



## Chapter 9: Comparing two means

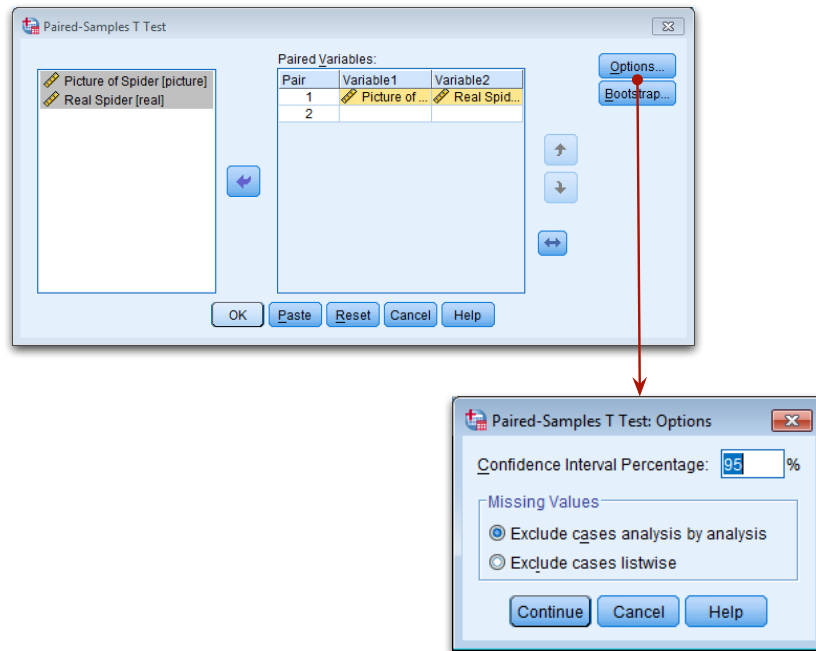
### Smart Alex's Solutions

#### Task 1

*Is arachnophobia (fear of spiders) specific to real spiders or will pictures of spiders evoke similar levels of anxiety? Twelve arachnophobes were asked to play with a big hairy tarantula spider with big fangs and an evil look in its eight eyes, and at a different point in time were shown only pictures of the same big hairy tarantula. The participants' anxiety was measured in each case. The data are in **Big Hairy Spider.sav**. Do a t-test to see whether anxiety is higher for real spiders than for pictures.*

#### Compute the test

We have 12 arachnophobes 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 onto doing the test itself. First, we need to access the main dialog box by selecting **Analyze Compare Means**  **Paired-Samples T Test...** (see below). 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 (*Cmd* on a Mac) on the keyboard and select the second (in this case **Real**). To transfer these two variables to the box labelled *Paired Variables* click on . 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. To run the analysis click on **OK**.



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

### Output from the dependent *t*-test

The resulting output produces three tables. **Error! Reference source not found.** The output below 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, which is the sample standard deviation divided by the square root of the sample size ( $SE = s/\sqrt{N}$ ), so for the picture condition  $SE = 9.2932/\sqrt{12} = 9.2932/3.4641 = 2.68$ .

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 |

**Error! Reference source not found.** The output 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). For these data the experimental conditions yield a fairly large correlation coefficient ( $r = .545$ ) but not a significant one because  $p > .05$ .

The next output (below) 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 *not* to be a chance result. First, the table tells us the mean difference between scores. 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. The test statistic,  $t$ , is calculated by dividing the mean of differences by the standard error of differences ( $t = -7/2.8311 = -2.47$ ). 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. 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. 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 arachnophobes than exposure to a picture,  $t(11) = -2.47, p = .031$ .

|        |                                 | Paired Differences |                |                 |   |       | t      | df | Sig. (2-tailed) |
|--------|---------------------------------|--------------------|----------------|-----------------|---|-------|--------|----|-----------------|
|        |                                 | Mean               | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference |       |        |    |                 |
|        |                                 |                    |                |                 | 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 interval 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. So, assuming this sample's confidence interval is one of the 95 out of 100 that contains the population value, 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.

### Calculating the effect size

We can compute the effect size in the same way as for the independent  $t$ -test. All we need is the value of  $t$  and the  $df$  from the SPSS output and 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 probably a substantive finding.

### Reporting the analysis

The SPSS **Error! Reference source not found.** output tells us that the value of  $t$  was  $-2.47$ , that this was based on 11 degrees of freedom, 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 with real spiders ( $M = 47.00$ ,  $SE = 3.18$ ) than with pictures of spiders ( $M = 40.00$ ,  $SE = 2.68$ ),  $t(11) = -2.47$ ,  $p = .031$ ,  $r = .60$ .

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 using the exact significance. Finally, note that we've reported the effect size at the end.

Try to avoid writing vague, unsubstantiated things like this:




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

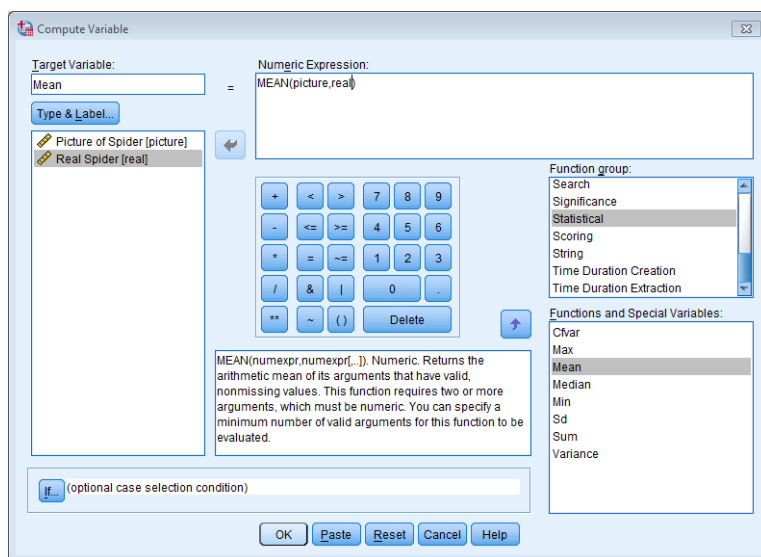
More scared than what? Where are the degrees of freedom? Was the result statistically significant? Was the effect important (what was the effect size)?

## Task 2

*Plot an error bar graph of the above data (remember to make the necessary adjustments for the fact the data are from a repeated-measures design).*

### Step 1: Calculate the mean for each participant

To correct the repeated-measures error bars, we need to use the *compute* command. To begin with, we need to calculate the average anxiety for each participant and so we use the *mean* function. Access the main *compute* dialog box by selecting **Transform**  **Compute Variable...**. Enter the name **Mean** into the box labelled *Target Variable* and then in the list labelled *Function group* select *Statistical* and then in the list labelled *Functions and Special Variables* select *Mean*. Transfer this command to the command area by clicking on . When the command is transferred, it appears in the command area as *MEAN(?,?)*; the question marks should be replaced with variable names (which can be typed manually or transferred from the variables list). So replace the first question mark with the variable **picture** and the second one with the variable **real**. The completed dialog box should look like the one below. Click on  to create this new variable, which will appear as a new column in the data editor.



Using the *compute* function to calculate the mean of two columns

## Step 2: Calculate the grand mean

Access the *descriptives* command by selecting **Analyze** **Descriptive Statistics** **Descriptives...**. The dialog box shown below should appear. The *descriptives* command is used to get basic descriptive statistics for variables, and by clicking on **Options...** a second dialog box is activated. Select the variable **Mean** from the list and transfer it to the box labelled **Variable(s)** by clicking on **➔**. Then use the *Options* dialog box to specify only the mean (you can leave the default settings as they are, but it is only the mean in which we are interested). If you run this analysis the output should provide you with some self-explanatory descriptive statistics for each of the three variables (assuming you selected all three). You should see that we get the mean of the picture condition, and the mean of the real spider condition, but it's actually the final variable we're interested in: the mean of the picture and spider condition. The mean of this variable is the grand mean, and you can see from the summary table that its value is 43.50. We will use this grand mean in the following calculations.

The image shows two dialog boxes from SPSS. The 'Descriptives' dialog box has 'Picture of Spider (p...', 'Real Spider (reat', and 'Mean [Mean]' selected in the 'Variable(s):' list. The 'Descriptives: Options' dialog box has 'Mean' checked under 'Dispersion', and 'Variable list' selected under 'Display Order'. Below the dialog boxes is a 'Descriptive Statistics' table.

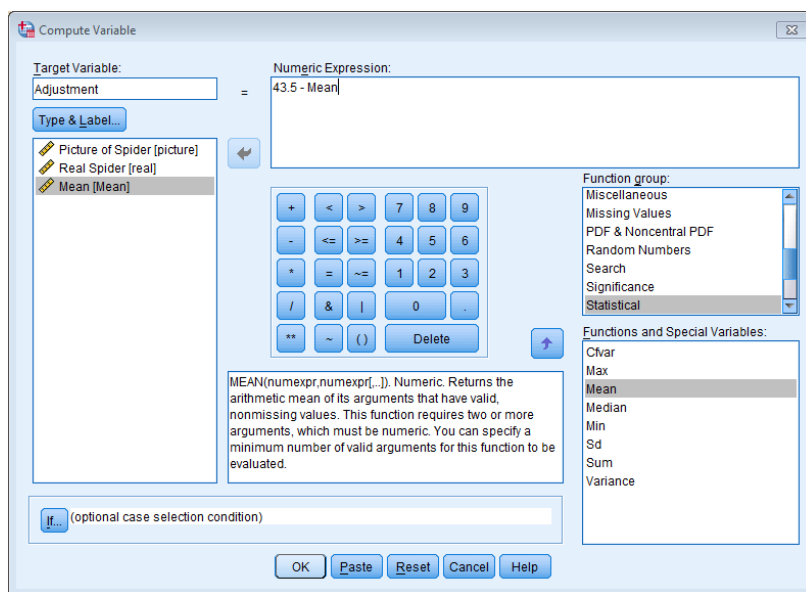
|                    | N  | Minimum | Maximum | Mean    | Std. Deviation |
|--------------------|----|---------|---------|---------|----------------|
| Picture of Spider  | 12 | 25      | 55      | 40.00   | 9.293          |
| Real Spider        | 12 | 30      | 65      | 47.00   | 11.029         |
| Mean               | 12 | 30.00   | 57.50   | 43.5000 | 8.94173        |
| Valid N (listwise) | 12 |         |         |         |                |

Dialog boxes and output for descriptive statistics

## Step 3: Calculate the adjustment factor

If you look at the variable labelled **Mean**, you should notice that the values for each participant are different, which tells us that some people had greater anxiety than others did across the conditions. The fact that participants' mean anxiety scores differ represents individual differences between different people (so it represents the fact that some of the participants are generally more scared of spiders than others). These differences in natural anxiety to spiders contaminate the error bar graphs, which is why, if we don't adjust the values that we plot, we will get the same graph as if an independent design had been used. Loftus and Masson (1994) argue that to eliminate this contamination we should equalize the means between participants (i.e., adjust the scores in each condition such that when we take the

mean score across conditions, it is the same for all participants). To do this, we need to calculate an adjustment factor by subtracting each participant's mean score from the grand mean. We can use the *compute* function to do this calculation for us. Activate the *compute* dialog box, give the target variable a name (I suggest **Adjustment**) and then use the command '43.5-mean'. This command will take the grand mean (43.5) and subtract from it each participant's average anxiety level (see below).






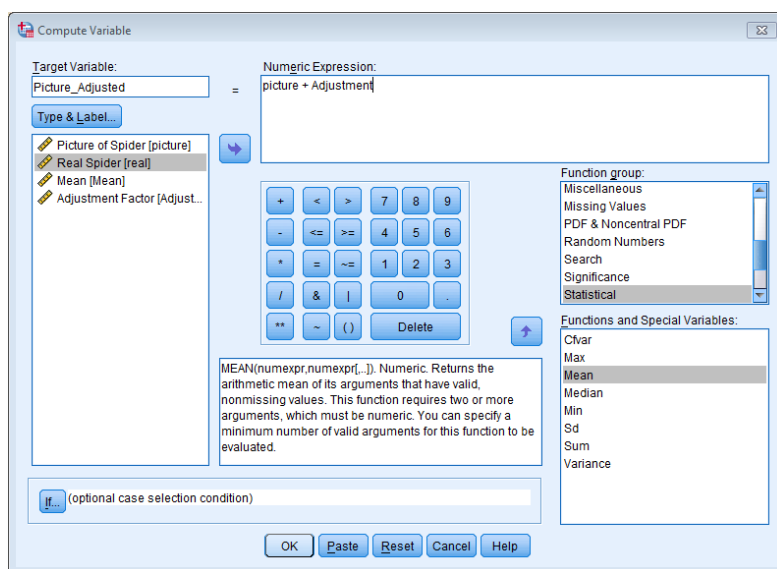
#### Calculating the adjustment factor

This process creates a new variable in the data editor called **Adjustment**. The scores in the **Adjustment** column represent the difference between each participant's mean anxiety and the mean anxiety level across all participants. You'll notice that some of the values are positive, and these participants are one's who were less anxious than average. Other participants were more anxious than average and they have negative adjustment scores. We can now use these adjustment values to eliminate the between-subject differences in anxiety.

#### Step 4: Create adjusted values for each variable

So far, we have calculated the difference between each participant's mean score and the mean score of all participants (the grand mean). This difference can be used to adjust the existing scores for each participant. First we need to adjust the scores in the **picture** condition. Once again, we can use the *compute* command to make the adjustment. Activate the *compute* dialog box in the same way as before, and then title our new variable **Picture\_Adjusted** (you can then click on **Type & Label...** and give this variable a label such as 'Picture Condition: Adjusted Values'). All we are going to do is to add each participant's score in the **picture** condition to

their adjustment value. Select the variable **picture** and transfer it to the command area by clicking on , then click on  and select the variable **Adjustment** and transfer it to the command area by clicking on . The completed dialog box is shown below. Now do the same thing for the variable **real**: create a variable called **Real\_Adjusted** that contains the values of **real** added to the value in the **Adjustment** column.

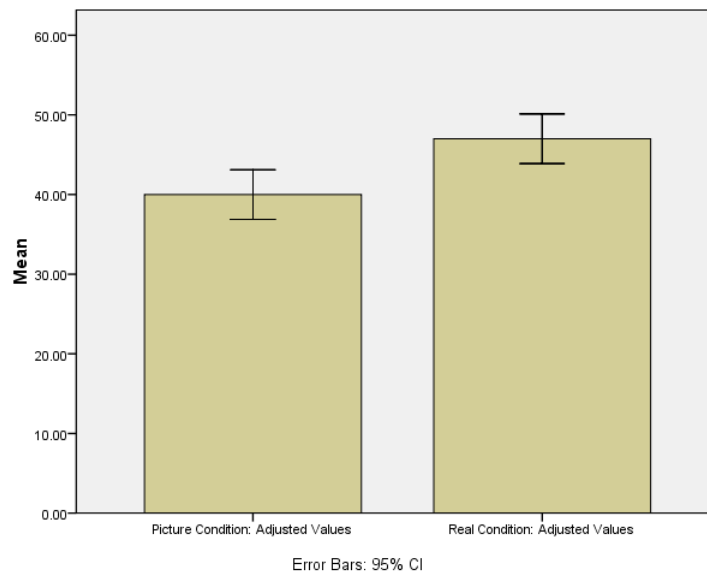


#### Adjusting the values of picture

Now, the variables **Real\_Adjusted** and **Picture\_Adjusted** represent the anxiety experienced in each condition, adjusted so as to eliminate any between-subject differences. If you don't believe me, then use the *compute* command to create a variable **Mean2** that is the average of **Real\_Adjusted** and **Picture\_Adjusted**. You should find that the value in this column is the same for every participant, thus proving that the between-subject variability in means is gone: the value will be 43.50, which is the grand mean.

The resulting error bar graph is shown below. The error bars don't overlap which suggests that the groups are significantly different (although we knew this already from the previous task).





Error bar graph of the adjusted values of Big Hairy Spider.sav

### Task 3

*One of my pet hates is 'pop psychology' books. They usually spout nonsense that is unsubstantiated by science and give psychology a very bad name. As part of my plan to rid the world of popular psychology I did a little experiment. I took two groups of people who were in relationships and randomly assigned them to one of two conditions. One group read the famous popular psychology book Women are from Bras and Men are from Penis, whereas another group read Marie Claire. I tested only 10 people in each of these groups, and the dependent variable was an objective measure of their happiness with their relationship after reading the book. The data are in the file **Penis.sav**. Analyse them with the appropriate t-test.*

SPSS output for the independent *t*-test

| Book Read              |   | N  | Mean    | Std. Deviation | Std. Error Mean |
|------------------------|---|----|---------|----------------|-----------------|
| Relationship Happiness | Women are from Bras, Men are from Penis | 10 | 20.0000 | 4.10961        | 1.29957         |
|                        | Marie Claire                            | 10 | 24.2000 | 4.70933        | 1.48922         |

|                        |                             | Levene's Test for Equality of Variances |      | t-test for Equality of Means |        |                 |                 |                       |   |         |
|------------------------|-----------------------------|---|------|------------------------------|--------|-----------------|-----------------|-----------------------|---|---------|
|                        |                             | F                                       | Sig. | t                            | df     | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference |         |
|                        |                             |   |      |                              |        |                 |                 |                       | Lower                                     | Upper   |
| Relationship Happiness | Equal variances assumed     | .491                                    | .492 | -2.125                       | 18     | .048            | -4.2000         | 1.97653               | -8.35253                                  | -.04747 |
|                        | Equal variances not assumed |   |      | -2.125                       | 17.676 | .048            | -4.2000         | 1.97653               | -8.35800                                  | -.04200 |

## Calculating the effect size

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

$$\begin{aligned}
 r &= \sqrt{\frac{(-2.125)^2}{(-2.125)^2 + 18}} \\
 &= \sqrt{\frac{4.52}{22.52}} \\
 &= .45
 \end{aligned}$$

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

## Reporting the results

When you report any statistical test you usually state the finding to which the test relates, and then, in brackets, report the test statistic (usually with its degrees of freedom), the probability value of that test statistic, and more recently the American Psychological Association is, quite rightly, requesting an estimate of the effect size. To get you into good habits early, we'll start thinking about effect sizes now, before you get too fixated on Fisher's magic .05. In this example we know that the value of *t* was  $-2.12$ , that this was based on 18 degrees of freedom, and that it was significant at  $p = .048$ . This can all be obtained from the SPSS [Error! Reference source not found](#).output above. We can also see the means for each group. Based on what we learnt about reporting means, we could now write something like:

- ✓ On average, the reported relationship happiness after reading *Marie Claire* ( $M = 24.20$ ,  $SE = 1.49$ ), was significantly higher than after reading *Women are from Bras and Men are from Penis* ( $M = 20.00$ ,  $SE = 1.30$ ),  $t(18) = -2.12$ ,  $p < .05$ ,  $r = .45$ .

## Task 4

Imagine Twaddle and Sons, the publishers of *Women are from Bras and Men are from Penis*, were upset about my claims that their book was as useful as a paper umbrella. They designed their own experiment in which participants read their book and one of my books (Field & Hole, 2003) at different times. Relationship happiness was measured after reading each book. They used a sample of 500 participants, but got each participant to take part in both conditions (in counterbalanced order and with a six-month delay). Does reading their wonderful contribution to popular psychology lead to greater relationship happiness compared to my tedious book about experiments? The data are in **Field&Hole.sav**.

### SPSS output

Paired Samples Statistics

|        |   | Mean    | N   | Std. Deviation | Std. Error Mean |
|--------|---|---------|-----|----------------|-----------------|
| Pair 1 | Women are from Bras, Men are from Penis | 20.0180 | 500 | 9.98123        | .44637          |
|        | Field & Hole                            | 18.4900 | 500 | 8.99153        | .40211          |

Paired Samples Correlations

|        |  | N   | Correlation | Sig. |
|--------|--|-----|-------------|------|
| Pair 1 | Women are from Bras, Men are from Penis & Field & Hole | 500 | .117        | .009 |

Paired Samples Test

|        |  | Paired Differences |                |                 |   | t      | df    | Sig. (2-tailed) |       |
|--------|--|--------------------|----------------|-----------------|---|--------|-------|-----------------|-------|
|        |  | Mean               | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference |        |       |                 |       |
|        |  |                    |                |                 | Lower                                     |        |       |                 | Upper |
| Pair 1 | Women are from Bras, Men are from Penis - Field & Hole | 1.5280             | 12.62807       | .56474          | .4184                                     | 2.6376 | 2.706 | 499             | .007  |

### Calculating the effect size

We know the value of  $t$  and the  $df$  from the SPSS **Error! Reference source not found.** output above and so we can compute  $r$  as follows:

$$\begin{aligned}
 r &= \sqrt{\frac{2.706^2}{2.706^2 + 499}} \\
 &= \sqrt{\frac{7.32}{506.32}} \\
 &= .12
 \end{aligned}$$

If you think back to our benchmarks for effect sizes, this represents a small effect (it is just above .1, the threshold for a small effect). Therefore, although this effect is highly statistically significant, the size of the effect is very small and so represents a trivial finding.

### Interpreting and writing the results

In this example, it would be tempting for Twaddle and Sons to conclude that their book produced significantly greater relationship happiness than our book. In fact, many researchers would write conclusions like this:

- The results show that reading *Women are from Bras and Men are from Penis* produces significantly greater relationship happiness than that book by smelly old Field and Hole. This result is highly significant.

However, to reach such a conclusion is to confuse statistical significance with the importance of the effect. By calculating the effect size we've discovered that although the difference in happiness after reading the two books is statistically very different, the size of effect that this represents is very small indeed. So, the effect is actually not very significant in real terms. A more correct interpretation might be to say:

- The results show that reading *Women are from Bras and Men are from Penis* produces significantly greater relationship happiness than that book by smelly old Field and Hole. However, the effect size was small, revealing that this finding was not substantial in real terms.

Of course, this latter interpretation would be unpopular with Twaddle and Sons who would like to believe that their book had a huge effect on relationship happiness.

## Task 5

*In Chapter 3 (Task 5) we looked at data from people who had been forced to marry goats and dogs and measured their life satisfaction as well as how much they like animals (**Goat or Dog.sav**). Conduct a t-test to see whether life satisfaction depends upon the type of animal to which a person was married.*

SPSS output for the independent *t*-test

|                       | Type of Animal Wife | N  | Mean  | Std. Deviation | Std. Error Mean |
|-----------------------|---------------------|----|-------|----------------|-----------------|
| Life Satisfaction (%) | Goat                | 12 | 38.17 | 15.509         | 4.477           |
|                       | Dog                 | 8  | 60.13 | 11.103         | 3.925           |

|                       | Levene's Test for Equality of Variances |      | t-test for Equality of Means |        |                 |                 |                       |   |        |
|-----------------------|---|------|------------------------------|--------|-----------------|-----------------|-----------------------|---|--------|
|                       | F                                       | Sig. | t                            | df     | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference |        |
|                       |   |      |                              |        |                 |                 |                       | Lower                                     | Upper  |
| Life Satisfaction (%) | 1.407                                   | .251 | -3.446                       | 18     | .003            | -21.958         | 6.372                 | -35.346                                   | -8.570 |
|                       |   |      | -3.688                       | 17.843 | .002            | -21.958         | 5.954                 | -34.475                                   | -9.441 |

## Calculating the effect size

We know the value of *t* and the *df* from the SPSS [Error! Reference source not found.](#) output above and so we can compute *r* as follows:

$$\begin{aligned}
 r &= \sqrt{\frac{(-3.446)^2}{(-3.446)^2 + 18}} \\
 &= \sqrt{\frac{11.87}{29.87}} \\
 &= .63
 \end{aligned}$$

If you think back to our benchmarks for effect sizes, this represents a large effect. Therefore, as well as being statistically significant, this effect is large and so represents a substantive finding.

## Reporting the results

- ✓ On average, the life satisfaction of men married to dogs ( $M = 60.13$ ,  $SE = 3.93$ ), was significantly higher than that of men who were married to goats ( $M = 38.17$ ,  $SE = 4.48$ ),  $t(18) = -3.45$ ,  $p < .01$ ,  $r = .63$ .

## Task 6

*What do you notice about the *t*-value and significance above compared to when you ran the analysis as a regression in Chapter 8, Task 2?*

Output from the independent  $t$ -test

|                       |                             | 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  |
| Life Satisfaction (%) | Equal variances assumed     | 1.407                                   | .251 | -3.446                       | 18     | .003            | -21.958         | 6.372                 | -35.346                                   | -8.570 |
|                       | Equal variances not assumed |   |      | -3.688                       | 17.843 | .002            | -21.958         | 5.954                 | -34.475                                   | -9.441 |

## Output from the linear regression

| Coefficients <sup>a</sup> |                     |                             |            |                           |       |      |
|---------------------------|---------------------|-----------------------------|------------|---------------------------|-------|------|
| Model                     |                     | Unstandardized Coefficients |            | Standardized Coefficients | t     | Sig. |
|                           |                     | B                           | Std. Error | Beta                      |       |      |
| 1                         | (Constant)          | 16.208                      | 9.452      |                           | 1.715 | .104 |
|                           | Type of Animal Wife | 21.958                      | 6.372      | .630                      | 3.446 | .003 |

a. Dependent Variable: Life Satisfaction (%)

Looking at the two outputs above, we can see that the results are the same. The main point I wanted to make here is that whether you run these data through regression or a  $t$ -test, the results are identical.

## Task 7

*In Chapter 2 we looked at hygiene scores over three days of a rock music festival (Download Festival.sav). Do a paired-samples  $t$ -test to see whether hygiene scores on day 1 differed from those on day 3.*

**Paired Samples Correlations**

|        |   | N   | Correlation | Sig. |
|--------|---|-----|-------------|------|
| Pair 1 | Hygiene (Day 1 of Download Festival) & Hygiene (Day 3 of Download Festival) | 123 | .458        | .000 |

**Paired Samples Statistics**

|        |                                      | Mean   | N   | Std. Deviation | Std. Error Mean |
|--------|--------------------------------------|--------|-----|----------------|-----------------|
| Pair 1 | Hygiene (Day 1 of Download Festival) | 1.6515 | 123 | .64390         | .05806          |
|        | Hygiene (Day 3 of Download Festival) | .9765  | 123 | .71028         | .06404          |

**Paired Samples Test**

|        |   | Paired Differences |                |                 |   | t      | df     | Sig. (2-tailed) |       |
|--------|---|--------------------|----------------|-----------------|---|--------|--------|-----------------|-------|
|        |   | Mean               | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference |        |        |                 |       |
|        |   |                    |                |                 | Lower                                     |        |        |                 | Upper |
| Pair 1 | Hygiene (Day 1 of Download Festival) - Hygiene (Day 3 of Download Festival) | .67496             | .70709         | .06376          | .54875                                    | .80117 | 10.587 | 122             | .000  |

**Calculating the effect size**

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

$$\begin{aligned}
 r &= \sqrt{\frac{10.587^2}{10.587^2 + 122}} \\
 &= \sqrt{\frac{112.08}{234.08}} \\
 &= .69
 \end{aligned}$$

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

**Reporting the results**

- ✓ On average, hygiene scores significantly decreased from day 1 ( $M = 1.65$ ,  $SE = 0.06$ ), to day 3 ( $M = 0.98$ ,  $SE = 0.06$ ) of the Download music festival,  $t(122) = 10.59$ ,  $p < .001$ ,  $r = .69$ .

## Task 8

Analyse the data in Chapter 6, Task 1 (whether men and dogs differ in their dog-like behaviours – **MenLikeDogs.sav**) using an independent t-test with bootstrapping. Do you reach the same conclusions?

**Group Statistics**

| Species            |                 | Statistic       | Bootstrap <sup>a</sup> |            |                             |          |          |
|--------------------|-----------------|-----------------|------------------------|------------|-----------------------------|----------|----------|
|                    |                 |                 | Bias                   | Std. Error | BCa 95% Confidence Interval |          |          |
|                    |                 |                 |                        |            | Lower                       | Upper    |          |
| Dog-Like Behaviour | Dog             | N               | 20                     |            |                             |          |          |
|                    |                 | Mean            | 28.0500                | -.0202     | 2.3704                      | 23.5990  | 32.6475  |
|                    |                 | Std. Deviation  | 10.98072               | -.42579    | 1.45425                     | 8.27237  | 12.47959 |
|                    |                 | Std. Error Mean | 2.45536                |            |                             |          |          |
|                    | Man             | N               | 20                     |            |                             |          |          |
|                    |                 | Mean            | 26.8500                | -.0197     | 2.2303                      | 23.1176  | 30.9093  |
|                    | Std. Deviation  | 9.90096         | -.41885                | 1.95349    | 6.58720                     | 12.55587 |          |
|                    | Std. Error Mean | 2.21392         |                        |            |                             |          |          |

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

**Independent Samples Test**

|                    |                             | Levene's Test for Equality of Variances |      | t-test for Equality of Means |        |                 |                 |                       |   |         |
|--------------------|-----------------------------|---|------|------------------------------|--------|-----------------|-----------------|-----------------------|---|---------|
|                    |                             | F                                       | Sig. | t                            | df     | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference |         |
|                    |                             |   |      |                              |        |                 |                 |                       | Lower                                     | Upper   |
| Dog-Like Behaviour | Equal variances assumed     | 1.149                                   | .291 | .363                         | 38     | .719            | 1.20000         | 3.30609               | -5.49284                                  | 7.89284 |
|                    | Equal variances not assumed |   |      | .363                         | 37.600 | .719            | 1.20000         | 3.30609               | -5.49518                                  | 7.89518 |

If we look at the significance value of  $t$  in the output above, we can see that it was .719. Because this value is above .05 this indicates a non-significant effect. Therefore we would conclude that men and dogs do not significantly differ in the amount of dog-like behaviour they engage in.

The **Error! Reference source not found.** output below shows the results of bootstrapping. You can see that the confidence interval ranged from  $-5.25$  to  $7.87$ , which implies that the difference between means in the population could be negative, positive or even zero. In other words, it's possible that the true difference between means is zero. Therefore, this bootstrap confidence interval confirms our conclusion that men and dogs do not differ in amount of dog-like behaviour.



Bootstrap for Independent Samples Test

|                    |                             | Mean Difference | Bootstrap <sup>a</sup> |            |                             |         |
|--------------------|-----------------------------|-----------------|------------------------|------------|-----------------------------|---------|
|                    |                             |                 | Bias                   | Std. Error | BCa 95% Confidence Interval |         |
|                    |                             |                 |                        |            | Lower                       | Upper   |
| Dog-Like Behaviour | Equal variances assumed     | 1.20000         | -.06309                | 3.29713    | -5.25160                    | 7.86818 |
|                    | Equal variances not assumed | 1.20000         | -.06309                | 3.29713    | -5.25160                    | 7.86818 |

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

### Calculating the effect size

We know the value of  $t$  and the  $df$  from the Independent Samples Test table above and so we can compute  $r$  as follows:

$$\begin{aligned}
 r &= \sqrt{\frac{0.363^2}{0.363^2 + 38}} \\
 &= \sqrt{\frac{0.132}{38.13}} \\
 &= .06
 \end{aligned}$$

If you think back to our benchmarks for effect sizes, this represents a tiny effect.

### Reporting the results

- ✓ On average, men ( $M = 26.85$ ,  $SE = 2.23$ ) engaged in less dog-like behaviour than dogs ( $M = 28.05$ ,  $SE = 2.37$ ). However, this difference, 1.2, BCa 95% CI [-5.83, 7.99], was not significant,  $t(38) = 0.36$ ,  $p = .72$ ,  $r = .06$ .

## Task 9

Analyse the data in Chapter 6, Task 2 (whether the type of music you hear influences goat sacrificing – **DarkLord.sav**) using an matched-samples t-test with bootstrapping. Do you reach the same conclusions?

**Paired Samples Statistics**

|        |            |                 | Statistic | Bootstrap <sup>a</sup> |            |                             |         |
|--------|------------|-----------------|-----------|------------------------|------------|-----------------------------|---------|
|        |            |                 |           | Bias                   | Std. Error | BCa 95% Confidence Interval |         |
|        |            |                 | Lower     |                        |            | Upper                       |         |
| Pair 1 | Message    | Mean            | 9.1563    | .0016                  | .6212      | 8.0000                      | 10.3155 |
|        |            | N               | 32        |                        |            |                             |         |
|        |            | Std. Deviation  | 3.54792   | -.07251                | .44241     | 2.72893                     | 4.23651 |
|        |            | Std. Error Mean | .62719    |                        |            |                             |         |
|        | No Message | Mean            | 11.5000   | -.0060                 | .8003      | 10.0313                     | 13.0000 |
|        |            | N               | 32        |                        |            |                             |         |
|        |            | Std. Deviation  | 4.38472   | -.13655                | .58753     | 3.25719                     | 5.07907 |
|        |            | Std. Error Mean | .77512    |                        |            |                             |         |

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

**Paired Samples Correlations**

|        |                      |    | N    | Correlation | Sig.  | Bootstrap for Correlation <sup>a</sup> |            |                             |
|--------|----------------------|----|------|-------------|-------|--|------------|-----------------------------|
|        |                      |    |      |             |       | Bias                                   | Std. Error | BCa 95% Confidence Interval |
|        |                      |    |      |             |       | Lower                                  | Upper      |                             |
| Pair 1 | Message & No Message | 32 | .283 | .116        | -.006 | .152                                   | -.016      | .560                        |

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

**Paired Samples Test**

|        |                      | Paired Differences |                |                 |   |         | t      | df | Sig. (2-tailed) |
|--------|----------------------|--------------------|----------------|-----------------|---|---------|--------|----|-----------------|
|        |                      | Mean               | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference |         |        |    |                 |
|        |                      |                    |                |                 | Lower                                     | Upper   |        |    |                 |
| Pair 1 | Message - No Message | -2.34375           | 4.79657        | .84792          | -4.07310                                  | -.61440 | -2.764 | 31 | .010            |

**Bootstrap for Paired Samples Test**

|        |                      | Mean     | Bootstrap <sup>a</sup> |            |                 |                             |         |
|--------|----------------------|----------|------------------------|------------|-----------------|-----------------------------|---------|
|        |                      |          | Bias                   | Std. Error | Sig. (2-tailed) | BCa 95% Confidence Interval |         |
|        |                      |          |                        |            |                 | Lower                       | Upper   |
| Pair 1 | Message - No Message | -2.34375 | .00756                 | .86017     | .019            | -4.19233                    | -.65625 |

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

**Error! Reference source not found.** The output above shows the results of bootstrapping. You can see that the confidence interval ranged from  $-4.19$  to  $-0.66$ ; because it does not cross zero, we can be confident that the effect in the population is unlikely to be zero and so implies that there is a significant difference between means in the population. Therefore, this bootstrap confidence interval confirms our conclusion that there is a significant difference between the number of goats sacrificed when listening to the song containing the backward message compared to when listening to the song played normally.

## Calculating the effect size

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

$$\begin{aligned} r &= \sqrt{\frac{(-2.76)^2}{(-2.76)^2 + 31}} \\ &= \sqrt{\frac{7.62}{38.62}} \\ &= .44 \end{aligned}$$

If you think back to our benchmarks for effect sizes, this represents a fairly large effect (it is close to .5, which is the point at which an effect is classed as being large). Therefore, as well as being statistically significant, this effect represents a substantive finding.

## Reporting the results

- ✓ Fewer goats were sacrificed after hearing the backward message ( $M = 9.16$ ,  $SE = 0.62$ ), than after hearing the normal version of the Britney song ( $M = 11.50$ ,  $SE = 0.80$ ). This difference, 2.34, BCa 95% CI [-4.19, -.66], was significant,  $t(31) = -2.76$ ,  $p < .05$ ,  $r = .44$ .

## Task 10

*Thinking back to Labcoat Leni's Real Research 3.1, test whether the number of offers was significantly different in people listening to Bon Scott compared to those listening to Brian Johnson, using an independent t-test and bootstrapping. Do your results differ from Oxoby (2008)? (The data are in **Oxoby (2008) Offers.sav**.)*

**Group Statistics**

|                 |   |                 | Statistic       | Bootstrap <sup>a</sup> |            |                             |       |       |
|-----------------|---|-----------------|-----------------|------------------------|------------|-----------------------------|-------|-------|
|                 |   |                 |                 | Bias                   | Std. Error | BCa 95% Confidence Interval |       |       |
|                 |   |                 |                 |                        | Lower      | Upper                       |       |       |
| Offer Made (\$) | Background Music                        |                 | N               | 18                     |            |                             |       |       |
|                 | Bonn Scott (It's a Long Way to the Top) |                 | Mean            | 3.28                   | -.01       | .28                         | 2.78  | 3.76  |
|                 |   |                 | Std. Deviation  | 1.179                  | -.040      | .154                        | .926  | 1.348 |
|                 |   |                 | Std. Error Mean | .278                   |            |                             |       |       |
|                 | Brian Johnson (Shoot to Thrill)         |                 | N               | 18                     |            |                             |       |       |
|                 |   |                 | Mean            | 4.00                   | .00        | .23                         | 3.59  | 4.42  |
|                 |   | Std. Deviation  | .970            | -.035                  | .124       | .765                        | 1.095 |       |
|                 |   | Std. Error Mean | .229            |                        |            |                             |       |       |

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

**Independent Samples Test**

|                 |                             | Levene's Test for Equality of Variances |      | t-test for Equality of Means |       |                 |                 |                       |   |       |
|-----------------|-----------------------------|---|------|------------------------------|-------|-----------------|-----------------|-----------------------|---|-------|
|                 |                             | F                                       | Sig. | t                            | df    | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference |       |
|                 |                             |   |      |                              |       |                 |                 |                       | Lower                                     | Upper |
| Offer Made (\$) | Equal variances assumed     | 1.029                                   | .317 | -2.007                       | 34    | .053            | -.722           | .360                  | -1.453                                    | .009  |
|                 | Equal variances not assumed |   |      | -2.007                       | 32.79 | .053            | -.722           | .360                  | -1.454                                    | .010  |

**Bootstrap for Independent Samples Test**

|                 |                             | Mean Difference | Bootstrap <sup>a</sup> |            |                             |       |
|-----------------|-----------------------------|-----------------|------------------------|------------|-----------------------------|-------|
|                 |                             |                 | Bias                   | Std. Error | BCa 95% Confidence Interval |       |
|                 |                             |                 |                        |            | Lower                       | Upper |
| Offer Made (\$) | Equal variances assumed     | -.722           | -.009                  | .355       | -1.403                      | -.075 |
|                 | Equal variances not assumed | -.722           | -.009                  | .355       | -1.403                      | -.075 |

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

The **Error! Reference source not found.** output above shows the results of bootstrapping. You can see that the confidence interval ranged from  $-1.40$  to  $-0.08$ , which does not cross zero (both limits are negative) and thus implies that the difference between means in the population is unlikely to be zero. However, the upper confidence interval ( $-0.08$ ) is very close to zero, which reflects the only borderline significant  $t$ -value.

## Calculating the effect size

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

$$\begin{aligned} r &= \sqrt{\frac{(-2.01)^2}{(-2.01)^2 + 34}} \\ &= \sqrt{\frac{4.04}{38.04}} \\ &= .33 \end{aligned}$$

If you think back to our benchmarks for effect sizes, this represents a medium effect.

## Reporting the results

- ✓ On average, more offers were made when listening to Brian Johnson ( $M = 4.00$ ,  $SE = 0.23$ ) than Bon Scott ( $M = 3.28$ ,  $SE = 0.28$ ). This difference, 0.72, BCa 95% CI  $[-1.40, -0.08]$ , was only borderline significant,  $t(34) = -2.01$ ,  $p = .05$ ; however, it produced a medium effect,  $r = .33$ .