8, 9, 11, 12 (mean = 10) and we changed three of these values to 7, 15 and 8, then the final value *must* be 10 to keep the mean constant. Therefore, if we hold one parameter constant then the degrees of freedom must be one less than the sample size. This fact explains why when we use a sample to estimate the standard deviation of a population, we have to divide the sums of squares by $N - 1$ rather than N alone.

$$\text{variance } (s^2) = \frac{\text{SS}}{N-1} = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{5.20}{4} = 1.3 \tag{2.1}$$

The variance is, therefore, the average error between the mean and the observations made (and so is a measure of how well the model fits the actual data). There is one problem with the variance as a measure: it gives us a measure in units squared (because we squared each error in the calculation). In our example we would have to say that the average error in our data (the variance) was 1.3 friends squared. It makes little enough sense to talk about 1.3 friends, but it makes even less to talk about friends squared! For this reason, we often take the square root of the variance (which ensures that the measure of average error is in the same units as the original measure). This measure is known as the standard deviation and is simply the square root of the variance. In this example the **standard deviation** is:

$$\begin{aligned} s &= \sqrt{\frac{\sum(x_i - \bar{x})^2}{N-1}} \\ &= \sqrt{1.3} \\ &= 1.14 \end{aligned} \tag{2.2}$$

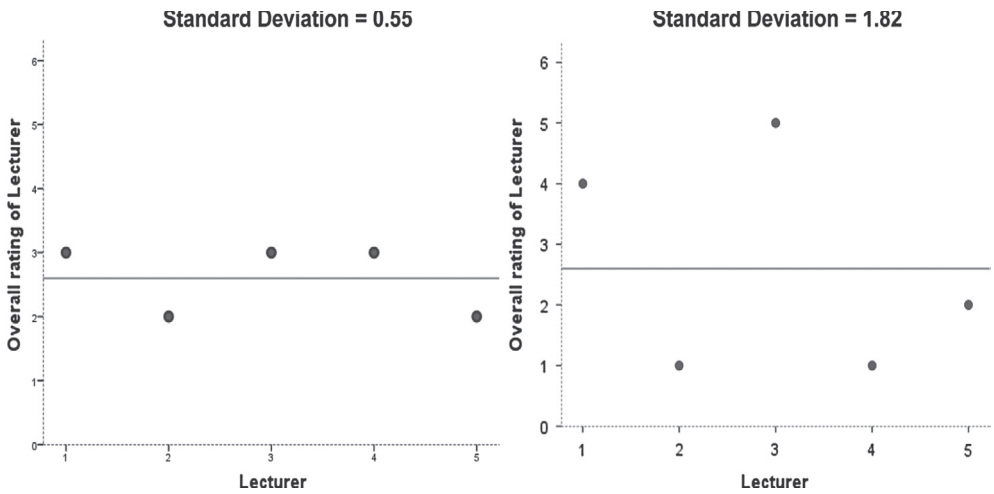


FIGURE 2.5 Graphs illustrating data that have the same mean but different standard deviations

The sum of squares, variance and standard deviation are all, therefore, measures of the 'fit' (i.e. how well the mean represents the data). Small standard deviations (relative to the value of the mean itself) indicate that data points are close to the mean. A large standard deviation (relative to the mean) indicates that the data points are distant from the mean (i.e. the mean is not an accurate representation of the data). A standard deviation of 0 would mean that all of the scores were the same. Figure 2.5 shows the overall ratings (on a 5-point scale) of two lecturers after each of five different lectures. Both lecturers had an average rating of 2.6 out of 5 across the lectures. However, the first lecturer had a standard deviation of 0.55 (relatively small compared to the mean). It should be clear from the graph that ratings for this lecturer were consistently close to the mean rating. There was a small fluctuation, but generally his lectures did not vary in popularity. As such, the mean is an accurate representation of his ratings. The mean is a good fit to the data. The second lecturer, however, had a standard deviation of 1.82 (relatively high compared to the mean). The ratings for this lecturer are clearly more spread from the mean; that is, for some lectures he received very high ratings, and for others his ratings were appalling. Therefore, the mean is not such an accurate representation of his performance because there was a lot of variability in the popularity of his lectures. The mean is a poor fit to the data. This illustration should hopefully make clear why the standard deviation is a measure of how well the mean represents the data.

SELF-TEST In section 1.7.2.2 we came across some data about the number of friends that 11 people had on Facebook (22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252). We calculated the mean for these data as 96.64. Now calculate the sums of squares, variance and standard deviation.

SELF-TEST Calculate these values again but excluding the extreme score (252).

2.4.3. Expressing the mean as a model ②

The discussion of means, sums of squares and variance may seem a side track from the initial point about fitting statistical models, but it's not: the mean is a simple statistical model

JANE SUPERBRAIN 2.3

The standard deviation and the shape of the distribution ①

As well as telling us about the accuracy of the mean as a model of our data set, the variance and standard deviation also tell us about the shape of the distribution of scores. As such, they are measures of dispersion like those we encountered in section 1.7.3. If the mean represents the data well then most of the scores will cluster close to the mean and the resulting standard deviation is small relative to the mean. When the mean is a worse representation of the data, the scores cluster more widely around the mean (think back to Figure 2.5) and the standard deviation is larger. Figure 2.6 shows two distributions that have the same mean (50) but different standard deviations. One has a large standard deviation relative to the mean ($SD = 25$) and this results in a flatter distribution that is more spread out, whereas the other has a small standard deviation relative to the mean ($SD = 15$) resulting in a more pointy distribution in which scores close to the mean are very frequent but scores further from the mean become increasingly infrequent. The main message is that as the standard deviation gets larger, the distribution gets fatter. This can make distributions look platykurtic or leptokurtic when, in fact, they are not.

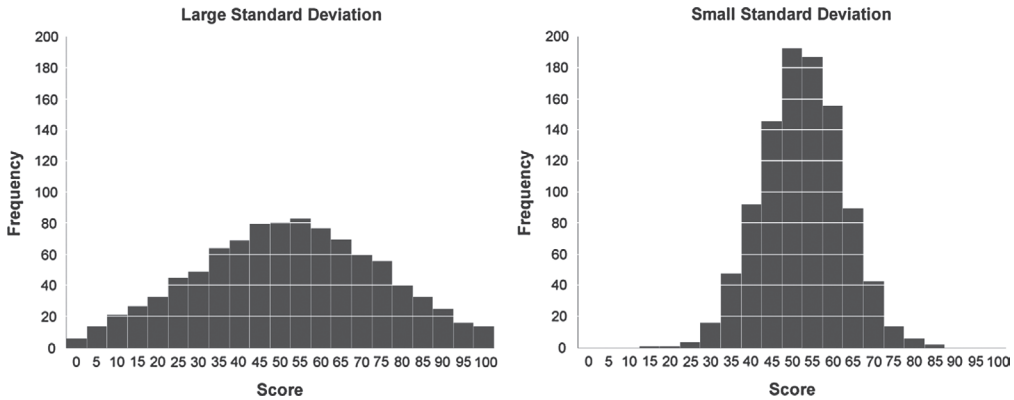


FIGURE 2.6 Two distributions with the same mean, but large and small standard deviations

that can be fitted to data. What do I mean by this? Well, everything in statistics essentially boils down to one equation:

$$\text{outcome}_i = (\text{model}) + \text{error}_i \tag{2.3}$$

This just means that the data we observe can be predicted from the model we choose to fit to the data plus some amount of error. When I say that the mean is a simple statistical model, then all I mean is that we can replace the word 'model' with the word 'mean' in that equation. If we return to our example involving the number of friends that statistics lecturers have and look at lecturer 1, for example, we observed that they had one friend and the mean of all lecturers was 2.6. So, the equation becomes:

$$\text{outcome}_{\text{lecturer1}} = \overline{X} + \varepsilon_{\text{lecturer1}}$$
$$1 = 2.6 + \varepsilon_{\text{lecturer1}}$$

From this we can work out that the error is 1 – 2.6, or –1.6. If we replace this value in the equation we get 1 = 2.6 – 1.6 or 1 = 1. Although it probably seems like I'm stating the obvious, it is worth bearing this general equation in mind throughout this book because if you do you'll discover that most things ultimately boil down to this one simple idea!

Likewise, the variance and standard deviation illustrate another fundamental concept: how the goodness of fit of a model can be measured. If we're looking at how well a model fits the data (in this case our model is the mean) then we generally look at deviation from the model, we look at the sum of squared error, and in general terms we can write this as:

$$\text{deviation} = \sum(\text{observed} - \text{model})^2 \tag{2.4}$$

Put another way, we assess models by comparing the data we observe to the model we've fitted to the data, and then square these differences. Again, you'll come across this fundamental idea time and time again throughout this book.

2.5. Going beyond the data ①

Using the example of the mean, we have looked at how we can fit a statistical model to a set of observations to summarize those data. It's one thing to summarize the data that you have actually collected but usually we want to go beyond our data and say something general about the world (remember in Chapter 1 that I talked about how good theories should say something about the world). It is one thing to be able to say that people in our sample responded well to medication, or that a sample of high-street stores in Brighton had increased profits leading up to Christmas, but it's more useful to be able to say, based on our sample, that all people will respond to medication, or that all high-street stores in the UK will show increased profits. To begin to understand how we can make these general inferences from a sample of data we can first look not at whether our model is a good fit to the sample from which it came, but whether it is a good fit to the **population** from which the sample came.

2.5.1. The standard error ①

We've seen that the standard deviation tells us something about how well the mean represents the sample data, but I mentioned earlier on that usually we collect data from samples because we don't have access to the entire population. If you take several samples from a population, then these samples will differ slightly; therefore, it's also important to know how well a particular sample represents the population. This is where we use the **standard error**. Many students get confused about the difference between the standard deviation and the standard error (usually because the difference is never explained clearly). However, the standard error is an important concept to grasp, so I'll do my best to explain it to you.

We have already learnt that social scientists use samples as a way of estimating the behaviour in a population. Imagine that we were interested in the ratings of all lecturers (so lecturers in general were the population). We could take a sample from this population. When someone takes a sample from a population, they are taking one of many possible

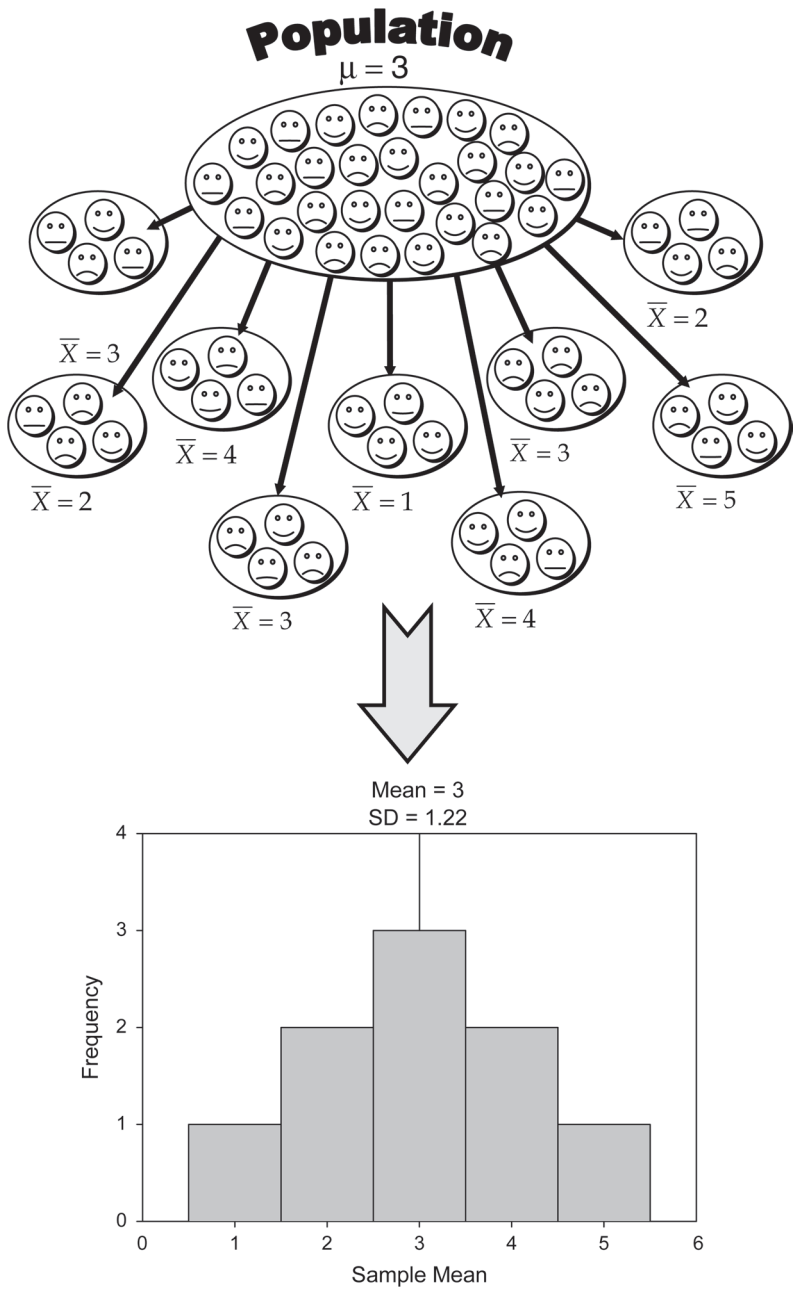


FIGURE 2.7 Illustration of the standard error (see text for details)

samples. If we were to take several samples from the same population, then each sample has its own mean, and some of these sample means will be different.

Figure 2.7 illustrates the process of taking samples from a population. Imagine that we could get ratings of all lecturers on the planet and that, on average, the rating is 3 (this is the *population mean*, μ). Of course, we can't collect ratings of all lecturers, so we use a sample. For each of these samples we can calculate the average, or *sample mean*. Let's imagine we took nine different samples (as in the diagram); you can see that some of the samples have the same mean as the population but some have different means: the first sample of lecturers were rated, on average, as 3, but the second sample were, on average,

rated as only 2. This illustrates **sampling variation**: that is, samples will vary because they contain different members of the population; a sample that by chance includes some very good lecturers will have a higher average than a sample that, by chance, includes some awful lecturers! We can actually plot the sample means as a frequency distribution, or histogram,[4] just like I have done in the diagram. This distribution shows that there were three samples that had a mean of 3, means of 2 and 4 occurred in two samples each, and means of 1 and 5 occurred in only one sample each. The end result is a nice symmetrical distribution known as a **sampling distribution**. A sampling distribution is simply the frequency distribution of sample means from the same population. In theory you need to imagine that we're taking hundreds or thousands of samples to construct a sampling distribution, but I'm just using nine to keep the diagram simple.[5] The sampling distribution tells us about the behaviour of samples from the population, and you'll notice that it is centred at the same value as the mean of the population (i.e. 3). This means that if we took the average of all sample means we'd get the value of the population mean. Now, if the average of the sample means is the same value as the population mean, then if we knew the accuracy of that average we'd know something about how likely it is that a given sample is representative of the population. So how do we determine the accuracy of the population mean?

Think back to the discussion of the standard deviation. We used the standard deviation as a measure of how representative the mean was of the observed data. Small standard deviations represented a scenario in which most data points were close to the mean, a large standard deviation represented a situation in which data points were widely spread from the mean. If you were to calculate the standard deviation between *sample means* then this too would give you a measure of how much variability there was between the means of different samples. The standard deviation of sample means is known as the **standard error of the mean (SE)**. Therefore, the standard error could be calculated by taking the difference between each sample mean and the overall mean, squaring these differences, adding them up, and then dividing by the number of samples. Finally, the square root of this value would need to be taken to get the standard deviation of sample means, the standard error.

Of course, in reality we cannot collect hundreds of samples and so we rely on approximations of the standard error. Luckily for us some exceptionally clever statisticians have demonstrated that as samples get large (usually defined as greater than 30), the sampling distribution has a normal distribution with a mean equal to the population mean, and a standard deviation of:

$$\sigma_{\overline{X}} = \frac{s}{\sqrt{N}} \tag{2.5}$$

This is known as the **central limit theorem** and it is useful in this context because it means that if our sample is large we can use the above equation to approximate the standard error (because, remember, it is the standard deviation of the sampling distribution).[6] When the sample is relatively small (fewer than 30) the sampling distribution has a different shape, known as a *t*-distribution, which we'll come back to later.

[4] This is just a graph of each sample mean plotted against the number of samples that have that mean – see section 1.7.1 for more details.

[5] It's worth pointing out that I'm talking hypothetically. We don't need to *actually* collect these samples because clever statisticians have worked out what these sampling distributions would look like and how they behave.

[6] In fact it should be the *population* standard deviation (σ) that is divided by the square root of the sample size; however, for large samples this is a reasonable approximation.

CRAMMING SAM'S TIPS **The standard error**

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

2.5.2. Confidence intervals ②

2.5.2.1. Calculating confidence intervals ②

Remember that usually we're interested in using the sample mean as an estimate of the value in the population. We've just seen that different samples will give rise to different values of the mean, and we can use the standard error to get some idea of the extent to which sample means differ. A different approach to assessing the accuracy of the sample mean as an estimate of the mean in the population is to calculate boundaries within which we believe the true value of the mean will fall. Such boundaries are called **confidence intervals.** The basic idea behind confidence intervals is to construct a range of values within which we think the population value falls.

Let's imagine an example: Domjan, Blesbois, and Williams (1998) examined the learnt release of sperm in Japanese quail. The basic idea is that if a quail is allowed to copulate with a female quail in a certain context (an experimental chamber) then this context will serve as a cue to copulation and this in turn will affect semen release (although during the test phase the poor quail were tricked into copulating with a terry cloth with an embalmed female quail head stuck on top).[7] Anyway, if we look at the mean amount of sperm released in the experimental chamber, there is a true mean (the mean in the population); let's imagine it's 15 million sperm. Now, in our actual sample, we might find the mean amount of sperm released was 17 million. Because we don't know the true mean, we don't really know whether our sample value of 17 million is a good or bad estimate of this value. What we can do instead is use an interval estimate: we use our sample value as the mid-point, but set a lower and upper limit as well. So, we might say, we think the true value of the mean sperm release is somewhere between 12 million and 22 million spermatozoa (note that 17 million falls exactly between these values). Of course, in this case the true value (15 million) does falls within these limits. However, what if we'd set smaller limits, what if we'd said we think the true value falls between 16 and 18 million (again, note that 17 million is in the middle)? In this case the interval does not contain the true value of the mean. Let's now imagine that you were particularly fixated with Japanese quail sperm, and you repeated the experiment 50 times using different samples. Each time you did the experiment again you constructed an interval around the sample mean as I've just described. Figure 2.8 shows this scenario: the circles represent the mean for each sample with the lines sticking out of them representing the intervals for these means. The true value of the mean (the mean in the population) is 15 million and is shown by a vertical line. The first thing to note is that most of the sample means are different from

[7] This may seem a bit sick, but the male quails didn't appear to mind too much, which probably tells us all we need to know about male mating behaviour.

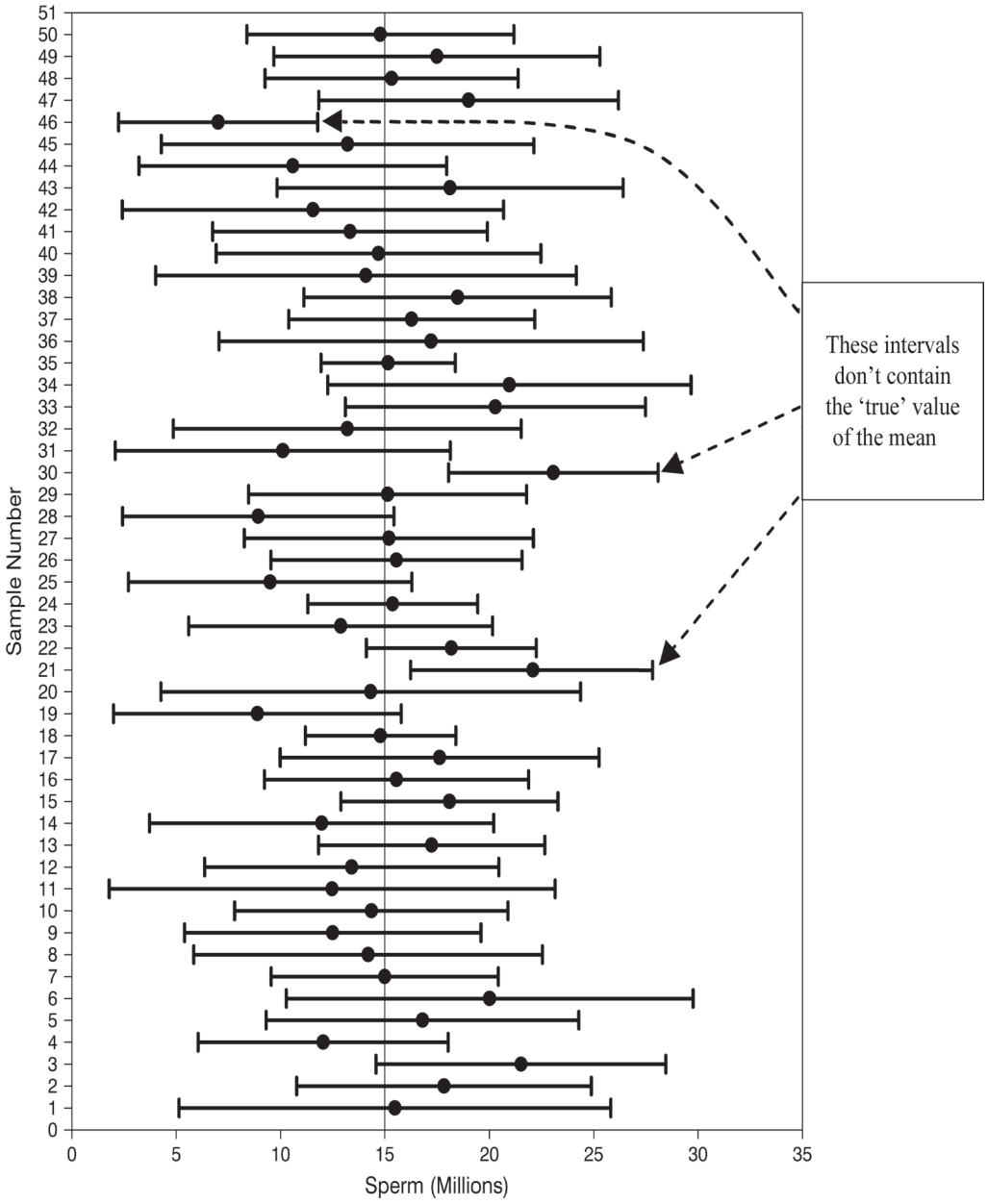


FIGURE 2.8 The confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 50 different samples (vertical axis)

the true mean (this is because of sampling variation as described in the previous section). Second, although most of the intervals do contain the true mean (they cross the vertical line, meaning that the value of 15 million spermatozoa falls somewhere between the lower and upper boundaries), a few do not.

Up until now I've avoided the issue of how we might calculate the intervals. The crucial thing with confidence intervals is to construct them in such a way that they tell us something useful. Therefore, we calculate them so that they have certain properties: in particular they tell us the likelihood that they contain the true value of the thing we're trying to estimate (in this case, the mean).

Typically we look at 95% confidence intervals, and sometimes 99% confidence intervals, but they all have a similar interpretation: they are limits constructed such that for

a certain percentage of the time (be that 95% or 99%) the true value of the population mean will fall within these limits. So, when you see a 95% confidence interval for a mean, think of it like this: if we'd collected 100 samples, calculated the mean and then calculated a confidence interval for that mean (a bit like in Figure 2.8) then for 95 of these samples, the confidence intervals we constructed would contain the true value of the mean in the population.

To calculate the confidence interval, we need to know the limits within which 95% of means will fall. How do we calculate these limits? Remember back in section 1.7.4 that I said that 1.96 was an important value of z (a score from a normal distribution with a mean of 0 and standard deviation of 1) because 95% of z-scores fall between –1.96 and 1.96. This means that if our sample means were normally distributed with a mean of 0 and a standard error of 1, then the limits of our confidence interval would be –1.96 and +1.96. Luckily we know from the central limit theorem that in large samples (above about 30) the sampling distribution will be normally distributed (see section 2.5.1). It's a pity then that our mean and standard deviation are unlikely to be 0 and 1; except not really because, as you might remember, we can convert scores so that they do have a mean of 0 and standard deviation of 1 (z-scores) using equation (1.2):

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

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

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

We rearrange these equations to discover the value of X:

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

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

$$\text{lower boundary of confidence interval} = \overline{X} - (1.96 \times \text{SE})$$
$$\text{upper boundary of confidence interval} = \overline{X} + (1.96 \times \text{SE})$$

As such, the mean is always in the centre of the confidence interval. If the mean represents the true mean well, then the confidence interval of that mean should be small. We know that 95% of confidence intervals contain the true mean, so we can assume this confidence interval contains the true mean; therefore, if the interval is small, the sample mean must be very close to the true mean. Conversely, if the confidence interval is very wide then the sample mean could be very different from the true mean, indicating that it is a bad representation of the population You'll find that confidence intervals will come up time and time again throughout this book.

2.5.2.2. Calculating other confidence intervals ②

The example above shows how to compute a 95% confidence interval (the most common type). However, we sometimes want to calculate other types of confidence interval such as a 99% or 90% interval. The 1.96 and –1.96 in the equations above are the limits within which 95% of *z*-scores occur. Therefore, if we wanted a 99% confidence interval we could use the values within which 99% of *z*-scores occur (–2.58 and 2.58). In general then, we could say that confidence intervals are calculated as:

$$\text{lower boundary of confidence interval} = \overline{X} - \left(z_{\frac{1-p}{2}} \times \text{SE}\right)$$

$$\text{upper boundary of confidence interval} = \overline{X} + \left(z_{\frac{1-p}{2}} \times \text{SE}\right)$$

in which p is the probability value for the confidence interval. So, if you want a 95% confidence interval, then you want the value of z for (1 – .95)/2 = .025. Look this up in the 'smaller portion' column of the table of the standard normal distribution (see the Appendix) and you'll find that z is 1.96. For a 99% confidence interval we want z for (1 – .99)/2 = .005, which from the table is 2.58. For a 90% confidence interval we want z for (1 – .90)/2 = .05, which from the table is 1.65. These values of z are multiplied by the standard error (as above) to calculate the confidence interval. Using these general principles we could work out a confidence interval for any level of probability that takes our fancy.

2.5.2.3. Calculating confidence intervals in small samples ②

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

$$\text{lower boundary of confidence interval} = \overline{X} - (t_{n-1} \times \text{SE})$$

$$\text{upper boundary of confidence interval} = \overline{X} + (t_{n-1} \times \text{SE})$$

The $n - 1$ in the equations is the degrees of freedom (see Jane Superbrain Box 2.3) and tells us which of the *t*-distributions to use. For a 95% confidence interval we find the value of t for a two-tailed test with probability of .05, for the appropriate degrees of freedom.

SELF-TEST In section 1.7.2.2 we came across some data about the number of friends that 11 people had on Facebook. We calculated the mean for these data as 96.64 and standard deviation as 61.27. Calculate a 95% confidence interval for this mean.

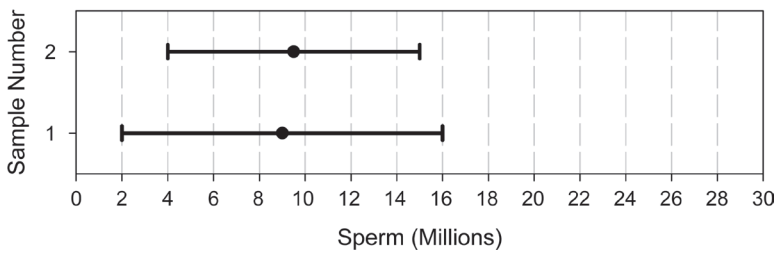
SELF-TEST Recalculate the confidence interval assuming that the sample size was 56

2.5.2.4. Showing confidence intervals visually ②

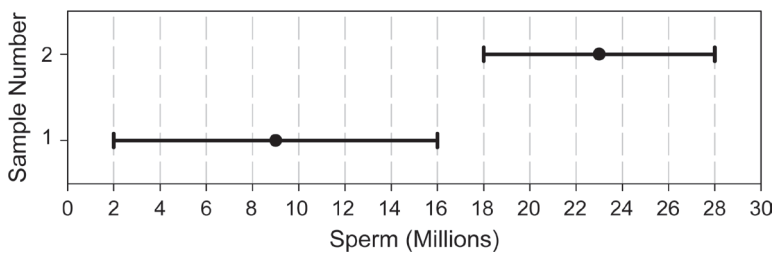
Confidence intervals provide us with very important information about the mean, and, therefore, you often see them displayed on graphs. (We will discover more about how to create these graphs in Chapter 4.) The confidence interval is usually displayed using something called an error bar, which just looks like the letter 'I'. An error bar can represent the standard deviation, or the standard error, but more often than not it shows the 95% confidence interval of the mean. So, often when you see a graph showing the mean, perhaps displayed as a bar (Section 4.6) or a symbol (section 4.7), it is often accompanied by this funny I-shaped bar. Why is it useful to see the confidence interval visually?

We have seen that the 95% confidence interval is an interval constructed such that in 95% of samples the true value of the population mean will fall within its limits. We know that it is possible that any two samples could have slightly different means (and the standard error tells us a little about how different we can expect sample means to be). Now, the confidence interval tells us the limits within which the population mean is likely to fall (the size of the confidence interval will depend on the size of the standard error). By comparing the confidence intervals of different means we can start to get some idea about whether the means came from the same population or different populations.

Taking our previous example of quail sperm, imagine we had a sample of quail and the mean sperm release had been 9 million sperm with a confidence interval of 2 to 16. Therefore, we know that the population mean is probably between 2 and 16 million sperm. What if we now took a second sample of quail and found the confidence interval ranged from 4 to 15? This interval overlaps a lot with our first sample:



The fact that the confidence intervals overlap in this way tells us that these means could plausibly come from the same population: in both cases the intervals are likely to contain the true value of the mean (because they are constructed such that in 95% of studies they will), and both intervals overlap considerably, so they contain many similar values. What if the confidence interval for our second sample ranges from 18 to 28? If we compared this to our first sample we'd get:



Now, these confidence intervals don't overlap at all. So, one confidence interval, which is likely to contain the population mean, tells us that the population mean is somewhere

between 2 and 16 million, whereas the other confidence interval, which is also likely to contain the population mean, tells us that the population mean is somewhere between 18 and 28. This suggests that either our confidence intervals both do contain the population mean, but they come from different populations (and, therefore, so do our samples), or both samples come from the same population but one of the confidence intervals doesn't contain the population mean. If we've used 95% confidence intervals then we know that the second possibility is unlikely (this happens only 5 times in 100 or 5% of the time), so the first explanation is more plausible.

OK, I can hear you all thinking 'so what if the samples come from a different population?' Well, it has a very important implication in experimental research. When we do an experiment, we introduce some form of manipulation between two or more conditions (see section 1.6.2). If we have taken two random samples of people, and we have tested them on some measure (e.g. fear of statistics textbooks), then we expect these people to belong to the same population. If their sample means are so different as to suggest that, in fact, they come from different populations, why might this be? The answer is that our experimental manipulation has induced a difference between the samples.

To reiterate, when an experimental manipulation is successful, we expect to find that our samples have come from different populations. If the manipulation is unsuccessful, then we expect to find that the samples came from the same population (e.g. the sample means should be fairly similar). Now, the 95% confidence interval tells us something about the likely value of the population mean. If we take samples from two populations, then we expect the confidence intervals to be different (in fact, to be sure that the samples were from different populations we would not expect the two confidence intervals to overlap). If we take two samples from the same population, then we expect, if our measure is reliable, the confidence intervals to be very similar (i.e. they should overlap completely with each other).

This is why error bars showing 95% confidence intervals are so useful on graphs, because if the bars of any two means do not overlap then we can infer that these means are from different populations – they are significantly different.

CRAMMING SAM'S TIPS **Confidence intervals**

A confidence interval for the mean is a range of scores constructed such that the population mean will fall within this range in 95% of samples.

The confidence interval is not an interval within which we are 95% confident that the population mean will fall.

2.6. Using statistical models to test research questions ①

In Chapter 1 we saw that research was a five-stage process:

1. Generate a research question through an initial observation (hopefully backed up by some data).
2. Generate a theory to explain your initial observation.
3. Generate hypotheses: break your theory down into a set of testable predictions.

4 Collect data to test the theory: decide on what variables you need to measure to test your predictions and how best to measure or manipulate those variables.

5 Analyse the data: fit a statistical model to the data – this model will test your original predictions. Assess this model to see whether or not it supports your initial predictions.

This chapter has shown that we can use a sample of data to estimate what's happening in a larger population to which we don't have access. We have also seen (using the mean as an example) that we can fit a statistical model to a sample of data and assess how well it fits. However, we have yet to see how fitting models like these can help us to test our research predictions. How do statistical models help us to test complex hypotheses such as 'is there a relationship between the amount of gibberish that people speak and the amount of vodka jelly they've eaten?', or 'is the mean amount of chocolate I eat higher when I'm writing statistics books than when I'm not?' We've seen in section 1.7.5 that hypotheses can be broken down into a null hypothesis and an alternative hypothesis.

SELF-TEST What are the null and alternative hypotheses for the following questions:

- ✓ 'Is there a relationship between the amount of gibberish that people speak and the amount of vodka jelly they've eaten?'
- ✓ 'Is the mean amount of chocolate eaten higher when writing statistics books than when not?'

Most of this book deals with *inferential statistics*, which tell us whether the alternative hypothesis is likely to be true – they help us to confirm or reject our predictions. Crudely put, we fit a statistical model to our data that represents the alternative hypothesis and see how well it fits (in terms of the variance it explains). If it fits the data well (i.e. explains a lot of the variation in scores) then we assume our initial prediction is true: we gain confidence in the alternative hypothesis. Of course, we can never be completely sure that either hypothesis is correct, and so we calculate the probability that our model would fit if there were no effect in the population (i.e. the null hypothesis is true). As this probability decreases, we gain greater confidence that the alternative hypothesis is actually correct and that the null hypothesis can be rejected. This works provided we make our predictions before we collect the data (see Jane Superbrain Box 2.4).

To illustrate this idea of whether a hypothesis is likely, Fisher (1925/1991) (Figure 2.9) describes an experiment designed to test a claim by a woman that she could determine, by tasting a cup of tea, whether the milk or the tea was added first to the cup. Fisher thought that he should give the woman some cups of tea, some of which had the milk added first and some of which had the milk added last, and see whether she could correctly identify them. The woman would know that there are an equal number of cups in which milk was added first or last but wouldn't know in which order the cups were placed. If we take the simplest situation in which there are only two cups then the woman has a 50% chance of guessing correctly. If she did guess correctly we wouldn't be that confident in concluding that she can tell the difference between cups in which the milk was added first from those in which it was added last, because even by guessing she would be correct half of the time. However, what about if we complicated things by having six cups? There are 20 orders in which these cups can be arranged and the woman would guess the correct order only 1 time in 20 (or 5% of the time). If she got the correct order we would be much more

JANE SUPERBRAIN 2.4

Cheating in research ①

The process I describe in this chapter works only if you generate your hypotheses and decide on your criteria for whether an effect is significant before collecting the data. Imagine I wanted to place a bet on who would win the Rugby World Cup. Being an Englishman, I might want to bet on England to win the tournament. To do this I'd: (1) place my bet, choosing my team (England) and odds available at the betting shop (e.g. 6/4); (2) see which team wins the tournament; (3) collect my winnings (if England do the decent thing and actually win).

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

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

Every time you change your hypothesis or the details of your analysis you appear to increase the chance of finding a significant result, but in fact you are making it more and more likely that you will publish results that other researchers can't reproduce (which is very embarrassing!). If, however, you follow the rules carefully and do your significance testing at the 5% level you at least know that in the long run at most only 1 result out of every 20 will risk this public humiliation.

(With thanks to David Hitchin for this box, and with apologies to him for turning it into a rugby example!)

confident that she could genuinely tell the difference (and bow down in awe of her finely tuned palette). If you'd like to know more about Fisher and his tea-tasting antics see David Salsburg's excellent book *The lady tasting tea* (Salsburg, 2002). For our purposes the take-home point is that only when there was a very small probability that the woman could complete the tea task by luck alone would we conclude that she had genuine skill in detecting whether milk was poured into a cup before or after the tea.

It's no coincidence that I chose the example of six cups above (where the tea-taster had a 5% chance of getting the task right by guessing), because Fisher suggested that 95% is a useful threshold for confidence: only when we are 95% certain that a result is genuine (i.e. not a chance finding) should we accept it as being true.[8] The opposite way to look at this is to say that if there is only a 5% chance (a probability of .05) of something occurring by chance then we can accept that it is a genuine effect: we say it is a **statistically significant** finding (see Jane Superbrain Box 2.5 to find out how the criterion of .05 became popular!).

[8] Of course, in reality, it might not be true – we're just prepared to believe that it is!

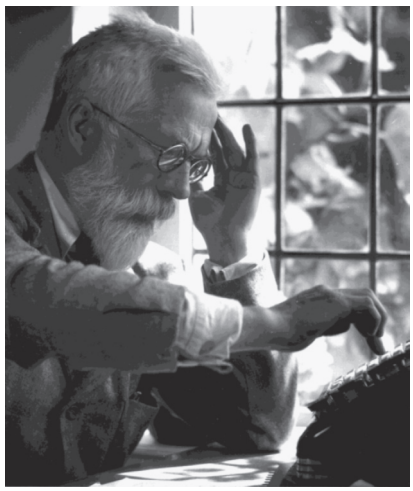


FIGURE 2.9 Sir Ronald A. Fisher, the cleverest person ever ($p < .0001$)

JANE SUPERBRAIN 2.5

Why do we use .05? ①

This criterion of 95% confidence, or a .05 probability, forms the basis of modern statistics and yet there is very little justification for it. How it arose is a complicated mystery to unravel. The significance testing that we use today is a blend of Fisher's idea of using the probability value p as an index of the weight of evidence against a null hypothesis, and Jerzy Neyman and Egon Pearson's idea of testing a null hypothesis *against* an alternative hypothesis. Fisher objected to Neyman's use of an alternative hypothesis (among other things), and Neyman objected to Fisher's exact probability approach (Berger, 2003; Lehmann, 1993). The confusion arising from both parties' hostility to each other's ideas led scientists to create a sort of bastard child of both approaches.

This doesn't answer the question of why we use .05. Well, it probably comes down to the fact that back in the days before computers, scientists had to compare their test statistics against published tables of 'critical values' (they did not have SAS to calculate exact probabilities for them). These critical values had to be calculated by exceptionally clever people like Fisher. In his incredibly influential textbook *Statistical methods for research workers* (Fisher, 1925)[9] Fisher produced tables of these critical values, but to save space produced tables for particular probability values (.05, .02 and .01). The impact of this book should not be underestimated (to get some idea of its influence 25 years after publication see Mather, 1951; Yates, 1951) and these tables were very frequently used – even Neyman and Pearson admitted the influence that these tables had on them (Lehmann, 1993). This disastrous combination of researchers confused about the Fisher and Neyman–Pearson approaches and the availability of critical values for only certain levels of probability led to a trend to report test statistics as being significant at the now infamous $p < .05$ and $p < .01$ (because critical values were readily available at these probabilities).

However, Fisher acknowledged that the dogmatic use of a fixed level of significance was silly: 'no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas' (Fisher, 1956).

The use of effect sizes (section 2.6.4) strikes a balance between using arbitrary cut-off points such as $p < .05$ and assessing whether an effect is meaningful within the research context. The fact that we still worship at the shrine of $p < .05$ and that research papers are more likely to be published if they contain significant results does make me wonder about a parallel universe where Fisher had woken up in a $p < .10$ kind of mood. My filing cabinet full of research with p just bigger than .05 are published and I am Vice-Chancellor of my university (although, if this were true, the parallel universe version of my university would be in utter chaos, but it would have a campus full of cats).

[9] You can read this online at http://psychclassics.yorku.ca/Fisher/Methods/.

2.6.1. Test statistics ①

We have seen that we can fit statistical models to data that represent the hypotheses that we want to test. Also, we have discovered that we can use probability to see whether scores are likely to have happened by chance (section 1.7.4). If we combine these two ideas then we can test whether our statistical models (and therefore our hypotheses) are significant fits of the data we collected. To do this we need to return to the concepts of systematic and unsystematic variation that we encountered in section 1.6.2.2. Systematic variation is variation that can be explained by the model that we've fitted to the data (and, therefore, due to the hypothesis that we're testing). Unsystematic variation is variation that cannot be explained by the model that we've fitted. In other words, it is error, or variation not attributable to the effect we're investigating. The simplest way, therefore, to test whether the model fits the data, or whether our hypothesis is a good explanation of the data we have observed, is to compare the systematic variation against the unsystematic variation. In doing so we compare how good the model/hypothesis is at explaining the data against how bad it is (the error):

$$\text{test statistic} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

This ratio of systematic to unsystematic variance or effect to error is a **test statistic**, and you'll discover later in the book there are lots of them: t, F and χ^2 to name only three. The exact form of this equation changes depending on which test statistic you're calculating, but the important thing to remember is that they all, crudely speaking, represent the same thing: the amount of variance explained by the model we've fitted to the data compared to the variance that can't be explained by the model (see Chapters 7 and 9 in particular for a more detailed explanation). The reason why this ratio is so useful is intuitive really: if our model is good then we'd expect it to be able to explain more variance than it can't explain. In this case, the test statistic will be greater than 1 (but not necessarily significant).

A test statistic is a statistic that has known properties; specifically we know how frequently different values of this statistic occur. By knowing this, we can calculate the probability of obtaining a particular value (just as we could estimate the probability of getting a score of a certain size from a frequency distribution in section 1.7.4). This allows us to establish how likely it would be that we would get a test statistic of a certain size if there were no effect (i.e. the null hypothesis were true). Field and Hole (2003) use the analogy of the age at which people die. Past data have told us the distribution of the age of death. For example, we know that on average men die at about 75 years old, and that this distribution is top heavy; that is, most people die above the age of about 50 and it's fairly unusual to die in your twenties. So, the frequencies of the age of demise at older ages are very high but are lower at younger ages. From these data, it would be possible to calculate the probability of someone dying at a certain age. If we randomly picked someone and asked them their age, and it was 53, we could tell them how likely it is that they will die before their next birthday (at which point they'd probably punch us!). Also, if we met a man of 110, we could calculate how probable it was that he would have lived that long (it would be a very small probability because most people die before they reach that age). The way we use test statistics is rather similar: we know their distributions and this allows us, once we've calculated the test statistic, to discover the probability of having found a value as big as we have. So, if we calculated a test statistic and its value was 110 (rather like our old man) we can then calculate the probability of obtaining a value that large. The more variation our model explains (compared to the variance it can't explain), the

bigger the test statistic will be, and the more unlikely it is to occur by chance (like our 110 year old man). So, as test statistics get bigger, the probability of them occurring becomes smaller. When this probability falls below .05 (Fisher's criterion), we accept this as giving us enough confidence to assume that the test statistic is as large as it is because our model explains a sufficient amount of variation to reflect what's genuinely happening in the real world (the population). The test statistic is said to be *significant* (see Jane Superbrain Box 2.6 for a discussion of what statistically significant actually means). Given that the statistical model that we fit to the data reflects the hypothesis that we set out to test, then a significant test statistic tells us that the model would be unlikely to fit this well if the there was no effect in the population (i.e. the null hypothesis was true). Therefore, we can reject our null hypothesis and gain confidence that the alternative hypothesis is true (but, remember, we don't accept it – see section 1.7.5).

JANE SUPERBRAIN 2.6

What we can and can't conclude from a significant test statistic ②

- **The importance of an effect:** We've seen already that the basic idea behind hypothesis testing involves us generating an experimental hypothesis and a null hypothesis, fitting a statistical model to the data, and assessing that model with a test statistic. If the probability of obtaining the value of our test statistic by chance is less than .05 then we generally accept the experimental hypothesis as true: there is an effect in the population. Normally we say 'there is a *significant* effect of …'. However, don't be fooled by that word 'significant', because even if the probability of our effect being a chance result is small (less than .05) it doesn't necessarily follow that the effect is important. Very small and unimportant effects can turn out to be statistically significant just because huge numbers of people have been used in the experiment (see Field & Hole, 2003: 74).
- **Non-significant results:** Once you've calculated your test statistic, you calculate the probability of that test statistic occurring by chance; if this probability is greater than .05 you reject your alternative hypothesis. However, this does *not* mean that the null hypothesis is true. Remember that the null hypothesis is that there is no effect in the population. All that a non-significant result tells us is that the effect is not big enough to be anything other than a chance finding – it doesn't tell us that the effect is zero. As Cohen (1990) points out, a non-significant result should never be interpreted (despite the fact that it often is) as 'no difference between means' or 'no relationship between variables'. Cohen also points out that the null hypothesis is *never* true because we know from sampling distributions (see section 2.5.1) that two random samples will have slightly different means, and even though these differences can be very small (e.g. one mean might be 10 and another might be 10.00001) they are nevertheless different. In fact, even such a small difference would be deemed as statistically significant if a big enough sample were used. So, significance testing can never tell us that the null hypothesis is true, because it never is!
- **Significant results:** OK, we may not be able to accept the null hypothesis as being true, but we can at least conclude that it is false when our results are significant, right? Wrong! A significant test statistic is based on probabilistic reasoning, which severely limits what we can conclude. Again, Cohen (1994), who was an incredibly lucid writer on statistics, points out that formal reasoning relies on an initial statement of fact followed by a statement about the current state of affairs, and an inferred conclusion. This syllogism illustrates what I mean:

  - If a man has no arms then he can't play guitar:
  - This man plays guitar.
  - Therefore, this man has arms.

  The syllogism starts with a statement of fact that allows the end conclusion to be reached because you can deny the man has no arms (the antecedent) by

denying that he can't play guitar (the consequent).[10] A comparable version of the null hypothesis is:

- If the null hypothesis is correct, then this test statistic cannot occur:
- This test statistic has occurred.
- Therefore, the null hypothesis is false.

This is all very nice except that the null hypothesis is not represented in this way because it is based on probabilities. Instead it should be stated as follows:

- If the null hypothesis is correct, then this test statistic is highly unlikely:
- This test statistic has occurred.
- Therefore, the null hypothesis is highly unlikely.

If we go back to the guitar example we could get a similar statement:

- If a man plays guitar then he probably doesn't play for Fugazi (this is true because there are thousands of people who play guitar but only two who play guitar in the band Fugazi):
- Guy Picciotto plays for Fugazi:
- Therefore, Guy Picciotto probably doesn't play guitar.

This should hopefully seem completely ridiculous – the conclusion is wrong because Guy Picciotto does play guitar. This illustrates a common fallacy in hypothesis testing. In fact significance testing allows us to say very little about the null hypothesis.

2.6.2. One- and two-tailed tests ①

We saw in section 1.7.5 that hypotheses can be directional (e.g. 'the more someone reads this book, the more they want to kill its author') or non-directional (i.e. 'reading more of this book could increase or decrease the reader's desire to kill its author'). A statistical model that tests a directional hypothesis is called a **one-tailed test**, whereas one testing a non-directional hypothesis is known as a **two-tailed test**.

Imagine we wanted to discover whether reading this book increased or decreased the desire to kill me. We could do this either (experimentally) by taking two groups, one who had read this book and one who hadn't, or (correlationally) by measuring the amount of this book that had been read and the corresponding desire to kill me. If we have no directional hypothesis then there are three possibilities. (1) People who read this book want to kill me more than those who don't so the difference (the mean for those reading the book minus the mean for non-readers) is positive. Correlationally, the more of the book you read, the more you want to kill me – a positive relationship. (2) People who read this book want to kill me less than those who don't so the difference (the mean for those reading the book minus the mean for non-readers) is negative. Correlationally, the more of the book you read, the less you want to kill me – a negative relationship. (3) There is no difference between readers and non-readers in their desire to kill me – the mean for readers minus the mean for non-readers is exactly zero. Correlationally, there is no relationship between reading this book and wanting to kill me. This final option is the null hypothesis. The direction of the test statistic (i.e. whether it is positive or negative) depends on whether the difference is positive or negative. Assuming there is a positive difference or relationship (reading this book makes you want to kill me), then to detect this difference we have to take account of the fact that the mean for readers is bigger than for non-readers (and so derive a positive test statistic). However, if we've predicted incorrectly and actually reading this book makes readers want to kill me less then the test statistic will actually be negative.

[10] Thanks to Philipp Sury for unearthing footage that disproves my point (http://www.parcival.org/2007/05/22/when-syllogisms-fail/).

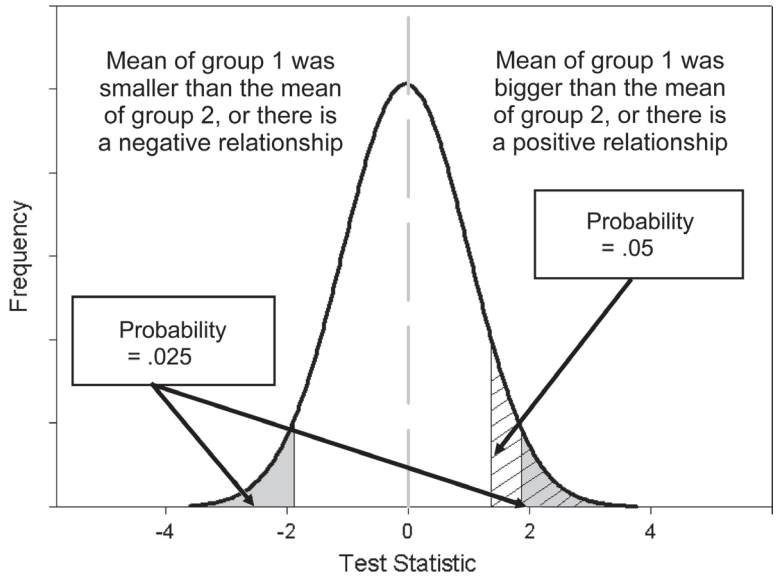


FIGURE 2.10 Diagram to show the difference between one- and two-tailed tests

What are the consequences of this? Well, if at the .05 level we needed to get a test statistic bigger than say 10 and the one we get is actually –12, then we would reject the hypothesis even though a difference does exist. To avoid this we can look at both ends (or tails) of the distribution of possible test statistics. This means we will catch both positive and negative test statistics. However, doing this has a price because to keep our criterion probability of .05 we have to split this probability across the two tails: so we have .025 at the positive end of the distribution and .025 at the negative end. Figure 2.10 shows this situation – the tinted areas are the areas above the test statistic needed at a .025 level of significance. Combine the probabilities (i.e. add the two tinted areas together) at both ends and we get .05, our criterion value. Now if we have made a prediction, then we put all our eggs in one basket and look only at one end of the distribution (either the positive or the negative end depending on the direction of the prediction we make). So, in Figure 2.10, rather than having two small tinted areas at either end of the distribution that show the significant values, we have a bigger area (the lined area) at only one end of the distribution that shows significant values. Consequently, we can just look for the value of the test statistic that would occur by chance with a probability of .05. In Figure 2.10, the lined area is the area above the positive test statistic needed at a .05 level of significance. Note on the graph that the value that begins the area for the .05 level of significance (the lined area) is smaller than the value that begins the area for the .025 level of significance (the tinted area). This means that if we make a specific prediction then we need a smaller test statistic to find a significant result (because we are looking in only one tail of the distribution), but if our prediction happens to be in the wrong direction then we'll miss out on detecting the effect that does exist! In this context it's important to remember what I said in Jane Superbrain Box 2.4: you can't place a bet or change your bet when the tournament is over. If you didn't make a prediction of direction before you collected the data, you are too late to predict the direction and claim the advantages of a one-tailed test.

2.6.3. Type I and Type II errors ①

We have seen that we use test statistics to tell us about the true state of the world (to a certain degree of confidence). Specifically, we're trying to see whether there is an effect in

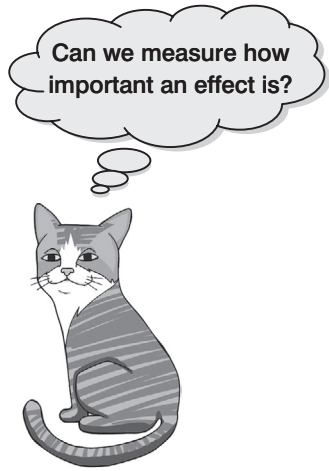
our population. There are two possibilities in the real world: there is, in reality, an effect in the population, or there is, in reality, no effect in the population. We have no way of knowing which of these possibilities is true; however, we can look at test statistics and their associated probability to tell us which of the two is more likely. Obviously, it is important that we're as accurate as possible, which is why Fisher originally said that we should be very conservative and only believe that a result is genuine when we are 95% confident that it is – or when there is only a 5% chance that the results could occur if there was not an effect (the null hypothesis is true). However, even if we're 95% confident there is still a small chance that we get it wrong. In fact there are two mistakes we can make: a Type I and a Type II error. A **Type I error** occurs when we believe that there is a genuine effect in our population, when in fact there isn't. If we use Fisher's criterion then the probability of this error is .05 (or 5%) when there is no effect in the population – this value is known as the **α-level**. Assuming there is no effect in our population, if we replicated our data collection 100 times we could expect that on five occasions we would obtain a test statistic large enough to make us think that there was a genuine effect in the population even though there isn't. The opposite is a **Type II error**, which occurs when we believe that there is no effect in the population when, in reality, there is. This would occur when we obtain a small test statistic (perhaps because there is a lot of natural variation between our samples). In an ideal world, we want the probability of this error to be very small (if there is an effect in the population then it's important that we can detect it). Cohen (1992) suggests that the maximum acceptable probability of a Type II error would be .2 (or 20%) – this is called the **β-level**. That would mean that if we took 100 samples of data from a population in which an effect exists, we would fail to detect that effect in 20 of those samples (so we'd miss 1 in 5 genuine effects).

There is obviously a trade-off between these two errors: if we lower the probability of accepting an effect as genuine (i.e. make α smaller) then we increase the probability that we'll reject an effect that does genuinely exist (because we've been so strict about the level at which we'll accept that an effect is genuine). The exact relationship between the Type I and Type II error is not straightforward because they are based on different assumptions: to make a Type I error there has to be no effect in the population, whereas to make a Type II error the opposite is true (there has to be an effect that we've missed). So, although we know that as the probability of making a Type I error decreases, the probability of making a Type II error increases, the exact nature of the relationship is usually left for the researcher to make an educated guess (Howell, 2006, gives a great explanation of the trade-off between errors).

2.6.4. Effect sizes ②

The framework for testing whether effects are genuine that I've just presented has a few problems, most of which have been briefly explained in Jane Superbrain Box 2.6. The first problem we encountered was knowing how important an effect is: just because a test statistic is significant doesn't mean that the effect it measures is meaningful or important. The solution to this criticism is to measure the size of the effect that we're testing in a standardized way. When we measure the size of an effect (be that an experimental manipulation or the strength of a relationship between variables) it is known as an **effect size**. An effect size is simply an objective and (usually) standardized measure of the magnitude of observed effect. The fact that the measure is standardized just means that we can compare effect sizes across different studies that have measured different variables, or have used different scales of measurement (so an effect size based on speed in milliseconds

could be compared to an effect size based on heart rates). Such is the utility of effect size estimates that the American Psychological Association is now recommending that all psychologists report these effect sizes in the results of any published work. So, it's a habit well worth getting into.



Many measures of effect size have been proposed, the most common of which are Cohen's *d*, Pearson's correlation coefficient *r* (Chapter 6) and the odds ratio (Chapter 18). Many of you will be familiar with the correlation coefficient as a measure of the strength of relationship between two variables (see Chapter 6 if you're not); however, it is also a very versatile measure of the strength of an experimental effect. It's a bit difficult to reconcile how the humble correlation coefficient can also be used in this way; however, this is only because students are typically taught about it within the context of non-experimental research. I don't want to get into it now, but as you read through Chapters 6, 9 and 10 it will (I hope!) become clear what I mean. Personally, I prefer Pearson's correlation coefficient, *r*, as an effect size measure because it is constrained to lie between 0 (no effect) and 1 (a perfect effect).[11] However, there are situations in which *d* may be favoured; for example, when group sizes are very discrepant *r* can be quite biased compared to *d* (McGrath & Meyer, 2006).

Effect sizes are useful because they provide an objective measure of the importance of an effect. So, it doesn't matter what effect you're looking for, what variables have been measured, or how those variables have been measured – we know that a correlation coefficient of 0 means there is no effect, and a value of 1 means that there is a perfect effect. Cohen (1988, 1992) has also made some widely used suggestions about what constitutes a large or small effect:

- **r = .10 (small effect)**: In this case the effect explains 1% of the total variance.
- **r = .30 (medium effect)**: The effect accounts for 9% of the total variance.
- **r = .50 (large effect)**: The effect accounts for 25% of the variance.

It's worth bearing in mind that *r* is not measured on a linear scale so an effect with $r = .6$ isn't twice as big as one with $r = .3$! Although these guidelines can be a useful rule of thumb to assess the importance of an effect (regardless of the significance of the test statistic), it is worth remembering that these 'canned' effect sizes are no substitute for evaluating an effect size within the context of the research domain that it is being used (Baguley, 2004; Lenth, 2001).

A final thing to mention is that when we calculate effect sizes we calculate them for a given sample. When we looked at means in a sample we saw that we used them to draw inferences about the mean of the entire population (which is the value in which we're actually interested). The same is true of effect sizes: the size of the effect in the population is the value in which we're interested, but because we don't have access to this value, we use the effect size in the sample to estimate the likely size of the effect in the population. We can also combine effect sizes from different studies researching the same question to get better estimates of the population effect sizes. This is called **meta-analysis** – see Field (2001, 2005b).

[11] The correlation coefficient can also be negative (but not below –1), which is useful when we're measuring a relationship between two variables because the sign of *r* tells us about the direction of the relationship, but in experimental research the sign of *r* merely reflects the way in which the experimenter coded their groups (see Chapter 6).

2.6.5. Statistical power ②

Effect sizes are an invaluable way to express the importance of a research finding. The effect size in a population is intrinsically linked to three other statistical properties: (1) the sample size on which the sample effect size is based; (2) the probability level at which we will accept an effect as being statistically significant (the α-level); and (3) the ability of a test to detect an effect of that size (known as the statistical **power**, not to be confused with statistical powder, which is an illegal substance that makes you understand statistics better). As such, once we know three of these properties, then we can always calculate the remaining one. It will also depend on whether the test is a one- or two-tailed test (see section 2.6.2). Typically, in psychology we use an α-level of .05 (see earlier) so we know this value already. The power of a test is the probability that a given test will find an effect assuming that one exists in the population. If you think back you might recall that we've already come across the probability of failing to detect an effect when one genuinely exists (β, the probability of a Type II error). It follows that the probability of detecting an effect if one exists must be the opposite of the probability of not detecting that effect (i.e. $1 - \beta$). I've also mentioned that Cohen (1988, 1992) suggests that we would hope to have a .2 probability of failing to detect a genuine effect, and so the corresponding level of power that he recommended was 1 – .2, or .8. We should aim to achieve a power of .8, or an 80% chance of detecting an effect if one genuinely exists. The effect size in the population can be estimated from the effect size in the sample, and the sample size is determined by the experimenter anyway so that value is easy to calculate. Now, there are two useful things we can do knowing that these four variables are related:

1. **Calculate the power of a test**: Given that we've conducted our experiment, we will have already selected a value of α, we can estimate the effect size based on our sample, and we will know how many participants we used. Therefore, we can use these values to calculate $1 - \beta$, the power of our test. If this value turns out to be .8 or more we can be confident that we achieved sufficient power to detect any effects that might have existed, but if the resulting value is less, then we might want to replicate the experiment using more participants to increase the power.
2. **Calculate the sample size necessary to achieve a given level of power**: Given that we know the value of α and β, we can use past research to estimate the size of effect that we would hope to detect in an experiment. Even if no one had previously done the exact experiment that we intend to do, we can still estimate the likely effect size based on similar experiments. We can use this estimated effect size to calculate how many participants we would need to detect that effect (based on the values of α and β that we've chosen).

The latter use is the more common: to determine how many participants should be used to achieve the desired level of power. The actual computations are very cumbersome, but fortunately there are now computer programs available that will do them for you (one example is G*Power, which is free and can be downloaded from a link on the companion website; another is nQuery Adviser but this has to be bought!). Also, Cohen (1988) provides extensive tables for calculating the number of participants for a given level of power (and vice versa). Based on Cohen (1992) we can use the following guidelines: if we take the standard α-level of .05 and require the recommended power of .8, then we need 783 participants to detect a small effect size ($r = .1$), 85 participants to detect a medium effect size ($r = .3$) and 28 participants to detect a large effect size ($r = .5$).

What have I discovered about statistics? ①

OK, that has been your crash course in statistical theory! Hopefully your brain is still relatively intact. The key point I want you to understand is that when you carry out research you're trying to see whether some effect genuinely exists in your population (the effect you're interested in will depend on your research interests and your specific predictions). You won't be able to collect data from the entire population (unless you want to spend your entire life, and probably several after-lives, collecting data) so you use a sample instead. Using the data from this sample, you fit a statistical model to test your predictions, or, put another way, detect the effect you're looking for. Statistics boil down to one simple idea: observed data can be predicted from some kind of model and an error associated with that model. You use that model (and usually the error associated with it) to calculate a test statistic. If that model can explain a lot of the variation in the data collected (the probability of obtaining that test statistic is less than .05) then you infer that the effect you're looking for genuinely exists in the population. If the probability of obtaining that test statistic is more than .05, then you conclude that the effect was too small to be detected. Rather than rely on significance, you can also quantify the effect in your sample in a standard way as an *effect size* and this can be helpful in gauging the importance of that effect. We also discovered that I managed to get myself into trouble at nursery school. It was soon time to move on to primary school and to new and scary challenges. It was a bit like using SAS for the first time!

Key terms that I've discovered

α-level
β-level
Central limit theorem
Confidence interval
Degrees of freedom
Deviance
Effect size
Fit
Linear model
Meta-analysis
One-tailed test
Population
Power
Sample
Sampling distribution
Sampling variation
Standard deviation
Standard error
Standard error of the mean (SE)
Sum of squared errors (SS)
Test statistic
Two-tailed test
Type I error
Type II error
Variance

Smart Alex's tasks

- **Task 1:** Why do we use samples? ①
- **Task 2:** What is the mean and how do we tell if it's representative of our data? ①

- **Task 3:** What's the difference between the standard deviation and the standard error? ①
- **Task 4:** In Chapter 1 we used an example of the time taken for 21 heavy smokers to fall off a treadmill at the fastest setting (18, 16, 18, 24, 23, 22, 22, 23, 26, 29, 32, 34, 34, 36, 36, 43, 42, 49, 46, 46, 57). Calculate the sums of squares, variance, standard deviation, standard error and 95% confidence interval of these data. ①
- **Task 5:** What do the sum of squares, variance and standard deviation represent? How do they differ? ①

- **Task 6:** What is a test statistic and what does it tell us? ①
- **Task 7:** What are Type I and Type II errors? ①
- **Task 8:** What is an effect size and how is it measured? ②
- **Task 9:** What is statistical power? ②

Answers can be found on the companion website.

Further reading

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. (A couple of beautiful articles by the best modern writer of statistics that we've had.)

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.)

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 has very clear introductions to sampling, confidence intervals and other important statistical ideas.)

Interesting real research

Domjan, M., Blesbois, E., & Williams, J. (1998). The adaptive significance of sexual conditioning: Pavlovian control of sperm release. *Psychological Science*, *9*(5), 411–415.