# Lesson 1: Hypothesis tests and Confidence Intervals for 2 independent samples

## Key Takeaways

By the end of this lesson you should be able to:

- Determine which case of independent samples we are dealing with
- Solve for and interpret a hypothesis test for two sample experiments
- Solve for and interpret a confidence interval for two sample experiments

## Section 10.1

Often experiments are performed to compare two (or more) objects, items, levels or categories to each other. Under this scenario we create a comparison between 2 (or more) samples, each having received a different characteristic of interest. For purposes of this course our interest will lie in comparing 2 different characteristics to each other. Hence, we now want to expand our use of Confidence Intervals and Hypothesis tests to the comparison of the means from two ***independent*** populations.

In such applications we are typically interested in two main points:

- Is there evidence of a difference in the means of two groups? This is typically answered using Hypothesis Tests and,
- To estimate the true difference in the means of the two groups using confidence intervals.

## Section 10.2: The Sampling Distribution of the Difference in Sample Means

### Notation

Let us start by expanding our notation. The easiest way to do this is to add a subscript 'i' to our previous symbols where i=1,2 depending on which sample we are referring to. And so we have:

Population parameters:

- Mean: $\boldsymbol{\mu_i}$ for $i = 1,2$
- Variance: $\boldsymbol{\sigma_i^2}$ for $i = 1,2$

Sample statistics:

- Sample of size $\boldsymbol{n_i}$ for $i = 1,2$
- Sample mean: $\boldsymbol{\bar{x}_i}$ for $i = 1,2$
- Sample Variance: $\boldsymbol{s_i^2}$ for $i = 1,2$

To describe the relationship between two population means, we will look at their difference $\boldsymbol{\mu_1 - \mu_2}$.

Inference procedures are based on the assumption that the observations come from (approximately) a normal distribution, or that sample sizes are large enough so that $\bar{X}_i$ is approximately normal, $i = 1,2$ by way of the central limit theorem.

### Building our Hypothesis Tests and Confidence Intervals

Recall we introduced our basic four steps in performing a hypothesis test. These are:

1. State your Hypothesis
2. Calculate the appropriate test statistic
3. Determine the significance
4. Make a conclusion

We must now adapt these steps to our new application.

To perform the hypothesis test and solve for the confidence interval we must first identify our parameter of interest and its appropriate sample estimator. Previously we stated that we are now trying to make inference on the difference in the population means, namely, $\mu_1 - \mu_2$. A natural estimator of $\mu_1 - \mu_2$ is $\boldsymbol{\bar{X}_1 - \bar{X}_2}$**.** It can be shown that if $\bar{X}_1$ and $\bar{X}_2$ are both normally distributed and ***independent*** then: $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

## Section 10.4: Hypothesis Tests and Confidence Intervals for $\boldsymbol{\mu_1 - \mu_2}$ when $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ are unknown

To define our hypothesis test steps and Confidence Interval under this application we need to further distinguish between two possible scenarios:

1. Case 1: **both $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ are unknown and $\boldsymbol{\sigma_1 \neq \sigma_2}$**
2. Case 2: **both $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ are unknown and $\boldsymbol{\sigma_1 = \sigma_2}$**

Under both scenarios the hypothesis stated is the same:

**Step 1** - State the Hypothesis

Again, the null hypothesis ($H_0$) captures the statement of no difference (i.e. both samples have the same average response). The alternate hypothesis ($H_a$) captures the statement of interest and can take on one of three forms.

| **Null Hypothesis, $H_0$** | **Alternate Hypothesis, $H_a$** | |
|---|---|---|
| $\boldsymbol{\mu_1 = \mu_2}$<br>This can also be expressed as:<br>$\boldsymbol{\mu_1 - \mu_2 = 0}$<br>Where '0' represents the hypothesized valued. | $\boldsymbol{\mu_1 \neq \mu_2}$ | Two-sided tests |
| | $\boldsymbol{\mu_1 > \mu_2}$ | One-sided Upper tail test |
| | $\boldsymbol{\mu_1 < \mu_2}$ | One-sided Lower tail test |

### Case 1: Hypothesis Tests for $\boldsymbol{\mu_1 - \mu_2}$ when both $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ are Unknown and $\boldsymbol{\sigma_1 \neq \sigma_2}$

**Referred to as the Welch Approximate t Procedure**

**Step 2** - Test Statistic: $\boldsymbol{t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$

**Step 3** – Critical value or p-value: Determined using the **t-table** as previously discussed, where the degrees of freedom are defined as,

$$\boldsymbol{d.f.} = \frac{\left(\frac{\boldsymbol{s_1^2}}{\boldsymbol{n_1}} + \frac{\boldsymbol{s_2^2}}{\boldsymbol{n_2}}\right)^{\mathbf{2}}}{\frac{\boldsymbol{s_1^4}}{\boldsymbol{n_1^2(n_1-1)}} + \frac{\boldsymbol{s_2^4}}{\boldsymbol{n_2^2(n_2-1)}}}$$

If performing the test by hand and using the t-table the degrees of freedom will be approximated as the smaller value between $\boldsymbol{n_1 - 1}$ and $\boldsymbol{n_2 - 1}$, stated as: $\mathbf{min}\ (\boldsymbol{n_1 - 1},\ \boldsymbol{n_2 - 1})$.

Next week, we will cover some shortcuts for these calculations in R.

**Step 4** – Conclusion as previously discussed

## Case 1: Confidence Interval for $\boldsymbol{\mu_1 - \mu_2}$ when $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ are unknown and $\boldsymbol{\sigma_1 \neq \sigma_2}$

Under this scenario, an approximate two-sided 100(1-$\alpha$)% CI for $\mu_1 - \mu_2$ is given by:

$$(\boldsymbol{\bar{x}_1 - \bar{x}_2}) \pm \boldsymbol{t_{\frac{\alpha}{2};df}} \sqrt{\frac{\boldsymbol{s_1^2}}{\boldsymbol{n_1}} + \frac{\boldsymbol{s_2^2}}{\boldsymbol{n_2}}}$$

## Example 1

A Bank manager has developed a new system to reduce the time customers spend waiting for teller service during peak hours. The manager hopes the new system will reduce the waiting times from the current nine to ten minutes to less than six minutes.

A random sample of 100 customers is drawn and their waiting time is recorded. It is found that their average waiting time is 5.46 mins with a standard deviation of 2.475 mins.

Use this information to test whether the average waiting time is now less than 6 mins.

**Soln**:

This is a 1 sample hypothesis test.

**Step 1:** $H_0: \mu = 6\ \ vs.\ \ H_a: \mu < 6$

**Step 2:** Notice that for this investigation the standard deviation is found by using the sample of 100 customers. Hence, we are working under the scenario that $\sigma$ is unknown which implies the t-test.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.46 - 6}{2.475/\sqrt{100}} = -2.1818$$

**Step 3:** To practice let us solve for the Critical Value and P-value.

Things to note:

- $\alpha = 0.05$
- We are performing a one-sided lower tail test
- $\sigma$ is unknown and so critical value and p-value will come from the t-distribution
- Degrees of freedom are n-1=99

First let us solve for the critical value:

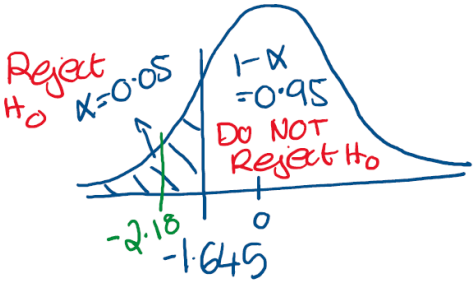
Critical value $=-t_{0.05;99} \approx -z_{0.05} = -1.645$

This critical value has very large degrees of freedom. As previously stated for such large degrees of freedom the z-distribution approximates the t-distribution very well and so we refer to the row labeled "∞" in our t-table to find the critical value.

Now let us solve for the p-value:

Since this is a one-sided lower tail test the p-value = $P(T_{99} < -2.18) \approx P(Z < -2.18) = 0.0146$. Again because of the large degrees of freedom we can jump in to the z-table for this.

**Step 4: Conclusion**




let us use both methods for practice:

Critical value:

here |-2.18|>|-1.645| and so we reject the null hypothesis at a 5% level of significance. Therefore, the data implies that the average waiting time is less than 6 minutes.

p-value:

here we have that p-value=0.0146 and is less than $\alpha = 0.05$ hence again we reject the null hypothesis at a 5% level of significance and conclude the same as above.

## Example 2

Consider again the Bank customer waiting time example. Under this example we have that the Bank manager adopted a new system in the hope of reducing waiting time during peak times. Previously we tested whether the average waiting time for the new system has been reduced to less than 6 mins.

Now we want to compare it directly to the average waiting time of the *current system* to determine whether it has in fact been effective at reducing waiting time. Two, independent random samples each of size 100 are drawn.

The first sample was subjected to the current waiting system and had an average waiting time of 8.79mins, with a standard deviation of 4.82mins.

The second sample was subjected to the new system and had an average waiting time of 5.14mins, with a standard deviation of 1.79mins.

Assuming that the population variances are not equal, test at a 5% level of significance whether the new system has in fact reduced the average waiting time.

**Soln**:

Let us start by summarising the information given to us in the question:

- Here we identify that we are working with two independent samples where the population standard deviations are unknown and NOT equal (i.e. $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ are unknown and $\boldsymbol{\sigma_1 \neq \sigma_2}$)
- Information on samples:

| Group 1: Current | Group 2: New |
|---|---|
| $n_1 = 100$ | $n_2 = 100$ |
| $\bar{x}_1 = 8.79$ | $\bar{x}_2 = 5.14$ |
| $s_1 = 4.82$ | $s_2 = 1.79$ |

**Step 1:** $H_0: \mu_1 = \mu_2 \; vs. \; H_a: \mu_1 > \mu_2$

Notice with the current system being labeled group 1 our alternate is testing that the average waiting time from the current system is larger than the new system.

**Step 2:** $t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(8.79 - 5.14) - 0}{\sqrt{\frac{4.82^2}{100} + \frac{1.79^2}{100}}} = 7.099$

**Step 3:** To practice let us solve for the Critical Value and P-value.

Things to note:

- $\alpha = 0.05$
- We are performing a one-sided upper tail test
- Degrees of freedom $= \boldsymbol{\min(n_1 - 1,\, n_2 - 1) = min(100 - 1, 100 - 1) = 99}$

First let us solve for the <u>critical value:</u>

Critical value $= t_{0.05;99} \approx z_{0.05} = 1.645$

This critical value has very large degrees of freedom. As previously stated for such large degrees of freedom the z-distribution approximates the t-distribution very well and so we refer to the row labeled "∞" in our t-table to find the critical value.

Now let us solve for the <u>p-value</u>:

Since this is a one-sided lower tail test the p-value = $P(T_{99} > 7.099) \approx P(Z > 7.099) \approx 0$. Again because of the large degrees of freedom we can jump into the z-table for this.

**Step 4: Conclusion**

Let us use both methods for practice:

<u>Critical value</u>:

here 7.099 > 1.645 and so we reject the null hypothesis at a 5% level of significance. Therefore, the data implies that on average the new system has a shorter waiting time.

<u>p-value</u>:

here we have that p-value$\approx 0$ and is less than $\alpha = 0.05$ hence again we reject the null hypothesis at a 5% level of significance and conclude the same as above.

### Example 3

Solve for a 90% confidence interval for the difference in average waiting time between the two systems in the bank. Interpret your interval.

Soln:

Let us start by listing the information given to us in the question. We have:

| **Group 1: Current** | **Group 2: New** |
|---|---|
| $n_1 = 100$ | $n_2 = 100$ |
| $\bar{x}_1 = 8.79$ | $\bar{x}_2 = 5.14$ |
| $s_1 = 4.82$ | $s_2 = 1.79$ |

Here we want to create a 90% confidence interval, i.e $90\% = (1 - 0.1)100\% \Rightarrow \alpha = 0.1$.

Hence, a 90% confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2};df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= (8.79 - 5.14) \pm t_{0.05;99} \times \sqrt{\frac{4.82^2}{100} + \frac{1.79^2}{100}}$$

$$= (2.8042; 4.4958)$$

Where $t_{0.05;99} \approx z_{0.05} = 1.645$

- Therefore are 90% confident that the true but unknown average difference lies between 2.8042 and 4.4958.
- Also notice that this confidence interval is entirely positive suggesting that on average the waiting time in the current system is larger than that of the new system.

### Case 2: Hypothesis Tests for $\mu_1 - \mu_2$ when both $\sigma_1$ and $\sigma_2$ are Unknown and $\sigma_1 = \sigma_2$

1. **Step 1**: State the Hypothesis:

| **Null Hypothesis, $H_0$** | **Alternate Hypothesis, $H_a$** | |
|---|---|---|
| $\mu_1 = \mu_2$<br>This can also be expressed as:<br>$\mu_1 - \mu_2 = 0$<br>Where '0' represents the hypothesized valued. | $\mu_1 \neq \mu_2$ | Two-sided tests |
| | $\mu_1 > \mu_2$ | One-sided Upper tail test |
| | $\mu_1 < \mu_2$ | One-sided Lower tail test |

**Step 2**- Test Statistic: $\boldsymbol{t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$

Where $\boldsymbol{s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

And represents the **Pooled Sample Variance.** This gives us an estimate of the common variance and is simply a weighted average of the two sample variances.

**Step 3** - Critical value or p-value: Determined using the **t-table** as previously discussed, where the degrees of freedom are defined as $\boldsymbol{df = n_1 + n_2 - 2}$

**Step 4** - Conclusion as previously discussed

## Case 2: Confidence Interval for $\boldsymbol{\mu_1 - \mu_2}$ when both $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ are Unknown and $\boldsymbol{\sigma_1 = \sigma_2}$

Under this scenario, an approximate two-sided 100(1-$\alpha$)% CI for $\mu_1 - \mu_2$ is given by:

$$\boldsymbol{(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2};\, n_1 + n_2 - 2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

## Example 4

A marketing research firm wishes to compare the prices charged by two supermarket chains- Miller's and Albert's. The research firm, using a standardised one-week shopping list, makes identical purchases at ten of each chain's stores. The stores for each chain are randomly selected, and all purchases are made during a single week.

Because the stores in each sample are different stores in different chains, it is reasonable to assume that the samples are independent, and we assume that the weekly expenses at each chain are normally distributed.

Based on the data collected it was found that the average weekly expense at Miller's was $121.92, with a standard deviation of 1.40 and at Albert's it was $114,81 with a standard deviation of 1.84. Assuming that the population variances are equal, test at a 10% level of significance the hypothesis that the average weekly expenses is different between the two stores.

**Soln**:

Let us start by summarising the information given to us in the question:

- Here we identify that we are working with two independent samples where the population standard deviations are unknown and equal (i.e. $\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ are unknown and $\boldsymbol{\sigma_1 = \sigma_2}$)
- Information on samples:

| Group 1: Current | Group 2: New |
|---|---|
| $n_1 = 10$ | $n_2 = 10$ |
| $\bar{x}_1 = 121.92$ | $\bar{x}_2 = 114.81$ |
| $s_1 = 1.40$ | $s_2 = 1.84$ |

**Step 1:** $H_0: \mu_1 = \mu_2 \; vs. \; H_a: \mu_1 \neq \mu_2$

**Step 2:** $t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(121.92-114.81)-0}{\sqrt{2.6728}\times\sqrt{\frac{1}{10}+\frac{1}{10}}} = 9.725$

Where $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{(10-1)(1.40)^2+(10-1)(1.84)^2}{10+10-2} = 2.6728$

**Step 3:** To practice let us solve for the Critical Value and P-value.

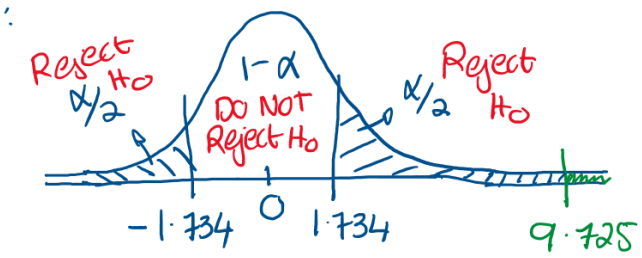
Things to note:

- $\alpha = 0.1$
- We are performing a two-sided test
- Degrees of freedom $= n_1 + n_2 - 2 = 10 + 10 - 2 = 18$

First let us solve for the critical value:
Critical value $= \pm t_{0.05;18} = \pm 1.734$

Now let us solve for the p-value:
Since this is a two-sided test the p-value = $2 \times P(T_{18} > 9.725) < 2 \times 0.0005 = 0.001.$

**Step 4: Conclusion**




let us use both methods for practice:
Critical value:
here 9.725 > 1.734 and so we reject the null hypothesis at a 10% level of significance. Therefore, the data implies that the average prices at the stores are different.

p-value:

here we have that p-value$< 0.001$ and is less than $\alpha = 0.10$ hence again we reject the null hypothesis at a 10% level of significance and conclude the same as above.

## Practice

Chapter 10: 3, 5, 6, 13, 14, 16 – 20, 25, 26, 28, 29, 32, 34

# Lesson 2: Hypothesis tests and Confidence Intervals for dependent (paired) data

## Key Takeaways

By the end of this lesson you should be able to:

- Solve for and interpret a hypothesis test for paired data
- Solve for and interpret a confidence interval for paired data

## Section 10.5: Paired-Difference procedures (for Dependent Samples)

### Independent vs. Dependent Groups

When comparing two population means to each other we need to first determine whether the two populations are related (dependent) or not related (independent) to each other. So far we have been working with independent groups.

- **Independent groups** occur when results from one group, **do not affect** the other. (i.e. the groups are said to be disjoint)
    - These groups are **non-paired**, example; males vs. females, smokers vs. Non-smokers.
- **Dependent groups** occur when our **pairs are related**, example; Twins, Matched pairs (matched based on certain criteria such as age e.t.c), Using the same unit (person) twice.

Dependent Samples lead to **Confounding.** To adjust for confounding we adopt a slightly different set-up when creating Confidence Intervals and Hypothesis Tests.

### Hypothesis Tests for a mean difference ($\mu_1 - \mu_2$) with paired data

Let $\boldsymbol{\mu_D = \mu_1 - \mu_2}$ represent the population mean difference. This will be estimated by

$$\boldsymbol{\bar{d} = \bar{X}_1 - \bar{X}_2}$$

which represents the ***difference*** of the ***sample means*** of the pairs.

**Step 1 –** State your hypothesis

| **Null Hypothesis, $H_0$** | **Alternate Hypothesis, $H_a$** | |
|---|---|---|
| $\boldsymbol{\mu_D = \mu_0}$ | $\boldsymbol{\mu_D \neq \mu_0}$ | Two-sided tests |
| | $\boldsymbol{\mu_D > \mu_0}$ | One-sided Upper tail test |
| | $\boldsymbol{\mu_D < \mu_0}$ | One-sided Lower tail tests |

**Step 2** – Solve for the appropriate test statistic $\boldsymbol{t = \frac{\bar{d} - \mu_0}{SE_{\bar{d}}}}$

Where $\boldsymbol{SE_{\bar{d}} = \frac{s_d}{\sqrt{n_d}}}$, where $\boldsymbol{s_d}$ is the standard deviation of **the differences** and $\boldsymbol{n_d}$ is the number of **differences**.

**Step 3** – Solve for the critical value or p value: Determined using the **t-table** as previously, where the degrees of freedom are defined around the number of pairs ($n_d$): $\boldsymbol{df = n_d - 1}$

**Step 4** – Make a conclusion as previously discussed.

## Confidence Interval for a mean difference ($\mu_1 - \mu_2$) with paired data

Under this scenario, an approximate two-sided 100(1-$\alpha$)% CI for the mean difference of paired data $\mu_D = \mu_1 - \mu_2$ is given by: $\boldsymbol{\bar{d} \pm t_{\frac{\alpha}{2};n_d-1} SE_{\bar{d}}}$

## Example 1

A supermarket chain wants to know if its "buy one, get one free" campaign increases customer traffic enough to justify the cost of the program. For each of 10 stores it selects two days at random to run the test. For one of those days, the program will be in effect. The chain wants to test the hypothesis that there is no mean difference in traffic against the alternative that the program increases the mean traffic. The results from the 10 stores are presented on the next slide.

| **Store #** | **Customer visits with Program (1)** | **Customer visits without program (2)** | Difference, $d_i$ |
|---|---|---|---|
| 1 | 140 | 136 | $140 - 136 = 4$ |
| 2 | 233 | 235 | $233 - 235 = -2$ |
| 3 | 110 | 108 | 2 |
| 4 | 42 | 35 | 7 |
| 5 | 332 | 328 | 4 |
| 6 | 135 | 135 | 0 |
| 7 | 151 | 144 | 7 |
| 8 | 33 | 39 | -6 |
| 9 | 178 | 170 | 8 |
| 10 | 147 | 141 | 6 |
| Mean | 150.1 | 147.1 | $\bar{d} = 150.1 - 147.1 = 3$ |
| Std. dev. | 86.98 | 86.33 | |

Using the information provided test the null hypothesis of no difference in average traffic at a 5% level of significance.

**Soln**:

Let us start by summarising the information given to us in the question:

- Here the data is paired based on the store location.
- In total we have 10 stores, i.e 10 pairs⇒ $n_d = 10$
- $\bar{d} = \frac{\sum_{i=1}^{10} d_i}{10} = 150.1 - 147.1 = 3$, because of the linearity properties of the mean.
- $s_d = \sqrt{\frac{\sum_{i=1}^{10}(d_i - \bar{d})^2}{10-1}} = 4.52155$

**Step 1:** $H_0: \mu_D = 0 \ \ vs. \ \ H_a: \mu_D \neq 0$

**Step 2:** $t = \frac{\bar{d}-\mu_0}{SE_{\bar{d}}} = \frac{3-0}{4.52155/\sqrt{10}} = 2.09814$

**Step 3:** To practice let us solve for the Critical Value and P-value.

Things to note:

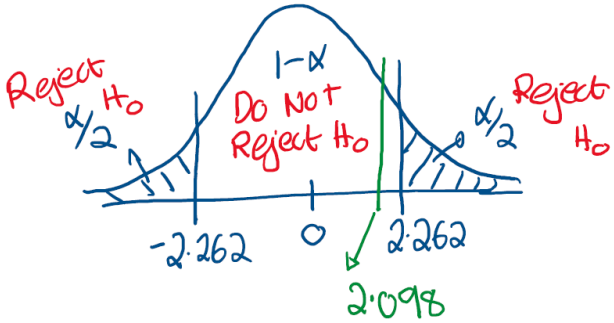
- $\alpha = 0.05$
- We are performing a two-sided test
- Degrees of freedom $= n_d - 1 = 10 - 1 = 9$

First let us solve for the critical value: Critical value $= \pm t_{0.025;9} = \pm 2.262$

Now let us solve for the p-value: Since this is a two-sided test the p-value = $2 \times P(T_9 > 2.09814)$.

From the t-table at 9 degrees of freedom the t-value of 2.09814 lies between the columns highlighted by $t_{0.025}$ and $t_{0.05}$. Hence, $0.025 < P(T_9 > 2.09814) < 0.05 \Rightarrow 0.05 < p - value < 0.1$

**Step 4: Conclusion** let us use both methods for practice:




Critical value: here 2.098 lies between the critical values and so we do NOT reject the null hypothesis at a 5% level of significance. Therefore, the data implies that the average difference is not different to 0.

p-value: here we have that 0.05< p-value$< 0.1$ and is less than $\alpha = 0.10$ hence again we do not reject the null hypothesis at a 10% level of significance and conclude the same as above.

Have you noticed something from these examples? The paired-difference procedure is the SAME as the one-sample procedure we learned previous, but now our data is the differences.

## Practice

Chapter 10: 9, 10, 31, 40, 41