CHAPTER FIVE

Making Inferences from Data

To introduce the issue that is at the heart of this chapter, we will use our imaginary experiment again. In this experiment we want to test the hypothesis that people who are in a good mood perform better on intellectual tasks than people who are in a neutral mood. To test this hypothesis, we create two conditions. In one condition a group of participants watch a movie excerpt with humorous content (the experimental condition), and in the other condition a different group of participants watch an excerpt with an emotionally neutral content (the control condition). We expect that participants in the experimental condition will perform better on an intellectual task than those in the control condition, because of a mood change induced by the experimental condition. The level of intellectual performance is measured by presenting participants with 10 logical problems. To decide if our hypothesis is correct we must count the number of logical problems solved by participants in each condition of the experiment.

Now, suppose that participants in the experimental condition generally solve more logical problems than participants in the control condition (e.g., the *mean* number of logical problems solved is higher in the experimental condition). This indicates that our hypothesis 'might' be correct. However, finding that participants in the experimental condition tend to solve more problems than those in the control condition is insufficient to lead us to the conclusion that mood *really* has an effect on intellectual performance. As we saw in Chapter 3, no meaningful conclusions can be drawn from the results of an experiment if we have not previously ensured that our experiment has *validity*. That means three things. First, we must make sure that our IV and DV really measure 'mood' and 'intellectual performance' (construct validity). Second, we must ensure that we are really observing the effects of the IV on the DV, and not those of systematic NVs (internal validity). This is extremely important because if there are systematic NVs affecting the DV, then we cannot claim to have a *true experimental design* (we will discuss this notion at greater length later in this chapter). Third, we must ensure that the effects of the IV on the DV that we observe can, as far as possible, be generalized to other people and situations (external validity). (See Figure 3.6 for a schematic illustration of what these three types of validity are about.)

At this point, suppose we find the differences that we were expecting, *and* we are confident that our experiment has validity. Will this be sufficient to conclude that people who have a good mood perform better than people in a neutral mood? Unfortunately it will not! The fact is that once we have reached this stage we have to do still more things with our data. This is because there are still *random* NVs that can affect our DV (intellectual performance) because they can *never* be completely eliminated. Therefore, we need to use procedures that can ensure that random NVs are not responsible for differences between scores in the experimental and control conditions. Put differently, we need to deal with our data in a way that will allow us to *infer* whether scores in the two conditions are sufficiently different to justify the conclusion that the hypothesis is correct. These procedures concern the domain of **statistical inference**.

Given that statistical inference is essentially about inferring that differences in scores between conditions are *not* due to random NVs, it is important to discuss the nature of NVs and the effects they may have on scores on the DV. In doing so, we will necessarily repeat some of the ideas already expressed in Chapter 3.

Random NVs and their Effects on Scores in the DV

To explain the nature of random NVs, consider again the example used in Chapter 3. Suppose that all the participants in our experiment come from the same university, and that on the day preceding the experiment they attended a party where they had several drinks and stayed until late. On the following day our participants might find it difficult to concentrate on intellectual tasks and, as a consequence, they might generally perform worse on the logical problems than in normal circumstances. Clearly, this implies that the scores on the DV would partly depend on the effects of the level of participants' concentration (that is, an NV).

Note that the effect of this NV would be potentially the same in both the experimental (good mood) and control (normal mood) conditions. This is because, all participants having attended the party, the intellectual performance of both those in the experimental condition and those in the control condition would have the same possibility of being influenced by tiredness. Therefore, it can be said that NVs are 'a nuisance' because they introduce **variability** into the data, which makes it harder to see the effects of an IV on scores on the DV.

Let us explain this notion more carefully. Imagine that the IV had no effect on the DV, and that there were no random NVs affecting our DV. Then all scores in both conditions would be the same (see Figure 5.1a for an example). If there were still no NVs, but the IV did have an effect, we would have two possible scores, one for each condition (see Figure 5.1b). In this case it is rather obvious that the IV has affected scores in the two conditions differently. If there were no random NVs, our data would always be clear like that, and there would be no need for inferential statistics. Unfortunately, that is cloud-cuckoo land. There are always potential random NVs and they make it harder to tell whether our IV has had an effect. In Figure 5.1c

(a)	**expt. condition**	**control condition**
	2	2
	2	2
	2	2
	2	2
	mean = 2	mean = 2

(b)	**expt. condition**	**control condition**
	4	2
	4	2
	4	2
	4	2
	mean = 4	mean = 2

(c)	**expt. condition**	**control condition**
	2	2
	5	1
	5	0
	4	5
	mean = 4	mean = 2

Figure 5.1 Effects of random NVs on scores: (a) no effect of IV and no random NVs; (b) effect of IV but no random NVs; (c) effect of IV and effect of random NVs

you can see an example of what random NVs can do to the scores in the two conditions of the experiment. Here, due to the effect of the IV, the difference between means for the two conditions is the same as in Figure 5.1b but, because there are also random NVs operating, this difference might have been caused by (1) the IV, (2) random NVs just happening to pile up in favour of the experimental condition on this particular occasion, or (3) some combination of the two. How can we decide whether we should be persuaded that the IV had an effect? The clue is in how much variability there is between scores *within* a condition. The variability within a condition cannot have been caused by the IV, because every participant within a condition received the same treatment (i.e., the same level of the IV). So, the variability within conditions must have been due to the chance effects of random NVs. Therefore, the more differences there are within conditions compared to the mean difference between conditions, the more likely it is that random NVs that caused the differences within each condition could also have caused the difference between conditions.

What we really need to know is, given the amount of variability among scores within each condition, just how likely it is that the obtained difference between means (say, in Figure 5.1c) might have been entirely due to the effects of random NVs. This is where a statistical test will help. It will tell us either that it is unlikely that random effects could account for the data we obtained, in which case we will infer that our IV probably did contribute to the difference, or that it is not that unlikely, in which case we will conclude that the difference between means might well have been due to the cumulative effects of random NVs just happening (by chance) to pile up in favour of the experimental condition on this particular occasion. In the latter case, we could not claim that our experiment had shown there to be an effect of the IV (look back at Figure 3.5 for a summary of the effects of systematic and random NVs).

But there is another very important point to consider: Whichever way the evidence points, it is never conclusive. For example, we might conclude that the evidence is not strong enough to persuade us that the IV had an effect, but that does not mean that the IV definitely had no effect. It might just not have been big enough to show up clearly against all of the variability produced by the random effects. At this point, we need to be clear that we do not reach a cut-and-dried conclusion that the IV definitely did or did not have an effect. Our conclusion is necessarily **probabilistic**. We conclude that random effects *probably were not* sufficient to have caused the difference between means (supporting an effect caused by the IV) or, alternatively, that random effects *probably were* sufficient to have caused the difference (not supporting an effect caused by the IV).

Also, remember that we cannot just assume that potential NVs will be random in their effects. We saw in Chapter 3 that NVs can have systematic effects; that means that they can affect only one condition of the experiment, thereby providing plausible explanations of a difference between scores in the two conditions of the experiment, which compete with the explanation that it is the IV that caused the difference. We also saw that potential systematic NVs should be controlled by holding them constant, effectively eliminating them as variables, and, when this is not possible, they should be controlled by *turning* them into random NVs. Once systematic NVs have been *made* random, then they can be dealt with using inferential statistics. (Figure 5.2 recapitulates the main points in this argument that were made in detail in Chapter 3.)

So, how do we ensure that potential systematic NVs are made random? We must use a procedure known as **random allocation**. However, at this point it is necessary to make it clear that this procedure is logically distinct from another important procedure, which is known as **random sampling**. Now, since random allocation is often confused with random sampling, and since, as we said, random sampling is an important issue, let us clarify what it is about, before we discuss random allocation.

Random sampling

In order to be able to say that the results of an experiment apply (can be generalized) to the population from which the sample of people who participate in the experiment is drawn, it is necessary that the participants are **representative** of the population. (This is the issue of external (population) validity introduced in Chapter 2.) How can we ensure that this is the case? In principle, a representative sample of participants can be obtained by random sampling from a defined population. For example, we might want to draw conclusions about all first year psychology students in our university. This would entail putting the names of all of those students in a metaphorical

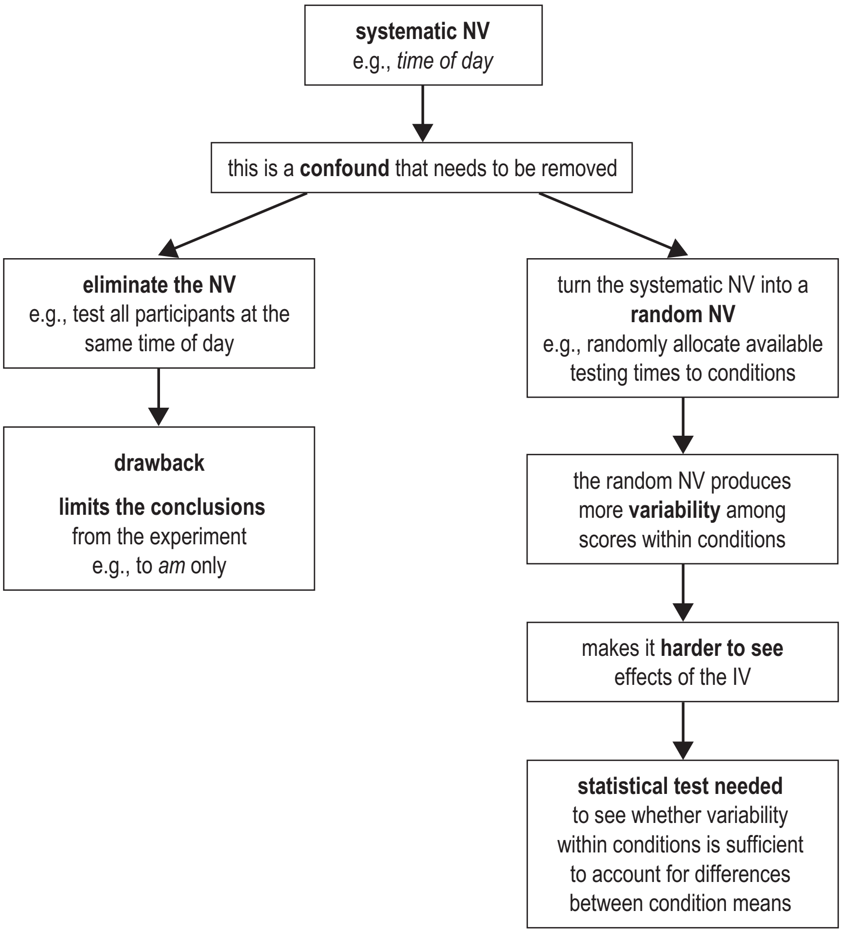
systematic NV
e.g., time of day
this is a confound that needs to be removed
eliminate the NV
e.g., test all participants at the same time of day
drawback
limits the conclusions from the experiment
e.g., to am only
turn the systematic NV into a random NV
e.g., randomly allocate available testing times to conditions
the random NV produces more variability among scores within conditions
makes it harder to see effects of the IV
statistical test needed
to see whether variability within conditions is sufficient to account for differences between condition means

Figure 5.2 Dealing with systematic (confounding) nuisance variables (NVs)

hat and pulling out the required number (the sample size) while blindfolded. A more realistic alternative would be to assign a unique number to each student and then use a random number table to select the required number of students. The point is that every student must have an equal chance of being selected. The reality is that random sampling from a defined population presents some difficulties (see Additional information (5.1)), with the result that it is more an ideal than a procedure that is commonly adhered to.

Having clarified what random sampling is about, let us return to the issue of random allocation, which, as we specified above, is a procedure used to ensure that potential systematic NVs are made random, thereby allowing us to infer whether the IV can account for the difference between DV means in the two conditions.

Additional information (5.1) – The unreality of random sampling in experiments

Random sampling procedures tend to be quite time consuming and are often difficult to implement. For these reasons, they are rarely carried out in practice. If a sample of first year psychology students were required, it is much more likely that an **opportunity sample** (e.g., volunteers from a class) would be used. The difficulty of obtaining a random sample from a population of all first year psychology students in UK universities would obviously be even more problematic, and imagine the difficulty of trying to obtain a random sample of all UK undergraduates. The broader the population of interest, the less likely it is that a random sample will be obtainable. In fact, the extent to which generalization to other people is possible is much more dependent on **replication** of the results (i.e., showing that we get similar results) with different participants, than on the existence of a random sample from a population. We would also like to be able to generalize our conclusions to other specific *situations* (e.g., viewing alone, with familiar others, with unfamiliar others, in a relaxed environment, in a formal environment etc.). In this case, the argument about generalization being based on replication applies even more strongly, because there is usually no attempt to randomly sample from a population of possible situations of interest.

Random allocation

The random allocation of experimental units (participants and test occasions) to conditions is the hallmark of a ‘true’ experiment. A true experiment is one in which the only thing that differs systematically between conditions is the experimenter’s manipulation of the level of the IV (e.g., mood-raising video versus control video). Other NVs, such as individual characteristics of participants and particular environmental conditions, will also affect the scores obtained on the DV, but we know that so long as these do not affect one condition in a systematically different way than they affect the other condition, they will not be a threat to the internal validity of the experiment. Provided that any potential NVs are *random* in their effects, that is, they have an equal chance of affecting scores in favour of either condition of the experiment, any systematic effect can only have been the effect of the IV.

Random allocation of participants (and test occasions) to conditions is a way of ensuring that the particular characteristics of participants (e.g., motivation, suggestibility, alertness etc.) have an equal chance of affecting the mean score in either condition. It is not the case that random allocation guarantees that the participant characteristics in the two conditions will be equal. In fact, that is extremely unlikely. More of the highly alert participants will probably be allocated to one or other condition, but the point is that it could be either condition that benefits.

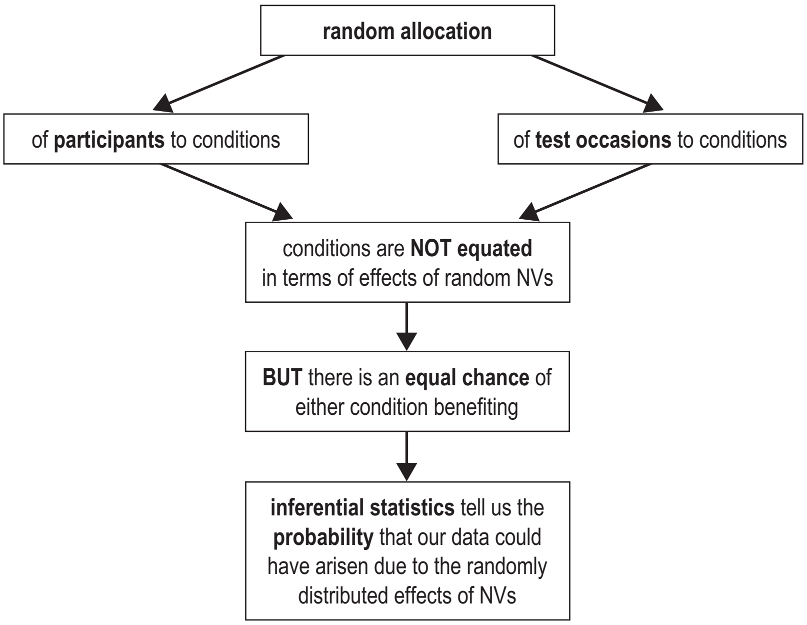
random allocation
of participants to conditions
of test occasions to conditions
conditions are NOT equated in terms of effects of random NVs
BUT there is an equal chance of either condition benefiting
inferential statistics tell us the probability that our data could have arisen due to the randomly distributed effects of NVs

Figure 5.3 Using random allocation

The same argument applies to the allocation of available testing times to conditions (see Figure 5.3). Unless participants will all be tested on the same occasion, each available testing occasion should be allocated at random (e.g., by tossing a coin) to one of the two conditions. The need for this aspect of random allocation is frequently overlooked. Thus, experiments are often, incorrectly, carried out with participants in the two conditions being tested in separate groups. That leaves open the possibility that any systematic effect that is inferred from the data could have been due to particular characteristics of the two test situations.

Additional information (5.2) – More on random allocation in experiments

Suppose we have a sample of 20 people who are going to take part in our experiment. We can expect that they will differ from one another in all sorts of ways that might affect their scores on our logical reasoning test (i.e., there will be NVs). Now suppose that there is really no effect of our IV (*type of video*). Then, if there were no NVs, we would expect everyone in the sample to get the same score. But there *are* NVs, so that the participants get different scores from one another, even though there is no effect of the IV. Now suppose that the scores obtained by the sample of 20 people were:

Person	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Score	23	26	17	24	31	19	33	25	27	21	21	18	17	12	30	22	16	23	20	17

Then, suppose that the first 10 people had been randomly allocated to the experimental (good mood) video condition and the remainder to the control (normal mood) condition. The means for the two conditions would have been:

Experimental mean	= 24.6
Control mean	= 19.6

We can see that the mean score in the experimental condition would have been higher, but remember that the differences were due solely to NVs and not at all to the different videos. Furthermore, if the people with the first 10 scores had been put in the control condition, the mean score in that condition would have been higher by the same amount. So, it is clear that, if there is no effect of the IV, random NVs are equally likely to result in a higher or lower mean for the experimental condition. That is not all – depending on which people were randomly allocated to which condition, the difference between means in the two conditions (in whichever direction) would vary. With most allocations, individual differences would tend to roughly balance out across conditions, so that the difference between means would be relatively small, but, with some allocations, the difference would be relatively large. For example, if it just happened that the ten people with the highest scores had all been in the experimental condition, the means for the two groups would have been:

Experimental mean	= 26.4
Control mean	= 17.8

This looks like quite a strong advantage for the experimental condition, but we know that in this case it was just due to NVs (e.g., individual differences among the sample) and, as usual, if the allocation had been reversed, the apparent advantage would have been for the control condition.

Random allocation is the best way to ensure that there is an equal chance of the advantage going to either condition. Whatever the outcome, we know that the effects of the NVs will be randomly distributed between conditions, so that if there is really no effect of the IV, any apparent effect we see will be due to chance. Our statistical analysis uses the fact that most differences between means caused by random NVs are relatively small to tell us how confident we can be that an obtained difference between means is large enough to make it unlikely to have been caused by random NVs (chance), and was therefore probably due to the systematic effect of the IV (provided there are no confounding variables) (see Figure 5.3).

The Process of Statistical Inference

At this point, we are ready to discuss the details of the process of statistical inference that allows us to decide whether, on the basis of the data we have collected, a hypothesis is *probably* correct or not. So, we will now discuss what you should do in order to decide whether differences between conditions are due to chance (i.e., the effects of random NVs), or to the effects of the IV, as predicted. Let us start by introducing some technical terms that will allow us to provide a formal explanation of this process.

Experimental hypothesis and null hypothesis

In Chapter 2, we discussed that what is meant by a hypothesis, in the context of an experiment, is a prediction about an effect of an IV on a DV. We will call this the **experimental hypothesis**. We can also talk about a contrary hypothesis, one that predicts that the IV will *not* have an effect on the DV. This is referred to as the **null hypothesis**.

Statistical significance

Another important concept is that of **statistical significance**. We talk about a difference between means being statistically significant when there is a low probability that it could have arisen as the result of *random error*, that is, the chance effects of random NVs. But what do we mean by a low probability? Total certainty that random error was responsible would be represented by a probability of '1' and total certainty that random error was not responsible would be represented by '0'. By convention, we take 'low probability' to be a 1 in 20 chance (that is 5 in 100, which is a probability of .05) or, if we are feeling more conservative, a 1 in 100 chance (which is a probability of .01) or, if we are feeling really conservative, a 1 in 1,000 chance (that is, 0.1 in 100, which is a probability of .001). These levels of confidence are described as **alpha (α) levels** and the α level you are willing to accept as evidence of an effect is supposed to be set before data are collected. Then, if the probability level obtained when a statistic is calculated (more on this later) is below the designated α level, we can conclude that the null hypothesis can be rejected and the effect of our IV is said to be *statistically significant*. (Note that researchers prefer to say that 'the null hypothesis can be rejected', rather than say that 'the experimental hypothesis can be accepted'; see Additional information (5.5) for an explanation of why this is the case). Thus, if α has been set at .05 and the obtained probability (p) when a statistic is calculated is .04, we can claim that the effect of our IV was statistically significant but, if p is .06, we have to conclude that the effect was not statistically significant (the effect is then usually described as being 'non-significant').

Complications (5.1) – Reporting the lowest *p*-value possible

Not everyone agrees that a pre-defined significance level (i.e., α level) should be set, and the null hypothesis rejected if the probability of obtaining the data when the null hypothesis is true is less than the pre-defined α level. There is a view that focusing narrowly on whether the probability is below (reject null hypothesis) or above (fail to reject null hypothesis) the critical α value is too crude. For example, with α set at .05, a probability of $p = .049$ would be reported as significant (i.e., $p < .05$), whereas a probability of $p = .051$ would be reported as non-significant (i.e., $p > .05$). As the consequences of finding an effect to be statistically significant or non-significant can be considerable – not least in determining whether a study is published in a journal – we might question the logic of this all-or-none decision. Another example may help you to see the problem. Suppose, once again, that α is set at .05. Then, probabilities of $p = .051$ and, say, $p = .87$ would both be reported as non-significant, with no distinction made between them. Similarly, probabilities of $p = .049$ and $p = .0001$ would both be reported as significant, again with no distinction made between them. An alternative is to focus on the actual value of the probability. In this view, $p = .0001$ would be reported as 'highly significant' or perhaps as 'significant ($p < .001$)', and $p = .051$ might be described as 'approaching significance'. Against this argument, some researchers regard it as 'suspect' to decide what level of significance to report after seeing the result of the analysis. Nonetheless, it is common for researchers to report the lowest conventional level (.05, .01, .001 etc.) of probability that their analysis permits. The justification claimed for this is that the probabilities are best treated as 'indicative' of levels of confidence rather than as rigid decisions. There does seem to be a gap between the classical (predetermined α level) approach expounded in most statistic texts and what many researchers actually do.

Imaginary distributions

Now, we come at last to an explanation of how the statistical decision is reached. First, you need to make an imaginative leap. We have only done the experiment once, of course, and we got a particular set of data, with a particular mean for each of the two conditions. If we could wipe out all memory of the experiment and do it again, we would almost certainly get a different set of data and different values for each of the condition means. Now, the imaginative leap. Imagine that we were able to repeat the experiment thousands of times in this way. Each time we would get

Additional information (5.3) – An intuitive rationale for the conventional alpha levels

What is the rationale for adopting $p < .05$, $p < .01$ etc. as the critical values that we use to decide whether to reject the null hypothesis? The answer is that they are really convenient, and ultimately, arbitrary, conventions. They do, however, map reasonably to our intuitive notions about chance events. Let's do a mind experiment about when you decide that some outcome is more likely to have resulted from some systematic effect than from the operation of chance (random effects). Imagine that I show you 10 coins and bet you that I can toss them so that more of them come down 'heads' than 'tails'. You take on the bet, reasoning that you have an even chance of winning, and we agree that I will give you 10p for every coin that comes down tails and you will give me 10p for every coin that comes down heads. I toss the first coin, it comes down heads and you hand over 10p. The same happens with the next coin, and the next, and so on. After how many heads in a row would you become *suspicious* that this was not a game of chance? After how many would you become *convinced* that something systematic was causing the run of heads – that I really did have the knack of tossing coins so that they came down heads or, more likely, that I had a set of weighted coins? When we have asked these questions to classes of students, there has always been a majority that become suspicious after five heads in a row and convinced after seven in a row. The probabilities of two, three, four etc. up to 10 heads in a row are shown in Figure 5.4. There you can see how students' intuitions map on to the conventional values of $\alpha = .05$ and $\alpha = .01$. The probability of a run of heads drops below .05 for the first time when there have been five in a row and below .01 for the first time when there have been seven in a row. Of course, if the stakes were higher you might be inclined to challenge me sooner or, if there was a penalty for an incorrect challenge, you might wait for the probability to drop lower. These are analogous to the deliberations that lead a researcher to choose a higher or lower α level for statistical significance.

No. heads in a row	Probability	Conventional significance levels
1st head	$p = .5$	
2nd head	$p = .5 \times .5 = .25$	
3rd head	$p = .25 \times .5 = .125$	
4th head	$p = .125 \times .5 = .063$	
5th head	$p = .063 \times .5 = .031$	← $\boldsymbol{p < .05}$
6th head	$p = .031 \times .5 = .016$	
7th head	$p = .016 \times .5 = .008$	← $\boldsymbol{p < .01}$
8th head	$p = .008 \times .5 = .004$	

Figure 5.4 An intuitive rationale for the conventional levels of statistical significance

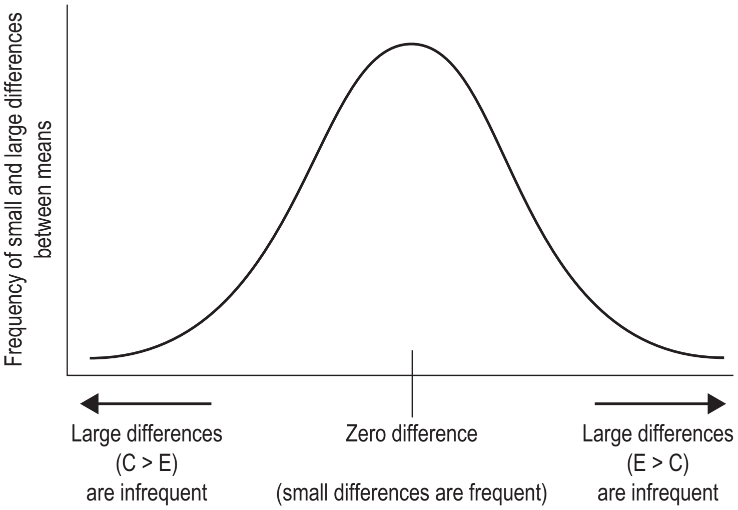

Figure 5.5 Hypothetical distribution of differences between means when the null hypothesis is true

different specific values for the experimental and control means. For either condition, we could imagine plotting the frequency with which different mean scores were obtained. These would be frequency distributions like those discussed in Chapter 4 and illustrated in Figure 4.3, but they would of course be imaginary (or hypothetical) distributions, because we did not really repeat the experiment thousands of times.

Now, imagine what the means for the two conditions would be like if the null hypothesis were true, that is, the difference between the means was entirely due to the chance effects of random NVs. Most often, the means would be very similar but sometimes the chance effects would happen to pile up in favour of one or other mean and, occasionally, the chance effects would pile up to create a really big difference between the means. Just as we could plot imaginary distributions for each mean, we could also plot the frequency of various sizes of *difference* between the two means – that is, a **hypothetical distribution of the differences between means**. If we did that, and the null hypothesis were true, we would be likely to see something like the distribution of means shown in Figure 5.5. This shows the frequency with which various values for the difference between means might be expected just on the basis of chance effects of random NVs; that is, it is based, not on real data, but on our understanding of chance effects, as they occur, for example, in coin-tossing experiments. The most frequent differences would be very close to zero, and the frequencies would decrease for progressively larger differences in either direction.

A COIN-TOSSING ANALOGY

Let's pursue the analogy of a coin-tossing experiment, to stand in for an experiment in which there happen to be *only* random (chance) effects operating. Suppose you and I each have 10 coins, which we can assume are 'unbiased', that is, they are equally likely to come down heads or tails when flipped. If we each flipped our 10 coins thousands of times and each time recorded the number of heads, the most frequently obtained difference between number of heads and tails would be zero, with small differences in either direction also being relatively frequent and large differences being infrequent. In fact, if the frequencies were plotted we would have an approximately *normal distribution* very like the one shown in Figure 5.5 (and earlier, in Figure 4.6). Suppose now that I provide each of us with a new set of 10 coins and bet you that I can use my telekinetic powers to make all of your coins come down one way and all of mine the other way. You take on the bet and we flip the coins and, lo and behold, all mine come down heads and all yours come down tails. When you get over the surprise, you will probably conclude that you just witnessed a systematic effect rather than a random (chance) effect, because you (rightly) believe that the chances of a difference of 10 heads between us, in the absence of some systematic biasing effect, would not be zero but would be extremely low. Of course, you will probably soon begin to entertain the unworthy thought that the systematic effect may have been biased coins rather than my telekinetic powers!

THE IMAGINARY DISTRIBUTION OF A NEW STATISTIC

Now, we have already explained that if the probability of getting a difference between means as great as that we obtained, just by chance (given the amount of variability among scores within each condition), is lower than the α value specified (e.g., .05), we should conclude that the difference is statistically significant (i.e., we should reject the null hypothesis at the .05 level of probability and conclude that the experimental hypothesis is supported). So, if the obtained difference between means is among the 5% largest possible differences in the distribution in Figure 5.5 (i.e., 2.5% largest in either direction), we conclude that the difference in means is statistically significant at the 5% level. This is actually a slight over-simplification. The 'difference between means' is a **statistic** – a value calculated from a sample of data, just as a mean of a sample of data is a statistic – but we use a slightly more complex statistic in practice, because, in this case, for example, we need to take account of the *variability among scores within conditions* as well as the difference between means for the two conditions. The reason why we take account of the variability of scores within conditions is that the bigger the effects of random NVs, the greater the variability they create among scores within each condition and the more plausible it becomes that the random NVs *alone* could account for the difference between means (i.e., without there being any effect of the IV).

There are a number of different statistics available. Which one it is appropriate to use depends on details of the experimental design and the type of data we are collecting. These statistics will be introduced in later chapters, but the point to hang on to now is that they are all used in the same way to summarize the data (just like the difference between means) in order to see whether the value of the statistic is extreme enough to make it unlikely (e.g., probability < .05) that it would have arisen just by chance when the null hypothesis is true. Just how each different statistic tells us what the probability is of chance having produced the difference between conditions will be dealt with in the following chapters.

Statistical inference – a concrete example

In order to make the argument about statistical inference more concrete, we will repeat it using the example of the experiment from Chapter 2 – that was the example about the effect of viewing a *mood-enhancing video* (compared with viewing a *neutral video*) on performance on a test of *logical reasoning*. Participants are randomly allocated to one of the video conditions and all participants are shown the relevant video and then tested on the set of logical problems at the same time in the same laboratory. This means that situational variables have been largely eliminated (not entirely, of course – one participant may have an uncomfortable chair or be sitting in a draught, for example). Individual difference variables, on the other hand, will clearly have an effect (some participants will simply be better than others at solving logical problems irrespective of their moods, some will be more motivated to do well, and so on).

Individual difference variables (and any remaining situational differences) will, however, function as random NVs. This is because participants were randomly assigned to conditions and (although this was not explicitly stated) participants should have been randomly allocated seating positions in the laboratory. When participants' scores have been recorded, the means for the two groups are obtained and a statistic is calculated from the data. An appropriate statistic in this case would be the independent groups (or unrelated) t-statistic. This statistic will be explained in Chapter 9. For the moment, all we need to know is that the value of the statistic gets bigger as the difference between the means increases and the variability among scores within each condition (due to the random NVs) decreases. The distribution of values of the statistic when the null hypothesis is true can be specified for each possible sample size (number of participants). The distributions for a small, medium and large number of participants (say N = 5, 30 and 100, per group) are shown in Figure 5.6. When, in the calculation of t, one mean is subtracted from the other to obtain the difference between means, the value will be positive or negative depending on the direction of the difference (which mean was larger).

You can see in Figure 5.6 that both large negative and large positive values of t will be rare when the null hypothesis is true (i.e., the tails of the distribution). If the value for t that we obtain in our experiment falls in one of the tail areas, we can conclude that the mean difference between problem scores in the two video conditions was statistically significant, that is, the null hypothesis (that the mean difference was due to chance effects of random NVs) can be rejected with a known maximum probability (the value at which α was set) of being mistaken. If the value of α (the level

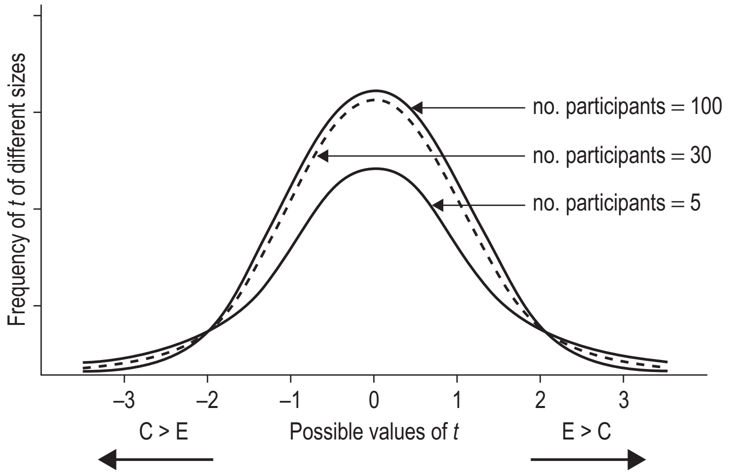

Figure 5.6 Hypothetical distribution of *t* for a small, medium and large number of participants when the null hypothesis is true

of statistical significance sought) was set at .05, and the obtained value of *t* falls within the .05 (5%) most extreme values in the tails (.025 at the end of each tail), the difference between means will be statistically significant at the 5% level. If the obtained value falls closer to the centre of the distribution than that (i.e., outside of the **rejection regions** in the tails), we will have to conclude that the null hypothesis cannot be rejected at the 5% level; that the difference between means is non-significant at that level of confidence. Figure 5.7 illustrates how statistical inferences about the null hypothesis and, indirectly, the experimental hypothesis are arrived at. The left-hand side shows a value of *t* that falls in one of the 'tails' of the hypothetical distribution and is therefore statistically significant. The right-hand side shows a value of *t* that falls outside of the tails (i.e., closer to the mean of the distribution) and is therefore non-significant. Tables giving the minimum size of *t* that will fall in the 5%, 1% or .1% most

Complications (5.2) – The truth about the null hypothesis

It is quite common for students (and researchers, for that matter) to refer to the probability of the null hypothesis being true. This is a misconception. The null hypothesis is either true or it is false. It refers to a 'state of the world'. There are no 'probabilities' associated with the truth or falsity of the null hypothesis. The probability that statements about statistical significance refer to is the probability that the data we obtained might have arisen just by chance when the null hypothesis is true.

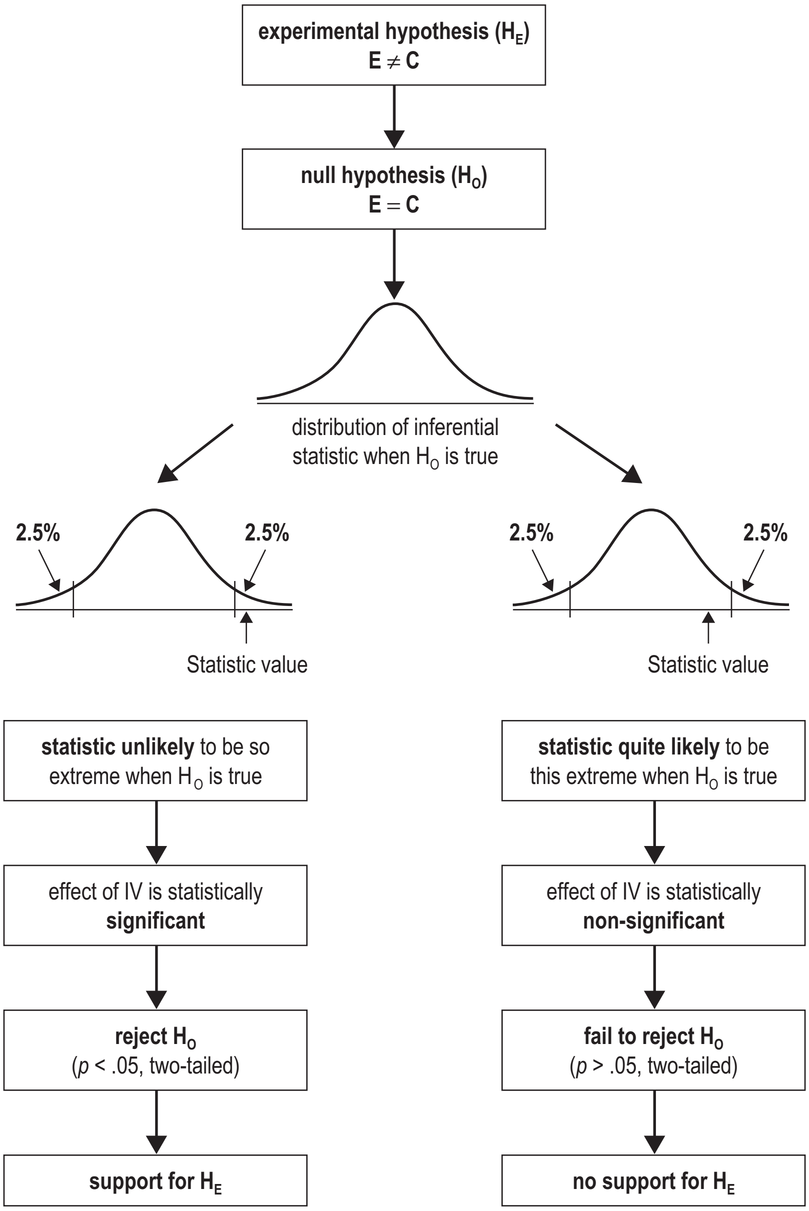
experimental hypothesis (H_E)
E ≠ C
null hypothesis (H_O)
E = C
distribution of inferential statistic when H_O is true
2.5%
2.5%
Statistic value
2.5%
2.5%
Statistic value
statistic unlikely to be so extreme when H_O is true
effect of IV is statistically **significant**
reject H_O
(p < .05, two-tailed)
support for H_E
statistic quite likely to be this extreme when H_O is true
effect of IV is statistically **non-significant**
fail to reject H_O
(p > .05, two-tailed)
no support for H_E

Figure 5.7 Experimental hypothesis and null hypothesis

extreme areas of the tails are available for different numbers of participants (i.e., the minimum *t* values needed for statistical significance at various α levels). Such a table is provided in Appendix 1 and also in Chapter 9, where *t*-tests will be considered in detail. Discussion of how to use the table will be held over to that chapter.

Additional information (5.4) – Statistical inferences from samples to populations

A formal treatment of the inference process we have been describing asks the question: How likely is it that our two samples of scores were drawn from populations with the same mean? In order to understand how an answer to this question is sought, we need to be clear that a population refers to all of the possible objects of a particular kind. It does not necessarily refer to people (e.g., the population of first year university students in the UK) or even to tangible entities of any kind (e.g., stars in the Milky Way, traffic lights in London). It can refer to an entirely imaginary set of all possible scores that could have been obtained by an infinite number of participants in an experiment. When we carried out our experiment, however, we only had a small number of participants and the scores they obtained are regarded as a **random sample of the population of scores** that we could have obtained if we had tested an infinite number of participants with the same characteristics as our sample. We can acknowledge that our sample of participants is unlikely to be a random sample from the population of possible participants that we are interested in, but this does not affect our (reasonable) assumption that the obtained scores are a random sample of the **imaginary distribution** of all possible scores (i.e., the imaginary population of scores). In fact, we have two samples of scores in our experiment, one sample for each condition. If the null hypothesis is true, the means of the populations of scores from which these samples are drawn will be equal; there will be, effectively, a single population of scores with a single mean. Still assuming that the null hypothesis is true, we have two random samples from the same population and the means of the samples will differ by chance alone. The means of some pairs of random samples (sets of scores obtained in particular experiments) would happen to differ a lot, so that they would *look like* samples drawn from two populations with different means. A statistical test ascertains the probability of getting, just by chance, two samples of scores that differ as much as those we obtained in our experiment. If the probability is below the value we set for α, we will conclude that it is unlikely that the two samples came from the same imaginary population, and that it is more likely that the null hypothesis is false and the samples came from two different populations of scores (one for each condition) with different means; that is, there was a statistically significant effect of our IV on our DV.

Additional information (5.5) – Why test the null hypothesis instead of the experimental hypothesis?

It does seem tortuous to set up a null hypothesis and subject it to a test to see if we can reject it rather than setting up the research hypothesis and testing it directly. The reason for testing the null hypothesis is that it is a **specific hypothesis** (the difference between the means of two populations of scores is zero). That allows us to construct an imaginary sampling distribution, showing the probabilities of obtaining a statistic of various sizes when that hypothesis is true. The research hypothesis, on the other hand, is a **non-specific hypothesis**. If you wanted to test the hypothesis that the difference between means is 1 on our DV scale, you could set up a sampling distribution and test that hypothesis, just as you could for any other specific difference between means. The problem is that you do not usually have a particular value in mind. To test all of the possible differences between means that you might wish to consider, you would need to set up a sampling distribution for each one. So, we end up setting up the specific null hypothesis, seeing if we can *reject* it at some level of probability, thereby allowing us to infer that the means do differ.

Statistical decision errors

When we make a statistical decision, we recognize that it might be mistaken. After all, the statistical inferences are all statements of probabilities rather than certainties. There are two ways we can be mistaken. First, we might reject the null hypothesis when it is in fact true, that is, there was no systematic effect – the difference between means in the two conditions was entirely attributable to random NVs. Because the difference between means created by the combined effects of the random NVs happened to be large – a difference that would happen, say, less than 5% of the time by chance – we were led to conclude that the difference was probably caused by a systematic effect (of the IV, we hope), whereas, assuming we set α equal to .05, it was in fact one of those 5% of occasions when a large difference was due to chance. This kind of mistake is called a **Type I error** and the probability of making it is known (i.e., the value at which we set α).

The other mistake we can make is to fail to reject the null hypothesis when it is in fact false, that is, there was a systematic effect but we failed to detect it. This may happen when the variability of scores within conditions is large relative to the difference between means, so that we are misled into concluding that random error was probably great enough to account for the difference between means. This kind of mistake is called a **Type II error** and the probability of it occurring is denoted by the symbol β. As with α, we can, in principle, set the value of β at a level that suits us. A level that, by convention, is often thought acceptable is .2 (20%). That is, we accept a 20% risk of failing to find a significant effect of the IV when it does in fact have an effect. The fact that most researchers are prepared to accept a considerably bigger risk of missing a real effect (20%) than of finding an effect that is really just due to chance (5%) reflects a general belief that the theoretical (and maybe, practical) consequences of concluding that an IV has an effect when it does not are more serious than the consequences of missing an effect.

Just as there are two ways of being mistaken, there are two ways of being right. You can correctly reject the null hypothesis – the probability of this outcome is $1 - \beta$ (that would be $1 - .2 = .8$, or 80%, using our example above). This value $(1 - \beta)$ is known as the **power** of the statistical test, that is, the likelihood of the test finding a significant effect when one does in fact exist. In practice, it is usually this *power* probability that is decided on by the researcher, and this automatically determines the value of β, the probability of a Type II error. The other way of being right is to correctly fail to reject the null hypothesis. The probability of this outcome is $1 - \alpha$

Complications (5.3) – What to conclude if you fail to reject the null hypothesis

If the statistical decision is to reject the null hypothesis, the inference is clear. The value of the statistic that was calculated (based, for example, on the difference between means and the variability of scores within conditions) is sufficiently extreme to persuade us that it is unlikely to have occurred by chance (random NVs) alone. We therefore conclude that the data probably arose at least partly as a result of an effect of the IV. In other words, we have found support for our experimental hypothesis.

If the statistical decision is to fail to reject the null hypothesis, the situation is less clear. Does that mean that the null hypothesis should be assumed to be true? The answer is 'no'. The null hypothesis might still be false, but the effect of the IV might be small relative to the effects of NVs and, therefore, hard to discern. The null hypothesis states that there will be zero difference between population means of the two conditions. With a small difference between population means, we would be unlikely to identify a significant difference from our sample data, unless we had extremely good control over random NVs and/or a very large sample of scores, in which case we might well be finding an effect that is too small to be of interest. This is not unlikely, since the null hypothesis is almost never *exactly* true.

Although people sometimes talk about accepting the null hypothesis when it cannot be rejected, it is probably safer to refer to '*failing to reject the null hypothesis*' or '*retaining the null hypothesis*' (i.e., provisionally).

Additional information (5.6) – The power of a test

The power of a test to find an effect when one actually exists depends on a number of factors

- Our ability to control random NVs. The more they are controlled (kept constant), the less variability there will be within each condition and the easier it will be to attribute a difference between means to the IV.
- The size of effect that we do not want to risk missing. The larger the effect, the easier it is to reach statistical significance. To some extent, a researcher can maximize an effect by selecting values for the IV that are relatively extreme. For example, it would be much easier to find an effect of age on time to run 50 metres if we compared 6-year-old and 10-year-old children than if we compared 6-year-olds with 6.1-year-olds! Similarly, we are more likely to find an effect of viewing different videos if the experimental one is really hilarious rather than mildly amusing.
- The α level set. Everything else being equal, it is easier to reach statistical significance with an α value of .05 than a value of .01.
- Whether a one- or two-tailed test is used (a distinction that we will explain in the next section). To anticipate, if you opt for a one-tailed test and your directional prediction is correct, a lower value of the statistic calculated will be needed for statistical significance at a given level of probability.
- Whether a parametric or non-parametric statistical test is used. This distinction will be discussed in subsequent chapters.
- The number of participants included in the experiment. The more participants per condition, the more powerful the test will be. It is beyond the scope of this book but, if you continue to study psychology, you will learn about how to get an estimate of how many participants you will need to achieve a given power.

(that would be $1 - .05 = .95$, or 95%, using our example above). This is the likelihood of the test failing to find a significant effect when one does not in fact exist. The relationship between the decision that is made when a statistical test has been carried out (reject or fail to reject the null hypothesis) and the reality of the situation (the null hypothesis is true or it is false) is illustrated in Figure 5.8.

One- and two-tailed tests

Usually a researcher has a view about the likely direction of a difference between means that will occur. It is likely, for example, that an experimenter conducting

Decision made using inferential statistic	The reality (H_0 is either true or false)	
	H_0 is true	H_0 is false
Reject H_0	Type I error probability = α	Correct decision probability = $1 - \beta$ = Power
Do not reject H_0	Correct decision probability = $1 - \alpha$	Type II error probability = β

Figure 5.8 Possible decisions about the null hypothesis (H_0)

the video-viewing experiment would expect scores to be higher in the '*funny*' video condition (E > C; where E stands for 'Experimental condition' and C for 'Control condition'). However, it is possible that people's moods might be worse after being shown the funny video, perhaps because they felt they were being 'manipulated'. If that happened and they scored lower on the logical reasoning test than those shown a neutral video, should the researcher conclude that the experiment showed an effect of the type of video? It all depends on the precise prediction that the experimenter made before collecting the data. If the researcher decided that, although a difference in favour of the funny video was expected, a difference in the opposite direction would be of interest, a **non-directional prediction** should be made; that is, the alternative to the null hypothesis (E = C: i.e., no significant difference) would be that there would be a significant difference in either direction (E > C or C > E). Then, if α was set at .05, we would be looking for an extreme positive or negative value of our statistic (independent groups *t*, in this case) at either end of the distribution of possible values when the null hypothesis is true; more specifically, a value among the .025 most extreme in either direction (see Figure 5.9a). If the value of the statistic falls in either tail (the *rejection regions*), we would conclude that the null hypothesis could be rejected at the 5% level and that there was a significant ($p < .05$) effect of the type of video viewed in a **two-tailed test** of the hypothesis. Sometimes, a two-tailed test is the only sensible option, as when you have two competing experimental conditions, rather than one experimental condition and one control condition.

If, on the other hand, the researcher decided that a difference in the non-expected direction would simply mean that the experiment had failed and was therefore of no interest, a **directional prediction** might be appropriate (e.g., E > C). In that case, if the *t*-statistic were among the .025 most extreme values in the 'wrong' tail (the one representing extreme differences in favour of the neutral video), the decision would be to fail to reject the null hypothesis and to conclude that the video effect was non-significant ($p > .05$) in a **one-tailed test** of the hypothesis. The gain from making the more specific directional prediction is that, if the difference between means is in the predicted direction, a lower value of the statistic (*t* in this example) will be needed

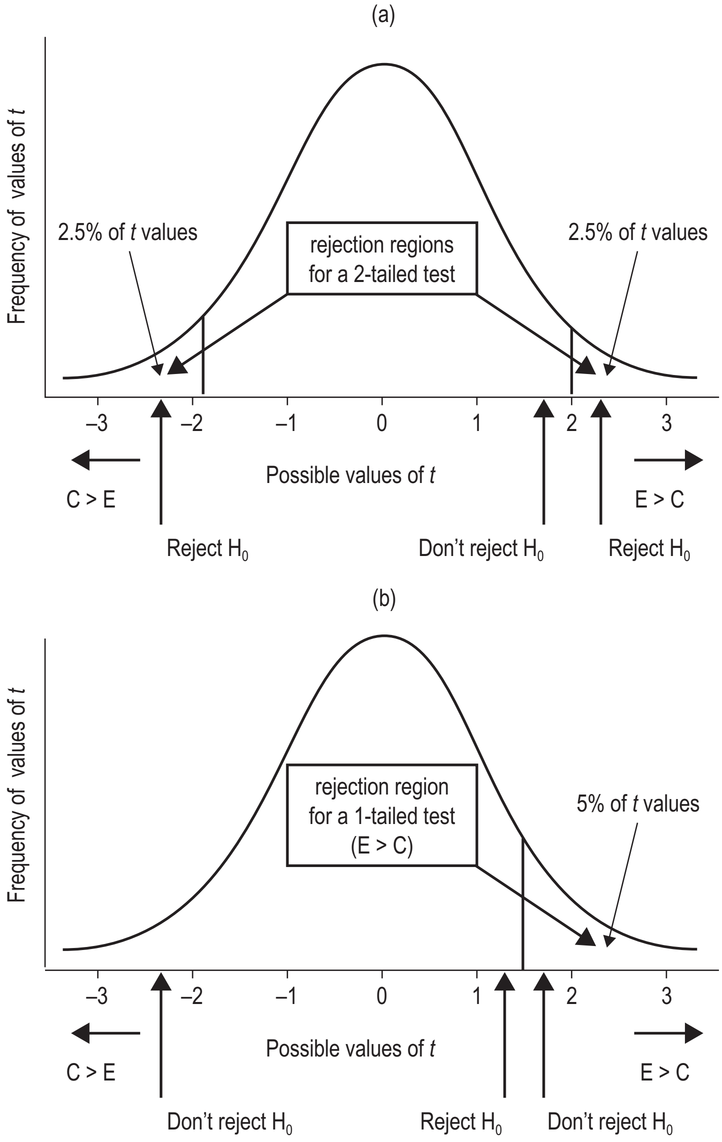

Figure 5.9 One- and two-tailed decisions: (a) two-tailed decision; (b) one-tailed decision

to achieve statistical significance. This is because the region for rejection of the null hypothesis will be the 5% most extreme values in the predicted direction (i.e., all .05 most extreme values are in one tail instead of being split .025 in each tail). This is illustrated in Figure 5.9b.

Complications (5.4) – The decision to do a one- or two-tailed test is 'set in stone'

It should be stressed that a decision to use a one-tailed test *must* be taken before the data have been collected. It is not acceptable to decide on a one-tailed test after you have seen what the data look like. In that case, a smaller value of the statistic would be needed to achieve significance, but it would be 'cheating'! Similarly, once a decision is made to do a one-tailed test, it would be unacceptable to change the decision (i.e., to do a two-tailed test) after it is seen that the difference went in the non-predicted direction. Once again, that would be 'cheating' – you would be looking at a rejection region of $p = .05$ in the originally predicted tail *plus* a rejection region of $p = .025$ in the originally non-predicted tail, so the real probability of the data being obtained when the null hypothesis is true would be .075, not .05! In view of the uncertainty about the stage at which a decision is made to opt for a one- or two-tailed test, some researchers take the view that the statistical test reported should *always* be two-tailed.

SUMMARY OF CHAPTER

- Knowing that an experiment has validity and that there are differences between DV scores in different conditions is not enough to infer that the IV has an effect on the DV. We still have to consider the possibility that these differences are determined by random NVs.
- Random NVs (i) do not pose a threat to the internal validity of an experiment; (ii) cannot be eliminated; (iii) increase the variability of scores on the DV within each condition; (iv) may occasionally pile up in favour of one condition to produce a large effect.
- In order to infer that differences in DV scores between the two conditions are so large that they cannot be due to the effects of random NVs (and therefore the hypothesis is correct), we make use of 'statistical inference'.
- Statistical inference consists of setting up a 'null hypothesis' (an hypothesis of 'no effect of the IV') and seeing whether it can be rejected as a likely explanation of any difference between scores in the two conditions. If it can be rejected at some level of confidence (probability), we infer that the difference between conditions is statistically significant at that level of probability.

- To test the null hypothesis we calculate a statistic from the data (different statistics are calculated depending on the research design) and we see whether it is among the most extreme values that would occur with a given probability (say, $p < .05$) if the null hypothesis were true and the experiment was repeated thousands of times.
- The statistical inference may be mistaken. We may find an effect when the null hypothesis is in fact true (Type I error), or we may fail to find an effect when the null hypothesis is in fact false (Type II error – this may mean that the experiment has insufficient 'power' to reveal an effect).
- If a directional prediction is made (e.g., 'scores will be higher in the experimental condition'), we can use a 'one-tailed' test, which requires a smaller value of the statistic to reach significance. If a non-directional prediction is made (i.e., 'scores in the two conditions will differ'), a two-tailed test must be used. The decision to use a one- or two-tailed test must be made before collecting the data.