

Copyright Notice

Staff and students of this University are reminded that copyright subsists in this extract and the work from which it was taken. This Digital Copy has been made under the terms of a CLA Licence which allows you to:

- Access and download a copy
- Print out a copy

This Digital Copy and any digital or printed copy supplied to or made by you under the terms of this Licence are for use in connection with this module or course of study. You may retain such copies after the end of the course, but strictly for your own personal use.

All copies (including electronic copies) shall include this Copyright Notice and shall be destroyed and/or deleted if and when required by the University.

Except as provided for by copyright law, no further copying, storage or distribution (including by e-mail) is permitted without the consent of the copyright holder.

The author (which term includes artists and other visual creators) has moral rights in the work and neither staff nor students may cause, or permit, the distortion, mutilation or other modification of the work, or any other derogatory treatment of it, which would be prejudicial to the honour or reputation of the author.

Title of extract: **Everything You Ever Wanted to Know about Statistics (Well Sort of)**

Name of Author: **Field, A.**

Title/Author of book/edition (if applicable): **Discovering Statistics Using SPSS for Windows. 2nd edition Field, A.**

Name of Visual Creator or other acknowledgement (if applicable):

Journal details –title/Vol/Year (if applicable):

Publisher: **Sage**

Pages scanned: **1 - 36**

Course of Study (Module name and number): **Quantitative Research Methods
DSAW7102**

CHAPTER 1

EVERYTHING YOU EVER WANTED TO KNOW ABOUT STATISTICS (WELL, SORT OF)

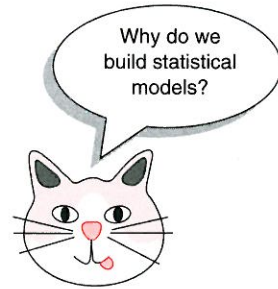
1.1. WHAT WILL THIS CHAPTER TELL US? ①

I realize that many people will use this book to ‘dip in’ to chapters that describe the statistical test that they’ve suddenly discovered they have to use. However, for any of those chapters to make sense you might find it useful to know a few important things about statistics. This chapter is an attempt to give you a brief overview of some important statistical concepts such as how we use statistical models to answer scientific questions.

1.2. BUILDING STATISTICAL MODELS ①

In the social sciences we are usually 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 seek to explain it by collecting data from the real world, and then using these data to draw conclusions about what is being studied. As statisticians our job is to take the available data and to use them in a meaningful way and this often 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, the 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: we build models of real-world processes in an attempt to predict how these processes operate under certain conditions. 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 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 and this is a term you will frequently come across.

Figure 1.1 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 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 (i.e. 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. if the bridge collapses in strong winds this could be due to the absence of a second supporting tower). So, using this model results in predictions that we can have some confidence in but not complete confidence. The final model (c) is completely different to the real-world situation. This model bears no structural similarities to the real bridge and so could be termed 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

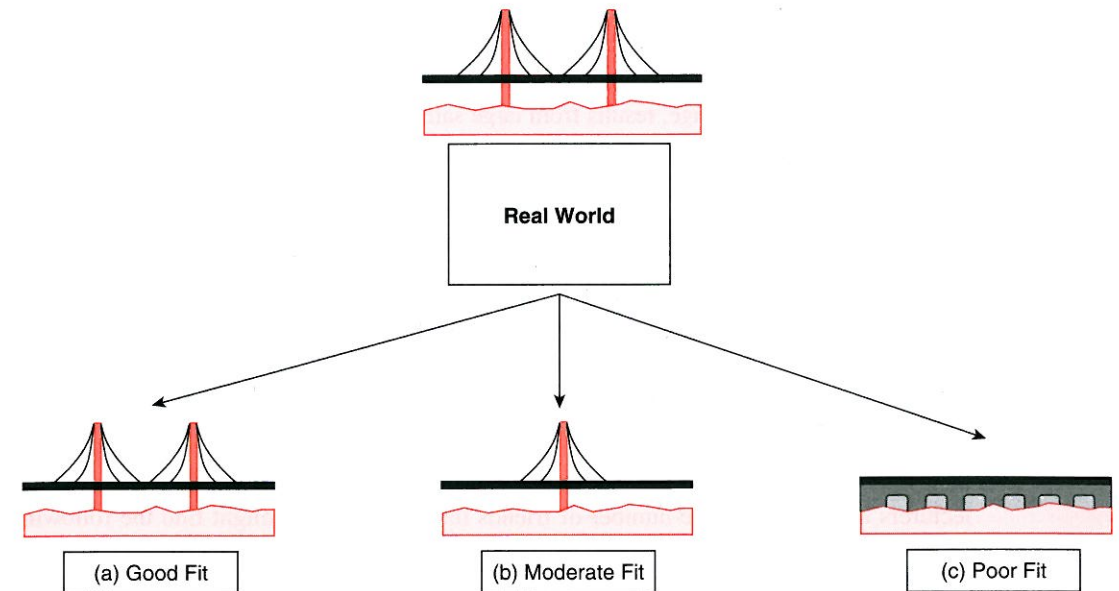


Figure 1.1 Fitting models to real-world data (see text for details)

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.

1.3. 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), but in either case 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 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-sized bridge. This metaphor can be extended to social 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, results from large samples should be fairly similar.

1.4. SIMPLE STATISTICAL MODELS ①

1.4.1. The mean, sums of squares, variance and standard deviations ①

One of the simplest models used in statistics is the mean. Some of you may have trouble thinking of the mean as a model, but in fact it is because it represents a summary of data. The **mean** is a hypothetical value that can be calculated for any data set; it doesn't have to be a value that is actually observed in the data set. 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) so the mean value is a *hypothetical* value. As such, the mean is a model created to summarize our data. Now, we can determine whether this is an accurate model by looking at how different our real data are from the model that we have created. One way to do this is to look at the difference between the data we observed and the model fitted. Figure 1.2 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 **deviance** 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).¹ For example, lecturer 1 had only 1 friend and so the difference is $x_1 - \bar{x} = 1 - 2.6 = -1.6$. You might notice that the deviance is a minus number, and this represents the fact that our model *overestimates* this lecturer's popularity: it 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:

total error = sum of deviances
= $\sum (x_i - \bar{x}) = (-1.6) + (-0.6) + (0.4) + (0.4) + (1.4) = 0$

¹ 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, $x_i = x_5 = 4$.

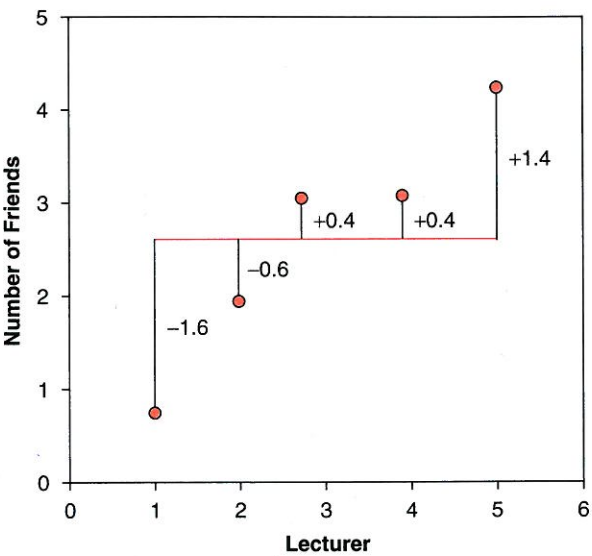


Figure 1.2 Graph showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends (the red line)

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,² 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:

sum of squared errors (SS) = $\sum (x_i - \bar{x})(x_i - \bar{x})$
= $(-1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2$
= $2.56 + 0.36 + 0.16 + 0.16 + 1.96$
= 5.20

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

² When you multiply a negative number by itself it becomes positive.

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 Box 8.2). This measure is known as the **variance** and is a measure that we will come across a great deal:

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

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:

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

The standard deviation is, therefore, a measure of 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 1.3 shows the overall ratings (on a five-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 of 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 of the data. This illustration should hopefully make clear why the standard deviation is a measure of how well the mean represents the data.

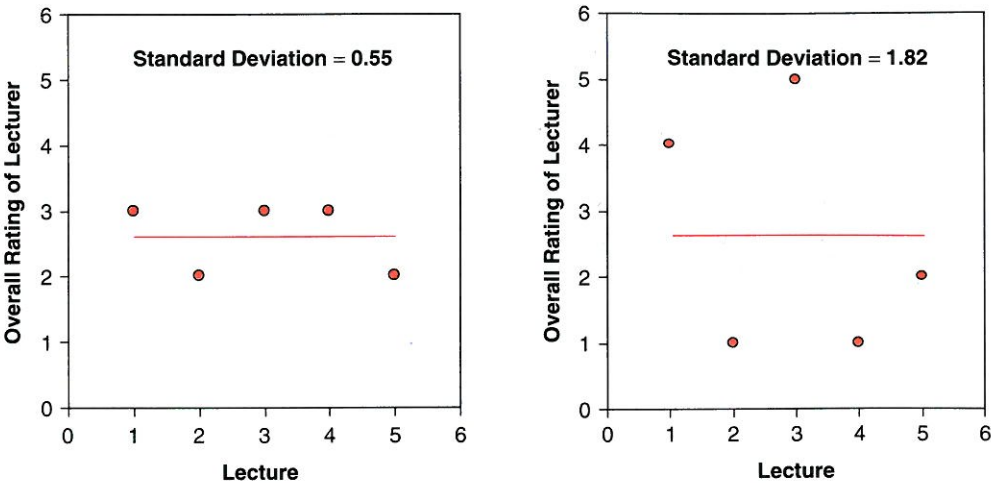


Figure 1.3 Graphs illustrating data that have the same mean but different standard deviations

The discussion of means, sums of squares and variance may seem a side-track from the initial point, but in fact the mean is probably one of the simplest statistical models that can be fitted to data. What do I mean by this? Well, everything in statistics essentially boils down to one equation:

$$\text{Outcome}_i = (\text{Model}_i) + \text{error}_i \tag{1.1}$$

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. With the mean, 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 he had one friend and the mean of all lecturers was 2.6. So, the equation becomes:

$$\begin{aligned} \text{Outcome}_{\text{lecturer 1}} &= \bar{X} + \varepsilon_{\text{lecturer 1}} \\ 1 &= 2.6 + \varepsilon_{\text{lecturer 1}} \end{aligned}$$

From this we can work out that the error is $1 - 2.6$, or -1.6 . If we replace this value into 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!

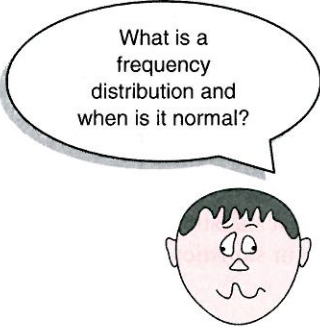
Likewise, the variance and standard deviation illustrate another fundamental concept: how the goodness-of-fit of a model can be measured. If we’re looking at how well a model

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:

$$\text{deviation} = \sum (\text{observed} - \text{model})^2 \tag{1.2}$$

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.

1.5. FREQUENCY DISTRIBUTIONS ①



Once you've collected some data a very useful thing to do is to plot a graph of how many times each score occurs. This is known as a **frequency distribution**, or **histogram**, which is simply a graph plotting values of observations on the horizontal axis, with a bar showing how many times each value occurred in the data set. Frequency distributions can be very useful for assessing properties of the distribution of scores. For one thing, by looking at which score has the tallest bar, we can immediately see the **mode**, which is simply the score that occurs most frequently in the data set.

Frequency distributions come in many different shapes and sizes. It is quite important, therefore, to have some general descriptions for common types of distributions. In an ideal world our data would be distributed symmetrically around the centre of all scores. As such, if we drew a vertical line through the centre of the distribution then it should look the same on both sides. This is known as a **normal distribution** and is characterized by the bell-shaped curve with which you might already be familiar. This shape basically implies that the majority of scores lie around the centre of the distribution (so the largest bars on the histogram are all around the central value). Also, as we get further away from the centre the bars get smaller, implying that as scores start to deviate from the centre their frequency is decreasing. As we move still further away from the centre our scores become very infrequent (the bars are very short). An example of a normal distribution is shown in Figure 1.4.

1.5.1. Properties of frequency distributions ①

There are two main ways in which a distribution can deviate from normal: (1) lack of symmetry (called **skewness**) and (2) pointyness (called **kurtosis**). Skewed distributions are not symmetrical and instead the most frequent scores (the tall bars on the graph) are clustered at one end of the scale. So, the typical pattern is a cluster of frequent scores at one end of the scale and the frequency of scores tailing off towards the other end of

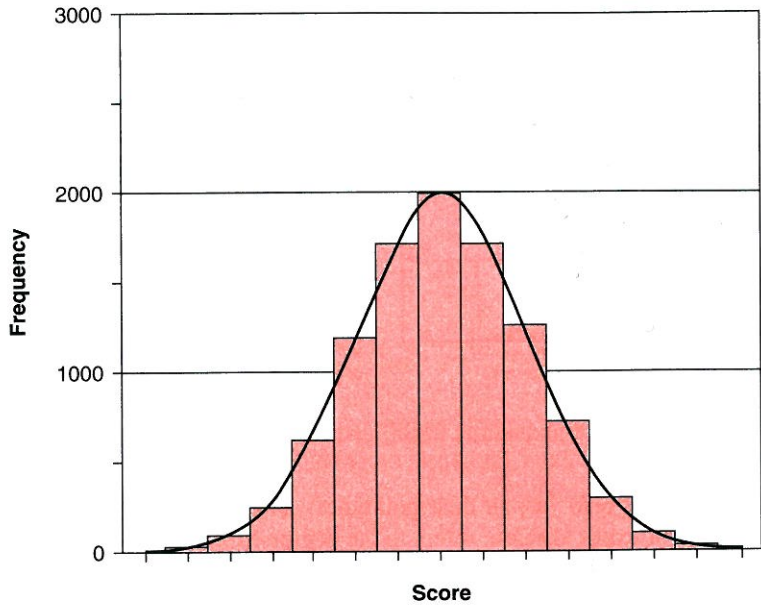


Figure 1.4 A 'normal' distribution (the curve shows the idealized shape)

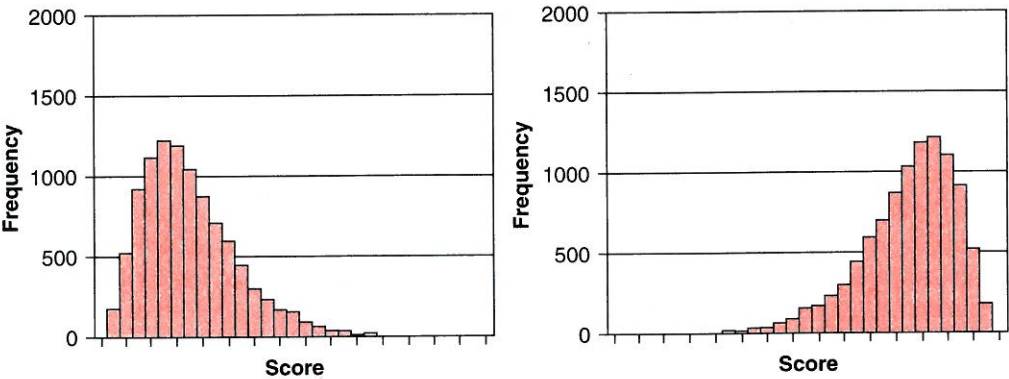


Figure 1.5 A positively (left) and negatively (right) skewed distribution

the scale. A skewed distribution can be either *positively skewed* (the frequent scores are clustered at the lower end and the tail points towards the higher or more positive scores) or *negatively skewed* (the frequent scores are clustered at the higher end of and the tail points towards the lower more negative scores). Figure 1.5 shows examples of these distributions.

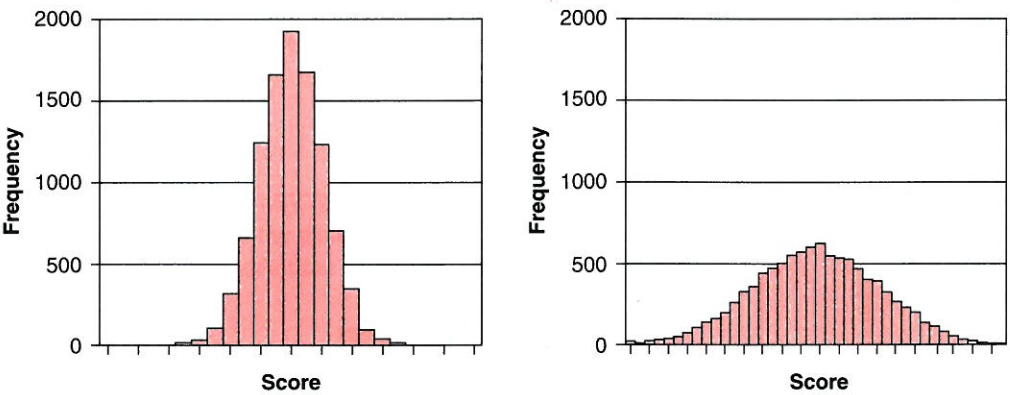


Figure 1.6 A leptokurtic (left) and platykurtic (right) distribution

Distributions also vary in their pointiness, or kurtosis. Kurtosis, despite sounding like some kind of exotic disease, refers to the degree to which scores cluster in the tails of the distribution. This characteristic is usually shown up by how flat or pointy a distribution is. A **platykurtic** distribution is one that has many scores in the tails (a so-called heavy-tailed distribution) and so is quite flat. In contrast, **leptokurtic** distributions are relatively thin in the tails and so look quite pointy. You can remember this by ‘the *leptokurtic* distribution *leapt* up in the air, and the *platykurtic* distribution was flat like a *plateau*’ (Figure 1.6), or the surrealists amongst you can remember it as ‘the platykurtic distribution had a duck-bill and the leptokurtic distribution compulsively stole things’. Ideally, we want our data to be normally distributed (i.e. not too skewed, and not too pointy or flat!).

In a normal distribution the values of skew and kurtosis are 0 (i.e. the distribution is neither too pointy, nor too flat, and is perfectly symmetrical). If a distribution has values of skew or kurtosis above or below 0 then this indicates a deviation from normal (see Chapter 3).

1.5.2. 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. 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 1.3) and the standard deviation is larger. Figure 1.7 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

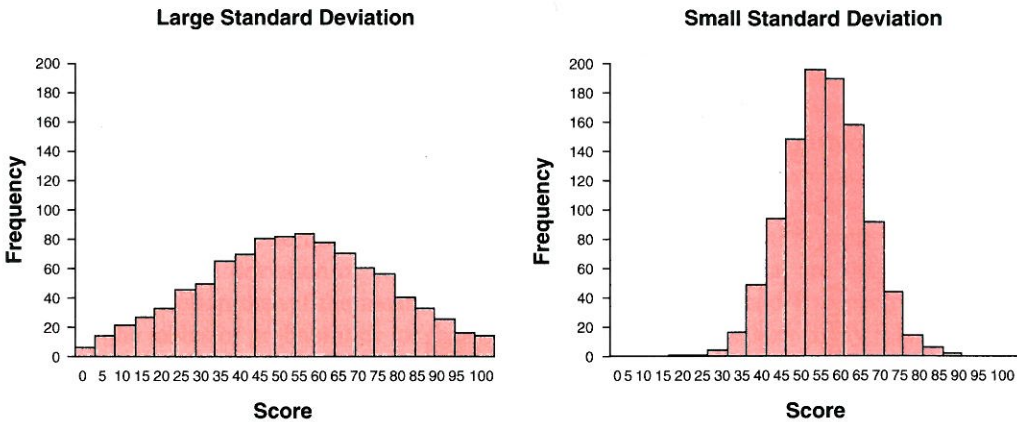


Figure 1.7 Two distributions with the same mean, but large and small standard deviations

but scores further from the mean become increasingly infrequent. The main message is that as the standard deviation gets larger, the distribution gets fatter.

1.5.3. What is the standard normal distribution? ①

Another way to look at frequency distributions is in terms of probability: they give us some idea of how likely a given score is to occur. Beachy Head is a large, windy cliff on the Sussex coast (not far from where I live) that has something of a reputation for attracting suicidal people, who seem to like throwing themselves off of it (and after several months of rewriting this book I find my thoughts drawn towards that peaceful chalky cliff top more and more often). Figure 1.8 shows a frequency distribution of some completely made up data of the number of suicides at Beachy Head in a year by people of different ages (although I made these data up they are roughly based on general suicide statistics such as those in Williams, 2001). There were 171 suicides in total and you can see that the suicidal lemmings were most frequently aged between about 25 and 30 (the highest bar), which is quite worrying for me (although it does mean that if I can just keep going for this year the urges will decrease next year!). The graph also tells us that, for example, very few people aged above 70 committed suicide at Beachy Head.

I said earlier that we could think of frequency distributions in terms of probability. To explain this, imagine that someone asked you ‘how likely is it that a 70 year old committed suicide at Beachy Head?’; what would your answer be? The chances are that if you looked at the frequency distribution you might respond ‘not very likely’ because you can see a couple of people out of the 171 suicides were aged around 70. What about if someone asked ‘how likely is it that a 30 year old committed suicide?’ Again, by looking at the graph, you might say ‘it’s actually quite likely’ because 32 out of the 171 suicides were by people aged around 30 (that’s more than one in every five people that committed suicide).

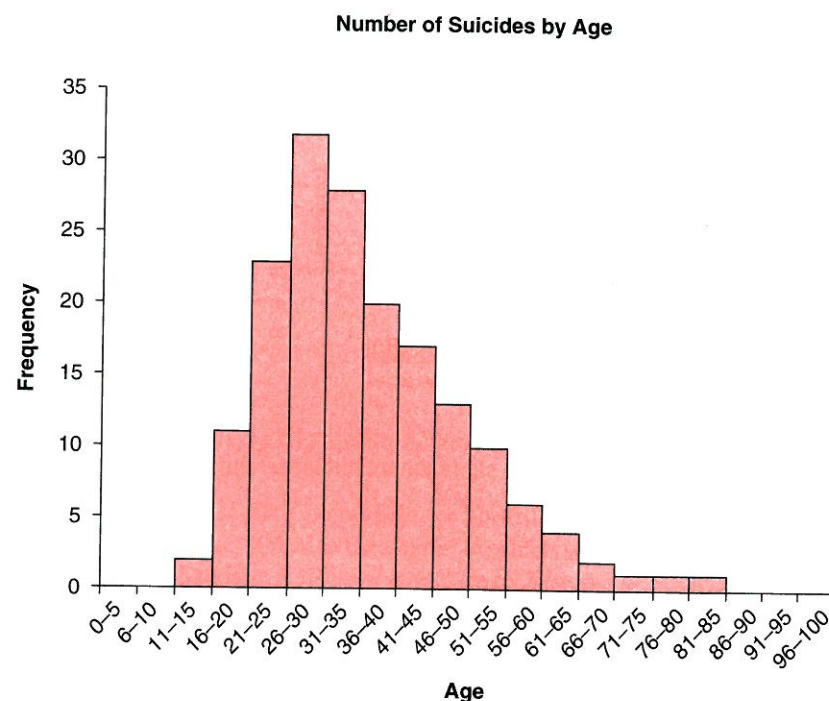


Figure 1.8 Frequency distribution showing the number of suicides at Beachy Head in year by age

So based on the frequencies of different scores it should start to become clear that we could use this information to estimate the probability that a particular score will occur. So, we could ask, based on our data, what's the probability of a suicide victim being aged 16–20? A probability value can range from 0 (there's no chance whatsoever of the event happening) to 1 (the event will definitely happen). So, for example, when I talk to my publishers I tell them there's a probability of 1 that I will have completed the revisions to this book by June 2003. However, when I talk to anyone else, I might, more realistically, tell them that there's a 0.1 probability of me finishing the revisions on time (or put another way, a 10% chance, or a 1 in 10 chance that I'll complete the book in time). In reality, the probability of me meeting the deadline is 0 (not a chance in hell) because I never manage to meet publishers' deadlines! If probabilities don't make sense to you then just ignore the decimal point and think of them as percentages instead (i.e. 0.1 probability that something will happen = 10% chance that something will happen).

I've talked in vague terms about how frequency distributions can be used to get a rough idea of the probability of a score occurring. However, we can be precise. For any distribution of scores we could, in theory, calculate the probability of obtaining a score of a certain size—it would be incredibly tedious and complex to do it, but we could. To spare our

sanity, statisticians have identified several common distributions. For each one they have worked out mathematical formulae that specify idealized versions of these distributions (they are specified in terms of a curved line). These idealized distributions are known as **probability distributions** and from these distributions it is possible to calculate the probability of getting particular scores based on the frequencies with which a particular score occurs in a distribution with these common shapes. One of these 'common' distributions is the normal distribution, which I've already mentioned. Statisticians have calculated the probability of certain scores occurring in a **normal distribution** with a mean of 0 and a standard deviation of 1. Therefore, if we have any data that are shaped like a normal distribution then if the mean and standard deviation are 0 and 1 respectively we can use the tables of probabilities for the normal distribution to see how likely it is that a particular score will occur in the data (I've produced such a table in the Appendix to this book).

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

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

The table of probability values that have been calculated for the standard normal distribution is shown in the Appendix. Why is this table important? Well, if we look at our suicide data, we can answer the question 'what's the probability that someone who threw themselves off of Beachy Head was aged 70 or older?' First we convert 70 into a z-score. Say the mean of the suicide scores was 36 and the standard deviation 13; the 70 will become $(70 - 36)/13 = 2.62$. We then look up this value in the column labelled 'smaller portion' (i.e. the area above the value 2.62). We should find that the probability is .0044, or, put another way, only a 0.44% chance that a suicide victim would be 70 years old or more. By looking at the column labelled 'bigger portion' we can also see the probability that a suicide victim was aged 70 or less! This probability is .9956, or, put another way, there's a 99.56% chance that a suicide victim was less than 70 years old!

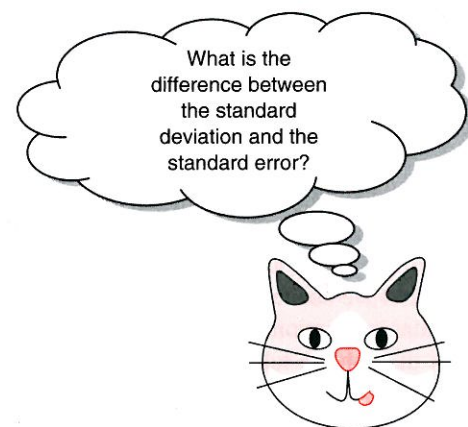
Hopefully you can see from these examples that the normal distribution and z-scores allow us to go a first step beyond our data in that from a set of scores we can calculate the probability that a particular score will occur. So, we can see whether scores of a certain size are likely or unlikely to occur in a distribution of a particular kind. You'll see just how useful this is in due course, but it is worth mentioning at this stage that certain z-scores are particularly important. This is because their value cuts off certain important percentages of the distribution. The first important value of z is 1.96 because this cuts off the top 2.5% of the distribution, and its counterpart at the opposite end (-1.96) cuts off the bottom 2.5% of the distribution. As such, taken together, these values cut off 5%

of scores, or, put another way, 95% of z-scores lie between -1.96 and 1.96 . The other two important benchmarks are ± 2.58 and ± 3.29 , which cut off 1% and 0.1% of scores respectively. Put another way, 99% of z-scores lie between -2.58 and 2.58 , and 99.9% of them lie between -3.29 and 3.29 . Remember these values because they'll crop up time and time again!

1.6. IS MY SAMPLE REPRESENTATIVE OF THE POPULATION? ①

1.6.1. The standard error ①

We've seen that the standard deviation and frequency distributions tell us something about how well the mean represents that data, but I mentioned earlier on that usually we collect data from samples rather than the entire population. I also mentioned that if you take



several samples from a population, then these samples would 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 samples. If we take several samples from the same population, then each sample has its own mean, and some of these sample means will be different (not every sample will have the same mean).

Figure 1.9 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*). 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 as 3, but the second sample were, on average, rated 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,³ 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 we'd take hundreds or thousands of samples to construct a sampling distribution, but I'm just using nine to keep the diagram simple! 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 know the accuracy of that average then 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.

Of course, in reality we cannot collect hundreds of samples and so we rely on approximations of the standard error (luckily for us lots of clever statisticians have calculated ways in which the standard error can be worked out from the sample standard deviation). Luckily, we don't need to understand why this approximation works. We can just trust that these people are ridiculously clever and know what they're talking about. The standard error can be calculated by dividing the sample standard deviation (s) by the square root of the sample size (N):⁴

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

³ This is just a graph of each sample mean plotted against the number of samples that has that mean—see Chapter 2 for more details.

⁴ In fact it should be the *population* standard deviation (σ) that is divided by the square root of the sample size; however, for large samples this is a reasonable approximation.

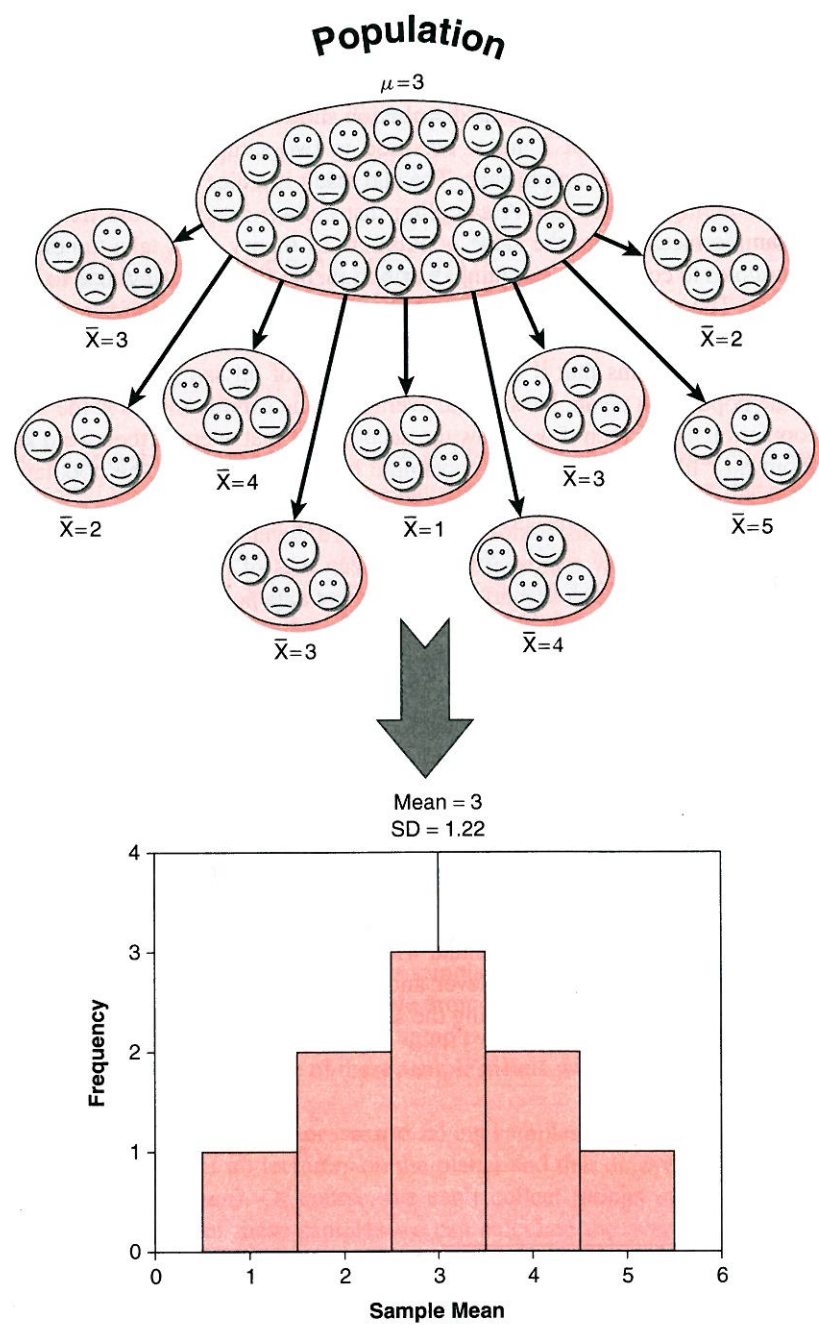


Figure 1.9 Figure illustrating the standard error (see text for details)

Cramming Samantha's Tip



To sum up, 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.

1.6.2. Confidence intervals ②

Remember that usually we're interested in using the sample mean as an estimate of the true value of the mean (i.e. 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 & 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 — sick eh?). 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 17 million falls exactly between these values). Of course, in this case the true value (15 million) does fall within these limits. However, what if we'd set smaller limits, what if we'd said we think the true value falls between 16 and 18 million (again, note that 17 million is in the middle)? In this case the interval does not contain the 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 1.10 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 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 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 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 1.10) 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 1.5.3 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 had a mean of 0 and a standard error of 1, then the limits of our confidence interval would be -1.96 and $+1.96$. You might also remember that we can convert scores into z -scores using this equation:

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

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

$$1.96 = \frac{X - \bar{X}}{s} \quad -1.96 = \frac{X - \bar{X}}{s}$$

What we need to know is the value of X in these equations, and to discover this we simply rearrange them:

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

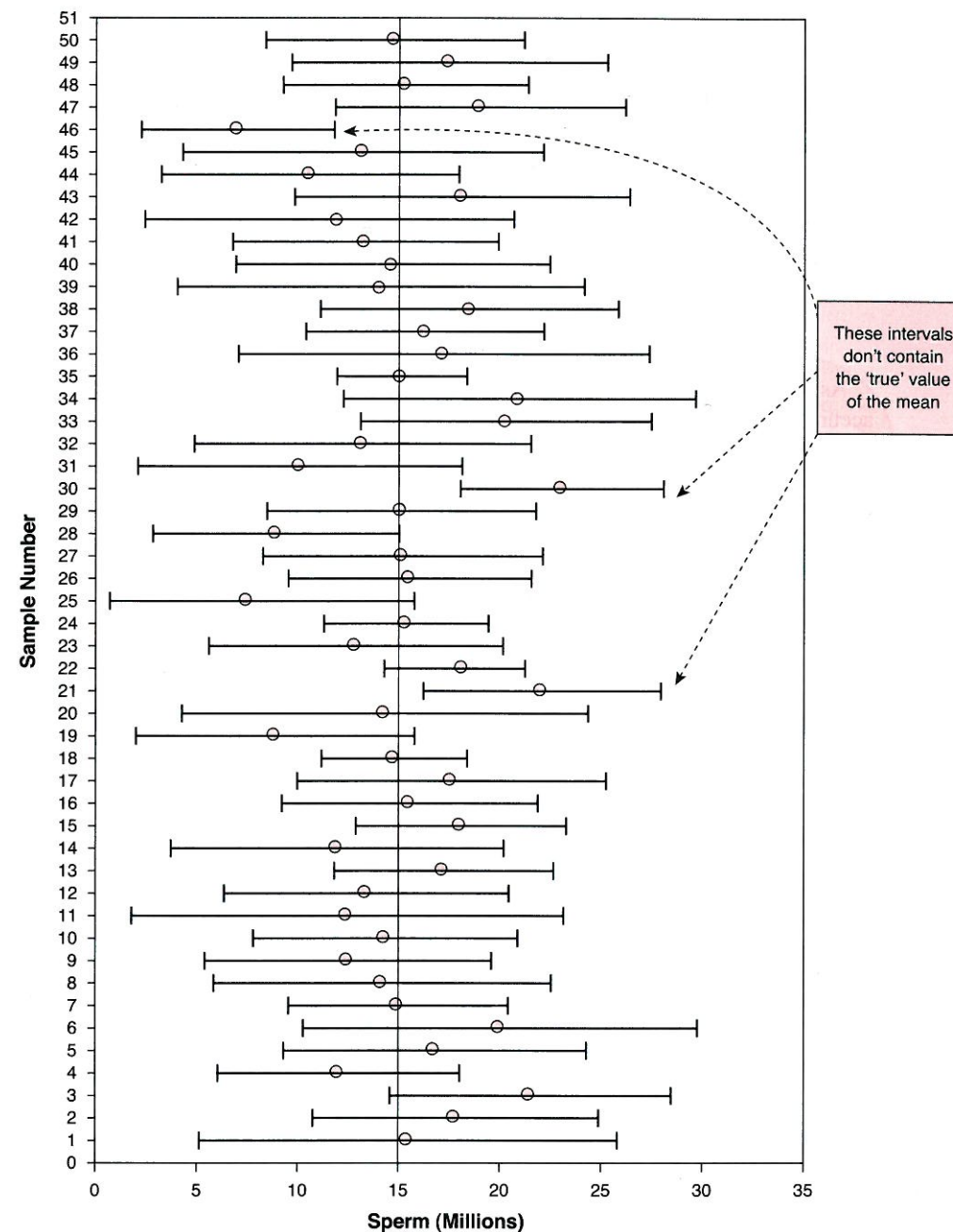


Figure 1.10 Figure showing confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 50 different samples (vertical axis)

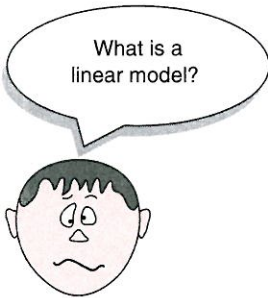
Therefore, the confidence interval can easily be calculated once the standard deviation (*s* in the equations above) and mean ($\bar{X}$ in the equations) are known. However, in fact we use the standard error and not the standard deviation because we're interested in the variability of sample means, 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.

Lower boundary of confidence interval = $\bar{X} - (1.96 \times SE)$

Upper boundary of confidence interval = $\bar{X} + (1.96 \times SE)$

As such, the mean is always in the centre of the confidence interval. If the mean accurately represents the true mean, then the confidence interval of that mean should be small: because 95% of confidence intervals contain the true mean, we can assume that 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.

1.7. LINEAR MODELS ①



The mean is an example of a statistical model, but you may well ask what other kinds of statistical models can be built. Well, if the truth is known there is only one model that is generally used, and this is known as the linear model. To some social scientists it may not be entirely obvious that my previous statement is correct, yet a statistician would acknowledge my sentiment much more readily. The reason for this is that there are a variety of different names given to statistical procedures that are based on the linear model. A classic example is that analysis of variance (ANOVA) and regression are identical systems (Cohen, 1968), yet they have different names and are used largely in different contexts (due to a divide in methodological philosophies—see Cronbach, 1957).

The word *linear* literally means 'relating to a line' but in statistical terms the line to which it refers is a straight one. A linear model is, therefore, 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. For example, in the chapter describing regression, it will become clear that two variables can be negatively related (this just means that as values of one variable increase, values of the other variable decrease). In such circumstances, the relationship may be summarized by a straight line.

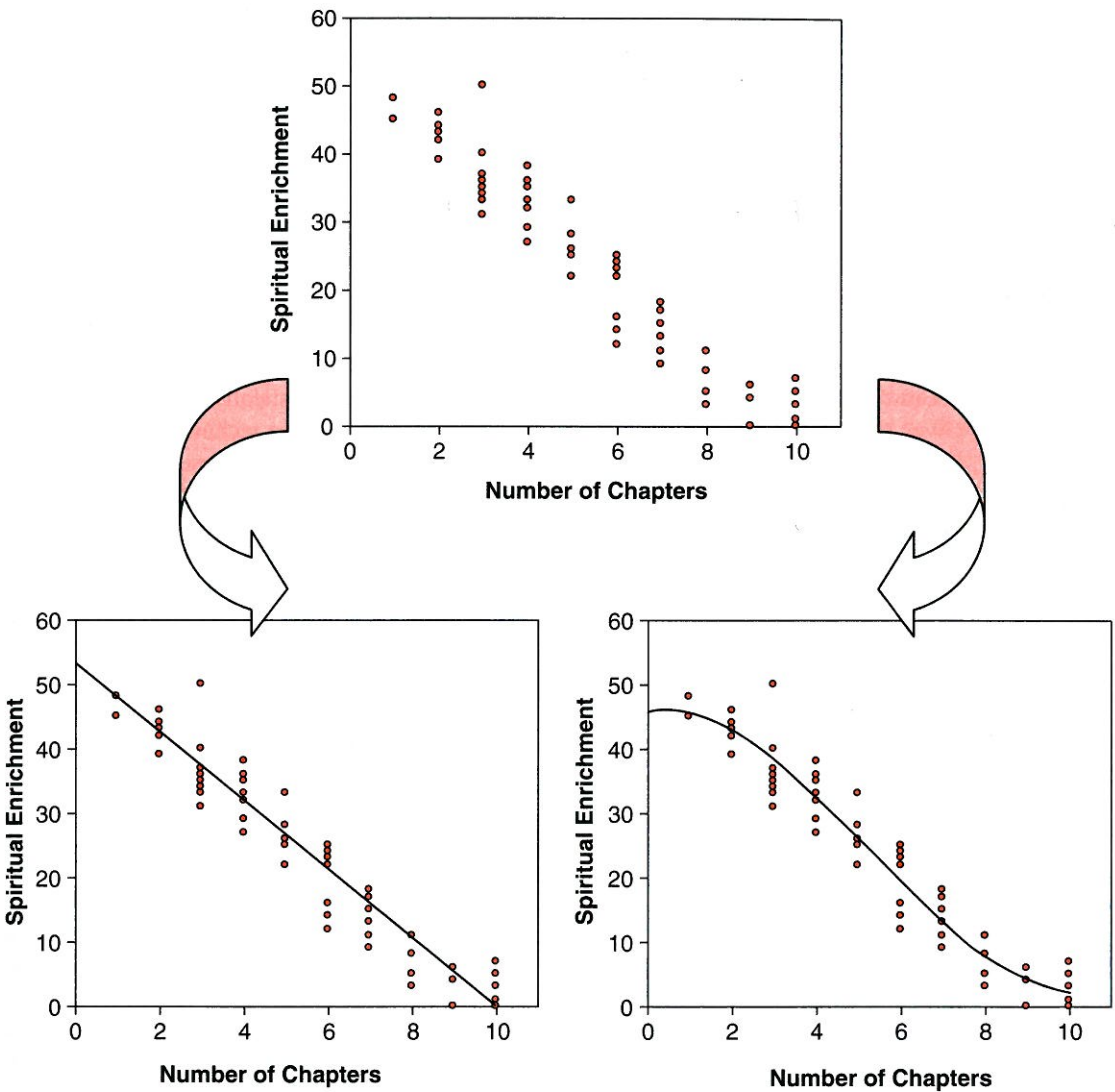


Figure 1.11 A scatterplot of some data with no model fitted to the data, with a linear model fitted, and with a non-linear model fitted

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. Figure 1.11 shows such a graph, and also shows the same graph but with a line that summarizes the

pattern of these data. A third version of the scatterplot is also included but has a curved line to summarize the general pattern of the data. As such, Figure 1.11 illustrates 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.

Most of the statistics used in the social sciences are based on linear models, which means that we try to fit straight line models to the data collected. This is interesting because most published scientific studies are ones with statistically significant results. Given that most social scientists are only ever taught how to use techniques based on the linear model, published results will be those that have successfully used linear models. Data that fit a non-linear pattern are likely to be wrongly ignored (because the wrong model will have been applied to the data, leading to non-significant results). This is why good researchers plot their data first: plots tell you a great deal about what models should be applied to data. It is possible, therefore, that some areas of science are progressing in a biased way, so, if you collect data that look non-linear, why not try redressing the balance and investigating different statistical techniques (which is easy for me to say when I don't include such techniques in this book!)?

1.8. HOW CAN WE TELL IF OUR MODEL REPRESENTS THE REAL WORLD? ①

So far, we've seen that we use samples to estimate what's happening in a larger population to which we don't have access. We've also seen that it's important to establish whether or not a model is a good or bad fit of the data, and whether it is representative of the population. I've explained all of this using the mean as an example, but what happens when our model is more complicated? Scientists are usually interested in more complex scenarios 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?'. In these cases, we fit models that are more complex: we're actually looking at detecting effects in our population and quantifying these effects. This section explains some of the ways in which we quantify effects and decide whether they are meaningful. Essentially, as we'll see, this is a four-stage process (but see Box 1.1):

1. Generate a hypothesis (or hypotheses)—this will usually be a prediction that some kind of effect exists in the population.
2. Collect some useful data.

3. Fit a statistical model to the data—this model will test your original predictions.
4. Assess this model to see whether it supports your initial predictions.

Scientists are usually interested in testing hypotheses; that is, testing the scientific questions that they generate. Within these questions, there is usually a prediction that the researcher has made. This prediction is called an **experimental hypothesis** (it is the prediction that your experimental manipulation will have some effect or that certain variables will relate to each other). The reverse possibility—that your prediction is wrong and that the predicted effect doesn't exist—is called the **null hypothesis**. A couple of examples are:

1. *Hamburgers make you fat*: the experimental hypothesis is that the more hamburgers you eat, the more you start to resemble a beached whale; the null hypothesis is that people will be equally fat regardless of how many hamburgers they eat.
2. *Cheese gives you nightmares*: the experimental hypothesis would be that those who eat cheese before bedtime have more nightmares than those who don't; the null hypothesis would be that those who eat cheese before bedtime have about the same number of nightmares as those who don't. (In case you're interested, I frequently eat cheese before bedtime, and most of the rest of the day in fact, and I don't have that many nightmares—except when I'm writing stats books and then I dream that I'm being taunted by a mean number.)

Box 1.1

Cheating in research ①

The four-stage process I describe in this chapter works only if you follow the steps without any cheating. Imagine I wanted to place a bet on who would win the Rugby World Cup (which is happening as I type). 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 (currently 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!

(Continued)

Box 1.1 (Continued)

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 four-stage 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 that 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 one 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!)

Most of this book deals with *inferential statistics*, which tell us whether the experimental 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 and see how well it fits the data (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 accept the experimental hypothesis. Of course, we can never be completely sure that either hypothesis is correct, and so we work with probabilities instead. Specifically we calculate the probability that the results we have obtained occurred by chance—as this probability decreases, we gain greater confidence that the experimental hypothesis is actually correct and that the null hypothesis can be rejected.

To illustrate this idea, Fisher⁵ (1925) 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 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

⁵ To find out more about Fisher visit <http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/rafreader.htm>

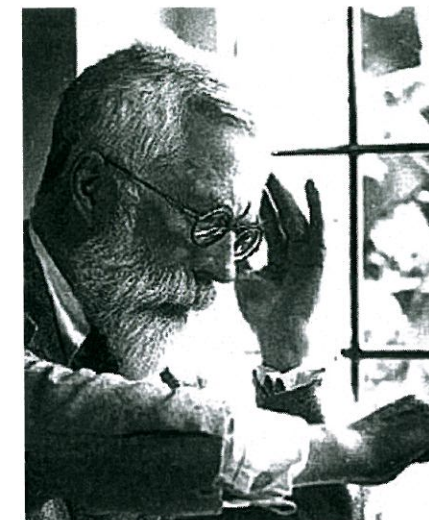


Figure 1.12 Ronald Fisher contemplating the consequences of setting $p = .05$

which the milk was added first from those in which it was added last, because most of us could perform fairly well on this task just by guessing. 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 only guess the correct order one time in 20 (or 5% of the time). If she got the correct order, we would probably be very 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 Field & Hole (2003), but 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 was added.

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 only when we are 95% certain that a result is genuine (i.e. not a chance finding) should we accept it as being true.⁶ The opposite way to look at this is to say that if there is only a 5% probability of something occurring by chance then we can accept that it is a true finding: we say it is a *statistically significant* finding (see Box 1.2 for a more specific discussion of what statistically significant actually means). This criterion of 95% confidence forms the basis of modern statistics and yet there is very little justification for it other than Fisher said so and he was a ridiculously clever bloke so we trust his judgement (Figure 1.12). Nevertheless, sometimes when I'm staring at my filing cabinet full of experimental results that were significant at only a 93% probability, I wonder how different my career would be had Fisher woken up that day in a 90% kind of a mood.

⁶ Of course, in reality, it might not be true—we're just prepared to believe that it is!

Research journals tend to have a bias towards publishing positive results (in which the experimental hypothesis is supported) and so had Fisher woken up in a 90% mood that morning, I'd have had a lot more 'successful' experiments and probably be vice-chancellor of my university by now (hmm, so maybe it's just as well for my university that he woke up in a 95% mood!).

1.8.1. Test statistics ①

So, how do we tell whether our model is a good representation of what's going on in the real world? Well, when we collect data, the data we get will vary (in fact we saw how to measure this variance in section 1.4.1). There are two types of variance (see also Chapter 7):

- **Systematic variation:** This variation is due to some genuine effect (be that the effect of an experimenter doing something to all of the participants in one sample but not in other samples, or natural variation between sets of variables). You can think of this as variation that can be explained by the model that we've fitted to the data.
- **Unsystematic variation:** This is variation that isn't due to the effect in which we're interested (so could be due to natural differences between people in different samples such as differences in intelligence or motivation). You can think of this as variation that can't be explained by the model that we've fitted to the data.

If we're trying to establish whether a model is a reasonable representation of what's happening in the population then we often calculate a test statistic. A test statistic is simply 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. An analogy used by Field and Hole (2003) is 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 (he says with a sigh of relief!). 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, say 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.

So, how do we calculate these test statistics? This depends on which statistic you're using (and as you'll discover later in the book there are lots of them: t , F and χ^2 to name only three). However, most of these statistics represent essentially the same thing:

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

The exact form of this equation changes from test to test, but essentially we're always comparing the amount of variance explained by the model we've fitted to the data against the variance that can't be explained by the model (see Chapters 5 and 7 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 always be greater than 1 (but not necessarily significant).

Given that we know how frequently different values of different test statistics occur by chance, once we've calculated a particular test statistic we can calculate the probability of getting this value. As we saw in the previous section, these probability values tell us how likely it is that our model, or effect, is genuine and not just a chance result. 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). Put another way, we accept our experimental hypothesis and reject our null hypothesis—however, Box 1.2 explains some common misconceptions about this process.

Box 1.2

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

1. **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

(Continued)

Box 1.2 (Continued)

significant just because huge numbers of people have been used in the experiment (see Field & Hole, page 60).

2. **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 experimental 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 0. As Cohen (1990) points out, a non-significant result should never be interpreted (despite the fact 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 1.6) 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!
3. **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). 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.

Box 1.2 (Continued)

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.

1.8.2. One- and two-tailed tests ①

When using statistical tests we either have a specific prediction about what will happen, such as 'the more someone reads this book, the more they want to kill its author', or we don't really know what will happen, such as 'reading more of this book could increase or decrease the reader's desire to kill its author'. The former example is directional: we've explicitly said that people will want to kill me more as they read more of this book. If we tested this hypothesis statistically, the test is called a **one-tailed test**. The second hypothesis is non-directional: we stated that the desire to kill me will change because of reading more of this book, but we haven't said whether that desire will increase or decrease. If we tested this hypothesis statistically, the test would be 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 (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 (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

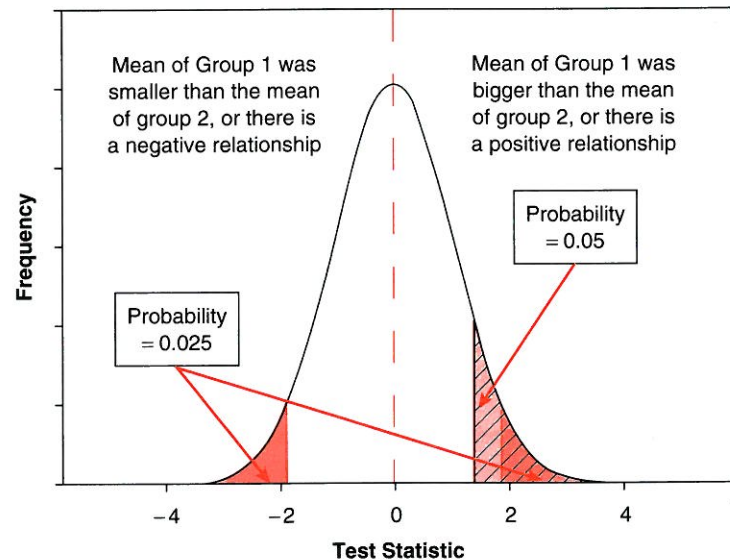


Figure 1.13 Diagram to show the difference between one- and two-tailed tests

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.

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 1.13 shows this situation—the tinted areas are the areas above the test statistic needed at a .025 level of significance. Combine the probabilities 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). Consequently, we can just look for the value of the test statistic that would occur by chance with a probability of .05. In Figure 1.13, the diagonal lines show the area above the positive test statistic needed at a .05 level of significance. The important point is 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), but if our prediction happens to be in the wrong direction then we'll miss detecting the effect that does exist! In this context, it's important to remember what I said in Box 1.1: 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.

1.8.3. 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 by chance. However, even if we're 95% confident there is still a small chance that we get it wrong. There are two mistakes we can make: a Type I and a Type II error. A **Type I error** occurs when we believe that there is a genuine effect in our population, when in fact there isn't. If we use Fisher's criterion then the probability of this error is .05 (or 5%) when there is no effect in the population—this value is known as the α -level. Assuming there is no effect in our population, if we replicated our data collection 100 times we could expect that on five occasions we would obtain a test statistic large enough to make us think that there is a genuine effect in the population even though there isn't. The opposite is a **Type II error**, which occurs when we believe that there is no effect in the population when, in reality, there is. This would occur when we obtain a small test statistic (perhaps because there is a lot of natural variation between our samples). In an ideal world, we want the probability of this error to be very small (if there is an effect in the population then it's important that we can detect it). Cohen (1992) suggests that the maximum acceptable probability of a Type II error would be .2 (or 20%)—this is called the β -level. That would mean that if we took 100 samples of data from a population in which an effect exists, we would fail to detect that effect in 20 of those samples (so we'd miss 1 in 5 genuine effects).

There is obviously a trade-off between these two errors: if we lower the probability of accepting an effect as genuine (i.e. make α smaller) then we increase the probability that we'll reject an effect that does genuinely exist (because we've been so strict about the level at which we'll accept that an effect is genuine). The exact relationship between the Type I and Type II error is not straightforward because they are based on different assumptions: to make a Type I error there has to be no effect in the population, whereas to make a Type II error the opposite is true (there has to be an effect that we've missed). So, although we know that as the probability of making a Type I error decreases, the probability of making a Type II error increases, the exact nature of the relationship is usually left for the researcher to make an educated guess (Howell, 2002, pp. 104–107 gives a great explanation of the trade-off between errors).

1.8.4. 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 Box 1.2. The first problem we encountered was knowing how important an effect is: just because a test statistic is significant

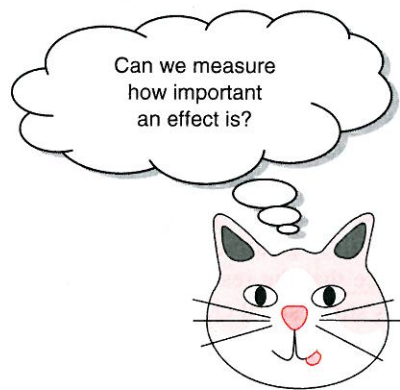
doesn't mean that the effect it measures is meaningful or important. The solution to this criticism is to measure the size of the effect that we're testing in a standardized way. When we measure the size of an effect (be that an experimental manipulation or the strength of relationship between variables) it is known as an **effect size**. An effect size is simply an objective and standardized measure of the magnitude of the 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). Many measures of effect size have been proposed, the most common of which are Cohen's d and Pearson's correlation coefficient r (see Field, 2001, and

Chapter 4). Many of you will be familiar with the correlation coefficient as a measure of the strength of relationship between two variables (see Chapter 4 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 4, 7 and 8 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).⁷

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 accepted 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.

⁷ The correlation coefficient can also take on minus values (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 experimenters coded their groups (see Chapter 4).



We can use these guidelines to assess the importance of our effects (regardless of the significance of the test statistic). However, r is not measured on a linear scale so an effect with $r = .6$ isn't twice as big as one with $r = .3$! 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.

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 (see Field, 2001).

1.8.5. Statistical power ②

We've seen that effect sizes are an invaluable way to express the importance of a research finding. The effect size in a population is intrinsically linked to three other statistical properties: (1) the sample size on which the sample effect size is based; (2) the probability level at which we will accept an effect as being statistically significant (the α -level); and (3) the ability of a test to detect an effect of that size (known as the **statistical power**, not to be confused with statistical powder, which is an illegal substance that makes you understand statistics better). As such, once we know three of these properties, then we can always calculate the remaining one. It will also depend on whether the test is one- or two-tailed (see section 1.8.2). Typically, in psychology we use an α -level of .05 (see earlier) so we know this value already. The power of a test is the probability that a given test will find an effect assuming that one exists in the population. If you think back you might recall that we've already come across the probability of failing to detect an effect when one genuinely exists (β , the probability of a Type II error). It follows that the probability of detecting an effect if one exists must be the opposite of the probability of not detecting that effect (i.e. $1 - \beta$). I've also mentioned that Cohen (1988, 1992) suggests that we would hope to have a .2 probability of failing to detect a genuine effect, and so the corresponding level of power that he recommended was $1 - .2$, or .8. We should aim to achieve a power of .8, or an 80% chance of detecting an effect if one genuinely exists. The effect size in the population can be estimated from the effect size in the sample, and the sample size is determined by the experimenter anyway so that value is easy to calculate. Now, there are two useful things we can do knowing that these four variables are related:

1. **Calculate the power of a test:** Given that we've conducted our experiment, we will have already selected a value of α , we can estimate the effect size based on our sample, and we will know how many participants we used. Therefore, we can use these values to calculate β , the power of our test. If this value turns out to be .8 or more

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

The latter use is the most 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 found on the CD-ROM, another is nQuery Adviser—see Field (1998b) for a review—but this has to be bought!). Also, Cohen (1988) provides extensive tables for calculating the number of participants for a given level of power (and vice versa). Based on Cohen (1992) we can use the following guidelines: if we take the standard α -level of .05 and require the recommended power of .8, then we need 783 participants to detect a small effect size ($r = .1$), 85 participants to detect a medium effect size ($r = .3$) and 28 participants to detect a large effect size ($r = .5$).

1.9. SOME CONCLUDING ADVICE

I just want to end this chapter with some general advice. Statistical procedures are just a way of number crunching and so even if you put rubbish into an analysis you will still reach conclusions that are statistically meaningful, but are unlikely to be empirically meaningful. There is a temptation to see statistics as some God-like way of determining the real truth, but statistics are merely a tool. My statistics lecturer used to say, quite rightly, 'if you put garbage in, you get garbage out' and to be honest I never really understood what he was going on about until people started presenting me with data sets containing millions of seemingly randomly connected variables that they wanted me to help them interpret. In a statistical analysis there is no substitute for empirical thinking! Bear that in mind throughout this book. Oh, and some final advice: never trust anything a psychologist says about statistics☺

1.10. WHAT HAVE WE DISCOVERED ABOUT STATISTICS? ①

OK, that's 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 boils 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. Now, onto SPSS!

1.11. KEY TERMS THAT WE'VE DISCOVERED

- α -level
- β -level
- Confidence interval
- Deviance
- Effect size
- Experimental hypothesis
 - Frequency distribution
 - Histogram
- Kurtosis
- Leptokurtic
- Linear model
- Mean
- Mode
- Normal distribution
- Null hypothesis
- One-tailed test
- Platykurtic
- Population
- Probability distribution
- Sample
- Sampling distribution
- Skewness
- Standard deviation
- Standard error
- Statistical power
- Sum of squared errors (SS)
- Systematic variation
- Test statistic
- Two-tailed test
- Type I error
- Type II error
- Unsystematic variation
- Variance
- z-score

1.12. SMART ALEX'S STATS QUIZ



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

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

3. What's the difference between the standard deviation and the standard error?①
4. What is a test statistic and what does it tell us?①
5. What are Type I and Type II errors?①
6. What is an effect size and how is it measured?②
7. What is statistical power?②

Some brief answers can be found in the file **Answers(Chapter 1).pdf** on the CD-ROM.

1.13. FURTHER READING

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 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.

Wright, D. B. (2002). *First steps in statistics*. London: Sage. Chapters 1, 4 and 5 are very clear introductions to sampling, confidence intervals and other important statistical ideas.