Chapter 7

Selecting samples for comparison

chapter 7

HYPOTHESIS TESTING with a single sample is used when we know about a particular population and wish to decide whether the sample comes from a different population or not. In most research we do not have details of any populations at all. All we know about are the samples we can obtain. In the majority of cases hypothesis testing is about comparing samples rather than comparing a sample mean with a sampling distribution. In the Cyadmine example considered in the previous chapter we had the details of a birth-weight population unaffected by the gas. More usually we will not have this information and can only collect a sample of babies affected by the gas and a sample of babies unaffected by it for comparison. We, of course, increase the problems of sample selection when we have two samples rather than one, as each is required to represent a population. Indeed, it is the fact that we want to use our samples to estimate populations that causes problems in sample selection, for we do not want to introduce biases that make our samples untypical of their population.

When we have two samples, not only do we wish them to represent their respective populations but we also want them to be comparable. For example, if we are comparing forty year old men and women on their degree of fitness we would not select women who were athletes and men who were taxi drivers as the samples are not comparable. Any difference in fitness could be due simply to occupation rather than gender. It is this problem of comparability we consider now.

Designing experiments to compare samples

The reason we undertake experiments is to test hypotheses. A major cause for concern is whether the experiment is really examining the hypothesis we wish it to test, to the exclusion of all others, or whether we have introduced a bias in some way. Poor sample selection can lead to an ambiguous experiment if we are unable to decide whether, say, a difference in fitness is due to occupation or gender.

Experimental variables

All experiments look at the effects of variables or factors, the terms are used synonymously. Variables are, not surprisingly, things that vary! Temperature, reaction time, teaching methods, gender, class, drinking habits, accuracy of performance are just a few examples.

In the simplest case of hypothesis testing we want to know whether a single score comes from a known population distribution or from a different population distribution. An example is comparing the reaction time of someone after a head injury with the population of reaction times from the uninjured population. We can also compare a sample mean with a known distribution of sample means. As an example we might compare the mean IQ score of a group of children taught by a new teaching method with the distribution of means of samples of the same size of children taught by the traditional method. In both these cases we need a known distribution.

More usually we will compare two or more samples of subjects to decide whether they come from the same or different populations, for example do men and women differ on their memory for faces? Note the word subject in this context simply refers to a member of a sample. A subject could be anything. Quite often it will be a person but it could be an animal (if we are studying rats learning mazes or dogs learning tricks) or indeed anything we want to study (bolts made by one machine in one sample and bolts made by another machine in a second sample). The use of the term 'subject' has been criticised in the study of psychology when referring to people who agree to take part in research. The modern terminology for such a person is participant as it is viewed as more respectful of these helpful individuals, without whom there would be little psychological research. However, in statistical analysis we refer to 'between subjects' and 'within subjects' for particular types of designs or calculations, so the term continues to have currency in this context. Where it is clear that it is people taking part in a study I will refer to them as participants rather than subjects.

In the examples we have considered so far each experiment has at least two factors. In the Cyadmine gas example we have the variable *Cyadmine*, varying between 'affected' and 'unaffected' and *birth-weight*, varying between the individuals we are measuring. In the memory experiment above we have *gender*, either 'men' or 'women', and *memory for faces*, which we vary along the scale devised to measure it.

In an experiment there can be one or more independent variables. These are the variables for which the experimenter selects the values in advance. With the variable *Cyadmine* we chose to look at two values: affected and

unaffected (rather than looking at, say, badly affected, moderately affected and slightly affected). With the variable *gender* we selected men and women (rather than boys and girls). The experimenter controls the values of the independent variables and the samples are selected so that they differ on these values.

As well as independent variables there is also the dependent variable in an experiment. This is the variable we measure and on which we obtain the scores. Whilst the researcher selects what factor will be the dependent variable in the experiment (birth-weight, reaction time, IQ score, memory for faces) the researcher cannot control the values of that variable. We do not know in advance what the scores will be on this variable. This is the point of performing the experiment. Let us consider another example of the two sample case: two groups of children engage in different methods of learning a second language. Is one method better than the other? In statistical terms we want to find a suitable dependent variable (such as amount learnt) that is dependent (i.e. influenced by) the independent variable (learning method) to see if the values of the dependent variables differ in our two samples to such an extent that we can conclude that the sample scores come from different distributions, and one method leads to a greater amount learnt than the other.

The problem of equivalent conditions

Experiments are all about predicting relationships between independent and dependent variables. A research hypothesis is a prediction that the dependent variable will vary with (depend on) changes in the independent variable.

Imagine we set up an experiment to test whether girls are better than boys at map reading. The first problem is deciding what we mean by 'map reading'. Reading a road map to get into town? Reading an ordnance survey map to cross a moor? There is not an easy answer to the question. We must make a choice and state it clearly. As we saw in Chapter 4, we must operationally define *map reading ability* for the purpose of our experiment, such as 'the time it takes a child to get from a specific church, across the fields to a specified post office, using an ordnance survey map only'. We have to attempt to arrange the conditions equally for the children, such as making sure that they are all unfamiliar with the route. And this highlights a second problem.

What if we find a difference between the boys and the girls on map reading ability: can we infer a relationship between gender and map reading

ability? Not necessarily; the reason being the difficulty of arranging equivalent conditions for the boys and girls. If the girls had undertaken the task in bright daylight and the boys in the dusk we would not be surprised if the boys were worse. In this case the independent variable *gender* was confounded by another variable *daylight*. Likewise, if all the boys were from an orienteering club and the girls had never seen a map before then a difference between them would not necessarily indicate a relationship of map reading ability to gender but to *experience*.

Confounding is an example of a systematic error. The experimental conditions are consistently different for the two samples due to other independent variables as well as the one under test. In addition to systematic errors influencing an experiment we also have random errors. These occur in an unsystematic way: a gust of wind makes it temporarily hard for one boy to read his map, a road is busy when one girl tries to cross but is quiet for another.

As it appears that we can never produce equivalent conditions for all the participants in the study should we abandon experimentation altogether? Unfortunately there is no research method without problems and there are ways of dealing with these difficulties. Systematic errors can be avoided when we are aware of them and it is the skill of the researcher to spot them. We can deliberately select our participants so that they are matched on a confounding variable. In our example, for each boy that has some map reading experience we match him with a girl who has had the same amount of experience. In this way the samples no longer differ on *experience* and it should no longer bias our results in favour of one sample. We can also monitor the *daylight* and make sure that the children perform in similar daylight conditions. By being a little more sophisticated in the design and operation of the experiment we can remove relevant systematic errors.

It is unlikely that we would match the children on hair colour as this is a factor we would not expect to influence this experiment. In matching we take account of only the factors we believe to be relevant. Again we can see it is one's expert knowledge of one's own discipline rather than statistical knowledge that guides these judgements. This is why an experiment should always be accurately reported, stating how the samples were matched. Another researcher might argue that an important confounding factor was not controlled for in the experiment.

We cannot control for random errors. However, our statistical tests are deliberately designed to help us decide if there is a difference between our samples above the level of any 'background noise' caused by these random errors, and we set a significance level to do this. We do not expect every

boy to get the same score, nor every girl. We expect a distribution of scores: not every boy runs at the same speed, not every girl trips up on the way. Random errors produce a distribution of scores across each sample. Statistical tests look for systematic differences between samples due to the independent variable above the random variation within a sample.

Related or independent samples

Sometimes, as in the map reading experiment, there are different participants in each sample. This is not surprising for the variable *gender*, as most children are either a boy or a girl, not both. In other experiments it is possible use the same participants in each sample. An example of this might be an experiment on the effect of temperature on reading comprehension where we test the participants' comprehension at two different temperatures. When a participant contributes a score to only one sample the experiment is called an unrelated, independent or between-subjects design and when the participants contribute a score to each sample the experiment is called a related, repeated measures or within-subjects design.

Consider an experiment where a researcher is trying to find out whether it is harder to understand the writing of Joseph Conrad (reputed to be difficult) compared to Charles Dickens. The researcher might select pieces written by the two authors (matched on length at least) and give them to a group of participants to read, followed by a comprehension test. This is a related design, as each person is in both samples. This has the advantage of matching the participants with themselves, so reducing possible errors due to differences between individuals (we will not have all the English enthusiasts in one sample). However, there are other problems to watch for. If the participants read the Dickens piece first followed by the Conrad they might perform worse on the Dickens, not due to comprehensibility, but because they read it first and it is not so fresh in their minds. We have introduced the confounding factor *memory time* into the experiment. To overcome this we must counterbalance the order of presentation, so half the participants read the Dickens first and half the Conrad. By this counterbalancing we will have controlled for confounding factors such as memory time, tiredness, boredom, experience of the test, etc.

The advantage of an independent design is that there are no carry-over effects from one sample to the next, whereas the disadvantage is that there may be systematic differences between the samples and therefore we must take care in our sample selection. In many cases we have to have an

independent design as we are testing an independent variable such as *gender* or *occupation* where participants can only be a member of one sample: people are normally working as either a doctor or a nurse but not both.

The interpretation of sample differences

Essentially, in designing experiments, we would like to select our subjects randomly from the largest population possible. If we do this then our results have the greatest generalisability. However we also have the greatest chance of confounding. Researchers compromise (as they must using any methodology) and lose some generalisability in favour of greater control over the variables involved. In the Cyadmine example we considered babies born in a single town where the gas cloud rested. This sample might not generalise to all Cyadmine-affected babies. Maybe there is something specific to the location that influenced the impact of the gas in some way. Yet this should not stop the researcher carrying out the test. Important information can still be found and it would also need to be demonstrated that the location does have an influence on the effects of Cyadmine.

Finally, in this section, we wish to design experiments that actually test out the hypothesis we are interested in! (It is amazing how many do not.) If we wish to test whether a reading scheme improves children's reading performance we cannot simply test them before and after they have taken part in the scheme. Any differences might be due to the fact that the children are older rather than the reading scheme as such. We have the confounding factor of *age*. To overcome this we match two groups of children on reading ability and then give one, the experimental group, the reading scheme but not the other, the control group. If the performance of the experimental group improves more than that of the control group then we may be able to relate it to the reading scheme as we have controlled for the effects of age by the selection of the control group.

In all experiments we are trying to establish relationships between the independent and dependent variables, controlling for extraneous variables that could influence this relationship. We must be careful when we do find a relationship that our interpretation is not in error. Experiments do not establish causal relationships, they only support or do not support testable hypotheses. For example, we might hypothesise that men and women differ on a certain factor. If we find a significant difference it supports our hypothesis but does not tell us why. The answer may be genetic, social or even a confounding factor that we have not taken account of.

The reason for undertaking experiments is to give us some systematic data on which to base our judgements and test our ideas. The more we learn about experimental methods the more sophisticated our judgements can be in assessing the worth of our findings. And it is the statistical analysis which helps us to decide what we have actually found out.

Chapter 8

Hypothesis testing with two samples

chapter 8

A TEACHER READ ABOUT a new reading scheme introduced in another country and wondered whether it could be used here. There were reports in the educational literature of the other country that the New Scheme resulted in better reading performance from the children. The problem was that these data were for another language. The teacher wanted to find out if the New Scheme was better than the Old Scheme currently being used in the classroom in this country.

The teacher decided to teach half the class on the New Scheme and half on the Old Scheme on the next class intake. The children were randomly allocated to the two schemes, to avoid biasing the samples due to factors such as intelligence. In this way the two samples were assumed to systematically differ only on the variable under study: *reading scheme*. The teacher can now compare the samples. Yet the teacher is not really interested in the samples as such but the population of children these samples are drawn from. Is the New Scheme better for children of this age rather than just this class? The question is whether the population distribution for the New Scheme is higher up a scale of reading performance than the distribution for the Old Scheme. This is a one-tailed prediction that the New Scheme will result in better performance than the Old Scheme. Unfortunately the teacher has no details of these populations, they are both unknown.

How can these samples be used to test the hypothesis? First of all we can ask whether the samples are representative of the populations we want to generalise to. How are the pupils selected for this school? What social groups do they come from? These factors might limit the generalisation. Second, we can look at the performance of the two samples on a test of reading. If the difference between the samples is small we might be sceptical of a difference in populations but if the difference is big we might decide that the finding indicates a likely difference in the populations. The problem we face is: how big must a difference be before we reject the null hypothesis and decide the samples really do come from populations with different distributions.

We can attack the problem in the following way. Let us assume that the two samples really do come from the same distribution, the null hypothesis is true and there is no difference in reading performance between the populations. What differences would we expect between two samples

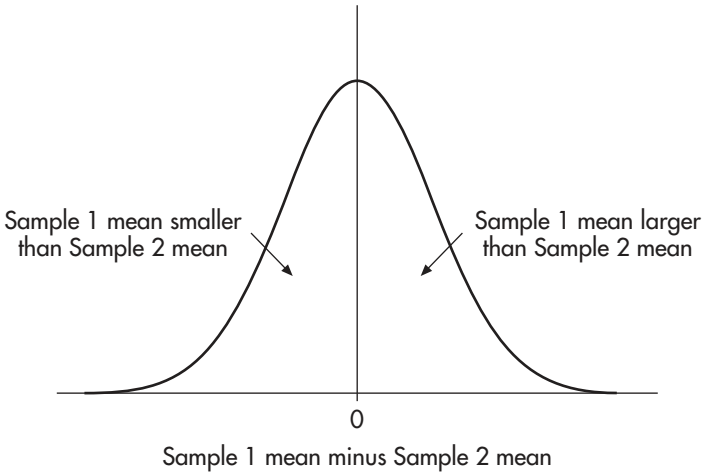

FIGURE 8.1 The distribution of the difference between sample means

simply by chance alone? We can find this out if we take the mean of every possible sample, of the size we are interested in, and compare it with the mean of every other possible sample of this size. These differences (in sample means) will tell us what differences we would expect when the null hypothesis is true. If we plot these differences we get the distribution of differences between sample means. Like the distribution of sample means this will tend to be a normal distribution as it is a sampling distribution. This will be especially the case if our sample size is large. The mean of this distribution will be zero because, when we take samples from the same distribution the differences will pile up around zero as there will be little or no difference between most sample means. Only occasionally will there be a large difference, say, when one sample has all the good readers and the other all the bad readers. The distribution of differences between sample means when the null hypothesis is true is shown in Figure 8.1.

Now, lo and behold, we have a known distribution: a normal distribution with a mean of zero. We also have a score to test: 'the difference in our sample means'. Hypothesis testing is all about comparing a score with a known distribution. If the probability is high that our difference in sample means comes from this distribution then the chances are that the null hypothesis is true. If there is a low probability of finding a difference such as ours from this distribution then the chances are that our samples come from different population distributions, and the null hypothesis can be rejected. All we need to do now is to construct a *z* score for the 'score' (the

difference between our sample means) and we can find the probability of this score coming from the 'known distribution' (the distribution of differences between sample means) to find the probability of the null hypothesis being true.

A z score needs a score, a mean and a standard deviation. Our 'score' is the difference in sample means. If we call the mean of Sample 1 $\overline{X}_1$ and the mean of Sample 2 $\overline{X}_2$ then the difference in sample means is $\overline{X}_1 - \overline{X}_2$. The mean and standard deviation of the distribution of differences in sample means, when the null hypothesis is true, are given the following symbols: $\mu_{\overline{X}_1-\overline{X}_2}$ and $\sigma_{\overline{X}_1-\overline{X}_2}$, respectively. (As it is the standard deviation of a distribution concerning sample means we must remember that $\sigma_{\overline{X}_1-\overline{X}_2}$ is a standard error. It gives us the standard distance of a difference in sample means from the mean of the differences in sample means.) And we so can write the following formula for z:

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - \mu_{\overline{X}_1-\overline{X}_2}}{\sigma_{\overline{X}_1-\overline{X}_2}}$$

Now we know that $\mu_{\overline{X}_1-\overline{X}_2} = 0$, so we can write z as follows:

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sigma_{\overline{X}_1-\overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\overline{X}_1-\overline{X}_2}}$$

All we need to do now is look up the z score in the standard normal table to find the appropriate probability value. The problem is that we do not know $\sigma_{\overline{X}_1-\overline{X}_2}$. We will have to estimate it. How do we estimate the standard error of the distribution of differences between sample means when the null hypothesis is true? We have to use our samples. We replace $\sigma_{\overline{X}_1-\overline{X}_2}$ in the formula with $s_{\overline{X}_1-\overline{X}_2}$, which is the standard error of the difference between our sample means. It may look a little different to the one we created in Chapter 6 but we have the t statistic once again, as an estimate of z by using sample information to estimate the population standard error. The difference is only in the appearance of the formula: we still have a 'score' $(\overline{X}_1 - \overline{X}_2)$ minus a mean (which in this case is zero) divided by an estimated standard error $(s_{\overline{X}_1-\overline{X}_2})$:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1-\overline{X}_2}}$$

Recall from Chapter 6 that we know all about the *t* distribution so we are able to find probability values in the tables. We must not forget that the *t* distribution is influenced by the degrees of freedom of the samples, as the larger the samples the closer the distribution approximates the normal distribution. We must work out the degrees of freedom of our samples if we are to compare our calculated *t* value to the correct *t* distribution.

We now have a statistic we can work out using the information from our samples and we will be able to use it to make decisions concerning the population distributions: just the thing for hypothesis testing using two samples. Essentially, $s_{\overline{X}_1-\overline{X}_2}$ is (an estimate of) how much we would expect our means to differ by chance (when they come from the same distribution) whereas $\overline{X}_1 - \overline{X}_2$ is the actual difference in means. $(\overline{X}_1 - \overline{X}_2)/s_{\overline{X}_1-\overline{X}_2}$ tells us how much bigger our difference in means is relative to the difference expected by chance alone. The larger this ratio the greater our confidence that the mean difference is not due to chance but due to two different population distributions. The tricky thing is working out $s_{\overline{X}_1-\overline{X}_2}$ but in subsequent sections we will see how this is done.

The assumptions of the two sample *t* test

The basic assumptions of the *t* test are the same whichever *t* test we are undertaking. We require the sampling distribution to be normally distributed so we usually assume that our samples come from normally distributed populations. Fortunately, the *t* test is robust so that even if the distributions are only vaguely normal: humped in the middle and tailing off to the sides, then the *t* test is still likely to be valid. This is especially true for large samples (greater than 30). Again, we must assume that the samples are randomly chosen from their populations so that we can use sample statistics (mean, standard deviation) as unbiased estimates of the population parameters. Finally, we assume that the two samples come from populations with equal variances (and equal standard deviations as one is simply the square root of the other) to allow us to use the sample information to estimate population standard deviations. Thus, we are assuming that any effect of the independent variable is to shift the distribution of the dependent variable along the scale (i.e. alter the population mean) but not change its shape (its variance, or standard deviation).

Related or independent samples

As we saw in the previous chapter, related samples involve subjects providing scores for both samples, whereas with independent samples each subject contributes a score to only one sample. The way we calculate the two sample *t* test depends on whether the two samples are related or independent: there are different formulae that take account of the various differences this entails. For example, if we have 10 subjects in our two samples, for related samples we require only 10 different subjects as they are used twice, whereas with independent samples we require 20 subjects, 10 for each sample. The details of the different formulae are shown below.

The related *t* test

We start with our formula for *t*:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

We can work out $\overline{X}_1 - \overline{X}_2$ easily enough. The difficulty is to work out $s_{\overline{X}_1 - \overline{X}_2}$. Recall that the standard deviation of a distribution of sample means is called a standard error of the mean. $s_{\overline{X}_1 - \overline{X}_2}$ is still a standard error, as it is still based on sample means, and so can be expressed as the standard deviation of the difference between the scores divided by the square root of the sample size:

$$s_{\overline{X}_1 - \overline{X}_2} = \frac{s_{X_1 - X_2}}{\sqrt{n}}$$

We now need to work out the standard deviation of the difference in sample scores, $s_{X_1 - X_2}$. The difference in sample scores is easy to calculate with related samples. For each subject we can calculate a difference score *d* simply by subtracting the subject's score in Sample 2 from their score in Sample 1: $d = X_1 - X_2$. We can legitimately do this as the samples are related. Consider the example of comparing the length of a night's sleep. If a person sleeps 8 hours on Monday and 7 hours on Tuesday the difference for that person is 1 hour of sleep. The difference score for the participant is $8 - 7 = 1$. We then find the standard deviation of the difference scores:

$$s_{X_1-X_2} = s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

Here we have the usual standard deviation formula. With n subjects in each sample there are n difference scores. We can now produce a formula for $s_{\bar{X}_1-\bar{X}_2}$, the standard error, by dividing the above formula by $\sqrt{n}$:

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n(n-1)}}$$

And now finally we have our formula for the two sample related t test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n(n-1)}}}$$

Note that, whilst the formula looks very different to the z formula, it is still a score $(\bar{X}_1 - \bar{X}_2)$ minus a population mean (0) divided by a standard deviation, although in this case it's rather a complex standard deviation: the estimate of the standard error of the difference in sample means.

A worked example

A teacher believed that the children in her class were better at their work in the morning than in the afternoon. She decided to test this out by using a mathematics test as this required the children to concentrate. If there was a post-lunch dip in performance the test should pick it up. She chose a random sample of 8 children from the class and gave them two tests matched on their difficulty. The samples were balanced on the two versions of the test, and at what time they were tested first, to control for carry-over effects. The tests gave a score out of 10, the higher the score the better the performance. The results were as follows:

Participant	*Morning*	*Afternoon*
1	6	5
2	4	2
3	3	4
4	5	4
5	7	3
6	6	4
7	5	5
8	6	3

This is a related two sample *t* test as all participants contributed a score to both samples.

We must now find the values to fit into the formula:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sum d^2 - \dfrac{(\sum d)^2}{n}}{n(n-1)}}}$$

We can now relabel the columns, with Sample 1 for Morning and Sample 2 for Afternoon and find the means ($\overline{X}_1$ and $\overline{X}_2$), the difference scores (d), the sum of the difference scores ($\sum d$), the square of the sum of the difference scores ($(\sum d)^2$), the squared difference scores (d^2), and the sum of the squared difference scores ($\sum d^2$). The number of participants in each sample is n.

Participant	*Sample 1* X_1	*Sample 2* X_2	*Difference* d	*Squared d* d^2
1	6	5	1	1
2	4	2	2	4
3	3	4	–1	1
4	5	4	1	1
5	7	3	4	16
6	6	4	2	4
7	5	5	0	0
8	6	3	3	9
$n = 8$	$\overline{X}_1 = 5.25$	$\overline{X}_2 = 3.75$	$\sum d = 12$ $(\sum d)^2 = 144$	$\sum d^2 = 36$

Inserting the figures into the t formula we get:

$$t = \frac{5.25 - 3.75}{\sqrt{\frac{36 - \frac{144}{8}}{8(8-1)}}} = \frac{1.50}{\sqrt{\frac{36-18}{56}}} = \frac{1.50}{\sqrt{0.321}} = \frac{1.50}{0.567} = 2.65$$

The degrees of freedom (df) for a related t test is always $n - 1$, so $df = 7$.

This is a one-tailed test as the prediction was that the children would perform better in the morning, and the prediction is that the scores in Sample 1 are larger than in Sample 2. As can be seen from the means this is the case but we need to test the significance of the difference. At the $p = 0.05$ level of significance, we find from the t distribution tables (Table A.2 in the Appendix) that $t = 1.895$, $df = 7$ for a one-tailed test.

The calculated value of t of 2.65 being greater than the table value of 1.895 allows us to reject the null hypothesis, at the $p = 0.05$ level of significance, and conclude that the pupils did perform significantly better on the mathematics test in the morning compared to the afternoon.

Sometimes we find that the calculated t has a minus sign. This simply indicates that the mean of Sample 1 is smaller than the mean of Sample 2. If we had found a minus sign in the above example we could have rejected the one-tailed prediction straight away as it would have meant better scores in the afternoon. If we had predicted that Sample 2 has the larger scores, or made a two-tailed prediction, we simply ignore the minus sign when comparing the calculated value with the table value.

The independent *t* test

We again start with our formula for t: $\frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$. The difficulty with independent samples is working out $s_{\overline{X}_1 - \overline{X}_2}$. How we do this is explained below. Now this does include some rather horrible formulae, so, if you wish, do not worry about following the derivation of the formula for the independent t test, feel free to skip the mathematics. If you understand the logic that we have to find a formula for $s_{\overline{X}_1 - \overline{X}_2}$ and that this formula, though rather cumbersome, is still an estimated standard error of the difference in sample means then that's fine.

We cannot produce difference scores as we did for the related t test. (If the samples are unrelated we cannot work out a difference score. If one

person sleeps 8 hours on Monday and another person sleeps 7 hours on Tuesday it is meaningless to subtract one from the other as they are from different participants.) Indeed we may have different numbers of subjects in the two samples (n_1 and n_2). We are helped out in this case by a mathematical finding called the *Variance Sum Law*, which provides us with a relationship between $s_{\bar{X}_1-\bar{X}_2}$ and the standard deviations of the two samples (s_1 and s_2):

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The importance of this is that we cannot work out $s_{\bar{X}_1-\bar{X}_2}$ but we can work out s_1 and s_2. Thus, we are able to produce a formula for the independent t that we can calculate.

Our problems are not over yet in developing the formula for t. We know that a sample standard deviation is a better estimate of the population parameter the larger the sample size and also that the t test assumes that the samples come from populations with equal standard deviations. From this we can infer that when we have samples of different sizes the larger one is likely to provide a better estimate of the population standard deviation than the smaller one. What we do is to weight the contribution of the two sample standard deviations by their sample size (more accurately, their variances by their degrees of freedom) and produce a population estimate based on the weighted average of the sample standard deviations, s_w:

$$s_w^2 = \frac{(n_1 - 1)s_1^2 - (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Now, instead of using the sample standard deviations in the formula for $s_{\bar{X}_1-\bar{X}_2}$ we replace them both with s_w:

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_w^2}{n_1} + \frac{s_w^2}{n_2}} = \sqrt{s_w^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

We now expand s_w in the formula:

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\left(\frac{(n_1 - 1)s_1^2 - (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Finally, we replace s_1 with $\sqrt{\dfrac{\sum X_1^2 - \dfrac{(\sum X_1)^2}{n_1}}{n_1 - 1}}$ and s_2 with $\sqrt{\dfrac{\sum X_2^2 - \dfrac{(\sum X_2)^2}{n_2}}{n_2 - 1}}$, the standard deviation formulae for two samples.

After a little tidying up, we obtain the formula for calculating $s_{\overline{X}_1 - \overline{X}_2}$:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{\sum X_1^2 - \dfrac{(\sum X_1)^2}{n_1} + \sum X_2^2 - \dfrac{(\sum X_2)^2}{n_2}}{(n_1 - 1) + (n_2 - 1)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

At last we are able to produce the formula for the two sample independent *t*:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\dfrac{\sum X_1^2 - \dfrac{(\sum X_1)^2}{n_1} + \sum X_2^2 - \dfrac{(\sum X_2)^2}{n_2}}{(n_1 - 1) + (n_2 - 1)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

This is unfortunately rather a large formula to calculate but I hope you can see how and why it was required by the above logic. Also on many occasions we can use computer programs to aid us in our calculations. As demonstrated below, *t* can be calculated without too much difficulty with just a calculator. But the point here is that, while the formula looks very different from the *z* formula, it is still an estimate of *z* being a 'score' ($\overline{X}_1 - \overline{X}_2$) minus a mean ($\mu_{\overline{X}_1 - \overline{X}_2} = 0$) divided by a standard deviation ($s_{\overline{X}_1 - \overline{X}_2}$).

It is important to recall that we are using the assumption that the two samples come from populations with equal variances (and hence equal standard deviations). If this is not the case it is inappropriate to average our standard deviations for estimation. Only if the larger sample variance is more than three times the other would we usually decide not to perform the test.

As the samples are unrelated, the degrees of freedom of the independent *t* test is the sum of the degrees of freedom of each sample: $(n_1 - 1) + (n_2 - 1)$.

A worked example

A new sleeping pill was being tested on a number of volunteers. It was predicted that it would have a differential effect on men and women. There were six men and eight women who agreed to take part in the experiment. Over a two week period they took either a placebo (a pill that had no effect) or the sleeping pill. Participants were not aware of which pill they were taking each night. The number of extra hours slept during the seven 'pill nights' compared to the seven 'placebo nights' was calculated. The men slept 4, 6, 5, 4, 5 and 6 extra hours and the women slept 3, 8, 7, 6, 7, 6, 7 and 6 extra hours. Is the prediction supported?

We must find the values to fit into the *t* formula:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{\sum X_1^2 - \frac{(\sum X_1)^2}{n_1} + \sum X_2^2 - \frac{(\sum X_2)^2}{n_2}}{(n_1 - 1) + (n_2 - 1)} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

I shall label the men as Sample 1 and the women as Sample 2.

Sample 1		*Sample 2*	
X_1	X_1^2	X_2	X_2^2
4	16	3	9
6	36	8	64
5	25	7	49
4	16	6	36
5	25	7	49
6	36	6	36
		7	49
		6	36
$n_1 = 6$		$n_2 = 8$	
$\sum X_1 = 30$	$\sum X_1^2 = 154$	$\sum X_2 = 50$	$\sum X_2^2 = 328$
$\overline{X}_1 = 5.0$		$\overline{X}_2 = 6.25$	
$(\sum X_1)^2 = 900$		$(\sum X_2)^2 = 2500$	

Inserting the figures into the t formula we get:

$$t = \frac{5.00 - 6.25}{\sqrt{\left(\frac{154 - \frac{900}{6} + 328 - \frac{2500}{8}}{(6-1)+(8-1)}\right)\left(\frac{1}{6}+\frac{1}{8}\right)}}$$

$$= \frac{-1.25}{\sqrt{\left(\frac{154 - 150 + 328 - 312.5}{5+7}\right)\left(\frac{1}{6}+\frac{1}{8}\right)}}$$

$$t = \frac{-1.25}{\sqrt{1.625 \times 0.292}} = \frac{-1.25}{\sqrt{0.474}} = \frac{-1.25}{0.688} = -1.82$$

The degrees of freedom, $df = (n_1 - 1) + (n_2 - 1) = (6 - 1) + (8 - 1) = 12$.

The minus sign simply indicates that Sample 2 (women) has the larger scores. As we are testing a two-tailed test we simply treat it as +1.82. From the tables of the t distribution $t = 2.179$, $df = 12$, $p = 0.05$ (from Table A.2 in the Appendix). As our calculated t value of 1.82 is not greater than the table value of 2.18 we cannot reject the null hypothesis: we have not found a significant difference in the extra sleep between men and women at the 5 per cent level of significance.

It is an interesting result however. Notice that the difference in means is 1.25 in favour of the women. The difference in means we would expect by chance is 0.688 (the bottom part of the t calculation). Even though this is not significant at $p = 0.05$ the actual probability is 0.0945, which is still quite small. There might actually be a genuine effect here 'bubbling under' but not quite strong enough to pick up in these data. If we had more participants or had made a one-tailed prediction we might have achieved significance. The reasons why this might be are explained in the next chapter.

Confidence intervals

We can work out confidence intervals for the differences in the mean values when we are comparing two samples. Recall from Chapter 6 that:

CI = Difference in means ± (critical t value × standard error of the difference in means)

For the example of the related *t* test given in this chapter, we calculate the 95 per cent confidence interval as follows:

$$95\%CI = (5.25 - 3.75) \pm (2.365 \times 0.567)$$

$$95\%CI = 1.50 \pm 1.341$$

$$95\%CI = (0.159, 2.841)$$

The critical *t* value (2.365) is found in the tables for $p = 0.05$ for a two-tailed test with $df = 7$. The standard error calculation (0.567) is the denominator in the formula for the calculated *t* value. Notice that the interval does not include the zero so we can confidently conclude that the difference between the sample means is not zero but a positive value.

For the example of the independent *t* test the 95 per cent confidence interval is calculcated thus:

$$95\%CI = (5 - 6.25) \pm (2.179 \times 0.688)$$

$$95\%CI = -1.25 \pm 1.499$$

$$95\%CI = (-2.749, 0.249)$$

The critical value of *t* of 2.179 is found from the tables at $p = 0.05$ for a two-tailed test, $df = 12$. Again, the standard error value (0.688) is taken from the *t* calculation. Notice that in this example the confidence interval includes the zero value. In this case we are not confident that the 'true' difference in the means is different from zero. Just as the *t* value did not reach significance so the confidence interval, whilst mostly below zero, still contains zero within it. Both analyses are telling us that we do not have enough evidence from these data to claim a difference in the sample means.

Details on how to undertake the two sample *t* test using the SPSS computer statistical package can be found in Chapter 7 of Hinton *et al.* (2004).

Chapter 9

Significance, error and power

chapter 9

Type I and Type II errors

Hypothesis testing is like digging for treasure on a treasure island. The significance level sets the probability that we have actually found treasure rather than made a mistake. We are very conservative here (that is why we only accept a 5 in 100 chance of making a mistake). We do not wander about picking up any old piece of rusting metal we chance upon and claim that we have found treasure. Our fellow treasure hunters would soon get fed up with us. We want to be sure that when we claim to have found treasure then we are correct. In hypothesis testing we do not want to make a Type I error: that is, claim that we have found a significant difference between the population distributions when there is not one. We do not want to claim that we have found treasure when we have not. That is why we set the significance level at a small probability level.

In the one-tailed prediction illustrated in Figure 9.1 we are saying that if the 'score' falls beyond the significance level then it belongs to a different distribution to the known distribution, the unknown distribution. You can see in this example, where the unknown distribution really is different to the known distribution, that a score beyond the significance level is more likely to come from the unknown distribution than the known distribution as more

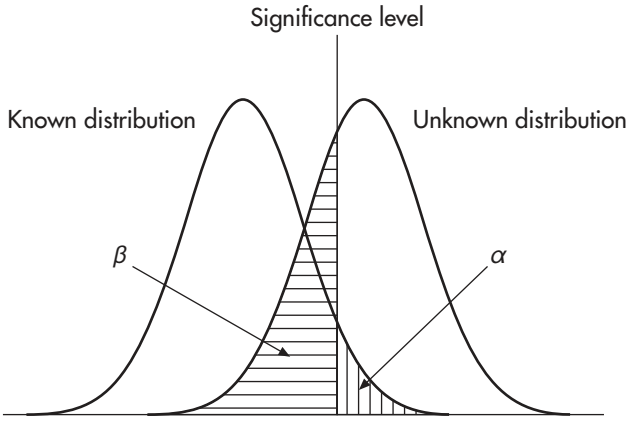

FIGURE 9.1 The risk of a Type I and Type II error

of it is beyond the significance level. However, there is still a small risk that such a score comes from the known distribution. The area labelled α is the size of this risk, the risk of a Type I error, which is the amount of the known distribution the 'wrong' side of the significance level. We specify the size of this risk by setting the significance level. By setting a significance level at $p = 0.05$ we are saying that only 5 per cent of the known distribution lies beyond it.

If the score falls below the significance level we accept the null hypothesis that the score comes from the known distribution. Looking again at Figure 9.1 we can see that 95 per cent of the known distribution lies below the significance level. Also there is more of the known distribution below the significance level than the unknown distribution, so the chances are that if a score lies this side of the significance level it comes from the known distribution and we are correct in accepting the null hypothesis.

We must be clear in understanding what 'accepting the null hypothesis' entails. All acceptance means is that we have not found a significant difference *in our experiment*. In fact some authors (Cohen, 1988, p. 16., see also Wilkinson and the Task Force on Statistical Inference, 1999) have argued that it is wrong to say that we accept the null hypothesis, rather we should always say 'we have failed to reject the null hypothesis' as this is a more accurate account of the situation – we have not found enough evidence to allow us to reject the null hypothesis. We have certainly not demonstrated that the null hypothesis is true. We can claim that we do not have the evidence to say it is not true, and there is a subtle difference between the statements. Again, if we dig for treasure on a desert island and do not find it, it does not mean that it is not there somewhere. When we 'accept the null hypothesis' we are only saying that we have not found a big enough difference for us to reject the possibility that the difference arose by chance. The probability of the difference arising by chance is too large for us to claim a genuine difference in the distributions. If we do not find treasure there are two possible reasons: one, there is no treasure there or, two, there is treasure but we have not found it. Similarly, if we do not find a significant difference when testing an hypothesis it could be that there really is no difference in the distributions or that there is a difference and we have missed it. In the former case all is well, we have not found a difference when there was not one to find. In the latter case we have committed a Type II error. We have not found a difference in the distributions by our test when there was a genuine difference to be found.

If a score falls below the significance level then we accept the null hypothesis that the score comes from the known distribution. However,

there is a risk that the score comes from the unknown distribution (as part of the unknown distribution lies below the significance level). The risk of making a Type II error is the amount of the unknown distribution below the significance level. This is the area labelled β in Figure 9.1. Note that the risk of a Type II error, β, may well be larger than α. Researchers do not want to make claims that turn out to be false so are happier to make Type II errors than Type I errors. We would prefer to miss out on the treasure occasionally rather than make a false claim. Most scientists publish their significant results and these results are scrutinised by others so it is deemed better to err on the side of caution rather than make a potentially embarrassing claim. It is tempting to think that all we need do to avoid a mistake is to set a very small value for α, say, 0.01 or 0.001. However, this would be an error as statistical testing is not just about keeping the risk of a Type I error low but also about the balance between the risk of Type I and Type II errors. As we shall see below (in the discussion of power) ignoring the risk of a Type II error could mean that our study, involving all the time and effort to carry it out, is simply not powerful enough to find the effects we are looking for – so we are wasting our time and effort.

Essentially we want a significance level that separates the known distribution from the unknown distribution. If we could find a position along the scale where all the known distribution fell to one side of the significance level and all the unknown distribution fell to the other side, then we would not make a Type I or Type II error as the significance level would separate the distributions perfectly. But, because of the overlap of the two distributions, some of the known distribution (α) falls the 'wrong' side of the significance level as does some of the unknown distribution (β). If a score falls below the significance level we 'accept the null hypothesis' as most of the known distribution ($1 - \alpha$) lies below it, with only β of the unknown distribution. If a score falls beyond the significance level we reject the null hypothesis as only α of the known distribution lies beyond it along with $1 - \beta$ of the unknown distribution. Although we risk these two types of error, we want the probability to favour the correct judgement.

Statistical power

For a moment let us assume that there really is treasure hidden on the desert island. With a good map and proper digging equipment there is an excellent chance of finding it. This is the analogy for a well-designed study, properly carried out. Yet without a map and only a child's bucket and spade the

chances of finding the treasure are slim. Similarly with the statistical analyses of our data. Some are very likely to find a difference in the two distributions whereas others may be unlikely to find it *even though it is really there*. The tests differ in their power. The power of a statistical test refers to its ability to find a difference in distributions when there really is one. In this case the unknown distribution is genuinely different to that of the known distribution. What are our chances of finding it? A score that actually comes from the unknown distribution will only be claimed to have come from the unknown distribution when that score is beyond the significance level. So we will correctly assign scores that belong to the part of the unknown distribution beyond the significance level. This is the whole of the unknown distribution excluding β. We call this area the power of the test.

The power of a test = $1 - \beta$

The power tells us the probability of finding the unknown distribution when it is really there. The more of the unknown distribution that lies beyond the significance level, the smaller β becomes and the larger $1 - \beta$. A more powerful statistical test is more likely to find a significant result than a less powerful test. Employing the treasure hunting analogy, a more powerful test is more likely to find the treasure when it is really there: it is the mechanical digger compared to the child's bucket and spade.

There is a problem that sometimes gets overlooked in statistical analysis. We do not want to use a test that is low in power as it is not likely to find a genuine difference in distributions. We may have constructed an excellent experiment only to fail to find a significant result due to the low power of our statistical test. Interestingly, 'power' became an increasingly important topic in statistical analysis in the latter part of the twentieth century, primarily due to the work of Jacob Cohen (e.g. Cohen, 1988), who has argued that much research has been carried out without a consideration of power in the design stage to the detriment of the research process. As a result of Cohen's work more researchers consider 'power' in the early stages of their research planning.

The power of a test

When undertaking research we want to have a good chance of finding an effect if there really is one to be found. In treasure hunting terms it would be helpful to know we are starting out with a mechanical excavator. Yet

there are many occasions when researchers set out with the statistical equivalent of buckets and spades. Clearly we want a powerful test but how can we achieve it?

The first thing to decide is what is the level of power we want? Crudely, just as we want α to be very small we also want $1 - \beta$ to be very large – the more powerful the test the better. But just like our consideration of α, we need to get the balance right. We want high power but not to the detriment of all other considerations. Cohen (1988) suggests that a power of 0.80 is a suitable value for a test of high power. As a result a power of 0.80 has become something of the conventional value for $1 - \beta$, just as 0.05 is the conventional value for α.

The problem is how do we design a study with the required power as many studies published in the journals have been shown to have much lower power than 0.80? The answer is that power is related to three factors that we can control: the size of α, the size of the effect we are looking for and, third, the size of the samples we select.

The choice of α level

The simplest way to increase the power of a test is to increase the size of α. We usually set the significance level at $p = 0.05$, that is $\alpha = 0.05$, but if we increase the level to say $p = 0.10$ or $p = 0.20$ then it has the effect of shifting more of the unknown distribution beyond the significance level. As α gets bigger β gets smaller and hence $1 - \beta$ gets bigger. However, while this reduces the risk of a Type II error it increases the chances of a Type I error. A significance level of $p = 0.10$ means that we will claim an effect erroneously ten times in a hundred rather than five in a hundred. And we don't want to do this for the reasons stated earlier: researchers would prefer to miss an effect than falsely claim one that could affect their reputation. Type I and Type II errors are inextricably linked, a reduction in one increases the other. Yet we can consider whether we really want to set a significance value as low as 0.01 or even 0.001. As Cohen (1988) points out, if we end up with such low power that the ratio of β to α is in the hundreds, then this implies we are stating that a Type II error is hundreds of times worse than a Type I error. If we don't really believe this, we may be happy to set our α value to a higher value (e.g. 0.05) and have a more powerful test.

However, there is a way of reducing β without increasing α: be more specific in our prediction. A one-tailed test is more powerful than a two-tailed test. In the latter case we have to consider both tails of the distribution

and we hedge our bets as to the position of the unknown distribution. For an overall significance level of 0.05 we must set the cut-off point at each tail at $p = 0.025$. It is like performing two one-tailed tests at the same time, one on each tail. If the unknown distribution really is higher than the known distribution we will only find it if it is beyond the $p = 0.025$ significance level. With a one-tailed test, we can focus on only one tail and at that tail α is twice the size (0.05) than for a two-tailed test. Shifting from a two-tailed to a one-tailed test increases $1 - \beta$. (We should note that this does make our one-tailed prediction more powerful but we now have *no power* in detecting the effect if the result goes the 'wrong way'.)

Effect size

A crucial factor affecting 'power' is the size of the effect we are looking for. If we look at Figure 9.1 we can see that the amount of overlap between the two distributions is the cause of our difficulty in setting a significance level with a low α and a high β. When there is a lot of overlap the risk of a Type II error, missing a genuine difference, increases. If the overlap between the distributions can be reduced, then β is reduced and we also reduce the chance of a Type II error and increase power. If there was no overlap between the distributions we would have no difficulty setting our significance level as we could position it between the two distributions. Sadly we will always have overlapping distributions but we can look at specifying how much overlap we have and designing studies to maximise their power.

Overlapping population distributions

The amount of overlap between two distributions depends on two factors: the difference between the population means and the size of the standard deviations. If the means are far apart then the overlap is less than when they are close together. Also if the standard deviations are small then the overlap is less than when they are large. (Recall that we always assume that the two distributions have the same standard deviation.) We can sum up the overlap by defining the effect size d (from Cohen, 1988). This is a standardised measure of the difference between the means in terms of standard deviation units. Using the label μ_1 as the mean of the known distribution and μ_2 as the mean of the unknown distribution, and σ as their standard deviation, we can

express the effect size, when predicting the one-tailed hypothesis that the unknown distribution will have the larger mean, as follows:

$$\text{Effect size, } d = \frac{\mu_2 - \mu_1}{\sigma}$$

For example, with $\mu_1 = 100$, $\mu_2 = 110$, and $\sigma = 15$, then the effect size d = 0.67. Just like the *z* score d is a standardised measure and does not depend on the measuring units we are using.

We need to know the size of the effect we are investigating in order to work out the power of our test *at the design stage* (a priori). You might think: how do I know the size of the effect before I have done the study? One source of information is past studies. If we were examining the speed of recognising different types of words we can look at the literature on the topic to see what other people have found in related studies. We can use these studies to get an estimate of the size of the effect we are looking for. If there is little background literature – you are studying a new area – then a pilot study might be worth carrying out to 'get a feel' for the type of results you might get.

Cohen (1988) makes the distinction between 'small' (d = 0.2), 'medium' (d = 0.5) and 'large' effects (d = 0.8) as helpful guide to evaluating the size of a predicted difference. He suggests that, rather than trying to work out a specific effect size by estimating means and standard deviations we can consider whether we expect a small, medium or large effect. He argues that, for new areas of research, effects are often small, partly because we may not have developed sophisticated measuring devices or experimental control leading to relatively large standard deviations. So, if we believe that the effect we are looking for is small then we can reasonably assume an effect size of 0.2. Cohen suggests that medium effects are 'visible to the naked eye' (Cohen, 1988, p. 26), meaning that we are aware of a difference such as that between experienced machine operators and novices as it is pretty clear to see but we want to examine it in detail. In cases like this we can assume a medium effect size of 0.5. Finally, there are the large effects which are blatantly obvious, or 'grossly perceptible' as Cohen (1988, p. 27) puts it, and uses as his example the height difference between 13 and 18 year old girls. If we believe that the effect we are looking for is large Cohen recommends that we select an effect size of 0.8.

In our example we do not have to estimate the effect size as I have stated the population means and standard deviations which we would not normally have. It is interesting to note that in Cohen's terms we are predicting a medium-to-large effect as d lies between 0.5 and 0.8.

Influencing effect size

You might be tempted to argue that you cannot change the effect size at all – surely a small effect is a small effect. However, if we consider for a moment what we actually mean by effect size then we can see how to influence it. A large effect size indicates only a small overlap between distributions whereas a small effect size indicates a large overlap of the distributions. What we need to do, therefore, to increase the power of a test, and increase the effect size, is to increase the difference between the means of the distributions or reduce their standard deviations.

The one major way to decrease the overlap between distributions is to *design your studies well*. It is very important to consider what a good design entails – essentially it is one that minimises error or random variability in the study and maximises the accuracy of measurement of the variables under study.

The more you reduce random variability in the study (by proper controls in the design and procedure) the greater will be the size of the effect. Imagine we are examining face recognition. We might study it in a natural setting such as an airport. However, we might choose to use computer displays with accurate timing and keypad responses in a quiet laboratory with no distractions in order to reduce the random variability in the study.

The effect of the sensitivity of the measuring device can crucially affect the power of a test. If we are investigating happiness then we might decide to use a more complex questionnaire than simply asking people if they are happy or not. Similarly, if we are testing a subtle effect such as speed of reading different passages of text then we may wish to use a more accurate time than a stopwatch. The reason for this is that the error in starting and stopping the stopwatch might be a second or two which could swamp an effect of only a few hundred milliseconds. If we can increase the accuracy of the measured times then we are more likely to find the effect (if there is one.)

Sample size

When we are studying samples to represent populations we use sampling distributions to represent our known and unknown distributions. The standard deviation of a sampling distribution, the standard error of the mean, decreases as the sample size increases. This is because the standard error is based on both the population standard deviation and the sample size:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

With a small sample size, such as 10, the standard error is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{10}} = \frac{\sigma}{3.16} = 0.32\sigma$$

The standard error here is just under one third of the population standard deviation. With a larger sample of, say, 50, the standard error becomes:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{50}} = \frac{\sigma}{7.07} = 0.14\sigma$$

This is just under a seventh of the population standard deviation. By increasing the sample size from 10 to 50 we have reduced the standard error by over a half (from a third to a seventh of the population figure). Increasing the sample size has reduced the spread of the distribution.

An increase in sample size has the effect of reducing the overlap between the distributions by reducing their standard deviations. As a result of this, more of a genuinely different unknown distribution ends up beyond the significance level (and there is an increase in power). Compare the distributions in Figure 9.2 with those of Figure 9.1. This shows the effect of

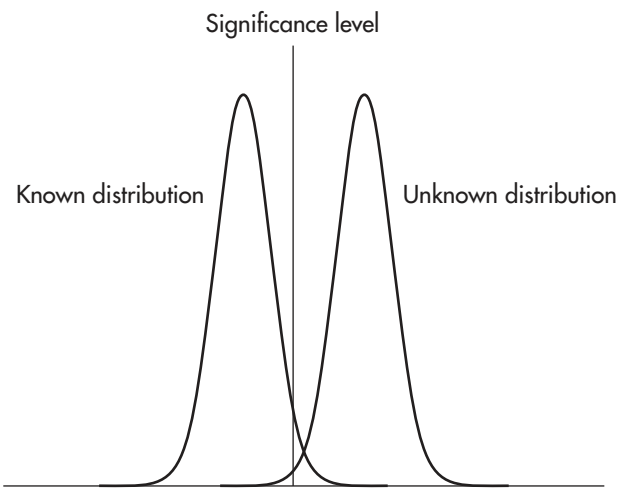

FIGURE 9.2 The effect of increasing the sample size on the overlap of the distributions

reducing the standard error by half as a result of an increase in sample size. The overlap is considerably reduced.

We can illustrate the effect of sample size on statistical power by the following example. Assume that the known population is a normal distribution with a mean of 100 and a standard deviation of 15. We will also assume that the unknown population is genuinely different with a mean of 110. In this case the population is not unknown any more so we would not need to perform any statistics as we know all we need to know – but this is for illustration only!

First we shall examine the situtation when a sample of 10 is used. The sampling distributions of the two populations will have means of 100 and 110 but their standard deviations will be the standard error: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{10}} = 4.74$. The $p = 0.05$ significance level cuts off the last 5 per cent of the known distribution. As the distribution is normally distributed we can use the z tables (Table A.1 in the Appendix) to find which z cuts off 0.05 of the distribution. This gives a value of $z = 1.65$. Remember than z is expressed in standard deviation units, so the significance level is 1.65 standard deviations above the mean of 100. The standard deviation, the standard error, of the known distribution is 4.74, so the significance level is therefore $1.65 \times 4.74 = 7.82$ above the mean of the known distribution, so is located at 107.82 on the scale.

We now perform a similar process in reverse on the unknown distribution to work out β and then $1 - \beta$. The significance level at 107.82 positions it 2.18 below the mean of the unknown distribution (110) on the scale. We convert this to standard deviation units to find z. As we assume the standard deviations of the two distributions are the same, the standard error of the unknown distribution is also 4.74, and the significance level is $\frac{2.18}{4.74} = 0.46$ standard deviations below the mean. When we look up this figure in the z tables we find that $p = 0.32$. There is 0.32 of the unknown distribution below the significance level, so $\beta = 0.32$ and the power of the test is $1 - \beta = 0.68$. There is 68 per cent of the unknown distribution above the significance level. So using a sample size of 10 gives a power of 0.68.

We can do the same calculations for a sample size of 50. In this case the standard error is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{50}} = 2.12$. The significance level is $1.65 \times 2.12 = 3.50$ above the mean of the known distribution, at 103.50. This is 6.50 below the mean of the unknown distribution, which gives a

z of $\frac{6.50}{2.12} = 3.07$. From the standard normal tables this gives $p = 0.0011$, so $\beta = 0.0011$ and the power of the test $(1 - \beta)$ is 0.9989. With a sample size of 50 we now have 99.89 per cent of the unknown distribution above the significance level. So changing the sample size from 10 to 50 has increased the power from 0.68 to 0.9989.

Choosing a sample size for a statistical test

An important decision for a researcher is deciding the appropriate number of participants for a study. This is where the work of Cohen (1988) is particularly helpful. As noted above, power is related to significance level, effect size and sample size. We can turn this relationship around and see that sample size is a function of significance level, power and effect size.

A researcher was investigating different visual displays for monitoring equipment for hospitals. Two different types of display were to be compared in the laboratory to see which one led to the fewest errors in reading the display. The researcher wanted to know how many participants to use. The researcher decided on a 0.05 level of significance for a two-tailed test. The level of power required was chosen as 0.8 and it was assumed that the effect size would be medium so 0.5 was specified as the effect size. A *t* test was to be carried out on the error data.

Can we carry out a calculation like the one in the above section to find the answer to our question? The answer is both yes and no. Yes, we can carry out a calculation to find the number of participants and no, it is not the same as the above section as that was worked out using population data which we do not have here. When we are comparing two samples we use the *t* distribution as the appropriate distribution for our analysis. However, there is a complication as the *t* distribution we usually employ for a *t* test calculation is based on samples drawn from the same distribution, i.e. when the samples come from the same population. This is the *t* distribution assuming *no effect*. Yet in our power analysis we are proposing an effect. So we have to use a special *t* distribution for our power analysis called a noncentral *t* distribution. In order to do this we need to calculate the noncentrality parameter δ which is quite easy as it is a function of d (the effect size) and the sample sizes ($\delta = \mathrm{d} \times \sqrt{\frac{n_1 \times n_2}{n_1 + n_2}}$,

where n_1 and n_2 are the sizes of the samples. With the noncentrality parameter (δ), the significance level (α) and the degrees of freedom (df) the power values can be calculated and, by a little reorganisation of the calculations, the sample size can be produced for a specific level of power.

Unfortunately, power analysis is not something to do by hand as we need to use distribution tables for the noncentral distribution. Cohen (1988) provided sets of useful tables that can be used to find the appropriate values. However, there are a number of easy-to-use software packages that can work out the power calculation and required sample sizes. A number of these are available free (for noncomercial use) e.g. GPOWER[6] which is very easy to use. The required values for the significance level, the effect size and the power are input and the output gives the required sample sizes. Using such a software package we can find that for an unrelated two-tailed test, using the 0.05 level of significance, examining a medium effect (d = 0.5) and seeking a power of 0.8 we need 128 participants in total or 64 in each sample.

Quite often we will find that in order to achieve the power required the sample sizes will be very large. If we had been examining a small effect (d = 0.2) in the above example we would have needed 788 participants in total or 394 in each sample. If we decide that it is not feasible to use groups of this size we can undertake a compromise power analysis. Rather than seek a sample size for a specific power (e.g. 0.8) we decide on the balance of risk we are willing to accept between a Type I and a Type II error: the ratio of β to α: where $q = \beta/\alpha$. If we decide that q = 3 is the balance of risk we can work with and we can afford to test 100 participants in each group then we can work out the power for this compromise. In this case the power is 0.5, so would be a test of medium power. We might be content with this compromise solution.

Finally, I have focused on the power calculations for a two sample *t* test. However, we can work out both an effect size and a noncentral distribution for a range of other statistics included in this book. So we can work out the power (or the sample size for a specific power) for the different tests we shall be considering. Fortunately, the software packages allow us to examine the power of different tests by including a menu where we simply select the statistical test we wish to perform. In the table below the 'conventional' effect size values for small, medium and large effects (from Cohen, 1988) are shown for a number of key statistical tests.

Test	*Effect size*	*Small effect*	*Medium effect*	*Large effect*
t test	d	0.2	0.5	0.8
Correlation	r	0.1	0.3	0.5
ANOVA	f	0.1	0.25	0.4
Multiple correlation and regression	f^2	0.02	0.15	0.35
Chi-square	w	0.1	0.3	0.5

Conclusion

Hypothesis testing involves making a decision concerning whether two distributions are the same or different. To make this decision we use a decision criterion, the significance level. Due to the overlap of the distributions the significance level cannot separate them completely when they are genuinely different to each other. As a result we end up with α of the known distribution and β of the unknown distribution the 'wrong' side of it. To limit the risk of Type I errors we set our significance level so that $\alpha = 0.05$, giving us a 5 in 100 chance, or smaller, of falsely rejecting the null hypothesis. We don't want to make Type I errors (and sometimes we are even more conservative, setting the significance level at $p = 0.01$, reducing the risk to 0.01).

This leaves β, the risk of making a Type II error. We do not have the same control over β as we do with α, as the distribution is unknown. Yet we do not want to use a test that is low in power, $1 - \beta$, as it reduces our chances of finding a real effect when it is there. Unfortunately, researchers do too often use tests of low power. To increase the power of our test we can do three things: design better studies, choose one-tailed tests, look for big effects, and increase the number of subjects.

When trying to decide whether the power of a test is adequate there are a couple of useful points to consider. Select the largest sample size you can *sensibly* test. If you have limited resources, time or access to subjects these restrictions may have priority. Then check the power of your test. If the power of your test is too low then you may be wasting your time continuing. However, consider the balance of risk of Type I and Type II

errors. You may wish to continue with the research as you have a reasonable compromise of α and β. If you find a significant effect then you do not need any more subjects. If the test yields no significant differences yet is unexpected, or approaches significance, then repeat the test when you can test more subjects. It is worth increasing sample size to increase both the power of the test and your confidence in the findings. The new subjects may confirm the previous results or produce a significant difference. One of the major ways of deciding whether a finding is worthwhile or not is to replicate (repeat) it. If a difference continues to be significant then other researchers are more likely to accept its validity.

To recap for a moment: all we are doing is trying to decide if a 'score' comes from one distribution or another. The overlap in distributions, when the distributions are different, makes it difficult to avoid the risk of error in setting our decision criterion, the significance level. We set the risk of a Type I error (α) by choosing the significance level. Yet we should not ignore β, the risk of a Type II error, as it is no fun trying to dig up treasure with a plastic bucket and spade. Increasing the power of a test reduces β and gives us a better chance of finding treasure when it really is there.