Hypothesis testing and statistical significance

5

CHAPTER OVERVIEW

In Chapter 4 we started you off on the road to using inferential statistics. In this chapter we will move a little further down the road and explain how we can apply our knowledge of probabilities and sampling distributions to testing the hypotheses that we set up in our research. More specifically, we will be explaining the following:

- the logic of hypothesis testing
- statistical significance and how it relates to probabilities
- how probability distributions form the basis of statistical tests
- the problems associated with basing conclusions on probabilities (i.e. Type I and Type II errors)
- one-tailed and two-tailed hypotheses
- how to choose the appropriate test to analyse your data.

5.1 Another way of applying probabilities to research: hypothesis testing

Suppose we were interested in examining the relationship between number of hours spent studying per week and exam grades. We would perhaps predict that the more time spent studying per week, the higher the exam grades. Here we have set up a prediction that we would then test by conducting a study. In this study we could randomly select a number of students and record how many hours they spent per week studying and find out if it is related to their final exam grade. According to our prediction, we would expect the population of scores to resemble that in the population illustrated in Figure 5.1. Here you can see there is a trend indicating that as the number of hours studied increases, so do exam grades. Let us assume that this is the pattern in the underlying population. One of the problems we face when conducting research is that when we select samples from populations we might not get a sample that accurately reflects that population. If you think back to Chapter 3, we explained that due to sampling error the samples might not resemble the population. Figure 5.1 illustrates three samples taken from the population presented therein. You should notice that even though there is a positive relationship in the population of scores, two of the samples do not reflect this. In fact, one of the samples actually suggests a negative relationship between hours studied and exam performance (as number of hours spent studying increases, exam performance decreases). Another of the samples suggests that there is no relationship between the two

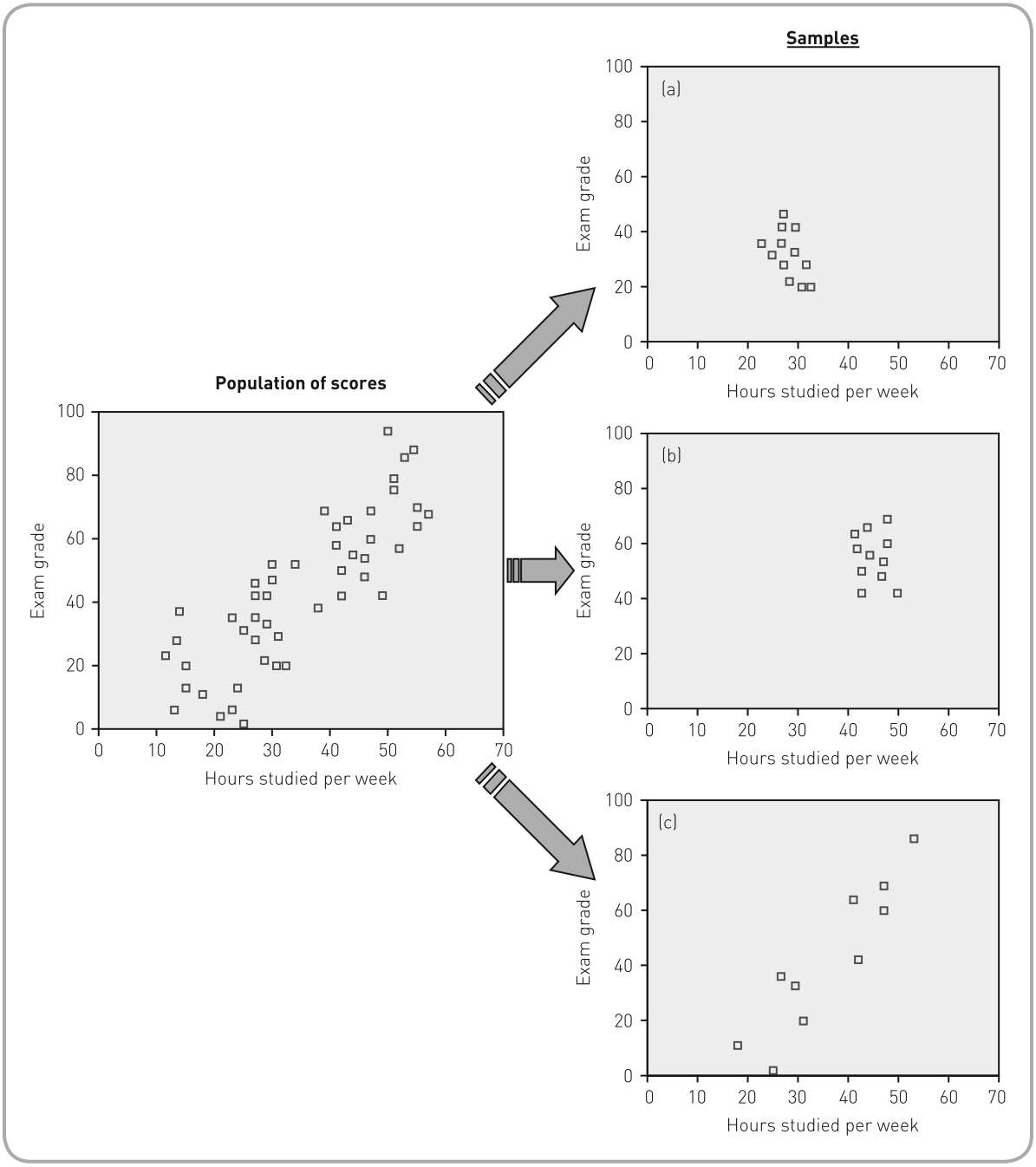

Figure 5.1 Scattergrams illustrating possible samples selected from a population with a positive relationship between number of hours spent studying and exam grades

variables. The remaining sample accurately reflects the underlying population by suggesting a positive relationship between the two variables. The point to note here is that, even though there is a relationship in the underlying population, the sample we select might not reflect this.

Now take a look at Figure 5.2. In this example there is no relationship between amount of time spent studying and exam performance in the underlying population. Again, we have presented three samples that have been selected from the population. Yet again, only one of the samples accurately reflects the population. The fact is that, due to sampling error, the

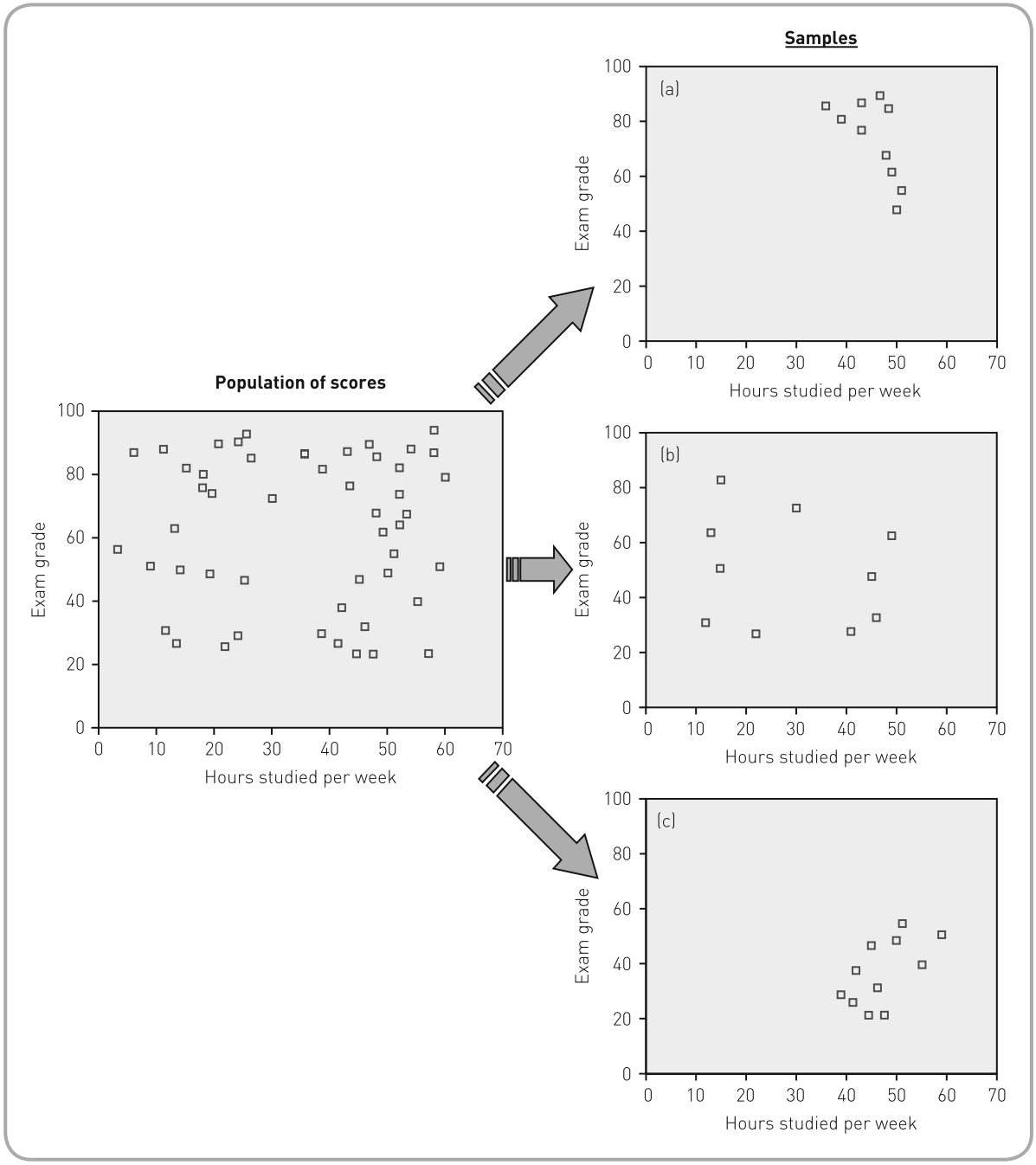

Figure 5.2 Scattergrams illustrating possible samples selected from a population with no relationship between number of hours spent studying and exam grades

samples we select might not be a true reflection of the underlying population. From any particular population, each of the patterns of sample scores we have presented will have a greater or lesser probability of being selected and this probability will depend on the size of the sample we select. Thus, for the population in Figure 5.1 we are more likely to get a pattern resembling that observed in sample (c) than those in samples (a) and (b), particularly with reasonably large sample sizes. And for the population presented in Figure 5.2 we are more likely to get the pattern resembling that in sample (b) than the ones in samples (a) and (c).

You need to be aware, though, that sometimes simply due to sampling error we are likely to get patterns of scores in our samples that do not accurately reflect the underlying population.

One of the problems we face when conducting research is that we do not know the pattern of scores in the underlying population. In fact, our reason for conducting the research in the first place is to try to establish the pattern in the underlying population. We are trying to draw conclusions about the populations from our samples. Essentially, we are in a situation akin to that illustrated in Figure 5.3. In this figure everything above the dashed line relates to what we have observed in our study and everything below the line is unknown to us. From the pattern of data we observe in our sample, we have to try to decide what the pattern may look like in the population. There may be an infinite number of possible patterns that reflect the population; however, we have given only two of these in the figure. From our sample we have to decide what we think the population is like. This is where we would use inferential statistical tests. Effectively what we do is observe the pattern of scores in the sample and decide which is the most plausible pattern in the population. Thus, given the pattern observed in the sample in Figure 5.3, we might argue that the pattern in population (b) is much more plausible than that shown in population (a). As is illustrated by Figures 5.1 and 5.2, however, the samples need not be an accurate reflection of the population. We therefore need some means of deciding, from the evidence presented by our sample data, what the most plausible pattern in the population might be.

Our statistical tests help us in this decision, but they do so in a way that is not very intuitive. What our statistical tests do is calculate a probability value, called the *p-value*. This probability tells us the likelihood of our obtaining our pattern of results due to sampling error if there is no relationship between our variables in the population. For example, they would tell us the probability of our obtaining the pattern of scores in the sample in Figure 5.3 if they came from population (a). If the pattern in our sample is highly unlikely to have arisen due to sampling error if the population resembles (a), we might reasonably conclude that the population resembles that in (b). You should note that this probability value is a conditional probability. It is the probability of obtaining your sample data *if* there was no relationship between the variables in the population.

Definition

The *p-value* is the probability of obtaining the pattern of results we found in our study *if* there was no relationship between the variables in which we were interested in the population.

Hypothesis testing is often seen as a competition between two hypotheses. It is seen as a competition between our research hypothesis (that there is a relationship between study hours and exam grade in the population) and something called the *null hypothesis* (that there is no relationship between the two variables in the population). Thus, the process of hypothesis testing resembles Figure 5.3. We need to decide between populations (a) and (b). In this case, population (a) represents the case if the null hypothesis were true and population (b) represents the case if the research hypothesis were true. The statistical tests we use tell us how likely it is that we would get our pattern of data if the null hypothesis were true. In Figure 5.3, we would probably find that the pattern of data in the sample would be highly unlikely to occur as the result of sampling error if they were drawn from a population resembling (a) where there is no relationship between hours spent studying per week and exam grade. In fact, the probability turns out to be less than 1 in 1000. In this case, it would make more sense to conclude that the data came from a population that resembles that illustrated in (b).

Now, let's have a look at the scenario represented by Figure 5.4. Remember that everything above the dashed line is what we observe from our study and everything below the line is unknown to us. Here you should be able to see that the sample appears to suggest that there

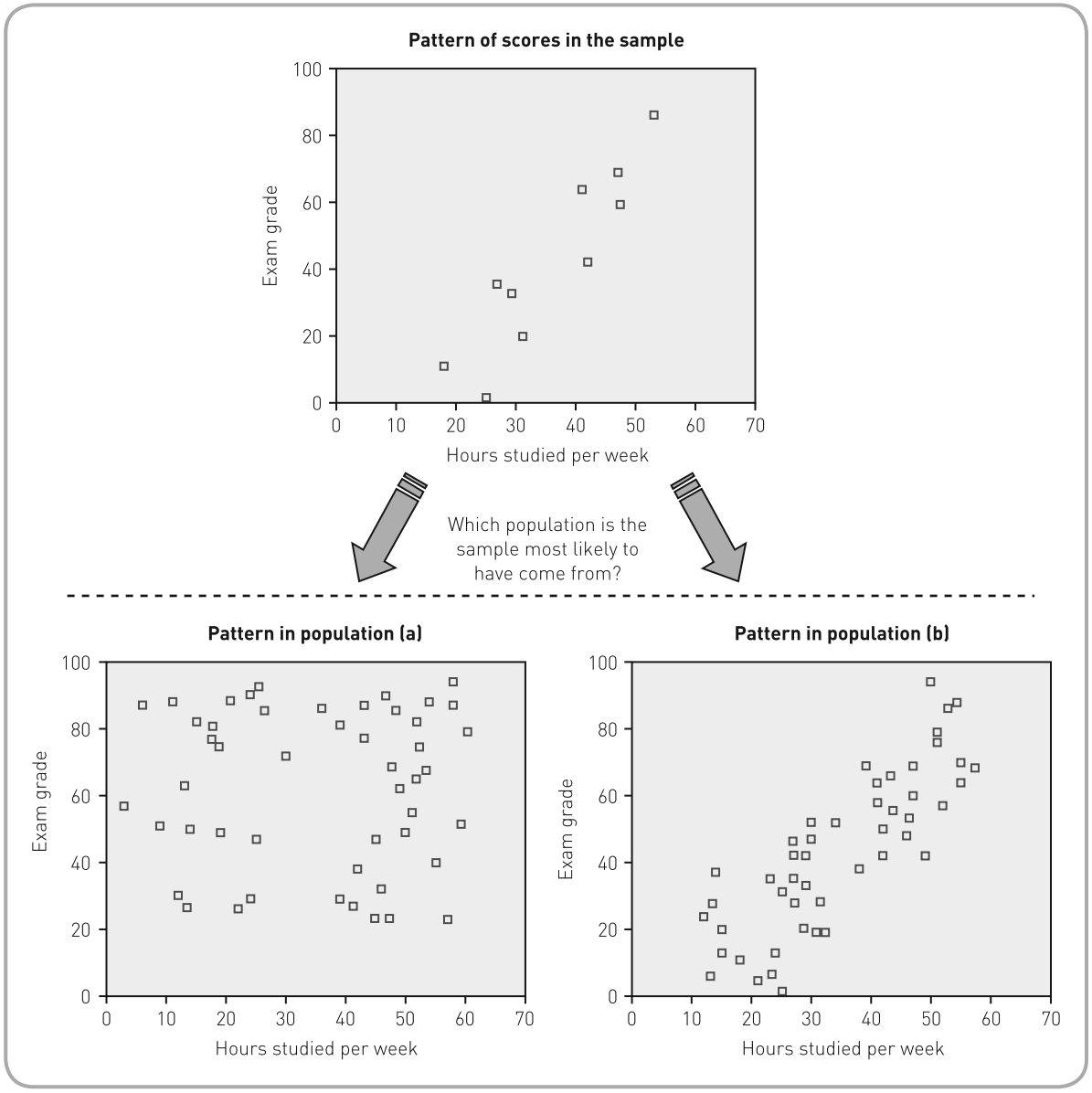

Figure 5.3 Scattergrams illustrating alternative underlying populations when a relationship is observed in a sample

is no discernible relationship between number of hours spent studying and exam grade. Intuitively, we would expect that this sample has come from a population resembling that shown in (a) rather than that shown in (b). However, again referring to Figure 5.1, you should be able to see that even when there is a relationship between two variables in the population we have the possibility that one will not be observed in our sample. This absence of a relationship in the sample would be the result of sampling error. So again in this case we could use inferential statistical tests to help us choose between the two hypotheses: the null hypothesis represented by population (a) or the research hypothesis represented by population (b). The statistical test would inform us of the probability that we would obtain the pattern in our sample illustrated in Figure 5.4 if the population resembled the pattern shown in (a): that is, if the null hypothesis were true. In this case we would find that there is a high probability of obtaining the pattern observed in our sample if the null hypothesis were true. In fact, there is

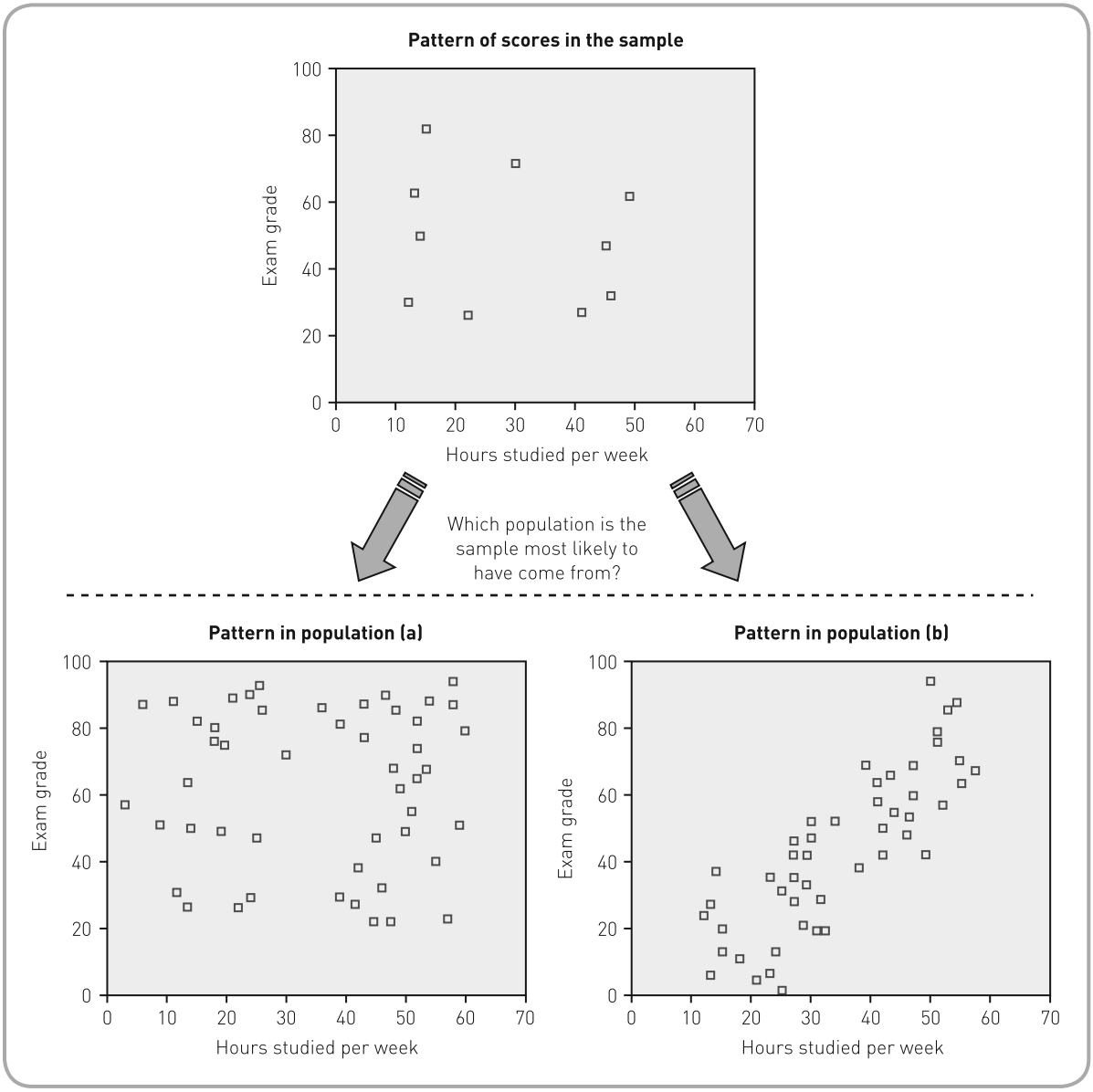

Figure 5.4 Scattergrams illustrating alternative underlying populations when no relationship is observed in a sample

a 61% probability of obtaining this pattern from a population resembling that shown in (a). In this case we would probably decide that the population does in fact resemble population (a) rather than population (b). There are other issues that we would need to address, however, before we could come to this conclusion, such as whether we had enough participants in our sample (see section 5.9 and Chapter 8).

5.2 Null hypothesis

We have just slipped a very important concept past you, which needs further explanation. The *null hypothesis* is very important to the process of hypothesis testing. We explained earlier

that the probability we calculate in statistical testing is based upon the assumption that there is no relationship between the two variables in the population. This assumption is the *null hypothesis*. If the research hypothesis (often called the *experimental* or *alternate* hypothesis) states that there will be a relationship between two variables, then the null hypothesis states that there is absolutely no relationship between the two variables. Similarly, if you are interested in comparing groups of people, where the research hypothesis states that there will be a difference between two groups, the null hypothesis states that there is no difference between them.

Definition

The *research hypothesis* is our prediction of how two variables might be related to each other. Alternatively, it might be our prediction of how specified groups of participants might be different from each other or how one group of participants might be different when performing under two or more conditions.

You may find when reading psychological journals that the authors suggest that the null hypothesis could not be rejected. This simply indicates that the statistical probability they calculated meant that it was likely that the null hypothesis was the more sensible conclusion. If you read about researchers rejecting the null hypothesis, it means that the probability of obtaining their findings if the null hypothesis were true is so small that it makes more sense to believe in the research hypothesis. As we indicated earlier in this section, this illustrates the competition between our null and research hypotheses. The importance of the null hypothesis is reflected by the fact that this whole approach to conducting research is called *null hypothesis testing (NHT)* or *null hypothesis significance testing (NHST)*.

Definition

The *null hypothesis* always states that there is no effect in the underlying population. By effect we might mean a relationship between two or more variables, a difference between two or more different populations or a difference in the responses of one population under two or more different conditions.

5.3 Logic of null hypothesis testing

If you understand the preceding sections, you should have no problems with grasping the general logic behind hypothesis testing, which is as follows:

- Formulate a hypothesis.
- Measure the variables involved and examine the relationship between them.
- Calculate the probability of obtaining such a relationship if there were no relationship in the population (if the null hypothesis were true).
- If this calculated probability is small enough, it suggests that the pattern of findings is unlikely to have arisen by chance and so probably reflects a genuine relationship in the population.

Put another way, if there is no real relationship in the population, you are unlikely to find a relationship in your randomly selected sample. Therefore, if you do find a relationship in your sample, it is likely to reflect a relationship in your population. It is important that you understand this, so take your time and ensure that you follow what we have just said.

Hypothesis testing is not restricted to the investigation of relationships between variables. If we are interested in studying differences between groups, we can also test hypotheses. The logic is broadly the same as that outlined for relationships above. For example, suppose we set up a study where we gave students two types of structured study, which differed only in the amount of time the students were required to study. In one group the students studied for 40 hours per week and the other group studied for 10 hours per week (this is the independent variable). We might hypothesise that the 40-hour group would achieve higher exam marks than the 10-hour group. This would be our research hypothesis. Our null hypothesis would be that there would be no difference between the two groups in their exam grades in the population. Once we have collected the data, we could then see if there is a difference between the two study groups. If a difference did exist, we would then need to work out the probability of obtaining such a difference by sampling error alone: that is, the probability of obtaining a difference of the size observed if the null hypothesis were true. If this probability is low, then it makes more sense to assume that the difference was due to the manipulation of the independent variable rather than to sampling error alone.

Activity 5.1

Take a look at this extract from the paper published by Ingram *et al*. (2009):

> Therefore, the aim of this study was to determine the efficacy of both COLD and contrast water immersion (CWI) as recovery methods in the 48 h period following simulated team sport exercise. It was hypothesised that COLD and CWI would be superior recovery methods compared to a control condition.

Try to work out what the null hypothesis would be in this case.

Discussion point

Criticisms against null hypothesis testing

Although null hypothesis testing is the dominant approach to research in psychology, today there is growing concern that it is inadequate in terms of providing useful insights into the variables that psychologists wish to investigate. For example, referring to hypothesis testing, Loftus (1991) says, 'I find it difficult to imagine a less insightful means of transiting from data to conclusions'. Loftus (1991, 1996) describes many problems associated with the use of hypothesis testing, but we will highlight two here. If you wish to read more, there are two references at the end of this chapter.

One of the main problems highlighted by Loftus relates to the null hypothesis. When we are looking for a difference between two conditions, we have to calculate the probability of obtaining our difference by chance if the null hypothesis is true. Remember, the null hypothesis states that there is *no* difference between the two conditions. The problem with the null hypothesis is that in few instances, in any science, will there be no difference between two conditions. It is quite unusual to find two things that are exactly equal, even in physics, and so to base our probability judgements on such a null hypothesis may be seriously misleading. This is just the gist of the point made by Loftus, but it serves to illustrate one of the criticisms raised by him.

The second problem that Loftus highlights is that, although we may report with some confidence that we have found a genuine difference between our two conditions and report the size of the difference, psychologists usually say very little about the underlying population means of the two conditions.

Loftus argues that hypothesis testing lures us away from thinking about the population means. He suggests that we can avoid this trap by routinely reporting confidence intervals in our research reports. For a more recent contribution to the debate concerning null hypothesis testing and confidence intervals, see Denis (2003).

Even though there are such criticisms levelled at the process of hypothesis testing, it does not mean that we should abandon this approach completely; rather, we need to have a thorough understanding of what it means to engage in hypothesis testing. This is what we hope to give you in this book. Therefore, alongside the statistical tests that help us test our hypothesis (e.g. the t-test) you should, as Loftus suggests, routinely report descriptive statistics and confidence intervals. One useful way of presenting confidence intervals is by generating *error bar charts* and presenting these in your reports. We have shown you what these are like earlier in this book (see Chapter 4).

Activity 5.2

Which of the following descriptions represents a good summary of the logic behind hypothesis testing?

(a) We measure the relationship between the variables from our sample data. If it is large, there must be a genuine relationship in the population.
(b) We measure the relationship between the variables from our sample and then find the probability that such a relationship will arise due to sampling error alone. If such a probability is large, we can conclude that a genuine relationship exists in the population.
(c) We measure the relationship between the variables from our sample and then find the probability that such a relationship will arise due to sampling error alone. If such a probability is small, we can conclude that a genuine relationship exists in the population.
(d) We measure a difference between our two conditions and then work out the probability of obtaining such a difference by sampling error alone if the null hypothesis were true. If the probability is small, we can conclude that a genuine difference exists in the population.

5.4 The significance level

Many of you, at this point, may be thinking that this is all well and good but how do we decide that the probability we calculate in null hypothesis testing is small enough for us to reject the null hypothesis? This is an excellent question and one that does not have a definitive answer. Most psychologists and indeed most reputable psychology journals use the convention that a probability of 5% is small enough to be a useful cut-off point. That is, given that the null hypothesis is true, if the probability of a given effect is less than 5% (0.05 or 1 in 20) then we have provided reasonable support for our research hypothesis. What this means is that, if you conduct a study 20 times, only once in those 20 studies would a relationship (or difference) as large as the one you observe come out by chance, if the null hypothesis were true. Given such a low probability, we can conclude with reasonable confidence that a real relationship (or difference) exists in the populations under investigation. The probability associated with each statistical test is often called the *p-value* or *alpha* (α). When this is printed on your SPSS output, it will be printed as a decimal and, as with all probabilities expressed as a decimal, it ranges from 0 to 1.

Definitions

The *p-value* for a particular inferential statistical test is the probability of finding the pattern of results in a particular study if the relevant null hypothesis were true. This is a conditional probability.

Alpha (α) is the criterion for statistical significance that we set for our analyses. It is the probability level that we use as a cut-off below which we are happy to assume that our pattern of results is so unlikely as to render our research hypothesis as more plausible than the null hypothesis.

In many journals you will typically see researchers reporting their findings as *significant* or *not significant*. On the assumption of the null hypothesis being true, if the probability of obtaining an effect due to sampling error is less than 5%, then the findings are said to be 'significant'. If this probability is greater than 5%, then the findings are said to be 'non-significant'. This way of thinking about your analysis has, however, come in for a good deal of criticism in recent years for the reasons discussed on pages 139–40.

Definition

When we find that our pattern of research results is so unlikely as to suggest that our research hypothesis is more plausible than the null hypothesis, we state that our findings are *statistically significant*. When we find that our pattern of data is highly probable if the null hypothesis were true, we state that our findings are *not significant*.

The conventional view today is that we should report exact probability levels for our test statistics (the exact p-value or α) and shift away from thinking in terms of whether or not the findings are statistically significant. Therefore, when reporting the results of your analyses you should report the exact probability values that are associated with your findings. We have presented the significant/non-significant view here so that you will know what it means when you come across such statements in journal articles.

We recommend that you use the 5% level of α as a rough guide to what has traditionally been seen as an acceptable probability of your findings being due to sampling error. Therefore, if you find that your p-value is a lot less than the 5% level, you can be reasonably confident that this is generally acceptable as indicating support for your research hypothesis. However, you should report the actual p-value and evaluate your findings in terms of effect size (see Chapter 8) and your error bar charts.

Activity 5.3

Suppose you have conducted a study looking for a difference between males and females on preference for action films. When you run your study, you find that there is a 0.005 probability of the difference you observe arising due to sampling error. How often is such a difference likely to arise by sampling error alone?

(a) 1 in 5000
(b) 1 in 2000
(c) 1 in 500
(d) 1 in 200
(e) 1 in 100

Suppose the probability was 0.01: which of the above is true in this situation?

5.5 Statistical significance

As suggested previously, when reading an article from a psychological journal or listening to eminent and not-so-eminent psychologists describe their research, you will often hear/read the word 'significant'. Psychologists say things like:

> Final year students who had taken a placement year achieved significantly higher marks in their final year.
>
> (Reddy and Moores, 2006)

> The panic disorder group did have significantly more emotional processing difficulties than the control groups . . .
>
> (Baker *et al*., 2004)

> . . . academic procrastination resulting from both fear of failure and task aversiveness was related significantly to worth of statistics, interpretation anxiety, test and class anxiety, computational self-concept, fear of asking for help, and fear of the statistics instructor.
>
> (Onwuegbuzie, 2004)

What are we to make of statements such as these? In everyday life we interpret the word 'significant' to mean considerable, critical or important. Does this mean that Reddy and Moores found considerably higher marks for those who had taken a placement year? Or that Baker *et al.* found a critical difference between their panic and control groups, or perhaps Onwuegbuzie found an important relationship between fear of failure and worth of statistics? In fact, they do not necessarily mean this. They are merely stating that what they found was *statistically significant.* Statistical significance is different from psychological significance. Just because a statistically significant difference is found between two samples of scores, it does not mean that it is necessarily a large or psychologically significant difference. For example, in the study cited there was a significant impact of a placement year on final-year marks. However, the placement year only accounts for between 3% and 5% of the differences between the two groups and this is not necessarily a psychologically significant difference (we will explain this further in Chapter 8).

As we have already explained, the probability we calculate in inferential statistics is simply the probability that such an effect would arise if there were no difference between the underlying populations. This does not necessarily have any bearing on the psychological importance of the finding. The psychological importance of a finding will be related to the research question and the theoretical basis of that research. One of the main problems with the p-value is that it is related to sample size. If, therefore, a study has a large number of participants, it could yield a statistically significant finding with a very small effect (relationship between two variables or difference between two groups). It is up to individual authors (and their audiences) to determine the psychological significance of any findings. Remember, *statistical significance does not equal psychological significance.*

Discussion point

Why report the exact p-value (α)?

There is quite a debate going on in psychology concerning the use of the alpha criterion of significance. The generally accepted criterion of significance ($p < 0.05$) is coming under increasing criticism. There is nothing intrinsically wrong with the 5% cut-off, yet it has been argued that the pursuance of this as the Holy Grail in psychology is distorting the legitimate goals of psychological research. The problem with the 5% criterion is that we are often led to believe that just because some effect is statistically significant then it is psychologically significant, or even that it is a large or important effect. In fact, if we look at this criterion logically, we can see the folly of this way of thinking. Suppose, for example, that you conducted a study looking at the relationship between statistics anxiety and procrastination. You find that, as statistics anxiety increases, so does procrastination. You find that the probability of obtaining such a relationship, if there really was no relationship in the population, is 4.9%. As this is less than the traditional 5%, you conclude that this is a real relationship between statistics anxiety and procrastination. You then conduct a follow-up study (being the good researcher that you are) and again find a relationship between statistics anxiety and procrastination. This time, however, you find that the probability of such a relationship, given that the null hypothesis is true, is 5.1%. What are we to make of this? Do you now conclude that there is no real relationship between statistics anxiety and procrastination? You can see that there is only 0.2% difference in the probability values between these two studies. So it does not really make sense to argue that the sizes of the relationship in the two studies are different. Yet, in all probability, the first of these would get published in a psychological journal and the second would not.

One of the big problems with the p-value is that it is related to sample size. We could have two studies where one has a very small p-value (say, 0.001) and one has quite a large p-value (say, 0.15). Yet we would not be able to say that the first study shows a large effect (strong relationship or large difference between conditions) and the second study a small effect. In fact it could be the reverse situation because it might simply be the case that the first study has a very large sample size and the second a small one. Even very small effects will lead to significant statistical tests with very large sample sizes.

How can we get around this problem? The best approach to this is to try to get a measure of the magnitude of the experimental effect: that is, to get information about the size of the relationship between statistics anxiety and procrastination. If you were looking for differences between groups, you would get a measure of the size of the difference between your groups. This is called the *magnitude of effect or effect size*. A more detailed description of effect size can be found in Chapter 8. The preferred course of action when reporting your research findings is to report the exact probability level and the effect size. For example, you should report the probability level (i.e. $p = 0.027$) and the effect size (e.g. $r = 0.70$, $r^2 = 0.49$ or $d = 0.50$). In this way, whenever someone reads about your research, he or she can get a fuller picture of what you have found. You should note that r is a correlation coefficient and indicates the strength of a relationship between variables (we explain this more in the next chapter); d is a measure of magnitude of effect used for differences between groups and is explained in Chapter 7. There is a very accessible discussion of effect sizes provided by Clark-Carter (2003).

5.6 The correct interpretation of the p-value

It is important to understand that the p-value is a conditional probability. That is, you are assessing the probability of an event's occurrence, given that the null hypothesis is true. The p-value that you will observe on any computer printout represents this probability. It does not represent the probability that the relationship you observed simply occurred by chance.

It represents the probability of the relationship occurring by chance if the null hypothesis were true. It is said to be a conditional probability. It is conditional upon the null hypothesis being true. A good discussion of the problems caused by misinterpreting what the p-value represents is given by Dracup (1995); however, we have summarised the main points in the discussion below. If you wish to read the original discussion, the reference is given at the end of the chapter.

Discussion point

Misinterpretation of the significance level (α)

Dracup (1995) has given a good discussion of the problems associated with the misinterpretation of the rationale behind hypothesis testing.

Many students new to statistics, and indeed those who perhaps should know better, equate the significance level (α) with the actual size of the experimental effect. The lower the significance level, the stronger, for example, the relationship between two variables. This is not what is meant by the significance of a finding. Alpha simply gives an indication of the likelihood of finding such a relationship if the null hypothesis were true. It is perhaps true that the stronger the relationship, the lower the probability that such a relationship would be found if the null hypothesis were true, but this is not necessarily so.

Dracup also highlights the fact that many statistical textbooks equate α with the probability that the null hypothesis is true. This is incorrect, as is clearly illustrated by Dracup. Alpha is the probability that we will get a relationship of an obtained magnitude if the null hypothesis were true. It is not the probability of the null hypothesis being true.

Related to this latter point, once someone has fallen into the trap of interpreting α as the probability of the null hypothesis being true, it is a relatively easy and convenient step to suggest that 1 – α must be the probability that the research hypothesis is true. Thus, if we set α at the traditional 5% level and find a significant relationship, these people would assume that there is a 95% probability that the research hypothesis is true. This is incorrect. In fact, we do not know what the probability is that the research hypothesis is correct; our α probability is conditional upon the null hypothesis being true and has nothing to do with the truth or falsity of the research hypothesis.

It is important to remember that what we have just explained about relationships is also relevant when looking for differences between groups. Thus, the p-value is the probability of finding a difference between two groups if the null hypothesis (no difference in the population) were true.

Activity 5.4

Imagine that you have conducted two separate studies and found a relationship between head size and IQ in study 1 and head size and shoe size in study 2. The probability of observing the relationship in study 1 by chance if the null hypothesis were true is found to be 0.04, whereas in study 2 the probability is 0.001. Which of these findings is the more important psychologically?

5.7 Statistical tests

Imagine you are investigating the relationship between number of hours spent studying and exam performance. Now suppose you have conducted a study and have found a pattern of scores similar to that given in the sample presented in Figure 5.3. How do you go about

calculating the probability that such a relationship is due to sampling error if the null hypothesis were true? This is where we need to use inferential statistical tests such as the Pearson product moment correlation coefficient (see Chapter 6). If you had conducted a study that examined the difference between two conditions of an independent variable, you would use a test such as the t-test to calculate your probability. In the rest of this section we hope to give a conceptual understanding of what statistical tests actually do.

When we look at the relationship between two variables (e.g. hours spent studying and exam grade), we are able to calculate a measure of the size or strength of the relationship (this is covered in more detail in the next chapter). Once we have a measure of the strength of a relationship, we need to find out the probability of obtaining a relationship of such strength by sampling error alone. In order to calculate the probability, we can make use of the probability distributions to which we introduced you in Chapter 4 (e.g. see page 100). Earlier we told you that the probability of obtaining any particular score from probability distributions is known. For example, the probability of obtaining a z-score of 1.80 or higher is only 3.8%. If we are able to convert the information we have about the strength of a relationship into a score from a probability distribution, we can then find the probability of obtaining such a score by chance. This would then give us an indication of the probability of obtaining the relationship we observe in our study by sampling error (by chance) if no such relationship really existed in the population. This is basically what significance testing involves. Converting the data from our samples into scores from probability distributions enables us to work out the probability of obtaining such data by chance factors alone. We can then use this probability to decide which of the null and experimental hypotheses is the more sensible conclusion. It should be emphasised here that these probabilities we calculate are based upon the assumption that our samples are randomly selected from the population.

Figure 5.5 shows the standard normal distribution and illustrates that the probability of obtaining scores in the extremes of the distribution is very small. You should remember that when looking at probability distributions the area under the graph represents probability. The larger the area above a positive score, the greater the probability of obtaining such a score or one larger. Similarly, the larger the area below a negative score, the greater the probability of obtaining that score or one smaller. Thus, once we have converted the degree of relationship between the variables into a score from a probability distribution, we can work out the

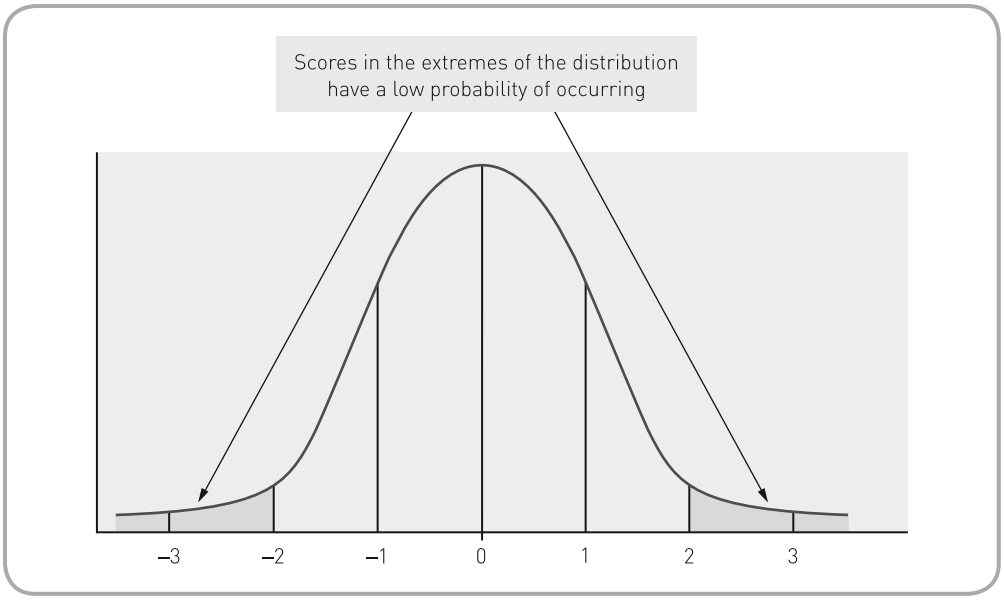

Figure 5.5 Diagram illustrating the extreme scores in a distribution

probability of obtaining such a score by chance. If the score is in either of the two regions indicated in Figure 5.5, then we can conclude that the relationship is unlikely to have arisen by chance – that is, it is unlikely to have been the result of sampling error if the null hypothesis were true.

Of course if we were investigating differences between groups we could also use probability distributions to find out the probability of finding differences of the size we observe by chance factors alone if the null hypothesis were true. In such a case we would convert the difference between the two groups of the independent variable into a score from a probability distribution. We could then find out the probability of obtaining such a score by sampling error if no difference existed in the population. If this probability is small, then it makes little sense to propose that there is no difference in the population and that the difference between our samples is the result of sampling error alone. It makes more sense to suggest that the difference we observed represents a real difference in the population. That is, the difference has arisen owing to our manipulation of the independent variable.

It is important to note that when we convert our data into a score from a probability distribution, the score we calculate is called the *test statistic*. For example, if we were interested in looking for a difference between two groups, we could convert our data into a t-value (from the t-distribution). This t-value is called our test statistic. We then calculate the probability of obtaining such a value by chance factors alone and this represents our p-value.

5.8 Type I error

Suppose we conducted some research and found that, assuming the null hypothesis is true, the probability of finding the effect we observe is small – as would be the case represented in Figure 5.3. In this case we would feel confident that we could reject the null hypothesis. Now suppose there really is no such effect in our population and we have stumbled across a chance happening. We have obviously made a mistake if we conclude that we have support for our prediction. Statisticians would say that in rejecting the null hypothesis in this case we have made a Type I (one) error.

Definition

A *Type I error* is where you decide to reject the null hypothesis when it is in fact true in the underlying population. That is, you conclude that there is an effect in the population when no such effect really exists.

If your p-value (α) is 5% then you will have a 1 in 20 chance of making a Type I error. This is because the p-value is the probability of obtaining an observed effect, given that the null hypothesis is true. It is the probability of obtaining an effect as a result of sampling error alone if the null hypothesis were true. We argued that if this is small enough then it is unlikely that the null hypothesis is true. But as the above case illustrates, we can be mistaken; we can make a Type I error. Therefore the p-value also represents the probability of your making a Type I error. If your p-value is 5%, it means that you have a 5% probability of making a Type I error if you reject the null hypothesis. Although this probability is small, it is still possible for it to occur. We can relate this to the National Lottery. There is only about a 1 in 14 million probability of your winning the Lottery if you pick one line of numbers. Even though this represents a tiny chance of winning, the possibility still exists, which is why people keep playing it. So beware, even if you find you have a p-value of only 0.001% there is still a very small probability of your making a Type I error if you decide to reject the null hypothesis.

5.10 Why set α at 0.05?

You may be wondering why we have a cut-off for α of 0.05. Who determined that 0.05 was an appropriate cut-off for allowing us to reject the null hypothesis, rather than say 0.2 or 0.001? Although this is a fairly arbitrary cut-off, there is a rationale behind it. Let us have a look at the situations where we set α at 0.2 and 0.001 respectively. If we set α at 0.2, we would be tolerating a Type I error in one case in every five. This is a very liberal criterion for significance. In one case in every five we would reject the null hypothesis when it is in fact true. On the positive side, we would be much less likely to make a Type II error. That is, we would be much less likely to accept the null hypothesis when it is false. With such a liberal criterion for significance, we are generally going to reject the null hypothesis more often and therefore are more likely to reject it when it is false (as well as more likely to reject it when it is true). This means a lower probability of a Type II error.

So, how about setting our α at 0.001? Here we are much less likely to make a Type I error. We are only likely to reject the null hypothesis when it is true one time in every thousand. This is a very conservative criterion for significance. On the face of it, this would appear to be a very good thing. After all, we don't want to incorrectly reject the null hypothesis, and so why not set a conservative criterion for significance? The problem here is that, although we reduce the probability of making a Type I error, we also increase the probability of not rejecting the null hypothesis when it is false. We increase the probability of making a Type II error. The reason for this is that with such a conservative criterion for significance, there are going to be fewer times when we reject the null hypothesis. Therefore, we are going to increase the likelihood of not rejecting the null hypothesis when it is false.

When setting our criterion for significance, we therefore need to strike the right balance between making Type I and Type II errors. In most situations an α of 0.05 provides this balance. You should note that there are sometimes other considerations which should determine the level at which you set your criterion for significance. For example, if we were testing a new drug, we should be much more conservative, as the consequence of making a Type I error could be very serious indeed. People may be given drugs that have nasty side-effects and yet not be effective in treating what they are supposed to treat. Another situation where you may want to set a different criterion for significance is where you conduct many statistical analyses on the same set of data. This is covered in more detail in Chapter 10 (see page 308).

5.11 One-tailed and two-tailed hypotheses

Earlier in this chapter we described a possible study investigating the relationship between number of hours spent studying per week and final examination grade (see section 5.1). We made the prediction (hypothesised) that, as hours of study increased, so would exam grades. Here we have made what we call a *directional hypothesis*. We have specified the exact direction of the relationship between the two variables: we suggested that, as study hours increased, so would exam grades. This is also called a *one-tailed hypothesis*. In this case we were sure of the nature of the relationship and we could thus make a prediction as to the direction of the relationship. However, it is often the case in psychology (and other disciplines) that we are not sure of the exact nature of the relationships we wish to examine. For example, suppose we wanted to investigate the relationship between anxiety and memory for negative information. Previous research in this area has yielded a number of contradictory findings. Mogg, Mathews and Weinman (1987) found that anxious individuals remember fewer negative words than non-anxious individuals, whereas Reidy (2004) found that anxious individuals tend to remember more negative than positive information. Here, then, we are not quite sure of the nature of the

relationship between anxiety and memory for negative words. We therefore would want to predict only that there was a relationship between the two variables without specifying the exact nature of this relationship. In making such a prediction, we are stating that we think there will be a relationship, but are not sure whether as anxiety increases memory for negative words will increase or decrease. Here we have made what we call a *bi-directional prediction*, better known as a *two-tailed hypothesis*.

Definition

A *one-tailed hypothesis* is one where you have specified the direction of the relationship between variables or the difference between two conditions. It is also called a *directional hypothesis*.

A *two-tailed hypothesis* is one where you have predicted that there will be a relationship between variables or a difference between conditions, but you have not predicted the direction of the relationship between the variables or the difference between the conditions. It is also called a *bi-directional hypothesis*.

You might be thinking to yourselves that these are bizarre terms to associate with these forms of hypotheses. Hopefully, all will become clear in the following explanation. To understand why we use the terms one- and two-tailed hypotheses you need to refer back to what we have taught you about distributions.

Previously we explained that a normal distribution and other probability distributions have *tails* at their extremes (see Figure 5.5). The probability of obtaining scores from these extremes (from the tails) is small compared with that of obtaining scores from the middle of the distribution (see Figure 5.6). For example, coming across a man who is 8 ft (244 cm) tall is highly unlikely, and this would thus be in the upper tail of the distribution of men's height.

You now need to think back to what we told you about statistical tests. We explained that we can use probability distributions to help us calculate the probability of a difference or a relationship occurring as a result of sampling error if one does not exist in the population. As an example, we showed you how we can use the standard normal distribution in such cases.

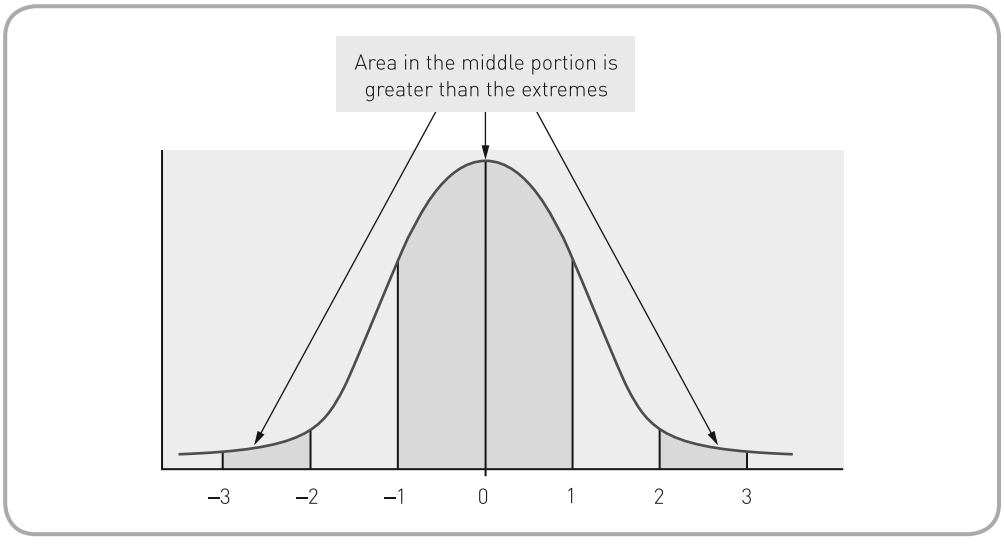

Figure 5.6 Scores in the extremes have lower probability of occurrence than scores in the middle of the distribution

5.12 Assumptions underlying the use of statistical tests

In the preceding sections and chapters of the book, we have introduced the basic concepts underlying statistical testing. In the remainder of the book we will be explaining a wide range of statistical tests suitable for a number of different research designs. However, these tests require that a number of assumptions be met before they can be legitimately applied to sample data.

Definition

Many statistical tests that we use require that our data have certain characteristics. These characteristics are called *assumptions*.

Most of the statistical techniques that we describe in this book make assumptions about the populations from which our data are drawn. Because population characteristics are called parameters (see Chapter 3), these tests are sometimes called *parametric tests*. Because the tests make these assumptions, we have to ensure that our data also meet certain assumptions before we can use such statistical techniques. The assumptions are described in the following sections.

Definition

Many statistical tests are based upon the estimation of certain parameters relating to the underlying populations in which we are interested. These sorts of test are called *parametric tests*. These tests make assumptions that our samples are similar to underlying probability distributions such as the standard normal distribution.

There are statistical techniques that do not make assumptions about the populations from which our data are drawn, but these are not used as frequently as the parametric tests. Because they do not make assumptions about the populations, they are often called *distribution-free tests*. We cover such tests in Chapter 16 of this book.

Definition

Where statistical tests do not make assumptions about the underlying distributions or estimate the particular population parameters, these are called *non-parametric* or *distribution-free* tests.

Assumptions underlying parametric tests

1. The scale upon which we measure the outcome or dependent variable should be at least *interval level*. This assumption means that any dependent variables that we have should be measured on an interval- or ratio-level scale or, if we are interested in relationships between variables, the variables of interest need to measured using either interval- or ratio-level scales of measurement.

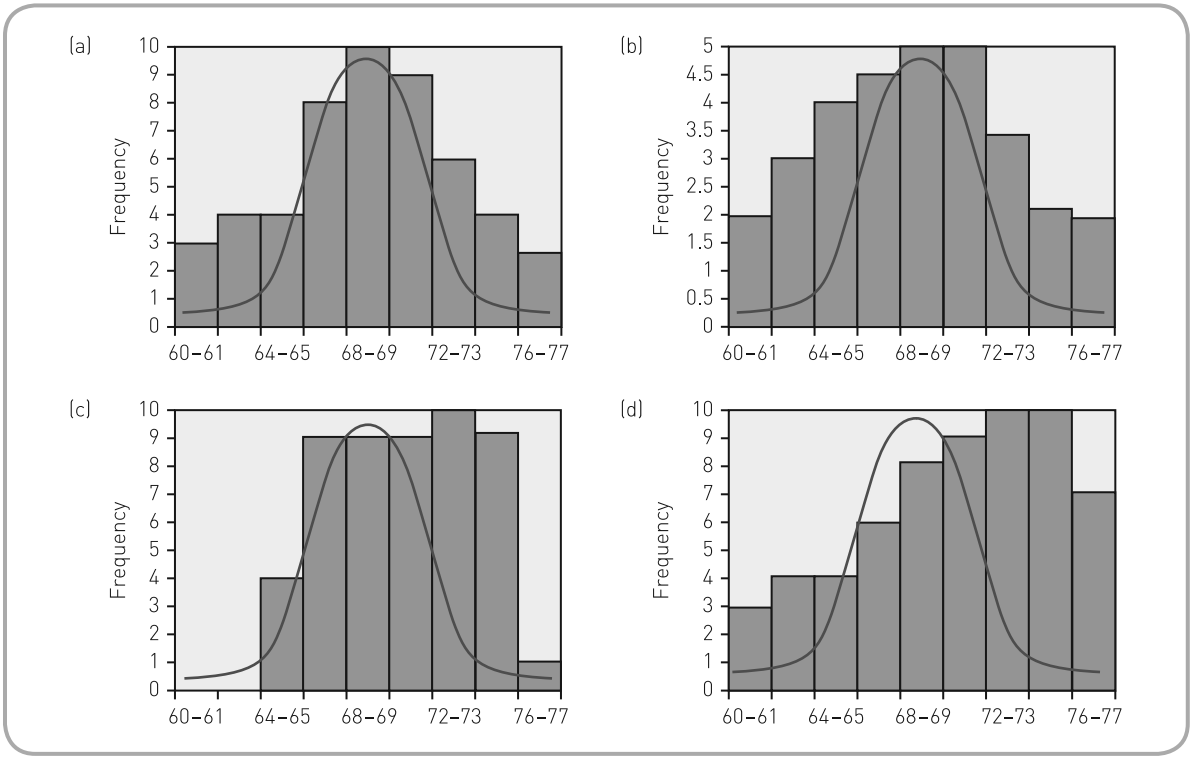

Figure 5.10 Examples of distributions which could be considered approximately normal ((a) and (b)) and those that probably cannot ((c) and (d))

2. The populations from which the samples are drawn should be *normally distributed*. Parametric tests assume that we are dealing with normally distributed data. Essentially this assumption means that we should always check that the data from our samples are roughly normally distributed before deciding to use parametric tests. We have already told you how to do this using box plots, histograms or stem and leaf plots. If you find that you have a large violation of this assumption, there are ways to transform your data legitimately so that you can still make use of parametric tests; as these are beyond the scope of this book, however, you should consult other more advanced texts. Howell (2002) gives a very good overview of such transformations. For your guidance, the distributions in Figure 5.10(a) and (b) are probably close enough to normal for you to use parametric tests. If your distributions are more like those in Figures 5.10(c) and (d), however, you should consider transforming your data.
3. The third assumption is that the variances of the populations should be approximately equal. This is sometimes referred to as the assumption of *homogeneity of variances*. If you remember, when we explained how to calculate the standard deviation in Chapter 3, we told you that you calculate the variance as a step on the way to calculating the standard deviation. More specifically, we informed you that the standard deviation is the square root of the variance. In practice, we cannot check to see if our populations have equal variances and so we have to be satisfied with ensuring that the variances of our samples are approximately equal. You might ask: *what do you mean by approximately equal?* The general rule of thumb for this is that, as long as the largest variance that you are testing is not more than *three times* the smallest, we have roughly equal variances. We realise that this is like saying that a man and a giraffe are roughly the same height, but this does illustrate the reasonable amount of flexibility involved in some of these assumptions. Generally, a violation of this

assumption is not considered to be too catastrophic as long as you have equal numbers of participants in each condition. If you have unequal sample sizes and a violation of the assumption of homogeneity of variance, you should definitely use a distribution-free test (see Chapter 16).

4. The final assumption is that we have no extreme scores. The reason for this assumption is easy to understand when you consider that many parametric tests involve the calculation of the mean as a measure of central tendency. If you think back to Chapter 3, you will recall that we explained that the mean is very sensitive to extreme scores and when these are present it is best to use some other measure of central tendency. If extreme scores distort the mean, it follows that any parametric test that uses the mean will also be distorted. We thus need to ensure that we do not have extreme scores. If you find that you have extreme scores, you should see Chapter 3 for a discussion of what to do about them.

Given that there are these assumptions underlying the use of parametric tests, you might ask: why bother with them? Parametric tests are used very often in psychological research because they are more *powerful* tests. That is, if there is a difference in your populations, or a relationship between two variables, the parametric tests are more likely to find it, provided that the assumptions for their use are met. Parametric tests are more powerful because they use more of the information from your data. Their formulae involve the calculation of means, standard deviations and some measure of error variance (these will be explained in the relevant chapters). Distribution-free or non-parametric tests, however, are based upon the rankings or frequency of occurrence of your data rather than the actual data themselves. Because of their greater power, parametric tests are preferred whenever the assumptions have not been grossly violated.

In this and previous chapters we have explained the important basic concepts for a good understanding of the most frequently used statistical tests. In addition to this we have presented you with a number of descriptive statistical techniques and some advice about when to use them. The preceding paragraphs have also presented advice on the criteria for choosing between various inferential statistical techniques. Before you move on to the nitty-gritty of the various inferential statistics, it is perhaps a good idea to review all such advice and therefore we present it here in summary form. Figure 5.11 gives a rough pictorial guide to the way your design will affect your choice of statistics. It should be stressed that this flowchart represents a general overview of the issues we have covered in the preceding chapters and should be used as such. Whenever you are uncertain as to which tests your data legitimately allow you to use, we recommend that you use the flowchart in conjunction with the advice given previously.

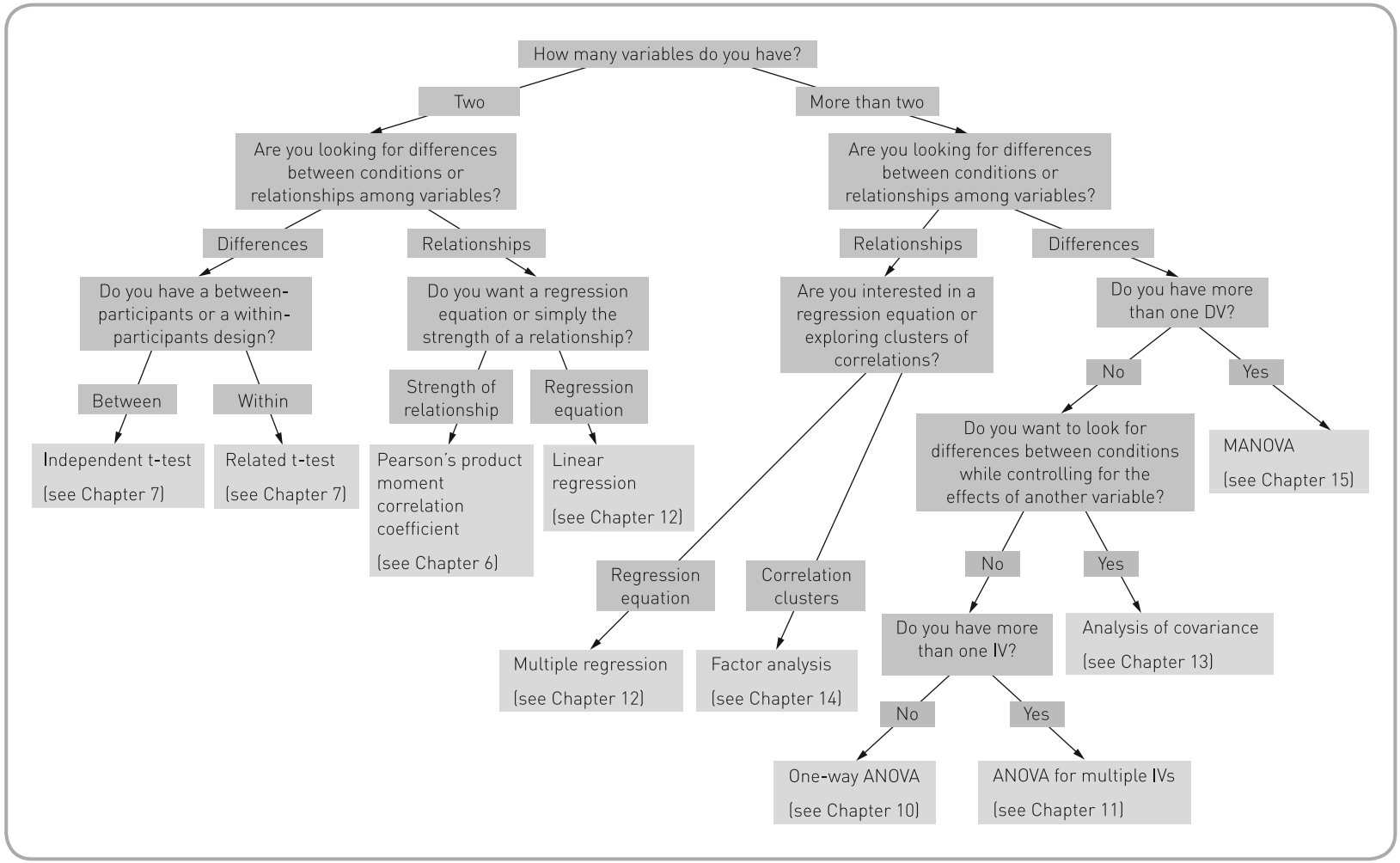

Figure 5.11 Flow diagram as a guide to choosing the most suitable test for the design of a study