Chapter 4

Introduction to hypothesis testing

In the book so far we have seen that frequency distributions can be described by choosing appropriate statistics, usually the mean and standard deviation. Furthermore, we can compare scores from different distributions by the use of standard scores. Finally, if scores are normally distributed we can find out additional information about probability values through the use of the standard normal distribution. Now we need to see how we can exploit this information to help us answer the questions we wish our research to answer. In this chapter we move from simply describing data to seeing how we can use it to test hypotheses.

Testing an hypothesis

An hypothesis is a supposition: we state something we suppose to be the case and then collect evidence that bears upon it. For example, we are sitting talking with a group of friends about intelligence and one friend, Peter, makes the surprising claim that his 'genius' is due to being *hothoused* as a boy. Everyone laughs at this claim of genius but he continues seriously. Hothousing, he explains, is where children are provided with lots of information even before they can speak. He tells us that his mother used to show him flashcards with pictures of different types of cars, buildings, and even politicians and describe to him what they were as he gurgled back. Children have untapped potential for learning at that age that is not exploited, he argues. He even begins to get some of the sceptics to start to be swayed by his view of the development of the intellect. Everyone is now interested so we decide we want to test out Peter's claim.

To do this we need to use a procedure called hypothesis testing. This procedure underlies all the statistical tests that we shall be looking at in this book. Hypothesis testing follows a logical sequence of stages from proposing the hypothesis to deciding whether to accept or reject it.

The first problem we face is putting the hypothesis in a form that we can test. There is no genius meter that we can attach to Peter to see if he gives a genius reading. We have to find a way of expressing our hypothesis in a form that can be tested. We might decide that intelligence can be measured by the ability to solve mathematical problems or write essays on the current political situation. On this occasion we decide to operationally define intelligence in terms of an Intelligence Quotient (IQ) test. Our operational definition is a redefinition of the original concept in terms of *something we can measure*, geniuses being those people who score very highly on an IQ test. You might believe that this is a poor definition of genius (given the criticisms of IQ tests) and you may be right. I would then demand that you provide a more appropriate measure so we could continue. This problem occurs often in research, different experimenters arguing for different operational definitions. Clearly, we must use our judgement to produce a suitable definition. In this case, Peter agrees that an IQ test is an acceptable measure of his genius.

Peter's argument is that hothousing enhanced his intellect; without the hothousing he would not be so intelligent. Similarly, the rest of us who have not had the advantage of being hothoused are not as intelligent as we would have been had we had it. Therefore, the hypothesis we are testing is that being hothoused (in the way Peter was) increases IQ. This is called the research hypothesis. Note that we are being very specific here, there may be different ways of being hothoused but we are only concerned with the form that Peter experienced.

To decide whether this hypothesis is true or not all we need to do is to compare two distributions: the distribution of IQ scores for everyone without the benefit of Peter's hothousing, which I'll call the 'usual-IQ' distribution, and the distribution of IQ scores for everyone with the benefit of Peter's hothousing, which I'll call the 'hothouse-IQ' distribution. If we find that the hothouse-IQ distribution is further up the IQ scale than the usual-IQ distribution, giving a higher mean, then we can say that hothousing does increase IQ scores. (We might not know why but we have shown that it does.) In Figure 4.1 the two distributions are positioned to show an effect of hothousing resulting in an IQ enhancement of 30 points, thus, in this example, the research hypothesis is supported as hothousing shifts the usual-IQ distribution up the scale to produce the hothouse-IQ distribution.

If we found that hothousing had no effect then the hothouse-IQ distribution would be identical to the usual-IQ distribution. As a final possibility, if hothousing actually resulted in a decrease of IQ then the hothouse-IQ distribution would be lower on the IQ scale than the usual-IQ distribution (to the left of it rather than the right as in Figure 4.1). Note that we have identified three possibilities here: the hothouse-IQ distribution can be higher, the same or lower than the usual-IQ distribution. Only if we found the first of these would we accept Peter's hypothesis, whereas if either of the other two occurred we would reject it.

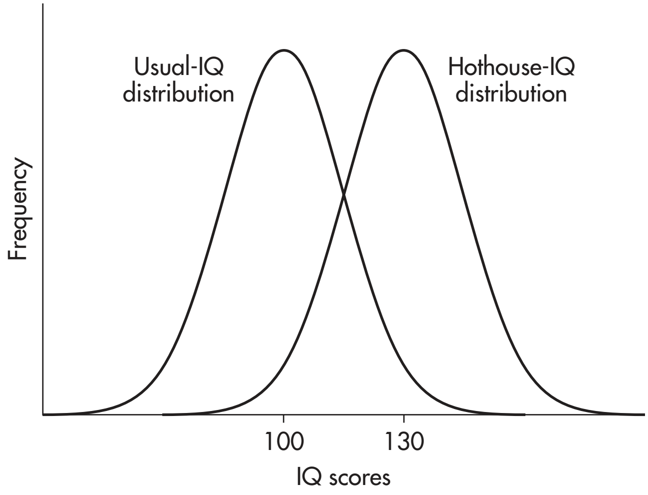

FIGURE 4.1 A hothousing effect of 30 IQ points

This is all apparently very easy but of course impossible! How are we going to find the hothouse-IQ distribution, given that this is the distribution of IQ scores for everyone after they had been hothoused as a child like Peter. The answer is we cannot. This distribution is something we simply cannot find out. Indeed, we can only find out one score from this distribution and that is Peter's score when we give him an IQ test.

Can we find out the usual-IQ distribution? It is simply too difficult to give everyone an IQ test, so what can we do? One assumption we can make is that IQ scores are normally distributed. If we do this then we will have a distribution we know a lot about. We can justify the assumption on the following grounds. First, as noted above, many human statistics are normally distributed so why not intelligence, and, second, believing this to be the case the creators of IQ tests deliberately constructed them to produce a normal distribution of scores with a mean of 100 ($\mu = 100$) and a standard deviation of 15 ($\sigma = 15$).

Note that we either have to test everyone we are interested in to find a particular distribution of scores (as in the examination example of Chapter 2) or make assumptions about the shape of the distribution. In the examination example there were only 100 scores but in many cases we are considering distributions comprising vast numbers of scores that are impossible to obtain, such as IQ scores for the adult population of the country. Hence we have to make assumptions about the distribution or else we cannot continue, and as we shall see in Chapter 5, assuming a normal distribution is often quite valid.

We now have one distribution we know about and one we do not. Unfortunately, without the hothouse-IQ distribution we are unable to test our research hypothesis. However, we are able to offer another hypothesis, the null hypothesis. The null hypothesis predicts that the two distributions are the same, that is, hothousing has no effect on IQ scores. Given that we know what the usual-IQ distribution looks like we can assume that the hothouse-IQ distribution is the same. If the null hypothesis is true then Peter's IQ score comes from the same distribution as the usual-IQ distribution.

We give Peter his IQ test and his score comes out at 120. We can find the position of this score in the distribution by finding the *z* score.

$$z = \frac{X - \mu}{\sigma} = \frac{120 - 100}{15} = 1.33$$

As we are assuming that the distributions are normal, we can look up this *z* score in the standard normal tables (Table A.1 in the Appendix) to find the probability of an IQ score higher than Peter's. A *z* score of 1.33 gives a probability of 0.0918. Given that we are assuming the distributions are the same, this means that 9.18 per cent of the usual-IQ distribution, who have not been hothoused, score higher than Peter who had. Can we use this evidence to support the null hypothesis that the distributions are the same or does the evidence supoprt the view that the distributions are different and Peter is from a distribution higher than the usual-IQ distribution? The fact that over 9 per cent of the usual-IQ population have higher IQ scores than Peter's doesn't convince me of the effect of hothousing. I would expect geniuses to be rarer than 9.18 per cent which is equivalent to 1 person in every 11 from the usual-IQ distribution scoring higher than Peter. On this evidence I accept the null hypothesis and say that we have not found evidence to support Peter's view of hothousing.

Now imagine that Peter had scored 145 instead of 120. This gives a *z* score of 3 and a probability of 0.0013 of a score higher than Peter's. This means that only 0.13 per cent of the usual-IQ population score are better than Peter. This small percentage, 0.13 per cent, tells us that only 1 person in every 769 from the usual-IQ distribution scores higher than Peter. On this evidence, if the two distributions are the same Peter is very unusual indeed. A score as high as Peter's score is so rare in the usual-IQ distribution that it seems more likely that it belongs to a different, higher, distribution. Here, the chances are that the null hypothesis is false. So I reject the null hypothesis and accept the hypothesis that Peter's score comes from a hothouse-IQ distribution higher up the IQ scale than the usual-IQ distribution.

Thus, hypothesis testing is a gamble on the basis of probabilities. If the probability of Peter's score coming from a distribution the same as the usual-IQ distribution is very low I reject the null hypothesis, if the probability is not very low I accept it.

If I accept the null hypothesis when the probability is 0.0918 and reject it when the probability is 0.0013 then where is my dividing line, at which probability do I switch from acceptance to rejection? The answer is: where ever I choose! However, it has been agreed for reasons discussed in Chapter 9, to conventionally reject the null hypothesis when the probability is less than or equal to 0.05 (written as: '$p < 0.05$' or 'significant at $p = 0.05$'). This means that when a score from the unknown distribution could only arise from the known distribution (i.e. the distributions are the same) with a chance of less than 5 times in 100 then we reject the null hypothesis and say that the score really does come from a different distribution. Essentially we are gambling on the probability that a score (such as Peter's IQ) comes from a unknown distribution (hothouse-IQ) identical to the known distribution (usual-IQ). When the chances are 1 in 20 or less (that is, a probability of 0.05 or less, as 1 divided by 20 = 0.05) we switch our gamble and bet that the distributions are different. Thus, the probability of 0.05 is called the significance level. If the probability of Peter's score is greater than or equal to the significance level we accept the null hypothesis and if it is lower than the significance level then we reject the null hypothesis.

The significance level of 0.05 means that we are more than 95 per cent certain that we are correct in accepting that the distributions are different. We are allowing ourselves to get it wrong, and claim there is a difference in the distributions when there is not, on 5 per cent or fewer occasions, as such an extreme score could only arise by chance (i.e. come from a distribution identical to the known distribution) 5 per cent or less of the time. Sometimes we want to be even more certain that we are correct in claiming a difference between the distributions. In these cases we take the significance level of $p = 0.01$, accepting only 1 chance in 100 or less that we have got it wrong. With this level of significance we can be 99 per cent or more certain that we have made the right choice in claiming a difference in the distributions.

A summary of the hypothesis testing

We tested the hypothesis that the hothousing Peter received produced his genius by the following steps:

1 We chose IQ as a measure of performance on which intelligence could be judged. This is our operational definition.
2 We set up a research hypothesis: hothousing of the form Peter experienced increases a person's IQ.
3 We set up the null hypothesis: hothousing of the form Peter experienced does not affect a person's IQ.
4 We cannot test the research hypothesis as we do not know both distributions. We can test the null hypothesis as we know the usual-IQ distribution and the null hypothesis assumes that the unknown hothouse-IQ distribution is the same.
5 We gave Peter the IQ test and obtained his score.
6 We worked out the probability of a score as high or higher than Peter's from the usual-IQ distribution by looking up the z score in the standard normal distribution table. We can only do this because we have assumed that the usual-IQ scores are normally distributed.
7 If the probability of a score as high or higher than Peter's is very small, smaller than the significance level, then we say that it is very rare for a score as high as Peter's score to come from a distribution the same as the known usual-IQ distribution and we reject the null hypothesis, concluding that the hothouse-IQ distribution is different, higher up the IQ scale. If the probability is not smaller than the significance level then we accept the null hypothesis and do not conclude that there is a difference in the distributions.

The logic of hypothesis testing

Despite the variety of statistical tests that we examine in this book they all follow the same basic logic. A research hypothesis predicts a difference in distributions whereas a null hypothesis predicts that they are the same. If we have the details of the two distributions we simply compare them. Usually we do not have these details. However, we can continue the analysis when one of the distributions is known and one unknown. One is known because we are able to make the assumption that it is normally distributed and we know about normal distributions. We select a significance level. This is our decision criterion for accepting or rejecting the null hypothesis. This is conventionally set at $p = 0.05$ or $p = 0.01$. We select the significance level before we collect the data. It is like betting on a horse race. We don't place a bet until we know the odds. We collect the data that provides a score from the unknown distribution. We look up the probability of a score such as this arising from the known distribution to decide whether to accept the null hypothesis and conclude that the distributions are the same. If the probability is lower than the significance level we reject the null hypothesis and say that the chances favour the score coming from a different distribution to the known distribution. If the probability is not lower than the significance level then we accept the null hypothesis.

One- and two-tailed predictions

Hypothesis testing is about deciding whether an unknown distribution is the same or different to a known distribution. There are three possible arrangements of the two distributions:

1 The unknown distribution is the same as the known distribution.
2 The unknown distribution is higher up the scale than the known distribution.
3 The unknown distribution is lower down the scale than the known distribution.

We always test the null hypothesis (1, above) that the distributions are the same but our research hypothesis can take a number of different forms. Our research hypothesis could predict 2 (above). In fact this was the prediction about the hothousing-IQ distribution, that it was higher up the scale than the usual-IQ distribution. Alternatively, we might predict 3, that the unknown distribution is lower than the known distribution. Imagine another friend David had a serious head injury through a car accident. In this case we might predict that this type of injury leads to a lower IQ than would have been achieved without it. Finally, there are occasions when we predict either 2 or 3. Here we are predicting that the unknown distribution will be different to the known distribution but leaving open the possibility that it will be higher or lower. A third friend Susan grew up eating her grandmother's special diet. We might predict that this diet affected her intellectual performance. However, we might not be sure whether to predict that the special diet improves IQ (maybe Susan was getting just the right mix of foods for intellectual growth) or reduces IQ (maybe Susan was missing out on important vitamins).

In the hothousing and brain injury examples we are predicting a direction to the difference in the distributions as the research hypothesis is stating in which direction along the scale the unknown distribution will be shifted in relation to the known distribution. These predictions are called one-tailed predictions. If you look back to Figure 4.1 you can see that the hothouse distribution is expected to overlap with only the higher end of the usual-IQ distribution, only one tail of the known distribution. If the hothouse-IQ distribution turned out to be the same as the usual-IQ distribution or even resulted in lower IQ scores then our hypothesis would not be supported. Only if the distribution is at the one-tail we are interested in, the upper end of the usual-IQ distribution, is our hypothesis supported (as in Figure 4.1). We infer this by observing whether Peter's IQ score occurs so far into the end of the upper tail (the top 5 per cent) of the usual-IQ distribution that we can claim that his score comes from a different distribution, higher up the scale.

The brain injury example is also a one-tailed prediction as it follows the same logic as the hothouse example, but here we are interested in the lower tail of the known distribution. Only if David's IQ falls into the bottom 5 per cent of the usual-IQ distribution would we accept the hypothesis that the brain-injury-IQ distribution is lower than the usual-IQ distribution.

The diet example is a two-tailed prediction as we are hedging our bets, we are saying that Susan's diet might have reduced her IQ or enhanced it. The diet-IQ distribution could overlap the lower tail of the usual-IQ distribution or the higher tail, either outcome supports our hypothesis of a difference in distributions. Only if the two distributions are the same do we accept the null hypothesis.

There are many instances where we are unable to make specific directional, one-tailed predictions. For example, in an experiment on stress and job satisfaction we might predict that a certain type of stress reduces job satisfaction as it produces anxiety. However, it could also increase job satisfaction if it results in interest and excitement. Where there is not enough evidence to decide which hypothesis to follow, the experimenter might decide to do a two-tailed test first of all, to see whether this type of stress has any effect at all, be it positive or negative. In this case any difference in the distributions would support the hypothesis.

Significance level and two-tailed predictions

When we undertake a one-tailed test we argue that if the test score has a probability lower than the significance level then it falls within the tail-end of the known distribution we are interested in. We interpret this as indicating that the score is unlikely to have come from a distribution the same as the

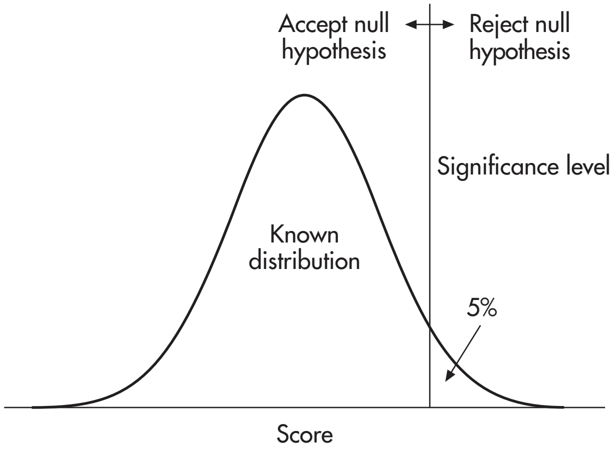

FIGURE 4.2 A one-tailed prediction and the significance level

known distribution but from a different distribution. If the score arises anywhere outside this part of the tail cut off by the significance level we reject the research hypothesis. This is shown in Figure 4.2. Notice that this shows a one-tailed prediction that the unknown distribution is higher than the known distribution. As an exercise try drawing this figure for a one-tailed prediction where the unknown distribution is predicted to be lower than the known distribution. (When you have tried this, look at Figure 6.1, which shows a prediction of this kind.)

With a two-tailed prediction, unlike the one-tailed, both tails of the known distribution are of interest, as the unknown distribution could be at either end. However, if we set our significance level so that we take the 5 per cent at the end of each tail we increase the risk of making an error. Recall that we are arguing that, when the probability is less than 0.05 that a score arises from the known distribution, then we conclude that the distributions are different. In this case the chance that we are wrong, and the distributions are the same, is less than 5 per cent. If we take 5 per cent at either end of the distribution, as we are tempted to do in a two-tailed test, we end up with a 10 per cent chance of an error, and we have increased the chance of making a mistake.

We want to keep the risk of making an error down to 5 per cent overall (our fixed amount of risk) as otherwise there will be an increase in our false claims of differences in distributions which can undermine our credibility with other researchers, who might stop taking our findings seriously (one mustn't cry wolf too often!). When we gamble on the unknown

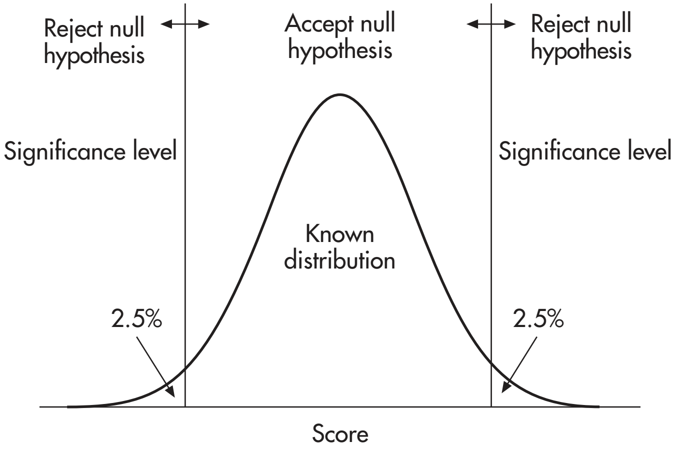

FIGURE 4.3 A two-tailed prediction and the significance level

distribution being at either tail of the known distribution, to keep the overall error risk to 5 per cent, we must share out our 5 per cent between the two tails of the known distribution, so we set our significance level at 2.5 per cent at each end. If the score falls into one of the 2.5 per cent tails we then say it comes from a different distribution. Thus, when we undertake a two-tailed prediction the result has to fall within a smaller area of the tail compared to a one-tailed prediction, before we claim that the distributions are different, to compensate for hedging our bets in our prediction. This is shown in Figure 4.3.

Hypothesis testing, as described here, where we are using a chosen significance level to make our decision is often referred to as significance testing. Whether we perform a one-tailed or a two-tailed test, the decision to reject (or not to reject) the null hypothesis depends on which side of the significance level our score falls. Significance testing has been extremely useful in analysing research findings, as I hope you appreciate from the example of Peter's 'genius' above. However, we need to be aware of its advantages and limitations, and these issues will be examined on page 71 and in Chapter 9.

Chapter 5

Sampling

chapter 5

Populations and samples

In the book so far we have been looking at what we call populations, that is the complete set of the things we are interested in. The frequency distributions have included all the scores we are interested in, such as the scores of all one hundred students who took the examination this year, the example from Chapter 2. A population need not be a collection of people, even though we are used to hearing the term used in this way, such as the population of Britain. A population can be a complete set of anything. In statistics it refers to a complete set of scores, such as the number of pages of each book in a library, the IQ scores of fifteen year old girls living in London, the number of goals scored in each football league match on a particular Saturday, the times to complete a jigsaw by members of the Robinson family, the number of food pellets eaten by each rat in an animal learning experiment. The population is simply every member of the particular category we wish to study.

Often, through the sheer size of the population we cannot study it all. In this case we select a sample. A sample is a subset of a population. Usually, we want to know about populations rather than samples yet we are almost always only able to test samples. This is the fundamental problem of statistical analysis. When and how can information from a sample give us information about a population? The following sections will deal with this key question. But first an example to illustrate the difficulty.

A doctor wishes to know the incidence of respiratory problems in British men over the age of 50 years. This is a large population and extraordinarily difficult to test them all. A sample must be tested instead. But the doctor is not interested in the sample *per se* but what it tells him or her about the population. If it is not possible to estimate details of the population from the sample it is not worth studying it. What this doctor, and researchers in general need to find is sample information that is useful for estimating details of the population.

Selecting a sample

One of the difficulties of using samples to represent populations is the selection of sample members. In most cases we want our sample to truly represent the population so we can generalise our findings to the population and claim the population will perform like the sample. If we have a sample with the same characteristics as the population we will have a representative sample. If the characteristics of the sample are different to those of the population then any findings based on the sample could be biased and not be generalisable to the population. Opinion pollsters will sometimes try to get a representative sample of the voting population to question, making sure that they have, for example, the same proportion of men and women in the sample as there are in the population.

Consider the example of respiratory problems. Most people would agree that a sample of men under 50 or a sample of women over 50 is clearly not representative of the population we wish to generalise to. However, will any group of men over 50 be acceptable? If we took all our men from a hill top village where the air is clear, or from a coal mining town polluted with coal dust we are likely to have a biased sample, as not all members of the population live in a hill top village or a coal mining town. We would need to take the sample from a range of locations, or from a place where there is not a specific bias due to the location. We would need to consider age as well. If our sample contained only men between 50 and 60 years old could we generalise to a population where there are many men older than 60 in the population?

Any difference between the sample and the population could lead to a problem of generalisation: location, age, occupation, class, whether they smoke or not and so on. It is almost impossible to obtain a truly representative sample, where every characteristic of the sample matches the population characteristics. Rather than giving up, researchers do the best they can with the available resources and try to be aware of any differences between the sample and population. Here the judgement is not entirely statistical but also depends on the researcher's expertise in the subject. A medical practitioner will know that certain factors are important with respect to respiratory problems, so will try to select a sample representative of the population on these key factors, such as whether the person is a smoker or not, but not on factors unlikely to be relevant to the study, such as a person's hair colour. It requires the professional judgement of the researcher (rather than statistical knowledge) to make the decision on which characteristics the sample must match the population on and which factors can be ignored.

An alternative way of selecting a sample to represent a population is through random selection. With a random sample the sample members are selected at random from the entire population, with each member of the population having an equal chance of being selected. If I take 100 ping pong balls and write the numbers 1 to 100 on them, put them in a sack, shake them up, then take five out without looking, I have a random sample of five numbers from the population of numbers 1 to 100. Similarly, if I am doing a survey, I might select names at random from the telephone directory to select people to send the survey to. I have no idea who those people will be, I am leaving it up to chance. By random selection I am not deliberately biasing my sample, so any differences between the sample and the population should be random and, therefore, not systematically influencing my data in any way.

However, even so-called random sampling might not be quite so random after all. If I am randomly selecting passers by in the street for a survey, I am excluding all those people not passing by. If I perform my survey at 3 pm then I will not get anyone whose occupation keeps them at work at this time. I may not have a random selection of the population I am interested in. Selecting numbers at random from a telephone book excludes all those people not listed in the directory. If my population is ‘people listed in the telephone book’ then it is fine, otherwise I need to be careful. Often it is hard to collect a truly random sample of the population we are interested in but, once again, we must do the best we can by deciding on the key factors and selecting randomly within these factors.

In many cases it is not possible to be truly representative or random but a good researcher will make it clear how the sample was selected so that other researchers can decide if there was a systematic bias on an important factor. Finally, there are a couple of useful points concerning a pragmatic approach to sampling that many researchers adopt.

1 This is the only sample I have, or am able to test, so even though there may be sampling problems I’ll test the sample anyway. If the results are interesting I can investigate further, aware of the potential difficulties in sampling.

It is called an opportunity sample when we simply select an available sample. There are many experiments in psychology that use samples of psychology students, who may not be representative of people in general. However, often they are available for testing and if it turns out that something intriguing comes up then other non-student samples can be tested as well. Furthermore we might decide that there is no serious reason to assume that the students will perform differently to the general population on this experiment.

2 If I don't find what I am interested in with a sample biased in my favour then it is not worth spending more resources finding a more representative sample.

If I am testing the hypothesis that people in Britain prefer the television to radio I might deliberately be perverse and select a group of people who have just bought a new radio. One might expect these people to be more favourable to radio than the general population and if I found that they preferred radio it would not be surprising. However, if I found that even these people preferred the television despite my bias in favour of radio in the sample selection then it is not unreasonable to infer that the general population would also prefer the television.

Sample statistics and population parameters

Statistics and parameters

At this point is worth explaining some terminology. To make the distinction between sample details and population details, the word statistic is used to refer to a sample figure and parameter for the population figure, so the sample mean is a statistic but the population mean is a parameter. (In the earlier chapters I have referred to a 'statistic' when I really should have been using the term 'parameter'. I did this because we are all familiar with the term statistic but not parameter. It is only at this point in the book that I believe the distinction should be made.) The term parameter for population characteristics explains why the tests we shall be looking at until Chapter 16 are referred to as parametric tests. In these tests we use sample statistics as estimators of population parameters. The two most important of these sample statistics are the sample standard deviation and the sample mean.

Sample standard deviation

Of the various measures of spread the mean absolute deviation and the standard deviation both use information from all the scores. However, it has been found that the sample mean absolute deviation is an unstable estimator of the population figure, that is, there is no consistent relationship between the sample statistic and the population parameter. On the other hand the standard deviation of a sample is a much more reliable estimator of the population value. Because of this, when we do not know the population standard deviation, we can use the sample standard deviation to estimate it. This is a key reason for the preference for the standard deviation in statistical analysis.

The formula for a standard deviation of a population was given in Chapter 1 and was designated by the symbol σ. However, if we apply that formula to the sample scores we end up with a sample standard deviation that underestimates the population value. To improve the estimate we change the formula and always calculate a sample standard deviation by the formula:

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}}$$

Notice we use 's' rather than σ to indicate it is a sample standard deviation rather than a population standard deviation. We also use the lower case 'n' for the sample size (the number of scores in the sample) and $\overline{X}$ for the mean of the sample (to distinguish it from the population parameter μ).

The reason why we use $n - 1$ instead of n in the formula is a little complicated but it helps when we consider the different purpose of the sample and population standard deviations. In the latter case we are simply seeking an average deviation and divide by the number of scores N. In the former case we are seeking a good estimate rather than an average. This estimate is more accurate when it is based not on the number of scores but on the degrees of freedom, $n - 1$. Degrees of freedom concern the scores that contain new information. As we have calculated the sample mean from the sample scores we have used up some of the information in the scores. The number of scores with new information, the degrees of freedom, is $n - 1$.

A simple example illustrates this fact. If I have a sample of four scores ($n = 4$) with a sample mean of 5, how many scores must I tell you before you can work out the rest? With 4 scores and a mean of 5 the total of the scores is 20. If we label the four scores as X_1, X_2, X_3, and X_4 then:

$$X_1 + X_2 + X_3 + X_4 = 20$$

I tell you that one score is 6, $X_1 = 6$, this gives us:

$$6 + X_2 + X_3 + X_4 = 20$$
$$X_2 + X_3 + X_4 = 14$$

The other three scores could be any three numbers that add up to 14, there is some freedom in what they could be. I now tell you that another score is 4, $X_2 = 4$:

$$4 + X_3 + X_4 = 14$$
$$X_3 + X_4 = 10$$

It is still not certain what the other two scores are, they still have some freedom, although now you know they add up to 10. The third score is 2, $X_3 = 2$. Given this information you can work out that the fourth score must be 8:

$$2 + X_4 = 10$$
$$X_4 = 8$$

There is no freedom for this last score to vary. The final score can only be 8 because we know that the mean is 5. As we started with the knowledge of the sample mean then only three ($n - 1$) of the scores give us any new information, so there are only three ($n - 1$) degrees of freedom in this sample.

In words, the sample standard deviation is the square root of the sums of squares divided by the degrees of freedom. We shall meet these terms often in our statistical analyses. The sums of squares, $\sum(X - \overline{X})^2$, requires us to calculate the sample mean first. However, we know that $\overline{X} = \dfrac{\sum X}{n}$ (which is the formula for the sample mean – add up all the scores in the sample and divide by the sample size). If we replace $\overline{X}$ by $\dfrac{\sum X}{n}$ in the sums of squares formula we end up with an equivalent formula for the sample standard deviation that does not require us to calculate the mean first:

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}}$$

In the formula $\sum X^2$ refers to the sum of the squared scores (we square each of the scores first then add them up), whereas $(\sum X)^2$ refers to the square of the sum of the scores (we add up the scores before we square the total).

Notice that dividing by the degrees of freedom, $n - 1$, rather than the sample size, n, makes less difference when the sample size is large but has a much larger effect when the sample size is small. Dividing by 99 rather than 100 will not change the calculation very much compared to dividing by 9 rather than 10. As we see below, small samples are not as good for estimating population values as large samples.

Sample mean

We also want to know what a central figure is in the population but when we only have a sample, rather than details of the population, we have to estimate it with a statistic from the sample. Of the various measures of central tendency (mode, median, mean), the sample mean is the best estimate of the population value, again for reasons of stability. But how good an estimate of μ is the sample mean $\bar{X}$? It depends a lot on the size of the sample, the larger the sample the better the sample mean is as an estimate of the population mean. It also depends on the specific sample that we pick. We can see this in the following example.

The population of IQ scores is normally distributed with a mean of 100 and a standard deviation of 15. If we took a sample of 20 people's IQ scores would our sample mean be 100? The answer is probably not. The reason is that we might have a sample with a number of clever people in it and so the sample mean would be higher than 100. Alternatively if we had some less able people in the sample the mean would be lower. So sample means will have a range of different values dependent on the scores we select for our sample.

Imagine for a moment that we were able to select every possible sample of 20 IQ scores and work out their sample means: what range of values would we get and with what frequency? What would be the mean of all these sample means?

So far we have only looked at the frequency distributions of scores, but now we are interested not in the individual scores but in the mean of every sample of size 20. If we plot this information as a frequency distribution, the curve determined by the number of sample means at each value, we get the distribution of sample means. It turns out that the distribution of sample means has some very interesting and useful characteristics.

First, we find that, as we obtain more samples, the mean of the sample means gets closer to the population mean. When we have selected all possible samples we find that the mean of sample means is the same as the population mean. Thus, if we collect the means of samples of 20 IQ scores, then the mean of all the sample means will be 100. We refer to the mean of sample means by the symbol $\mu_{\overline{X}}$. We use the Greek letter μ to show that it is still a population mean and the subscript $\overline{X}$ to show that it is the mean of a population of sample means.

Second, the distribution of sample means will tend to be a normal distribution. If the population of scores is normally distributed then the distribution of sample means will definitely be normally distributed. Even if the population of scores is not normally distributed the distribution of sample means will still look rather like a normal distribution with a hump in the middle and tails off to either side. The larger the samples we select the closer the distribution approaches the normal distribution. This can be shown by a mathematical proof, called the central limit theorem. Even though the distribution of scores is not normally distributed, the distribution of sample means will end up as a normal distribution as long as the samples are large enough. When the sample size is 30 or more the sampling distribution is almost exactly a normal distribution, regardless of whether the original distribution was normally distributed or not. This is an extremely useful piece of information for our statistical analysis as we now see.

Third, as the distribution of sample means is either normally distributed or approximately normally distributed, we can work out the probability of finding a sample with a particular mean value by calculating a z score for our sample mean and looking up the probability in the standard normal distribution tables.

Finally, we can easily work out the standard deviation of the distribution of sample means by a simple formula using the standard deviation of the individual scores. We call this new standard deviation the standard error of the mean and refer to it by the symbol $\sigma_{\overline{X}}$. The standard error provides us with the standard deviation of a sample mean from the population mean.

$$\text{Standard error, } \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and n is the sample size.

The standard error of the mean is precisely that, the standard distance, or error, that a sample mean is from the population mean. In our statistical tests we want to know how good an estimate the sample mean is of the population mean. The standard error tells us just that. Notice, as the sample size (n) gets larger so the standard error gets smaller. Again this illustrates that larger samples give better estimates of the population mean than smaller samples.

The distribution of sample means turns out to be something we now know a lot about without having to laboriously calculate means for all the samples. The distribution of the sample means will be a normal distribution (or similar to it) with a mean, $\mu_{\bar{X}}$, the same as the population mean, μ, and a standard deviation, $\sigma_{\bar{X}}$, the standard error of the mean, equal to the population standard deviation divided by the square root of the sample size.

In the IQ example the distribution of sample means for samples of 20 scores will be a normal distribution with a mean of 100 and a standard error of $\frac{15}{\sqrt{20}}$, which is 3.35. As we have a normal distribution and we know its mean and standard deviation we can calculate z scores and work out probability values, just as we did for a score and a population in previous chapters, but now we can do it with a sample mean and a population of sample means (the sampling distribution of the mean).

Summary

To recap, we want to know about populations rather than samples but usually we can only test samples. We want our sample to tell us about the population. We therefore have to be careful in selecting our sample because we would like to generalise from the sample to the population.

The sample mean and the sample standard deviation are the best estimates of the population parameters but we use degrees of freedom rather than sample size in calculating them as that improves their estimation. Larger samples provide better estimations of population figures than smaller samples. Degrees of freedom make more of a difference to the estimation when the sample size is small than when it is large.

We can compare our sample to the population by calculating the sampling distribution of the mean. This tells us what the distribution of sample means would look like if we took every sample the same size as our own (n) from the population and worked out their means. The sampling distribution of the mean turns out to be a distribution we know about because it is almost certainly normally distributed and has a mean the same as the population mean and a standard deviation, the standard error of the mean, equal to the population mean divided by the square root of the sample size.

As the distribution is normal and we know its mean and standard deviation we can calculate *z* scores and work out probability values. This is exactly what we need for hypothesis testing.

We shall see in the following chapters how the distribution of sample means is extremely useful to hypothesis testing when we consider a sample rather than a single score.

Chapter 6

Hypothesis testing with one sample

An example

There was a leak of the gas Cyadmine[4] at a chemical works and the gas cloud hung over the town of Newtoncastle for a number of days before dispersing into the atmosphere. There were some complaints of sore throats amongst the townspeople but the chemical company assured the public that there are no adverse effects of Cyadmine on the human body. However, a scientist who worked on the Cyadmine project has gone on record as saying that Cyadmine could have an effect on pregnant women and their unborn children. The company has dismissed the scientist's claim as nonsense, noting that the scientist was unable to specify what problems could arise. There is not a universal confidence in the chemical company and there is some concern in the affected areas especially from parents of young children. A doctor in the large maternity hospital has been keeping an eye on babies born in the nine months after the cloud passed over the town. She has noted that the babies appear healthy on all the usual checks but is suspicious that the Cyadmine could have affected their birth-weights as many of the babies appear rather small at birth. The doctor is worried about any long-term effects and wants to test whether the 'Cyadmine babies' are smaller at birth than usual. Essentially, the doctor is making a one-tailed prediction: the distribution of the birth-weights of the Cyadmine affected population will be different to the distribution of the birth-weights of the unaffected population with the overlap of the distributions occurring at the lower end of the unaffected distribution.

To test this hypothesis we need details of the two birth-weight populations. Comparing the two distributions will tell us whether there is a difference between the two, specifically whether the mean of the Cyadmine-affected population is lower on the birth-weight scale than the unaffected population. The problem is collecting the details of the two populations.

We may be lucky here, in that medical records are very detailed and let us assume in this case that there are detailed records of birth-weights. We find from the records that, for babies born in this country, the mean birth-weight is 3.2 kg and the standard deviation is 0.9 kg. These are the details we take for the population unaffected by Cyadmine.

The problem now is to collect details of the Cyadmine-affected population. Essentially what we want to know is how the unaffected population would be affected by Cyadmine were they to be affected by it, as the doctor's prediction is that the effect of Cyadmine is to shift everyone's birth-weight down the scale by a fixed amount. We can never get details of this population, all we have are the babies of Newtowncastle who were in the womb at the time of the leak. This is only a sample of the second population. Not only that, but our sample is not necessarily representative or random. We are unable to select freely from the Cyadmine-affected population. Our sample could be influenced by other factors as well as, or instead of, Cyadmine, such as a hospital inducing babies early, which might also lead to lower birth-weights.

We decide to select one hundred of these babies, balancing home births and hospital births, selecting a range of foetal ages when the cloud appeared, and so forth, to try to select a sample that will not be systematically influenced by factors such as hospital practice, foetal age, etc. We may not be able to account for all systematic differences, bar the Cyadmine effects, between the sample and the unaffected population but we can do our best to control for key confounding variables (see Chapter 7 for further explanation of 'confounding'). If we do find differences between Cyadmine babies and unaffected babies it will be worth investigating further to ascertain whether it is really due to Cyadmine or some other reason. If we find no difference we might decide we need investigate no further.

We obtain the birth-weights for the sample of Cyadmine babies and calculate the sample mean. This turns out to be 3.0 kg. Can we compare this mean with the population mean for the unaffected babies? The answer is no, because we are not comparing like with like and this allows for the possibility of bias. To explain this, let us consider the unaffected population for a moment. Not all babies have the same birth-weight, some will be lighter than others due to the normal spread of birth-weights. It is quite possible that if you selected a sample of unaffected babies you might find their sample mean lower than the population mean. By chance we might have selected a group of babies with relatively low birth-weights despite the fact that they come from a population with a higher mean birth-weight – we could have just selected small babies. (I'm sure that you can see that, equally, by chance, we might select a sample with a mean birth-weight higher than the population mean.) Even though our sample of Cyadmine-affected babies gave a sample mean lower than the unaffected population mean, we cannot take this as evidence for the effect of Cyadmine on birth-weight. It might not be due to a difference in populations but simply due to the nature of sampling.

If we cannot compare our sample mean with a population mean what can we do? Recall that we can compare a score with a population of scores, so we need to compare a sample mean with a population of sample means. If we select all possible samples of size 100 from the unaffected population and work out their sample means we can create a distribution of sample means. In this way we are creating a 'known distribution', the distribution of the mean for samples of size 100 from the unaffected population and an 'unknown distribution', the distribution of the mean for samples of size 100 from the affected population. Now we can compare these two populations of sample means. If they are different, with the affected distribution having a smaller mean, this will support the doctor's hypothesis. Unfortunately, we don't have the details of these populations yet, in fact we only have one value from the unknown population: the mean of our sample of 100 affected babies.

Do we have details of the distribution of sample means for samples of size 100 from the unaffected population? Here the answer is yes. Fortunately, as we saw in the previous chapter, we don't need to go out and select every possible sample of size 100 from the unaffected population as we know about sampling distributions – with a sample size greater than 30 the distribution of sample means will almost certainly be a normal distribution. Also, the mean of a sampling distribution, $\mu_{\overline{X}}$, is the same as the population mean, μ, so it will be 3.2. And the standard deviation of the sampling distribution, $\sigma_{\overline{X}}$, the standard error, will be the population standard deviation ($\sigma = 0.9$) divided by the square root of the sample size ($n = 100$), so will be $\dfrac{0.9}{\sqrt{100}} = 0.09$.

We have now created a logically identical framework for hypothesis testing to the one we had in Chapter 4. We have a 'score' from an unknown distribution, in this case our affected sample mean of 3.0, and we have a known distribution, the sampling distribution of unaffected samples of the same size. The distribution is known to be normally distributed with a mean of 3.2 and a standard deviation of 0.09. All we need to do is choose a significance level for the doctor's hypothesis, find the z score, look up the probability and make our decision as to whether the affected sample comes from the same distribution as the unaffected samples or a lower one.

We find out how likely it is to get a sample of 100 unaffected babies with a sample mean of 3.0 by working out the z score. Recall that a z score is a score minus a population mean divided by the population standard deviation. Here the sample mean $\overline{X}$ is our 'score', the mean of the sampling distribution, $\mu_{\overline{X}}$, and standard error, $\sigma_{\overline{X}}$, are the mean and standard deviation of the distribution we are interested in, so

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3.0 - 3.2}{\frac{0.9}{\sqrt{100}}} = \frac{-0.2}{0.09} = -2.22$$

We can look up the probability of the z score in the standard normal distribution tables as our sampling distribution is normally distributed. Remember the minus sign simply tells us that the score is lower than the mean of the distribution. From the Table A.1 in the Appendix a z score of 2.22 gives a probability of 0.0132. Thus, the probability of obtaining a sample mean as low or lower than 3.0 kg from a sample of 100 unaffected babies is only 0.0132. This is well within the bottom 5 per cent of the unaffected sampling distribution, well below the significance level of $p = 0.05$. We can conclude that a sample mean of 3.0 kg is so rare in the unaffected population that our affected sample mean of 3.0 kg indicates that the affected distribution is not the same as the unaffected distribution, and we reject the null hypothesis, concluding that Cyadmine-affected babies do have a lower birth-weight than unaffected babies. This is shown graphically in Figure 6.1.

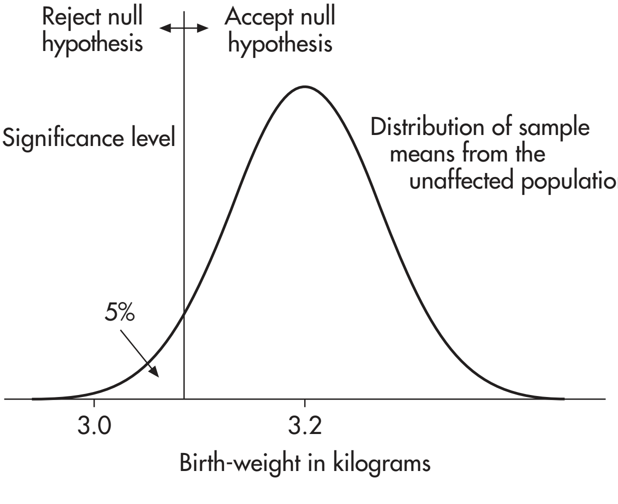

FIGURE 6.1 Hypothesis testing with a sample of Cyadmine-affected babies

In summary

When we have a sample from an unknown population we cannot compare it to a known population. We must find the sample mean, $\overline{X}$. Then we find the sampling distribution of the mean for all samples of the same size from the known population. This distribution is usually a normal distribution with a mean, $\mu_{\overline{X}}$, equal to the population mean μ, and a standard deviation (or standard error), $\sigma_{\overline{X}}$, equal to the population standard deviation, μ, divided by the square root of the sample size, $\sqrt{n}$.

Using this information in our example we tested the hypothesis that the unknown distribution is lower on the scale than the known distribution. As the known distribution is a normal distribution we worked out a z score to find the probability of finding a sample mean from the known distribution as small or smaller than the sample mean from the unknown distribution. As the probability was smaller than the significance level we rejected the null hypothesis and concluded that the unknown distribution is lower on the scale than the known distribution: Cyadmine-affected babies do have a lower birth-weight than unaffected babies.

When we do not have the known population standard deviation

The average number of purchases in a supermarket is 25 items. The company would like to increase this figure and introduces an advertising campaign to encourage shoppers to buy more products in the store. In the week after the campaign a sample of 50 shoppers are tested to see if the number of purchases has increased.

The following number of purchases were recorded:

30	44	19	32	25	30	16	41	28	45
28	20	18	31	15	32	40	42	29	35
34	22	30	27	36	26	38	30	33	24
15	48	31	27	37	45	12	29	33	23
20	32	28	26	38	40	28	32	34	22

The mean number of purchases for this sample is 30 items and the sample standard deviation is 8.43.

Has the advertising campaign had an effect? As we saw above we cannot compare the sample mean of the post-advertisement shoppers

(30 items) with the population mean of the pre-advertisement shoppers (25 items) as one is a sample and the other a population. To compare a sample mean with a distribution of sample means we must calculate the sampling distribution of samples of size 50 from the pre-advertisement shoppers. This distribution has a mean of $\mu_{\bar{X}} = 25$ ($\mu_{\bar{X}} = \mu$, the same as the population mean) and a standard error of $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{50}}$, where σ is the standard deviation of the pre-advertisement population.

Our sampling distribution is almost certainly normally distributed so we can look up a z score in the standard normal distribution table to find the probability of finding a sample mean as large as 30 from the pre-advertisement shoppers.

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \frac{30 - 25}{\dfrac{\sigma}{\sqrt{50}}}$$

Unfortunately, we are stuck, as in this case we do not know σ, the standard deviation of the pre-advertisement shopper population. In order to continue we have to make an estimate of σ. We assume that the effect of the advertising campaign is to shift the whole distribution of purchases up the scale: that is, after the campaign the population mean is higher (people buy more items) but that the standard deviation stays the same (the spread in the number of purchases stays the same). The only standard deviation we have is the post-advertisement sample standard deviation, s. Sample standard deviations are quite stable estimates of the population figure so we could use this to estimate the post-advertisement population standard deviation. As we are assuming that the post-advertisement population has the same standard deviation as pre-advertisement we can use our sample standard deviation, s, as an estimate of the pre-advertisement population standard deviation. (We are predicting that the effect of the advertisement will be to shift the distribution up the scale but not change the shape of the distribution in any way, so the standard deviation will remain the same.) In order to use our sample standard deviation as an estimate of the population parameter we must assume that our sample is not biased in any way, such as made up only of wealthy shoppers, or it will not be a good estimate. So we assume that our sample is randomly chosen from the post-advertisement population. If this assumption is met then our sample standard deviation should be a reasonable estimate of the pre-advertisement population figure.

To distinguish the fact that we do not have the population standard deviation σ but are using s as an estimate, instead of calling the statistic z, we call the new statistic t:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

As we saw in the previous chapter a sample standard deviation has the following formula:

$$s = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}}$$

replacing s by its formula in the formula for t we get a new formula for t:

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n(n - 1)}}}$$

Notice that t is influenced by the degrees of freedom of the sample ($n - 1$). This is because t is not the same as z but an estimate of it. When the degrees of freedom are small the t distribution is similar to a normal distribution but flatter and more spread out. As the degrees of freedom get larger the t distribution gets rapidly closer to a normal distribution and when the degrees of freedom are infinite it is identical to the normal distribution. Figure 6.2 shows three t distributions for 1, 10 and infinity degrees of freedom. Even at 10 degrees of freedom the t distribution is very similar to a normal distribution and at 30 degrees of freedom and above the differences are so small as to be irrelevant.

We always look up a z score in the standard normal distribution tables. We cannot do this with t as it is not a normal distribution. However, like the standard normal distribution tables, the values of the t distribution have been worked out. Indeed these have been worked out for the different t distributions corresponding to the different degrees of freedom. We can look up our calculated value of t in the table for the appropriate distribution and find the probability of this value arising from the known distribution.

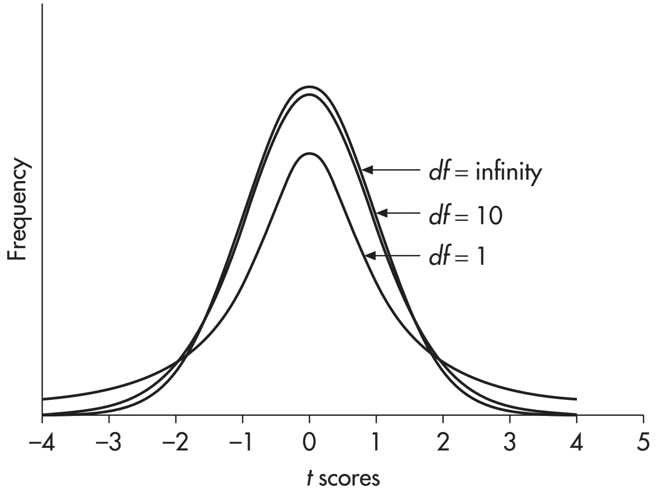

FIGURE 6.2 Examples of the *t* distribution

We can compare this value with our significance level and make a decision whether to accept or reject the null hypothesis. Thus we are able engage in hypothesis testing with a sample even when we do not know the standard deviation of the known population.

In order to perform a *t* test we have to make three assumptions:

1 The known population is normally distributed. This is important as (like a *z* score) we need our sampling distribution to be normally distributed. If it is not then the *t* distribution in the table might not provide us with the appropriate figures for our decision on the significance of the *t* value we calculate. However, it is often stated that the *t* test is 'robust': this is statistical jargon for saying that even if the underlying sampling distribution is not normal the *t* test might still provide a reasonably good figure for comparison. Certainly when the sample size is 30 or more the sampling distribution will be very close to normal, whatever the underlying population distribution.

2 The sample is randomly selected from the (unknown) population. We want our sample standard deviation to be an unbiased estimate of the population standard deviation, and hence a suitable estimate to use. Otherwise it will affect our calculation of *t*.

3 The standard deviation of the unknown population is the same as the known population. Only if we make this assumption can we take

the sample standard deviation as an estimate of the standard deviation of the known population.

Returning now to our example, the above assumptions are reasonable to make here, as long as we have no reason to believe our sample 50 shoppers was selected in a biased manner. We can now calculate *t* to find the probability of finding a pre-advertisement sample with a mean as large as 30 items:

$$t = \frac{\bar{X} - \mu}{\sqrt{\dfrac{\sum X^2 - \dfrac{(\sum X)^2}{n}}{n(n-1)}}} = \frac{30 - 25}{\sqrt{\dfrac{48482 - 45000}{2450}}} = \frac{5}{1.1922} = 4.19$$

We must also work out the degrees of freedom: $n - 1 = 50 - 1 = 49$.

If we look at the table of *t* values given in Table A.2 of the Appendix we note that, unlike the standard normal table, it does not have the figures for the whole distribution. Otherwise we would have table after table, giving all the values for each different *t* distribution. What the table shows is the key values for each distribution, where the key values are the values of *t* at the significance levels we commonly choose, i.e. which *t* value cuts off exactly 5 per cent and which value cuts off 1 per cent of the tail of the *t* distribution.

We have a one-tailed test (we are predicting the advertising campaign will result in more purchases). Using a significance level of $p = 0.05$, we look down the $p = 0.05$ column and along the row for 49 degrees of freedom and we find the *t* value is not there! There is a figure of 1.684 for 40 degrees of freedom and 1.671 for 60 degrees of freedom. The reason for this is that, again, if every figure was listed the column would go on for ever. We can see that there is there not much difference in these values (0.013) so we know roughly what our value for 49 degrees of freedom will be: somewhere between the two (1.671 and 1.684). We can find it out by a process called linear interpolation, which is easier than it sounds! Between 40 and 60 is a gap of 20 and between 1.684 and 1.671 there is a gap of 0.013. So for every degree of freedom between 40 and 60 the difference in the table is 0.013/20, which is 0.00065. For 9 degrees of freedom the gap is 9×0.00065, which is 0.00585. Therefore 49 degrees of freedom has a table *t* value of $1.684 - 0.00585$, which is 1.67815. (If you don't want to do a linear interpolation just take the larger of the two values in the table: 1.684.)

As a *t* value of 1.67815 cuts off exactly 5 per cent of the tail of the *t* distribution with 49 degrees of freedom, our value of *t* being larger will cut off less of the tail and the probability of getting a *t* value of 4.19 from the known distribution is less than 5 per cent, so we can reject the null hypothesis and argue for a difference in the two distributions.

More simply we can conclude that, as our calculated value of *t* of 4.19, with 49 degrees of freedom, is larger than the table value of 1.67815 for a one-tailed test, at the $p = 0.05$ level of significance there is a significant increase in the number of items purchased after the advertising campaign. Notice that it is also significant at the more conservative $p = 0.01$ level of significance and we would usually report the finding at the smaller significance level to indicate how unlikely it is that the effect could have occurred by chance. (See if you can work out by linear interpolation the table value of *t* for 49 degrees of freedom for the $p = 0.01$ level of significance. You should get a value of *t* of 2.40815.)

Confidence intervals

The *t* test is a test of significance and we seek evidence for a statistically significant difference between populations based on the sample information we have. An alternative approach is to use the sample information to estimate the population parameters. Now you may say that we have already done that by using our sample mean value as an estimate of the population value. That is true but we can be a little more sophisticated by working out a confidence interval for the mean. Rather than choosing a single value for the population mean we can specify a range of values within which we are confident that the value lies. We choose a level of confidence, usually either 95 per cent or 99 per cent confident, and then work out the range of values. With a 95 per cent confidence interval we are saying that if we worked out the confidence interval for 100 different samples from a population then 95 per cent of those confidence intervals would contain the population mean value. So our confidence interval is a good estimate of where the true mean lies.

In the above example we can work out the 95 per cent confidence interval quite easily as we use the information we produced for the *t* calculation to work it out. This is because for the *t* test the confidence interval (CI) is specified as follows:

$$\text{CI} = \text{Sample mean} \pm (\text{critical } t \text{ value} \times \text{standard error of the mean})$$

In this case the critical t value is the one that 'captures' the central 95 per cent of the distribution, leaving only 5 per cent outside the range, so is the two-tailed t value from the tables at $p = 0.05$, as this cuts off 0.025 from each end of the distribution.[5] We have 49 degrees of freedom so we can now find the critical t value from the tables, which is 2.0116 (by linear interpolation between the values for $df = 40$ and $df = 60$). We know the sample mean is 30 and we know that the (estimated) standard error of the mean is 1.1922 (as it is the bottom part of the t test formula).

So we have:

$$95\%\text{CI} = 30 \pm 2.0116 \times 1.1922 = 30 \pm 2.3982$$

which gives

$$95\%\text{CI} = (27.6018, 32.3982)$$

This gives us a helpful indication of the position of the true population mean. The narrower the confidence interval the more specific our estimate of the population mean. Here we are confident that the population mean lies between 27.6018 and 32.3982. Even the lowest of the two limits, 27.6018, is still well above the 25 value for the pre-advertising purchases.

We can extend our confidence interval analysis to give the confidence interval of the difference between our post- and pre-advertisement mean values ($\bar{X} - \mu$). We use the same formula but replace the sample mean with the difference in means:

$$\text{CI} = \text{Difference in means} \pm (\text{critical } t \text{ value} \times \text{standard error of the difference in means})$$

The critical t value and the standard error are the same as in the previous calculation and we know the value of μ so:

$$95\%\text{CI} = (30 - 25) \pm (2.0116 \times 1.1922)$$

$$95\%\text{CI} = (2.6018, 7.3982)$$

This provides us with a range of values that we are confident (95 per cent of the time) contains the real difference in the populations. Notice that in the 'worst case' (the lower limit) we still predict 2.60 more purchases after the advertisement so we can be confident that it has had an effect. Had the lower limit been zero or even negative we would not be able to assume a definite effect of the advertisement as the true difference could have been zero.

The general structure of a confidence interval

The above confidence intervals have been worked out using the sample statistic (the mean, or the difference in means), sample information (the standard error) and the appropriate statistical distribution for the data (the *t* distribution). We can calculate confidence intervals for many statistical analyses using the same structure as above, but we write the general statement as follows:

$$\begin{aligned}\text{CI} = \text{Value of statistic} &\pm \text{critical value of appropriate distribution}\\ &\times \text{standard error of the statistic}\end{aligned}$$

We then need to select the appropriate statistic, critical value and standard error to calculate the confidence interval. As we saw above, we work out the statistic and the estimate of the standard error from our data, choose the level of confidence we want (e.g. 90 per cent or 95 per cent) and then select the correct critical value for that confidence level.

Significance and confidence intervals

Significance tests and confidence intervals are both attempting to answer the same question: what does our sample information tell us about the population values and what can we conclude from it? In the first case, a significance test, we are seeking whether the sample statistic exceeds a particular criterion (the $p = 0.05$ significance level) to claim statistical significance (and reject the null hypothesis). In the second case, confidence intervals, we are seeking to find the range within which we can be confident that the population value lies. If we look at confidence intervals of a difference we can examine this range in relation to zero to give us an indication of whether we think the difference is important or not. If the confidence interval contains zero then the difference for the population values could well be zero and hence any difference we found in the sample means is not important.

Significance tests have been traditionally used in data analysis in a number of fields of study. However, confidence intervals are increasingly used.

This is because significance tests provide an 'either–or' outcome – either the null hypothesis is rejected or it is not at a particular significance level – whereas the confidence interval provides a range of values that provide a useful estimate of the size of the difference.

In a real sense significance tests and confidence intervals are complementary in that together they reveal a clearer picture of the data than they might on their own. In many cases (with a highly significant finding, for example) the conclusion is clear but where the finding is 'close' to significance (with a probability of 0.06, for example, which we would say is not significant) confidence intervals can help us evaluate the worth of further investigation, particularly if, as we shall see in Chapter 9, there are a number of factors that influence our statistical outcome.

Hypothesis testing with one sample: in conclusion

The same logic applies whether we are testing a sample or we are testing an individual score. However, with a sample the 'score from the unknown distribution' becomes the sample mean from the unknown sampling distribution and the 'known distribution' we compare it to is the distribution of sample means from the known population for samples of the same size. Once we have the details of the 'score' and the 'known distribution', then the procedures are identical: we work out the z score and find the probability in order to decide whether to accept or reject the null hypothesis. It is a little more complicated if we do not have the standard deviation of the known population but as long as we make the appropriate assumptions we can use the sample standard deviation to estimate it. We then calculate t instead of z. As the t distributions have all been worked out we can look up the critical value of t, with the appropriate degrees of freedom, for our chosen level of significance. If our calculated value is larger than the table value we can reject the null hypothesis.

Confidence intervals provide an alternative way of representing our findings as they provide a range of values within which we are confident that the population value lies. We may choose this as an alternative to our significance test or as supplementary information to it.