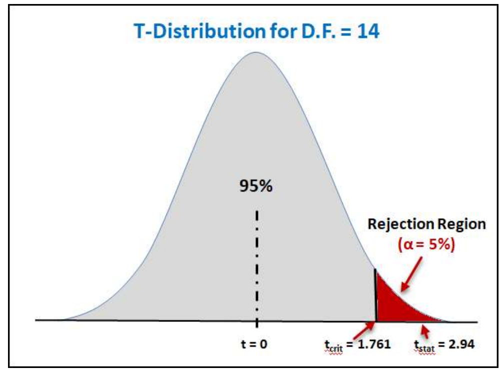
Introduction to Hypothesis Testing



It is the mark of a truly intelligent person to be moved by statistics. - George Bernard Shaw

A statistical tool that often moves people is the **Hypothesis Test** which is our next topic within **Inferential Statistics**!

Remember - anytime you see **Inferential Statistics**, you need to immediately recognize that *you're about to take **sample data** and somehow **make a conclusion** about the **population** associated with that sample*.

With Hypothesis Testing we're *using sample data to **test an assertion or a hypothesis** associated with the population being sampled from.*

Specifically, **hypothesis testing** is defined as a *statistical process used to make a decision between a null hypothesis and alternative hypothesis based on information in a sample.*

This chapter starts with a **refresher of inferential statistics**, specifically about **sampling distributions** which are an important concept in hypothesis testing.

We then move on to the **6-step process for performing a hypothesis test** which is the meat & potatoes of the chapter.

From here we go into the **specific population parameters & situations** where a hypothesis test can be conducted:

- The Population **Mean** using the **Normal Distribution** & **T Distribution**
- The Population **Variance** & Standard Deviation using the **Chi-Squared** Distribution & **F-Distribution**
- The Population **Proportion** using the **Normal Distribution**

We wrap up the chapter with a few additional concepts like:

- **Type I and Type II Errors** in Hypothesis Testing
- The **Power** of a Hypothesis Test
- **Statistical Significance** versus **Practical Significance**
- Using The **P-Value** Method

Refresher on Inferential Statistics for Hypothesis Testing

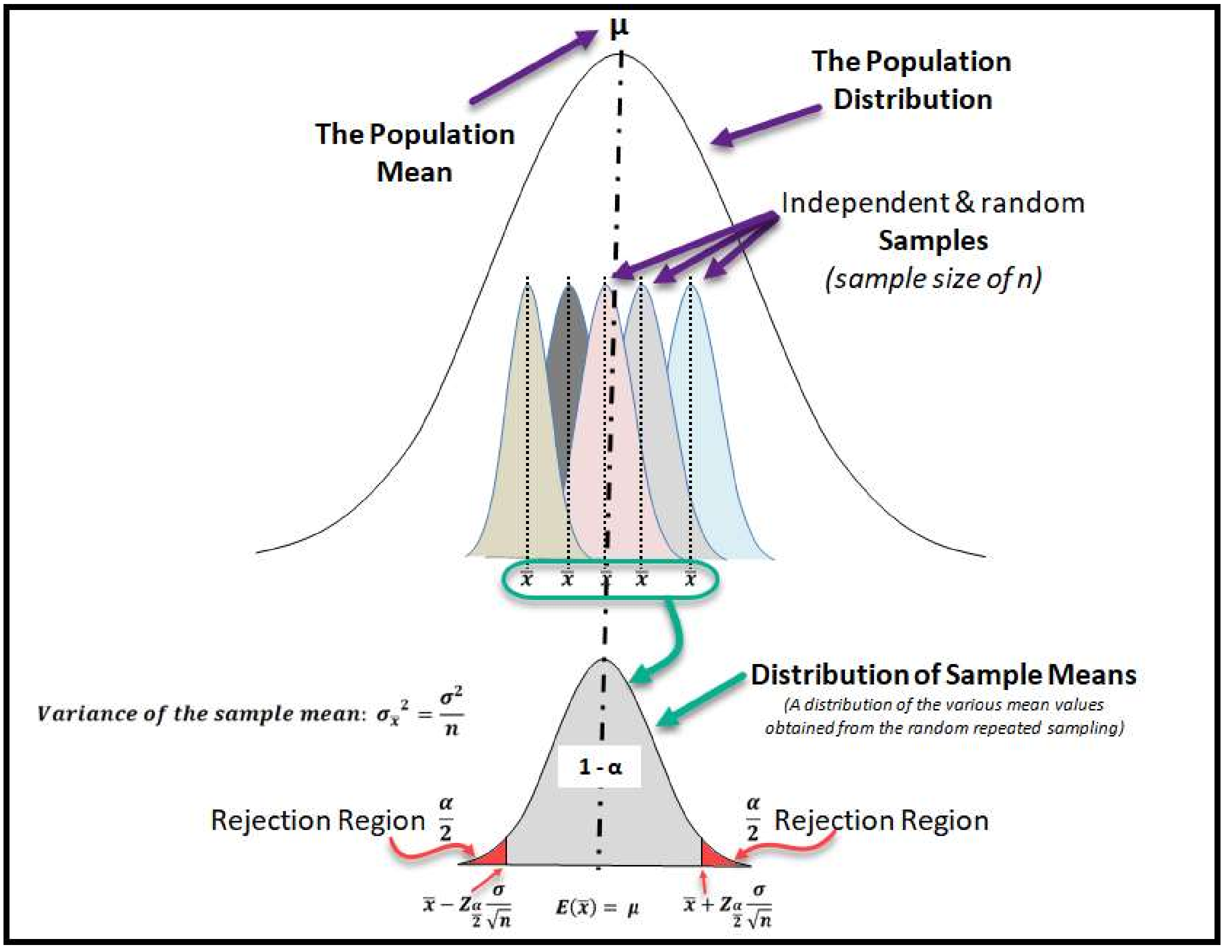
Alright, let's quickly review the idea of inferential statistics so you can see how hypothesis testing fits into the big picture.

Ok, remember that in **inferential statistics we're using sample data to make inferences about the population.**

I've shown that in the image below where we start with our **population distribution** where we often **do not know any of the population parameters** like the population mean or the population variance.

*However, we can use **sample data** in order to **test hypotheses** about these **population parameters**.*

We are able to test these hypotheses using the distribution of our sample statistics (sample mean, sample variance, etc); for example, below is the **Distribution of Sample Means**.



This **distribution of sample means** has a mean value that is assumed to be equal to the population mean. This assumption is a reflection of the fact that the sample mean is an **unbiased estimator** of the population mean.

The **distribution of sample means** has a **Variance** that is equal to the population variance divided by n (the sample size):

$$Variance\ of\ sample\ mean: V(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

The beauty of this sample distribution is that we can now take a sample and see where that sample mean falls on this distribution in terms of its probability of occurrence & use that to test our hypothesis about population parameters.

Make sense?

The Big Assumption of Hypothesis Testing

Hypothesis testing is based on a huge assumption that you must familiarize yourself with before you start.

This assumption is analogous to the starting assumption within the U.S. Criminal Justice System*: **Innocent until proven guilty.***

Every criminal trial in the U.S. begins with the judge and jury assuming the defendant is innocent. That's the **null hypothesis** in criminal justice, that the defendant is not guilty (innocent).

Hypothesis testing is a lot like the criminal justice system, we start every test assuming the null hypothesis is true.

The **alternative hypothesis** in Criminal Justice is that the defendant is guilty.

Prosecutors must use evidence (data) to prove that the null hypothesis is false, and that the alternative hypothesis is true.

This is exactly like hypothesis testing where *we start based on the assumption that the null hypothesis is true, then we must collect and analyze data in an attempt to reject that null hypothesis.*

I'll say it again because it's a very important assumption.

Hypothesis testing is based on the initial assumption that the null hypothesis is true, and **the null hypothesis can only be rejected if significant evidence suggests that it is not true**.

Now the question is, what type of evidence (data in our case) is needed to reject the null hypothesis in favor of the alternative hypothesis.

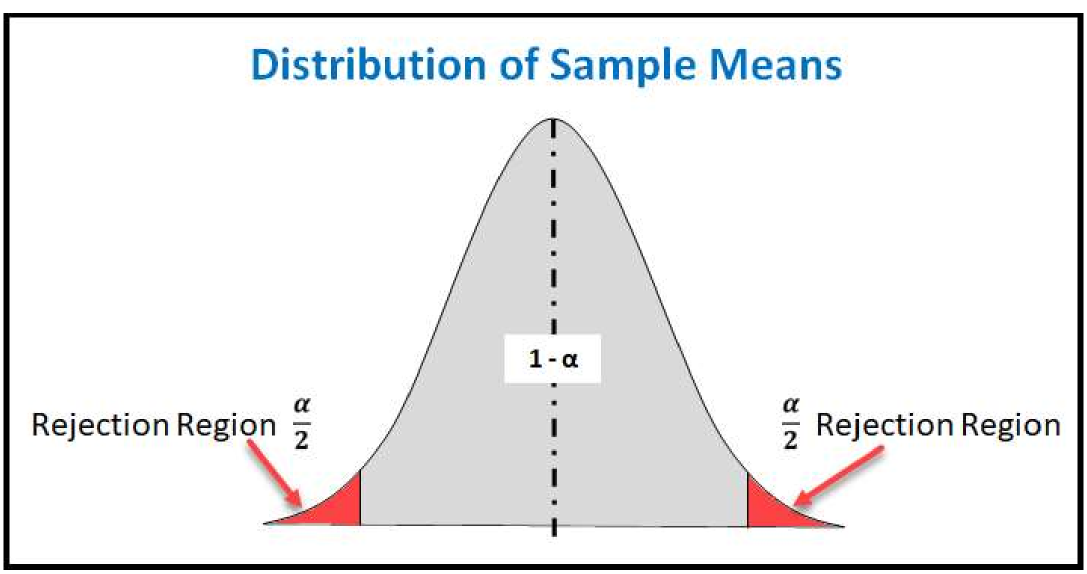
If the null hypothesis is correct (like we assume) then the sample we're measuring is likely to take a value near the expected value (μ_0, σ_0, etc), and is unlikely to take a value that's far away from the expected value.

What hypothesis testing does is ***it reflects the probability of getting the sample mean that was measured***.

If the probability is really small, then the null hypothesis is probably not true, and we can reject it in favor of the alternative hypothesis.

This is why I started with the section on a refresher on inferential statistics and specifically covered the idea of a sampling distribution (Image below).

Hypothesis testing makes use of the sampling distribution to assess the probability of the sample mean, thus providing a tool to distinguish between the null and alternative hypothesis.



The 6 Step Process of Hypothesis Testing

Ok, let's jump into the formal 6 step process for hypothesis testing.

A quick comment - this process works for whatever population parameter you're attempting to test, so pay attention!

- **Step 1.** Identify the **Null Hypothesis** (H_0) and the **Alternative Hypothesis** (H_a).
- **Step 2.** Choose the **Significance Level**, α.
- **Step 3.** Determine the **Rejection Region** for the Statistic of Interest (Mean, Variance, Proportion, etc).
- **Step 4.** Calculate the **Test Statistic** from your Sample Data
- **Step 5.** **Compare** the Test Statistic against the Rejection Criteria and **Make a Conclusion**
- **Step 6.** State the Decision in Terms of the **Original Problem Statement**

Let's use the following example to walk through each of these steps in the process:

> We've recently made a design change to our motor that we believe will result in an increase in the horsepower of the motor. Historically, our motor has been rated at 100 horsepower, but we believe our design change will be greater than that.
>
> H*ow can we use a hypothesis test to prove that our new design has more horsepower than the previous design?*

Step 1 - Identify the Null Hypothesis (H_0) & Alternate Hypothesis (H_a)

As we discussed above, the null hypothesis is important because it will be assumed to be true until it can be proven to be false.

This null hypothesis, called **H_0, is always a statement about the value of a population parameter.**

The null hypothesis typically represents the status quo or what has historically been true, and **will always contain the equals sign**.

There are 3 different ways to write your null hypothesis, and they're dependent on the situation being evaluated.

Left Tail Test	Two Tail Test	Right Tail Test
$H_0: \mu \geq \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$
$H_a: \mu < \mu_0$	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$

Each of these three null hypotheses comes with a unique **alternative hypothesis** that is **always mutually exclusive** from the null hypothesis.

The alternate hypothesis, usually called H_a or H_1, is a statement about the population parameter that is contradictory to the null hypothesis and is accepted as true only if there is convincing evidence in favor of it.

The alternative hypothesis usually contains the statement that we are attempting to prove and support.

Remember the **null hypothesis** generally represents the status quo, so our null hypothesis for the motor example should be based on the assumption that the design change did not improve horsepower.

Because we want to prove that our design change will improve the horsepower of, the **alternative hypothesis** that we'd like to prove is that the new mean value (**μ**) is greater than the historical mean value (**μ_0**).

So the appropriate **null & alternative hypothesis** is a **right tail test**:

Right Tail Test: $\quad H_0: \mu \leq \mu_0 \quad \& \quad H_a: \mu > \mu_0$

Step 2 - Choose the Significance Level (α)

So we've now got a null hypothesis and an alternative hypothesis, but how do we make a decision as to which hypothesis is correct?

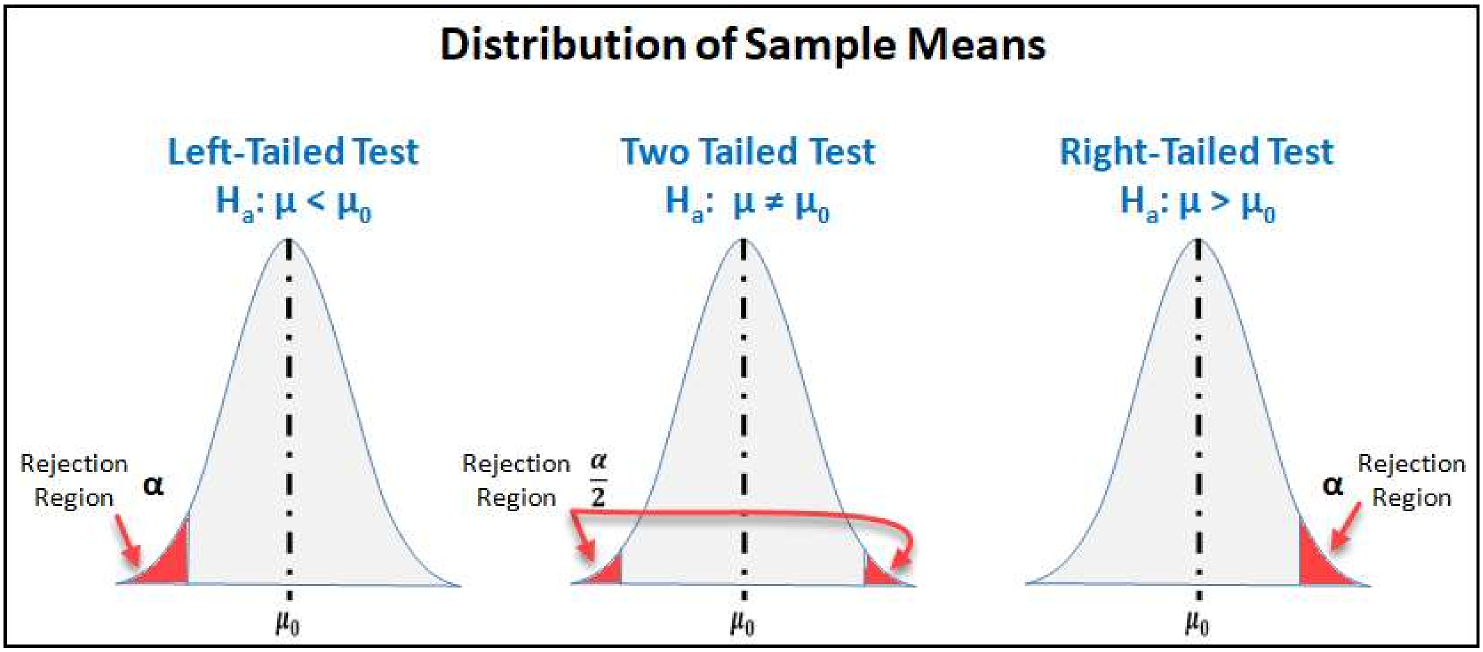
The criteria for judging between the null and alternative hypothesis can be stated as such: **if the value of the sample mean would be highly unlikely to occur if H_0 were true then we reject H_0 in favor of H_a.**

The way we judge a result to be "highly unlikely" is based on the significance level we choose.

Typically, the significance level is quite small, with 5% being the most common significance level used. Other common significance levels include 1%, 10%, etc.

You can see that graphically here where **the significance level determines the rejection criteria for our hypothesis test**.

That area in red represents 5% of the distribution, because that's the significance level we've chosen.



The image above shows what the **significance level & rejection region** looks like for the 3 different possible null hypotheses.

In addition to helping us discern between the null and alternative hypothesis, the **significance level is also equal to the probability of a type I error.**

A type I error is the risk that you'll reject the null hypothesis when it is actually true. We will discuss this more below, but it's an important point to note here.

When picking your significance level you have to understand the level of risk you're willing to accept that your hypothesis test might be erroneous.

By the way, this level of significance is the same α we used when creating confidence intervals.

In our specific example, we're going to use the typical **significance level of 5%**, and because our example is a **right tail test**, it would match the image on the far right where the entire rejection region would fall in the right tail.

Step 3 - Determine the Rejection Criteria for the Statistic of Interest

What we've learned so far is that our rejection region is a combination of the significance level, the type of test (one-tailed v. two tailed) and the statistic of interest being tested.

The last criteria needed to determine the rejection criteria is the sample size associated with your sample data.

Technically, if you're using the z-distribution, the sample size doesn't matter, but when using the t-distribution it does matter, and for our example we're going to use the t-distribution.

For our example, let's sample 15 motors and test their horsepower.

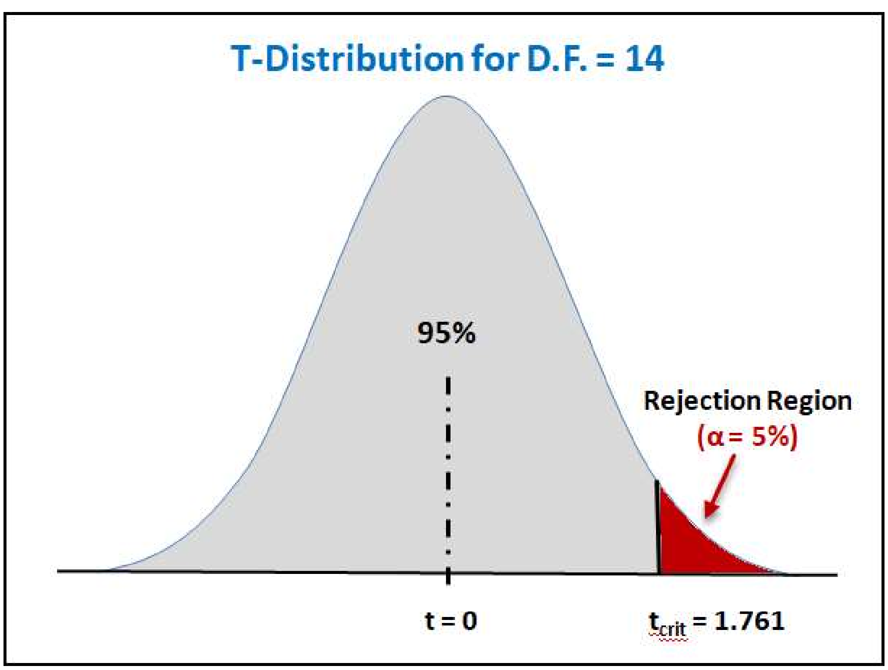
To find the rejection criteria for our example, we use the **t-distribution tables** where based on our **significance level (5%) and sample size (15)** we can look up the **critical t-value** to be **t_{crit} = 1.761.**

Critical Values of the Student T's Distribution at Various Degrees of Freedom and Alpha Levels (α)

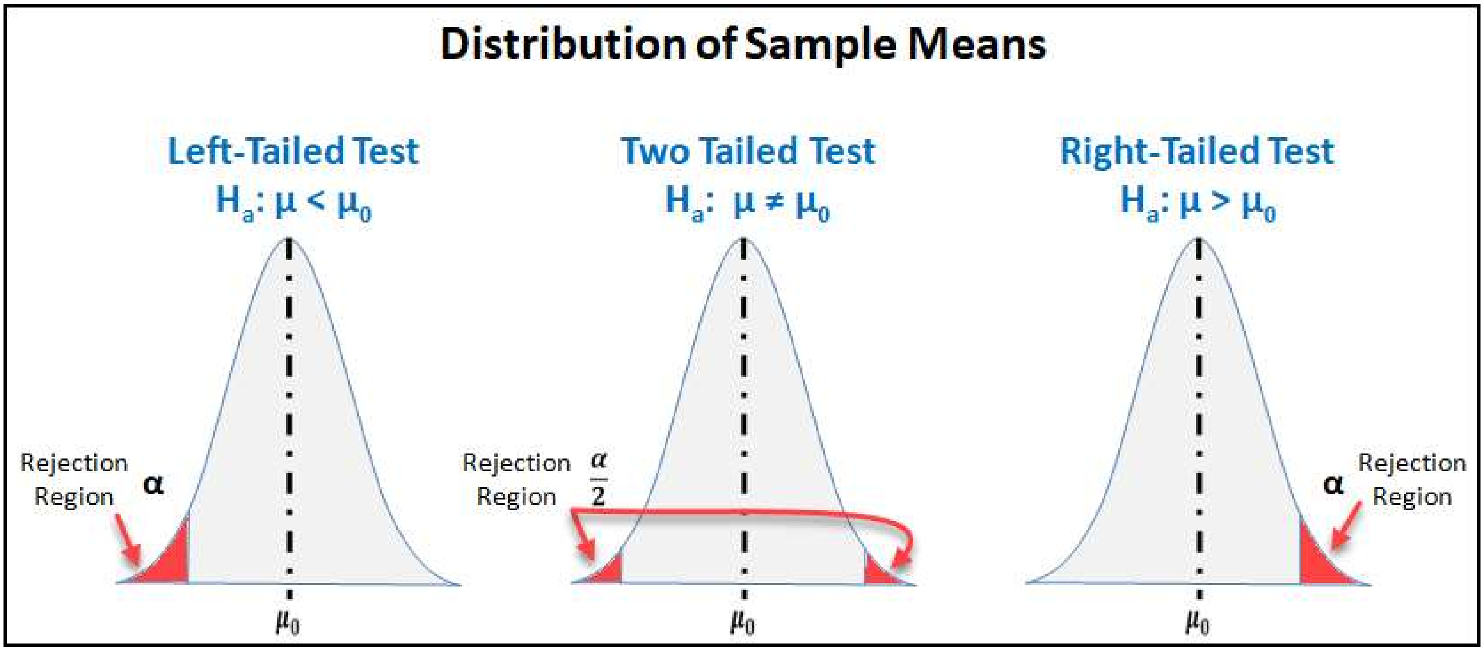
df (v)	α = 0.1	α = 0.05	α = 0.025	α = 0.01	α = 0.005	α = 0.001
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733

You can see this **t_{crit}** value below where we've got a **t-distribution** (degrees of freedom = 14) and **the area to the right of t = 1.761 form the rejection region for our example**.

The way you interpret this rejection region is that any **t-value greater than our critical t-value would be highly unlikely to occur if H_0 were true** (because it's so far away from the mean). Thus, providing the evidence we need to reject H_0 in favor of the H_a.



Before we jump into the next section though, I wanted to show you the rejection region for each of the three different null hypothesis again.



On the left is the **left tail test** where the full **alpha risk** is placed within the left tail of the distribution.

The **right tailed test** on the right-hand side is similar in that the full alpha risk is placed within the right tail of the distribution.

For the two-tailed test, the alpha risk is being equally distributed between the two tails of the test.

This distribution of risk is the same regardless of whether you're using the z-distribution or the t-distribution.

Heck, this logic even still applies when we're using the chi-squared distribution which is used for hypothesis testing of the variance & standard deviation.

Step 4 - Calculate the Test Statistic from Your Sample Data

Once you've determined your critical rejection values you can then calculate your test statistic from your sample data.

Let's look at our sample data from the 15 motors we tested whose horsepower values are:

101.5, 99.5, 103.5, 102, 100.5, 99.0, 104.0, 102.5, 100.5, 101.0, 98.5, 101.5, 102.5, 101.5, 100.0

From this data set we can calculate our sample mean as 101.2 HP, and our sample standard deviation as 1.58.

Using this data, we can **calculate the t-statistic** as:

$$t - statistic = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{101.2 - 100}{\frac{1.58}{\sqrt{15}}} = \frac{1.2}{0.41} = 2.94$$

Step 5 - Compare the Test Statistic against the Rejection Criteria and Make a Conclusion

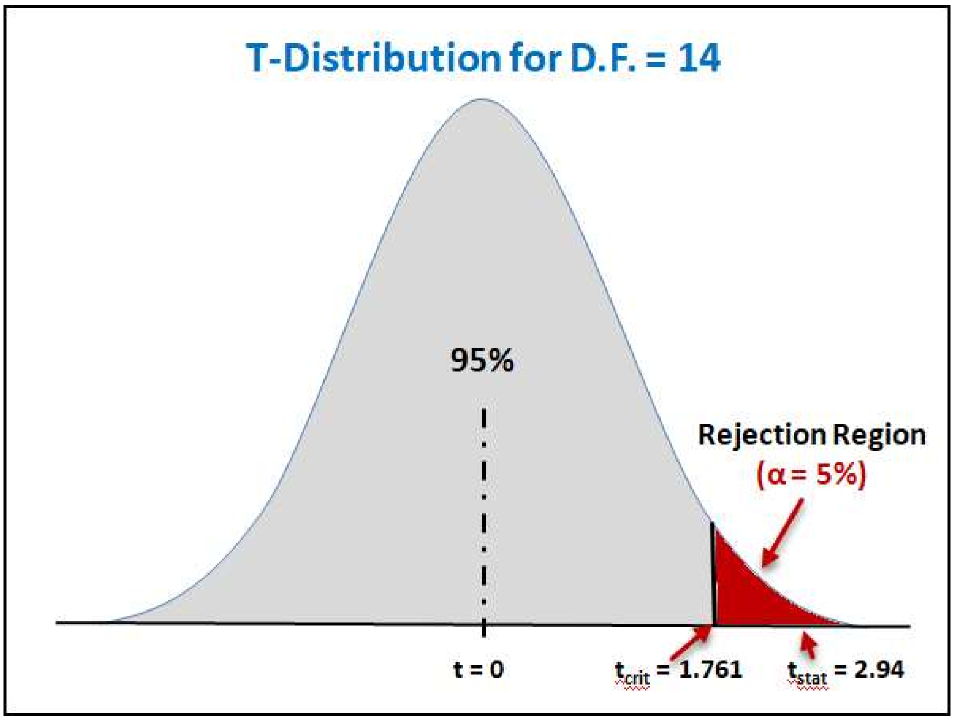
The end result of a hypothesis test is a choice of one of the following **2 possible conclusions**:

- Rejection of the null hypothesis, and therefore acceptance of the alternative hypothesis
- A failure to reject the null hypothesis and therefore fail to accept the alternative hypothesis

This is all based on a comparison of your test statistic against the rejection criteria.

In our specific example, the **critical t-value** was **1.761** which captures the alpha risk of 5% for a 1-tail distribution and 14 degrees of freedom.

The **t-statistic** associated with our sample data was **2.94** meaning that our t-statistic was greater than our t-critical, which looks like this:



Because our test statistic falls within the rejection region, then we can reject the null hypothesis and therefore accept the alternative hypothesis.

This result is considered a "**strong claim**". This means that the sample data was **significant enough** to reject the starting assumption that the null hypothesis was true in favor of the alternative hypothesis.

Let's say, for example, that our t-statistic came in at 1.000. This is less than the critical t-value and does not fall within the rejection region. Had this been the case, we would fail to reject the null hypothesis.

This result, of *failing to reject the null hypothesis* is considered a **weak claim**. Which means that there was likely not enough evidence to reject the null hypothesis, but we certainly didn't prove that the null hypothesis was true.

A common miss-interpretation in hypothesis testing is when people **"accept" the null hypothesis**. Generally, instead of accepting the null hypothesis, statisticians prefer to say that we "fail to reject" the null hypothesis.

Failing to reject the null hypothesis implies that the sample data was not sufficiently persuasive for us to accept the alternative hypothesis.

Usually in hypothesis testing the data is not sufficiently persuasive enough to completely accept the null hypothesis as true.

Step 6 - State the conclusion in terms of the original problem statement

The final step in the process is to restate the results of the hypothesis test back into the form of the original problem statement.

In our example it might look something like this:

The null hypothesis was that the new motor design would be equal to or less than the historical average of 100. The alternative hypothesis is that the new motor design would demonstrate an increase in horsepower when compared to the historical average of 100.

The hypothesis testing performed demonstrated, at the 5% significance level, that we can reject the null hypothesis and accept the alternative hypothesis that the new motor design demonstrates an increase in horsepower.

You get the point.

Hypothesis Testing for the Population Mean

Ok, when performing hypothesis testing for the **population mean**, there are two options.

The first option is to use the **normal distribution and the Z-transformation**. This option can only be used when the population standard deviation is known or when the sample size is greater than 30.

The second option is to use the **t-distribution**. This option must be used if your sample size is less than 30, or the population standard deviation is unknown.

The t-distribution can be safely used in situation as the t-distribution approximates the normal distribution when the sample size is greater than 30.

Hypothesis Testing for Means Using the Z-Score (Known Population Standard Deviation)

Ok, let's start with the first option, when we use the **Normal Distribution & the Z-transformation** to perform your **hypothesis test**.

This type of hypothesis test is used with testing the population mean.

Remember, this option only works when you know the population standard deviation and when your sample size is greater than 30.

The 6 step process outlined above fully applies to this situation. Where this situation becomes unique is in step 3 & 4, when you're determining your rejection criteria and determining your test statistic (z-statistic).

Let's use an example to show you how these uniqueness's work.

Example of Hypothesis Testing for Means Using the Z-Score

You're filling bottles of glue and your historical fill height is 3.45 inches. You sampled 30 units to confirm that your recent batch meets the specifications.

You know the population standard deviation is 0.50 inches, and the sample mean you measured is 3.43 inches. Using a level of significance of 5%, test the hypothesis that the true average fill height is 3.4h inches. Assume the fill height is normally distributed.

Step 1: State the null & alternative hypothesis.

Because our fill height might have shifted in either direction (up or down), we want to perform a two tailed test, so the null Hypothesis would be that our sample mean value is equal to the known population mean (3.45 inches).

$$H_0: \mu = \mu_0 \qquad\qquad H_a: \mu \neq \mu_0$$

The alternative hypothesis is that the mean value is different that the historical

Step 2: Choose the Significance Level, α.

The significance level is in the problem statement, it's 5%.

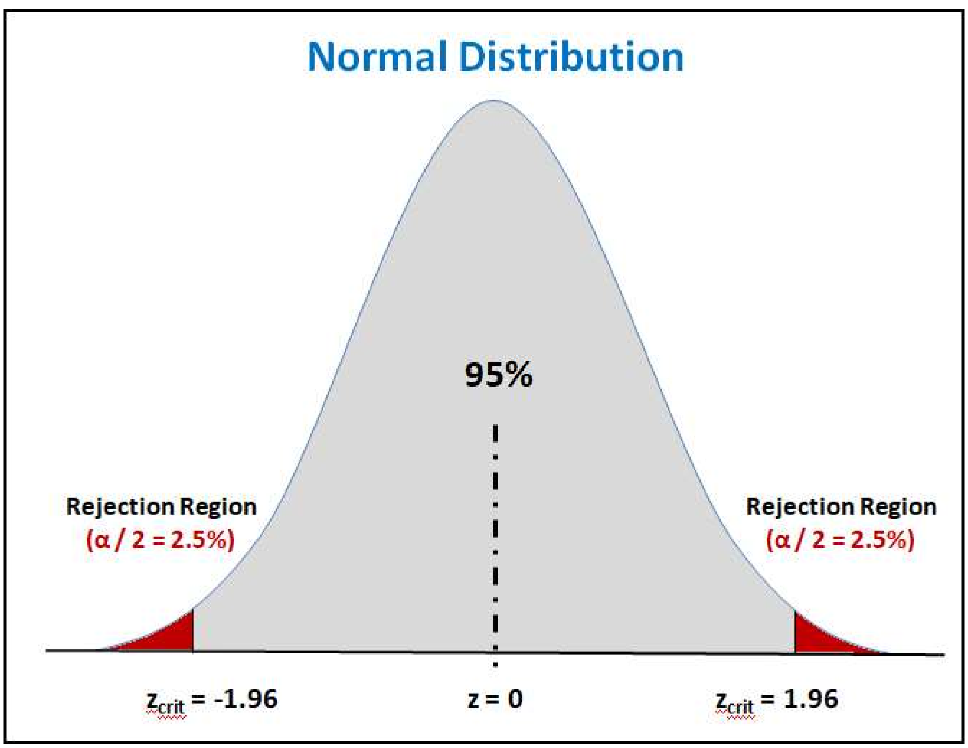
Step 3: Determine the Rejection Region for the Statistic of Interest

Ok, so this is where we run into something new compared to the 6-step process identified above.

First, is that our hypothesis test is now a 2-sided test as opposed to a 1-sided this. This means that our alpha risk is distributed between both tails of the normal distribution.

The critical z-score can be found using the NIST Z-Table. Remember that we're looking for the Z-score that captures 47.5% of the area of the distribution. This is at **Z = 1.96 and Z = -1.96**.

Graphically, this looks like this:



Step 4: Calculate the Test Statistic from your Sample Data

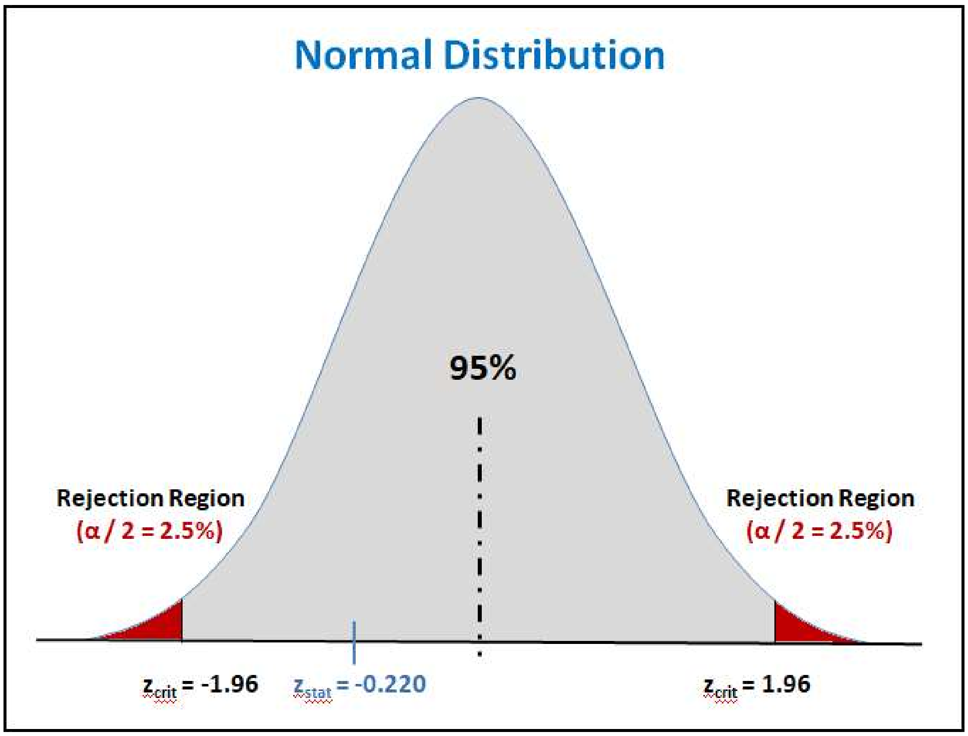
This is another unique situation with the normal distribution. In the example above we calculated the T-statistic, and now with the normal distribution, we're going to calculate the z-statistic.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3.43 - 3.45}{\frac{0.50}{\sqrt{30}}} = -0.220$$

Step 5: Compare the Test Statistic against the Rejection Criteria and Make a Conclusion

Let's compare our test statistic (-0.220) against our rejection criteria (-1.960, 1.960).

Our test statistic is not greater than our rejection criteria, therefore, **we fail to reject the null hypothesis** which graphically looks like this:



Step 6: State the Decision in Terms of the Original Problem Statement

Based on the sample data, we fail to reject the null hypothesis whereby the fill height of the glue bottles is not statistically significantly different than the historical fill height.

This makes sense as our measured sample mean is very close to the historical population mean, and well within the expected amount of variability given the standard deviation of 0.50 inches.

Hypothesis Testing for Means Using the T-Distribution

Alright, let's jump to the second option for hypothesis testing of means, which is the use of the t-distribution.

Remember, this option must be used if your sample size is less than 30, or the population standard deviation is unknown.

The t-distribution can be safely used in situation as the t-distribution approximates the normal distribution when the sample size is greater than 30.

The 6-step process outlined above fully applies to this type of hypothesis test.

Example of Hypothesis Testing for Means Using the T-Distribution

We've got a car engine whose emission have historically averaged 10.0 ppm. We recently made a change to our muffler and we believe that the emissions will be reduced.

We've tested 5 new engines measured the following emissions in ppm: **9.9, 10.1, 9.8, 9.9 and 9.9.**

Use a hypothesis test at a 10% significance level to determine if the population mean is less than or equal to 10.0ppm.

Step 1 - Identify the Null Hypothesis (H_0) and the Alternative Hypothesis (H_a).

In this instance, because we wish to assess if our muffler has reduced emissions, which means we should pick a left tail test whose null and alternative hypothesis look like this:

$$\textbf{Left Tail Test:} \quad H_0: \mu \geq \mu_0 \quad \& \quad H_a: \mu < \mu_0$$

Step 2 - Choose the Significance Level, α.

The significance level is there in the problem statement, it's 10%.

Step 3 - Determine the Rejection Region for the Statistic of Interest (Mean, Variance, Proportion, etc).

Because we've only sampled 5 units, we must use the t-distribution. When we combine the degrees of freedom (4) with our significance level (10%), we can look up the critical t-value in the **T-distribution Table** to be **1.533**.

Because we're performing a left tail test, that translates to a critical t-value of -**1.533.**

Step 4 - Calculate the Test Statistic from your Sample Data

Using our sample data (9.9, 10.1, 9.8, 9.9 and 9.9), we can calculate a sample mean of 9.92 ppm, and a sample standard deviation of 0.11 ppm.

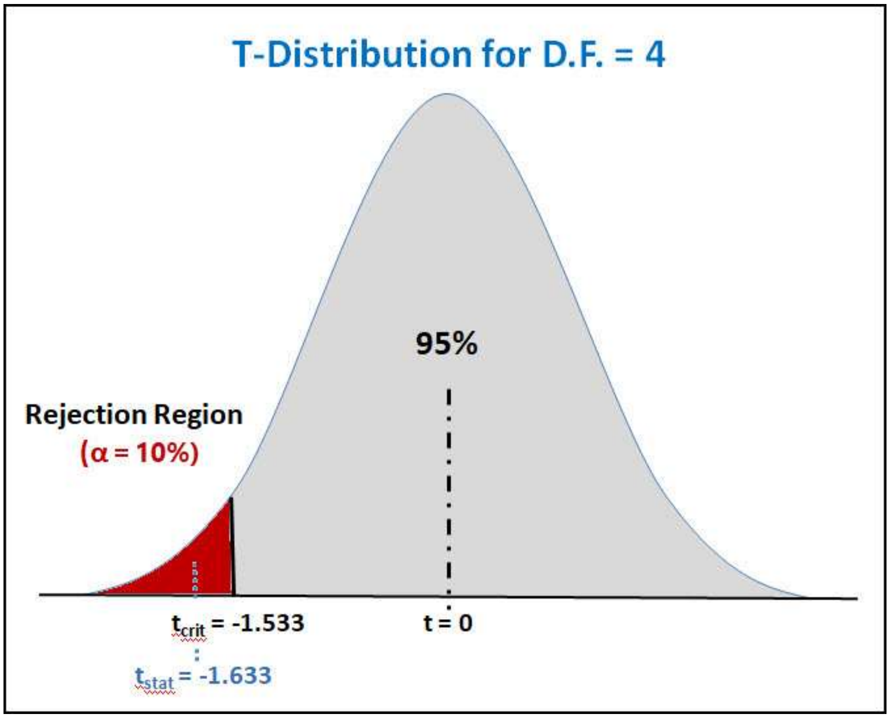
We can then calculate our test statistic:

$$t - statistic = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{9.92 - 10.0}{\frac{0.11}{\sqrt{5}}} = \frac{-0.08}{0.049} = -1.633$$

Step 5 - Compare the Test Statistic against the Rejection Criteria and Make a Conclusion

The next step is to compare our test statistic (-1.633), against our critical t-value that defines our rejection criteria (-1.533).

Graphically this looks like this:



Because our test statistic is more negative than our critical t-value, we can reject the null hypothesis and accept the alternative hypothesis.

Step - 6 State the Decision in Terms of the Original Problem Statement

This hypothesis test allows us to reject our null hypothesis that the new engines will perform similarly to our historical average.

We can reject this null hypothesis in favor of the alternative hypothesis that our new motors will demonstrate better (lower) performance in terms of emissions.