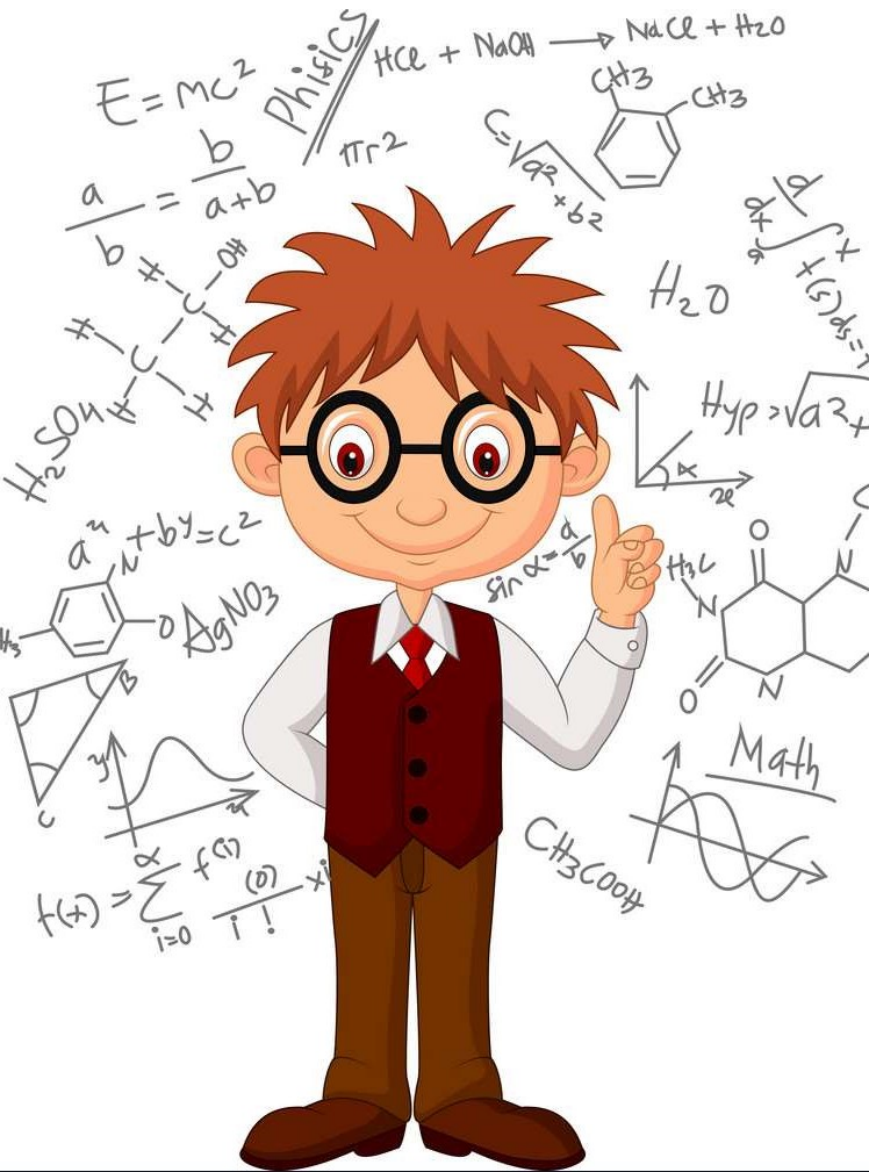
Statistics for Data Analytics

Lecturer: Marina Iantorno

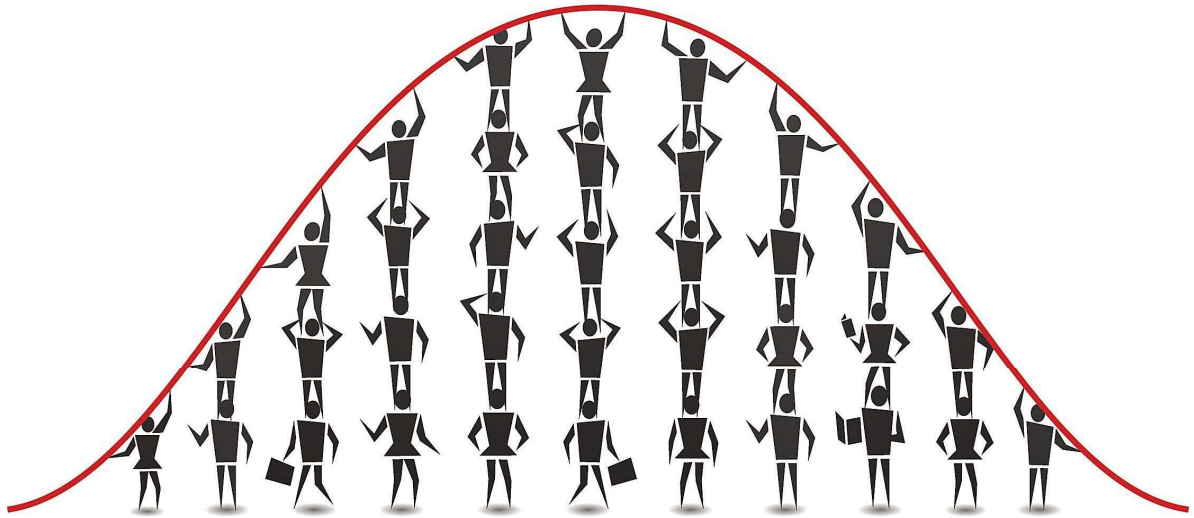
E-mail: miantorno@cct.ie

1st Term 2021/2022

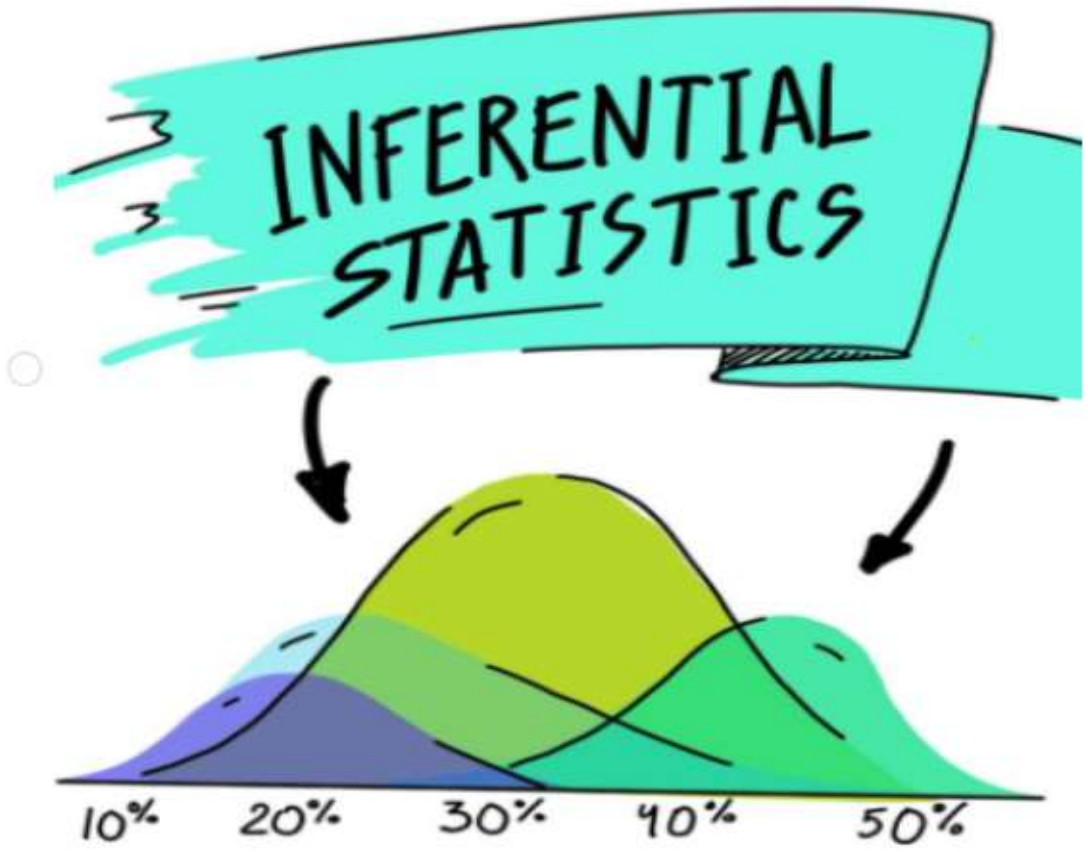


In today's class we will cover:

- T-of Student Distribution
- Hypothesis Test



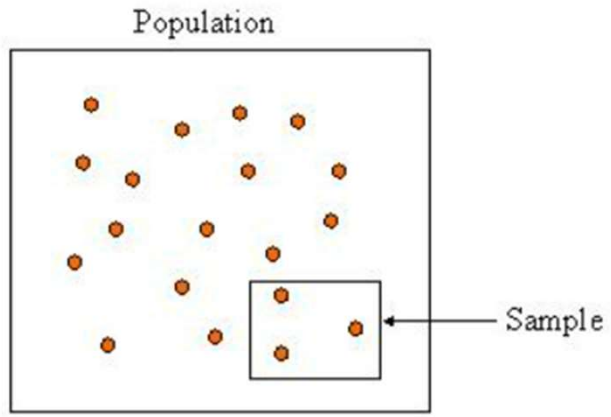
Normal Distribution


INFERENTIAL
STATISTICS
10%
20%
30%
40%
50%

Inferential Statistics

In statistics we usually work with samples from a population. As you may imagine, every sample is different. Suppose that we want to analyse something about Ireland and we picked two samples, it would not be the same to pick 1000 people and then another 1000 people. The values of the variable in place will likely change, and as analyst it is our duty to perform the best estimation for the population basing our analysis on the samples.

Population	Sample
μ	$\bar{x}$



Inferential Statistics

When we want to test or infer the value of the population mean, a new distribution appears to help us in the analysis: Student's T. This distribution has a behaviour very similar to the Normal, but it has one more component that is called "degrees of freedom". In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. This is also a symmetric distribution.

Population	Sample
μ	$\bar{x}$

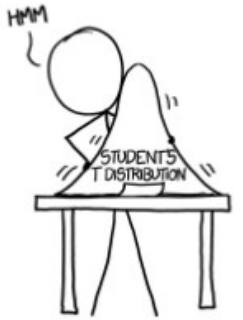
Inferential Statistics

The main difference between the T and the Normal distribution is that T works with sample data while Normal includes the population data.

The formula to calculate T is as follow:

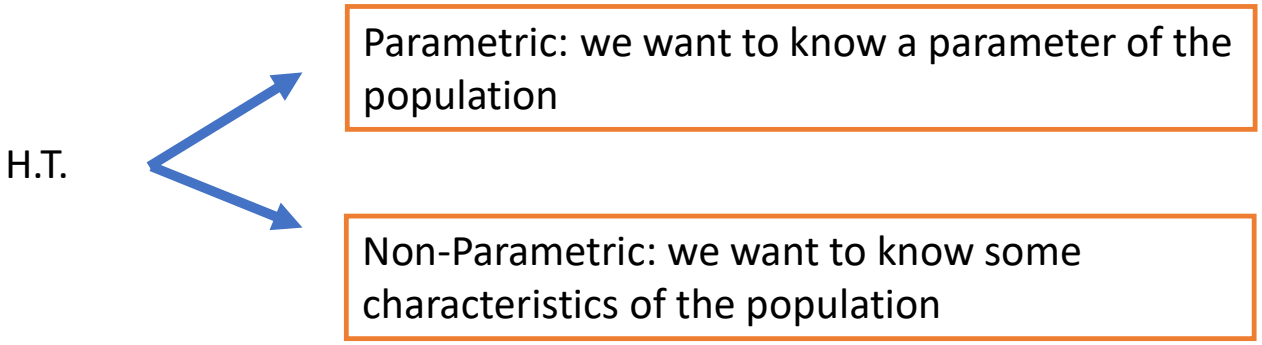
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

Population	Sample
μ	$\bar{x}$



Hypothesis Test

When we do research, we learn about the population based on the data we collected from the sample. The method in which we select samples to learn more about characteristics in a given population is called **hypothesis testing** (HT). Hypothesis testing is really a systematic way to test claims or ideas about a group or population.

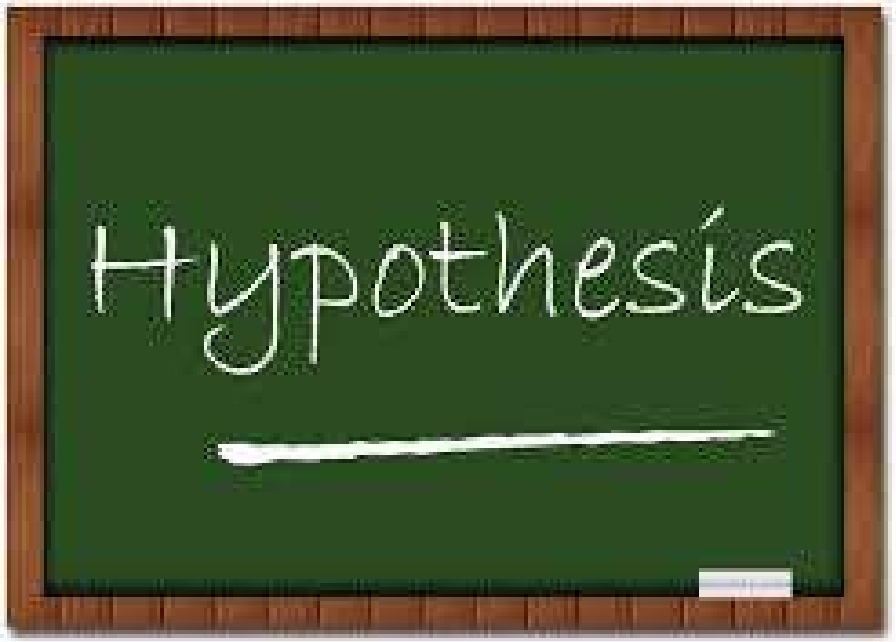


Hypothesis Test

The first step is to have a premise or claim that we want to test.

We will have two possibilities:

- Null Hypothesis: it always states that nothing is going on. We start assuming that our Null Hypothesis is true (this is always the accepted fact). Remember the presumption of innocence: s/he will be considered innocent until proven guilty.
- Alternative Hypothesis: it states that there is a difference in the information we have or were given.



Case 1: HT when σ is known

Hypothesis Test

Jack tells Mary that his average drive of a golf ball is 275 yards. Jean is skeptical and asks for substantiation. To that end, Jack hits 25 drives. The mean resulted only 264.4 of Jack's 25 drives. Jack still maintains that, on average, he drives a golf ball 275 yards and that his (relatively) poor performance can reasonably be attributed to chance. At the 5% significance level and consider that the standard deviation is such driving distances is 20 yards, do the data provide sufficient evidence to conclude that Jack's mean driving distance is less than 275 yards?

X = driving distance (in yards)

Population	Sample	HT

Hypothesis Test

Jack tells Mary that his average drive of a golf ball is 275 yards. Jean is skeptical and asks for substantiation. To that end, Jack hits 25 drives. The mean resulted only 264.4 of Jack's 25 drives. Jack still maintains that, on average, he drives a golf ball 275 yards and that his (relatively) poor performance can reasonably be attributed to chance. At the 5% significance level and consider that the standard deviation is such driving distances is 20 yards, do the data provide sufficient evidence to conclude that Jack's mean driving distance is less than 275 yards?

X = driving distance (in yards)

Population	Sample	HT
$\sigma = 20$	$n = 25$	$\alpha = 0.05$
	$\bar{x} = 264.4$	

Hypothesis Test

X = driving distance (in yards)

Step 1: Hypothesis

H_0: $\mu = 275$

H_1: $\mu < 275$

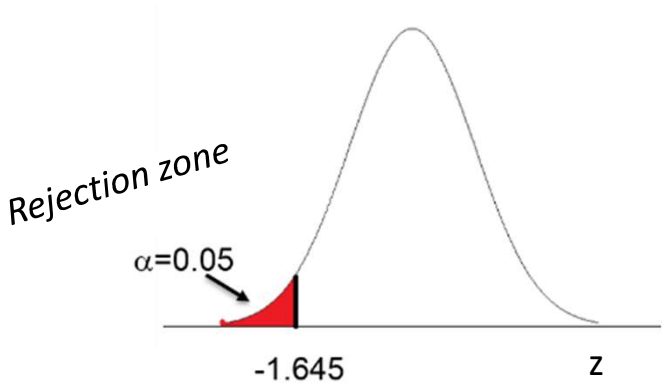
Step 2: Formula

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Normal

Step 3: Critical values

- Table: Normal
- Sign of H1: <
- $\alpha = 0.05$



Hypothesis Test

X = driving distance (in yards)

Step 4: Decision Rule

We reject H_0 if $z \leq -1.645$

We accept H_0 if $z > -1.645$

Step 5: z Calculation

$$z = \frac{264.4 - 275}{\frac{20}{\sqrt{25}}} = -2.65$$

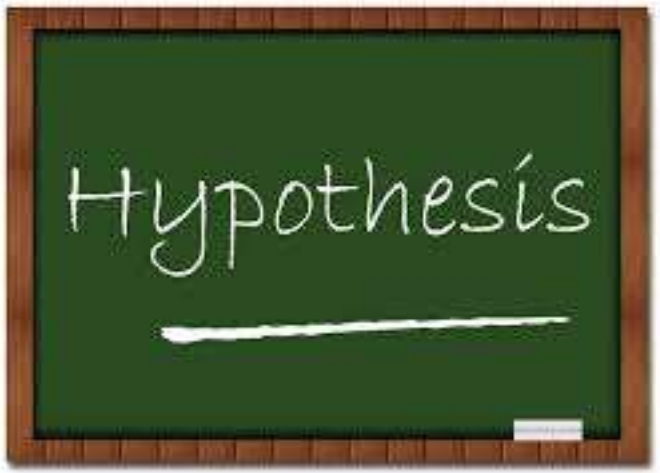
Step 6: Result and conclusion

$z = -2.65$

$z < -1.645$ → We reject H_0

Step 7: Interpretation

There is enough evidence to say that Jack's mean driving distance is less than his claimed 275 yards.



Case 2: HT when σ is unknown

Hypothesis Test

Acid rain from the burning of fossil fuels has caused many of the lakes around the world to become acidic. The biology in these lakes often collapses because of the rapid and unfavorable changes in water chemistry. A lake is classified as nonacidic if it has a pH greater than 6. A study was in place to measure the pH of 15 lakes and the scientists got an average of 6.6 and a standard deviation of 0.68. At a 5% significance level, could you conclude that, on average, these lakes are nonacidic?

X = pH of the lakes

Population	Sample	HT

Hypothesis Test

Acid rain from the burning of fossil fuels has caused many of the lakes around the world to become acidic. The biology in these lakes often collapses because of the rapid and unfavorable changes in water chemistry. A lake is classified as nonacidic if it has a pH greater than 6. A study was in place to measure the pH of 15 lakes and the scientists got an average of 6.6 and a standard deviation of 0.68. At a 5% significance level, could you conclude that, on average, these lakes are nonacidic?

X = pH of the lakes

Population	Sample	HT
	$n = 15$	$\alpha = 0.05$
	$S = 0.68$	
	$\bar{x} = 6.6$	

Hypothesis Test

X = pH of the lakes

Step 1: Hypothesis

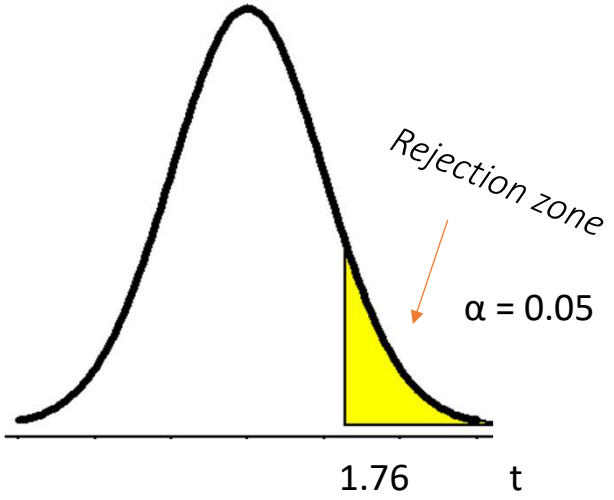
H_0: $\mu = 6$

H_1: $\mu > 6$

Step 2: Formula

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n-1}}}$$

T-student (V = n-1)



Step 3: Critical values

- Table: T-Student
- Sign of H1: >
- α = 0.05 and V = 14

Hypothesis Test

X = pH of the lakes

Step 4: Decision Rule

We reject H_0 if $t > 1.76$

We accept H_0 if $t < 1.76$

Step 5: t Calculation

$$t = \frac{6.6 - 6}{\frac{0.68}{\sqrt{14}}} = 3.30$$

Step 6: Result and conclusions

$t = 3.3$

$t > 1.76 \rightarrow$ We reject H_0

Step 7: Interpretation

We have enough evidence to say that on average, these lakes are nonacidic.

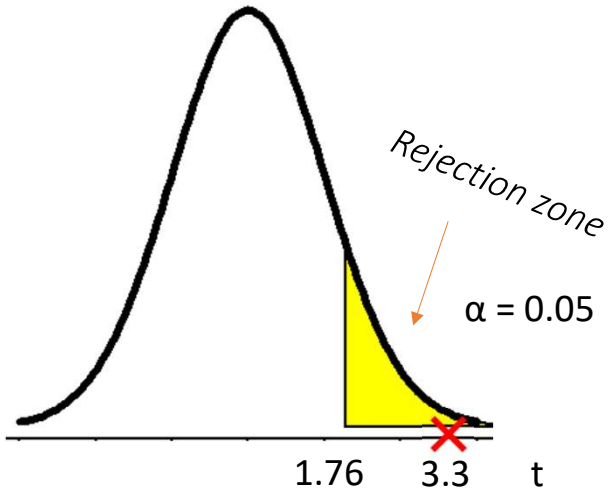
Hypothesis Test

If we follow the steps, we can reach to the results with no issues at all, but what would happen if we wanted to take a short cut?

Here arises a new concept that is called: “p-value”.

Imagine that we come back to step 3, and we identify the rejection zone and let’s remember the real t that we got in step 5 (3.3). Our real t with these values is 3.3, which means that we could “reduce” our alpha level.

When we calculate the p-value we are “moving” the rejection zone.



p-value = $P_{H_0}(t > 3.3) = 0.0026$

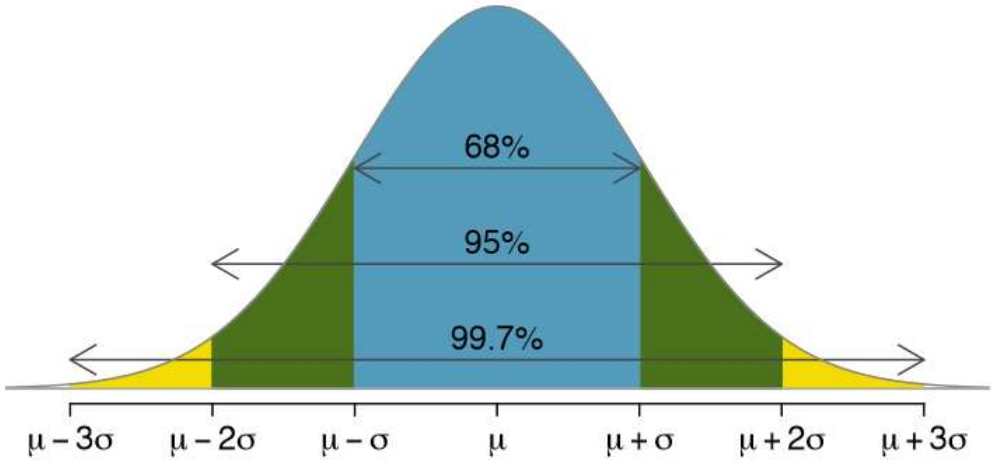
$\alpha = 0.01$ → we could reduce our significance level to 1%

P-value is a new possible α value. In other words, if your p-value is less than your α then, you can Reject H_0, but if your p-value is greater than or equal to α, you can’t reject H_0

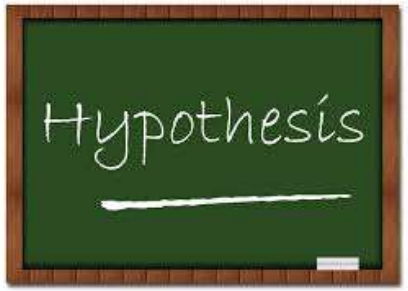
Hypothesis Test

When we calculate p-value we need to check what is stated on H1

H1 : $\mu > \mu_0$	p-value = $P_{H_0}(z > z\text{-obs})$ or $P_{H_0}(t > t\text{-obs})$
H1: $\mu < \mu_0$	p-value = $P_{H_0}(z < z\text{-obs})$ or $P_{H_0}(t < t\text{-obs})$
H1: $\mu \neq \mu_0$	p-value = $P_{H_0}(z \neq \lvert z\text{-obs} \rvert)$ or $P_{H_0}(t \neq \lvert t\text{-obs} \rvert)$



HT for two populations



HT for 2 populations when σ is unknown

HT for two populations

Case 2

To decide between investing in one asset or another, an analyst calculated the average return and risk of each of them. In the last 9 days, asset A had an average variation of 0.9% with a deviation of 0.193%, while asset B, which was not listed in one of the days, provided an average variation of 1.18% with a standard deviation of 0.245%. Verify if there is a significant difference between the assets with a confidence of 90%.

Population 1	Population 2	Sample 1	Sample 2	HT

HT for two populations

Case 2

To decide between investing in one asset or another, an analyst calculated the average return and risk of each of them. In the last 9 days, asset A had an average variation of 0.9% with a deviation of 0.193%, while asset B, which was not listed in one of the days, provided an average variation of 1.18% with a standard deviation of 0.245%. Verify if there is a significant difference between the assets with a confidence of 90%.

Population 1	Population 2	Sample 1	Sample 2	HT
		$n = 9$	$n = 8$	$\alpha = 0.10$
		$\bar{x} = 0.9$	$\bar{x} = 1.18$	
		$S = 0.193$	$S = 0.245$	

HT for two populations

Step 1: Hypothesis Statement

$H_0 : \mu 1 = \mu 2$

$H_1 : \mu 1 \neq \mu 2$

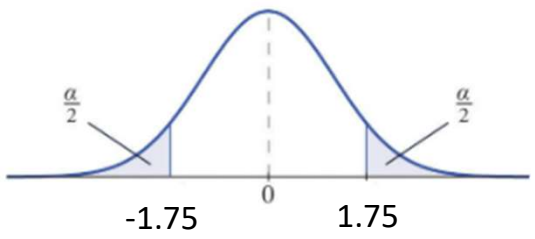
Step 2: Formula

$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\bar{s}^2}}$	T de Student	G.L : $n_1 + n_2 - 2$ $\bar{s}^2 = \frac{n_1 s_1^{\,2} + n_2 s_2^{\,2}}{n_1 + n_2}$

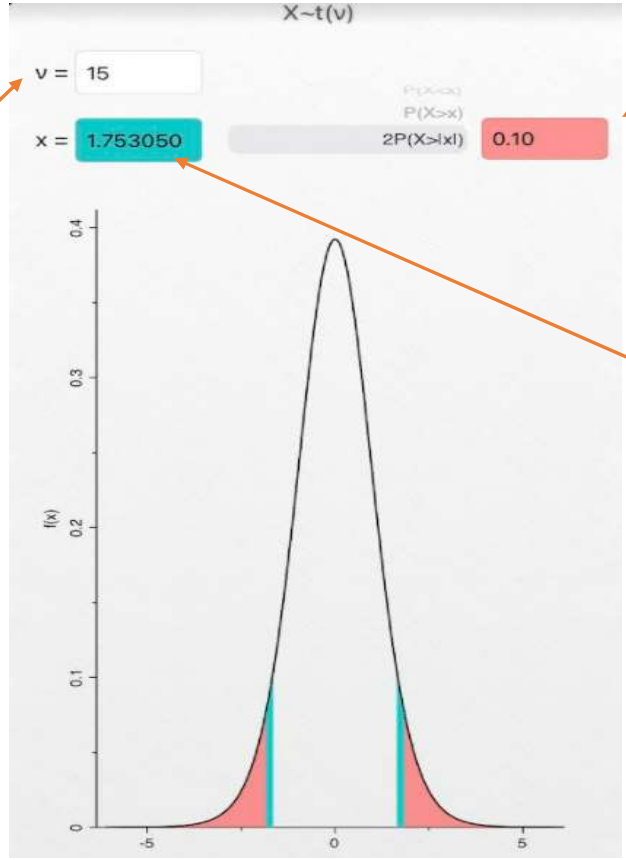
HT for two populations

Step 3: Critical Values

- Table: T-Student
- V = 8 + 9 – 2 = 15
- α = 0.10



Degrees of freedom here



Significance Level here

Your result will appear here

HT for two populations

Step 4: Decision Rule

We reject H_0 if $t < -1.75$ or $t > 1.75$

We accept H_0 if $-1.75 \leq t \leq 1.75$

Step 5: t Calculation

Auxiliar calculation

$$\overline{s}^2 = \frac{n_1 s_1^{\ 2} + n_2 s_2^{\ 2}}{n_1 + n_2}$$

$$\overline{S}^2 = \frac{9 * 0.193^2 + 8*0.245^2}{8 + 9} = 0.0479$$

$$t = \frac{(0.9 - 1.181) - (mu1 - mu2)}{\sqrt{\left[\frac{1}{8} + \frac{1}{9}\right] * 0.0479}} = -2.642$$

HT for two populations

Step 6: Result and conclusions

$T < -1.75$ → We reject H_0

Step 7: Interpretation

There is enough evidence at 10% significance level to say that there are significant differences between the Asset A and B.

THAT'S ALL FOR TODAY

THANK YOU