T-TESTS:

When to use a t-test:

The simplest experimental design is to have two conditions: an "experimental" condition in which subjects receive some kind of treatment, and a "control" condition in which they do not. We want to compare performance in the two conditions. Sometimes, the difference between the two conditions is very clear-cut: our experimental treatment has made a clear difference to subjects' behaviour. More often in psychology, the difference between the conditions is not so obvious; in these circumstances, use of a t-test can help us to decide whether the difference between the conditions is "real" or whether it is due merely to chance fluctuations from one time of testing to another. The t-test enables us to decide whether the mean of one condition is really different from the mean of another condition.

There are two versions of the t-test:

(a) dependent-means t-test (also known as the **"matched pairs"** or **"repeated measures"** t-test): use this when the same subjects participate in both conditions of the experiment.

(b) independent-means t-test (also known as an **"independent measures"** t-test): use this when you have two different groups of subjects, one group performing one condition in the experiment, and the other group performing the other condition.

In both cases, we have one independent variable (the thing we manipulate in our experiment), with two levels (the two different conditions of our experiment). We have one dependent variable (the thing we actually measure).

Examples where the t-test is appropriate:

(a) Differences between extraverts and introverts in performance on a memory test. (The independent variable (I.V.) is "personality type", with two levels - introversion and extraversion - and the dependent variable (D.V.) is the memory test score). An independent-measures t-test would be appropriate here.

(b) The effects of alcohol on reaction-time performance. (The I.V. is "alcohol consumption", with two levels - drunk and sober - and the D.V. is reaction-time performance). A repeated-measures t-test could be used here; each subject's reaction time could be measured twice, once while they were drunk and once while they were sober.

Rationale behind the t-test:

In essence, both types of t-test are similar in principle to the z-score.

(a) We have two sample means. These differ to a greater or lesser extent.

(b) We have some idea of what sort of difference we *believe* exists between the means of the two populations from which we think these samples have come. Under the null hypothesis (that our experimental manipulation has had no effect on our subjects), we would expect the two population means to be identical (i.e., to show *no* difference).

(c) We compare the difference we actually *have* obtained, to the difference (*no* difference) that we would *expect* to obtain. If we have found a very big difference between our two sample means, there are two possibilities. One possibility is that the two sample means could be a poor reflection of the means of the populations that they are supposed to represent (i.e., our samples are atypical ones). The other possibility is that the two sample means are in fact *good* reflections of their parent populations, and that *it is our initial assumption that the two populations are identical that is at fault*. The bigger the difference between our two sample means, the less plausible the first interpretation becomes, and the more we are inclined to believe the second interpretation.

(d) All this is complicated by the fact that, as with the z-score, you have to take into account the spread of possible differences between sample means that could be obtained.

$$t = \frac{\begin{array}{c}\textit{observed}\text{ difference}\\ \text{between sample means}\end{array} - \begin{array}{c}\text{difference between}\\ \text{population means, under}\\ \text{the null hypothesis}\end{array}}{\begin{array}{c}\text{estimate of the standard error of the difference}\\ \text{between two sample means}\end{array}}$$

Repeated measures versus Independent Measures:

The formula for working out the t-test differs according to whether we have a repeated measures design or an independent measures design. Consider what happens in these two types of experiment.

(a) The repeated-measures design:

In a repeated-measures experiment, we have two conditions. Let's call them A and B. Each subject does both A and B. If we did nothing at all to the subjects, we would have two measures of each subject's performance: what they did during condition A and what they did during condition B. We would expect each subject to behave fairly similarly on both occasions, because most of the characteristics of each subject (age, intelligence, memory, sex, motivation, arousal level, etc, etc) remain the same on both occasions. As long as our performance measure is reliable, and whatever we are measuring remains reasonably stable from one time of testing to another, there should be a strong relationship between a subject's performance in condition A and their performance in condition B. Subjects who scored highly on condition A would probably also score highly on condition B. Subjects who scored poorly on condition A would also probably score badly on condition B.

If we now do something to subjects in condition A that we don't do to them during condition B, the only major difference between A and B is in terms of what we have done. Any difference in performance between A and B is probably due to this experimental manipulation.

(b) The Independent Measures Design:

Now think about what happens when we use an independent-measures design. Now we have two groups of *different* subjects: one group does condition A, and an entirely different group of people does condition B. Even if we treated both groups in exactly the same way, we would expect to see more variation between performance in conditions A and B, because they are being performed by different people. Things such as age, intelligence, etc., that remained constant across the two conditions in the repeated-measures design, now differ between our two conditions in a haphazard way.

With a repeated-measures design, differences between the two conditions can be caused in only two ways: because of differences in what we did to the subjects, or because of any other differences in how a subject performs from one occasion to the next. The latter will probably be fairly minor, compared to the effects of our experimental manipulation.

With an independent-measures design, differences between the two conditions can also be caused in only two ways: because of differences in what we did to the subjects, or because subjects in one group behave differently from people in the other group (because they are different people with different abilities, motivation, arousal, etc). The differences between subjects may well introduce considerable random variation into the performance within each group and possibly between the two groups as well.

With both the repeated- and independent-measures designs, we can look at the scores that we obtain in our two conditions, and see that variation in them comes from two sources. We have *systematic* variation between the two conditions (systematic because we do something to *all* subjects in one condition that we do not do in the other condition) and *unsystematic* variation between the two conditions. We are likely to find less unsystematic variation with the repeated measures design than with the independent measures design, because in the former the variation stems from differences in how the same individuals behave from occasion to occasion, whereas in the latter the variation comes from differences between different people. We would expect to find more difference between two individuals than between two performances by one and the same individual.

What this means in practice is that the effect of our experimental manipulation is likely to show up more easily with a repeated measures design than with an independent measures design; the repeated measures design is more *sensitive* to the effects of our experimental manipulation. In a repeated measures design, the effect of our manipulation has to show up against a background of "noise" introduced by fluctuations in individuals' performance from one time to another, fluctuations which are relatively minor. With an independent measures design, the effects of our experimental manipulation have to show up against a background that is inherently "noisier", because it consists of differences between individuals, and these are likely to be larger than the differences between the same individuals at two different times.

The need for randomisation:

With the repeated measures design, it is important that all subjects do not do the two conditions in the same order; i.e., all subjects should not do condition A followed by condition B, or vice versa. In the case of the independent-measures design, it is important that subjects are randomly assigned to group A or group B.

This random allocation to conditions is important because the t-test works by comparing the systematic variation produced by our experimental manipulation, to the random variation produced by other factors (performance fluctuations from one testing session to another, in the case of the repeated-measures design, or random fluctuations in performance between the two groups of subjects, in the case of the independent-measures design). How can we distinguish between these two sources of variation in performance? The answer lies in how systematic they are.

First, let's consider the repeated-measures design: we have done something to *all* of the subjects in one condition that we didn't do to *all* of them in the other. In theory, our manipulations can therefore produce *only* systematic differences between the two conditions.

A subject might still perform differently in the two conditions for a variety of reasons unrelated to what we did to them: they might perform better on the second test simply because they are now more practiced at the task, or they might perform worse on the second test because they are bored or tired. We can't *eliminate* these kinds of effects; however, what we can do is *prevent them from producing systematic effects on our data*. If half of the subjects do condition A first and condition B second, and the other half of the subjects do B then A, these sorts of differences should largely cancel out. Subjects doing condition A followed by condition B might well be more bored while they are doing B, and this boredom might well affect their performance; however, the effects of boredom on subjects doing these conditions in the opposite order (i.e., B followed by A) should ensure that boredom affects performance in both conditions reasonably equally - or at least, not in any systematic way.

Similar arguments apply in the case of the independent-measures design. Where different subjects do different conditions of our experiment, differences in performance between the conditions can arise from a variety of factors. The subjects in the two conditions differ *systematically* from each other in terms of what we have done to them (i.e., we have given everybody in one group some treatment that we have not given to everyone in the other group). There are lots of other sources of differences between the two groups that can give rise to variations in performance, but as long as subjects are allocated randomly to the groups, the effects of these other factors should be *unsystematic* in nature. If they produce systematic effects, we would end up with results which were impossible to interpret, because we would have no way of distinguishing between the systematic effects of our experimental manipulation and the systematic effects of these other factors.

An example should make this clear. Imagine we wanted to do an experiment which involved measuring the effects of alcohol consumption upon reaction times. Reaction times tend to differ quite a lot between individuals. Some people naturally have fast reaction times, and others naturally have slow reaction times. If we randomly allocate people to one group or the other, by chance both groups in our experiment are likely to contain a mixture of people, some with relatively fast reaction times and others with relatively slow reaction times.

Random allocation of subjects to one condition or the other cannot *eliminate* natural variations in reaction times, but it should prevent them from *systematically* affecting our data. If on the other hand, we placed all of the people with quick reaction times in one group and all of the people with slow reaction times into the other group, we would be introducing another systematic influence on our results. Suppose we gave alcohol to the group of people with slow reaction times, and gave no alcohol to the group who had fast reaction times. If we found differences in reaction times between the two groups, we would have no way of distinguishing the systematic difference between the two groups that arises from their inherent difference in reaction times from the systematic difference between the two groups that stems from the effects of the alcohol. In practice, there are many such factors ("confounding variables") which might systematically affect our results; even with the most careful selection of subjects, it is hard to be certain that one has eliminated all of them, and so the best safeguard against their effects is to allocate subjects randomly to conditions.

Repeated Measures t-test, step-by-step:

Imagine we are interested in the effects of Prozac on driving performance. We take ten people from New Zealand, take them to a sheep farm, and test their driving performance twice: test A is given immediately after they have taken Prozac, and test B is given while they have no drugs in their system. (Five subjects would be tested in the order A then B, and five would be tested in the order B then A). Each subject thus provides us with two scores (one for each condition). The question we want to answer is: does Prozac significantly affect driving ability? Here are the data:

Number of sheep hit during a 30-minute driving test:

Subject:	**Score on test A**	**Score on test B**	**Difference, D**
1	28	25	3
2	26	27	-1
3	33	28	5
4	30	31	-1
5	32	29	3
6	30	30	0
7	31	32	-1
8	18	21	-3
9	22	25	-3
10	24	20	4
			ΣD = 6

$$t = \frac{\overline{D} - \mu D(hypothesised)}{S_{\overline{D}}}$$

1. Add up the differences:

$\Sigma D = 6$

2. Find the mean difference.

$$\overline{D} = \frac{\sum D}{N} = \frac{6}{10} = 0.6$$

3. Get the estimate of the population standard deviation (the standard deviation of the differences).

$$S_D = \sqrt{\frac{\sum\left(D - \overline{D}\right)^2}{n-1}}$$

4. Get the estimate of the population standard error (the standard error of the differences between two sample means).

$$S_{\overline{D}} = \frac{S_D}{\sqrt{n}}$$

5. Get the hypothesised difference between the population means. Generally, our null hypothesis is that there is no difference between the means of the populations from which our two samples of behaviour have come, and so the hypothesised difference is 0.

$\mu_D (\text{hyp.}) = 0$

6. Work out "t".

$$t = \frac{0.6 - 0}{0.92} = 0.65$$

7. The "degrees of freedom" (d.f.) are the number of subjects minus one:

d.f. = n - 1 = 10 - 1 = 9

8. Look up the critical value of t (e.g., in the table that's on my website: part of this is reproduced later in this document).

If we are predicting *a* difference between tests A and B (i.e., merely saying that A and B *differ*) we find the critical value of t for a "two-tailed" test. With 9 d.f., the critical value of t for a two-tailed test is 2.262.

If, on the other hand, we have a "directional" hypothesis (i.e., we are predicting that A is *bigger* than B, or A is *smaller* than B), we find the critical value of t for a "one-tailed" test. For 9 d.f., the critical value of t for a one-tailed test would be 1.833.

If our obtained t is bigger (or equal to) the critical t-value, we "reject the null hypothesis" - i.e., we conclude that the difference between our sample means (as represented by t) is so large that it is unlikely to have arisen merely by chance. In other words, there is probably a "real" difference between our two conditions.

In this case, we have an obtained t of 0.65. This is much smaller than 2.262 (or 1.833 for that matter). We would therefore conclude that there was no significant difference between performance on the two tests; the observed difference between the two tests is so small, that we had best assume that it has arisen by chance.

Another example of the repeated-measures t-test:

Ten participants take a test of motor coordination, once after drinking a pint of beer and once without drinking alcohol. Their times (in seconds) to complete the task are given. Perform a related samples t-test to test whether drinking beer makes you slower at the task.

1. Calculate the difference score (D) for each subject. Then find the sum of these difference scores.

	CONDITION				
Participant	With Beer	Without Beer	Difference (D)	$D-\overline{D}$	$(D-\overline{D})^2$
1	12.4	10.0	2.4	0.8	0.64
2	15.5	14.2	1.3	- 0.3	0.09
3	17.9	18.0	- 0.1	- 1.7	2.89
4	9.7	10.1	- 0.4	- 2	4
5	19.6	14.2	5.4	3.8	14.44
6	16.5	12.1	4.4	2.8	7.84
7	15.1	15.1	0.0	- 1.6	2.56
8	16.3	12.4	3.9	2.3	5.29
9	13.3	12.7	0.6	- 1	1
10	11.6	13.1	- 1.5	- 3.1	9.61
		ΣD	16.0	$\sum(D-\overline{D})^2$	48.36

2. Calculate the mean difference score ($\overline{D}$)

$$\frac{\sum D}{n}$$

(The sum of the difference scores divided by the number of difference scores)

$$\frac{16}{10} = 1.6 \qquad \overline{D} = 1.6$$

3. Calculate the standard deviation of the difference scores (use the numbers calculated in the table)

$$S_D = \sqrt{\frac{\sum\left(D - \overline{D}\right)^2}{n-1}}$$

$$S_D = \sqrt{\frac{48.36}{9}}$$

$$S_D = \sqrt{5.373}$$

$$S_D = 2.318$$

4. Find the standard error of the mean of the difference scores

$$S_{\overline{D}} = \frac{S_D}{\sqrt{n}}$$

$$S_{\overline{D}} = \frac{2.318}{\sqrt{10}}$$

$$S_{\overline{D}} = \frac{2.318}{3.162}$$

$$S_{\overline{D}} = 0.733$$

5. The hypothesised difference between the two conditions is zero. μ_D (hyp.) = 0

6. Find *t*. Take the mean difference score (Step 2), and divide it by the standard error of the mean of the difference scores (Step 4).

$$t = \frac{1.6 - 0}{0.733} \qquad t = 2.183$$

7. Calculate the degrees of freedom. d.f. = number of participants – 1

$d.f.=10-1$ $d.f.=9$

8. Use a table of critical t-values (see the one at the back of this document)

The critical t-value at the *p* = .05 significance level, for a two-tailed test, is: 2.262. Our t-value (from the experiment) was: 2.183. In order for this to be significant, it must be LARGER than the critical t-value derived from the table. Therefore, it can be concluded that the result is non-significant.

In other words, we can't be sure that the observed difference between the two conditions hasn't occurred merely by chance.

We would report this as follows:

"There is no significant difference between the times taken to complete the task with or without alcohol (t(9) = 2.183, p>.05)".

Think about the design. How could you design this experiment to ensure getting a fair result? Firstly, you have to take order effects into consideration – so half of the participants would complete the 'with beer' condition first and the 'without beer' second. The other half would complete the test in the reverse order.

Secondly, you have to take the effect of the beer into consideration – so you'd have to allow for sobering up time.

Independent Means t-test, step-by-step:

Imagine we did our Prozac experiment in a different way; now, instead of using the same subjects twice, we use two separate groups of subjects. Ten subjects do condition X (a driving test immediately after taking Prozac) and a different ten subjects do condition Y (a driving test without any drugs). Again, the question is: does Prozac significantly affect driving ability? Here are some of the data (I've only shown the data for subjects 1, 2 and 10 in each condition).

Group X: **Group Y:**

	X	**X^2**		**Y**	**Y^2**
Subject 1	38	1444	Subject 1	47	2209
Subject 2	35	1225	Subject 2	45	2025
. . .	. . .	. . .	. . .	. . .	. . .
Subject 10	41	1681	Subject 10	48	2304

Here is the formula for an independent-means t-test. You are probably thinking, "wow, it's a biggy". It is in fact one of the most intimidating-looking formulae known to man (or woman for that matter). However, if you break it down into its constituents, it's quite easy to work it out.

$$t = \frac{\left(\overline{X} - \overline{Y}\right) - \left(\mu X - \mu Y\right) hypothesised}{\sqrt{\dfrac{\sum\left(X - \overline{X}\right)^2 + \sum\left(Y - \overline{Y}\right)^2}{\left(N_X - 1\right) + \left(N_Y - 1\right)} * \left(\dfrac{1}{N_X} + \dfrac{1}{N_Y}\right)}}$$

Step 1:

N_X = the number of subjects in condition X. Here, N_X = 10.
N_Y = the number of subjects in condition Y. Here, N_Y = 10.

Step 2:

Add together all the X scores, to get the sum of X:

$\Sigma X = 389$

Step 3:

Divide the result of Step 2 (i.e., ΣX) by the number of subjects who did condition X, to get the mean of the X scores:

$$\overline{X} = \frac{\sum X}{N_X} = \frac{389}{10} = 38.9$$

Step 4:

Add together all the Y scores, to get the sum of Y:

$\Sigma Y = 442$

Step 5:

Divide ΣY by the number of subjects who did condition Y, to get the mean of the Y scores:

$$\overline{Y} = \frac{\sum Y}{N_Y} = \frac{442}{10} = 44.2$$

Step 6:

Square each X score, and then add together all these squared scores, to get ΣX^2.

$\Sigma X^2 = 15245$

Step 7:

Do the same with the Y scores, to get ΣY^2:

$\Sigma Y^2 = 19610$

Step 8:

Add together all the X scores (which we have already done in step 2 above) and then square this sum:

$(\Sigma X)^2 = 389^2$

Step 9:

Do the same with the Y scores:

$(\Sigma Y)^2 = 442^2$

Step 10:

Find $\sum\left(X-\overline{X}\right)^2$

$$\sum\left(X-\overline{X}\right)^2=\sum X^2-\frac{\left(\sum X\right)^2}{N_X}=15245-\frac{389^2}{10}=113$$

Step 11:
Likewise,

find $\sum\left(Y-\overline{Y}\right)^2$

$$\sum\left(Y-\overline{Y}\right)^2=\sum Y^2-\frac{\left(\sum Y\right)^2}{N_Y}=19610-\frac{442^2}{10}=74$$

Step 12:
For most purposes, you can regard (µX - µY) hyp. as being zero.

Step 13:
We have now worked out most of the bits of the t-test formula.

$$t=\frac{(38.9-44.2)-(0)}{\sqrt{\frac{113+74}{9+9}*\left(\frac{1}{10}+\frac{1}{10}\right)}}=\frac{-5.3}{\sqrt{\frac{187}{18}*(0.1+0.1)}}$$

$$t=\frac{-5.3}{\sqrt{10.389*0.2}}=\frac{-5.3}{\sqrt{2.078}}=\frac{-5.3}{1.442}$$

t = -3.68 with (n_X -1) + (n_Y - 1) degrees of freedom.

The final step is to compare our obtained value of t to the critical value of t for 18 d.f., found in a table of critical t-values in exactly the same way as described earlier for the dependent-means t-test. The critical value of t in this case (for a two-tailed test) is 2.878. Our obtained t is *bigger* than this critical value, and so we conclude that there is a statistically significant difference between the two groups: driving ability is affected by Prozac.

Another example of the independent-means t-test:

The effects of work schedule on worker productivity (output) as a function of work schedule were investigated in two different factories. In the first factory the employees are on a fixed shift system, while in the second factory the workers have a rotating shift system. Under the fixed shift system, a worker always works the same shift, while under the rotating shift system, a worker rotates through three shifts. Using the scores below. determine if there is a significant difference in worker productivity between the two groups of workers.

There are two conditions (fixed shift and rotating shift) and each participant took part in only one of the two conditions. The data are participant scores measured on a ratio scale (the higher the score, the greater the worker's productivity, as measured in terms of number of items produced per hour). Therefore we can use an independent means t-test.

$$t = \frac{\left(\overline{X} - \overline{Y}\right) - \left(\mu X - \mu Y\right) hypothesised}{\sqrt{\dfrac{\sum\left(X - \overline{X}\right)^2 + \sum\left(Y - \overline{Y}\right)^2}{\left(N_X - 1\right) + \left(N_Y - 1\right)} * \left(\dfrac{1}{N_X} + \dfrac{1}{N_Y}\right)}}$$

X is the condition in which we have scores for workers' output on fixed shifts.
Y is the condition in which we have scores for workers' output on rotating shifts.

The first step is to complete a table:

Fixed Shift	output		Rotating Shift	output	
Worker	X	X^2	Worker	Y	Y^2
1	79	6241	1	63	3969
2	83	6889	2	71	5041
3	68	4624	3	46	2116
4	59	3481	4	57	3249
5	81	6561	5	53	2809
6	76	5776	6	46	2116
7	80	6400	7	57	3249
8	74	5476	8	76	5776
9	58	3364	9	52	2704
10	49	2401	10	68	4624
11	68	4624	11	73	5329
Total	775	55837	Total	662	40982

Now we can begin to assign numbers to more parts of the formula.

$$t=\frac{\left(\overline{X}-\overline{Y}\right)-\left(\mu X-\mu Y\right)hypothesised}{\sqrt{\frac{\sum\left(X-\overline{X}\right)^2+\sum\left(Y-\overline{Y}\right)^2}{\left(N_X-1\right)+\left(N_Y-1\right)}*\left(\frac{1}{N_X}+\frac{1}{N_Y}\right)}}$$

Therefore:

N_X is the number of participants in condition X = 11

N_y is the number of participants in condition Y = 11

$\sum X$ is the sum of all scores in condition X = 775

$\sum Y$ is the sum of all scores in condition Y = 662

$\left(\sum X\right)^2$ is the sum of all scores in condition X squared: 775 x 775 = 600625

$\left(\sum Y\right)^2$ is the sum of all scores in condition Y squared: 662 x 662 = 438244

$\sum X^2$ is the sum of the squared scores in condition X = 55837

$\sum Y^2$ is the sum of the squared scores in condition Y = 40982

$\overline{X}$ is the mean of all scores in condition X: 775 / 11 = 70.45

$\overline{Y}$ is the mean of all scores in condition Y: 662 / 11 = 60.18

$\left(\mu X-\mu Y\right)hypothesised$ is usually regarded as equalling 0

We also need to calculate:

$$\sum\left(X-\overline{X}\right)^2$$

$$=\sum X^2-\frac{\left(\sum X\right)^2}{N_X}$$

$$=55837-\frac{600625}{11}$$

= 55837 - 54602.27

= 1234.73

And:

$$\sum\left(Y-\bar{Y}\right)^2$$

$$=\sum Y^2-\frac{\left(\sum Y\right)^2}{N_Y}$$

$$=40982-\frac{438244}{11}$$

= 40982 - 39840.36

= 1141.64

So, if we put all of this into the equation:

$$\frac{(70.45-60.18)-(0)}{\sqrt{\frac{1234.73+1141.64}{(11-1)+(11-1)}*\left(\frac{1}{11}+\frac{1}{11}\right)}}$$

$$=\frac{10.27}{\sqrt{\frac{2376.37}{20}*0.182}}$$

$$=\frac{10.27}{\sqrt{118.82\times 0.182}}$$

$$=\frac{10.27}{\sqrt{21.625}}$$

$$= \frac{10.27}{4.650}$$

$$= 2.209$$

We also need to calculate the degrees of freedom, represented by the formula:

$$\left(\mathrm{N}_x - 1\right) + \left(\mathrm{N}_y - 1\right)$$

$$\left(11 - 1\right) + \left(11 - 1\right)$$

10 + 10 = 20

So, there are 20 degrees of freedom.

Therefore, t(20)= 2.209.

We now need to compare this obtained value of t to the critical value of t (found in the table of critical values using 20 degrees of freedom). If the obtained t value is LARGER than the critical value, it can be concluded that there is a statistically significant difference between the two groups.

	Table of critical values of t:							
	One Tailed Significance level:							
	0.1	**0.05**	**0.025**	**0.005**	**0.0025**	**0.0005**	**0.00025**	**0.00005**
	Two Tailed Significance level:							
df:	**0.2**	**0.1**	**0.05**	**0.01**	**0.005**	**0.001**	**0.0005**	**0.0001**
2	1.89	2.92	4.3	9.92	14.09	31.6	44.7	100.14
3	1.64	2.35	3.18	5.84	7.45	12.92	16.33	28.01
* * **18**	1.33	1.73	2.1	2.88	3.2	3.92	4.23	4.97
19	1.33	1.73	2.09	2.86	3.17	3.88	4.19	4.9
20	1.33	1.72	**2.09**	**2.85**	3.15	3.85	4.15	4.84

For a two-tailed test:
The critical value from the table at the .05 level is 2.09
The critical value from the table at the .01 level is 2.85

The obtained t(20) = 2.21, which is significant at the .05 level.

The question asked whether there was a significant difference between the two groups in terms of worker productivity (output) From this we can see that there is a significant difference between the two groups. In other words, the type of shift work carried out by the workers does affect the overall mean level of productivity.

Requirements for performing a t-test:

t-tests are examples of *parametric* tests; they are based on the assumption that our data possess certain characteristics. If the data do not have these characteristics (or "parameters") then the t-test may not give meaningful results. These assumptions are as follows:

(a) The frequency of obtaining scores in each group or condition should be roughly normally distributed. (Obviously, if you have a sample of ten or fifteen subjects, you are unlikely to get a perfect bell-shaped curve. However, if when you do a rough plot of how often each score crops up, you find that the scores are markedly skewed to one side of the mean or the other, then the results of a t-test on those data should be viewed with caution).

(b) The data should consist of measurements on an interval or ratio scale.

(c) The two groups or conditions should show "homogeneity of variance". In other words, the spread of scores within each group should be roughly comparable.

It has to be acknowledged that t-tests are "robust" with respect to violations of these assumptions, as long as the samples are not too small and there are equal numbers of subjects in both groups (in the case of the independent-means t-test).

If the data do not satisfy the requirements for a t-test, consider using a non-parametric test which does not make the three assumptions mentioned above. The Mann-Whitney test can be used instead of an independent-means t-test, and the Wilcoxon test can be used instead of a repeated-measures t-test.

Table of critical values of t:								
	One Tailed Significance level:							
	0.1	**0.05**	**0.025**	**0.005**	**0.0025**	**0.0005**	**0.00025**	**0.00005**
	Two Tailed Significance level:							
d.f.	**0.2**	**0.1**	**0.05**	**0.01**	**0.005**	**0.001**	**0.0005**	**0.0001**
2	1.89	2.92	4.3	9.92	14.09	31.6	44.7	100.14
3	1.64	2.35	3.18	5.84	7.45	12.92	16.33	28.01
4	1.53	2.13	2.78	4.6	5.6	8.61	10.31	15.53
5	1.48	2.02	2.57	4.03	4.77	6.87	7.98	11.18
6	1.44	1.94	2.45	3.71	4.32	5.96	6.79	9.08
7	1.41	1.89	2.36	3.5	4.03	5.41	6.08	7.89
8	1.4	1.86	2.31	3.36	3.83	5.04	5.62	7.12
9	1.38	1.83	2.26	3.25	3.69	4.78	5.29	6.59
10	1.37	1.81	2.23	3.17	3.58	4.59	5.05	6.21
11	1.36	1.8	2.2	3.11	3.5	4.44	4.86	5.92
12	1.36	1.78	2.18	3.05	3.43	4.32	4.72	5.7
13	1.35	1.77	2.16	3.01	3.37	4.22	4.6	5.51
14	1.35	1.76	2.14	2.98	3.33	4.14	4.5	5.36
15	1.34	1.75	2.13	2.95	3.29	4.07	4.42	5.24
16	1.34	1.75	2.12	2.92	3.25	4.01	4.35	5.13
17	1.33	1.74	2.11	2.9	3.22	3.97	4.29	5.04
18	1.33	1.73	2.1	2.88	3.2	3.92	4.23	4.97
19	1.33	1.73	2.09	2.86	3.17	3.88	4.19	4.9
20	1.33	1.72	2.09	2.85	3.15	3.85	4.15	4.84
21	1.32	1.72	2.08	2.83	3.14	3.82	4.11	4.78
22	1.32	1.72	2.07	2.82	3.12	3.79	4.08	4.74
23	1.32	1.71	2.07	2.81	3.1	3.77	4.05	4.69
24	1.32	1.71	2.06	2.8	3.09	3.75	4.02	4.65
25	1.32	1.71	2.06	2.79	3.08	3.73	4	4.62
26	1.31	1.71	2.06	2.78	3.07	3.71	3.97	4.59
27	1.31	1.7	2.05	2.77	3.06	3.69	3.95	4.56
28	1.31	1.7	2.05	2.76	3.05	3.67	3.93	4.53
29	1.31	1.7	2.05	2.76	3.04	3.66	3.92	4.51
30	1.31	1.7	2.04	2.75	3.03	3.65	3.9	4.48