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

# 2

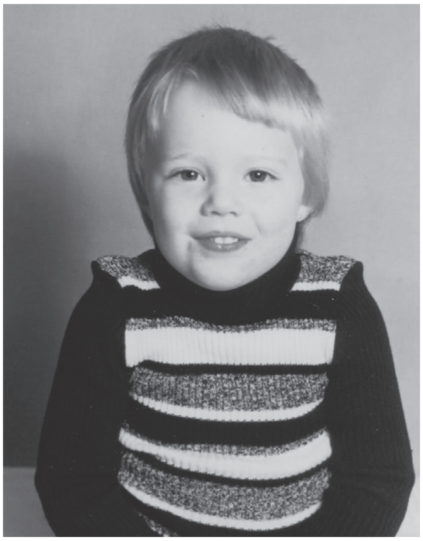
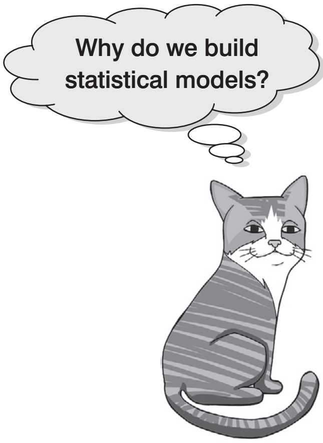


**FIGURE 2.1**
The face of innocence … but what are the hands doing?

## 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

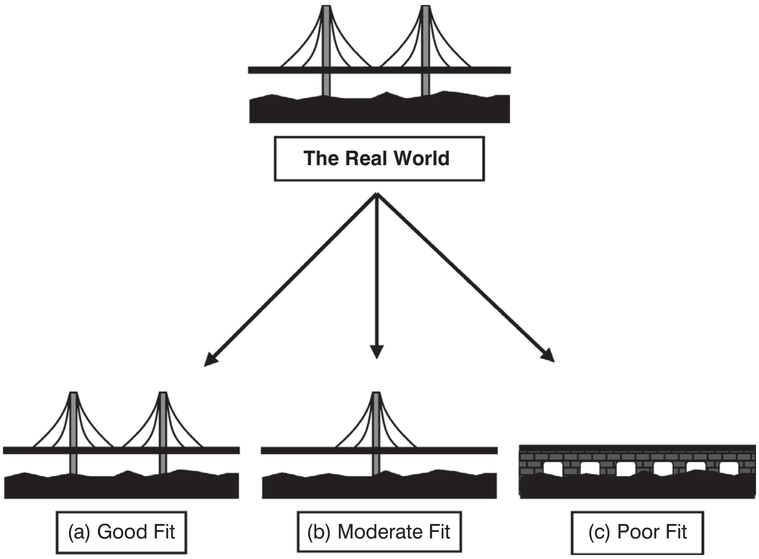


**FIGURE 2.2** Fitting models to real-world data (see text for details)

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

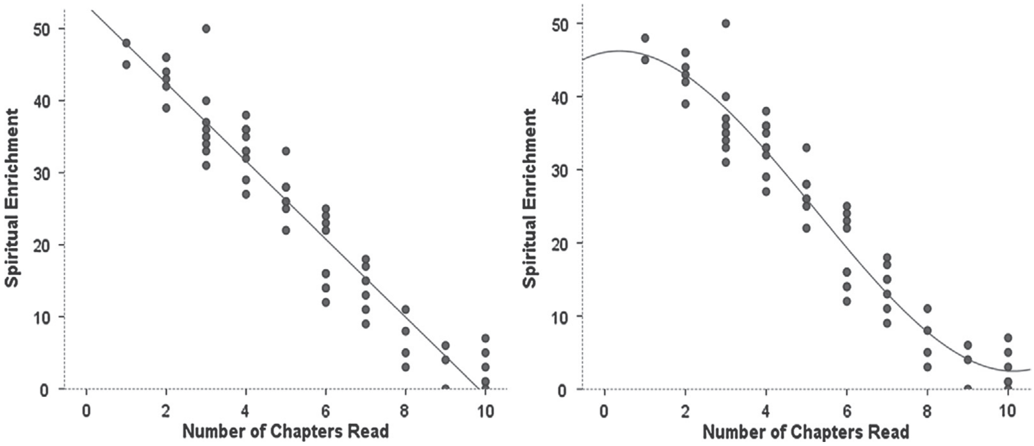


**FIGURE 2.3** A scatterplot of the same data with a linear model fitted (left), and with a non-linear model fitted (right)

(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 *i*th 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, $x_i = x_5 = 4$.

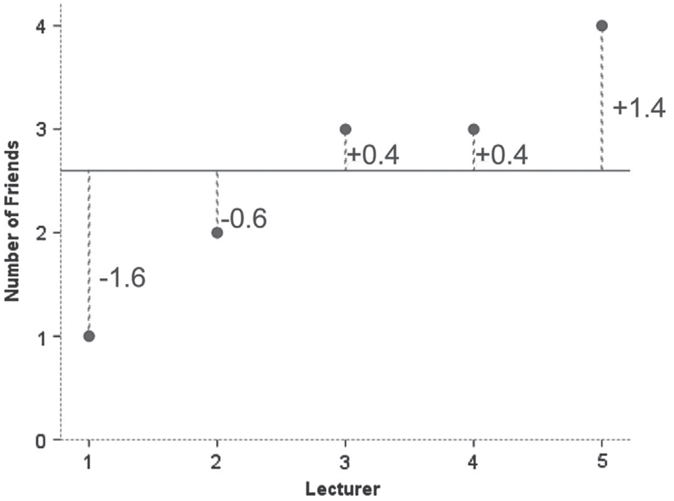


**FIGURE 2.4** Graph showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends

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:

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

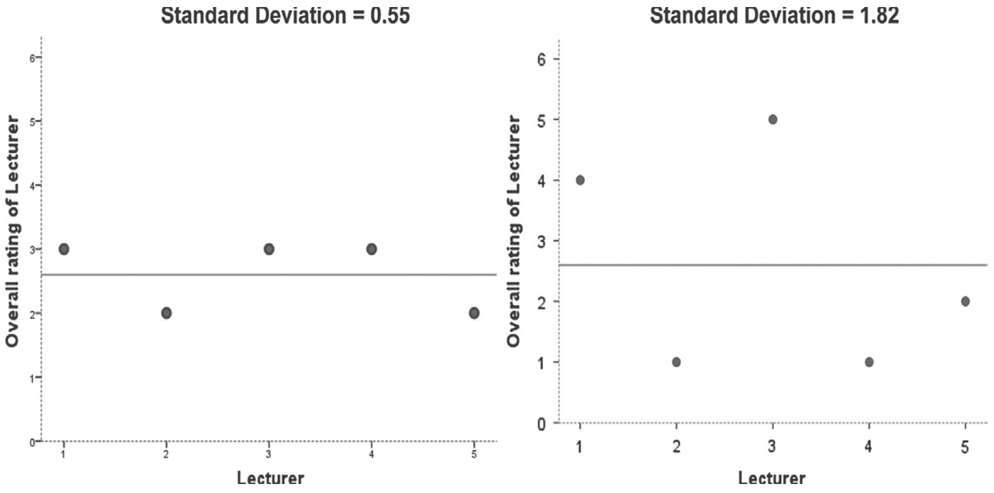


**FIGURE 2.5** Graphs illustrating data that have the same mean but different standard deviations

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.

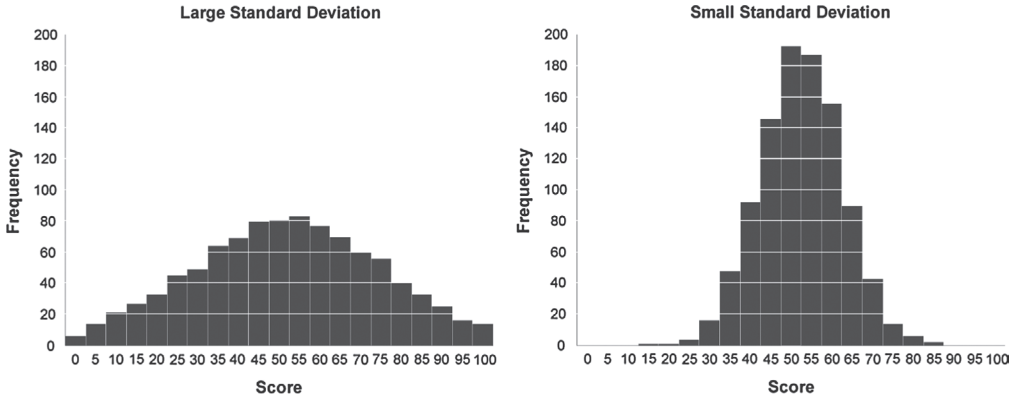


**FIGURE 2.6** Two distributions with the same mean, but large and small standard deviations

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:

$$\begin{aligned}\text{outcome}_{\text{lecturer1}} &= \overline{\text{X}} + \varepsilon_{\text{lecturer1}}\\ 1 &= 2.6 + \varepsilon_{\text{lecturer1}}\end{aligned}$$

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

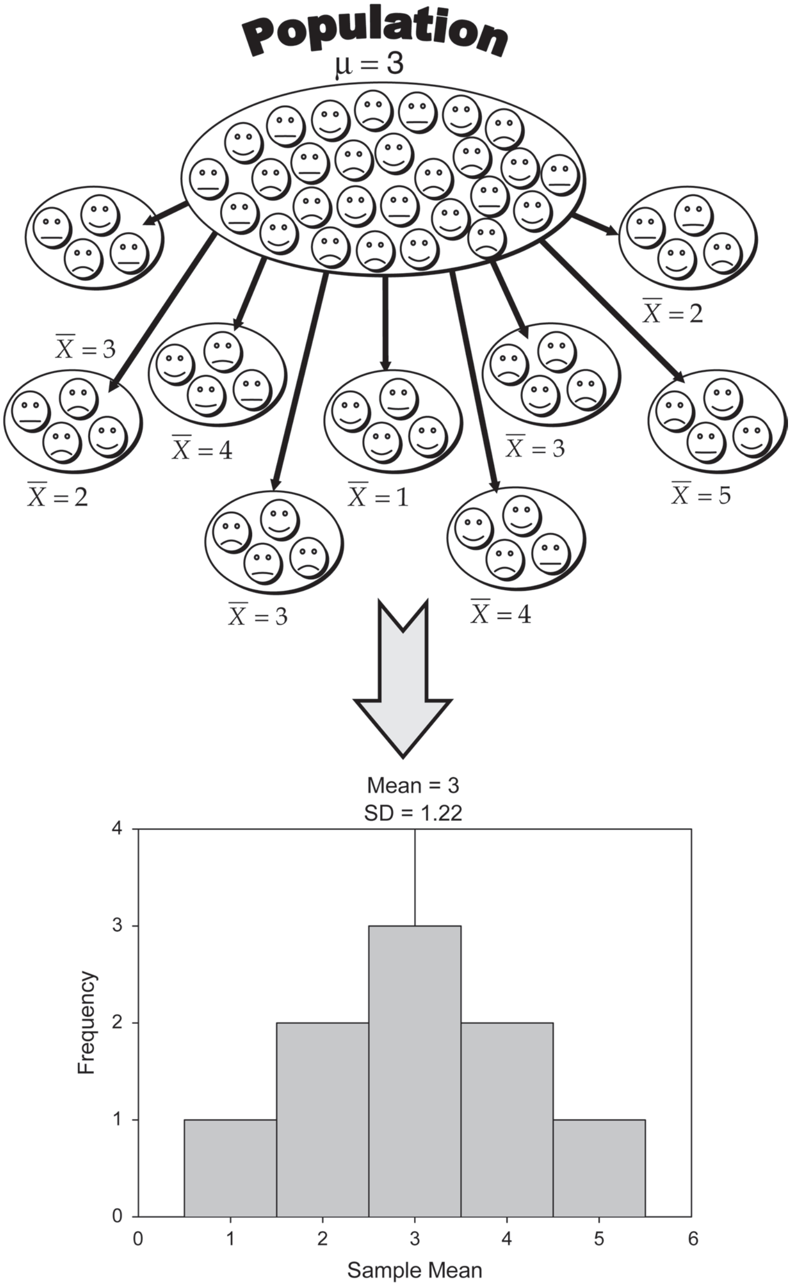


**FIGURE 2.7** Illustration of the standard error (see text for details)

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*, $\mu$). 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 ($\sigma$) 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**

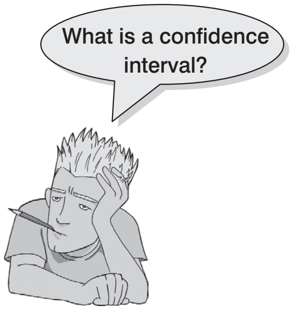
**The standard error is the standard deviation of sample means.** As such, it is a measure of how representative a sample is likely to be of the population. A large standard error (relative to the sample mean) means that there is a lot of variability between the means of different samples and so the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

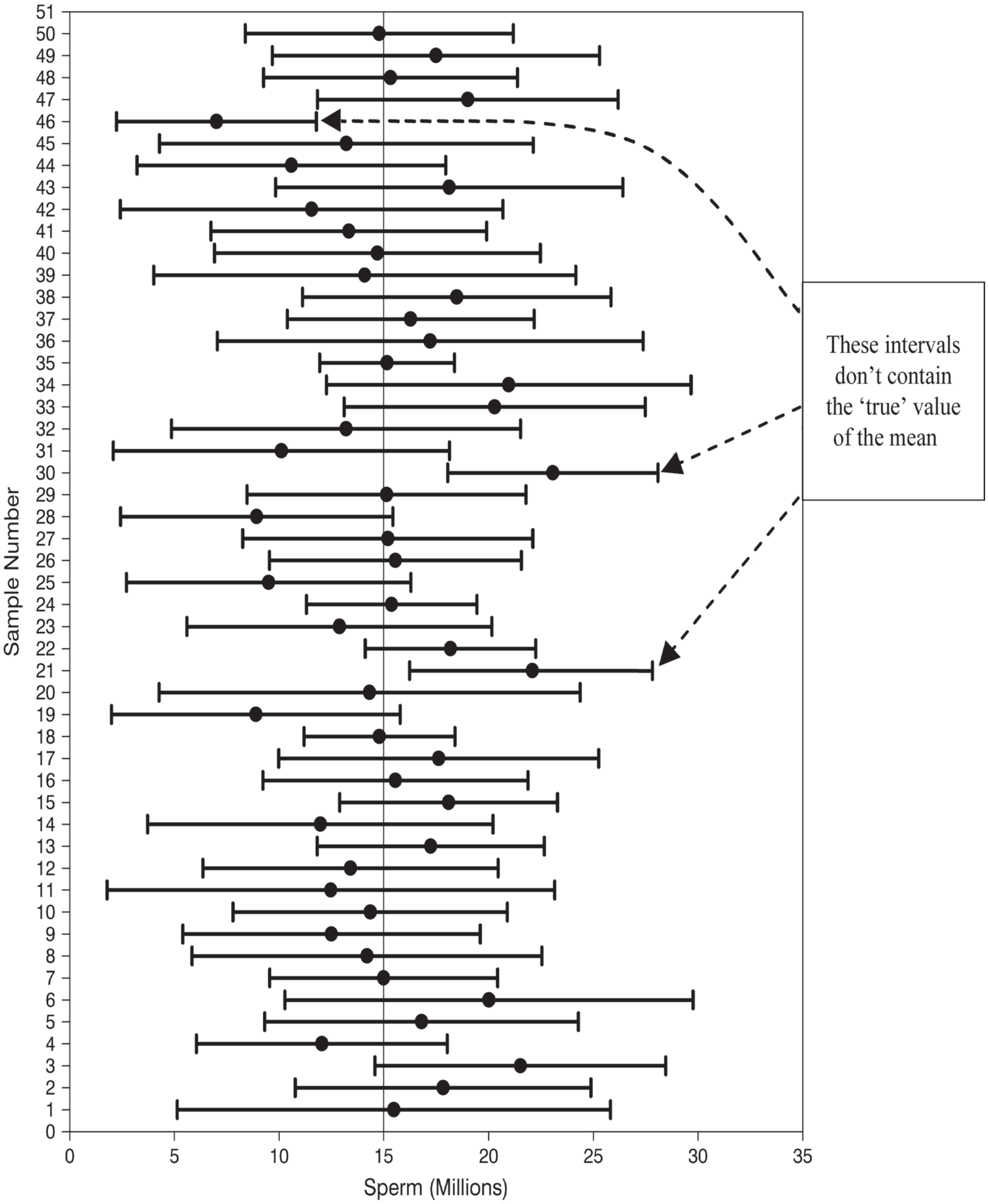


**FIGURE 2.8** The confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 50 different samples (vertical axis)

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*:

$$\begin{aligned} 1.96 \times s &= X - \overline{X} & \qquad -1.96 \times s &= X - \overline{X} \\ (1.96 \times s) + \overline{X} &= X & \qquad (-1.96 \times s) + \overline{X} &= X \end{aligned}$$

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.

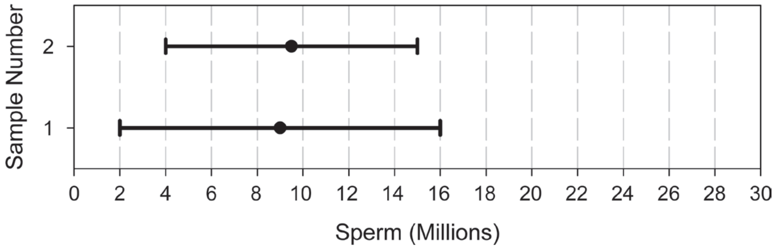
SELF-TEST Recalculate the confidence interval assuming that the sample size was 56

### 2.5.2.4. Showing confidence intervals visually ②

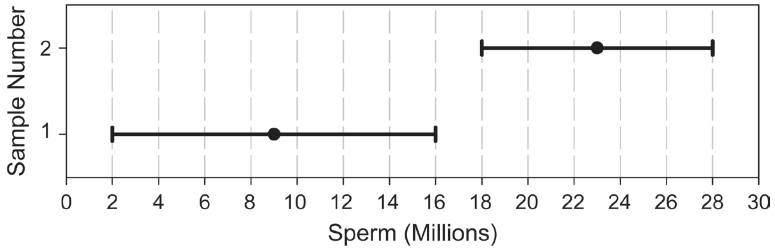
Confidence intervals provide us with very important information about the mean, and, therefore, you often see them displayed on graphs. (We will discover more about how to create these graphs in Chapter 4.) The confidence interval is usually displayed using something called an error bar, which just looks 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!

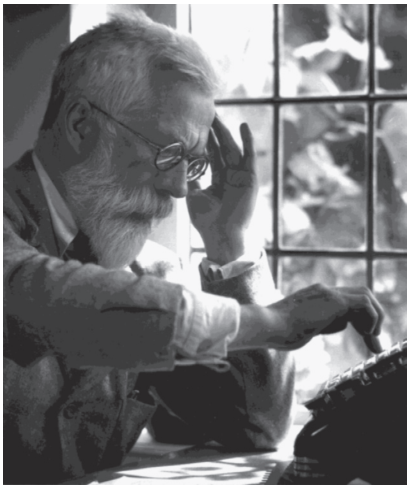


**FIGURE 2.9** Sir Ronald A. Fisher, the cleverest person ever ($p < .0001$)

## 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 $\chi^2$ 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/).

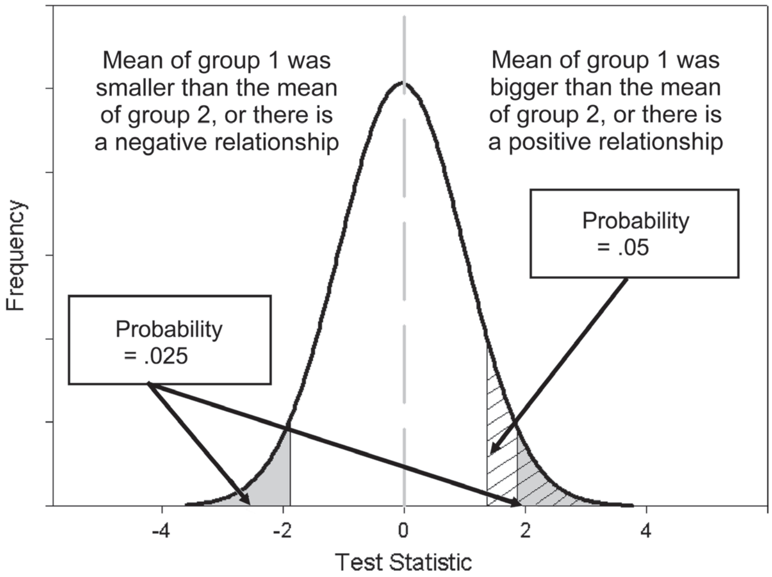


**FIGURE 2.10** Diagram to show the difference between one- and two-tailed tests

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

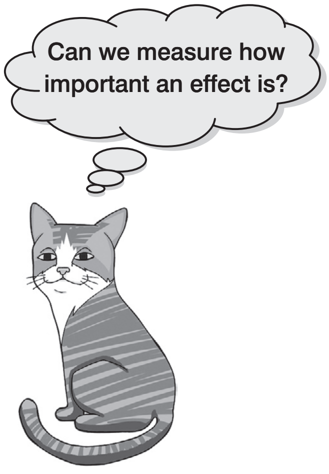
We have seen that we use test statistics to tell us about the true state of the world (to a certain degree of confidence). Specifically, we're trying to see whether there is an effect in our population. There are two possibilities in the real world: there is, in reality, an effect in the population, 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 **$\alpha$-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 **$\beta$-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 $\alpha$ 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 $\alpha$-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 $\alpha$-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 ($\beta$, 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 $\alpha$, 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 $\alpha$ and $\beta$, 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 $\alpha$ and $\beta$ 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 $\alpha$-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

$\alpha$-level
$\beta$-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.