Using the t-Test

This guide has been adapted from an undergraduate topic on Attention and the brain, part of a Level 2 module on Practical science (S288). You will find out why statistical tests are important for assessing data gathered through experimental research, and also the limitations of such methods.

1 Beyond reasonable doubt

If you have been carrying out investigations you may have been studying differences between conditions or groups. However, a simple difference in experiment performance results between two conditions is not enough by itself to draw conclusions. The question you need to ask is whether those differences are reliable. That is; they are likely to occur again if the experiment was repeated.

Anyone who plays or watches sport, or listens to live music, knows that an individual's performance varies day by day. Similarly, in an experiment, the performance of participants (human or otherwise) varies each time the experimental task is run, and this occurs irrespective of any change in the independent variable. There are also profound individual differences between participants purely in terms of their basic abilities on experimental tasks. This is the case in animals as well as humans. Thus, even where participants are allocated to conditions randomly (if they can be) in an experiment, and tested in the absence of the independent variable, it would be unreasonable to expect exactly the same scores on the dependent variable by the participants in each condition.

What this means in practice is that the scores of participants in one condition will naturally differ from the scores of participants in another condition, irrespective of any effect of the independent variable.

Let's consider an example. Your duration of sleep is probably different from one night to the next. This random or chance fluctuation in performance goes on continuously. Now imagine you are interested in determining the effect on sleep duration of having a malted drink at bedtime. For your experimental results to be meaningful you would need to distinguish between any difference in the duration of your sleep (the dependent variable) that was due to the malted drink (the independent variable), and any variation that was due to chance fluctuation.

It always happens that we assess the effects of the independent variable against a background of inherent variation in performance of the dependent variable. This is the 'doubt' in the title to this section: that maybe the independent variable had no effect whatsoever on the dependent variable, and the results of the experiment are just chance fluctuation. To get beyond this element of doubt and to demonstrate a relationship between two events, the effect of the independent variable on the dependent variable has to be over and above chance fluctuation. This is where statistical analysis comes in.

This guide briefly discusses the idea of how all variation in data can be observed and analysed statistically in order to gauge the independent variable's effect on the dependent variable. It then focuses specifically on a statistical test called a '*t*-test', which gauges whether differences in the dependent variable for two conditions are over and above that which would be expected due to natural random fluctuation, and shows you how to perform such an analysis using some calculative software available for this unit.

2 Examining distribution of experimental data

There is no great need to understand in detail the mathematics behind any particular statistical test, just as there is no need to understand how a word processor works, so long as you can use it appropriately. However, it is helpful to be able to visualise what the statistical tests are doing, and that is the purpose of this section.

Graphical interpretation of frequency data

A useful starting point for thinking about statistical analysis and *t*-tests is what we call the 'normal distribution'. To illustrate this idea, imagine that 2100 adults were asked how many hours they slept each night. The results are shown in Figure 1. The variable measured (sleep duration) is plotted on the horizontal axis. The number of people stating a particular sleep duration, or, put another way, the frequency with which a particular sleep duration was stated, is plotted on the vertical axis. The resulting plot is a therefore known as a frequency distribution graph.

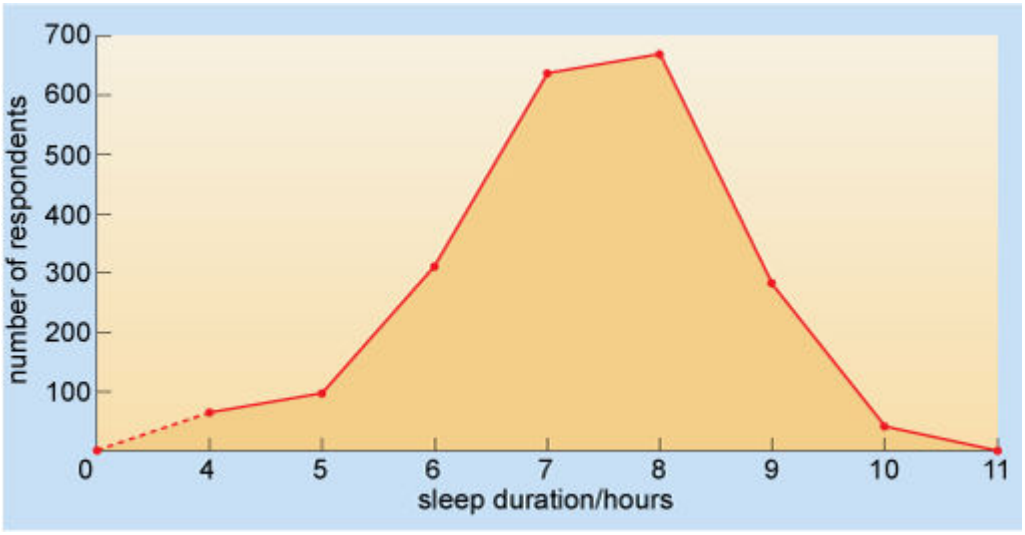

Figure 1 Daily sleep duration in hours as reported by a sample of 2100 adults

The graph shows clearly what you might have predicted: most people sleep between seven and eight hours each night; very few people sleep for less than four hours or more than ten hours. The shape of this graph, with few observations at either extreme of the sleep-duration scale and the bulk of observations in the middle of the range of results, is characteristic of a normal distribution.

The classic normal distribution is shown in Figure 2 for comparison. Note in particular that it is symmetrical about the midpoint on the horizontal axis (the vertical dashed line on the graph). The midpoint is the mean or average value of hours that participants reported sleeping for.

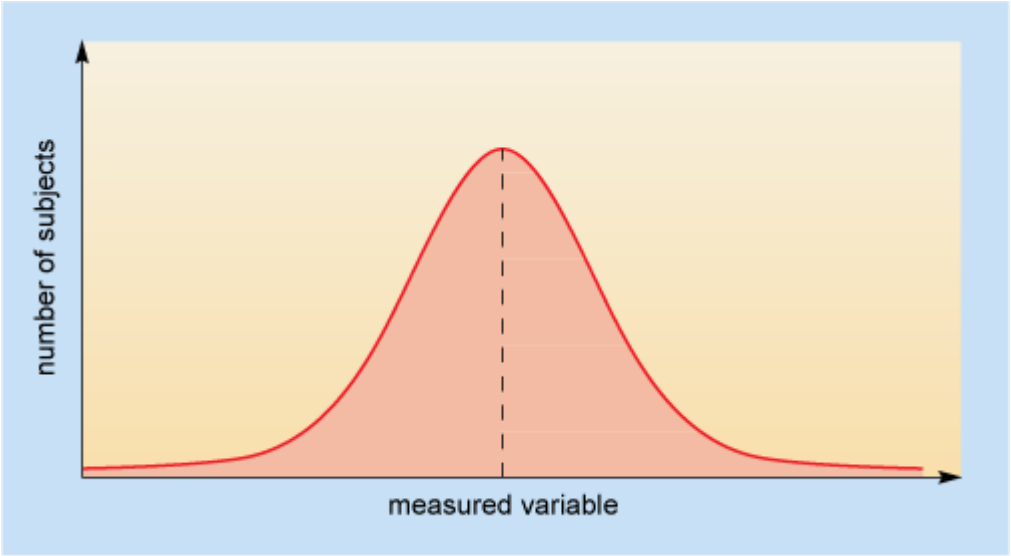

Figure 2 The classic normal distribution

This normal distribution can be used in two main ways. First, it can be used to identify and define observed values that are extreme or abnormal.

- Question: Which mathematical way of describing the spread of values should be used in conjunction with the mean?

- Answer: The standard deviation

In a normal distribution, 95% of the observed values lie within 1.96 standard deviations on either side of the mean. (For simplicity, 1.96 is usually rounded up to 2.) Put another way, this means that only 5% of the observed values are more than two standard deviations above or below the mean.

- Question: Why has the phrase ‘above or below’ been included in the previous sentence?

- Answer: Because the normal distribution is symmetrical, extreme values can be either higher or lower than the mean.

- Question: What percentage of observed values are two standard deviations below the mean?

- Answer: There are 2.5% of observed values two standard deviations below the mean. Likewise there are 2.5% of observed values two standard deviations above the mean, giving a total of 5% above or below that value overall.

Any value that is above or below the mean by more than two standard deviations is, by convention, defined as being abnormal.

The normal distribution can be used to compare data from two conditions. It is this comparison that is at the heart of a number of statistical tests, often referred to as parametric tests. The *t*-test is one such parametric test.

For example, imagine three experiments, 1, 2 and 3, each consisting of two conditions, A and B, in which reaction time is measured in 2000 participants. The results from each experiment are plotted in Figures 3a, b and c.

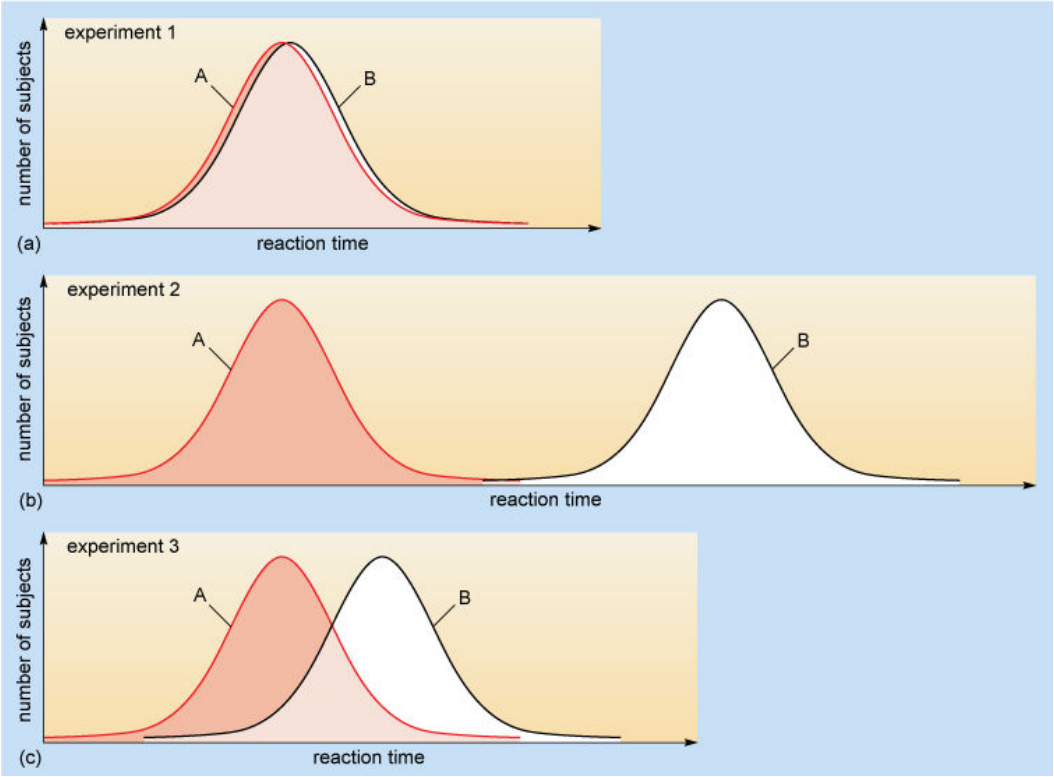

Figure 3 Hypothetical results from Experiments 1, 2 and 3

In Experiment 1 the data from participants in Condition B virtually match those from Condition A; the frequency distributions almost totally overlap (Figure 3a). It is obvious that there is virtually no difference between the conditions; the independent variable therefore had no obvious effect.

In Experiment 2, the data from participants in Condition B are totally different from those of participants in Condition A; the frequency distributions do not overlap at all (Figure 3b). Clearly, there is a difference between the conditions; the independent variable therefore had a very marked effect on the performance of participants in Condition B.

While results do occasionally fit the patterns shown in Experiments 1 and 2, the usual pattern is that depicted by Experiment 3. Here, there is overlap between the data from the two conditions (Figure 3c). Condition B data appear to have a slightly higher mean value for the dependent variable, but the effect of the independent variable on the performance of the participants is small. The question is, are the data from the two conditions really different because of the effect of the

independent variable in the experiment, or is the difference just due to chance variation?

Working with data on a smaller scale

Now that you can visualise the problem of distinguishing between results due to chance variation and those due to a real effect of the independent variable, we can consider the more usual situation in experiments, where the sample is not thousands, but perhaps just tens of participants. In this situation there are relatively few data points; and it simply is not possible to plot a meaningful frequency distribution graph. This does not matter, however. As long as it is sensible to calculate a mean and a standard deviation from the data then that defines what the frequency distribution would look like, were you to plot it, and is also sufficient for statistical purposes.

Consider the data in Table 1, which derives from an experiment in which positive air pressure was used to affect the duration of deep sleep. You do not have to concern yourself with details of the experiment from which the data were derived, other than the fact there were two conditions: C1 in which treatment was not given, and C2 in which air-pressure treatment was given.

- Question: Which is the control, and which the experimental condition in this experiment?

- Answer: C1 is the control condition and C2 is the experimental condition.

Table 1 Duration of sleep (in minutes) of participants with breathing difficulties

Participant rank	Condition C1 deep sleep/min	Participant rank	Condition C2 deep sleep/min
1	1.8	1	1.9
2	1.9	2	3.1
3	2.1	3	3.6
4	2.2	4	4.7
5	2.4	5	5.2
6	2.4	6	5.9
7	2.6	7	6.4
8	2.6	8	9.8
9	2.6	9	11.3
10	3.2	10	11.4
11	3.3	11	11.9
12	3.5	12	12.5
13	3.8	13	16.4
14	4.0	14	16.5

Participant rank	Condition C1 deep sleep/min	Participant rank	Condition C2 deep sleep/min
15	4.1	15	16.8
16	5.1	16	17.7
17	5.7	17	19.4
18	6.1	18	24.6
19	8.9	19	28.4
20	9.1	20	41.1
21	9.2	21	43.7
22	10.3	22	46.4
23	10.6	23	52.9
24	12.1	24	73.2
25	13.3	25	125.0
26	42.5	26	158.2
27	59.6	27	174.3

You could suggest the following from looking at these data:

- Some participants in condition C1 – who had no treatment – spent longer in deep sleep than many participants in condition C2 – who did receive the treatment. Participant 22 (C1) spent longer in deep sleep than Participants 1, 2, 3, 4, 5, 6, 7 and 8 (C2) for instance.
- The shortest period of deep sleep in the two conditions was very similar: 1.8 minutes versus 1.9 minutes.
- More participants in C2 spent longer in deep sleep. Only six participants in C1 exceeded ten minutes of deep sleep, compared with 19 participants in C2.

The null hypothesis for this experiment would predict no significant difference in deep-sleep duration between those participants in the experimental and control conditions. In other words, the difference in the overall range of deep-sleep duration across the two conditions could have arisen by chance. Statistical tests provide a way of assessing how likely this is.

3 Probability in statistics

Statistical tests are a mathematical way of calculating precisely how likely it is that a particular experimental result arose by chance. Two general points are very important when considering such tests and you should bear them in mind.

- First, statistics never produce definite answers; they deal with probability rather than certainty. Indeed, you will never know whether a particular set of results actually arose by chance or not; you will only know the probability that they arose by chance.

- Second, statistical logic asks how likely it is that the results arose by chance, or, more perversely, how likely it is that there is no effect of the independent variable.

In other words, statistical inference is based on the null hypothesis, not the experimental hypothesis. Specifically, this kind of statistical testing works out the likelihood of your result if the null hypothesis is true (i.e. if there is no effect). We then simply define reliable effects as those that are sufficiently unlikely if the null hypothesis is true.

Consider the hypothetical sleep experiment mentioned in Section 1, which was interested in investigating whether a bedtime malted drink affects the duration of sleep. Now, to think about what the statistics are going to do, suppose the drink has no actual effect on the test condition (i.e. those that did have it before bed).

- Question: What kind of result would you expect from the two conditions in this experiment if this was the case?

- Answer: You would expect the results under the two conditions to be very similar. That is; the frequency distributions from each condition would look almost identical.

If the actual data in the two conditions are similar, the results could well have arisen by chance, and the null hypothesis is accepted. If the actual data in the two conditions are dissimilar (i.e. the result probably did not arise by chance) then you can reject the null hypothesis. In order to infer how similar or dissimilar the results from the two conditions are you must calculate the probability that the results arose by chance.

Expressing probabilities in statistical tests

Probabilities in statistical tests are expressed as '*p* values'. These give the probability that the data put into the statistical test arose by chance. A *p* value of 1 means that it is an absolute certainty that the results arose by chance, and a *p* value of 0 means that it is impossible that the results arose by chance.

In practice, *p* values of 1 or 0 are very rare, but the nearer the *p* value is to 0, the less likely the results are to have arisen by chance. For example, if $p = 0.001$ then there is only a one-in-one-thousand probability that the results arose by chance. Such a result would mean the null hypothesis could be rejected, leaving the experimental hypothesis to explain the data.

In many sciences it is customary to say that *p* has to be less than 0.05 before the null hypothesis can be rejected. When $p = 0.05$, there is a one-in-twenty probability that the results arose by chance, and when the value is less than this the result is said to be statistically reliable (or statistically significant).

There is no mathematical or logical reason why 0.05 is chosen as the cut-off point for rejecting a null hypothesis, and indeed some experimenters prefer the value of 0.01. Historically, though, $p < 0.05$ has come to be accepted as a reasonable

threshold level of reliability. Actual levels of probability are usually quoted in research papers, allowing the reader to apply their own level of reliability in evaluating the results.

3.1 Limitations of statistical analysis

So far we have identified a number of ways that statistical analysis can help you to determine the likelihood that variation in a dependent variable is caused by the effect of the independent variable rather than by chance. This is obviously a very useful thing to be able to make a judgement on, but statistical tests are not without their limitations, too.

One thing to remember about your analyses is that statistical procedures only deal with the numbers fed into them: they do not know the meaning of the numbers. A statistical test will churn around any set of numbers fed in using a suitable format, and the delivery of a statistic at the end of this process does not sanctify the data. Just because the data were analysed does not mean that either the analysis itself, or its outcome, was meaningful.

Also, bear in mind that rejecting the null hypothesis does not automatically mean that the results went in the direction that you predicted (for example that one group performed better than the other one). Therefore you still need to spend time looking at your data to think about what the results mean. Get into the habit of describing your data carefully (was the mean value for one group larger than that for the other?) and thinking about what the outcome of your analyses might mean in terms of the original question your experiment set out to answer. A good experimental report will always spend lots of time considering what the data show in relation to the original experimental question.

You should now work though the guide provided below on The Student's t-test and how to use the accompanying software.

4 The Student's *t*-test

The Student's *t*-test is a robust, well-documented test that compares the means of two sets of data. The end result of the test is a single value of '*t*', which is a measure of the extent to which the two sets of data overlap. Remember, where the data extensively overlap, the independent variable had little effect on the participants. This is the case if *t* is small. If *t* is large, the two sets of data only partially overlap which means that the independent variable had a (mathematically) noticeable effect on the participants. Put another way:

- If *t* is small, the null hypothesis (of no difference between the participants' performance in the two conditions) is accepted.
- If *t* is large, the null hypothesis is rejected.

Specific values of *t* are converted into a probability, so that the rather vague phrases about extensive and partial overlap used above become actual numbers, albeit still

only probabilities. The computation of *t* is usually left to computer software packages; one such multimedia package is the *t*-test calculator available in The OpenScience Lab.

Requirements for doing a *t*-test

One value that must be known before you can apply a statistical test is the number of degrees of freedom. Degrees of freedom is an important, if elusive, mathematical term and usually its numerical value is one less than the number of participants in each condition and this is the method of calculation you will use here.

Degrees of freedom are a measure of how many items in a set of data need to be specified before all the items in that set are known. For example if, in a two-participant trial, the mean score is 25, and you know that Participant 2 scored 18, then Participant 1 must have scored 32 because no other value would give a mean for the two participants of 25. This trial therefore has 1 degree of freedom (2 – 1 = 1). In the case of an experiment that has two groups and the participants in each group are different, the degrees of freedom are calculated for each group as reported above, and then added together. As an example, consider an experiment which has seven participants in Group A and eight participants in Group B. In this case the degrees of freedom would be (7–1) + (8–1) = 13.

When running the *t*-test calculator software the results that need to be reported are the degrees of freedom, the value of *t*, and the *p* value calculated by the test. Here is a fictitious example. Let's imagine an experiment had 24 participants in each group. The degrees of freedom are therefore (24–1) + (24–1) = 46. The researcher calculates a *t* value from the data and the value returned by the calculator is 2.36. They also deduce a *p* value of 0.003. The researcher should then report these results in either of the following formats:

$t(46) = 2.36, p = 0.003$

or

$t(46) = 2.36, p<0.05$

The number in brackets is the degrees of freedom in the investigation, the value after the first equal sign is the value of *t* and the number after the second equal sign is the value of *p*. In the next section we discuss specifically how to use the Student's *t*-test calculator provided in The OpenScience Laboratory in order to obtain such values for your investigations. Note that you may need to run several different *t*-tests in an investigation; one for each measure that you have chosen to record.

4.1 Using the *t*-test calculator

The *t*-test calculator that you will need to use can be accessed from the following link:

- Click here to access the t-test calculator software

It can also be accessed from The OpenScience Laboratory *www.opensciencelab.ac.uk*. Below are some screenshots taken from the *t*-test calculator.

The opening screen tells you that it is possible to use the *t*-test calculator:

- If you want to test for differences in population means
- Provided that your measurements are at the interval level
- Provided that your measurements can be assumed to come from normally distributed populations.

You should first check if your data meets these requirements before proceeding to entering your experimental data by clicking on the 'Data' tab. Figure 4 shows you how to do this.

Figure 4 Entering data in to the *t*-test calculator

Once you have entered your data you need to click on the 't test' tab and note your degrees of freedom. Recall that these are calculated as one less than the number of

participants in each condition. So if you had eight participants in one condition and seven in another your degrees of freedom would be 13 (as shown on the screen shot example in Figure 5).

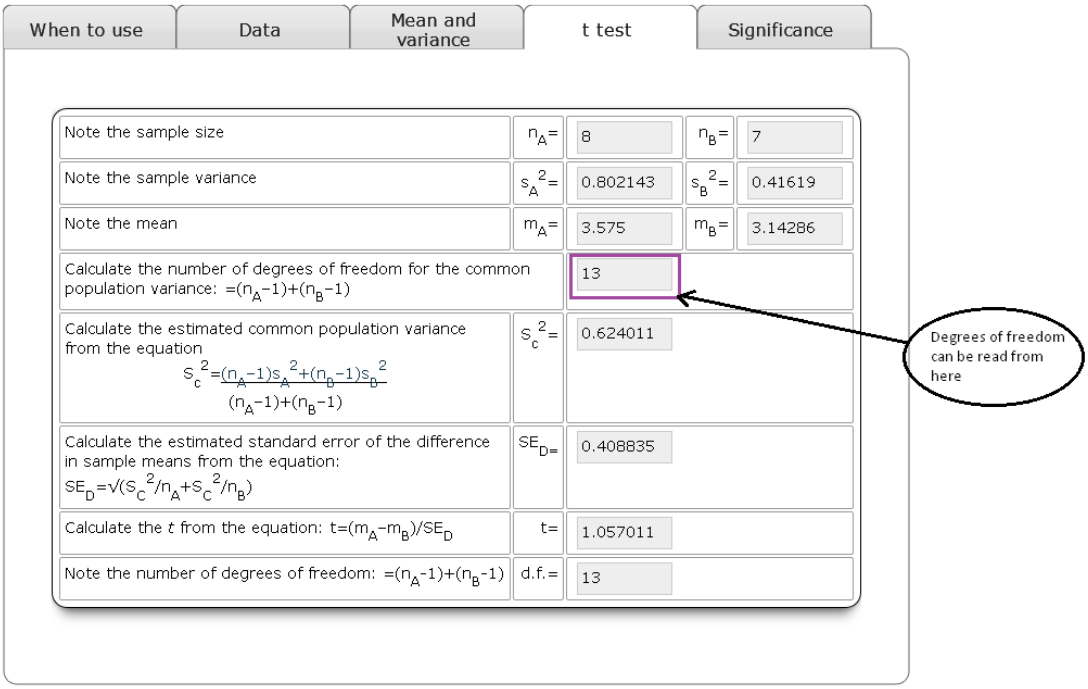

Figure 5 Recording your degrees of freedom

Next you need to decide what your critical value of *p* will be, and this information is found on the 'Significance' tab. As you can see in Figure 6, a value of 0.05 has been suggested, but you can amend this if you wish. Once you have decided on this value you then need to calculate your critical value of *t* using Table 2, and enter this value into the *t*-test calculator as shown in Figure 6.

Experimental hypotheses can take two forms: they can be directional (also called 'one-tailed'), predicting an effect in a particular direction, or they can be bi-directional (or 'two-tailed'), simply predicting that there will be a difference between the experimental and the control condition, but not specifying what the effect will be.

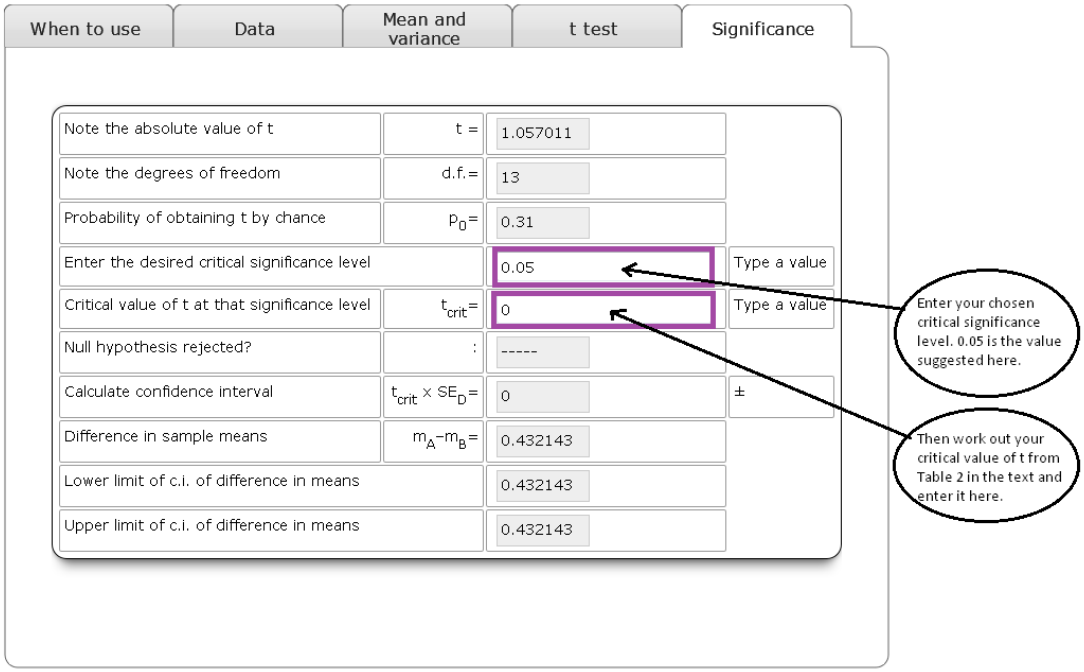

Figure 6 Entering your critical values of *p* and *t*

Table 2 Critical Values of *t*

Probability *p*	**Directional**	**0.05**	**0.025**	**0.005**	**0.0005**
	Bi-directional		**0.05**	**0.01**	**0.001**
Degree of freedom					
4		2.13	2.78	4.60	8.61
5		2.02	2.57	4.03	6.87
6		1.94	2.45	3.71	5.96
7		1.89	2.36	3.50	5.41
8		1.86	2.31	3.36	5.04
9		1.83	2.26	3.25	4.78
10		1.81	2.23	3.17	4.59
11		1.80	2.20	3.11	4.44
12		1.78	2.18	3.05	4.32
13		1.77	2.16	3.01	4.22
14		1.76	2.14	2.98	4.1
15		1.75	2.13	2.95	4.07
16		1.75	2.12	2.92	4.01
17		1.74	2.11	2.90	3.97
18		1.73	2.10	2.88	3.92
19		1.73	2.09	2.86	3.88
20		1.72	2.09	2.85	3.85
21		1.72	2.08	2.83	3.82

Probability *p*	Directional	0.05	0.025	0.005	0.0005
22		1.72	2.07	2.82	3.79
23		1.71	2.07	2.81	3.77
24		1.71	2.06	2.80	3.75
25		1.71	2.06	2.79	3.73
26		1.71	2.06	2.78	3.71
27		1.70	2.05	2.77	3.69
28		1.70	2.05	2.76	3.67
29		1.70	2.05	2.76	3.66
30		1.70	2.04	2.75	3.65
35		1.69	2.03	2.72	3.59
40		1.68	2.02	2.70	3.55
45		1.68	2.01	2.69	3.52
50		1.68	2.01	2.68	3.50
55		1.67	2.00	2.67	3.48
60		1.67	2.00	2.66	3.46
65		1.67	2.00	2.65	3.45
70		1.67	1.99	2.65	3.43
75		1.67	1.99	2.64	3.42
80		1.66	1.99	2.64	3.42
85		1.66	1.99	2.63	3.41
90		1.66	1.99	2.63	3.40
95		1.66	1.99	2.63	3.40
100		1.66	1.98	2.63	3.39

Once you have entered the relevant values you can click on the ‘Enter’ button on your computer and the *t*-test will run its calculation and show you whether the null hypothesis can be rejected (Figure 7) on the grounds of statistical reliability.

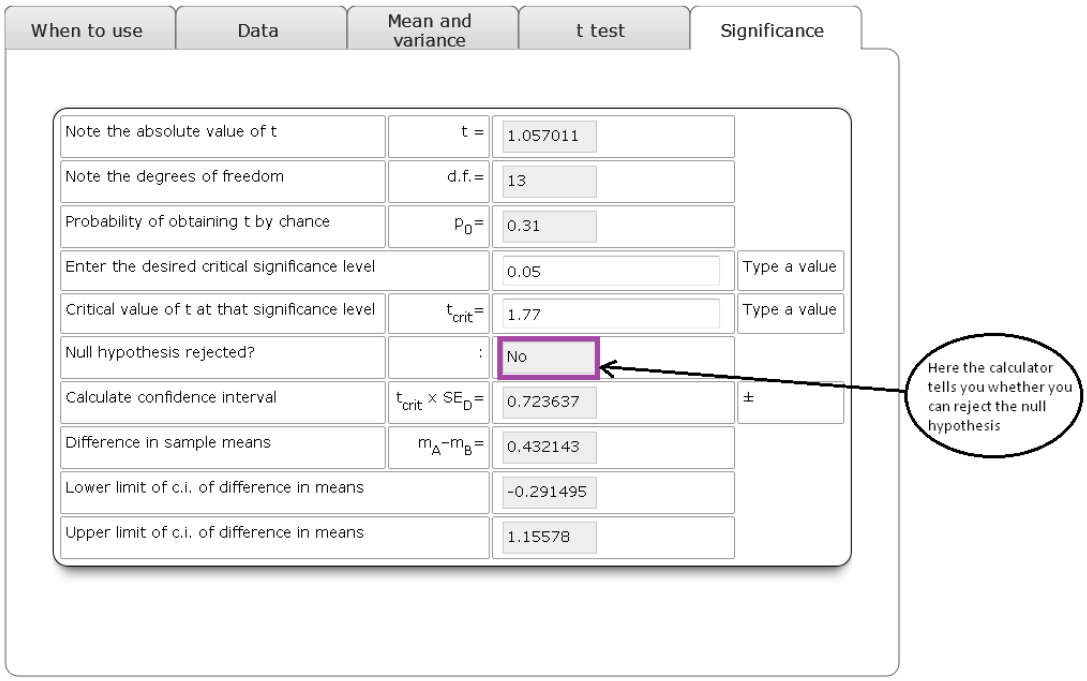

Figure 7 Deciding if the null hypothesis can be rejected

Finally you should go to the 'Mean and variance' tab and note down the calculated values of the means and variance from the *t*-test calculator, and use the variance to work out what your standard deviations are (Figure 8).

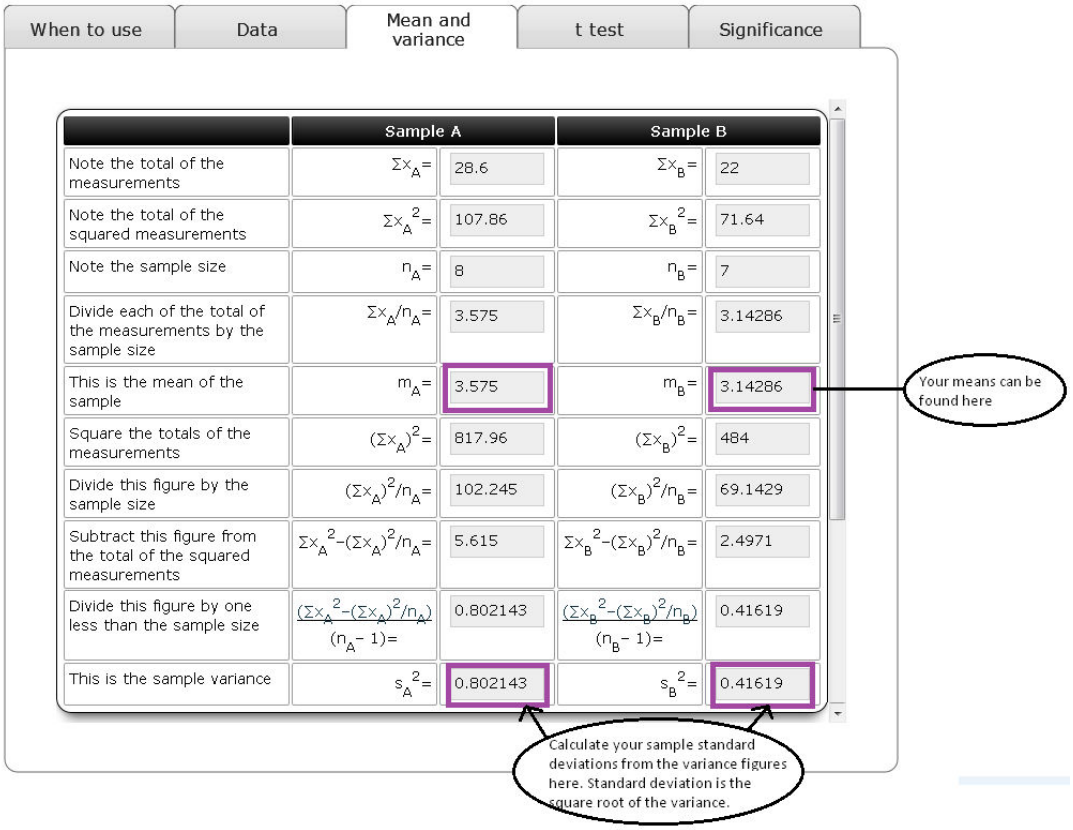

Figure 8 Working out your group means and standard deviations

You must piece all of this information together for each measure in your investigations in order to reach a suitable conclusion on your original experimental question. Can the null hypothesis be accepted or rejected?

5 Summary

The following list recaps the main points made in this guide:

- The data obtained from participants in one condition of an experiment are usually not completely different from those obtained from participants in another condition: the data are said to overlap.
- Provided it is sensible to calculate a mean and standard deviation, then it is possible to carry out a statistical test to establish the probability that the data from the two conditions came from the same population.
- If the calculated probability is small, then the data probably came from different populations; a reliable difference has been established.
- The low probability also means that the null hypothesis can be rejected.
- Conversely, if the calculated probability is large, then the data probably did not come from different populations and a reliable difference cannot be established.
- A high probability also means the null hypothesis cannot be rejected.
- Any outcome should always be interpreted alongside the observed pattern of the data obtained to establish what this means in the context of the original experimental question.