2 Sample *t*-Test (1 tailed, equal variance)

Suppose we have two samples of ceramic sherd thickness collected from an archaeological site, where the two samples are easily distinguishable by the use of different styles to decorate the slip. However, the samples seem to be roughly similar in thickness (mm) suggesting that they might have had a similar function.

Therefore, we are interested in testing the hypothesis that mean difference in sherd thickness between the two samples is statistically zero, or in other words that there is statistically no difference between the sherd thickness of sample 1 and sample 2. As sample 1 seems to be slightly thicker on average than sample 2, we state the alternative hypothesis that sample 1 is thicker than sample 2. As such, let $\overline{Y}_1$ = the mean sherd thickness of sample 1, and $\overline{Y}_2$ = the mean sherd thickness of sample 2. Formally, we state the hypothesis at the *a* = 0.05 level:

$$H_O : \overline{Y}_1 - \overline{Y}_2 = 0$$
$$H_A : \overline{Y}_1 - \overline{Y}_2 \geq 0$$

Here are the data:

Sample 1

19.7475	30.5562	11.0734	18.1730	11.7280
19.8387	14.5291	19.4998	18.8374	12.2898
12.6873	14.7627	18.3869	17.9287	21.0552
17.6973	14.3439	10.7374	15.3563	21.4184
19.0878	12.5745	18.0030	18.6004	25.5953

Sample 2

17.4715	9.8613	23.4827	13.6029	18.1181
20.0386	19.6289	24.9357	17.8812	20.2681
12.6012	9.7741	19.9265	16.4178	14.7372
20.4401	15.1119	7.9955	5.1385	22.5915
22.4969	17.4448	17.6675	7.0984	16.7546

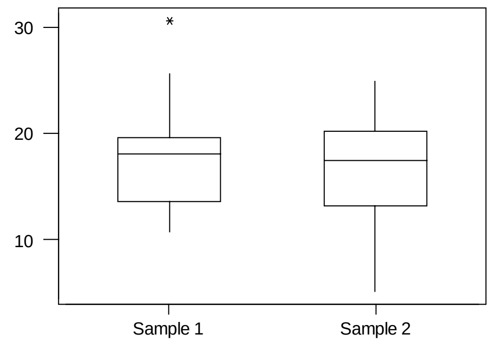
The descriptive statistics output is:

Descriptive Statistics

Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
Sample 1	25	17.380	18.003	17.096	4.615	0.923
Sample 2	25	16.46	17.47	16.58	5.29	1.06

Variable	Min	Max	Q1	Q3
Sample 1	10.737	30.556	13.516	19.624
Sample 2	5.14	24.94	13.10	20.15

From the descriptive stats we see that the means and medians are acceptable, and that the boxplots suggest both distributions are close enough to normal for us to use a parametric hypothesis test.



We see from the output and the boxplot that the variances (or standard deviations) are roughly equal. As we are interested in the statistics describing the variance *between* the two samples, we need to take into account the standard errors of both samples. As we have determined the variances are equal, we POOL the standard errors (se_P) using one of the following equations:

$$se_P = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{s_1^2 + s_2^2}{n}} \tag{1}$$

, as n is the same for both samples, or

$$se_p = \frac{s_1 + s_2}{2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{2}$$

, which should give an equivalent answer.

The *t*-test statistic is now calculated as:

$$t_{STAT} = \frac{\overline{Y}_1 - \overline{Y}_2}{se_p} \tag{3}$$

And the degrees of freedom used to look up the t_{CRIT} are $df = n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$.

If we wish to establish confidence limits around the mean of the difference between the two samples we calculate:

$$L_L = \left(\overline{Y}_1 - \overline{Y}_2\right) - t_{CRIT}\, se_p \tag{4}$$

$$L_L = \left(\overline{Y}_1 - \overline{Y}_2\right) + t_{CRIT}\, se_p \tag{5}$$

So, in terms of our example let us first find our t_{CRIT} from the table. We see that if df = 48 (25+25-2), and a = 0.05 = between 2.021 and 2.000, so let's call it 2.01.

Now, to calculate our pooled standard error let's use equation 2 (as this is the algorithm MINITAB uses). Plugging in the numbers we find:

$$se_p = \frac{4.615 + 5.29}{2}\sqrt{\frac{1}{25} + \frac{1}{25}} = 4.953\sqrt{\frac{2}{25}} = 1.401$$

So, if we want to calculate our t_{STAT} we plug the numbers into equation 3 and we get:

$$t_{STAT} = \frac{17.38 - 16.46}{1.401} = 0.657$$

Comparing our t_{STAT} to our t_{CRIT} we find that $0.657 < 2.01$, leading us to the conclusion that at the a = 0.05 level, we fail to reject the null hypothesis that there is no statistical difference between the sample means.

The above hypothesis test is a little abstract as the results are based on a distribution of values of differences between the samples, and so it can sometimes be hard to visualize exactly what is going on! However, a conceptually easier way to perform the same hypothesis test is to establish confidence limits. What we are doing conceptually is putting confidence limits around the difference between the two means to see whether a value of zero is a valid estimate of the difference between the two. Remember this is what our null hypothesis asks.

So, plugging our numbers into equations 4 and 5 we find:

$$L_L = (17.38 - 16.46) - 2.01(1.401) = -1.90$$
$$L_U = (17.38 - 16.46) + 2.01(1.401) = 3.74$$

As the lower bound is a negative number, and the upper bound is a positive number, the confidence encompass zero indicating that zero (our hypothesized difference) is a valid estimate of the mean difference between the two samples. This conclusion is obviously the same as the previous test, and we would treat our null hypothesis in the same way.

In MINITAB we follow these procedures:

Enter your two samples in two columns

>STAT

>BASIC STATS

>2 SAMPLE t

>Choose SAMPLES IN DIFFERENT COLUMNS

>Choose the alternative hypothesis (in this case GREATER THAN)

>Leave the confidence level at 95%

>Choose ASSUME EQUAL VARIANCES

>GRAPHS

>BOXPLOTS

>OK

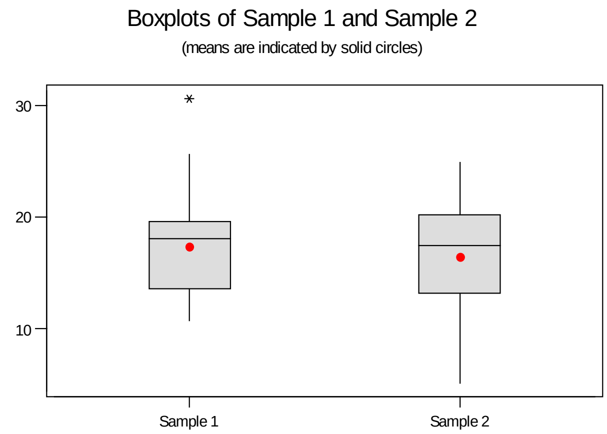
MINITAB gives both results in the output:

Two Sample T-Test and Confidence Interval

```
Two sample T for Sample 1 vs Sample 2
           N      Mean     StDev   SE Mean
Sample 1  25     17.38      4.62      0.92
Sample 2  25     16.46      5.29       1.1

95% CI for mu Sample 1 - mu Sample 2: ( -1.90,  3.7)
T-Test mu Sample 1 = mu Sample 2 (vs >): T= 0.66  P=0.26  DF=  48
Both use Pooled StDev = 4.97
```

And the boxplots look like:



To check that our math fits our computer output we see that the *Pooled StDev* in the output = 4.97 (we got 4.953, the difference due to rounding errors), and the *T* score in the output = 0.66 (we got 0.657 or 0.66). Finally, looking at our confidence limits we see the output gives us a 95% CI = (-1.90, 3.7), whereas we got (-1.90, 3.74). We could not manually calculate the *p* value, but the output tells us it is 0.26, which again fits our failure to reject the null hypothesis.

Although doing the *t*-test manually requires some mathematical trickery, it is essential for us to understand the decisions we are making when we choose different options in a software package as one small mistake coming from a basic misunderstanding can invalidate the whole test.

There are other variations to the *t*-test that we shall follow in the same way as we did here.