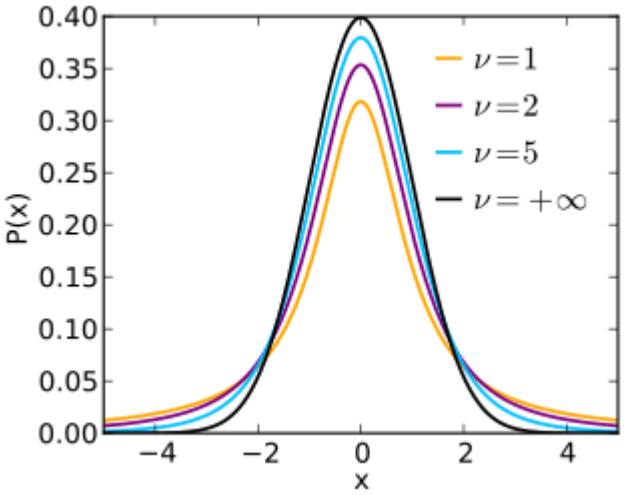
Critical Reasoning 17 - *t*-tests



Four t-distributions for each of four different degrees of freedom (v). 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)

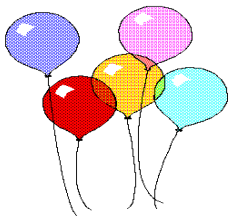
In Critical Reasoning 15 we looked at statistical hypothesis testing in general. We also learned about some *z*-tests in particular but decided that it was not practical in most instances as we seldom know the value of the population standard deviation (σ). Using the standard error ($\sigma/\sqrt{n}$) 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 Critical Reasoning 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 UNISA study guide to undergraduate Psychology Statistics both as a curriculum guide and model of clear explanation.

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 and 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). Suppose that we ascertain one (or more) population parameters; that leaves *n* - 1 (or less) independent 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

Degrees of Freedom: A Pictorial Example



If there are five balloons of different colours and there are *n*=5 children who much 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.

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 http://www.statsdirect.com/help/content/parametric_methods/single_sample_t.htm

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 but 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 the 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: $(x_1 - \overline{x}, \ldots, x_n - \overline{x})$ (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 120mmHg, but is it significant? 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 must reject the null hypothesis in favour of the alternative hypothesis.

Clearly the mean systolic blood pressure of about 130mmHg as measured by the matric girls in this class is significantly higher than that of the population mean of 120mmHg. 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** which are the result of some preparative or investigative procedure cannot be so easily discounted. This practical was probably the first time that many of the girls had used a sphygmomanometer, which is moreover a rather delicate instrument. Neither were the instruments were 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 even holding one's breath for 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 known population mean and an unknown standard deviation is very much real and robust. More often than not however, we may not even have any parametric data and

instead simply wish to compare two samples, one with another. 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 hand, we would be dealing with two independent samples. (p. 112) Similarly, if we were running a clinical trial for a new potential 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. Samples that are systematically dependent in this way are also known as **correlated**, a term we shall expand upon in the study unit on regression. (*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 decides to 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, we can state the statistical hypothesis that will have to be tested as following:

H_0: $\mu_1 = \mu_2$

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:

H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 > 0$

Because the consultants really want to convince the larger company of the efficacy of their workshop training they decide to set the bar for testing their hypothesis at a fairly conservative 1% confidence level. In other words $\alpha = 0.01$.

If we look at H_1 in the bottom row above, the consultants essentially wish to test for a statistically significant difference between the two means ($\mu_1 - \mu_2$). 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 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. Also note that 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. That the data for our two populations are *normally distributed* and that they have the *same variance*, and
2. That the samples are independent. (Kruger and Janeke, p. 115)

We know that both samples were selected at random form the parent population and 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 below 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: $\mu_1 - \mu_2 = > 0$ so we should go ahead and test it. Had we found a difference the other way round or no difference at all we would have had to stop at this point. But let us proceed by calculating 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 in our example. Some older programs however always return a two-tailed value, in which case; simply divide the returned *p*-value by two. 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 in favour of our 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 controls.

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 $t_{\overline{d}}$ - 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. We have found the following study notes (available here) from the University of North Texas to be succinct and well explained. They inform the following discussion.

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:

$$d = x_2 - x_1$$

It is these d's that we use to conduct our t-test. According to the notes above,

- 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 where every 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:

$H_0: \overline{D} = 0$

$H_1: \overline{D} < 0$

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 & 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, never the less we do not know, if and to what extent, it will be significant. Our psychotherapist therefore decides to set her α at a respectable 0,05. Next, she proceeds to calculate the mean and standard deviation:

$$\overline{d} = \frac{-2-4-2-1-3+1-2-1-3+0}{10} = -1{,}7$$

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). Had this statistic been greater than or equal to zero, our psychotherapist would have had to stop here, but this was not the case. The next step is to find a suitable *t*-test formula for dependent samples. As Kruger & Janeke 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 (*l.c.*)

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

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. (See above) 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) in favour of 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, a 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, which she continues to regard as unethical, she instead offers them the standard treatment of one controlled release nicotine patch of the same strength per day to be worn each day for the duration as the experiment. 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 hypothesis is:

$H_0: \overline{D}_t = \overline{D}_c$

$H_1: \overline{D}_t < \overline{D}_c$

The null hypothesis (H_0) states that the mean difference scores will be no different in the treatment group and 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 can 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) in favour of 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 people in the hypnotherapy treatment group experienced a significantly larger overall drop in their scores gauging their reported urge to smoke than 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 ($\Delta \approx 0{,}5$) effect, 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 for the purposes of elucidation, however in the real world, data are seldom so compliant. In part 2 of the task below you will be asked to provide some constructive criticism of the 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.
- 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 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)

Task:

By now you would have discovered your own preferred electronic means by which to perform *t*-tests therefore there are no calculations required in the following two tasks. Instead you will have to use your own imagination and statistical insight to propose meaningful and practicable techniques for investigating the following scenarios:

I.

The Stanford Marshmallow Experiment is the name given to a series of experiments led by the Stanford Psychologist Walter Mischel in the 1960's and 70's. Each was essentially a test of self-

[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.)"

mastery in the form of delayed gratification. Typically a child would be given one marshmallow and told he could have two later (± 15 minutes) if he saved (did not eat) the first one while the experimenter left the room. In a series of follow up studies those children who could wait the longest for the reward "tended to have better life outcomes as measured by SAT scores, educational attainment, body mass index (BMI), and other life measures." (Wikipedia: Stanford marshmallow experiment)

Image by J. Adam Fenster / University of Rochester

This now classic experiment has been repeated in numerous contexts and with various rewards but always with the same result: Children who evinced the greatest delayed gratification went on to achieve greater success in later life. Suppose that you are satisfied with Mischel's findings but wonder if there is any heredity factor at work in children's tendency towards delayed gratification. *I.e.* is there some genetic factor that might explain why, if at all, more closely related children might show similar waiting times compared to more distantly related children? Your task is to propose an experiment and a method of statistical analysis that could potentially answer this question. You have no access to a genetics laboratory, only several bags of delicious marshmallows and stack of mini paper plates on which to place them. You also have a small table and chair, a stopwatch, a laptop with a spreadsheet program installed, as well as a video camera. Pretend for a moment that you intend to set up a stall at a twins' convention and that you intend to invite twins as well as their non-twin siblings and any other children who wish to participate in your experiment, one at a time. Parents may watch their child from behind a screen if they wish but may not communicate with them during the experiment. No child may be kept waiting for more than 15 minutes.

II.

A Hypnotherapeutic Treatment for (Cigarette) Smoking: Critique the experimental method(s) of the imaginary psychotherapist in the body of the text above in a constructive manner. What were some of the obstacles to the study? What would you have done differently and why?

Feedback:

I.

The Stanford Marshmallow Experiment: There are two variables of interest here: the independent variable of degree of genetic relation and the dependent variable that requires no operationalisation *i.e.* waiting time (seconds). There is however an additional nuisance variable, that of age. Because younger children tend to be more impulsive this variable must be controlled for in the form of a matched study. *I.e.* children will have to be matched as closely as possible for age before comparing their waiting times. As far as degree of genetic relation is concerned, children are genetically related to different degrees: monozygotic or "identical" twins are genetically nearly identical; dizygotic or

"fraternal" twins are as genetically similar to each other as full siblings, and children of unrelated parents are relatively distantly related. (For simplicity's sake we shall not draw a distinction between monoamniotic and diamniotic twins that did or did not share the same amniotic sac, respectively.)

You hypothesise that there *is* a hereditary component to the observable trait of waiting time. You expect the waiting time of pairs of "identical" twins to be most similar compared to the waiting time of unrelated pairs of children whose waiting times should be least similar, on the whole. The waiting times of full siblings and "fraternal" twins should occupy an intermediary level of similarity.

If we simply subtract the waiting times of matched pairs, one from the other, some of the answers will be positive while others will be negative and so will tend to cancel each other out. If, instead we take the absolute value of the differences, we will always end up with a positive answer. We can express this absolute value of the difference in waiting time between any two paired children as:

$$d = |t_1 - t_2|$$

We can then express the mean difference in waiting times for pairs drawn from within the three differently related groups as:

$$\overline{d}_i\ ,\overline{d}_s\ ,\overline{d}_u$$

where the subscripts i, s, u stand for "**i**dentical", "full **s**ibling or fraternal twin" and "**u**nrelated" respectively. If we use an upper case D to refer to the respective populations then your hypothesis is essentially:

H_0: $\overline{D}_i = \overline{D}_s = \overline{D}_u$ or $\{\overline{D}_i - \overline{D}_s = 0$ and $\overline{D}_s - \overline{D}_u = 0\}$

H_1: $\overline{D}_i < \overline{D}_s < \overline{D}_u$

The null hypothesis (H_0) states that the mean difference in waiting times for pairs drawn from all three groups should be the same, while the alternative hypothesis (H_1) states that the mean difference in waiting times for pairs that are genetically identical should be less than the mean difference in waiting times for pairs that are related as full siblings or fraternal twins, which in turn should be less than the mean difference in waiting times for pairs that are unrelated.

If the mean differences in waiting times for all three groups are the same or very close or you should simply stop you analysis and write up your findings. If however the mean differences in waiting time are aligned as they are in the alternative hypothesis (or perhaps *very unexpectedly* in the opposite direction), you should proceed with some sort of test, remembering to set your α at a suitable value. (Say 0,1 or 0,05.) Since the children in your study will be matched for age, the most obvious candidate for a statistical test in this case should be the $t_{\overline{d}}$ test for comparing the differences between means of two dependent groups at a time. Also recall that *t*-tests in general require that:

1. The samples are more or less normally distributed, and
2. They have the same or similar standard deviations (variance)

You may use the formula for $t_{\overline{d}}$ that we met above:

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

Alternatively if you test the null hypotheses in the difference form as at right, above you will be comparing a difference between means against a constant $\overline{D}$ of zero, therefore the formula above simply becomes:

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

From here you can compute first a *t*-value and thence a *p*-value for $\overline{d}_i$ *vs.* $\overline{d}_s$ followed by $\overline{d}_s$ *vs.* $\overline{d}_u$. Remember that although we are dealing with matched samples, the means that you wish to test are *derived from individual differences*, therefore the degrees of freedom that you need to enter into your computer program or statistic calculator each time is simply *n* - 1 *df*. If one or both *p*-values are significant you should proceed to test the size of the effect in each case using Cohen's *d*, where

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_p}$$

and where s_p is the pooled standard deviation (of two groups taken together, first for $\overline{d}_i$ *vs.* $\overline{d}_s$ and then for $\overline{d}_s$ *vs.* $\overline{d}_u$.) So long as your *p*-values are significant, even if your effect sizes are small (say, ≈0,2 standard deviations), you shall have made an interesting discovery.

Now that you know what is to be done with your data, we can turn to the more fundamental question how it is to be correctly and unbiasedly sampled. One way to anonymise your data and thus mitigate bias is to assign each participant a case number as they come to your stall, from, say C001 to C120. Then, next to each case number you can type in an age, and adjacent to that, one or more encoded relations, if any. *E.g.* 'S004' for 'full sibling to C004' or perhaps 'I002' for 'identical to C002', or just '0' for 'unknown relation to anyone else in the study.' To the right of the relationship code you can record his or her waiting time.

If you follow this procedure and intend using a spreadsheet to capture your data there should be 4 columns and *n* rows with the case numbers recorded in column 1, the ages in column 2, the relationship codes in column 3 and the waiting times in column 4. In order to prepare your data for matching you should then perform a sort by column 2 (either ascending or descending) so that your cases are sorted by age. This will greatly simplify the task of matching cases in order to populate the various samples sets according to their relations, i, s and u because those of similar or identical age will be adjacent to one another on the spreadsheet after the sort.

Next you can create additional columns (two each) for your data sets i, s and u which will contain the waiting times of two matched individuals for each relationship type and an adjacent column for the *difference* in waiting times. The contents of these "difference columns" will provide you with the d_{i1} to d_{in}, d_{s1} to d_{sn} and d_{u1} to d_{un} for use in the statistical analysis above.

If you are more "old-school" inclined you may want to use index cards which are easy to sort by hand and can be placed side-by-side on a flat surface to compare waiting times. However once you

have compiled your three sets of d_1 to d_n's you will have to use some sort of statistical or spreadsheet program to handle the calculations above.

II.

A Hypnotherapeutic Treatment for (Cigarette) Smoking: The example proposed by way of an illustrated 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 participant 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 and 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.