CHAPTER NINE

Tests of Significance for Interval Data

In this chapter we discuss two important statistical tests that are used for testing hypotheses in an experimental design in which there are two conditions and the DV is measured on an interval scale. These tests are the **independent groups *t*-test**, which is appropriate when each condition of the experiment uses different participants, and the **related *t*-test**, which is suitable for an experiment in which the same participants are employed in both conditions (i.e., a repeated measures design) or there are matched pairs of participants. In addition, we will discuss the **one-sample *t*-test**, which is used when just one set of scores is obtained.

Remember that the fact that the DV has been measured by means of an interval scale is not in itself sufficient to warrant the use of a *t*-test. A *t*-test can be used only when parametric assumptions are met (see Chapter 6 for a detailed discussion on parametric and non-parametric data). More precisely, if a non-parametric test is called for and you have different participants in the conditions of the experiment, you should use the Mann–Whitney *U* test, while if the same participants or matched pairs of participants are used in the two conditions, then you should use the Wilcoxon Matched Pairs *T* test. These two non-parametric tests have been thoroughly discussed in Chapter 8.

Interval Data

We offered a detailed discussion of the nature of interval measurement in Chapter 6. However, we will briefly summarize this notion here. Basically, there are various ways of measuring a variable, and they vary in terms of the properties that the measurement scales possess. In an interval scale, not only do larger numbers mean more of whatever is being measured (which is a characteristic shared with ordinal scales) but, in addition, the intervals between numbers represent equal differences of the measured variable at all points of the scale. For instance, suppose that we are investigating the effect of noise on people's memory for words; we may set up an experiment in which we ask participants to read out 50 words either in a noisy environment or in a quiet one, and then we count the number of words that are remembered, expecting that participants who read the words out in the 'noise' condition would remember fewer words than participants in the 'quiet' condition. Clearly, the DV is the number of words remembered, and people can score from 0 (no words remembered at all) to 50 (all words remembered). In this case (provided that the words are all of comparable memorability), we may safely assume not only that, say, a score of 24 stands for a better memory than a score of 20, but also that the interval between 24 and 20 is broadly the same as that between 14 and 10 (or between 44 and 40, between 9 and 5, and so on).

The Independent-Groups *t*-test

Let us start by reminding you of the hypothetical experiment discussed in Chapter 2. This will allow us to introduce the first of the tests presented in this chapter, that is the 'independent groups *t*-test'.

Selecting a test for our 'mood' experiment

In the fictitious experiment used in Chapter 2, we wanted to test the following hypothesis: 'when people have a positive mood, their intellectual performance will be enhanced'. We therefore proposed to design an experiment involving two conditions. In the experimental condition a group of participants watch a movie excerpt with a funny content; in the control condition a different group of participants watch an excerpt with an emotionally neutral content. This should lead participants in the experimental condition to have a better mood than those in the control condition. To measure the level of intellectual performance we proposed to ask participants to solve 10 logical problems. (See Figure 2.1 for an overview of the experimental design.) Obviously, if our hypothesis is correct we should find that participants in the experimental condition (mood enhanced) tend to solve a higher number of logical problems (i.e., to have a better intellectual performance) than participants in the control condition (mood unaltered).

In Chapter 4 we presented a table (see Table 4.1) showing a hypothetical set of scores produced by respondents in both the experimental and the control conditions – remember that each score represents the number of logical problems solved by a specific participant. In that chapter we also calculated the mean score produced by respondents in each condition of the experiment. This was 6.8 in the experimental condition and 5.4 in the control condition. We can also tell you that the standard deviation was 1.3 in the experimental condition and 1.5 in the control condition. The means indicate that, as predicted, participants in the experimental condition did, overall, solve more problems than participants in the control condition. However, as we have often emphasized, the difference between means in itself cannot be used to infer that our hypothesis is correct: in order to make a proper inference we need to use a statistical test.

Which test should we use in order to see whether our hypothesis is correct or not? A careful look at Table 6.1 will help you to decide. We already know that, concerning the nature of the research question, we are looking for a causal effect of one variable (mood) on another (intellectual performance), which leads directly to the prediction of a difference between means. We also know that, concerning the type of experimental design that is used, we have allocated different participants to the two conditions and that, therefore, we have used an 'independent groups design'. (The reason why we used this type of design was discussed in Chapter 3, particularly in the section entitled '*Participant variables*'.) At this point the question we need to ask ourselves is concerned with the type of measurement scale that we have used. Did we use a nominal, ordinal, interval or ratio scale? Clearly, our scale was not a nominal one, because different scores do not refer only to different qualitative characteristics of the respondent, but refer to different degrees of our DV, that is 'intellectual performance'. Next, we can ask whether our scale was limited to an ordinal level. The answer is that it was not, because we may confidently assume that intervals between numbers are broadly equivalent at all points of our scale. So, at this point we know that we have at least an interval scale.

So, regarding our experiment on mood and performance, we now know that we are looking for a difference between means, that our design is an independent groups one, and that our DV is measured on at least an interval scale. If you look at Table 6.1 you will realize that we have only two options in terms of the statistical test to be selected: either an independent groups *t*-test or a Mann–Whitney *U* test. To choose between these two tests we need to decide whether parametric assumptions are reasonably met or not. In Chapter 6 we offered some useful rules of thumb on how to decide whether the parametric assumptions are met (see Figure 6.1 for a schematic summary of these rules). On the basis of these rules we can be reasonably confident that in our experiment the parametric assumptions are met. This is because (i) the variance of the samples in the two conditions ($1.3^2 = 1.69$ and $1.5^2 = 2.25$) does not differ substantially, and because (ii) the frequency distribution of scores in each condition is reasonably close to a normal distribution (see Figure 4.1 for histograms and Figure 4.3 for frequency polygons for the data in the conditions of our experiment). So, the statistical test we should use to ascertain whether our hypothesis is correct is the independent groups *t*-test (consider that this specific test may also be defined as an 'unmatched *t*-test', a '*t*-test for two independent means', a '*t*-test for unrelated samples' and an 'independent samples *t*-test'; so, don't worry if other books use one of these labels: what they mean is always the same thing!).

The logic of the independent groups t*-test*

Once you know that what you need is the independent groups *t*-test, all you have to do is to enter your data into a computer package and use the appropriate procedure to run the test (see SPSS operations and output (9.1) for how to run this test using SPSS). The package will perform a series of calculations on your data, based on a specific mathematical formula. Here we will explain the rationale behind this formula, but we will not explain its mathematical details, as this is beyond the scope of this book. (See Formulae (9.1) if you want to see what one version of the formula for the *t*-test looks like.)

The independent *t*-test focuses on two things. First, it looks at the *difference between* the mean of the scores obtained by participants in the experimental condition and the mean of the scores obtained by participants in the control condition. Second, the *t*-test is interested in the variability of the scores obtained by participants *within* each condition. What the formula does is to contrast the difference between the means obtained in the two conditions with the general variability of the scores within each condition. The *t*-value represents an indicator of this contrast, which is summarized in the following verbal formula:

$$t = \frac{\text{difference between the mean scores in the two conditions}}{\text{general variability of scores within each condition}}$$

Technically, the 'general variability of scores within each condition' is defined as the 'standard error of the differences between means', but don't worry about that at this stage. Just remember that the denominator in the equation is an estimate of how 'spread out' scores are within the two conditions. Broadly speaking, the more different the two means and the less variable the scores within each condition, the higher the value of *t*. On the contrary, the less different the two means and the more variable the scores within each condition, the lower the *t*-value. Clearly, when the means in the two conditions are very different and the scores within each group have little variability, the difference between the two means is probably due to the fact that participants in the two conditions were exposed to different levels of the IV (i.e., that the difference is not due to chance). On the other hand, when the means are very similar and the scores within each group have high variability, it is quite likely that random variability would be sufficient to produce a small preponderance of higher scores in one condition, and we can be almost certain that the difference between the means is due to chance. That also implies that the higher the value of *t* the smaller the probability that the difference between the means is due to chance (i.e., random NVs).

Formulae (9.1) – The independent groups *t*-test

The formula for the independent groups *t*-test varies depending on whether the number of participants in the two groups is equal or not. The simplest version of the formula is the one that holds only when group sizes are equal, and that is the formula given below:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2 + s_2^2}{n}}}$$

where the symbols mean:

$\bar{X}_1$ = mean of scores in condition 1
$\bar{X}_2$ = mean of scores in condition 2
s_1^2 = sample variance of condition 1 (see Formulae (4.2))
s_2^2 = sample variance of condition 2
n = number of participants in each condition

So, the arithmetical operation that you see in the 'numerator' (i.e., the expression on the top) calculates the difference between the mean obtained by participants in the experimental condition and the mean obtained by those who were in the control condition. On the other hand, the arithmetical operations that you can see in the 'denominator' (i.e., the expression underneath) calculate the general degree of variability of the scores obtained by participants within each condition (broadly speaking, this is equivalent to the average standard deviation within the two conditions).

If the number of participants in the two groups differs, the formula becomes a bit more complicated because the two sample variances in the denominator have to be weighted according to their sample sizes. The denominator in that formula will thus provide a **weighted average** of the two sample variances, usually referred to as a **pooled variance estimate**.

As you are unlikely ever to need to do these calculations by hand, you do not need to worry about the details of the formulae for computing *t*, and we are not even going to show you the formula that applies when sample sizes are unequal. The formula that is used when sample sizes are equal will suffice as an illustration in case you are interested in how the calculations are done.

If you do ever calculate a *t*-statistic yourself, or you are given a *t*-value without being told anything about the probability of it having arisen by chance (i.e., its statistical significance), provided you know the sample sizes of the two groups, you can use a statistical table (as in Statistical Table 9.1, a partial version of which is shown here) to see whether the *t*-value is large enough to be statistically significant. The table gives the critical values of *t* (i.e., the minimum value needed for statistical significance at various levels of probability) for different sample sizes.

To use the table, you need to know the calculated *t*-value (let's suppose it is $t = 2.62$) and the degrees of freedom (*df*s) for your *t*-statistic. The concept of degrees of freedom was explained briefly in Chapter 4 (see Complications (4.2)). As two standard deviations are computed on the way to calculating *t* for an independent groups design (i.e., one for each group), two degrees of freedom are lost. So, instead of referring directly to the total sample size ($n_1 + n_2 = N$), the table specifies the *df* for the calculation of *t* (always $N - 2$ for an independent groups design, because one *df* is lost for each group). As an example of using the table, suppose you had carried out

Statistical Table 9.1 Critical values of *t*. *t* is significant when it **equals or exceeds** the table value (partial table – full version in Appendix 1)

	level of significance for a one-tailed test						
	.10	.05	.025	.01	.005	.001	.0005
	level of significance for a two-tailed test						
df	.20	.10	.05	.02	$\boxed{.01}$	.002	.001
1	3.08	6.31	12.71	31.82	63.66	318.31	
2	1.89	2.92	4.30	6.96	9.92	22.33	
3	1.64	2.35	3.18	4.54	5.84	10.22	
28	1.31	1.70	2.05	2.47	2.76	3.41	
29	1.31	1.70	2.05	2.46	2.76	3.40	
30	1.31	1.70	2.04	2.46	$\boxed{2.75}$	3.39	
40	1.30	1.68	2.02	2.42	2.70	3.31	
60	1.30	1.67	2.00	2.39	2.66	3.23	
120	1.29	1.66	1.98	2.36	2.62	3.16	
2000	1.28	1.65	1.96	2.33	2.58	3.09	

Source: The entries in this table were computed by Pat Dugard, a freelance statistician.
For an independent groups (between Ss) test, $df = N - 2$ (where N is the total number of scores in both groups)
For a related (within Ss or matched pairs) test, $df = N - 1$ (where N is the number of pairs of scores)

an experiment with 20 participants in an experimental condition and 15 participants in a control condition. Then $N = 20 + 15 = 35$ and the *df*s are $35 - 2 = 33$. So, you look for 33 in the left-hand column of the table. If, as in this case, the required *df* value is not shown in the table, the cautious solution is to select the nearest *smaller* value ($df = 30$, in this example), and enter the table at that row. Next, you need to decide what alpha level (level of significance) you want to test for. Suppose you are interested in a two-tailed test at alpha = .01. You should look down that column third from the right and locate the critical value of *t* at its intersection with the row where $df = 30$. The critical value is 2.75. As the obtained *t*-value (2.62) is not as big as the critical value (2.75), you should conclude that the difference between the experimental and control means did not reach statistical significance at the 1% level (i.e., $p > .01$) in a two-tailed *t*-test.

Note that had you been testing a directional hypothesis (say, the experimental mean is greater than the control mean), you might have decided to use a one-tailed test and would have been looking down a different column (fourth from the right) and the critical value for *t* would have been 2.46. In that case, provided the experimental mean was indeed greater than the control mean, you would have concluded that the predicted difference in favour of the experimental condition was statistically significant at the 1% level ($p < .01$) in a one–tailed test.

We remind you (see Complications (5.1)) that, rather than just looking down the column containing critical values for a pre-selected alpha level, in practice, some researchers scan the columns to see what is the lowest level of significance that can be reported for the *t*-value they obtained. Thus, with 20 *df*s and *t*-values of 2.12 and 3.12 for two related experiments, a researcher might refer to Statistical Table 9.1 and report that the effect in the first experiment was significant at the 5% level ($p < .05$) and that the effect in the second experiment was significant at the 1% level ($p < .01$), both in two-tailed tests. We also remind you (see Complications (5.4)) that no such exploratory strategy can be used to decide whether to report a one- or two-tailed level of significance. The decision to use a one- or two-tailed test must *always* be made in advance.

On the subject of one- and two-tailed tests, Statistical Table 9.1 (in common with several of the other statistical tables in this book) is particularly useful for the way it makes clear the relationship between critical values for the statistic and significance levels for one- and two-tailed tests. Looking at the top of the table, you can see, for example, that any value of *t* that is significant at the 10% level ($p < .10$) in a two-tailed test will be significant at the 5% level ($p < .05$) in a one-tailed test. The general rule is: whatever the probability that the obtained value of a statistic can be attributed to chance in a two-tailed test, the probability that it can be attributed to chance in a one-tailed test will be half of that (see 'One- and two-tailed tests' in Chapter 5 for an explanation of this rule). An example of an analysis using an independent groups *t*-test is given in 'SPSS Operations and Output (9.1)'.

SPSS operations and output (9.1) – Computing an independent groups *t*-test

The data we will use in this example analysis are those shown, albeit in a different layout, in Table 4.1. To perform an independent groups *t*-test in SPSS, you must devote one column to the IV and one to the DV. In the column concerning the IV (which, with reference to our fictitious experiment, we might label as 'mood') you specify which condition each participant in the experiment belongs to (usually coded as 1 and 2). In the column about the DV (which we could label as 'perform') you specify the scores produced by all participants in the experiment. Then proceed as follows:

(i) Click on *Analyze*, from the menu at the top of the screen. Then click on *Compare means*, and then on *Independent Samples T-test*.

(ii) Move the DV from the rectangular box on the left side of the window into the box called *Test variable*.

(iii) Move the IV from the rectangular box on the left side of the window into the box called *Grouping variable*.

(iv) Click on *Define groups* and then type in the numbers used in your data file to refer to each condition (i.e., each independent group of participants) in your experiment. For instance, regarding our fictitious experiment, if we had used 1 = good mood and 2 = neutral mood we would type 1 in the *Group 1* box and 2 in the *Group 2* box.

(v) Click on *Continue* followed by *OK*.

(vi) If you want to look at the shapes of the distributions of scores in the two conditions to see whether they are approximately normal, click on *Data*, then *Split File*. Click on the radio button *Organize output by groups* and move the IV into the *Groups Based on* box and click *OK*.

(vii) Click on *Graphs*, then *Histogram*. In the box on the left, select the DV and move it into the *Variable* slot, then click on *Display normal curve*, followed by OK.

The output includes the following (we have not reproduced the histograms because they can be seen in Figure 4.1):

Group Statistics

	MOOD	N	Mean	Std. Deviation	Std. Error Mean
PERFORM	good mood	20	6.8000	1.2814	.2865
	neutral mood	20	5.3500	1.4609	.3267

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
PERFORM	Equal variances assumed	.511	.479	3.337	38	.002	1.4500	.4345	.5703	2.3297
	Equal variances not assumed			3.337	37.365	.002	1.4500	.4345	.5699	2.3301

The SPSS output (9.1) refers to the data set (see Table 4.1) based on our imaginary 'mood and intellectual performance' experiment. The mean and standard deviation for each condition of the experiment can be seen in the '*Group Statistics*' table. In the table called '*Independent Samples Test*', you can see, among other things, the value of *t* (which in this case is 3.337), the *df* (38) and the probability of obtaining that specific value of *t* by chance in a two-tailed test of the hypothesis (which is .002) – see column labelled 'Sig. (2-tailed)'.

Note that the result of another test (using the statistic, *F*, which is not dealt with in this book) is presented towards the left of the 'Independent Samples Test' box. This is a test to see whether the parametric assumption of 'equality of variances' is met. If the variances differ significantly, you should look across the 'Equal variances not assumed' row of the table. Otherwise, as in this case, you should look across the 'Equal variances assumed' row. The probability (.002) is the same in both rows in this example because the variances in the two conditions are very similar (see squares of SDs from the 'Group Statistics' box – i.e., $1.2814^2 = 1.64$ and $1.4609^2 = 2.13$). Note that the *df* in the 'Equal variances not assumed' row is slightly reduced from *df* = 38. This is how the program makes allowance for non-equality of variances, though the reduction in *df* is too small in this case to affect the probability (.002). If you ever need to report the *df* for an 'Equal variances not assumed' solution, you should round the *df* down to a whole number, in this case from 37.365 to 37.

Finally, note that SPSS often produces additional output that exceeds your current needs (e.g., 'Std. Error Mean', ' Standard Error Difference' and '95% Confidence Interval of the Difference' in the output above), and that at present you can safely ignore.

Drawing inferences about our 'mood' experiment by using a t*-test*

As we explained in Chapter 5, experimental psychologists normally accept a value of t that has less than a 5% probability of being obtained by chance, as an indication that the experimental hypothesis is supported. So, if we submit the data collected in our experiment on the effects of mood on performance to a t-test, we end up with $t = 3.34$ (note that it is usually sensible to report statistical values to a maximum of two decimal places). The probability of obtaining this specific t-value by chance, in a study involving two groups of 20 participants (i.e., *df* = 38), are two in one-thousand, or, if you prefer, 0.2% (experimental psychologists and statisticians express this idea as $p = 0.002$, as explained in Chapter 5). Obviously, this probability is less than 5% (and, indeed, less than 1%), and so, provided we had a valid experimental design, we can infer that our manipulation probably had a strong effect, in the sense that participants in the experimental condition (good mood) performed better than participants in the control condition (neutral mood). Therefore, we may conclude that our experimental hypothesis – that 'when people have a positive mood, their intellectual performance will be enhanced' – was supported (or, technically speaking, the null hypothesis can be rejected).

Remember that $t = 3.34$ is not necessarily associated with $p = 0.002$. This is so in our experiment given the specific number of participants in each condition. However, with a different number of participants per condition, this t-value would be associated with a different value of p. This is because in that case our experiment would have different degrees of freedom (reported as '*df*' in the SPSS output). Basically, given the same value of t, the more the degrees of freedom, the smaller the value of p. Putting it another way, the more degrees of freedom, the smaller will be the value of t needed for a given level of statistical significance.

You should also remember another thing. We are assuming that our hypothesis is correct on the basis of the values of t and p, but this is only because we know that the mean score in the experimental condition was bigger than the mean score in the control condition (thereby showing that intellectual performance was better under good mood). But consider that, had the means been the other way around (i.e., 5.4 in the experimental condition and 6.8 in the control condition), you would have obtained the same *t*-value (except that it would have been a negative value, i.e., $t = -3.337$) and the same *p*-value. However, in this case, the difference between the scores on intellectual performance in the two conditions would not, as indicated by the negative value of *t*, have been in the predicted direction! It goes without saying that, in this case, the hypothesis would have probably been wrong, and the null hypothesis could not have been rejected. In other words, our hypothesis was a directional hypothesis (i.e., we predicted not only that the two conditions would produce significantly different scores, but also the direction of this difference), therefore, only a difference in favour of the scores in the experimental condition will allow us to reject the null hypothesis. (See Chapter 5 for more information on the notion of directional hypotheses.) If we had decided, before collecting our data, to carry out a one-tailed test of the directional hypothesis, we would need to halve the two-tailed probability provided in the SPSS output, i.e., the one-tailed probability would be .001. If we had opted for a one-tailed test, therefore, even a *p*-value of .10 or less in the SPSS output would have been sufficient for us to report a significant one-tailed effect.

Additional information (9.1) – Effect size

If you use a great many participants, your experiment has a *high power*, and it is unlikely to miss a real effect even if it is very small (see Chapter 5, specifically in the section on ‘Statistical decision errors’ and in Additional information (5.6) for a discussion of ‘power’). A very small effect that is picked up because the experiment has high power may be of limited theoretical or practical significance. Therefore, in addition to the usual information about statistical significance, it is useful to know whether the effect that was found is a large or small effect. Indeed, an increasing number of psychology journals now insist that information about **effect size** is reported along with the usual information about statistical significance. There are several measures of effect size in use. An intuitively meaningful one – in relation to the parametric analyses discussed in this chapter – is that defined as the difference between means in units of standard deviation. This is known as a standardized measure of the kind discussed in Chapter 4 in the section on ‘*z*-scores’. The point about standardized measures is that they provide a stable interpretation of differences between measures regardless of the scale on which they are measured. For example, knowing that the mean number of problems solved in the two conditions of the mood experiment were 6.80 and 5.35 does not give us much idea of whether this is a ‘big’ or a ‘small’ difference. If we tell you, however, that the two means differ by 1.06 standard deviations and that an approximate guide to effect size (see Cohen in the ‘brief list of recommended books’) is that a difference of .2 of a SD is a small effect, a difference of .5 SD is a medium effect and a difference of .8 SD is a large effect, you can immediately conclude that the effect we found was quite large, as well as being statistically significant. The calculation of effect size (signified by '*d*'), is straightforward:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{SD}$$

where *SD* is the mean standard deviation of the two groups. Remember that equal variances (and of course SDs) in the populations from which the samples are drawn is a required assumption for parametric tests and the average of the two sample SDs is the best estimate we have of the joint population SD. (Note that if the numbers of participants in the two groups are not the same, a *weighted* average of the SDs has to be calculated – see Howell in the 'brief list of recommended books'.)

Reporting results (9.1) – The independent groups *t*-test

In a report of the experiment (assuming that a two-tailed test had been decided on) the result could be described as follows:

> In an independent groups *t*-test of the hypothesis that positive mood would result in higher intellectual performance than neutral mood, the difference in number of problems solved was in the predicted direction (positive mood mean = 6.80; neutral mood mean = 5.35) and was statistically significant (t (df = 38) = 3.34; $p < .05$; two-tailed). The effect size was given by $d = 1.06$.

Recall that some researchers would report the *lowest* conventional level of statistical probability reached (i.e., $p < .01$), rather than a predetermined alpha level (e.g., $p < .05$).

The Related (Repeated Measures) *t*-test

An imaginary experiment

Let us now consider another fictitious experiment. Suppose that we are studying 'spider phobia'. Then suppose that we have a theory according to which, because there are many venomous spiders, during evolution the human species has developed an 'adaptive' fear of spiders. This theory also holds that, because there are many more venomous spiders among the hairy than among the non-hairy ones, humans will find hairy spiders more scary than non-hairy ones. This theory sees phobia of spiders as just an exaggerated expression of this 'natural' fear, and therefore it predicts that, although spider-phobic people are afraid of all spiders, they tend to fear hairy spiders more than non-hairy ones. Now, in order to test this hypothesis, we can recruit 20 individuals who have been diagnosed as spider phobic, and show them a series of, say, 10 three-dimensional, very vivid pictures of different spiders, five of which belong to hairy species and five of which belong to non-hairy ones. While participants observe the different pictures, we can assess their level of anxiety by recording their pulse-rate. (Obviously, the order of presentation of pictures should be counterbalanced, as it is possible that the spiders observed later would elicit less anxiety than the ones observed earlier because of habituation.) Clearly, we expect that pulse-rate will be higher when participants see pictures of hairy spiders then when they see pictures of non-hairy ones. Finally, suppose that we find that, on average, participants' pulse-rate is 108.50 beats per minute (with a standard deviation of 10.20) when exposed to pictures of hairy spiders, and 105.75 beats per minute (with a standard deviation of 9.16) when exposed to pictures of non-hairy spiders.

Selecting a test for the 'spider phobia' experiment

Although the means indicate, as predicted, that viewing hairy spiders produces higher pulse-rates than viewing non-hairy ones, which statistical test should be used to test statistically the hypothesis that spider-phobic people find hairy spiders more scary than non-hairy ones? As usual we must start by deciding what we are testing for. We are clearly testing for a difference between conditions, as we want to know whether seeing different types of spiders produces different emotional responses. Second, we need to consider the type of research design we are using. Basically, we have two conditions, one in which participants are shown hairy spiders, and one in which the *same* participants are shown non-hairy spiders. Now, this is clearly a repeated measures design, as the same people are employed in both conditions of the experiment. Third, we must decide what kind of scale we have used to measure the DV, which is about the level of fear prompted by the view of spiders. We can consider our scale as an interval scale, as the intervals between the various levels of pulse-rate can be understood as being broadly the same. Finally, we need to know whether parametric assumptions are met. Given that the variability of the scores in the conditions is similar, as indicated by the standard deviations (10.20 and 9.16), and providing that the distribution of the pulse-rates for each type of spider was broadly similar to a normal distribution, we can say that parametric assumptions are reasonably met. At this point we can choose our statistical test; we are looking for a difference between conditions, we have a repeated measures design, we have used an interval measurement scale, and parametric assumptions are met: the test to be used is the related *t*-test! Had the parametric assumptions not been reasonably met, the test of choice would be the Wilcoxon Matched Pairs *T* test (see Table 6.1).

Additional information (9.2) – Control conditions

When we first introduced the notion of experimental and control conditions in Chapter 2, we explained that not all experiments include a control condition. This is the case in the spider phobia experiment. Here, the presentation of non-hairy spiders pictures is intended to produce an effect on participants (unlike the neutral condition in the mood experiment), and so it cannot be strictly defined as a 'control' condition. Therefore, conditions should not be labelled as experimental and control conditions; instead, they should be given descriptive names (e.g., *hairy* and *non-hairy*). Whether or not it is referred to as such, a control condition is one that involves a treatment that is the same as in the experimental condition in all respects except the critical one that is the subject of the hypothesis. A clear example of this is in treatment (e.g., drug) evaluation studies, where the control condition involves the giving of a *placebo* (something that appears the same as the treatment, but lacks the active ingredient).

At this point you should proceed as usual. That is, you enter your data into a datafile and analyse it using SPSS (see SPSS operations and output (9.2)). With alpha set at .05, a probability of $p < .05$ will allow you to reject the null hypothesis (see SPSS operations and output (9.2) to find the t-value and the two-tailed probability of obtaining that t-value by chance).

The logic of the related (repeated measures) t*-test*

As with the independent groups t-test, once you know that what you need is the related t-test, you only need to enter your data into a computer package and use the appropriate procedure to run the test (see SPSS operations and output (9.2) for how to run a related t-test using SPSS). The rationale behind the formula for a related t-test is basically the same as for the independent groups t-test. The principal difference is that the scores in the two conditions are converted to a single set of *difference* scores. For each participant, the score in one condition is subtracted from the score in the other condition. The analysis is then carried out on these difference scores. If the null hypothesis is true, we should expect the positive and negative differences to approximately balance out. So, under the null hypothesis, the mean of the differences is predicted to be zero. Technically, the mean of the hypothetical *population* of difference scores, of which our difference scores are a sample, is hypothesized to be zero.

Additional information (9.3) – Analysis of data from a matched pairs design

A matched pairs design is also analysed using a related *t*-test. Here there are different participants in the two conditions, but each participant in the 'hairy' condition has been matched with a participant in the 'non-hairy' condition. The variable on which they have been matched will be one, such as 'severity of phobia', which is likely to have a strong effect on the DV (pulse-rate) regardless of which condition the participant is in. Thus, a severe phobic is likely to have higher pulse-rates than a less severe phobic both in the hairy and non-hairy conditions. For this reason, each matched pair of participants can be treated as though they were the same participant being exposed to both conditions and the difference between their scores in the two conditions can be used in a related *t*-test, just as for a repeated measures design where it really is the same participant being exposed to both conditions. (See Additional information (6.1) for a discussion of the extent to which participant NVs are effectively controlled in a matched pairs design.)

As with the independent groups *t*-test, the related *t*-test contrasts two things. In this case, it contrasts the extent to which the mean of the sample of difference scores deviates from a population mean of zero with the variability within the sample of difference scores. Again, the *t*-value is an indicator of this contrast and may be summarized in the following verbal formula:

$$t = \frac{\text{difference between sample and population means of difference scores}}{\text{variability of difference scores within the sample}}$$

Technically, the 'variability of difference scores within the sample' is defined as the 'standard error of the mean' of the difference scores, but, once again, you do not need to be concerned about that at this stage. You just need to know that the denominator in the equation is an estimate of how 'spread out' the difference scores are. As with the independent groups *t*-test, the greater the difference in the numerator and the smaller the variability in the denominator, the higher the value of *t*. In this case, it can be inferred that the higher the value of *t* the smaller the probability that the deviation of the mean of difference scores from zero is due to chance (i.e., random NVs). The statistical formula can be seen in Formulae (9.2).

Formulae (9.2) – The related (repeated measures) *t*-test

$$t = \frac{\bar{D} - 0}{\frac{s_D}{\sqrt{n}}}$$

where the symbols mean:

$\bar{D}$ = mean of difference scores
s_D = standard deviation of difference scores (standard error of their mean)
n = number of difference scores (number of participants when repeated measures or number of matched pairs)

So, the arithmetical operation that you see in the numerator calculates the difference between the mean of the participants' difference scores and the hypothetical population mean of zero. On the other hand, the arithmetical operations that you can see in the denominator calculate the degree of variability of the sample of difference scores.

Once again, as you are unlikely ever to need to do these calculations by hand, you do not need to worry about the details of the formulae for computing *t*.

If you ever calculate a related *t*-statistic yourself, or you are given a related *t*-value without being told anything about the probability of it having arisen by chance (i.e., its statistical significance), provided you know the number of pairs of scores (usually, the total number of participants, but it would be the number of matched pairs of participants if you were using a matched pairs design), you can use the same statistical table (Statistical Table 9.1) to see whether the *t*-value is large enough to be statistically significant. The critical values of *t* given in the table are interpreted in the usual way (i.e., they are the minimum values needed for statistical significance at various levels of probability) for different sample sizes. Note, however, that the degrees of freedom for a related *t*-test are $N - 1$, where N is the number of pairs of scores (i.e., difference scores). This is because only one standard deviation has to be computed on the way to obtaining a related *t*-value (see Formulae (9.2)).

SPSS operations and output (9.2) – Computing a related *t*-test

To perform a related *t*-test in SPSS, you must create two columns, one for each condition of the experiment. That is, under one column you will include scores produced by each participant in one condition, and under the other column you will type scores produced by the same (or matched) participants in the other condition. For instance, concerning our 'spider phobia' experiment, you may create a column called 'hairy' (meaning 'hairy spiders') and a column called 'nonhairy' (meaning 'non-hairy' spiders) and enter participants' average

Paired Samples Statistics

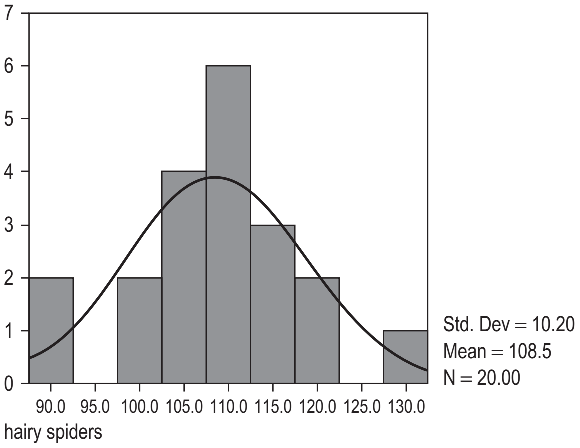
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	hairy spiders	108.5000	20	10.1955	2.2798
	non-hairy spiders	105.7500	20	9.1587	2.0479

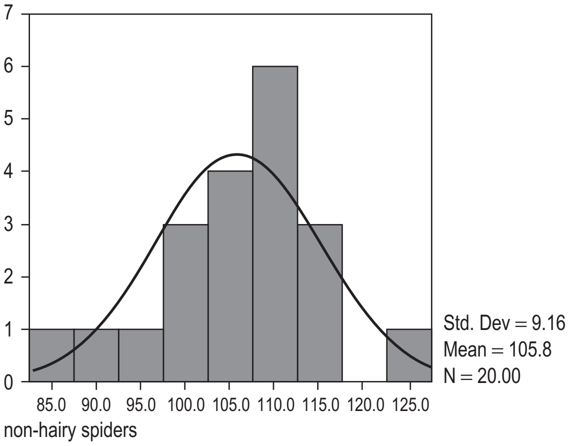
Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	hairy spiders & non-hairy spiders	20	.754	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	hairy spiders – non-hairy spiders	2.7500	6.8509	1.5319	–.4563	5.9563	1.795	19	.089





pulse-rate when seeing each type of spider in the relevant column (by 'average' we mean that, given that participants' pulse-rate was taken five times for each type of spider, you need to calculate the mean of these five measures of pulse-rate). Then proceed as follows:

1. Click on *Analyze*, from the menu at the top of the screen. Then click on *Compare means*, and then on *Paired Samples T-test*.
2. Click on the two variables that you want to compare (e.g., hairy and nonhairy). As they are highlighted, move them into the box called *Paired variables* by clicking on the arrow button.
3. Click on *OK*.
4. In order to look at distributions of scores in the two conditions, click on *Graphs*, then *Histogram*. In the box on the left, select one of the conditions and move it into the *Variable* slot, then click on *Display normal curve*, followed by OK. Repeat with the other condition entered in the *Variable* slot.

The SPSS output (9.2) refers to a data set (see Table 9.1) based on imaginary results that emerged from our hypothetical 'spider phobia' experiment. The mean and standard deviation for each condition of the experiment can be seen in the '*Paired Samples Statistics*' table. In the table called '*Paired Samples Test*', you can see, among other things, the value of *t*, which in this case is 1.795, the *df*, which is 19 and the probability of

Table 9.1 Hypothetical data for a repeated measures design: pulse-rates (beats per minute) of spider phobics when viewing hairy and non-hairy spiders

Participant	IV: type of spider	
	hairy	non-hairy
1	110	113
2	115	111
3	110	103
4	111	104
5	103	98
6	111	107
7	130	125
8	89	91
9	116	110
10	121	117
11	112	112
12	104	87
13	119	108
14	100	99
15	104	110
16	104	100
17	98	113
18	115	110
19	110	103
20	88	94

obtaining that specific value of *t* in a test of a two-tailed hypothesis, which is .089 (see column labelled 'Sig. (2-tailed)'.

The information in the 'Paired Samples Correlations' table simply tells you that there is a fairly strong positive relationship between scores in the two conditions. Measures of relationship (or *correlation*) between variables will be discussed in detail in Chapter 10. For the moment, just remember that a positive correlation implies that people who have relatively high pulse-rates in the 'hairy' condition tend also to have relatively high pulse-rates in the 'non-hairy' condition, and those with relatively low pulse-rates in the 'hairy' condition tend also to have relatively low pulse-rates in the 'non-hairy' condition. This suggests that a repeated measures design was a good choice because participant differences have a marked effect on the DV and are therefore well worth controlling (see Chapter 3 under the heading 'Controlling participant nuisance variables').

As the histograms indicate that both distributions of scores approximate to normal distributions and their variances do not differ greatly ($10.20^2 = 104.04$ and $9.16^2 = 83.91$), the assumptions required for a parametric test may be considered met. Note that for the related *t*-test, SPSS does not provide any adjustment for non-equal variances and it would be particularly unwise to use the parametric related *t*-test on data with substantially different variances in the two conditions. In that case, the alternative (non-parametric) Wilcoxon Matched Pairs *T* Test should certainly be preferred.

Recall that it is desirable to report *effect size* as well as statistical significance information. In this case, effect size (d) is given by the difference between the means for the two conditions divided by the standard deviation of the difference scores, which is given in the 'Paired Samples Test' table as 6.8509. So $d = (108.50 - 105.75)/6.85 = .40$. Using Cohen's convention, this is a small to medium effect.

Reporting results (9.2) – related *t*-test

The way the result would be described in a report of the experiment (assuming that alpha was set at .05 and a two-tailed test had been decided on) would be something like this:

> In a related *t*-test of the hypothesis that spider phobics would have higher pulse-rates when viewing hairy spiders than when viewing non-hairy spiders, the difference was in the predicted direction (hairy mean = 108.50; non-hairy mean = 105.75) but was statistically non-significant in a two-tailed test (t ($df = 19$) = 1.80; $p > .05$). The effect size was $d = .40$.

Note that, even though a directional prediction was made, it is not 'wrong' to decide on a two-tailed test. But note, also, that had a one-tailed test been decided on before the data were collected, the difference would have reached statistical significance because the one-tailed probability would have been .089/2 = .044 (i.e., $p < .05$).

The One-Sample *t*-test

In some circumstances you may want to obtain scores on a DV in just one condition in order to compare the mean of those scores with some already known mean of another set of scores on that DV. As only one sample of scores is used in the analysis, the parametric test used is known as the 'one-sample *t*-test'. The usual parametric assumptions apply even though one of the means was obtained on a previous occasion, probably by a different researcher.

An example of an occasion when it might be appropriate to collect a single set of scores would be when you wanted to know whether the mean reading age of a group of 10-year-old children, who had been selected for participation in a remedial reading experiment, was significantly lower than that of the population of 10-year-old children in the schools from which the sample came. Provided that the reading test to be used had previously been administered to a representative sample of the population of 10-year-olds in the schools in question, the test could now be administered to the 'remedial' sample and a one-sample *t*-test could be used to test the difference between their mean reading age and that of the representative sample of 10-year-olds. Another example would be when it was required to establish that a group of people who were to take part in an experiment were 'typical' of the population from which the sample was drawn; that is, that the sample mean did not differ significantly from the known *population mean* on the DV (e.g., intelligence; reading age; reaction time). Another possibility would be that you intended to do an experiment similar to one you had seen reported and would like some assurance that the participants you had selected were similar to the participants in the reported experiment in some important respect. You might want the assurance so that you could make comparisons between the results of the reported experiment and the one that you would be carrying out.

The logic of the one-sample t*-test*

Essentially, the one-sample *t*-test is a more *general version* of the related *t*-test, because the known mean that the single set of scores is compared to can be any value. In the first example above, if the mean reading age of a representative sample of 10-year-olds in the relevant schools were 9.5 years, that would be the mean against which the scores of the 'remedial' sample would be tested. In another scenario, the mean IQ (intelligence quotient) of the UK population of 10-year-olds is often *set* at 100, with 'more intelligent' children having scores ranging above 100 and 'less intelligent' children having scores ranging below 100. If we wanted to establish whether a sample of 10-year-olds whom we intended to use in an experiment were *typical* of the UK population of 10-year-olds, we would test the IQs of our sample against the known mean of 100.

In the case of the related *t*-test, the situation is more constrained. Then, we are always working with a sample of *difference scores* and the hypothetical mean difference against which they are compared is always *zero*. In this sense the related *t*-test can be seen as a special case of the one-sample *t*-test, because the comparison mean for the related *t*-test is always the same – zero.

Another way of construing the relation between these two tests is to think of the sample of scores used in the one-sample *t*-tests as a set of *difference scores*, in which the second score to be subtracted from the first is zero. It is as though the first condition contained the single sample scores and the second condition contained a column of zeros. So, the formula for one-sample *t* (see Formulae (9.3)) is effectively the same as that for related *t*, the difference being that the value in the numerator that is subtracted does not have to be zero. This means that in SPSS it is necessary to specify the value of the population mean that is to be subtracted from the sample mean.

Formulae (9.3) – The one-sample *t*-test

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

where the symbols mean:

$\bar{X}$ = mean of scores in the single sample
μ = known mean of population from which the sample is drawn
s = standard deviation of scores in the single sample
n = number of scores (i.e., participants) in the single sample

Remember that the related *t*-test and the one-sample *t*-test are effectively doing the same job. In fact, if you wished, you could compute related *t* using the SPSS procedure for a one-sample *t*-test (the procedure for a one-sample *t*-test is shown in SPSS operations and output (9.3)). All you would need to do would be to specify the value of the 'Test Value' (i.e., the population mean) as zero. Conversely, you could compute one-sample *t* using the SPSS procedure for a related *t*-test (see SPSS operations and output (9.2)). In this case, all you would need to do would be to enter the sample scores under the first condition and a column of zeros under the second condition.

An illustrative set of IQ scores is provided for a sample of 15 10-year-olds in Table 9.2. In this example, the reason for calculating a one-sample *t*-value might be to establish whether the mean of the sample scores differs significantly from the known 10-year-old population mean of 100.

Table 9.2 Hypothetical data for a one sample design: IQ scores of a sample of 10-year-old children

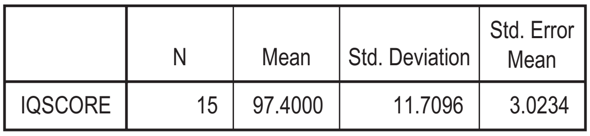
Participant	IQ score
1	95
2	87
3	101
4	96
5	105
6	116
7	102
8	81
9	97
10	90
11	123
12	86
13	95
14	104
15	83

SPSS operations and output (9.3) – Computing a one-sample *t*-test

To perform a one-sample *t*-test, you must enter the sample scores in one column and then proceed as follows:

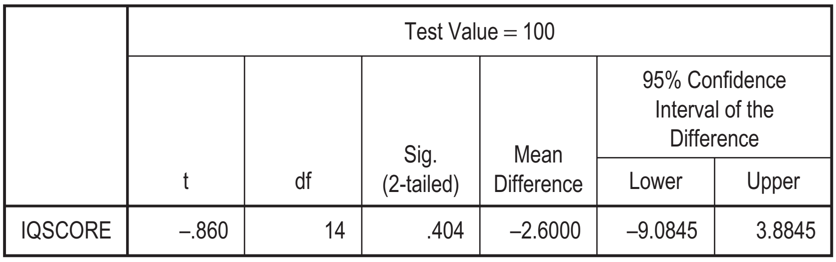
(i) Click on *Analyze*, from the menu at the top of the screen. Then click on *Compare Means*, and then on *One-Sample T test*.
(ii) Move the DV scores for the sample into the *Test Variable(s)* box.
(iii) Enter '100' in the *Test Value* box and click OK.

One-Sample Statistics



	N	Mean	Std. Deviation	Std. Error Mean
IQSCORE	15	97.4000	11.7096	3.0234

One-Sample Test



	Test Value = 100					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
IQSCORE	–.860	14	.404	–2.6000	–9.0845	3.8845

The SPSS output (9.3) refers to the imaginary IQ data set in Table 9.2. The mean (97.40) and standard deviation (11.71) of the sample of DV scores are shown in the table called '*One-Sample Statistics*'. The '*One-Sample Test*' table shows the value of t (–.860). This is a negative value because the sample mean is lower than the population mean of 100. That table also shows the df ($N - 1 = 14$) and the two-tailed probability of obtaining the specific value of t when the null hypothesis is true.

Once again, we should report effect size as well as statistical significance information. In this case, effect size (d) is given by the difference between the mean of the sample and the population mean divided by the standard deviation of the population (which is set at 15 by the IQ test constructors). So $d = (97.4 - 100)/15 = -.17$ (the minus sign can be ignored). Using Cohen's convention, this is a very small effect.

Reporting results (9.3) – One-sample *t*-test

The way the result would be described in a report of the study (assuming that alpha was set at .05 and a two-tailed test had been decided on) would be something like this:

> In a one-sample t-test of the hypothesis that the mean IQ of a sample of 10-year-olds would differ from the population mean of 100, the sample mean of 97.40 did not differ significantly from the population mean (t (df = 14) = .86; $p > .05$; two-tailed). It is therefore reasonable to treat the sample as representative of the population with respect to IQ. The effect size was given by $d = .17$.

SUMMARY OF CHAPTER

- In an experiment, when the DV has been measured on an interval scale (or a scale intermediate between ordinal and interval) and parametric assumptions are met, the statistical test to be used is a t-test.
- With an independent groups design the specific t-test to use is the independent groups t-test. This contrasts the difference between the mean scores in the two conditions of the experiment with the general variability of the scores within each condition. The t-value represents an indicator of this contrast. The higher the value of t, the lower the probability that the observed difference between means emerged by chance. If the probability of obtaining a specific value of t is less than 5%, and the difference between means is in the right direction, the null hypothesis can be rejected.
- When the experiment is based on a repeated measures design, the t-test to be used is the related t-test. This contrasts the mean difference between

participants' scores in two conditions with a difference of zero assumed for a hypothetical population of difference scores when the null hypothesis is true.

- When the experiment is based on a matched pairs design, we use the related *t*-test again. This test contrasts the mean difference between matched pairs of participants' scores in two conditions with a population difference of zero.
- When a single sample design is used, the *t*-test to be used is the one-sample *t*-test. This is similar to the related *t*-test, but in this case the mean of the single sample of scores is contrasted with the known mean of the population of scores from which the sample was drawn.