2 Sample *t*-Test (unequal sample sizes and unequal variances)

Like the last example, below we have ceramic sherd thickness measurements (in cm) of two samples representing different decorative styles from an archaeological site. However, this time we see that the sample sizes are different, but we are still interested in seeing whether the average thickness is statistically significant between the two samples or not.

Let $\overline{Y}_1$ = the sample mean of sherd thickness from sample 1, and $\overline{Y}_2$ = the sample mean of sherd thickness from sample 2. We wish to test the hypothesis at the $a = 0.05$ level (95%) that there is no statistical difference between the mean values of sample 1 and 2. Formally, we state:

$$H_o : \overline{Y}_1 - \overline{Y}_2 = 0$$
$$H_a : \overline{Y}_1 - \overline{Y}_2 \neq 0$$

If the data are normally distributed (or close enough) we choose to test this hypothesis using a 2-tailed, 2 sample *t*-test, taking into account the inequality of variances and sample sizes.

Below are the data:

Sample 1

19.7146	19.3516	20.8439	18.6316	23.7872
22.8245	29.1662	28.8265	22.4471	28.4952
26.3348	21.5908	23.8161	27.8443	27.9284
25.4338	25.0997	27.0340	25.3329	22.2871
20.8310	18.0220	23.5834	26.6790	13.2098

Sample 2

40.0790	24.2808	34.6926	37.1757	26.5954
18.5252	23.5064	30.9565	29.3769	19.7374
35.8091	39.7922	29.9376	40.7894	33.9418
26.5560	21.4682	23.9296	39.6987	30.6148
31.3332	13.1078	27.6245	27.1912	26.8967
39.6987	25.3269	37.2205	27.3089	28.4069
25.1476	30.2518	33.9531	36.1267	30.6148
29.6046	39.1803	32.0166	28.7846	33.8551

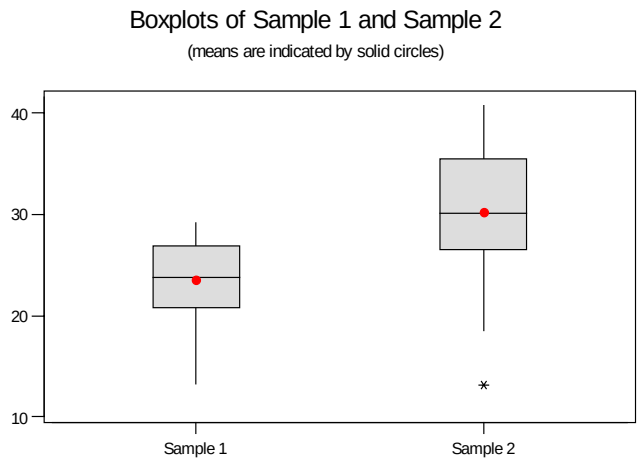
So first of all we need to look at our data, so we run the descriptive stats option in MINITAB and choose to present the samples graphically using a couple of boxplots.

Descriptive Statistics

Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
Sample 1	25	23.565	23.787	23.771	3.960	0.792
Sample 2	40	30.28	30.09	30.52	6.49	1.03

Variable	Min	Max	Q1	Q3
Sample 1	13.210	29.166	20.837	26.857
Sample 2	13.11	40.79	26.57	35.53

Looking at the descriptive stats output we see that the mean of sample 1 is smaller than sample 2, but also that the standard deviation of sample 1 is smaller than sample 2. So straight away we know we cannot assume equal variances as we did in the last example. We notice that the sample sizes are also different; we are also going to have to deal with this issue when calculating our degrees of freedom (v or df). However, we notice that the means are very similar to the medians in both samples, and the boxplots suggest that the data is close enough to normal to go ahead with the parametric test, the *t*-test.



So, as we know by now, as we are dealing with 2 samples we need to take into account the measures of dispersion of both samples, though in this case we know we cannot just take the average of the two (as we did in the last example) because the variations are very unequal. There is a standard method to deal with this contingency as, understandably, this situation arises much of the time in the real world. We use what is known as the *Satterthwaite Approximation*:

$$SE_S = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (1)$$

With this equation we see that we can take into account both unequal variances and unequal sample sizes at the same time, and as such, the Satterthwaite approximation

gives a weighted average of the standard errors. When the errors are equal, the Satterthwaite approximation gives roughly the same answer as the pooled variance procedure.

If we wish to calculate a *p* value and compare it to our *a*, the *t*-test statistic is now calculated in the same way as before:

$$t_{STAT} = \frac{\overline{Y}_1 - \overline{Y}_2}{se_p} \qquad (2)$$

However, we have to calculate our degrees of freedom to find our t_{CRIT}, and this is a little more complex this time as the sample sizes are unequal. In this case, the equation used to estimate *v* is:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}\right]} \qquad (3)$$

Okay, this looks ugly…and it is. We do not have to be concerned with the derivation of this equation, or even why exactly it works, we just have to plug in our numbers and chug through the equation when the time comes. We can check manual calculations with the MINITAB output as MINITAB uses this algorithm if we chose the right option when running the test (more later).

So let's plug and chug equation 3. To get the variances we square our standard deviations from the MINITAB output and plug the numbers in:

$$v = \frac{\left(\frac{15.6816}{25} + \frac{42.1641}{40}\right)^2}{\left[\frac{(15.6816)^2}{24} + \frac{(42.1641)^2}{39}\right]} = \frac{(0.627139 + 1.054102)^2}{(0.016388 + 0.028491)} = \frac{2.82657}{0.044878} \approx 63$$

You see that we round the result of this equation to the nearest integer, which is 63. To find our t_{CRIT} we then look up $v = 63$, $a = 0.05$ in the table and find our $t_{CRIT} = 2.000$ (approximately).

So, to calculate our standard error using the Satterthwaite approximation we plug and chug equation 1:

$$SE_S = \sqrt{\frac{15.6816}{25} + \frac{42.1641}{40}} = \sqrt{0.62726 + 1.0541} = 1.2967$$

Let us first calculate our t_{STAT} using equation 2:

$$t_{STAT} = \frac{23.565 - 30.28}{1.2967} = -5.18$$

We can see straight off that –5.18 standard errors is far way from the mean, in fact we calculated our t_{CRIT} to be + or -2.000 telling us already that we are going to end up rejecting our null hypothesis in favor of the alternative.

Let's now calculate our confidence limits (basic equations not shown):

$$L_L = (23.565 - 30.28) - 2.000 * 1.2967 = -9.308$$
$$L_U = (23.565 - 30.28) + 2.000 * 1.2967 = -4.122$$

We see that both bounds are negative numbers indicating that they do not encompass zero, therefore the hypothesis that there is no difference between the two samples is not supported by the data; we reject the null hypothesis in favor of the alternative, at the a = 0.05 level. The fact that both bounds are negative is a result of sample 1's mean being much smaller than sample 2 in addition to their variances.

Now let's run the test in MINITAB to check our results and see how our manual math skills held up against MINITAB's algorithms.

To perform this test 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 NOT EQUAL)

>Leave the confidence level at 95%

>DO NOT Choose ASSUME EQUAL VARIANCES; MINITAB will use the Satterthwaite approximation as a default

>OK

The output from MINITAB should look like:

Two Sample T-Test and Confidence Interval

```
Two sample T for Sample 1 vs Sample 2
           N      Mean     StDev   SE Mean
Sample 1  25     23.56      3.96      0.79
Sample 2  40     30.28      6.49       1.0

95% CI for mu Sample 1 - mu Sample 2: ( -9.31,  -4.1)
T-Test mu Sample 1 = mu Sample 2 (vs not =): T= -5.18  P=0.0000  DF=  62
```

First, you will see that MINITAB does not explicitly give us the Satterthwaite approximation of the standard error, but we will be able to tell if we were correct if the rest of the numbers turn out well.

Looking for the degrees of freedom (*df*) we see that MINITAB got 62, whereas we got 63. By the time we are dealing with 60-odd degrees of freedom the critical values do not change that much so the error is acceptable; in fact the error comes from rounding error in our manual calculations as we can only input so many decimal places, whereas MINITAB can use dozens. This brings up an important point; rounding errors get magnified when we start multiplying and squaring values so always use as many decimal places as you are given in such cases. The closeness of our degrees of freedom to the

MINITAB output lets us know that MINITAB probably used only one or two more decimal places.

The t_{STAT} in the output (T) is –5.18, the exact value we got manually indicating that our calculation of the Satterthwaite approximation was good, and as we expected, the *p* value is highly significant, therefore as $p < a$ we reject the null hypothesis in favor of the alternative. Remember that as we are dealing with a 2 Tailed *t*-test, the actual *a* value we are comparing our *p* value is $a/2 = 0.025$, as there are equal areas of rejection on both sides of our *t* distribution. Even so, we are still left with the same result.

Finally, looking at the confidence limits in the output, we see that our manual calculations are exactly the same as the output.

Our archaeological conclusion would be that the mean thickness of the sherds from sample 1 are much less than the mean thickness of sherds from sample 2, and the statistics indicate that they (most probably) do not belong to the same statistical population of vessels, under the assumption that vessel thickness is an accurate proxy measure of functional difference. Not only are they stylistically different, they are also most probably functionally different.