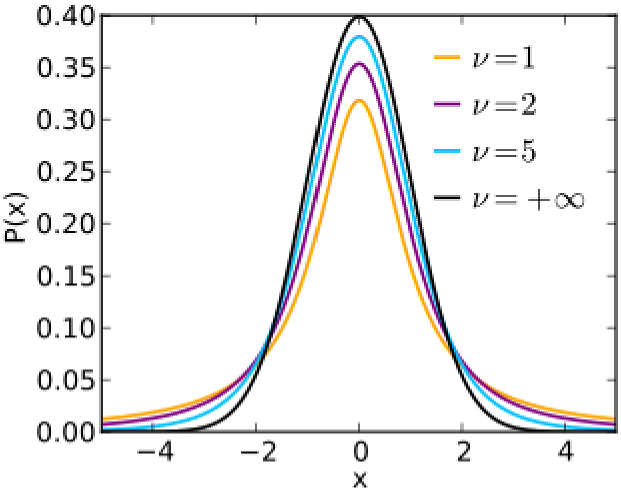
Critical Reasoning 17 - *t*-tests



Four t-distributions for each of four different degrees of freedom (ν)*. The value of the probability density function* (*Px*) *is shown on the vertical axis with the number of standard deviations shown below.* (*Source Wikipedia: Student's t-distribution*)

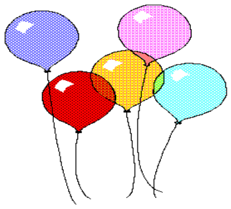
A ***t*-test** is a statistical test that is used to determine the significance of the difference between the means of two groups. Recall that in Critical Reasoning 15 we looked at statistical hypothesis testing in general. We also learned about some z-tests in particular, but decided that they were not practical in most instances as we seldom know the value of the population standard deviation (σ). Using the standard error ($\sigma/\sqrt{n}$) we were able to compare the mean of a single set of measurements to a given constant in a number of idealised examples. In this study unit and the following alternating ones we introduce, first t-tests and then other specific tests as alternative means of hypothesis testing under certain conditions. Once again, we have relied on Professors Kruger and Janeke's (2012) UNISA study guide to undergraduate Psychology Statistics. Finally, we have added a section to this study unit that sets out a nonparametric parallel to t-tests, namely the Mann–Whitney *U* test (not part of the undergraduate syllabus for the Humanities).

The *t*-distributions

If we are in a situation where we do not know the population standard deviation (σ), we might decide to estimate it using the sample variance (s^2). In practice, especially where n is small, this leads to an underestimation of σ^2 so that so that the associated z-value is slightly larger that it would have been had we known the true population variance. The unintended consequence of which is that we are more, rather than less, likely of making a Type I error. In order to compensate for this William Sealy Gosset, under the pseudonym 'Student', developed a series of probability distributions that have collectively become known as the t-distribution. (Kruger & Janeke, 2012 p. 103)

"Whereas the normal distribution describes a full population, t-distributions describe samples drawn from a full population; accordingly, the t-distribution for each sample size is different." (Wikipedia: Student's t-distribution) The size of this difference depends on a quantity known as **degrees of freedom** (*df* or ν) *i.e.* the number of independent values, that have the freedom to vary in a sample. Suppose that we ascertain one (or more) population parameters; that leaves n - 1 (or less) independent

Degrees of Freedom: A Pictorial Example



If there are five balloons of different colours and there are n = 5 children who must each select one, then there are n - 1 = 4 degrees of freedom of choice because the last child to pick a balloon will have to settle for whatever colour is left.

pieces of unknown information "free" to vary in the final calculation. Thus in a t-test for a single sample there are simply n - 1 degrees of freedom. In a t-test for two independent samples, on the other hand, there are n_1 + n_2 - 2 degrees of freedom. You will not be expected to compare three of more samples at the same time at undergraduate level. However three or more groups of samples can always be compared two-by-two at a time in a round robin fashion until each of the groups has been compared with all of the others. This is of course highly time-consuming for larger numbers of samples.

Fortunately, with the aid of modern statistical packages (including *Excel,*) we do not need to look up the relevant p-value for each t-test by hand because the value is calculated for us directly. That does not mean that we do not have to know how t-tests work or how the various t-statistics are calculated. Consider first the special case of a sample mean compared to a population mean without knowing the standard deviation of the population. This can be explained by way of an example adapted from statsdirect.com

The Single Sample *t*-Test

Suppose that you are back at the High School for girls as their matric Biology teacher as in a previous example. This class has 20 students and you are teaching them how to use a sphygmomanometer to measure each other's resting blood pressure. As you will be aware, this measurement involves both a diastolic (minimum arterial pressure) and a systolic (peak arterial pressure) value measured in millimetres of Mercury (mmHg). Once again the girls write up their results on the white board and you discuss their measurements and implications with your class, comparing them to the population mean resting values for healthy 18year olds of 120 mmHg (systolic) over 70mmHg (diastolic).

Suppose further, that you and your class want to know whether their measured sample of values are representative of the general population when you do not know the respective population standard deviations. You decide that two separate t-tests for single samples, one for the systolic and another for the diastolic values are in order. Since the statistical procedure for both t-tests will be the same, you decide to perform only one set of calculations for the systolic sample, leaving the t-test for the diastolic sample as an assignment. Here is a table of the class' systolic blood pressures measured in mmHg.

128	127	118	115
144	142	133	140
132	131	111	132
149	122	139	119
136	129	126	128

As always, before we proceed with any testing, we must clearly state our hypotheses. What we want to know is whether our sample of matric girls' resting systolic blood pressure (BPsys) is different, either way, from that of the general population of healthy 18 year olds. Therefore our alternative hypothesis will be non-directional. Thus:

H_0: BP Sys = 120

H_1: BP Sys ≠ 120

We must also state at the outset at what level of confidence we wish to test our hypothesis. Suppose we choose a 95% confidence level which is associated with an α of 0,05.

Next, consider the formula for the t-test statistic for a single sample:

$$t_{\overline{x}} = \frac{\overline{x} - \mu}{s_{\overline{x}}}$$

where $\overline{x}$ is the sample mean and μ is the population mean. Note that this looks just like the formula for the one sample z-test, introduced in Critical Reasoning 15, except that we have replaced the unknown standard deviation (σ) with the standard error of the sample mean ($s_{\overline{x}}$) which, recall, is given by:

$$s_{\overline{x}} = \frac{s}{\sqrt{n}}$$

Therefore, substituting this quantity into the formula above yields:

$$t_{\overline{x}} = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

In this form we will only have to do a once off substitution at the end of the calculation for $t_{\overline{x}}$, which means we are less likely to make a mistake by leaving something out along the way. First though, we must calculate the sample mean ($\overline{x}$) and the sample standard deviation (s) using a scientific calculator or spreadsheet program like *Excel*.

Note that there is a slight difference in the formulae for the population standard deviation (σ) that we met in Critical Reasoning 13 compared to that for a sample standard deviation (s) that we require now. Whereas:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Besides obviously having to replace the μ above by $\overline{x}$ below, the n - 1 in the denominator of the latter corresponds to the number of degrees of freedom in the vector deviations of the mean: $(x_1$ - $\overline{x}, \dots, x_n$ - $\overline{x})$. (Wikipedia: Standard deviation)

If the programme (or calculator) that you are using is any good, you should be prompted to make a choice between calculating σ or s. If you do not know how to do these calculations, there are some excellent tutorials that can be found by typing the "How to calculate … using…" specific question into your preferred search engine.

For our systolic blood pressure sample we found that,

$\overline{x}$ = 130.05 and s = 9.960316

As for the population mean (μ) and the sample size (n), we already have this information:

μ = 120 and n = 20

Our sample mean of ≈ 130 mmHg is definitely larger than the population mean of 120 mmHg, but is it significantly larger? To find out we must proceed with our testing. If we substitute these values into the equation above we get:

$$t_{\overline{x}} = \frac{130.05 \; - \; 120}{9.960316/\sqrt{20}} = 4.512404$$

Also we have *df* = 20 - 1 = 19 degrees of freedom which we need to know when we look up the associated p-value. Some programs will return this value automatically as part of the t-test operation; otherwise there are numerous reliable online t-value calculators which can be found using your preferred search engine. We used one such calculator to find the following p-values associated with our t-statistic:

One-tailed probability (right tail):	0.00011919
Two-tailed probability:	0.00023838

Because the hypothesis we wish to test is non-directional, we must compare the latter probability with our α = 0,05. Obviously, the *p* ≈ 0.002 above is much smaller than our α = 0,05. As you will recall, this p-value is the chance of obtaining the result that we did *under the null hypothesis*. Therefore we reject the null hypothesis and decide to fall back on the alternative hypothesis.

Clearly, the mean systolic blood pressure of about 130 mmHg as measured by the matric girls in this class is significantly higher than that of the population mean of 120 mmHg. What could this result mean? It could be that this sample of girls really *is* overall slightly more hypertensive (having an elevated blood pressure) than their heathy 18 year old counterparts in the general population. We have already dismissed the possibility that these measurements were a fluke or due to chance as indicted by our very low p-value.

The influence of one more artefacts cannot be so easily discounted. An **artefact**, in the sciences, is something that is observed that is not naturally present, but occurs as the result of some preparative or investigative procedure. This practical was probably the first time that many of the girls had used a sphygmomanometer, which is a rather delicate instrument. Nor were the instruments calibrated at the beginning of the practical. Also the accurate measurement of blood pressure requires the subject to sit quite still, while breathing normally. Any fidgeting, talking or holding one's breath for even a moment can result in an elevated reading. Clearly, not all of these factors were controlled for.

All of which beside, this example is purely fictitious but the method of using a t-test to compare a single sample to a known population mean and an unknown standard deviation is very much real and robust. More often than not however, we may simply wish to compare two samples, without making any assumptions about underlying statistical distributions in the data. To do so we must first decide whether the samples we wish to compare are independent or dependent.

Independent *vs.* Dependent Samples

Two (or more) samples are considered **independent** if the composition of each sample in no way systematically affects the composition of the other(s). In other words, there must be no obvious relationship between such groups. If, as Kruger and Janeke suggest, we wish to compare a construct such as "self-esteem" between two randomly sampled groups of men on the one hand *vs.* women on the other, we would be dealing with two independent samples. (p. 112) Similarly, if we were running a clinical trial for a potential new drug, we would want to compare two randomly sampled independent groups, one which receives the actual drug *vs.* another (the control group) that receives a placebo.

On the other hand, two (or more) samples are considered **dependent** if the composition of one is systematically related to that of the other group(s). Samples that are systematically dependent in this way are also known as correlated, a term we shall expand upon in Critical Reasoning 19. (*l.c.*) At first blush one might think that dependent samples are a bad thing, given that in the experimental situation we are at pains to eliminate (or isolate) all but a few intervening relations among samples. But consider the following investigations:

1. A psychotherapist believes she has developed a new hypnotherapeutic technique for helping patients quit smoking. She believes it would be unethical to divide a group of subjects desperate to quit into two, only to subject half of them to fake hypnotherapy over a number of sessions. Instead the therapist decides to provide her new therapeutic technique to all the subjects in her study, noting on a ten-point scale how intensely each would rate their desire to smoke, both before and after the treatment. So, although she only has one group of subjects, in effect she has two samples: a before treatment sample and an after treatment sample. Then of course, every subject will be represented twice, once in each sample. And because everybody is systematically related to him or herself both before and after, the two samples will be dependent.

2. A psychologist wishes to test whether gender makes any difference to mathematical ability at undergraduate level. Previous studies have found that there is a difference; however our psychologist believes that these studies have failed to take intelligence into account. To him it is obvious that smarter (higher I.Q.) female students will, on the whole, tend to outperform their duller (lower I.Q.) male counterparts on tests for mathematical ability and *vice versa*. As far as he is concerned, the variable I.Q. is simply muddying the water. (Statisticians refer to such variables as "**nuisance variables**" that have to be controlled for.) Our psychologist decides that in order to control for I.Q. he will pair off students, as closely as possible, from each sample according to their I.Q. Thus the first two members from each sample (male and female) will be those with similarly highest I.Q. scores. The next pair will be those

> with the second similarly next to highest I.Q. scores and so on down to the last pair from each sample who will have the similarly lowest IQ scores. These two samples are now dependent because they have been deliberately "matched" (one-for-one) for I.Q. scores as closely as possible, hence the term "**matched samples**". So although our psychologist has not eliminated I.Q. as a variable he has removed the "nuisance" effect that such a variable might have had on the outcome of the study, had the samples had not been matched in this way.

As Kruger and Janeke point out, dependent samples must always be of the same size (n) because each member of the one sample is matched to a counterpart in the other sample. Indeed, in the case of before and after samples, each person in the before sample is matched with him or herself in the after sample. If, on the other hand, someone was present in the before group but not in the after group or *vice versa* they cannot be counted in either sample without compromising the study. In the case of independent samples however, they need not be the same size; although they may be. (*l.c.*)

The authors also warn not to confuse "the notion of dependent versus independent *samples* with the distinction between dependent and independent *variables*… While the latter refers to the relationships among variables - how one may affect the other - in the case of samples it is a relationship among the groups from which the data were collected (*i.e.* where the variables were measured) that is of concern". (*l.c.* original emphasis)

There are two specific t-tests, one for independent and another for dependent samples which we shall explain by means of examples below; however they can only be used correctly in the appropriate context, so deciding whether your samples are independent or dependent is the first crucial step.

The Independent Two-Sample t_c - test for Difference between Means

Consider the following example which is slightly modified from that of Kruger and Janeke. (p. 112 ff.) A company of industrial psychology consultants has been offering workshops that it claims improves participants' sensitivity towards gender issues in the work place. This, they claim, leads to increased productivity, reduced gender based conflict in the work place, as well as increased job satisfaction. Another much larger company is considering employing the consultants to run workshops for its employees but is demanding "proof" that they are effective, as claimed.

The consultants propose the following study by way of demonstration: Two groups of 20 employees are randomly selected from the company's population of employees and booked into a hotel for one day. Group 1 (the treatment group) is enrolled in a full day of workshops followed by tea and cake at 4pm. Group 2 (the control group) is encouraged to enjoy the hotel facilities for the day and is asked to assemble at a separate venue in the hotel at 4pm for tea and cake.

The day before the workshops both groups are tested on a variety of standardised psychometric tests including *inter alia* those for gender sensitivity and perceived job satisfaction. One month later, both groups are subject to the same battery of tests. Because both groups were selected at random they represent independent samples of the company's parent population of employees.

There are three core research questions that the larger company wants answered:

- Did the workshops increase employee's gender sensitivity?
- Did the workshops decrease the number of incidents of gender based conflict?
- Did the workshops result in higher perceived job satisfaction among employees?

The company of industrial psychologists will have to answer all three questions with data from the study to back up their claims. We shall concentrate on the first question by way of explanation.

If, for the moment, we regard the two groups as different populations, then the independent variable has two values or levels: those that received the training (call them population 1) and those that received no training (call them population 2). In order to answer the research question, these two populations must now be compared with respect to their dependent variable *i.e.* gender sensitivity scores. If we assume that both populations had the same mean gender sensitivity scores prior to the training, then the psychologists will have to demonstrate that population 1, who received the training, had significantly higher mean gender sensitivity scores one month after the training than those in group 2.

If we let μ_1 stand for the mean post training gender sensitivity score for group 1 and μ_2 for the mean post training gender sensitivity score for group 2, then we can state the statistical hypothesis that will have to be tested as following:

$$\mathrm{H}_0 : \mu_1 = \mu_2$$

$$\mathrm{H}_1 : \mu_1 > \mu_2$$

Another way of expressing this is to take μ_2 to the left both above and below, so that we get:

$$\mathrm{H}_0 : \mu_1 - \mu_2 = 0$$

$$\mathrm{H}_1 : \mu_1 - \mu_2 > 0$$

Because the consultants have a financial interest in convincing the larger company of the efficacy of their workshop training they decide to set the bar for testing their hypothesis at the 1% confidence level. In other words, at α = 0.01.

If we look at H_1in the bottom row above, the consultants essentially wish to test for a statistically significant difference between the two means (μ_1 - μ_2). According to Kruger and Janeke, "Statisticians have determined that the distribution of the difference between two normally distributed variables also produces a normally distributed variable… Furthermore, they have found that as long as the two standard deviations (of the two groups being compared) do not differ significantly, we can estimate the standard deviation of the pooled means ($\sigma_{\overline{x_1}-\overline{x_2}}$) as follows:" (p. 114)

$$\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

A few paragraphs up we regarded "the two groups as different populations." This of course is not literally true: The two populations we have in mind are actually two different samples from the larger company's parent population. And since we do not know the population standard deviations

($\sigma_1$and σ_2) we will have to substitute them for the sample standard deviations ($s_1$and s_2) which we can work out from our raw data. Then the formula above becomes:

$$s_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Analogous to what we did in Critical Reasoning 15, where we divided the difference between the sample and the population mean by the standard error to obtain a z-distribution, so here we can divide the difference between the two sample means by the sample standard deviation of the pooled means ($s_{\overline{x_1}-\overline{x_2}}$) above to obtain a t-distribution, thus:

$$t_c = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Note that the subscript c, as in t_c, is neither a variable nor a constant. It is simply a letter used to distinguish this kind of t-test for independent samples from other sorts of t-tests. If you are anxious that the formulae in this study unit are becoming increasingly complicated, please be assured that you will not have to memorise them for exam purposes in the Humanities, not even at honours level. All of them will be provided for you on a separate data sheet; however you will be required to know which formula to use in the appropriate context.

Before we can proceed with our t_c test we must be sure that the following two assumptions about our data are true:

1. The data for our two populations are *normally distributed* and that they have the *same variance*, and
2. The samples are independent. (Kruger & Janeke, p. 115)

We know that both samples were selected at random form the parent population, so we can assume that they are independent of each other. Short of drawing histograms for each data set, we have no reason to suspect that they are *not* normally distributed. Besides which, when we have calculated their respective standard deviations we can compare them directly. (Recall variance is simply the square or the standard deviation.)

In the following table we have used the same descriptive statistics from Kruger and Janeke's example (*l.c.*) for our purposes. Remember that group 1 received the training; group 2 did not:

Table of descriptive statistics for gender sensitivity scores

Group	Sample size (n)	Mean ($\overline{x}$)	Std. deviation (s)	Minimum	Maximum
1	20	10.65	3.20	4.0	15.0
2	20	6.15	3.18	4.0	15.0

As we can see, the difference in standard deviations is very small; therefore the first assumption about the equivalence of variance is as close as makes no significant difference. What should stand out immediately is the relatively large difference between the two sample means:

$\overline{x}_1 - \overline{x}_2$ = 10.65 – 6.15 = 4.5

This is consistent with our alternative hypothesis H_1: μ_1 - μ_2 = > 0 therefore we should proceed with testing. First we calculate the value of t_c by means of substitution:

$$t_c = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(10.65 - 6.15)}{\sqrt{\frac{(3.20)^2}{20} + \frac{(3.18)^2}{20}}}$$

$$= \frac{(4.50)}{\sqrt{\frac{10.24}{20} + \frac{10.1}{20}}} = \frac{4.50}{\sqrt{\frac{20.35}{20}}}$$

$$= \frac{4.50}{\sqrt{1.0175}} = \frac{4.50}{1.0087}$$

$$t_c = 4.4612$$

As in the previous example, we must use a computer program or an online calculator to find the p-value that corresponds to our t_c above. Note that there are n_1 + n_2 - 2 degrees of freedom in this example *i.e.* 20 + 20 - 2 = 38 *df*. Also remember to select the option to calculate the value for a one tailed hypothesis as is the case for our example. Some older programs however always return a two-tailed value, in which case simply divide the returned p-value by 2. The p-value we found for our t_c was less than 0.00001. This is an astonishingly small p-value for a result in the Humanities, much smaller than our α = 0.01. Therefore we reject the null hypothesis and decide to fall back on the alternative hypothesis. If our samples were representative, then we have given as near a convincing demonstration that our workshop training method for improving gender sensitivity in the workplace is actually effective compared to a control group.

In fact we can go one step further and quantify the size of this effect by using Cohen's d. (See Critical Reasoning 15.) Recall:

$$d = \frac{estimated\ mean\ difference}{estimated\ standard\ deviation} = \frac{\overline{x}_1 - \overline{x}_2}{s_p}$$

where s_p is the pooled standard deviation (of both groups taken together). That value was calculated as s_p = 3,888, which we can substitute into the formula above as follows:

$$d = \frac{10.65 - 6.15}{3.888} = \frac{4.50}{3.888} = 1.157$$

Recall that any effect size above 0.8 standard deviations is considered "large" therefore our d of 1.157 by comparison is impressive.

The Dependent Two-Sample $t_{\bar{d}}$ - test for Difference between Means

We have already explained the difference between dependent and independent samples. When we are dealing with subjects that are either matched, related (naturally or otherwise) or even self-related (such as in repeat measure tests), we are dealing with two (or more) groups of scores that are meaningfully dependent in some way. This requires a different statistical strategy. Instead of comparing means as we did with independent groups, we are interested in a **difference score** d for each pair of subjects. This is simply the difference in scores for each matched pair:

$$d = x_2 - x_1$$

We use these d's to conduct our t-test. Note that:

- The population of difference scores has μ = 0 and a standard deviation (σ) which we can estimate.
- If there is no difference between the pairs, then the mean of the difference scores will be equal to zero, for which we can use the following notation:

$$\mu_D = 0 \text{ or } \mu_2 - \mu_1 = 0$$

Also note that with two matched samples, the freedom of values to vary in half of the sample is constrained by the values in the other half of the sample, therefore the degrees of freedom for two matched samples is only:

$$df = \frac{1}{2}n - 1$$

Example: Consider the case of the afore mentioned psychotherapist who believes she has developed a new hypnotherapeutic technique for helping patients to quit smoking. Suppose she selects n = 10 recruits from among the general public who have expressed an interest in an online advertisement that promises to help them quit smoking. Before she begins she asks all participants to jot down on a ten-point scale how intensely each of them would rate their desire to smoke that day: 1 for no desire at all, to 10 for the most irresistible desire possible. After three free sessions of hypnotherapy, over a three week period, she waits a further week before asking them to again rate their desire to smoke on the same ten-point scale. Here are her results (borrowed from a similar example available here:)

Before	9	10	7	5	7	5	9	6	8	7
After	7	6	5	4	4	6	7	5	5	7

Because our psychotherapist is dealing with dependent samples, she is not directly interested in the before and after scores but in their difference (d). She therefore creates a further row of the table above with each difference score d = $x_{after} - x_{before}$. Thus:

d	-2	-4	-2	-1	-3	1	-2	-1	-3	0

Before testing her results she states her directional hypothesis as follows:

$$\mathrm{H}_0 \colon \overline{D} = 0$$

$$\mathrm{H}_1 \colon \overline{D} < 0$$

The null hypothesis (H_0) states that the mean difference scores will show no change, while the alternative hypothesis (H_1) states that the mean difference scores will show a decrease. As Kruger and Janeke (p. 119) point out, $\overline{D}$ refers to the population mean difference scores, whereas our psychotherapist is technically referring to her *sample* mean difference scores, $\overline{d}$. Similarly, when calculating the standard deviation of the *sample* difference scores, it should be denoted by $s_{\overline{d}}$.

Just running your eye along the bottom row should tell you which hypothesis is going to be favoured; nevertheless we do not know if, and to what extent, it will be significant. Our psychotherapist therefore decides to set her α at 0.05. Next, she proceeds to calculate the mean and standard deviation of difference scores:

$$\overline{d} = \frac{-2-4-2-1-3+1-2-1-3+0}{10} = -1.7$$

and

$$s_{\overline{d}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - (-1.7))^2} = 1.49$$

Note that the *sample* mean difference score $\overline{d}$ of -1.7 is in the right direction of the alternative hypothesis (H_1). The next step is to find a suitable t-test formula for dependent samples. As Kruger and Janeke (p. 119) point out:

> It so happens that we are familiar with this particular test statistic already! Since we are in fact comparing a single mean (the difference score) with a specific constant (zero), this is just an application of the t-test for one sample when the population standard deviation (σ) is unknown (*i.e.* the $t_{\overline{x}}$ text statistic [above]). All we need to do is substitute the sample mean ($\overline{x}$) with the mean of the differences $\overline{d}$. So the test statistic is
>
> $$t_{\overline{d}} = \frac{\overline{d} - \overline{D}}{\frac{s_{\overline{d}}}{\sqrt{n}}}$$

We can make one further simplification to this formula because the value of $\overline{D}$ for the null hypothesis that we want to test is zero. Therefore we have:

$$t_{\overline{d}} = \frac{\overline{d}}{\frac{s_{\overline{d}}}{\sqrt{n}}}$$

This formula can be used in general for t-tests for two matched or dependent samples where the null hypothesis can be expressed in the difference form above. Substituting our previously calculated values into this equation gives us:

$$t_{\overline{d}} = \frac{-1.7}{\frac{1.49}{\sqrt{10}}} = -3.61$$

Next we need to look up the single tailed p-value associated with our $t_{\overline{d}}$ above. Because we are dealing with a mean derived from individual differences, the degrees of freedom are simply n - 1 = 9, which we also need to enter into our calculator or program. We did so and found the associated p-value to be 0.00283. This is well below the α of 0.05; therefore we reject the null hypothesis (H_0) and decide to fall back on the alternative hypothesis (H_1). It seems that the treatment developed by our psychotherapist is effective. In fact, we can go one step further and quantify the size of this effect using Cohen's d, as follows:

$$d = \frac{\overline{d} - \overline{D}}{s_{\overline{d}}} = \frac{-1.7 - 0}{1.49} = -1.14$$

As in the previous example, the size of this effect is considerable. Note that we are only interested in the absolute size of the value of this calculation (in standard deviations) which here is larger than 0.8 typical for a large effect size. (See Critical Reasoning 15.)

Using Differences Scores to Compare Two Independent Groups

The method of using difference scores can also be applied to two independent groups sampled both before and after a treatment, where one sample receives an experimental treatment while the other, the control group, receives either a placebo or a standard treatment. This can be illustrated by way of a further example.

Further example: Suppose that our psychotherapist is sceptical about her remarkable result above. Could there be some outside factor at play that might explain her findings? She decides to retest her method. This time round, she decides to recruit a control group but instead of offering them no treatment at all, she offers them a standard treatment of one controlled release nicotine patch of the same strength per day to be worn each day for the duration of the trial. Instead of repeating the experimental group's hypnotherapy sessions, our psychotherapist decides to keep their earlier results on file and just asks the control group rate their desire to smoke on the same ten-point scale, both before and after the nicotine patch treatment. She now has four sets of data to process: the before and after ratings for the treatment group (on file) as well as the new before and after ratings for the control group.

Instead of trying to analyse four means simultaneously, our psychotherapist instead decides to compare the *difference* between the before and after scores in both the experimental and control groups. She reasons that if her new treatment method is more effective than the standard treatment she will see a greater decline (more negative difference) in the experimental group's reported desire to smoke at the end of their study period compared to that of the control group at the end of their study period. If we represent the mean difference scores (for d = $x_{after} - x_{before}$) of the treatment and control groups populations as $\overline{D}_t$ and $\overline{D}_c$ respectively, then her hypotheses are:

$$\mathrm{H}_0 \colon \overline{D}_t \;\; = \overline{D}_c$$

$$\mathrm{H}_1: \overline{D}_t < \overline{D}_c$$

The null hypothesis (H_0) states that the mean difference scores will be no different in the treatment group *vs*. the control group; while the alternative hypothesis (H_1) states that the mean difference scores will be the smaller (more negative) in the treatment group compared with the control group. The psychotherapist now tabulates her updated findings as follows, (after Kruger & Janeke p. 122.)

Table of descriptive statistics for d: change in desire to smoke

Treatment Group	Sample size (n)	Mean differences ($\overline{d}$)	Std. dev. of differences ($s_{\overline{d}}$)
1. (Hypnotherapy)	10	-1.7	1.49
2. (Standard patches)	10	-0.9	1.36

To begin with, the mean difference scores are in the right direction: The experimental group's mean difference score is more negative (-1.7) than that of the control group (-0.9); therefore we decide to proceed with testing the hypothesis. Also notice that the standard deviations of the difference scores are very close. This is an assumption required for the type of test we intend to implement. Although both groups were sampled twice (before and after their treatment) and are thus self-dependent, the groups themselves were independently sampled of each other and thus bare no meaningful connection to each other. We can therefore use the two sample t_c test for independent samples to compare the two treatment groups above. Instead of:

$$t_c = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

we can replace $\overline{x}_1$ and $\overline{x}_2$ with $\overline{d}_1$ and $\overline{d}_2$, and $s_1$and s_2 with s_{d1} and s_{d2} respectively, so that our formula becomes:

$$t_c = \frac{(\overline{d}_1 - \overline{d}_2)}{\sqrt{\frac{s_{d1}^2}{n_1} + \frac{s_{d2}^2}{n_2}}}$$

Note that we have not tinkered with the mathematics of the formula, only replacing difference between means with difference between *mean differences*, and standard deviations with *standard deviations of differences*. Now it is simply a matter of substituting the information in table above:

$$t_c = \frac{(-1.7) - (-0.9)}{\sqrt{\frac{1.49^2}{10} + \frac{1.36^2}{10}}} = \frac{-0.8}{\sqrt{\frac{2.2201}{10} + \frac{1.8496}{10}}}$$

$$= \frac{-0.8}{\sqrt{0.2201 + \ 0.18496}} = \frac{-0.8}{\sqrt{0.40506}} = \frac{-0.8}{0.63644}$$

$$t_c = -1.2567$$

As in the previous examples, we must use a computer program or an online calculator to find the p-value that corresponds to our t_c above. Note that there are n_1 + n_2 - 2 degrees of freedom in this example *i.e.* 10 + 10 - 2 = 18 *df*. Also note that we are looking for a one tailed p-value because our psychotherapist's hypothesis is directional. The p-value we found for our t_c was less than 0.00001. This is well below the α of 0.05; therefore we reject the null hypothesis (H_0) and decide to fall back on the alternative hypothesis (H_1), which states that the mean difference scores are smaller (more negative) in the hypnotherapy treatment group compared with the standard treatment group. Therefore the subjects in the hypnotherapy treatment group experienced a significantly larger over-all drop in their scores gauging their reported urge to smoke compared to those in the standard treatment group at the end of the experiment.

As before, we can go one step further and quantify the size of this effect, this time by using Glass' Δ instead of Cohen's d (because we do not know the *pooled* standard deviation, although we do know the standard deviation of the control group (s_{d2}) (See Critical Reasoning 15.) Then,

$$\Delta = \frac{\overline{d}_1 - \overline{d}_2}{s_{d2}}$$

Substituting the relevant variables we get,

$$\Delta = \frac{(-1.7) - (-0.9)}{1.36} = \frac{-0.8}{1.36}$$

$$\Delta = -0.59$$

This would be judged as a 'medium' size effect (Δ ≈ 0.5), so not only can we be highly confident that the new hypnotherapy treatment for smoking is significantly more effective than the standard control; the size of the effect is neither trivial nor incredulously enormous either. Of course, this is a fictional example and all figures were preselected by way of example; however in the real world, data are seldom so compliant. In the task below you will be asked to provide some constructive criticism of this example as if it were a real study supervised by yourself.

General Strategy for Addressing Problems Involving t-Tests

- Are the mean and standard deviation of the population(s) known? If so would a z-test not be more appropriate? When is as t-test appropriate?
- t-tests are predicated on the assumption of normality of distribution and equality of variance (or standard deviations); however these requirements are not rigid. So long as the distribution(s) are approximately normal[1] and the variance (or standard deviations) are similar, such a test should be acceptable *ceteris paribus*.
- A single sample t-test is appropriate when there is one sample which is being compared against a constant.
- A two sample t-test is appropriate when there are two (or more) samples to compare that are either independent or dependent.

[1] Recall the discussion about the Central Limit Theorem in Critical Reasoning 13 *i.e.* "Regardless of the shape, mean or standard deviation of the parent population, the distribution of the sampling means approaches a normal distribution as n increases. (In fact, it approaches very close to normal with an n of as low as 30.)"

- The factors that determine a t-value are the size of the mean(s), the standard deviation(s) and the sample size(s). Know which formula to use in the appropriate context.
- State a research hypothesis for a research scenario requiring the comparison of one or more statistical populations. Formulate the appropriate statistical hypotheses, and test the sample results for statistical significance. (Both the t-value and the degrees of freedom are required for the conversion to the corresponding p-value.)
- Know how to calculate the effect size (Cohen's d or Glass' Δ) when two sample means are being compared. What does the calculated valve indicate? (Kruger & Janeke, p. 124 edited)

Parametric *vs*. Nonparametric tests

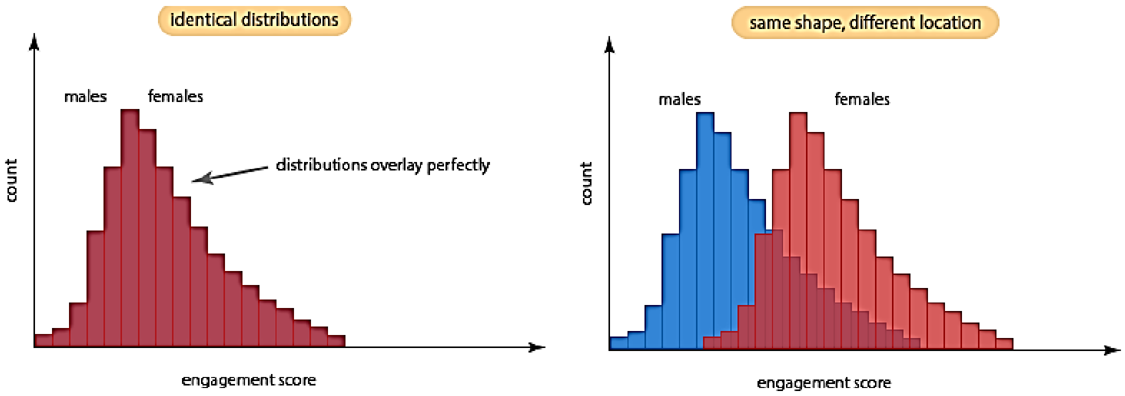
Hitherto, we have only considered examples of **parametric tests** which assume that there is some underlying statistical distribution(s) of the data. Such tests assume that several conditions must be met in order for the test to be reliable. *E.g.* Student's t-test for independent samples assumes that each sample is normally distributed and that sample variances are the same. **Nonparametric tests**, by contrast, do not rely on any assumptions about underlying statistical distribution(s) of the data. Thus, they can thus be applied independently of the assumptions for parametric tests.

Nonparametric test are more **robust** than parametric tests in that the former can be legitimately applied in a broader range of circumstances. Parametric test however have more statistical power than their nonparametric equivalents. In other words, a parametric test is more able to lead to the rejection of the null hypothesis (H_0). As a rule of thumb, the p-value associated with a parametric test will be lower than the p-value associated with the nonparametric equivalent test calculated from the same data set. (XLSTAT SUPPORT CENTER)

Mann–Whitney *U* test

The Mann-Witney *U* test is a nonparametric test of the null hypothesis that is used to compare differences between two independent groups when the dependent variable is measured on an ordinal scale and may or not be normally distributed. The Mann-Witney *U* test is often presented as a nonparametric alternative t-tests for independent samples; although this is not always the case. The Mann-Witney *U* test requires that the following assumptions be met:

- The dependent variable should, at least, be measured on an ordinal scale (such that one can rank observations).
- The independent variable should consist of two categorical groups.
- All the observations from both groups should be independent of each other.
- The data need not be normally distributed but at least they should be the same "shape". See diagram below.



Note that the two distributions on the left are identical and therefore overlap perfectly, while the two distributions on the right are the same shape but have a different location. (© *Laerd Statistics*, 2013)

There are mathematical and software tools for assessing this; however for the purposes of our present introduction, an inspection of data in the form of histograms will suffice.

The statistical hypotheses are as follows:

- Under the null hypothesis H_0, the distributions of both populations are equal.
- Under the alternative hypothesis H_1, the distributions of both populations are not equal.
- The Mann-Witney *U* test is only consistent when, under the alternative hypothesis H_1, the probability of an observation from population X exceeding an observation from population Y is different (larger, or smaller) from the probability of an observation from Y exceeding an observation from X; *i.e.*

$$\mathrm{P}(X > Y) \neq \mathrm{P}(Y > X) \text{ or } \mathrm{P}(X > Y) + 0.5 \cdot \mathrm{P}(X = Y) \neq 0.5$$

The Mann-Witney *U* test involves calculating a statistic *U*, whose distribution under the null hypothesis is known. The direct method, below, may be used for small samples where we tabulate the distribution; however for sample sizes larger than ≈ 20, we may approximate using the normal distribution. (Wikipedia: Mann–Whitney U test)

The direct method is used for comparing two small sets of observations, where *U* corresponds to the number of wins out of all pairwise contests. For data sets 1 and 2, count the number of times each observation in set 1 wins over any observations in set 2. If there are any **ties**, where the values are the same, count 0.5. Then U_1 is given by the sum of wins and ties for set 1; conversely U_2 is given by the sum of wins and ties for set 2. (Wikipedia: Mann–Whitney *U* test)

Suppose that you are dissatisfied with Aesop's fable of a single tortoise beating a single hare in a race. To find out whether the outcome of the fable can be extended to tortoises and hares in general, you select 6 tortoises and 6 hares at random and race them together. Standing at the finish line you record the order in which each finishes the race, writing a 'T' for a tortoise or an 'H' for a hare, thus:

T H H H H H T T T T T H

Using the direct method you take each tortoise in turn and count the number of times it beats a hare. You get {6; 1; 1; 1; 1; 1} which sums to U_1 = 11. Alternatively, you could have taken each hare in turn and counted the number of times it beats a tortoise. Then you would have got {5; 5; 5; 5; 5; 0} which sums to U_2 = 25. Note that the sum of U_1 and U_2 is 36, which is 6 x 6. We shall interpret the significance of the *U* statistic at the end of the second method of calculation below. (Wikipedia: Mann–Whitney *U* test)

The indirect method for larger samples involves several steps:

1. Beginning with 1, numerically rank all observations from both sets together, from smallest to largest. Where there are groups of tied values, assign a rank equal to the midpoint of the unadjusted rankings. *E.g.* in the set {3; 5; 5; 5; 5; 8}, the unadjusted rank would simply be {1; 2; 3; 4; 5; 6}. Therefore when we adjust the ranks of first set, we get {1; 3.5; 3.5; 3.5; 3.5; 6} because 3.5 is the midpoint between 1 and 6, *i.e.* (1 + 6)/2 = 3.5.

2. Next, add up all the ranks from the set that comprises observations from sample 1. This then also determines the sum of ranks from sample 2 because the sum of all ranks is *N*(*N* + 1)/2, where *N* is the total number of observations.

3. Now, *U* is given by:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

where n_1 is the sample size for sample 1 and R_1 is the sum of ranks of sample 1. (Note that it does not matter which of the two samples is chosen as sample 1.) Similarly, *U* is given by:

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

where n_2 is the sample size for sample 2 and R_2 is the sum of ranks of sample 2. Note that by substitution, the sum of $U_1$and U_2 is

$$U_1 + U_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2}$$

Since we know that the sum of all ranks $R_1 + R_2 = N(N+1)/2$ and that $N = n_1 + n_2$, the formula above, reduces to

$$U_1 + U_2 = n_1 n_2$$

Thus, the maximum value of *U* is the product of the sample sizes $n_1 n_2$. (Wikipedia: Mann–Whitney *U* test)

The Significance of the Mann–Whitney *U* Test Statistic: Like any test statistic, the value of the Mann–Whitney *U* test statistic is associated with a p-value which must be calculated or looked up in a table. Before doing so, we should clearly set out the null and alternate hypotheses, whether or not the latter is directional, as well as the level of significance, α. The sample sizes n_1 and n_2 will also be required. The smaller valve of U_1 and U_2 should be used when consulting significance tables. (Wikipedia: Mann–Whitney *U* test)

The following table displays critical values of the Mann–Whitney *U* test for values of n up to 20 for both two-tailed and one-tailed testing at the α = 0.5 and 0.1 levels of significance. If a test with a larger n or a different level of significance is called for, a quick online search should yield the required table. Alternatively, the Mann–Whitney *U* test is supported by many statistical packages which calculate the statistics of interest directly. However these are also dependent on the data meeting the necessary assumptions and the hypotheses being set out correctly. The following example can be calculated ether manually or with the aid of a suitable statistical package.

Consider another tortoise and hare race with 19 participants of each species. Those first to last past the finishing line are recorded as follows:

H H H H H H H H H T T T T T T T T T **T H** H H H H H H H H H T T T T T T T T T

If we were to simply compare medians, in bold, we might conclude that the median race time of tortoises, in position 19, beats that of hares, in position 20. However, if we compute the value of *U* using either method we will see that U_1 = 100 because each of 10 tortoises beats each of 10 hares, *i.e.* U_1 = 10 x 10. (Note this is smaller than U_2 for each of 18 hares who beats each of 18 tortoises.) We make no assumption about the direction of our (alternative) hypothesis; therefore looking up the smaller value of U_1 for a two-tailed test with n_1 and n_2 = 20, we see that this is associated with a p-value of <0,01. This level of significance would probably be acceptable to a biologist, but not to a particle physicist. (Wikipedia: Mann–Whitney *U* test)

There is more to learn about the Mann–Whitney *U* test, especially for larger samples; however it is very unlikely that this will be covered at an undergraduate or honours level in the humanities. (See Wikipedia: Mann–Whitney *U* test for further discussion.) The important points to grasp are that t-tests are suitable for parametric data but that a Mann–Whitney U test should be considered as a parallel test for non-parametric data that are at least ordinal. Finally, you should be familiar with the assumptions for both tests and the meaning of both test statistics.

Task:

Consider again the Hypnotherapeutic Treatment for Quitting (Cigarette) Smoking above**:** Constructively critique the experimental method(s) of the imaginary psychotherapist. What were some of the obstacles to the study? What would you have done differently and why?

Feedback:

The example proposed by way of explanation is problematic in several ways. Firstly the sample size of 10 experimental *vs*. 10 control treatment is far too small to recommend a new clinical treatment. At best this example could be used as a pilot study to justify further investigation. Secondly, the nature of hypnotherapy is a dynamic (and mercurial) feedback process between therapist and patient that cannot be standardised or quantified in the manner required for basic statistical analysis. Thirdly, as they say, “The road to Hell is paved with good intentions”. Asking participants to subjectively rate their desire to smoke on a particular day is problematic in itself because most addicts have rather poor insight into their cravings, whether they be for sugar, pornography or cigarettes. The second worst case scenario could be a cohort of “statistically cured” addicts who report relatively low cravings at the end of the experiment but who go on to smoke a pack-a-day anyway. The hypnotherapist should have chosen a more objective scale of measurement such as the *actual* number of cigarettes smoked per day before and after the treatment.

References:

KRUGER, P. & JANEKE, H. C. (2012) *Psychological Research - Study Guide for PYC3704*. UNISA Department of Psychology

The next Critical Reasoning study unit concerns the logic of set theory.