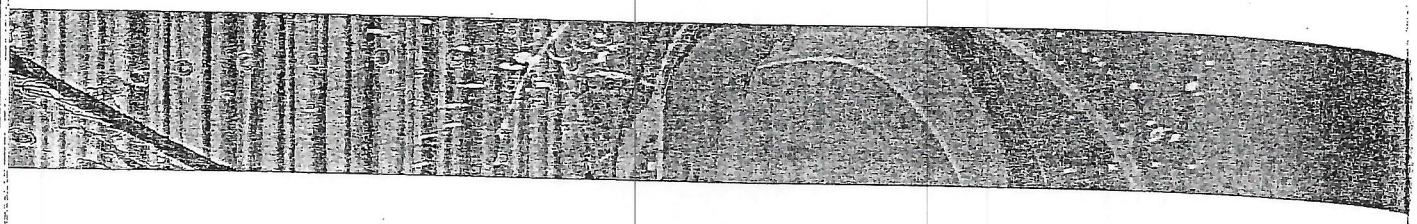
9 Comparing two means



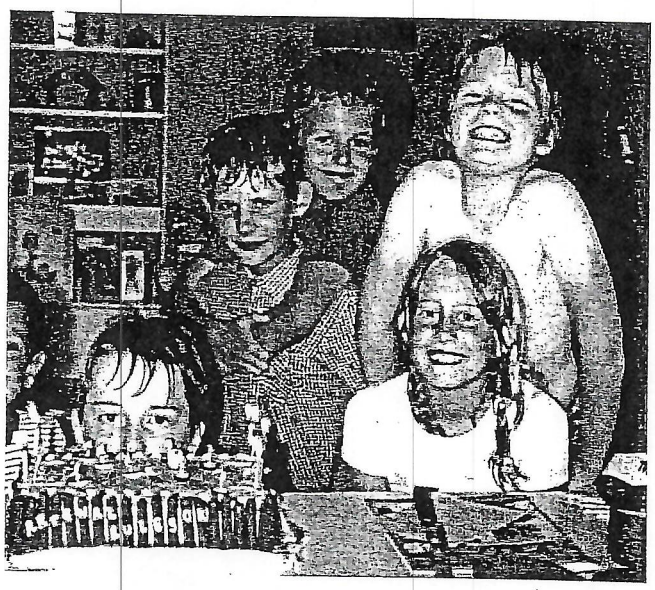


FIGURE 9.1
My (probably) 8th birthday. L–R: My brother Paul (who still hides behind cakes rather than have his photo taken), Paul Spreckley, Alan Palsey, Clair Sparks and me

9.1. What will this chapter tell me? ①

Having successfully slayed audiences at holiday camps around the country, my next step towards global domination was my primary school. I had learnt another Chuck Berry song ('Johnny B. Goode'), but also broadened my repertoire to include songs by other artists (I have a feeling 'Over the edge' by Status Quo was one of them).[1] Needless to say, when the opportunity came to play at a school assembly I jumped at it. The headmaster tried to have me banned,[2] but the show went on. It was a huge success (I want to reiterate my

[1] This would have been about 1982, so just before they became the most laughably bad band on the planet. Some would argue that they were *always* the most laughably bad band on the planet, but they were the first band that I called my favourite band.

[2] Seriously! Can you imagine, a headmaster banning a 10 year old from assembly? By this time I had an electric guitar and he used to play hymns on an acoustic guitar; I can assume only that he somehow lost all perspective on the situation and decided that a 10 year old blasting out some Quo in a squeaky little voice was subversive or something.

Skip to 9.3.1 - Pg 325

earlier point that 10 year olds are very easily impressed). My classmates carried me around the playground on their shoulders. I was a hero. Around this time I had a childhood sweetheart called Clair Sparks. Actually, we had been sweethearts since before my new-found rock legend status. I don't think the guitar playing and singing impressed her much, but she rode a motorbike (really, a little child's one) which impressed *me* quite a lot; I was utterly convinced that we would one day get married and live happily ever after. I was utterly convinced, that is, until she ran off with Simon Hudson. Being 10, she probably literally did run off with him – across the playground. To make this important decision of which boyfriend to have, Clair had needed to compare two things (Andy and Simon) to see which one was better; sometimes in science we want to do the same thing, to compare two things to see if there is evidence that one is different to the other. This chapter is about the process of comparing two means using a *t*-test.

9.2. Looking at differences ①

Rather than looking at relationships between variables, researchers are sometimes interested in looking at differences between groups of people. In particular, in experimental research we often want to manipulate what happens to people so that we can make causal inferences. For example, if we take two groups of people and randomly assign one group a programme of dieting pills and the other group a programme of sugar pills (which they think will help them lose weight) then if the people who take the dieting pills lose more weight than those on the sugar pills we can infer that the diet pills caused the weight loss. This is a powerful research tool because it goes one step beyond merely observing variables and looking for relationships (as in correlation and regression).[3] This chapter is the first of many that looks at this kind of research scenario, and we start with the simplest scenario: when we have two groups, or, to be more specific, when we want to compare two means. As we have seen (Chapter 1), there are two different ways of collecting data: we can either expose different people to different experimental manipulations (*between-group* or *independent* design), or take a single group of people and expose them to different experimental manipulations at different points in time (a *repeated-measures* design).

9.2.1. A problem with error bar graphs of repeated-measures designs ①

We also saw in Chapter 4 that it is important to visualize group differences using error bars. We're now going to look at a problem that occurs when we graph repeated-measures error bars. To do this, we're going to look at an example that I use throughout this chapter (not because I am too lazy to think up different data sets, but because it allows me to illustrate various things). The example relates to whether arachnophobia (fear of spiders) is specific to real spiders or whether pictures of spiders can evoke similar levels of anxiety. Twenty-four arachnophobes were used in all. Twelve were asked to play with a big hairy tarantula spider with big fangs and an evil look in its eight eyes. Their subsequent anxiety was measured. The remaining twelve were shown only pictures of the same big hairy tarantula and again their anxiety was measured. The data are in Table 9.1 (and spiderBG.sav if you're having difficulty entering them into SPSS yourself). Remember that each row in the data editor represents a

[3] People sometimes get confused and think that certain statistical procedures allow causal inferences and others don't (see Jane Superbrain Box 1.4).

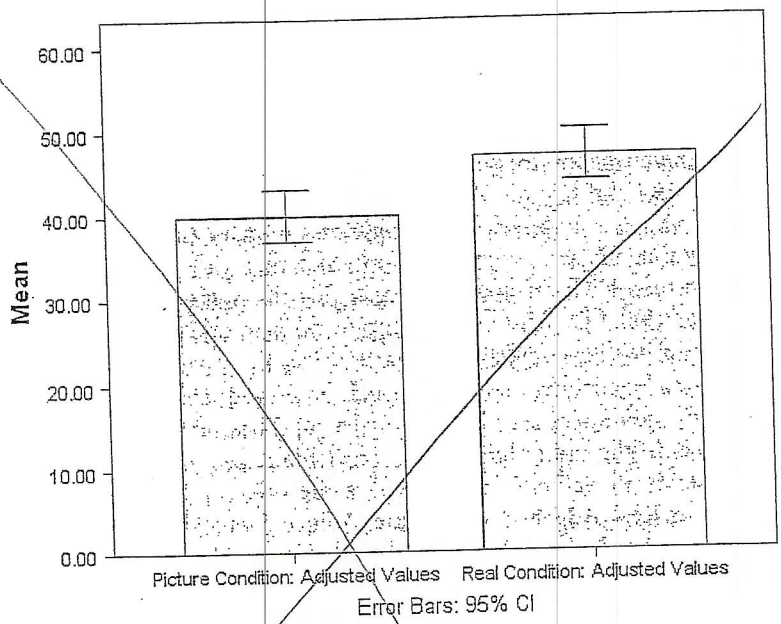


FIGURE 9.7
Error bar graph of the adjusted values of **spiderRM.sav**

The resulting error bar graph is shown in Figure 9.7. Compare this graph to the graphs in Figure 9.2 – what differences do you see? The first thing to notice is that the means in the two conditions have not changed. However, the error bars have changed: they have got smaller. Also, whereas in Figure 9.2 the error bars overlap, in this new graph they do not. In Chapter 2 we discovered that when error bars do not overlap we can be fairly confident that our samples have not come from the same population (and so our experimental manipulation has been successful). Therefore, when we plot the proper error bars for the repeated-measures data it shows the extra sensitivity that this design has: the differences between conditions appear to be significant, whereas when different participants are used, there does not appear to be a significant difference. (Remember that the means in both situations are identical, but the sampling error is smaller in the repeated-measures design.) I expand upon this point in section 9.6.

9.3. The *t*-test ①

We have seen in previous chapters that the *t*-test is a very versatile statistic: it can be used to test whether a correlation coefficient is different from 0; it can also be used to test whether a regression coefficient, *b*, is different from 0. However, it can also be used to test whether two group means are different. It is to this use that we now turn.

The simplest form of experiment that can be done is one with only one independent variable that is manipulated in only two ways and only one outcome is measured. More often than not the manipulation of the independent variable involves having an experimental condition and a control group (see Field & Hole, 2003). Some examples of this kind of design are:

- Is the movie *Scream* 2 scarier than the original *Scream*? We could measure heart rates (which indicate anxiety) during both films and compare them.
- Does listening to music while you work improve your work? You could get some people to write an essay (or book!) listening to their favourite music, and then write a different essay when working in silence (this is a control group). You could then compare the essay grades!

I mentioned in section 2.6.1 that most test statistics can be thought of as the 'variance explained by the model' divided by the 'variance that the model can't explain'. In other words, effect/error. When comparing two means the 'model' that we fit to the data (the effect) is the difference between the two group means. We saw also in Chapter 2 that means vary from sample to sample (sampling variation) and that we can use the standard error as a measure of how much means fluctuate (in other words, the error in the estimate of the mean). Therefore, we can also use the standard error of the differences between the two means as an estimate of the error in our model (or the error in the difference between means). Therefore, we calculate the *t*-test using equation (9.1) below. The top half of the equation is the 'model' (our model being the difference between means is bigger than the expected difference, which in most cases will be 0 – we expect the difference between means to be different to zero). The bottom half is the 'error'. So, just as I said in Chapter 2, we're basically getting the test statistic by dividing the model (or effect) by the error in the model. The exact form that this equation takes depends on whether the same or different participants were used in each experimental condition:

$$t = \frac{\text{observed difference between sample means} - \text{expected difference between population means (if null hypothesis is true)}}{\text{estimate of the standard error of the difference between two sample means}} \quad (9.1)$$

9.3.2 Assumptions of the *t*-test ①

Both the independent *t*-test and the dependent *t*-test are *parametric tests* based on the normal distribution (see Chapter 5). Therefore, they assume:

- The sampling distribution is normally distributed. In the dependent *t*-test this means that the sampling distribution of the *differences* between scores should be normal, not the scores themselves (see section 9.4.3).
- Data are measured at least at the interval level.

The independent *t*-test, because it is used to test different groups of people, also assumes:

- Variances in these populations are roughly equal (*homogeneity of variance*).
- Scores are independent (because they come from different people).

These assumptions were explained in detail in Chapter 5 and, in that chapter, I emphasized the need to check these assumptions before you reach the point of carrying out your statistical test. As such, I won't go into them again, but it does mean that if you have ignored my advice and haven't checked these assumptions then you need to do it now! SPSS also incorporates some procedures into the *t*-test (e.g. Levene's test, see section 5.6.1, can be done at the same time as the *t*-test). Let's now look at each of the two *t*-tests in turn.

9.4. The dependent *t*-test ①

If we stay with our repeated-measures data for the time being we can look at the dependent *t*-test, or paired-samples *t*-test. The dependent *t*-test is easy to calculate. In effect, we use a numeric version of equation (9.1):

$$t = \frac{\overline{D} - \mu_D}{s_D/\sqrt{N}} \quad (9.2)$$

Equation (9.2) compares the mean difference between our samples ($\overline{D}$) to the difference that we would expect to find between population means (μ_D), and then takes into account the standard error of the differences ($s_D/\sqrt{N}$). If the null hypothesis is true, then we expect there to be no difference between the population means (hence $\mu_D = 0$).

9.4.1 Sampling distributions and the standard error ①

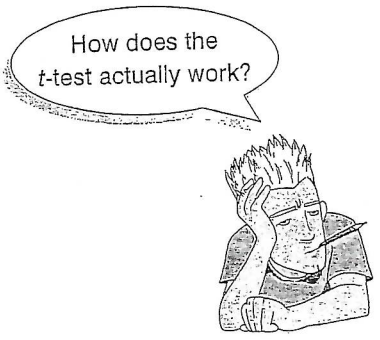
In equation (9.1) I referred to the lower half of the equation as the standard error of differences. The standard error was introduced in section 2.5.1 and is simply the standard deviation of the sampling distribution. Have a look back at this section now to refresh your memory about sampling distributions and the standard error. Sampling distributions have several properties that are important. For one thing, if the population is normally distributed then so is the sampling distribution; in fact, if the samples contain more than about 50 scores the sampling distribution should be normally distributed. The mean of the sampling distribution is equal to the mean of the population, so the average of all possible sample means should be the same as the population mean. This property makes sense because if a sample is representative of the population then you would expect its mean to be equal to that of the population. However, sometimes samples are unrepresentative and their means differ from the population mean. On average, though, a sample mean will be very close to the population mean and only rarely will the sample mean be substantially different from that of the population. A final property of a sampling distribution is that its standard deviation is equal to the standard deviation of the population divided by the square root of the number of observations in the sample. As I mentioned before, this standard deviation is known as the standard error.

We can extend this idea to look at the *differences* between sample means. If you were to take several pairs of samples from a population and calculate their means, then you could also calculate the difference between their means. I mentioned earlier that *on average* sample means will be very similar to the population mean: as such, most samples will have very similar means. Therefore, most of the time the difference between sample means from the same population will be zero, or close to zero. However, sometimes one or both of the samples could have a mean very deviant from the population mean and so it is possible to obtain large differences between sample means by chance alone. However, this would happen less frequently.

In fact, if you plotted these differences between sample means as a histogram, you would again have a sampling distribution with all of the properties previously described. The standard deviation of this sampling distribution is called the **standard error of differences**. A small standard error tells us that most pairs of samples from a population will have very similar means (i.e. the difference between sample means should normally be very small). A large standard error tells us that sample means can deviate quite a lot from the population mean and so differences between pairs of samples can be quite large by chance alone.

9.4.2 The dependent *t*-test equation explained ①

In an experiment, a person's score in condition 1 will be different to their score in condition 2, and this difference could be very large or very small. If we calculate the differences between each person's score in each condition and add up these differences we would get the total amount of difference. If we then divide this total by the number of participants we get the average difference (thus how much, on average, a person's score differed in condition 1 compared to condition 2). This average difference is $\overline{D}$ in equation (9.2) and it is an indicator



of the systematic variation in the data (i.e. it represents the experimental effect). We need to compare this systematic variation against some kind of measure of the 'systematic variation that we could naturally expect to find'. In Chapter 2 we saw that the standard deviation was a measure of the 'fit' of the mean to the observed data (i.e. it measures the error in the model when the model is the mean), but it does not measure the fit of the mean to the population. To do this we need the standard error (see the previous section, where we revised this idea).

The standard error is a measure of the error in the mean as a model of the population. In this context, we know that if we had taken two random samples from a population (and not done anything to these samples) then the means could be different just by chance. The standard error tells us by how much these samples could differ. A small standard error means that sample means should be quite similar, so a big difference between two sample means is unlikely. In contrast, a large standard error tells us that big differences between the means of two random samples are more likely. Therefore it makes sense to compare the average difference between means against the standard error of these differences. This gives us a test statistic that, as I've said numerous times in previous chapters, represents model/error. Our model is the average difference between condition means, and we divide by the standard error which represents the error associated with this model (i.e. how similar two random samples are likely to be from this population).

To clarify, imagine that an alien came down and cloned me millions of times. This population is known as Landy of the Andys (this would be possibly the most dreary and strangely terrifying place I could imagine). Imagine the aliens were interested in spider phobia in this population (because I am petrified of spiders). Everyone in this population (my clones) will be the same as me, and would behave in an identical way to me. If you took two samples from this population and measured their spider fear, then the means of these samples would be the same (we are clones), so the difference between sample means would be zero. Also, because we are all identical, then all samples from the population will be perfect reflections of the population (the standard error would be zero also). Therefore, if we were to get two samples that differed even very slightly then this would be very unlikely indeed (because our population is full of cloned Andys). Therefore, a difference between samples must mean that they have come from different populations. Of course, in reality we don't have samples that perfectly reflect the population, but the standard error gives an idea of how well samples reflect the population from which they came.

Therefore, by dividing by the standard error we are doing two things: (1) standardizing the average difference between conditions (this just means that we can compare values of t without having to worry about the scale of measurement used to measure the outcome variable); and (2) contrasting the difference between means that we have against the difference that we could *expect* to get based on how well the samples represent the populations from which they came. If the standard error is large, then large differences between samples are more common (because the distribution of differences is more spread out). Conversely, if the standard error is small, then large differences between sample means are uncommon (because the distribution is very narrow and centred around zero). Therefore, if the average difference between our samples is large, and the standard error of differences is small, then we can be confident that the difference we observed in our sample is not a chance result. If the difference is not a chance result then it must have been caused by the experimental manipulation.

In a perfect world, we could calculate the standard error by taking all possible pairs of samples from a population, calculating the differences between their means, and then working out the standard deviation of these differences. However, in reality this is impossible. Therefore, we estimate the standard error from the standard deviation of differences obtained within the sample (s_D) and the sample size (N). Think back to section 2.5.1 where we saw that the standard error is simply the standard deviation divided by the square root of the sample size; likewise the standard error of differences ($\sigma_{\bar{D}}$) is simply the standard deviation of differences divided by the square root of the sample size:

$$(\sigma_{\bar{D}}) = \frac{s_D}{\sqrt{N}}$$

If the standard error of differences is a measure of the unsystematic variation within the data, and the sum of difference scores represents the systematic variation, then it should be clear that the *t*-statistic is simply the ratio of the systematic variation in the experiment to the unsystematic variation. If the experimental manipulation creates any kind of effect, then we would expect the systematic variation to be much greater than the unsystematic variation (so at the very least, *t* should be greater than 1). If the experimental manipulation is unsuccessful then we might expect the variation caused by individual differences to be much greater than that caused by the experiment (so *t* will be less than 1). We can compare the obtained value of *t* against the maximum value we would expect to get by chance alone in a *t*-distribution with the same degrees of freedom (these values can be found in the Appendix); if the value we obtain exceeds this critical value we can be confident that this reflects an effect of our independent variable.

9.4.3 The dependent *t*-test and the assumption of normality ①

We talked about the assumption of normality in Chapter 5 and discovered that parametric tests (like the dependent *t*-test) assume that the sampling distribution is normal. This should be true in large samples, but in small samples people often check the normality of their data because if the data themselves are normal then the sampling distribution is likley to be also. With the dependent *t*-test we analyse the *differences* between scores because we're interested in the sampling distribution of these differences (not the raw data). Therefore, if you want to test for normality before a dependent *t*-test then what you should do is compute the differences between scores, and then check if this new variable is normally distributed (or use a big sample and not worry about normality!). It is possible to have two measures that are highly non-normal that produce beautifully distributed differences!

SELF-TEST Using the spiderRM.sav data, compute the differences between the picture and real condition and check the assumption of normality for these differences.

9.4.4 Dependent *t*-tests using SPSS ①

Using our spider data (spiderRM.sav), we have 12 spider-phobes who were exposed to a picture of a spider (picture) and on a separate occasion a real live tarantula (real). Their anxiety was measured in each condition (half of the participants were exposed to the picture before the real spider while the other half were exposed to the real spider first). I have already described how the data are arranged, and so we can move straight on to doing the test itself. First, we need to access the main dialog box by selecting Analyze ▸ Compare Means ▸ Paired-Samples T Test... (Figure 9.8). Once the dialog box is activated, you need to select pairs of variables to be analysed. In this case we have only one pair (Real vs. Picture). To select a pair you should click on the first variable that you want to select (in this case Picture), then hold down the *Ctrl* key on the keyboard and select the second (in this case Real). To transfer these

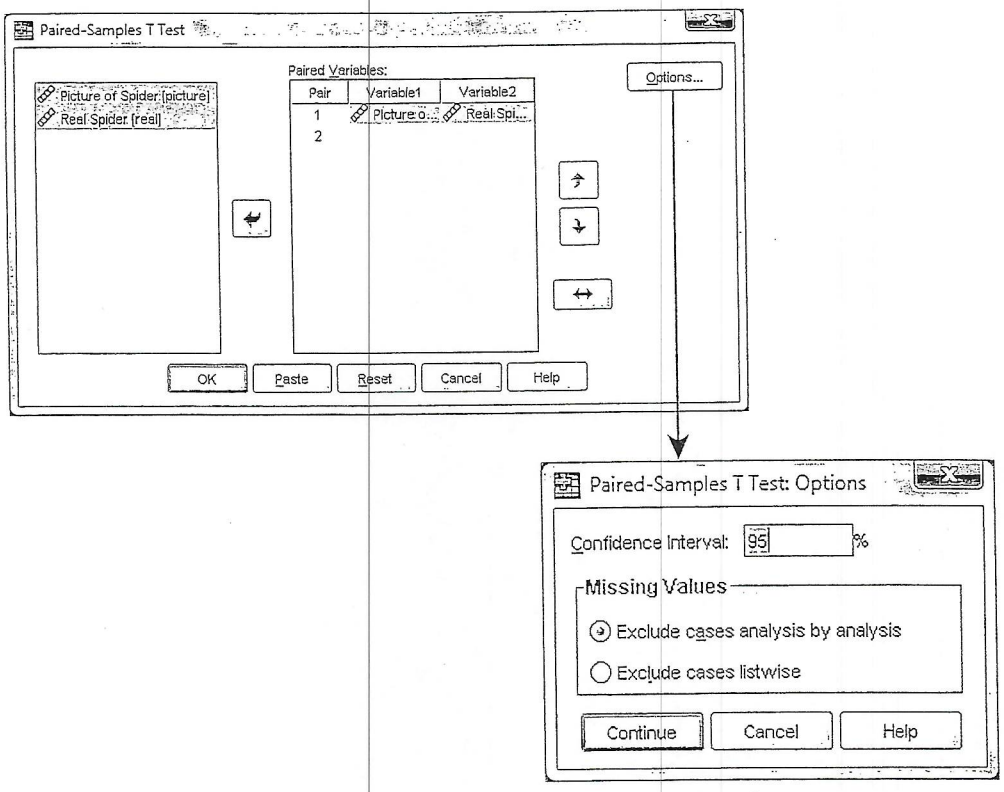


FIGURE 9.8
Main dialog box for paired-samples *t*-test

two variables to the box labelled *Paired Variables* click on . (You can also select each variable individually and transfer it by clicking on , but the method using the *Ctrl* key to select both variables is quicker.) If you want to carry out several *t*-tests then you can select another pair of variables, transfer them to the variables list, then select another pair and so on. In this case, we want only one test. If you click on Options... then another dialog box appears that gives you the chance to change the width of the confidence interval that is calculated. The default setting is for a 95% confidence interval and this is fine; however, if you want to be stricter about your analysis you could choose a 99% confidence interval but you run a higher risk of failing to detect a genuine effect (a Type II error). You can also select how to deal with missing values (see SPSS Tip 6.1). To run the analysis click on OK.

9.4.5 Output from the dependent *t*-test ①

The resulting output produces three tables. SPSS Output 9.1 shows a table of summary statistics for the two experimental conditions. For each condition we are told the mean, the number of participants (N) and the standard deviation of the sample. In the final column we are told the standard error (see section 9.4.1), which is the sample standard deviation divided by the square root of the sample size ($SE = s/\sqrt{N}$), so for the picture condition $SE = 9.2932/\sqrt{12} = 9.2932/3.4641 = 2.68$.

SPSS Output 9.1 also shows the Pearson correlation between the two conditions. When repeated measures are used it is possible that the experimental conditions will correlate (because the data in each condition come from the same people and so there could be some constancy in their responses). SPSS provides the value of Pearson's r and the two-tailed significance value (see Chapter 6). For these data the experimental conditions yield a fairly large correlation coefficient ($r = .545$) but are not significantly correlated because $p > .05$.

SPSS OUTPUT 9.1

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Picture of Spider	40.00	12	9.293	2.683
	Real Spider	47.00	12	11.029	3.184

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Picture of Spider & Real Spider	12	.545	.067

SPSS Output 9.2 shows the most important of the tables: the one that tells us whether the difference between the means of the two conditions was large enough to *not* be a chance result. First, the table tells us the mean difference between scores (this value, i.e. $\bar{D}$ in equation (9.2), is the difference between the mean scores of each condition: $40 - 47 = -7$). The table also reports the standard deviation of the differences between the means and more important the standard error of the differences between participants' scores in each condition (see section 9.4.1). The test statistic, t, is calculated by dividing the mean of differences by the standard error of differences (see equation (9.2): $t = -7/2.8311 = -2.47$). The size of t is compared against known values based on the degrees of freedom. When the same participants have been used, the degrees of freedom are simply the sample size minus 1 ($df = N - 1 = 11$). SPSS uses the degrees of freedom to calculate the exact probability that a value of t as big as the one obtained could occur if the null hypothesis were true (i.e. there was no difference between these means). This probability value is in the column labelled *Sig.* By default, SPSS provides only the two-tailed probability, which is the probability when no prediction was made about the direction of group differences. If a specific prediction was made (e.g. we might predict that anxiety will be higher when a real spider is used) then the one-tailed probability should be reported and this value is obtained by dividing the two-tailed probability by 2 (see SPSS Tip 9.1). The two-tailed probability for the spider data is very low ($p = .031$) and in fact it tells us that there is only a 3.1% chance that a value of t this big could happen if the null hypothesis were true. We saw in Chapter 2 that we generally accept a $p < .05$ as statistically meaningful; therefore, this t is significant because .031 is smaller than .05. The fact that the t-value is a negative number tells us that the first condition (the picture condition) had a smaller mean than the second (the real condition) and so the real spider led to greater anxiety than the picture. Therefore, we can conclude that exposure to a real spider caused significantly more reported anxiety in spider-phobes than exposure to a picture, $t(11) = -2.47$, $p < .05$. This result was predicted by the error bar chart in Figure 9.7.

SPSS OUTPUT 9.2

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	Picture of Spider - Real Spider	-7.000	9.807	2.831	-13.231	-.769	-2.473	11	.031

Finally, this output provides a 95% confidence interval for the mean difference. Imagine we took 100 samples from a population of difference scores and calculated their means ($\bar{D}$) and a confidence interval for that mean. In 95 of those samples the constructed confidence intervals contains the true value of the mean difference. The confidence interval tells us the boundaries within which the true mean difference is likely to lie.[4] So, assuming this sample's confidence interval is one of the 95 out of 100 that contains the population value,

[4] We saw in section 2.5.2 that these intervals represent the value of two (well, 1.96 to be precise) standard errors either side of the mean of the sampling distribution. For these data, in which the mean difference was –7 and the standard error was 2.8311, these limits will be $-7 \pm (1.96 \times 2.8311)$. However, because we're using the t-distribution, not the normal distribution, we use the critical value of t to compute the confidence intervals. This value is (with $df = 11$ as in this example) 2.201 (two-tailed), which gives us $-7 \pm (2.201 \times 2.8311)$.

SPSS TIP 9.1 One- and two-tailed significance in SPSS ①

Some students get a bit upset by the fact that SPSS produces only the two-tailed significance much of the time and are confused by why there isn't an option that can be selected to produce the one-tailed significance. The answer is simple: there is no need for an option because the one-tailed probability can be ascertained by dividing the two-tailed significance value by 2. For example, if the two-tailed probability is .107, then the one-tailed probability is .107/2 = .054.

we can say that the true mean difference lies between −13.23 and −0.77. The importance of this interval is that it does not contain zero (i.e. both limits are negative) because this tells us that the true value of the mean difference is unlikely to be zero. Crucially, if we were to compare pairs of random samples from a population we would expect most of the differences between sample means to be zero. This interval tells us that, based on our two samples, the true value of the difference between means is unlikely to be zero. Therefore, we can be confident that our two samples do not represent random samples from the same population. Instead they represent samples from different populations induced by the experimental manipulation.

9.4.6 Calculating the effect size ②

Even though our t-statistic is statistically significant, this doesn't mean our effect is important in practical terms. To discover whether the effect is substantive we need to use what we know about effect sizes (see section 2.6.4). I'm going to stick with the effect size r because it's widely understood, frequently used, and yes, I'll admit it, I actually like it! Converting a t-value into an r-value is actually really easy; we can use the following equation (e.g. Rosenthal, 1991; Rosnow & Rosenthal, 2005).[5]

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

We know the value of t and the df from the SPSS output and so we can compute r as follows:

$$r = \sqrt{\frac{-2.473^2}{-2.473^2 + 11}} = \sqrt{\frac{6.116}{17.116}} = .60$$

If you think back to our benchmarks for effect sizes this represents a very large effect (it is above .5, the threshold for a large effect). Therefore, as well as being statistically significant, this effect is large and so represents a substantive finding.

[5] Actually, this will overestimate the effect size because of the correlation between the two conditions. This is quite a technical issue and I'm trying to keep things simple here, but bear this in mind and if you're interested read Dunlap, Cortina, Vaslow, and Burke (1996).

9.4.7. Reporting the dependent *t*-test ①

There is a fairly standard way to report any test statistic: you usually state the finding to which the test relates and then report the test statistic, its degrees of freedom and the probability value of that test statistic. There has also been a recent move (by the American Psychological Association among others) to recommend that an estimate of the effect size is routinely reported. Although effect sizes are still rather sporadically used, I want to get you into good habits so we'll start thinking about effect sizes now. In this example the SPSS output tells us that the value of t was –2.47, that the degrees of freedom on which this was based was 11, and that it was significant at $p = .031$. We can also see the means for each group. We could write this as:

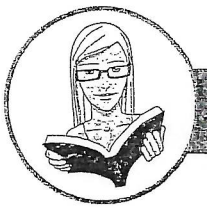
✓ On average, participants experienced significantly greater anxiety to real spiders ($M = 47.00$, $SE = 3.18$) than to pictures of spiders ($M = 40.00$, $SE = 2.68$), $t(11) = -2.47$, $p < .05$, $r = .60$.

Note how we've reported the means in each group (and standard errors) in the standard format. For the test statistic, note that we've used an italic t to denote the fact that we've calculated a t-statistic, then in brackets we've put the degrees of freedom and then stated the value of the test statistic. The probability can be expressed in several ways: often people report things to a standard level of significance (such as .05) as I have done here, but sometimes people will report the exact significance. Finally, note that I've reported the effect size at the end – you won't always see this in published papers but that's no excuse for you not to report it!

Try to avoid writing vague, unsubstantiated things like this:

✗ People were more scared of real spiders ($t = -2.47$).

More scared than what? Where are the *df*? Was the result statistically significant? Was the effect important (what was the effect size)?



CRAMMING SAM'S TIPS

- The dependent *t*-test compares two means, when those means have come from the same entities; for example, if you have used the same participants in each of two experimental conditions.
- Look at the column labelled *Sig.* If the value is less than .05 then the means of the two conditions are significantly different.
- Look at the values of the means to tell you how the conditions differ.
- SPSS provides only the two-tailed significance value; if you want the one-tailed significance just divide the value by 2.
- Report the *t*-statistic, the degrees of freedom and the significance value. Also report the means and their corresponding standard errors (or draw an error bar chart).
- If you're feeling brave, calculate and report the effect size too!

9.5. The independent *t*-test ①

9.5.1 The independent *t*-test equation explained ①

The independent *t*-test is used in situations in which there are two experimental conditions and different participants have been used in each condition. There are two different equations that can be used to calculate the *t*-statistic depending on whether the samples contain an equal number of people. As with the dependent *t*-test we can calculate the *t*-statistic by using a numerical version of equation (9.1); in other words, we are comparing the model or effect against the error. With the dependent *t*-test we could look at differences between pairs of scores, because the scores came from the same participants and so individual differences between conditions were eliminated. Hence, the difference in scores should reflect only the effect of the experimental manipulation. Now, when different participants participate in different conditions then pairs of scores will differ not just because of the experimental manipulation, but also because of other sources of variance (such as individual differences between participants' motivation, IQ, etc.). If we cannot investigate differences between conditions on a *per participant* basis (by comparing pairs of scores as we did for the dependent *t*-test) then we must make comparisons on a *per condition* basis (by looking at the overall effect in a condition – see equation (9.3)):

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\text{estimate of the standard error}} \tag{9.3}$$

Instead of looking at differences between pairs of scores, we now look at differences between the overall means of the two samples and compare them to the differences we would expect to get between the means of the two populations from which the samples come. If the null hypothesis is true then the samples have been drawn from the same population. Therefore, under the null hypothesis $\mu_1 = \mu_2$ and therefore $\mu_1 - \mu_2 = 0$. Therefore, under the null hypothesis the equation becomes:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{estimate of the standard error}} \tag{9.4}$$

In the dependent *t*-test we divided the mean difference between pairs of scores by the standard error of these differences. For the independent *t*-test we are looking at differences between groups and so we need to divide by the standard deviation of differences between groups. We can still apply the logic of sampling distributions to this situation. Now, imagine we took several pairs of samples – each pair containing one sample from the two different populations – and compared the means of these samples. From what we have learnt about sampling distributions, we know that the majority of samples from a population will have fairly similar means. Therefore, if we took several pairs of samples (from different populations), the differences between the sample means will be similar across pairs. However, often the difference between a pair of sample means will deviate by a small amount and very occasionally it will deviate by a large amount. If we could plot a sampling distribution of the differences between every pair of sample means that could be taken from two populations, then we would find that it had a normal distribution with a mean equal to the difference between population means ($\mu_1 - \mu_2$). The sampling distribution would tell us by how much we can expect the means of two (or more) samples to differ. As before, the standard deviation of the sampling distribution (the standard error) tells us how variable the differences between

sample means are by chance alone. If the standard deviation is high then large differences between sample means can occur by chance; if it is small then only small differences between sample means are expected. It, therefore, makes sense that we use the standard error of the sampling distribution to assess whether the difference between two sample means is statistically meaningful or simply a chance result. Specifically, we divide the difference between sample means by the standard deviation of the sampling distribution.

So, how do we obtain the standard deviation of the sampling distribution of differences between sample means? Well, we use the variance sum law, which states that the variance of a difference between two independent variables is equal to the sum of their variances (see, for example, Howell, 2006). This statement means that the variance of the sampling distribution is equal to the sum of the variances of the two populations from which the samples were taken. We saw earlier that the standard error is the standard deviation of the sampling distribution of a population. We can use the sample standard deviations to calculate the standard error of each population's sampling distribution:

$$\text{SE of sampling distribution of population 1} = \frac{s_1}{\sqrt{N_1}}$$

$$\text{SE of sampling distribution of population 2} = \frac{s_2}{\sqrt{N_2}}$$

Therefore, remembering that the variance is simply the standard deviation squared, we can calculate the variance of each sampling distribution:

$$\text{variance of sampling distribution of population 1} = \left(\frac{s_1}{\sqrt{N_1}}\right)^2 = \frac{s_1^2}{N_1}$$

$$\text{variance of sampling distribution of population 2} = \left(\frac{s_2}{\sqrt{N_2}}\right)^2 = \frac{s_2^2}{N_2}$$

The variance sum law means that to find the variance of the sampling distribution of differences we merely add together the variances of the sampling distributions of the two populations:

$$\text{variance of sampling distribution of differences} = \frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}$$

To find out the standard error of the sampling distribution of differences we merely take the square root of the variance (because variance is the standard deviation squared):

$$\text{SE of the sampling distribution of differences} = \sqrt{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)}$$

Therefore, equation (9.4) becomes:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)}} \tag{9.5}$$

Equation (9.5) is true only when the sample sizes are equal. Often in the social sciences it is not possible to collect samples of equal size (because, for example, people may not complete an experiment). When we want to compare two groups that contain different numbers of participants then equation (9.5) is not appropriate. Instead the pooled variance estimate *t*-test

is used which takes account of the difference in sample size by *weighting* the variance of each sample. We saw in Chapter 1 that large samples are better than small ones because they more closely approximate the population; therefore, we weight the variance by the size of sample on which it's based (we actually weight by the number of degrees of freedom, which is the sample size minus 1). Therefore, the pooled variance estimate is:

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

This is simply a weighted average in which each variance is multiplied (weighted) by its degrees of freedom, and then we divide by the sum of weights (or sum of the two degrees of freedom). The resulting weighted average variance is then just replaced in the *t*-test equation:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

As with the dependent *t*-test we can compare the obtained value of t against the maximum value we would expect to get by chance alone in a *t*-distribution with the same degrees of freedom (these values can be found in the Appendix); if the value we obtain exceeds this critical value we can be confident that this reflects an effect of our independent variable. One thing that should be apparent from the equation for t is that to compute it you don't actually need any data! All you need are the means, standard deviations and sample sizes (see SPSS Tip 9.2).

The derivation of the *t*-statistic is merely to provide a conceptual grasp of what we are doing when we carry out a *t*-test on SPSS. Therefore, if you don't know what on earth I'm babbling on about then don't worry about it (just spare a thought for my cat: he has to listen to this rubbish all the time!) because SPSS knows how to do it and that's all that matters!

SPSS TIP 9.2 Computing *t* from means, SDs and Ns ③

Using syntax, you can compute a *t*-test in SPSS from only the two group means, the two group standard deviations and the two group sizes. Open a data editor window and set up six new variables: **x1** (mean of group 1), **x2** (mean of group 2), **sd1** (standard deviation of group 1), **sd2** (standard deviation of group 2), **n1** (sample size of group 1) and **n2** (sample size of group 2). Type the values of each of these in the first row of the data editor. Open a syntax window and type the following:

```
COMPUTE df = n1+n2-2.
COMPUTE poolvar = (((n1-1)*(sd1 ** 2))+((n2-1)*(sd2 ** 2)))/df.
COMPUTE t = (x1-x2)/sqrt(poolvar*((1/n1)+(1/n2))).
COMPUTE sig = 2*(1-(CDF.T(abs(t),df))) .
Variable labels sig 'Significance (2-tailed)'.
EXECUTE .
```

The first line computes the degrees of freedom, the second computes the pooled variance, s_p^2, the third computes *t* and the fourth its two-tailed significance. All of these values will be created in a new column in the data

editor. The line beginning 'Variable labels' simply labels the significance variable so that we know that it is two-tailed. If you want to display the results in the SPSS Viewer you could type:

```
SUMMARIZE
  /TABLES= x1 x2 df t sig
  /FORMAT=VALIDLIST NOCASENUM TOTAL LIMIT=100
  /TITLE='T-test'
  /MISSING=VARIABLE
  /CELLS=NONE.
```

These commands will produce a table of the variables **x1**, **x2**, **df**, **t** and **sig** so you'll see the means of the two groups, the degrees of freedom, the value of *t* and its two-tailed significance.

You can run lots of *t*-tests at the same time by putting different values for the means, SDs and sample sizes in different rows. If you do this, though, I suggest having a string variable called **Outcome** in the file in which you type what was being measured (or some other information so that you can identify to what the *t*-test relates).

I have put these commands in a syntax file called **Independent t from means.sps**. My file is actually a bit more complicated because it calculates an effect size measure (Cohen's *d*). For an example of how to use this file see Labcoat Leni's Real Research 9.1.

LABCOAT LENI'S REAL RESEARCH 9.1

You don't have to be mad here, but it helps ☺

BOARD, B. J., & FRITZON, K. (2005). *PSYCHOLOGY, CRIME & LAW, 11*, 17–32.

In the UK you often see the 'humorous' slogan 'You don't have to be mad to work here, but it helps' stuck up in work places. Well, Board and Fritzon (2005) took this a step further by measuring whether 39 senior business managers and chief executives from leading UK companies were mad (well, had personality disorders, PDs). They gave them The Minnesota Multiphasic Personality Inventory Scales for DSM III Personality Disorders (MMPI-PD), which is a well-validated measure of 11 personality disorders: Histrionic, Narcissistic, Antisocial, Borderline, Dependent, Compulsive, Passive–aggressive, Paranoid, Schizotypal, Schizoid and Avoidant. They needed a comparison group, and what better one to choose than 317 legally classified psychopaths at Broadmoor Hospital (a famous high-security psychiatric hospital in the UK).

The authors report the means and SDs for these two groups in Table 2 of their paper. Using these values and the syntax file **Independent t from means.sps** we can run *t*-tests on these means. The data from Board and Fritzon's Table 2 are in the file **Board and Fritzon 2005.sav**. Use this file and the syntax file to run *t*-tests to see whether managers score higher on personality disorder questionnaires than legally classified psychopaths. Report these results. What do you conclude?

Answers are in the additional material on the companion website (or look at Table 2 in the original article).

9.5.2. The independent *t*-test using SPSS ①

I have probably bored most of you to the point of wanting to eat your own legs by now. Equations are boring and that is why SPSS was invented to help us minimize our contact with them. Using our spider data again (spiderBG.sav), we have 12 spider-phobes who

were exposed to a picture of a spider and 12 different spider-phobes who were exposed to a real-life tarantula (the groups are coded using the variable **group**). Their anxiety was measured in each condition (**anxiety**). I have already described how the data are arranged (see section 9.2), so we can move straight on to doing the test itself. First, we need to access the main dialog box by selecting Analyze Compare Means Independent-Samples T Test... (see Figure 9.9). Once the dialog box is activated, select the dependent variable from the list (click on **anxiety**) and transfer it to the box labelled *Test Variable(s)* by clicking on ▸. If you want to carry out *t*-tests on several dependent variables then you can select other dependent variables and transfer them to the variables list. However, there are good reasons why it is not a good idea to carry out lots of tests (see Chapter 10).

Next, we need to select an independent variable (the grouping variable). In this case, we need to select **group** and then transfer it to the box labelled *Grouping Variable*. When your grouping variable has been selected the Define Groups... button will become active and you should click on it to activate the *Define Groups* dialog box. SPSS needs to know what numeric codes you assigned to your two groups, and there is a space for you to type the codes. In this example, we coded our picture group as 0 and our real group as 1, and so these are the codes that we type. Alternatively you can specify a *Cut point* in which case SPSS will assign all cases greater than or equal to that value to one group and all the values below the cut point to the second group. This facility is useful if you are testing different groups of participants based on something like a median split (see Jane Superbrain Box 9.1) – you would simply type the median value in the box labelled *Cut point*. When you have defined the groups, click on Continue to return to the main dialog box. If you click on Options... then another dialog box appears that gives you the same options as for the dependent *t*-test. To run the analysis click on OK.

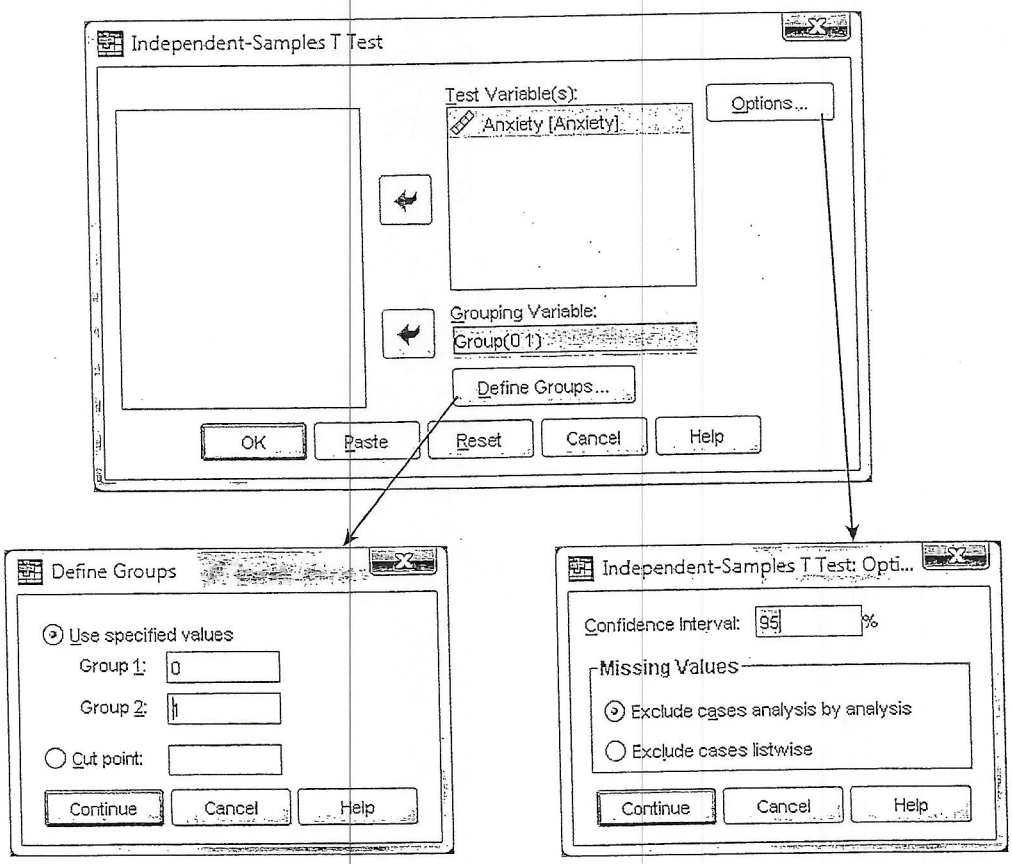


FIGURE 9.9 Dialog boxes for the independent-samples *t*-test

JANE SUPERBRAIN 9.1

Are median splits the devil's work? ②

Often in research papers you see that people have analysed their data using a 'median split'. In our spider phobia example, this means that you measure scores on a spider phobia questionnaire and calculate the median. You then classify anyone with a score above the median as a 'phobic', and those below the median as 'non-phobic'. In doing this you 'dichotomize' a continuous variable. This practice is quite common, but is it sensible?

MacCallum, Zhang, Preacher, and Rucker (2002) wrote a splendid paper pointing out various problems on turning a perfectly decent continuous variable into a categorical variable:

1 Imagine there are four people: Peter, Birgit, Jip and Kiki. We measure how scared of spiders they are as a percentage and get Jip (100%), Kiki (60%), Peter (40%) and Birgit (0%). If we split these four people at the median (50%) then we're saying that Jip and Kiki are the same (they get a score of 1 = phobic) and Peter and Birgit are the same (they both get a score of 0 = not phobic). In reality, Kiki and Peter are the most similar of the four people, but they have been put in different groups. So, median splits change the original information quite dramatically (Peter and Kiki are originally very similar but become very different after the split, Jip and Kiki are relatively dissimilar originally but become identical after the split).

2 Effect sizes get smaller: if you correlate two continuous variables then the effect size will be larger than if you correlate the same variables after one of them has been dichotomized. Effect sizes also get smaller in ANOVA and regression.

3 There is an increased chance of finding spurious effects.

So, if your supervisor has just told you to do a median split, have a good think about whether it is the right thing to do (and read MacCallum et al.'s paper). One of the rare situations in which dichotomizing a continuous variable is justified, according to MacCallum et al., is when there is a clear theoretical rationale for distinct categories of people based on a meaningful break point (i.e. not the median); for example, phobic versus not phobic based on diagnosis by a trained clinician would be a legitimate dichotomization of anxiety.

9.5.3. Output from the independent *t*-test ①

The output from the independent *t*-test contains only two tables. The first table (SPSS Output 9.3) provides summary statistics for the two experimental conditions. From this table, we can see that both groups had 12 participants (column labelled *N*). The group who saw the picture of the spider had a mean anxiety of 40, with a standard deviation of 9.29. What's more, the standard error of that group (the standard deviation of the sampling distribution) is 2.68 ($SE = 9.293/\sqrt{12} = 9.293/3.464 = 2.68$). In addition, the table tells us that the average anxiety level in participants who were shown a real spider was 47, with a standard deviation of 11.03 and a standard error of 3.18 ($SE = 11.029/\sqrt{12} = 11.029/3.464 = 3.18$).

SPSS OUTPUT 9.3

Group Statistics

	Spider or Picture?	N	Mean	Std. Deviation	Std. Error Mean
Anxiety	Picture	12	40.00	9.293	2.683
	Real Spider	12	47.00	11.029	3.184

The second table of output (SPSS Output 9.4) contains the main test statistics. The first thing to notice is that there are two rows containing values for the test statistics: one

row is labelled *Equal variances assumed*, while the other is labelled *Equal variances not assumed*. In Chapter 5, we saw that parametric tests assume that the variances in experimental groups are roughly equal. Well, in reality there are adjustments that can be made in situations in which the variances are not equal. The rows of the table relate to whether or not this assumption has been broken. How do we know whether this assumption has been broken?

We saw in section 5.6.1 that we can use Levene's test to see whether variances are different in different groups, and SPSS produces this test for us. Remember that Levene's test is similar to a *t*-test in that it tests the hypothesis that the variances in the two groups are equal (i.e. the difference between the variances is zero). Therefore, if Levene's test is significant at $p \leq .05$, we can gain confidence in the hypothesis that the variances are significantly different and that the assumption of homogeneity of variances has been violated. If, however, Levene's test is non-significant (i.e. $p > .05$) then we do not have sufficient evidence to reject the null hypothesis that the difference between the variances is zero – in other words, we can assume that the variances are roughly equal and the assumption is tenable. For these data, Levene's test is non-significant (because $p = .386$, which is greater than .05) and so we should read the test statistics in the row labelled *Equal variances assumed*. Had Levene's test been significant, then we would have read the test statistics from the row labelled *Equal variances not assumed*.

SPSS OUTPUT 9.4

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
Anxiety	Equal variances assumed	.782	.386	-1.681	22	.107	-7.000	4.163	-15.634	1.634
	Equal variances not assumed			-1.681	21.385	.107	-7.000	4.163	-15.649	1.649

Having established that the assumption of homogeneity of variances is met, we can move on to look at the *t*-test itself. We are told the mean difference $(\overline{X}_1 - \overline{X}_2 = 40 - 47 = -7)$ and the standard error of the sampling distribution of differences, which is calculated using the lower half of equation (9.5):

$$\sqrt{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)} = \sqrt{\left(\frac{9.29^2}{12} + \frac{11.03^2}{12}\right)}$$
$$= \sqrt{(7.19 + 10.14)}$$
$$= \sqrt{17.33}$$
$$= 4.16$$

The *t*-statistic is calculated by dividing the mean difference by the standard error of the sampling distribution of differences ($t = -7/4.16 = -1.68$). The value of *t* is then assessed against the value of *t* you might expect to get by chance when you have certain degrees of freedom. For the independent *t*-test, degrees of freedom are calculated by adding the two sample sizes and then subtracting the number of samples ($df = N_1 + N_2 - 2 = 12 + 12 - 2 = 22$). SPSS produces the exact significance value of *t*, and we are interested in whether this value is less than or greater than .05. In this case the two-tailed value of *p* is .107, which is greater than .05, and so we would have to conclude that there was no significant difference between the means of these two samples. In terms of the experiment, we can infer that spider-phobes are made equally anxious by pictures of spiders as they are by the real thing.

Now, we use the two-tailed probability when we have made no specific prediction about the direction of our effect (see section 2.6.2). For example, if we were unsure whether a real spider would induce more or less anxiety, then we would have to use a two-tailed test. However, often in research we can make specific predictions about which group has the highest mean. In this example, it is likely that we would have predicted that a real spider would induce greater anxiety than a picture and so we predict that the mean of the real group would be greater than the mean of the picture group. In this case, we can use a one-tailed test (for more discussion of this issue see section 2.6.2). The one-tailed probability is .107/2 = .054 (see SPSS Tip 9.1). The one-tailed probability is still greater than .05 (albeit by a small margin) and so we would still have to conclude that spider-phobes' anxiety when presented with a real spider was not significantly different to spider-phobes who were presented with a picture of the same spider. This result was predicted by the error bar chart in Figure 9.2.

9.5.4 Calculating the effect size ②

To discover whether our effect is substantive we can use the same equation as in section 9.4.6 to convert the t-statistics into a value of r. We know the value of t and the df from the SPSS output and so we can compute r as follows:

$$r = \sqrt{\frac{-1.681^2}{-1.681^2 + 22}}$$
$$= \sqrt{\frac{2.826}{24.826}}$$
$$= .34$$

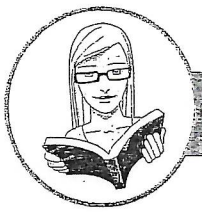
If you think back to our benchmarks for effect sizes this represents a medium effect (it is around .3, the threshold for a medium effect). Therefore, even though the effect was non-significant, it still represented a fairly substantial effect. You may also notice that the effect has shrunk, which may seem slightly odd given that we used exactly the same data (but see section 9.6)!

9.5.5 Reporting the independent t-test ①

The rules that I made up, erm, I mean, reported, for the dependent t-test pretty much apply for the independent t-test. The SPSS output tells us that the value of t was −1.68, that the number of degrees of freedom on which this was based was 22, and that it was not significant at $p < .05$. We can also see the means for each group. We could write this as:

- ✓ On average, participants experienced greater anxiety to real spiders (M = 47.00, SE = 3.18) than to pictures of spiders (M = 40.00, SE = 2.68). This difference was not significant $t(22) = -1.68$, $p > .05$; however, it did represent a medium-sized effect r = .34.

Note how we've reported the means in each group (and standard errors) as before. For the test statistic everything is much the same as before except that I've had to report that p was greater than (>) .05 rather than less than (<). Finally, note that I've commented on the effect size at the end.



CRAMMING SAM'S TIPS

- The independent *t*-test compares two means, when those means have come from different groups of entities; for example, if you have used different participants in each of two experimental conditions.
- Look at the column labelled *Levene's Test for Equality of Variance*. If the *Sig.* value is less than .05 then the assumption of homogeneity of variance has been broken and you should look at the row in the table labelled *Equal variances not assumed*. If the *Sig.* value of Levene's test is bigger than .05 then you should look at the row in the table labelled *Equal variances assumed*.
- Look at the column labelled *Sig.* If the value is less than .05 then the means of the two groups are significantly different.
- Look at the values of the means to tell you how the groups differ.
- SPSS provides only the two-tailed significance value; if you want the one-tailed significance just divide the value by 2.
- Report the *t*-statistic, the degrees of freedom and the significance value. Also report the means and their corresponding standard errors (or draw an error bar chart).
- Calculate and report the effect size. Go on, you can do it!

9.6. Between groups or repeated measures? ①

The two examples in this chapter are interesting (honestly!) because they illustrate the difference between data collected using the same participants and data collected using different participants. The two examples in this chapter use the same scores in each condition. When analysed as though the data came from the same participants the result was a significant difference between means, but when analysed as though the data came from different participants there was no significant difference between group means. This may seem like a puzzling finding – after all the numbers were identical in both examples. What this illustrates is the relative *power* of repeated-measures designs. When the same participants are used across conditions the unsystematic variance (often called the error variance) is reduced dramatically, making it easier to detect any systematic variance. It is often assumed that the way in which you collect data is irrelevant, but I hope to have illustrated that it can make the difference between detecting a difference and not detecting one. In fact, researchers have carried out studies using the same participants in experimental conditions, then repeated the study using different participants in experimental conditions, then used the method of data collection as an independent variable in the analysis. Typically, they have found that the method of data collection interacts significantly with the results found (see Erlebacher, 1977).

9.7. The *t*-test as a general linear model ②

A lot of you might think it's odd that I've chosen to represent the effect size for my *t*-tests using *r*, the correlation coefficient. In fact you might well be thinking 'but correlations show relationships, not differences between means'. I used to think this too until I read a fantastic paper by Cohen (1968), which made me realize what I'd been missing; the complex, thorny, weed-infested and large Andy-eating tarantula-inhabited world of statistics suddenly turned into a beautiful meadow filled with tulips and little bleating lambs all jumping for joy at the wonder of life. Actually, I'm still a bumbling fool trying desperately to avoid having the blood

sucked from my flaccid corpse by the tarantulas of statistics, but it was a good paper! What I'm about to say will either make no sense at all, or might help you to appreciate what I've said in most of the chapters so far: all statistical procedures are basically the same, they're just more or less elaborate versions of the correlation coefficient!

In Chapter 7 we saw that the *t*-test was used to test whether the regression coefficient of a predictor was equal to zero. The experimental design for which the independent *t*-test is used can be conceptualized as a regression equation (after all, there is one independent variable (predictor) and one dependent variable (outcome)). If we want to predict our outcome, then we can use the general equation that I've mentioned at various points:

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

If we want to use a linear model, then we saw that this general equation becomes equation (7.2) in which the model is defined by the slope and intercept of a straight line. Equation (9.6) shows a very similar equation in which A_i is the dependent variable (outcome), b_0 is the intercept, b_1 is the weighting of the predictor and G_i is the independent variable (predictor). Now, I've also included the same equation but with some of the letters replaced with what they represent in the spider experiment (so, A = anxiety, G = group). When we run an experiment with two conditions, the independent variable has only two values (group 1 or group 2). There are several ways in which these groups can be coded (in the spider example we coded group 1 with the value 0 and group 2 with the value 1). This coding variable is known as a *dummy variable* and values of this variable represent groups of entities. We have come across this coding in section 7.11:

$$\begin{aligned} A_i &= b_0 + b_1 G_i + \varepsilon_i \\ \text{Anxiety}_i &= b_0 + b_1 \text{group}_i + \varepsilon_i \end{aligned} \tag{9.6}$$

Using the spider example, we know that the mean anxiety of the picture group was 40, and that the group variable is equal to 0 for this condition. Look at what happens when the group variable is equal to 0 (the picture condition): equation (9.6) becomes (if we ignore the residual term):

$$\begin{aligned} \bar{X}_{\text{Picture}} &= b_0 + (b_1 \times 0) \\ b_0 &= \bar{X}_{\text{Picture}} \\ b_0 &= 40 \end{aligned}$$

Therefore, b_0 (the intercept) is equal to the mean of the picture group (i.e. it is the mean of the group coded as 0). Now let's look at what happens when the group variable is equal to 1. This condition is the one in which a real spider was used, therefore the mean anxiety ($\bar{X}_{\text{Real}}$) of this condition was 47. Remembering that we have just found out that b_0 is equal to the mean of the picture group ($\bar{X}_{\text{Picture}}$), equation (9.6) becomes:

$$\begin{aligned} \bar{X}_{\text{Real}} &= b_0 + (b_1 \times 1) \\ \bar{X}_{\text{Real}} &= \bar{X}_{\text{Picture}} + b_1 \\ b_1 &= \bar{X}_{\text{Real}} - \bar{X}_{\text{Picture}} \\ &= 47 - 40 \\ &= 7 \end{aligned}$$

b_1, therefore, represents the difference between the group means. As such, we can represent a two-group experiment as a regression equation in which the coefficient of the independent variable (b_1) is equal to the difference between group means, and the intercept (b_0) is equal to the mean of the group coded as 0. In regression, the t-test is used to ascertain whether the regression coefficient (b_1) is equal to 0, and when we carry out a t-test on grouped data we, therefore, test whether the difference between group means is equal to 0.

SELF-TEST To prove that I'm not making it up as I go along, run a regression on the data in **spiderBG.sav** with **group** as the predictor and **anxiety** as the outcome. **Group** is coded using zeros and ones and represents the dummy variable described above.

The resulting SPSS output should contain the regression summary table shown in SPSS Output 9.5. The first thing to notice is the value of the constant (b_0): its value is 40, the same as the mean of the base category (the picture group). The second thing to notice is that the value of the regression coefficient b_1 is 7, which is the difference between the two group means (47 – 40 = 7). Finally, the t-statistic, which tests whether b_1 is significantly different from zero, is the same as for the independent t-test (see SPSS Output 9.4) and so is the significance value.[6]

SPSS OUTPUT 9.5
Regression analysis of between-group spider data

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	40.000	2.944		13.587	.000
	Condition	7.000	4.163	.337	1.681	.107

a. Dependent Variable: Anxiety

This section has demonstrated that differences between means can be represented in terms of linear models and this concept is essential in understanding the following chapters on the general linear model.

9.8. What if my data are not normally distributed? ②

We've seen in this chapter that there are adjustments that can be made to the t-test when the assumption of homogeneity of variance is broken, but what about when you have non-normally distributed data? The first thing to note is that although a lot of early evidence suggested that t was accurate when distributions were skewed, the t-test *can be* biased when

[6] In fact, the value of the t-statistic is the same but has a positive sign rather than negative. You'll remember from the discussion of the point–biserial correlation in section 6.5.5 that when you correlate a dichotomous variable the direction of the correlation coefficient depends entirely upon which cases are assigned to which groups. Therefore, the direction of the t-statistic here is similarly influenced by which group we select to be the base category (the category coded as 0).

the assumption of normality is not met (Wilcox, 2005). Second, we need to remember that it's the shape of the sampling distribution that matters, not the sample data. One option then is to use a big sample and rely on the central limit theorem (section 2.5.1) which says that the sampling distribution should be normal when samples are big. You could also try to correct the distribution using a transformation (but see Jane Superbrain Box 5.1). Another useful solution is to use one of a group of tests commonly referred to as *non-parametric tests*. These tests have fewer assumptions than their parametric counterparts and so are useful when your data violate the assumptions of parametric data described in Chapter 5. Some of these tests are described in Chapter 15. The non-parametric counterpart of the *dependent t-test* is called the *Wilcoxon signed-rank Test* (section 15.4), and the independent *t*-test has two non-parametric counterparts (both extremely similar) called the *Wilcoxon rank-sum test* and the *Mann–Whitney test* (section 15.3). I'd recommend reading these sections before moving on.

A final option is to use robust methods (see section 5.7.4). There are various robust ways to test differences between means that involve using trimmed means or a bootstrap. However, SPSS doesn't do any of these directly. Should you wish to do these then plugin for SPSS. Look at the companion website for some demos of how to use the R plugin.

What have I discovered about statistics? ①

We started this chapter by looking at my relative failures as a human being compared to Simon Hudson before investigating some problems with the way SPSS produces error bars for repeated-measures designs. We then had a look at some general conceptual features of the *t*-test, a parametric test that's used to test differences between two means. After this general taster, we moved on to look specifically at the dependent *t*-test (used when your conditions involve the same entities). I explained how it was calculated, how to do it on SPSS and how to interpret the results. We then discovered much the same for the independent *t*-test (used when your conditions involve different entities). After this I droned on excitedly about how a situation with two conditions can be conceptualized as a general linear model, by which point those of you who have a life had gone to the pub for a stiff drink. My excitement about things like general linear models could explain why Clair Sparks chose Simon Hudson all those years ago. Perhaps she could see the writing on the wall! Fortunately, I was a ruthless pragmatist at the age of 10, and the Clair Sparks episode didn't seem to concern me unduly; I just set my sights elsewhere during the obligatory lunchtime game of kiss chase. These games were the last I would see of women for quite some time ...

Key terms that I've discovered

Dependent *t*-test
Grand mean
Independent *t*-test
Standard error of differences
Variance sum law