Examining Differences between Means: The t-test

Key goals for this chapter

- Introduce the t-test – a statistical procedure that allows researchers to establish how likely it is that observed differences between two means would have arisen by chance.
- Show how different types of t-test are used to examine different types of data.
- Explain the process of interpreting the results of t-tests, and the different ways that this can be done.

In the previous chapter we dealt with some statistical inferences under conditions where we knew quite a few things about the dependent variable. We used two examples: coin tossing and IQ test scores. At the risk of revealing trade secrets, we should point out that these two examples are pretty common in this sort of book.

The reasons for the popularity of these examples are worth looking at: the example of coin tossing is widely used because it relates to people’s experiences in their everyday lives, but it is also an example where we can directly calculate an exact probability for each possible outcome. However, this state of affairs is pretty rare in psychological statistics. As we made clear in Chapter 7, most of the time we first have to calculate a test statistic (involving dividing an information term by an error term) and then identify the probability associated with that statistic by working it out or looking it up in a table.

The example of IQ scores is also widely used in books such as this one, in order to provide an introductory example of statistical inference. In part this is because IQ is a psychological variable that most people will have heard of before, but an even better reason is that we know that in the population the mean IQ is 100 and the standard deviation is 15. How do we know this? You might think that there is a mysterious process at work here, but in fact...
we know these things because the people who design IQ tests actually go to some trouble to ensure that IQ tests yield a mean of 100 and a standard deviation of 15. Once the test designer has ensured that the test is a valid and reliable measure of intelligence through extremely extensive sampling of the population and rigorous comparison with other measures, the mean of the test is reset to 100 and the standard deviation is set to 15 (following the process called standardization that we explained in Chapter 7). When all this is done the results of the test are then said to be IQ scores.

But this is actually quite an unusual set of circumstances. This extensive process of measurement and validation has been carried out for only a limited number of psychological variables, and in many cases there are still residual disputes about the validity of these measures. This means that, most of the time, psychological researchers do not know the population parameters of the measures they use. Instead they have to estimate these parameters. Along these lines, you will remember that we discussed how to estimate the population standard deviation in Chapter 6.

If we can estimate the population standard deviation we are in a position to conduct a *t*-test. A *t*-test is a test for differences between means under conditions where the population standard deviation is not known and therefore has to be estimated. In this chapter we first discuss the case where we are comparing the mean of a single sample with a known value. This is very similar to the case in the previous chapter where we compared the mean IQ of a sample of children in a particular class with the population mean of 100. We then deal with within-subjects experiments where we compare the mean response of a sample at one point in time, or on one measure, with the sample’s mean response at another time or on another measure. Finally, we discuss the two-sample or between-subjects *t*-test. Here we compare means of two different samples of people. At that stage we will finally be able to answer the sort of question that we posed at the start of the previous chapter: How certain can we be that the scores for two groups of people on some psychological variable are really different?

### Student’s *t*-distribution

The *t*-distribution describes a sampling distribution used to test the difference between means when the population standard deviation (σ) is unknown. It has a mean, or expected value, of 0. The reason why the expected value is 0 is that the sampling distribution of differences between means is based on a statistical model where we imagine that we are continuously drawing two random samples from the same population and comparing their means. Because the two samples are drawn from the same population, their means should both be estimates of the same population mean. Thus, saying that the expected value of the sampling distribution of the difference between the means is 0 is just a technical way of saying that we would expect no difference between the means if they are in fact estimates of the same
population mean. If we find a large enough difference between the means of the two samples we draw, we start to question whether both means are actually estimates of the same population mean. We would go through the same process if we saw a coin come down heads or tails too many times and started to wonder if the coin was fair.

Put simply, the *t*-distribution describes the statistical model of what happens when you take random samples of a certain size from the population and subtract one mean from the other and divide the result by an estimate of the standard error. We are able to compare empirical results with the results we obtain from this statistical model and, what is more, the *t*-distribution tells us how likely the observed difference between means would be if that *t*-distribution’s statistical model did apply. In the same way that the binomial distribution can tell us how likely a run of heads is with a fair coin, the *t*-distribution can tell us what the chance is of drawing two random samples from the same population with a difference between them of a certain size or bigger. However, what the model can never do is tell us whether the two samples are from the same population or from different populations. That is a decision researchers have to make for themselves, guided by the results of the statistical test.

The shape of the *t*-distribution changes as the sample size increases. This is once again a consequence of the law of large numbers. As we take larger and larger samples we obtain a statistic that actually becomes closer and closer to the normal distribution. This is largely a result of the fact that as samples become larger we obtain better estimates of the unknown population standard deviation. Indeed, with infinitely large samples *t* becomes normally distributed and equivalent to *z*.

There is thus a different *t*-distribution for every different sample size. Each *t*-distribution shows what the probability of any given difference between means is, expressed in standard error units. For example, if we found a difference between mean heights for a sample of three men and a sample of three women of 10 cm and a standard error of 5 cm then this would correspond to 2.0 *t*-units (i.e., 10/5). This would cut off the region of the *t*-distribution that is marked in Figure 8.1(a). We can see that only a small proportion of all the possible differences between the means lies beyond this point.

What this distribution does is map out all the possibilities that could occur if you were to calculate the mean for random samples of three people from the population and subtract the means for another random sample of three people, and express the differences between the means in standard error units. We can contrast this to the *t*-distribution for two samples of 30 people (Figure 8.1(b)).

The overall shapes of these distributions are pretty similar, but you can see that the tails of the sampling distribution as a whole are much smaller with the larger sample size. This is reflected in the fact that the area beyond 2.0 *t*-units is even smaller than it was in Figure 8.1(a), so that it contains an even smaller proportion of possible differences between means.

Rather than corresponding to sample size exactly, however, the sampling distribution of *t* depends on something called the **degrees of freedom**. The degrees of freedom are the
8. Differences between Means

Figure 8.1  Sampling distribution of the difference between the means for two samples of size 3 (a) and 30 (b)
Test Yourself 8.1*

Which of the following statements is/are true about the t-distribution?

a. Its shape changes with the number of degrees of freedom.

b. Its expected value is 0.

c. It can be used to test differences between means providing that the population standard deviation is known.

d. Both (a) and (b).

e. All of the above.

The correct answer is (d). Answers (a) and (b) are both true of the t-distribution, but (c) is wrong – and therefore (e) too – because it is not necessary to know the population standard deviation in order to use the t-distribution. Indeed, the t-distribution was specifically developed to apply to sampling situations where the population standard deviation is unknown.

number of scores in a given set that are free to vary. To explain what this means, think of a waiter in a restaurant who forgets which customer ordered what at a table of five people. Here he has four degrees of freedom because if he can remember what four people ordered, he must get the fifth right. Put another way, if the waiter just guesses which meal to hand to one person after another he is only ‘free’ to make mistakes to a certain degree and that degree is four (if he gets the first four people’s meals wrong then he knows that the meal belongs to the fifth). If he has one table of five people and one of two – providing that he knows which table placed the orders – then he has five degrees of freedom. That is,

df = (N_{table1} - 1) + (N_{table2} - 1) = 4 + 1

Returning to Figure 8.1, the top figure is actually the t-distribution for 4 degrees of freedom and the bottom figure is t for 58 degrees of freedom. The calculation of the number of degrees of freedom varies a bit depending on what sort of t-test you are doing, so we explain the calculation for each case below. The issue of degrees of freedom is the major difference between the z-distribution and the t-distribution. There is only one z-distribution for z-tests, the standard normal distribution, but for t-tests there is a different distribution for each different number of degrees of freedom.

degrees of freedom (df) In the calculation of a given statistic, the number of scores that are free to vary.

sampling distribution of the difference between means The hypothetical distribution of differences between the means of two random samples drawn from
the same population. It has a mean of 0 and a standard deviation that decreases with the square root of the sample size.

**t-distribution** The hypothetical distribution of expected differences between the means of randomly chosen samples drawn from the same population divided by an estimate of the standard error.

### Comparing the results for a single sample with a specific value

Often psychologists collect a set of data from a sample and then want to know whether the results are greater or less than some specified value. Let us imagine that 10 students do a multiple-choice test in psychology. Imagine, too, that the mean number of correct responses is 25 out of 100 and that there are five choices for each question. We might want to decide whether the students are performing better on the test than they would if they had just randomly chosen an answer. Given that there are five choices, the expected number of correct answers is 20. Some example data are shown in Table 8.1.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Exam score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
</tr>
</tbody>
</table>

To judge whether students are performing better than chance we construct the test statistic \( t \) as a ratio of information over error (in fact we could do an exact probability calculation in this case as we did in the coin-tossing example, but we chose not to do that here).

We can use a computer program such as IBM SPSS Statistics software (SPSS) to help us make this judgement. To start with, open up the application SPSS on your computer and choose the option to type in new data.
Research Methods and Statistics in Psychology

Figure 8.2

Figure 8.3
8. Differences between Means

Type in the data exactly as shown in Figure 8.2. Include one row for each participant with the participant number in the first column and the score in the second column. This is a very common pattern used in data analysis in psychology and other human sciences where we have rows representing a set of scores for a participant and columns representing variables. In this case the first variable is not real data, it is simply a record-keeping variable that we create for the purpose of keeping track of which person has which score. Note too that, to enable you to check your work, these data are also provided in the Supplementary Online Material [Chapter8_OneSamplettest.sav].

When you have done that let us rename ‘VAR00001’ by changing it to ‘Participant’ and ‘VAR00002’ to ‘TestScore’ (with no space before ‘Score’). We are now actually ready to use SPSS to perform the statistical analysis, but before starting choose a helpful name and save the data file somewhere you will be able to keep it safe (noting that the habit of saving data files carefully and using logical names to identify them is a very good one to get into).

After saving the data go to the ‘Analyze’ menu of SPSS and select ‘Compare Means → One-Sample T Test’. When you do you should see the display shown in Figure 8.3.

The variable that we want to test here is called TestScore. We want to be able to judge whether exam scores are higher than the expected (chance) value of 20. Thus, our test variable is ’TestScore’ so we need to drag that variable from the left to the right in order to select it for the analysis. We also have to enter the test value of 20 into the text box labelled ‘Test Value’. The window should now look like Figure 8.4.

![Figure 8.4](image-url)
If it does, click on ‘OK’ and you should see the display in Figure 8.5.

T-Test

SPSS here provides us first with some descriptive statistics that tell us that there are 10 participants ($N = 10$) with a mean exam score of 25.0 and that the standard deviation is 3.68 and the standard error of the mean is 1.16 (this is just the standard deviation divided by the square root of the sample size).

The next table contains the actual results of the analysis (or inferential statistics). Here we have compared a mean score of 25.0 with the test value of 20. We want to know if that difference is big enough to be taken seriously or so small that it can be treated as a chance variation. In other words, is this class of 10 students good or are they just lucky?

The results of the analysis give us a value for $t$ of 4.294. This has been calculated by dividing the difference between the mean (25; $\overline{X}$ in the equation below) and the test value (20; value in the equation below) by the standard error of the mean (1.16; $s_{\overline{X}}/\sqrt{N}$ in the equation below). This division involves the ratio of information to error. Here the information term tells us that the mean is five points bigger than the test value and the error term tells us that the average amount of variation expected in the mean – for a sample of 10 people drawn from a population with that standard deviation – is around 1.16 points.

$$ t(9) = \frac{\overline{X} - \text{value}}{s_{\overline{X}}/\sqrt{N}} $$

$$ = \frac{25 - 20}{3.68/\sqrt{10}} $$

$$ = 4.29 $$

Is this $t$-value big enough for us to decide one way or another? If the $t$-value were 0 then that would tell us that there was absolutely no difference between the sample mean and
The mean height of a sample of 100 men in 2014 is 180 cm, with a standard deviation of 10 cm. The mean for the same population in 1914 was 170 cm. Imagine that we want to use the 2014 sample to test whether the height of the population has changed over the intervening 100 years. Which of the following statements is true?

a. The number of degrees of freedom for the test in this case is 99.
b. The t-value is 10.0.
c. The t-value corresponds to a very small probability that a random process of taking samples of 100 men from a population similar to the 1914 population would produce a sample of 180 cm or taller.
d. All of the above.
e. Answers (a) and (b) only.

The correct answer is (d). Answers (a), (b) and (c) are all correct here. The degrees of freedom are equal to \( n - 1 = 100 - 1 = 99 \), so (a) is correct. The t-value is \( \frac{180 - 170}{10/\sqrt{100}} = 10 \), so (b) is correct. A t-value of 10.00 far exceeds the t-value for 99 degrees of freedom associated with a probability level of .001, so (c) is correct. In other words, there is less than a 1 in 1000 chance that a random process of taking samples from the 1914 population would produce a difference this size or larger.
The table puts the same information in another way by showing us the confidence intervals for the difference. We will come back to confidence intervals later in the chapter, but for now just think of confidence intervals as a range of expected values.

**Within-subjects t-tests**

The within-subjects *t*-test is used when we are comparing means based on sets of data that are collected in pairs from the same participants. Let us go back to the example used in Chapter 2 where we talked about the process of testing a ‘physical reinforcement theory’. This simple theory explained the observation that ‘absence leads the heart to wander’ by proposing that attraction to another person is dependent on physical reinforcement achieved through interaction with them. Let us imagine we do an experiment to test this theory where participants have to make attractiveness ratings of two people – one with whom they have interacted (Person I, for Interaction) and one with whom they have not (Person N, for No interaction). In this experiment physical reinforcement has been manipulated within subjects because all participants are in both conditions. Our theory predicts that attraction will be greater for ratings of Person I than Person N. The dependent variable is ratings of the attractiveness of these two people on a nine-point scale from 1 to 9 (where 1 = ‘do not like at all’ and 9 = ‘like a great deal’). Suppose our data were as in Table 8.2.

**Table 8.2**

<table>
<thead>
<tr>
<th>Participant</th>
<th>Target person</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>N</td>
</tr>
<tr>
<td>1.00</td>
<td>7.00</td>
<td>6.00</td>
</tr>
<tr>
<td>2.00</td>
<td>6.00</td>
<td>6.00</td>
</tr>
<tr>
<td>3.00</td>
<td>5.00</td>
<td>2.00</td>
</tr>
<tr>
<td>4.00</td>
<td>6.00</td>
<td>4.00</td>
</tr>
<tr>
<td>5.00</td>
<td>5.00</td>
<td>3.00</td>
</tr>
<tr>
<td>6.00</td>
<td>7.00</td>
<td>5.00</td>
</tr>
<tr>
<td>7.00</td>
<td>3.00</td>
<td>5.00</td>
</tr>
</tbody>
</table>

Note that these are *pairs* of data. For example, Participant 1 responded ‘7’ when rating the person he or she had interacted with (Person I) and ‘5’ when rating the person he or she had not interacted with (Person N).

Just by looking at these results we can see that the pattern of data is generally consistent with our hypothesis, but the question is whether the ratings of I and N are *sufficiently different* to infer that they reflect a real underlying difference. As we have already noted several times, inferential statistics cannot answer this question directly, but they can tell
8. Differences between Means

us how likely it is that two sets of responses as different as this would have been drawn randomly from the same population. To find out, we can use the skills with SPSS from the earlier example:

Figure 8.6

1. Open SPSS and create a new data file (the Type in Data choice).
2. Create three variables in the data file named ‘Participant’, ‘RateI’ and ‘RateN’ and enter the data as in Figure 8.6 (this is also provided in the Supplementary Online Material [Chapter8_WithinSSttest.sav]).
   Click on ‘Variable View’ and then ‘Variable Label’ so you can give the variables meaningful labels such as ‘Rating of Person I’ and ‘Rating of Person N’.
3. After saving the data file in a personal folder or on a USB drive go to the ‘Analyze’ menu of SPSS and select ‘Compare Means → Paired-Samples T Test’. When you do you should see the display in Figure 8.7.
The variables we want to compare are ‘Rating of Person I’ and ‘Rating of Person N’. So drag ‘Rating of Person I’ to be ‘Variable 1’ in ‘Pair’ ‘1’ and ‘Rating of Person N’ to be ‘Variable 2’ and then click on ‘OK’.

The first table that is generated gives us the descriptive statistics, in the form shown in Figure 8.8.

The second table (Figure 8.9) gives us another descriptive statistic that tells us the strength of the (linear) relationship between the variables. We discuss this more in the next chapter, but for the time being we simply want to know if the value in the correlation cell is greater than .3 (if it is not then we are not using the best method to answer the question).
8. Differences between Means

The third table (Figure 8.10) gives us the result of the \( t \)-test. The value tells us that the mean difference between the ratings of Person I and Person N is 1.14 (I is rated higher than N) and that the variability in these differences can be expressed as a standard deviation of 1.67616 or a standard error of the mean of .63353. This corresponds to a \( t \)-value with 6 degrees of freedom (df) of 1.804 that is associated with a \( p \)-value of .121.

![Figure 8.10](image)

SPSS has calculated the \( t \)-value using the following formula:

\[
t = \frac{\bar{D}}{S_D/\sqrt{N}}
\]

(\text{information term})

(\text{error term})

So, substituting relevant values from the above tables, we can see this was calculated as follows:

\[
t = \frac{1.14285}{1.67616/\sqrt{7}}
\]

= \frac{1.14285}{0.617}

= 1.804

The corresponding \( p \)-value tells us that if there really were no differences between the measures (i.e., they were both measuring the same thing), then a difference as large as that which has been obtained could be expected 121 out of every 1000 times. That is, it is quite unlikely but not extremely unlikely.

**Between-subjects \( t \)-tests**

The above case dealt with a situation where we had two sets of data from the same sample of participants. So, by using difference scores, we could use one mean and one standard deviation to represent both sets of scores. Let us now consider a between-subjects case where we are comparing means obtained from different groups of participants. In essence, our decision as to whether those means are likely to have been drawn from the same population comes down to a statement about the difference between two means relative to the variation of scores around each of those means and the number of responses on which the means are based.
This may sound complex, but the underlying logic is quite simple and can be illustrated in the following scenario. In this diagram each 'e' is a response on a scale from 1 to 7 by a person in an experimental condition, and each 'c' a response on the same scale by a person in a control condition; the sample size varies between the three cases:

<table>
<thead>
<tr>
<th>Case 1</th>
<th>c</th>
<th>c</th>
<th>c</th>
<th>c</th>
<th>e</th>
<th>e</th>
<th>e</th>
<th>Mean c = 2.5 Mean e = 5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 2</th>
<th>c</th>
<th>c</th>
<th>e</th>
<th>e</th>
<th>Mean c = 2.5 Mean e = 5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 3</th>
<th>cc</th>
<th>cc</th>
<th>ee</th>
<th>ee</th>
<th>Mean c = 2.5 Mean e = 5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

In which of these cases would you be most confident that the responses of people in the two conditions reflected a real underlying difference between groups? Your confidence should be greatest in Case 3. This is for two reasons. First, in Case 3 the dispersion of the two sets of responses is smaller than in Case 1 (and the overlap in the response distributions is smaller). Second, Case 3 provides more evidence of a real difference than Case 2 because it is based on more data (i.e., the sample size is larger). So here, even though the means are the same in each case, we would be more confident that the population mean for the experimental condition really was greater than that for the control condition in Case 3 because the difference in means relative to the amount of error is greater.

Going back to our example of research on ‘physical reinforcement theory’ that was exploring the basis of attraction between people, let us consider a new experiment. In this experiment, instead of each participant rating two target people (the within-subjects case), let us imagine that some participants in an experimental condition interact with and then rate a target person and that other participants in a control condition do not interact with the target person they are asked to rate. Participants might then rate their attraction to the target person they are asked to consider on a nine-point scale (where 1 = ‘do not like at all’ and 9 = ‘like a great deal’). Here we would obtain two independent sets of ‘attractiveness ratings’, one from participants in the experimental condition and one from participants in the control condition. Let us imagine that the data looked like Table 8.3.

Our physical reinforcement theory predicts that participants in the experimental condition should like the target person more. This would appear to be the case: the mean rating of the target’s attractiveness is higher in the experimental condition ($\bar{X}_1 = 6.4$) than in the control condition ($\bar{X}_2 = 4.4$).
8. Differences between Means

Table 8.3

<table>
<thead>
<tr>
<th>Participant</th>
<th>Attractiveness rating</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>Experimental</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>Experimental</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>Experimental</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>Experimental</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>Experimental</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>Experimental</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>Experimental</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>Experimental</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>Experimental</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>Experimental</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>Control</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>Control</td>
</tr>
<tr>
<td>13</td>
<td>6</td>
<td>Control</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>Control</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>Control</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>Control</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>Control</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>Control</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>Control</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>Control</td>
</tr>
</tbody>
</table>

The question, though, is whether this difference is big enough to allow us to conclude that the two means are unlikely to have been drawn from the same population. In other words, what we really want to know is whether the two samples are drawn from the same population (with the same population mean) and are just different by chance or whether the samples are drawn from two different populations. As you may be sick of reading by now, we can never tell this exactly. So instead we determine the chance that a difference this large could be obtained by drawing random samples from the same population. To do this we can go back to SPSS and create a new data file entering the numbers from the table above (again, this is also available in the Supplementary Online Material [Chapter8_BetweenSStest.sav]).

This time, in addition to scores we have a categorical variable, condition, that has two levels: experimental condition and control condition. We need to tell SPSS which of our participants are in each condition. Let us use the coding ‘1 = Experimental’ and ‘2 = Control’.

We now need to put some labels on the values to help us understand the output and so that we remember how we coded these variables the next time we look at them. After entering the data as shown above, go to ‘Variable View’ and click on ‘VAR00003’. First of all rename this ‘Condition’ and then click on ‘Value Labels’. Type in ‘Experimental Group’ for the value ‘1’,
and ‘Control Group’ for value ‘2’. After renaming ‘VAR00001’ as ‘Participant’ and ‘VAR00002’ as ‘AttrScore’ and giving it a clear variable label such as ‘Attractiveness Rating’ and then saving the data file, we are ready to go.

Now select ‘Compare Means Æ Independent-Samples T Test’. Select ‘AttrScore’ as the ‘Test Variable’ and ‘Condition’ as the ‘Grouping Variable’, and then select ‘Group 1’ to have the value ‘1’ and ‘Group 2’ to have the value ‘2’. Click on ‘OK’ and SPSS will produce two tables. The first of these (shown in Figure 8.11) contains descriptive statistics. The second (shown in Figure 8.12) contains the inferential statistics.

**Figure 8.11**

**Figure 8.12**

SPSS gives us two versions of the t-test, one assuming that the samples are drawn from populations with the same variance and one that relaxes that assumption. We will come to this later, but in this case the results are the same for the two methods so we will choose the method assuming equal variance.

SPSS has calculated a t-value of 2.900 based on 18 degrees of freedom that is associated with a p-value of .010. SPSS has calculated this t-value using the following formula:

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}}} \tag{information term}
\]

\[
= \sqrt{\frac{2.38 \left( \frac{1}{10} + \frac{1}{10} \right)}{0.69}} = 2.90
\tag{error term}
\]

In this case:
8. Differences between Means

The \( p \)-value of .010 implies that there is 1 chance in 100 that a random process could have produced a difference between means this large.

Before moving on, it is worth noting that one of the most confusing things about \( t \)-tests is the wide variety of different terms that are used to describe the different types of tests. If you were to read another textbook after this one, you might be forgiven for thinking that there are lots of different types of \( t \)-test. In fact there are only two types that you really have to worry about – the within-subjects \( t \)-test and the between-subjects \( t \)-test. We have chosen to use these names because they relate directly to the methodological terms that psychologists use, and because they match the usage of these terms in other statistical tests (in particular, analysis of variance, an advanced procedure for comparing more than two means, also referred to as ANOVA; see Chapter 10).

Table 8.4 is intended to help you work out what the different types of \( t \)-test are when you see the terms used in other books and in the research literature.

### Table 8.4 Terms used to describe \( t \)-tests

<table>
<thead>
<tr>
<th>Terms for within-subjects ( t )-tests</th>
<th>Terms for between-subjects ( t )-tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent ( t )-test</td>
<td>Two-sample ( t )-test</td>
</tr>
<tr>
<td>Related-samples ( t )-test</td>
<td>Independent-groups ( t )-test</td>
</tr>
<tr>
<td>Paired samples ( t )-test (this is what SPSS calls it)</td>
<td>Independent-samples ( t )-test (this is what SPSS calls it)</td>
</tr>
<tr>
<td>Repeated-measures ( t )-test</td>
<td></td>
</tr>
<tr>
<td>Correlated-samples ( t )-test</td>
<td></td>
</tr>
</tbody>
</table>

**Test Yourself 8.3**

Which of the following is *not* relevant to a between-subjects \( t \)-test?

a. The pooled variance estimate.

b. A difference score \( D \).

c. An information term.

d. A sampling distribution of the differences between means.

e. A probability that a difference of the same size or larger could be obtained by a random process.

The correct answer is (b). The difference score \( D \) is used in the within-subjects \( t \)-test where we deal with the difference between pairs of the scores from the same participant.
The degree to which two distributions share the same range of values. The overlap of the two distributions in Figure 8.13 is shown by the shaded region.

A pooled variance estimate is an estimate of the variance of the population made by averaging the variances of the two samples to provide the error term for the between-subjects t-test.

A weighted average is an average where the result is modified to take account of differing sample sizes. A weighted average tends towards a given sample mean in

Test Yourself 8.4***

If an experimenter conducts a t-test to see whether the responses of participants in a control group differ from those of an experimental group, which of the following outcomes will yield the highest t-value?

a. If there are 10 participants in each condition and the difference between the mean responses of the control group and the experimental group is 2 and both have standard deviations of 1.

b. If there are 10 participants in each condition and the difference between the mean responses of the control group and the experimental group is 2 and both have standard deviations of 2.

c. If there are 20 participants in each condition and the difference between the mean responses of the control group and the experimental group is 1 and both have standard deviations of 1.

(d) If there are 20 participants in each condition and the difference between the mean responses of the control group and the experimental group is 2 and both have standard deviations of 1.

e. If there are 20 participants in each condition and the difference between the mean responses of the control group and the experimental group is 2 and both have standard deviations of 2.

The correct answer is (d). The value of test is largest when you have small standard deviation, large samples and a large difference between means. All these conditions apply in answer (d). The largest standard deviation gives a large difference between means. The pooled standard deviation is 1 and the largest difference between means is 2. Based on these principles, we can conclude that the value of test is largest in (d).

The correct answer is (d). The value of test is largest when you have small standard deviation, large samples and a large difference between means.
Differences between Means

proportion to the sample’s size relative to the size of other samples being compared. If one group of 10 people has a mean score of 10 and another group of 100 people has a mean score of 20, the weighted average will therefore be much closer to the mean of the larger sample. In this case it is not 15 but \[
\frac{(10 \times 10) + (100 \times 20)}{110} = 19.09.
\]

The controversy about what to do with t-values

At this stage you might be feeling that the process of statistical inference is a bit clumsy. After all, we are interested in two possibilities: whether the results we have obtained are due to chance or whether they represent a real difference between two means. These possibilities seem pretty obvious to most people when they are explained, but confusion arises because inferential statistics never allow us to reach either conclusion with complete certainty. Statistical tests never allow us to say that our results are due to a random process. All they can ever do is tell us how likely they would be if they were due to a random process. Similarly, the results of statistical tests can never tell us that our findings are not due to chance.

The inability of statistics to answer particular questions can be confusing and frustrating for both students and experienced researchers – not least because in order to convey accurately the truth of what statistical tests are telling us we have to use very awkward forms of expression. Indeed, you probably encountered some frustration while reading the preceding pages. Consider the way we described the results of the within-subjects t-test: ‘That is, it is quite unlikely but not really unlikely.’ We can hear you saying ‘Can’t you just tell me whether it is one or the other? Is it unlikely or not? Or could you at least tell me whether it is unlikely enough?’.
However, attempts to simplify the way we express statistical ideas lead to inaccuracy. Because the temptation to simplify statistical test results is quite strong, many psychologists and statisticians have argued that statistical tests are poorly understood and poorly used. This problem can be traced back to the fact that statistical tests do not answer exactly the questions that research scientists want to answer. They only answer related questions.

When you decided where to study or which mobile phone to buy you may have asked other people about their experiences or have read reviews. You may have used this information to make your decision, but your decision is not the same as that information.

RESEARCH BITE 8.1
The prudence of a Bayesian approach

One major limitation of the hypothesis-testing approach is that it does not allow researchers to test the plausibility of the null hypothesis or to weigh evidence in favour of this against that in favour of the alternative hypothesis. One approach that can do these things involves using Bayesian statistics. Although beyond the scope of this text, these are becoming increasingly popular in many areas of psychology. A recent study by Ruud Wetzels and colleagues (2011) also shows that in around 70% of cases where a hypothesis-testing approach would lead one to conclude that there is positive evidence against the null hypothesis, a Bayesian approach would characterize that evidence as relatively weak.

Reference

Photo 8.1 Did this card player cheat, or is he just lucky? It is important to remember that statistics cannot answer such questions. In this case they could only allow us to say exactly how lucky this person was if they were playing fair – and then to base our judgement on this information.
8. Differences between Means

Doing a statistical test is more like getting the advice of a trusted friend (‘yep, that’s a good phone’) than actually making a purchase decision.

Another useful analogy might be an election result and a public opinion poll a month before the election. The question that everybody wants to know the answer to is who will win the election. We cannot answer that question until the election but, as we saw in Chapter 6, if an opinion poll is conducted properly, we can use its results as a guide upon which to base our expectations. Statistical tests play the same role. They cannot tell us if there is a real difference but they give us a basis for making a reasonable decision about whether there is a difference.

Problems with reporting and interpreting statistics have been around for a long time. But recently in psychology there has been a great deal of argument about exactly how to use inferential statistics (e.g., as reflected in the initial report of the Task Force on Statistical Inference of the Board of Scientific Affairs of the American Psychological Association, released December 1996, 1998). For the purposes of this chapter, the controversy relates to what to do with $t$-values once we have obtained them. The next few sections offer choices to students and instructors about how to conduct statistical tests and make statistical inferences.

We suggest that all students should be familiar with the hypothesis-testing approach and with the alternative approaches. This follows the advice of many psychological statisticians who argue that we should compensate for the limitations of the hypothesis-testing approach by employing (or at least considering) other methods (Cohen, 1994, 1995; Hammond, 1996; Judd, McClelland, & Culhane, 1995; Syvantek & Ekeberg, 1995; Wilkinson et al., 1999).

Handling the results of $t$-tests: The hypothesis-testing approach

Once we have obtained the value of $t$ and worked out the corresponding probability value we have some more decisions to make. What we have effectively done so far is turn an observed difference between means into a probability value. We then have to decide whether that probability is small enough for us to conclude that the two means were unlikely to have been drawn from the same population. You might reasonably ask whether this process has got us very far. After all, we only did the $t$-test so that we could make a definitive statement about whether the difference between the means was large or small, and now it seems that we have to decide whether the $p$-value is big or small.

The answer is that we have indeed made some progress because, rather than simply talking about differences between means, we now have a probability value that is comparable across research situations. The probability value is comparable because it depends on both the difference between the means and the amount of dispersion or random error in the sample.

The hypothesis-testing or significance-testing approach to $t$-tests is an attempt to come up with definitive answers to the question of whether we have observed an informative difference between means or not. The logic underpinning this approach is to develop a statistical model that assumes that both means are based on random samples from the same population (i.e., the samples share the same population mean and standard deviation). The sampling
distribution provided by the $t$-distribution is just a mathematical way of expressing this statistical model of no difference between means, and, as we have discussed, the expected value of the $t$-distribution is zero. That is, if the means are both drawn from the same population then there should be no difference.

This statistical model that there should be no difference between the means is called the **null hypothesis.** In all psychological research the null hypothesis is really a statement about what we would expect if there is nothing dramatic going on in our research (and in the population to which it relates). That is, it is a statement of what should occur if we are simply observing a random process. The null hypothesis is often called $H_0$ (pronounced 'H nought').

When testing for a difference between means using a $t$-test, the null hypothesis is

$$H_0: \mu_1 = \mu_2 = \mu$$

What this says is that the null hypothesis assumes that, $\mu_1$, the population mean associated with Sample 1 (in the between-subjects case) or with Variable 1 (in the within-subjects case), is the same as $\mu_2$, the population mean associated with Sample 2 or Variable 2, and that they are both equal to the same value ($\mu$). This model also assumes that the standard deviations of each sample are estimates of the population standard deviation (and should therefore be similar).

The model is just like imagining that we have programmed a computer to repeatedly produce means for pairs of samples of a given size from a population with a certain mean and standard deviation. Imagine that we then get the computer to calculate the $t$-values obtained and draw these differences as a frequency graph. If the null hypothesis is true then our data should look like a typical distribution of $t$-values randomly generated by the computer (i.e., a normal distribution). In fact, because random variables are so well behaved in the long run, we do not need to write such a computer program to do statistics – we can use equations and look up values in a table.

When we use a hypothesis-testing approach to test for differences between means we contrast this null hypothesis model with the alternative that there actually is a difference between the means. This is called the **alternative hypothesis** or $H_1$ and is a statement to the effect that something dramatic is going on in our research – reflecting something going on in the population from which our samples are drawn. In this case we assert that the two means are really drawn from two different populations with different means. When testing for a difference between means using a $t$-test, the general form of the alternative hypothesis is therefore

$$H_1: \mu_1 \neq \mu_2$$

This says that the alternative hypothesis is that $\mu_1$, the population mean associated with Sample 1 (in the between-subjects case) or with Variable 1 (in the within-subjects case), is *not* the same as the population mean associated with Sample 2 or Variable 2. An alternative hypothesis like this is normally based on a theory which provides the rationale for our research.

The logical process of the hypothesis-testing approach involves testing between the null hypothesis and the alternative hypothesis. We do not accept the null hypothesis when inferen-
Differences between Means

8. Differences between Means

tial uncertainty is high, but we reject it and accept the alternative hypothesis when inferential uncertainty is low. In other words, we never say that the null hypothesis is true, but under certain circumstances we conclude that it is probably false, and so accept the alternative hypothesis. Researchers and students often forget the fact that, regardless of the results of statistical tests, we can reject or not reject the null hypothesis but we are never entitled to accept the null hypothesis. Again, this is because statistics can never prove that a random process has produced the results. We do not know whether the null hypothesis is actually true or whether our study was just not sensitive enough to detect a real difference between means.

In order to make the decision about whether we will reject the null hypothesis or not, we take our \( t \)-distribution and divide it into three regions. A large region in the middle contains most possible values of \( t \) and two small regions in the tails of the distribution contain only a few extreme values of \( t \). These small regions are called the rejection regions. If the obtained value of \( t \) falls in one of these rejection regions then we say that the difference between means is statistically significant. In other words, statistical significance is another way of saying that inferential uncertainty is low. The process is shown in Figure 8.14.

![The \( t \)-distribution showing rejection regions](image)

**Figure 8.14** The \( t \)-distribution showing rejection regions

You will note that for each tail there is a value that cuts off the rejection region. These values are called critical values. If the obtained \( t \)-value falls in the tail beyond either critical value, it falls in the rejection region. Under these conditions we reject the null hypothesis (hence the name ‘rejection regions’) and conclude that the difference between means is significant – thereby accepting the alternative hypothesis.
How large should these rejection regions be? Well, that depends on how much inferential uncertainty you are prepared to live with. Those psychologists who use significance levels tend to follow the suggestion that a 5% or less chance that a random process could produce the same results is sufficiently low to conclude that the results were not produced by a random process. This figure is commonly used in psychological research, but people also use 1% and .1% chances. We call these values **alpha levels** (or, to use the Greek letter for alpha, $\alpha$ levels). Alpha levels correspond to the size of the rejection region, and because we can never be completely sure that our results are not due to chance, an alpha level is essentially just a researcher’s statement about the level of inferential uncertainty that he or she is prepared to live with. Rather than use percentages as alpha levels we quote them as probability values, so a rejection region of 5% corresponds to an alpha level of .05. When the obtained $t$-value exceeds the critical value and the value falls in the rejection region, we say ‘$p$ is less than $\alpha$’ where $\alpha$ is an alpha level such as .05 (so here we would write $p < .05$).

As an analogy to the process of working with alpha levels, imagine that you and a friend want to drive to a town 500 kilometres away. Your car is not particularly new and lately you have heard a few rattles coming from somewhere near the engine. Your natural inclination is to assume that the rattles mean nothing, that they are just randomly generated noise that you need not pay attention to. This, in effect, is a null hypothesis – a statement of the view that the noise does not amount to much. On the other hand, you at least entertain the possibility that there may be something to the noise after all. Maybe the engine is about to self-destruct. This, in effect, is an alternative hypothesis – a statement of the view that the rattle is not just a random noise, but a signal that you should pay attention to and probably do something about.

Faced with these two possibilities (the only logical alternatives) you then have to decide which one to go with. It is not an easy decision to make, but one way to do this would be to decide on an alpha level for your car before you set out on your trip. This amounts to a statement of the amount of uncertainty about whether the car is fit for the journey that you are prepared to accept – a statement of when you are going to take the mysterious noise seriously. The alpha level you set will depend on a large number of factors including the importance of the journey, the cost of a breakdown and the weather. You might be prepared to start the journey if you thought that the chance that there was really something wrong with the engine was less than 10% (an alpha level of .10). However, your friend might only be prepared to travel with you if there was no more than a 1% chance of the engine breaking down (an alpha level of .01). Obviously, in everyday life we do not go around quantifying uncertainty like this in numbers. We express our everyday doubts by saying things like ‘I’m pretty sure we’ll make it’ or ‘I don’t want to seem rude, but I’d feel happier taking the bus’.

In reflecting on this process, it is important to point out that alpha levels are criteria that researchers set before they conduct research in much the same way that high jumpers set the height of the bar before they run up and make a jump. If their obtained result yields a value that falls in the rejection region associated with a given alpha level then they reject the null hypothesis. How much they exceed the critical value by is not particularly important. On the other hand, if their obtained value falls outside the rejection region then they do not reject the null
8. Differences between Means

hypothesis. Again this decision should not be affected by the distance between the obtained and critical values. The point here is that if you adopt the hypothesis-testing approach then you either clear the bar (the critical value associated with your alpha level) or you do not.

There are just two further complications to the process we have described so far. You will note that there are two rejection regions in Figure 8.14. This is because $t$-values can be positive or negative. This occurs because the first mean can be larger than the second, or the second can be larger than the first. Because the obtained value of $t$ can fall in either tail, we call this a **two-tailed test**.

The total rejection region is divided into two tails. If $\alpha = .05$ (5%) then each of the two rejection regions will be 2.5% (i.e., $\alpha/2$) of the total distribution. Under some circumstances, however, we can take a slightly different approach where we test a different alternative hypothesis. Think for a moment about $H_1$:

$$H_1: \mu_1 \neq \mu_2$$

This can be re-expressed in the following way:

$$H_{1a}: \mu_1 > \mu_2$$

and

$$H_{1b}: \mu_1 < \mu_2$$

That is, if the two means are not equal then one has to be larger than the other. Sometimes we might want to test both those possibilities, but under other circumstances this does not make sense. If we compare the heights of men and women we might want to know whether men are taller or not, but it is not particularly useful or relevant to test the idea that men are shorter than women. Thus, under some circumstances we only have

$$H_1: \mu_1 > \mu_2$$

That is, we hypothesize that one mean will be greater than the other. If this is the case, rather than using two rejection regions, it can be appropriate to use a single rejection region at one end of the distribution, as shown in Figure 8.15. You will note that the critical value cutting off the rejection region is closer to zero (i.e., smaller) in this case. This test with only one rejection region is called a **one-tailed test** or a **directional test**.

Opinions about when one-tailed tests should be used vary, but they are most appropriate under circumstances where researchers know that one mean cannot possibly be larger than the other. So unless they have a good reason not to, researchers should use the two-tailed test described above, where there is a rejection region in each tail. Using one-tailed tests to make non-significant results become significant is particularly inappropriate – this is equivalent to high jumpers lowering the height of the bar when they have just failed to clear it with their one and only attempt.
A second complication is that sometimes when we make a statistical inference we find that we have made a mistake. Either we have said that there is a difference between means when there is not really a difference, or we have said that there is no difference when there is. When we say there is a difference between means when there is none – in other words, when we reject the null hypothesis when it is actually true – this is called a Type I error. The second type of error – where we do not reject the null hypothesis when it is false – is called a Type II error.

In the language that we have been using to deal with issues of statistical inference, these two types of error relate to the appropriateness of a researcher’s certainty. A Type I error occurs when researchers are wrong in saying that something is going on. A Type II error occurs when researchers are wrong in saying that something is not going on.

As you can probably see, these two types of error are actually interrelated. This is because as researchers seek to reduce the likelihood of making a Type I error (by reducing their alpha level and being prepared to tolerate less inferential uncertainty), their chances of making a Type II error are increased. In fact both errors arise because we make all-or-none decisions about information which is not clear cut but is based on probability. However, as we have been at pains to point out, if the data were clear cut we would not need to use statistics in the first place.

The chance of making these errors is something that we have to take very seriously. An acute problem arises where certain statistical tests or methodological circumstances are more
likely to lead to one type of error than the other. Because we are always liable to make some errors we need to know whether those errors will be of a particular kind so that we can correct them. Furthermore, some types of error are more undesirable than others in particular circumstances. For example, if clinical psychologists were testing the effectiveness of a therapy that has the possibility of harmful side-effects, they would want to avoid Type I errors because they would not want to expose clients to these risks if the therapy did not actually help.

We are more likely to make Type II errors when the sample size or the effect is small, when we are investigating effects of a small size, where there is a lot of random error, or where the alpha level is very small. It is worth looking at the terms in the equation for the t-statistic to see why this is the case. Type II errors can only occur when the obtained test statistic is smaller than the critical value of t—that is, when we do not reject the null hypothesis. Type II errors are therefore associated with factors that reduce the value of the calculated t-value and with those that increase the critical t-value. A small value of calculated t will be obtained when there is a small sample size, a large error term or a small information term (i.e., a small difference between means). A large critical value of t will be obtained when the alpha level (not the p-value) is very small.

When the likelihood of making a Type II error is inflated for one or more of these reasons, the t-test will be unlikely to reveal any relationships that may be present in the data. This is analogous to looking at stars with a very weak telescope. To avoid these problems we can use a more powerful test, in the same way that an astronomer would use a more powerful telescope. Powerful tests have a greater ability to reveal relationships that are present in the data. It is thus possible to distinguish between tests in terms of their power. Saying that a test is powerful is the same as saying that the test has a low chance of producing a Type II error. Formally, we can therefore define power as 1 − β, where β (the Greek letter beta) is the probability of a Type II error.

In general, we should use the most powerful test of a hypothesis that is appropriate for the data we are dealing with (a point we will return to later when we discuss the assumptions of statistical tests). A test that is not powerful is called a conservative test. In the example of testing the effectiveness of a therapy with harmful side-effects, we would probably wish to use a conservative test. However, in order to identify those side-effects in the first place, we should definitely use a powerful test. We can summarize the process of hypothesis testing as follows:

1. Decide on an alpha level before you commence the research. This is the amount of inferential uncertainty that you see as acceptable.
2. Calculate the t-statistic and obtain the corresponding probability value using SPSS.
3. If the probability value is less than the alpha level then reject the null hypothesis and conclude that the difference between means is statistically significant. If the probability value is greater than the alpha level you should conclude that the result is not statistically significant and should make no inferences about differences between the means.
When you report the results you should do so in the following way (assuming an alpha level of .05). To report a significant result:

The difference between means was significant, $t(18) = 2.90, p = .010$

To report a non-significant result:

The difference between means was not significant, $t(6) = 1.804, p = .121$

Obviously in these examples the values of $t$ and $p$ and the degrees of freedom vary. When doing your own research you need to use the values that you have calculated for the data you are working with.

Note here that you should report the $p$-value that corresponds to the highest $t$-value in the relevant row of the table that has been exceeded (or the exact $p$-value if, for example, it is available from a computer program). Some textbooks advise researchers simply to report whether the obtained $p$-value is less than the alpha level (e.g., with an $\alpha$ of .05 to report ‘$p < .05$’ even when $p$ is much smaller than .001). However, we advise against this practice simply because it reduces the amount of information available to the reader about the inference that is being made.

The whole process of using the hypothesis-testing approach to conduct $t$-tests is summarized in Online Appendix B. Examples of how to write these tests up formally are presented there, too.

---

Test Yourself 8.5**

Which of the following must be true of a statistically significant result of a $t$-test?

a. The probability that a difference at least as large as the observed difference could be produced by a particular random process will be less than the alpha level.

b. The obtained value of $t$ will exceed the alpha level.

c. The rejection regions must be significantly different from each other.

d. Both (a) and (c).

e. Alpha must be set at .05.

The correct answer is (a). A significant $t$-test yields a $p$-value that is less than a predetermined alpha level. Answer (b) is wrong because the value of $t$ must exceed the critical values. Answer (c) is wrong because the means are not significantly different from each other. Answer (d) is wrong because the rejection regions must be significantly different from each other. Answer (e) is wrong because alpha is a predetermined alpha level. Answer (a) is correct because it states that the probability of a difference at least as large as the observed difference will be less than the alpha level.
alpha level The largest probability that the researcher is prepared to accept in order to reject the null hypothesis. The probability of a random process producing a result as big as the one obtained has to be smaller than this for a statistical test to be significant. It is equal to the probability of making a Type I error and should be set before research is carried out.

alternative hypothesis The hypothesis that the research reveals an effect (e.g., that two samples are drawn from different populations and that there is an informative difference between their means). Also referred to as $H_1$. This is normally the hypothesis that the research was designed to test.

critical value The value associated with a statistical test (e.g., $t$) that cuts off the rejection region. If the obtained value exceeds this value then the test result will be significant. This is because the probability of the obtained value will be less than the alpha level.

directional test A one-tailed statistical test, where the alternative hypothesis specifies the direction of any effect (e.g., hypothesizing that one particular mean will be greater than another).

$H_0$ The null hypothesis that the research reveals no effect.

$H_1$ The alternative hypothesis that the research reveals an effect.

hypothesis-testing approach The process of making decisions as to whether the pattern of obtained results (e.g., differences between means) is due to chance ($H_0$) or to significant effects ($H_1$).

null hypothesis The hypothesis that the research reveals no effect (e.g., that two samples are drawn from the same population and that there is no informative difference between their means). Also referred to as $H_0$. This is the hypothesis that statistical procedures allow researchers to reject.

one-tailed test A directional test with a single rejection region. Such a test has a rejection region whose area is twice as large as that for a two-tailed test, and is based on the alternative hypothesis being directional (e.g., hypothesizing that one particular mean is larger than the other).

$p$-value The probability, as revealed by a statistical test, that a random process (involving taking random samples of a particular size from a particular population) could produce some outcome.

power The ability of a statistical test to reveal effects that exist in the data. Increasing power reduces inferential uncertainty. Formally, power is equal to $1 - \beta$, where $\beta$ is the probability of a Type II error.

rejection region The area of the sampling distribution associated with a statistical test (e.g., the $t$-distribution) that is cut off by the critical value(s). The size of these
regions depends upon the **alpha level**. For example, an alpha level of .05 corresponds to a rejection region (or regions) taking up 5% of the total distribution.

**significance-testing approach** Another name for the **hypothesis-testing approach**.

**statistical significance** An outcome where the probability that an effect at least as large as that observed could be produced by a random process is less than a predetermined **alpha level**. Where a result is statistically significant it is implausible that a random process could have produced the effect.

**two-tailed test** A test with two **rejection regions** that is based on the **alternative hypothesis** that research reveals an effect (e.g., that two means are different). The nature of any departure from randomness (e.g., which mean will be larger) is not specified.

**Type I error** Making the statistical inference that an effect exists (e.g., that there is a difference between means) when the effect is actually due to a random process. Here the **null hypothesis** is rejected when it is actually true.

**Type II error** Making the statistical inference that there is no effect (e.g., that there is no difference between means) when the effect does actually exist. Here the **null hypothesis** is not rejected when it is actually false.

### Other ways of handling the results of t-tests: Probability-level, confidence-interval and effect-size approaches

Despite its popularity, the hypothesis-testing or significance-testing approach is only one way to make decisions about inferential uncertainty. There are several alternatives that are preferred by statisticians who do not like the all-or-nothing nature of the hypothesis-testing approach. In this section we present three alternative methods for dealing with inferential uncertainty. All of them are acceptable in different areas of the psychological research literature. However, different instructors will have preferences about which one they want students to use in their work.

**The probability-level approach**

The **probability-level approach** is the technique that we presented in the sections earlier where we first explained how to do a t-test. This involves simply calculating and reporting both a t-value and an associated probability level. If this approach is taken, the results are not described as being significant or non-significant – or at least hard-and-fast significance levels are not used. Instead the researcher simply states the probability that a random process could have produced a difference as large as the one obtained. The disadvantage of this method is that it is a bit clumsier to express in words. One advantage is that other people can
look at your results and use another approach if they choose, so it is best where possible to provide an exact $p$-value (e.g., writing '$p = .014$' rather than '$p < .05$') even if you adopt the hypothesis-testing approach. If the $p$-value is very small (i.e., less than .001) it is best to write '$p < .001$', rather than '$p = .000$', just to capture the idea that the value is not exactly zero.

The confidence-interval approach

When we estimate parameters we can do one of two things. We can estimate a parameter by suggesting a particular value. This is called a point estimate. However, when we do this we still have a degree of uncertainty about the correspondence between this point estimate and the true parameter value. To deal with this uncertainty we could therefore specify a range of values around the point estimate within which we expect the true value to lie. Such a range of possible values is called a confidence interval (or sometimes an interval estimate). Confidence intervals can be constructed for a variety of parameters, including the mean and the difference between means.

When examining differences between means, the confidence-interval approach involves reporting a probable range of differences between population means. This is the likely range of differences that would be obtained by taking random samples from the same population and subtracting one mean from the other.

The size of the confidence interval depends on the amount of inferential uncertainty that researchers are prepared to accept and the amount of random error present. The less uncertainty they are prepared to live with, the larger the confidence interval will be. As a result, more and more of the possible differences between means that would be obtained from a random process will lie within the confidence interval.

The probability of including the true value in the confidence interval is used to name the confidence interval. A 95% confidence interval for the sample mean would contain the true population mean 95% of the time if many samples of the same size as the empirical sample were drawn randomly from that population. A 95% confidence interval for a difference between means would contain the true difference between means 95% of the time.

We can set the size of the confidence interval in SPSS by clicking on ‘Options...’ (Figure 8.16) and then setting the ‘Confidence Interval Percentage’ (Figure 8.17).

In the SPSS examples here we have confidence intervals for differences between means. These are obtained by taking the $t$-value that corresponds to the proportion of the sampling distribution that we wish to exclude. A 95% confidence interval excludes 5% of the possible differences between means so we want the $t$-value that corresponds to that proportion for the given number of degrees of freedom. We can look these up in a set of statistical tables, and for the example of the between-subjects $t$-test with $df = 18$, the critical value for a 95% confidence interval is 2.101. When we multiply this number by the standard error of the difference between the means we obtain the amount we have to add and subtract from the observed difference (i.e., 2) to create the confidence interval. This value is $2.1009 \times 0.68961 = 1.44881$. The lower bound (limit) of the 95% confidence interval is thus $2 - 1.48881 = 0.55119$ and the upper bound is $2 + 1.48881 = 3.48881$. 

8. Differences between Means
In other words, if we repeatedly took a sample of 10 people from one population and another sample of 10 people from a population with the same variance and a mean score two points higher we would expect (i.e., be confident) the obtained differences between the samples to be between (i.e., within the interval) 0.55119 and 3.48881 on 95% of occasions. We write this (rounded to two decimal places) as CI_{.95} = [0.55, 3.49].
8. Differences between Means

Why is this range of values useful? The answer is that knowing what to expect gives us some basis for deciding whether the results we obtained are remarkable or surprising. The confidence interval does not contain the value 0.0. That tells us that if two samples did come from the same population with the same mean then this is a remarkable result.

The last of these steps is contentious because many would argue that it simply reproduces the logic of the $t$-test (see Frick, 1995). This is because there is a simple relationship between these confidence intervals and $p$-values. A difference with a $p$-value of less than .05 will lie outside the 95% confidence interval, a difference with a $p$-value of less than .01 will lie outside the 99% confidence interval, and so on. For this reason confidence intervals can be seen as a means of rewriting statistical tests, because if a difference is significant with $p < .05$ then it will also lie outside the 95% confidence interval. This is because the $t$-test and the confidence interval involve the same three terms: the difference between the means, the standard error and the critical value. However, the two approaches use these terms in different ways. As Figure 8.18 illustrates, the confidence interval corresponds to the region of the $t$-distribution that is between the tails that would be cut off by the critical values. So, a 95% confidence interval is given by

![Figure 8.18 The $t$-distribution showing confidence interval](image)

Figure 8.18 The $t$-distribution showing confidence interval
Research Methods and Statistics in Psychology

confidence interval includes all the possible outcomes that would lead us not to reject the null hypothesis at an alpha level of .05.

For these data any difference between the means that is greater than 1.45 or less than −1.45 will lie outside the 95% confidence interval. Here the difference between means was 2.0, so this difference lies outside the 95% confidence interval. In other words, we can conclude that 95% of differences between means produced by a random process would not be this large.

Confidence intervals are a very good way of making the level of inferential uncertainty associated with a given test clear. By alerting the researcher to the range of possible outcomes associated with the process of doing the research, the calculation of a confidence interval puts results in perspective by comparing our obtained result directly with a range of possible results that would be produced by a random process. This avoids the trap that can occur when researchers using the hypothesis-testing approach lose sight of the idea that their results are being treated as one possible sample from a specified population and are being used to estimate parameters of that population. Nevertheless, most of the thorny issues associated with statistical inference are as evident in the use of confidence intervals as they are when using the hypothesis-testing approach.

However, one additional benefit of the confidence-interval approach is that even when an observed difference between means lies within a specified confidence interval, the confidence interval itself provides information about how large an effect would need to be in order to make us certain that something was going on in our research. This information can be very useful as it may give researchers insight into the ability of their tests, measures or manipulations to reveal effects. As we discussed in Chapter 4, this is an important methodological consideration.

The effect-size approach

One of the things about the approaches to the results of $t$-tests that we have discussed so far is that inferential uncertainty is reduced by increasing the sample size. This follows again from the law of large numbers. This means that we can do exactly the same study twice (where the second study is a replication) and obtain a significant difference in the first study and not in the second, simply because the second study has fewer participants – even though the differences between the means and the variances are the same in each case. So, the fact that a result is significant does not mean that it is large enough to be important or interesting. The important point here is that statistical significance is not the same as psychological significance. This is true partly because if the sample size is large enough then just about any difference can be significant. The mistaken belief that a statistically significant finding in a psychological study necessarily tells us something important about the psychological processes or states under investigation (i.e., that statistical significance implies psychological significance) can be referred to as the significance fallacy. This fallacy is made all the more potent because it is one to which researchers easily succumb – particularly if they are under pressure to make as much of their research as possible.

To address this problem many psychologists who use statistics prefer to base their interpretation of results on an effect-size approach. Effect sizes are not affected by the sample size and simply tell us whether the difference between the means is large compared
8. Differences between Means

with the amount of dispersion (or descriptive uncertainty). These effect sizes help us to compare results across studies.

A number of proposals have been made about how to measure effect size. One measure is $d$. This is essentially the difference between means divided by the standard deviation. Unlike the $t$-statistic, the standard deviation is not divided by the square root of the sample size, so the size of $d$ does not depend on the sample size. The measure of effect size that we will use here is $r$, largely because we use this statistic again in the next chapter. The formula for $r$ is

$$r = \sqrt{t^2 / (t^2 + df)}$$

This value of $r$ can vary between 0 and 1. And as with all measures of effect size, the value of $r$ can be compared across studies. It is very important to do this, but in areas where there is no prior research we can follow the ‘rule of thumb’ used by Cohen (1988): if $r$ is at least .1 it can be considered to be a weak effect, if it is at least .3 it is a moderate effect, and if it is at least .5 it is a strong effect. In the examples we used previously, the size of the effect for the between-subjects $t$-test in the ‘physical reinforcement theory’ experiment would be presented as follows:

$$r_{between} = \sqrt{t^2 / (t^2 + df)}$$
$$= \sqrt{2.90^2 / (2.90^2 + 18)}$$
$$= \sqrt{8.41 / 26.41}$$
$$= .56$$

This is a strong effect. In the within-subjects example used to examine the same theory, the effect size would look like this:

$$r_{within} = \sqrt{t^2 / (t^2 + df)}$$
$$= \sqrt{1.43^2 / (1.43^2 + 6)}$$
$$= \sqrt{2.04 / 8.04}$$
$$= .50$$

This too is a strong effect.

Our use of these examples makes it clear that the different approaches to statistical tests lead to different conclusions about the importance or interest of studies – remember that the hypothesis-testing approach would have suggested that the within-subjects test results yielded no effect (by any conventional choice of alpha level). Results can be highly statistically
RESEARCH BITE 8.2
When small effects are a big deal

When researchers use the effect-size approach to assess the impact that an independent variable has on a dependent variable there is a general assumption that bigger effects are better. Indeed, where the effect is small, readers may conclude that the experimental effect that has been demonstrated is trivial and unimpressive. However, in an influential paper, Deborah Prentice and Dale Miller (1992) argue that this need not be the case. In particular, this is because small effects can be very impressive if the manipulation of the IV was quite minimal or if the DV is hard to influence. They therefore argue that information about effect size always needs to be evaluated relative to the specifics of a particular experimental design.

Reference

Test Yourself 8.6**

Researchers conduct a t-test and obtain a p-value of .0012. Which of the following is an appropriate conclusion on the basis of the information provided?

a. The result is significant.

b. The effect size will be large.

c. Both (a) and (b).

d. A 99% confidence interval for the differences between the means will include the observed difference.

e. A 95% confidence interval for the differences between the means will not include the observed difference.

The correct answer is (e). We cannot say whether a result is significant or not unless an alpha level is specified.
8. Differences between Means

significant but of small size, or of large size but not significant. This is because effect sizes are not estimates of inferential uncertainty. They do not tell us whether the results could have been due to chance, they simply tell us the size of those results relative to error but ignoring sample size. To address the issue of inferential uncertainty, effect sizes need to be accompanied by other measures. Before taking the results of a piece of research that showed a large effect size too seriously, most researchers would want to be confident that the results were not a chance occurrence. In other words, effect sizes are very useful but they complement rather than replace the other approaches. Our advice is that the results of *t*-tests should always be accompanied by measures of effect size.

**confidence interval** An estimated range of values for a population parameter. The size of the confidence interval is expressed as a percentage. If sampling is random, a 95% confidence interval has a 95% chance of containing the population parameter.

**confidence-interval approach** An approach to inferential statistics based on the use of **confidence intervals** rather than **point estimates**.

**effect size** A measure of the size of an effect that ignores the sample size.

**effect-size approach** An approach to inferential statistics based on the calculation of **effect sizes** rather than probability levels.

**point estimate** A single value used to estimate a population parameter. For example, the sample standard deviation is a point estimate of the population standard deviation.

**probability-level approach** A procedure whereby the researcher calculates the probability that a random process could produce a difference as big as that observed. No hard-and-fast inferences about statistical significance are made.

**psychological significance** An outcome of research which produces important knowledge about psychological processes or states. This should not be confused with statistical significance.

**significance fallacy** The mistaken belief that a statistically significant result in a psychological study constitutes a **psychologically significant** finding.

**Some notes of caution**

All of the procedures described in this chapter are based on statistical models. Statistical models allow us to make inferences so long as the principles they are based on are applicable. These principles are called **assumptions**. Assumptions are the foundations upon which everything else we do with statistics is based. If the assumptions do not hold (where we say there is a **violation of assumptions**) then the process of doing inferential statistics can be very risky because we can make the wrong inferences.

Given that assumptions are so important, you may well ask why we are talking about them now, towards the end of the chapter. That is a very good question, and the answer is that
we have, in fact, been talking about assumptions all the time up to this point, without calling them assumptions. This may become clearer when we point out that the assumptions of the between-subjects *t*-test are that our samples are *independent* random samples from a population with a normal distribution, and that the standard deviations of both samples are estimates of the same underlying population standard deviation.

Let us go through these assumptions one at a time. There is only one new term that has been introduced here, and that is ‘independent’. **Independence** refers to things being separate and standing on their own. Events are independent if the occurrence of one has no effect on the probability of the other. Coin tosses are independent events. If we toss a coin and obtain a head then the chance of the next coin toss being a head is the same regardless of what has occurred in the past. This point is often misunderstood. One way in which it is misunderstood is in the *gambler’s fallacy* – the belief that luck has to change (see Question 7.7 at the end of the previous chapter). Samples are independent if the probability of one participant’s inclusion in the sample is not affected by the probability of any other person being sampled. In other words, the samples are selected randomly from the population.

Whenever researchers do anything to bias the selection procedure – for example, if they deliberately select participants to be in one particular group (rather than randomly allocating participants) or if they use the same participants twice in a between-subjects *t*-test – the

---

**RESEARCH BITE 8.3**

The skewing effects of neglecting skew

Many statistical tests rely on variables being normally distributed. However, many variables that psychologists study violate this assumption. Winnifred Louis, Ken Mavor, and Deborah Terry (2003) make the point that this is true in the case of extreme behaviours because strong norms mean that few people engage in these. When norms are strong and behaviours are skewed to be very common (e.g., wearing clothes) or very rare (e.g., joining a hate group), analyses which include norms as predictors of behaviour tend to produce misleading results. Specifically, they underestimate the norms’ importance relative to factors that are more normally distributed (e.g., personality). This can lead to faulty theoretical conclusions.

**Reference**

8. Differences between Means

assumption of independence is violated. Under these conditions \( t \)-tests can give misleading results. For within-subjects \( t \)-tests the assumption of independence is relaxed. This is for the obvious reason that if the same sample is used for both measures these samples cannot be independent. However, the participants that comprise the sample still need to be independent of each other.

The use of \( t \)-tests also assumes that the samples are drawn from a normally distributed population. This is known as the assumption of normality. This is a very important idea in those areas of psychological statistics that rely on normal theory tests. We will provide no detailed justification for this idea here, but just point out that the mathematical solutions to statistical problems become much easier if we are able to assume that the variables are normally distributed. It is easy to see why this would be the case given the advantages of a normal distribution that we have already identified. If we know that a variable is normally distributed and we know its mean and variance, we know a great deal about it. Not surprisingly, it is easier to develop statistical tests for variables that are well understood in this way.

Finally, a third assumption is that the samples must both have standard deviations that are estimates of the same population standard deviation. This is called the assumption of equal variance. Where this is not true we cannot be sure that our procedures for deciding the level of inferential uncertainty will be accurate. If the standard deviations are not equal then we should treat the results of the \( t \)-test with caution. In fact, there are correction procedures that can be used under these circumstances involving the transformation of variables (see Chapter 6).

When we do between-subjects \( t \)-tests, SPSS provides Levene’s test of the equality of variance as shown in Figure 8.19. If the probability (marked ‘\( \text{Sig.} \)’) is very small it is worth considering using the version of the \( t \)-test that does not assume equal variances in the line below.

![Figure 8.19](image-url)

As we mentioned in Chapter 6, one thing that can contribute a great deal to the amount of variance in a distribution is the presence of one or more outliers. This means that if we have outliers in one sample and not in the other, the assumption of equal variance is likely to be violated. The effect of this will normally be to reduce the power of the test. The only way to check whether outliers could be affecting your results is to look at your data. This could involve constructing a frequency graph.

For example, the bar charts for the between-subjects \( t \)-test that we used as an example earlier are shown in Figure 8.20. If instead the distribution of data from the experimental condition had been like that in Figure 8.21 (i.e., if the response of Participant 1 had been ‘1’ rather than ‘4’),
Figure 8.20  Bar charts for the between-subjects t-test carried out above (see p. 209): (a) control condition; (b) experimental condition
we would have evidence of an outlier. The presence of this outlier increases the standard deviation of the data in this condition from 1.65 to 2.28. As a result, the assumption of equal variance could be violated and, if it were, the power of a *t*-test would be dramatically reduced. That is because a larger standard deviation makes it more difficult to demonstrate differences between the means. Outliers like this can be present for all sorts of reasons, including mistakes in recording the data. As we noted in Chapter 6, outliers can sometimes be corrected by appropriate statistical transformation, but sometimes it is better to remove or recode individual outliers.

Another point that we glossed over earlier concerns the advantages that we get from using the various types of tests. The within-subjects *t*-test is especially useful because it is a very powerful test: it is good at finding differences where they exist. The power of this test comes from using the same people on two different measures. More specifically, because each person is used as their own baseline for assessing the difference, this rules out a lot of the random variation that comes from using different people. This has the effect of reducing the amount of error so that any information is more likely to be revealed. However, if the correlation between the variables is less than .3 (see Chapter 9) then the power advantage of the within-subjects *t*-test disappears.

Figure 8.21  Bar chart showing evidence of outlier
Researchers conduct a t-test to compare two groups and find that one of the groups has a much larger standard deviation than the other. Which of the following statements is true?

a. The variances are robust.
b. The assumption of equal variance may have been violated.
c. The researchers should make sure that their distributions are free of parameters.
d. The standard deviations are not normal.
e. Both (b) and (d).

The correct answer is (b). Answers (a), (c) and (d) are all nonsense answers involving mixing up other terms from this section. In particular, with respect to (a), tests can be robust, not variances, and with respect to (d), distributions are normal, not standard deviations. If (d) is wrong, then (e) must also be wrong.

Where the assumptions of the t-test do not hold we must proceed with caution. By far the best option is to use non-parametric and distribution-free tests. These do not make the same restrictive assumptions as the t-test and we will consider them in detail in Chapter 11. There is a tendency in psychology to assume that statistical tests are robust to violations of assumptions (particularly the assumption of equal variance). However, these assumptions are considerably less robust than is commonly believed. In general, it is highly desirable to look for violations of assumptions in your data – for example, by using a frequency graph to explore its distribution. One of the major benefits of this inspection process is that problems or other interesting features of the data are likely to reveal themselves. There is little to lose and much to gain from looking carefully at data before subjecting them to statistical tests.

**assumption of equal variance** An assumption of the t-test procedure that the variances of the samples that are compared are both estimates of the same population variance and thus are similar.

**assumption of independence** An assumption of the t-test procedure that sampling of the underlying population is independent and can be treated as random.

**assumption of normality** An assumption of the t-test procedure that the samples are random samples drawn from a normally distributed population.

**assumptions** The conditions that must be satisfied before we can use a statistical test with confidence. The assumptions of the t-test are that the samples are independent and randomly sampled from a normally distributed population that has the same variance for all samples.
distribution-free tests  Tests that make no assumptions about the theoretical distribution of variables in the population from which a sample is drawn (i.e., the sampling distribution).

gambler’s fallacy  The mistaken idea that random events are not independent. A gambler may believe that a long run of good or bad luck has to change. The gambler’s fallacy arises from a misunderstanding of the law of large numbers. The idea that a random process will behave in a predictable way on average over a long run of observations can be misunderstood to imply that there is ‘a law of averages’ that serves to change the probability of random events based on past events. However, coins, dice and other things that generate random outcomes do not have memories.

independence  A term with numerous meanings in statistics. Two events are independent where the probability of one occurring does not depend upon the probability of the other occurring.

non-parametric tests  Generally similar to distribution-free tests, but, technically speaking, these tests do not involve hypotheses related to a population parameter.

normal theory tests  Statistical tests, such as the t-test, that assume that samples are drawn from a population that is normally distributed.

robust test  A statistical test that tends to yield valid results even when the assumptions underpinning that test are violated.

violation of assumptions  Threats to interpretation of the results of statistical tests that arise when the conditions required for the test to work properly do not hold.

Overview

This chapter began with a consideration of how to compare means when the population standard deviation is unknown. We covered three situations where the t-test can be used: for testing the difference between the mean of a single sample and some hypothesized population mean; for testing the difference between the means for a single sample on two different variables; and for testing the difference between two samples on a single variable.

In each case we find a value for t that corresponds to a probability that the results are due to random error. This probability corresponds to the level of inferential uncertainty. However, once we have obtained that probability we have some additional decisions to make.

Having found a t-value and a probability level, there are several choices as to how to treat, interpret and use these. The traditional approach has been to perform significance tests. To understand much of the literature that has already been published in psychology you will need to have some familiarity with this approach. Nowadays, however, important alternatives are available. These techniques include the use of confidence intervals and effect sizes.
Research Methods and Statistics in Psychology

### CHECKLIST
Revisiting the key goals for this chapter

- I understand how to conduct a *t*-test – a statistical procedure that allows me to establish how likely it is that differences between two means of the size we have observed could be produced by a random process (drawing random samples of specified size from a specified population).
- I know how different types of *t*-test are used to examine different types of data.
- I understand the process of interpreting the results of *t*-tests, and the different ways that this can be done.

### Further reading
Wilkinson and colleagues (1999) summarize the findings of the American Psychological Association’s Task Force on Statistical Inference, and we recommend that you read this (it can be found online at: www.apa.org/science/leadership/bsa/statistical/tfsi-followup-report.pdf). Smithson’s (2002) book also provides a detailed explanation of various alternatives to the hypothesis-testing approach, focusing in particular on confidence intervals and effect sizes.


### *t*-tests: A checklist for research evaluation and improvement

**Table 8.5**

<table>
<thead>
<tr>
<th>Potential problem</th>
<th>Question to ask</th>
<th>Potential improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferential uncertainty about an individual sample mean</td>
<td>Do the researchers want to compare a sample mean with another score when the population mean and standard deviation are unknown?</td>
<td>Subtract the comparison score from the sample mean and divide by the standard error of the mean (the sample standard deviation divided by the square root of the sample size). This procedure creates a <em>t</em>-score and – having calculated the degrees of freedom (the sample size minus one) – statistical tables or a computer program such as SPSS can be used to provide information about the probability of observing by chance a <em>t</em>-score as large as the one obtained.</td>
</tr>
</tbody>
</table>
### Differences between Means

<table>
<thead>
<tr>
<th>Potential problem</th>
<th>Question to ask</th>
<th>Potential improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferential uncertainty about means from the same sample</td>
<td>Do the researchers want to compare means associated with data that have been obtained from the same sample on two occasions when the population mean and standard deviation are unknown?</td>
<td>Following the procedures described in this chapter, use a statistical package such as SPSS to conduct a within-subjects <em>t</em>-test. This will generate a <em>t</em>-score and provide information about the probability (<em>p</em>) of observing by chance a <em>t</em>-score as large as the one obtained.</td>
</tr>
<tr>
<td>Inferential uncertainty about means from different samples</td>
<td>Do the researchers want to compare means associated with data that have been obtained from two independent samples when the population mean and standard deviation are unknown?</td>
<td>Following the procedures described in this chapter, use a statistical package such as SPSS to conduct a between-subjects <em>t</em>-test. This will generate a <em>t</em>-score and provide information about the probability (<em>p</em>) of observing by chance a <em>t</em>-value as large as the one obtained.</td>
</tr>
<tr>
<td>Significance fallacy</td>
<td>If the researchers are using a hypothesis-testing approach to <em>t</em>-tests, have they distinguished appropriately between psychological and statistical significance?</td>
<td>Consider significance levels in conjunction with effect sizes (e.g., compare <em>p</em> and <em>r</em>). Be alert to the dangers of making too much of statistically significant <em>t</em>-values when effect sizes are low.</td>
</tr>
<tr>
<td>Violation of the assumption of independence</td>
<td>Do the means that are being compared relate to independent sets of observations?</td>
<td>If sample scores are not independent (and dependent scores are not being compared in a within-subjects <em>t</em>-test) it can be misleading to compare them using a <em>t</em>-test (or any other statistical test) if, in effect, the same piece of data is being counted more than once. Accordingly, if possible, any source of dependence (e.g., the fact that the same participant responded twice) should be removed before collecting data. If this is not possible, the nature of the violation should be noted and results treated with caution.</td>
</tr>
<tr>
<td>Violation of the assumption of normality</td>
<td>Do the means that are being compared relate to observations that are normally distributed?</td>
<td>If the distribution of sample scores is non-normal (e.g., severely skewed or U-shaped) and the cell sizes are unequal, or there are reasons to believe that the mean is a misleading indicator of central tendency, you should seriously consider using a non-parametric test (Chapter 11) or correct the distribution by transformation (see Chapter 6).</td>
</tr>
</tbody>
</table>

(Continued)
### Potential Problem Question to Ask Potential Improvement

<table>
<thead>
<tr>
<th>Potential problem</th>
<th>Question to ask</th>
<th>Potential improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violation of the assumption of equal variance</td>
<td>Do the means that are being compared relate to samples that have equal variance?</td>
<td>If sample variances are unequal (e.g., one is more than four times bigger than the other) you should seriously consider using a non-parametric test (see Chapter 11). This is only worth doing if the size of one sample is much larger than the other.</td>
</tr>
<tr>
<td>Outliers</td>
<td>Is the distribution of scores (especially its shape and variance) dramatically affected by the presence of outliers?</td>
<td>Examine the distribution of scores to see if it contains outliers. If it does, see if this can be corrected by an appropriate statistical transformation of the sample data. If this is not possible, consider (a) removing individual outliers or (b) recoding them to the same value as the nearest non-outlier. Ensure you report any transformations that you make.</td>
</tr>
</tbody>
</table>

---

### Test Yourself 8.8*

If an experimenter were to conduct an experiment in which participants were randomly assigned to either a control condition or an experimental condition, which of the following statements would be true?

- a. It will be appropriate to analyse results using a between-subjects *t*-test.
- b. Any statistical analysis will be based on pairs of responses.
- c. If a *t*-test is performed to analyse the results, it will have *n* − 1 degrees of freedom.
- d. The experiment involves two related samples.
- e. Both (a) and (c).

### Test Yourself 8.9**

John, a second-year psychology student, is using the hypothesis-testing approach and an alpha level of .05 to examine a difference between two means. He discovers that this difference is associated with a *t*-value of 3.46. If the critical *t*-value with α = .05 is 2.056 what should he conclude?

- a. That the difference between the means is statistically significant.
- b. That the alpha level is too high.
- c. That the alpha level is not high enough.
- d. That the experiment did not contain enough participants to draw a strong conclusion.
- e. That no conclusion can be made about the nature of the underlying populations.
8. Differences between Means

Test Yourself 8.10**

Which of the following suggests that the assumptions underlying a between-subjects $t$-test have been violated?

a. Evidence that the dependent variable is normally distributed.
b. Evidence that the samples being compared have unequal variances.
c. Evidence that the manipulation of the independent variable had no effect.
d. Evidence that sampling was random, and that scores were independent.
e. None of the above.

Test Yourself 8.11***

Which of the following increases the likelihood of a Type II error when conducting a $t$-test?

a. A high alpha level.
b. A large sample size.
c. High power.
d. Low random error.
e. A small difference between means.
Research Methods and Statistics in Psychology

Discussion/essay questions

a. Why is the hypothesis-testing approach to statistical inference controversial?

b. Has a reliance on hypothesis-testing techniques had a detrimental impact on the study of psychology?

c. Discuss the relative advantages and disadvantages of a confidence-interval approach to statistical inference.

d. In what way does the presence of outliers affect the results of t-tests?

Exercises

a. An experimenter conducts a study in which participants try to remember a number of digits. Some participants do this after consuming a litre of full-strength beer, others do it after drinking a litre of full-cream milk. The data from the study were as follows:

i. What form of t-test is appropriate to analyse these data?

ii. What value of t does the appropriate test yield?

iii. Is this effect statistically significant with $\alpha$ set at .05?

b. Imagine that the data in Table 8.6 were obtained from examination of the same participants under different conditions (i.e., that the researcher collected pairs of data from each participant).

Table 8.6

<table>
<thead>
<tr>
<th>Participant number</th>
<th>Milk</th>
<th>Beer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>
8. Differences between Means

<table>
<thead>
<tr>
<th>i.</th>
<th>What form of $t$-test is appropriate to analyse the data?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ii.</td>
<td>What value of $t$ does the appropriate test yield?</td>
</tr>
<tr>
<td>iii.</td>
<td>Is this effect statistically significant with $\alpha$ set at .05?</td>
</tr>
</tbody>
</table>

| c. | A parapsychologist wants to know whether a person claiming to have psychokinetic powers can produce higher numbers throwing red dice than throwing blue dice. The results are analysed by $t$-test. Why is it more appropriate to test this hypothesis by comparing scores obtained when the person throws many dice at the same time and adds them up, rather than just comparing the scores when a single die is thrown? [Hint: What shape would the distributions of scores be in each case?] |

| d. | A psychotherapist consults you (as an expert psychological statistician) to see whether an expensive new therapy she has developed is working and whether it has any side-effects. For this purpose she provides you with two sets of scores for 200 patients. One set of scores measures psychological functioning pre- and post-treatment and the other measures the presence of self-reported side-effects pre- and post-treatment. |

| i. | What form of $t$-test is appropriate to analyse these data? |
| ii. | What alpha levels would you recommend? |
| iii. | In the course of your discussion with the therapist, you find out that the therapy was administered in groups. Why does this make it problematic to analyse the data using a $t$-test? |

| e. | A personality researcher asks you whether 10 students in his class have a mean score significantly greater than the population norm on a standard personality test (with alpha set at .05). The population norm on the test is 100, and the students’ scores are 111, 101, 94, 131, 120, 98, 114, 113, 132 and 109. What answer would you give him? |