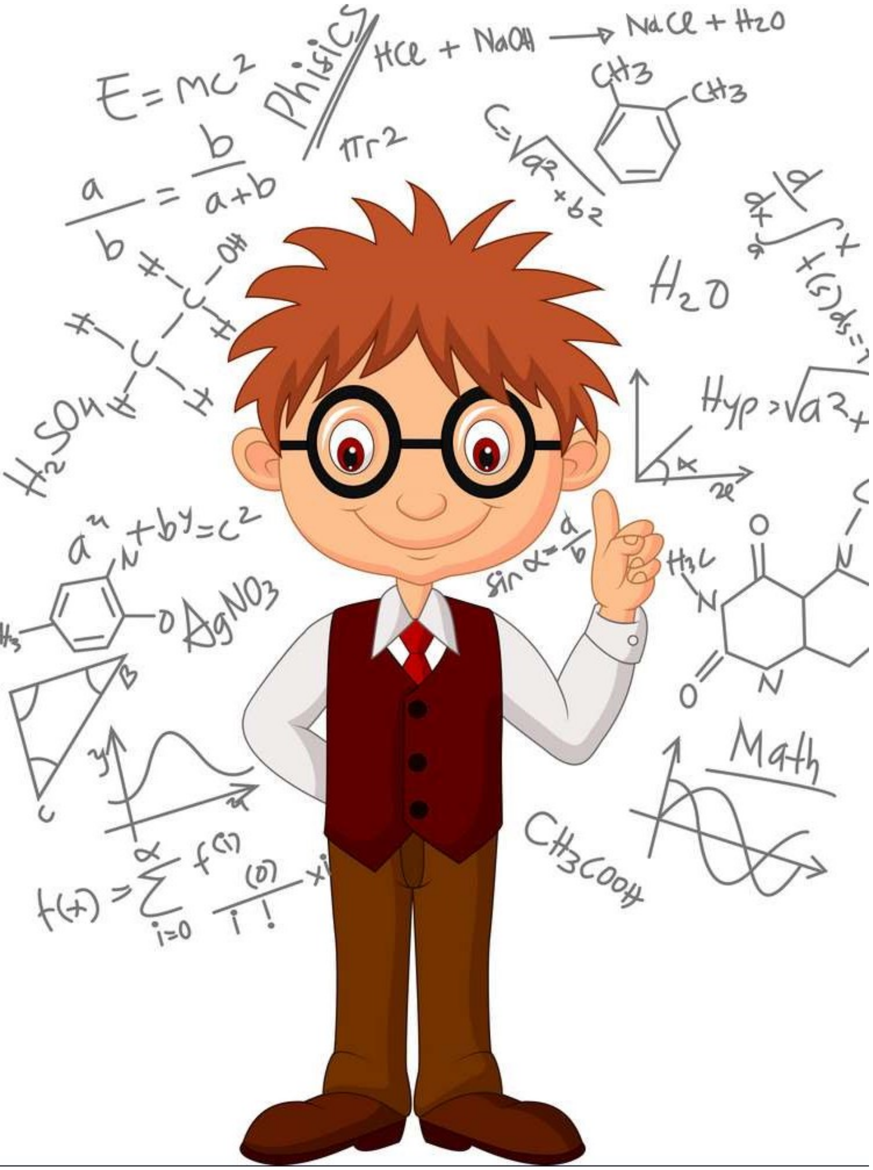
Statistics for Data Analytics

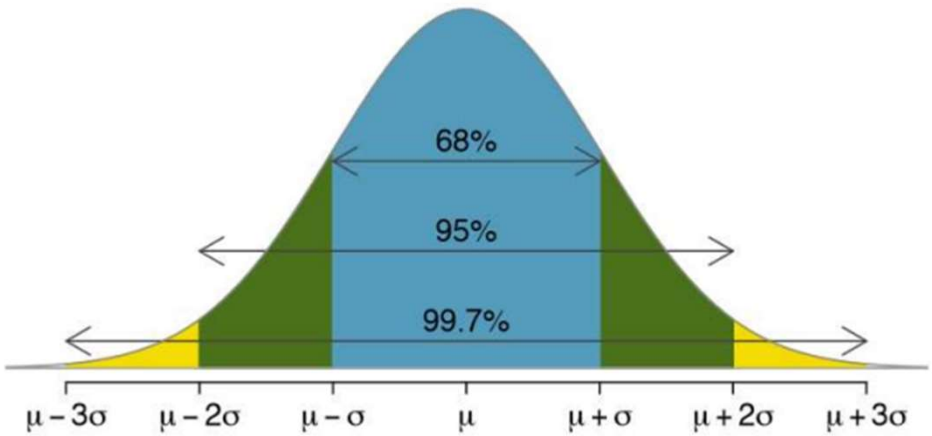
Lecturer: Marina Iantorno
E-mail: miantorno@cct.ie

CCT College Dublin
Computing • IT • Business



In today's class we will cover:

- ❑Confidence Intervals
- ❑Practice



Confidence Intervals

Confidence Intervals

So far we covered the idea of identifying whether the population mean is equals, less or greater than certain values, but what happens if we don’t have value to compare the mean with? Here is when we start working with Confidence Intervals. This technique will help us to identify a range where the population mean is likely to be located.

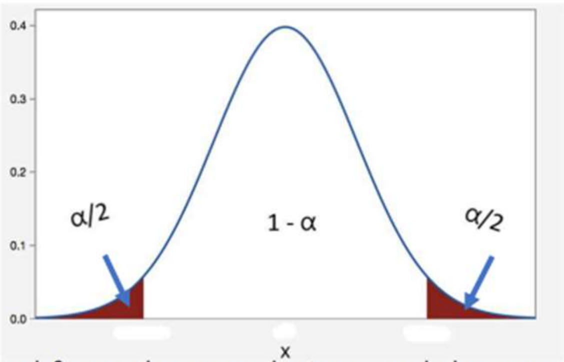
Why do we try to find an interval and not an exact values?

Well, remember that we are trying to find information about the population based on the sample. The confidence level represents the chance that if you were to repeat your sample-taking over and over, you would get a range of likely values that actually contains the real population parameter.

Confidence Intervals

For example, suppose we want to know the average time that students at the College practice Statistics at home per week. We take a random sample of 1000 students and find that the average time a student spends practicing at home is 2.5 hours. You hope and assume that the average for the whole population is close to 2.5 hours but it is probably not exact. The solution then, is to set a margin of error from the mean and appears again the significance level of the analysis.

Confidence Intervals

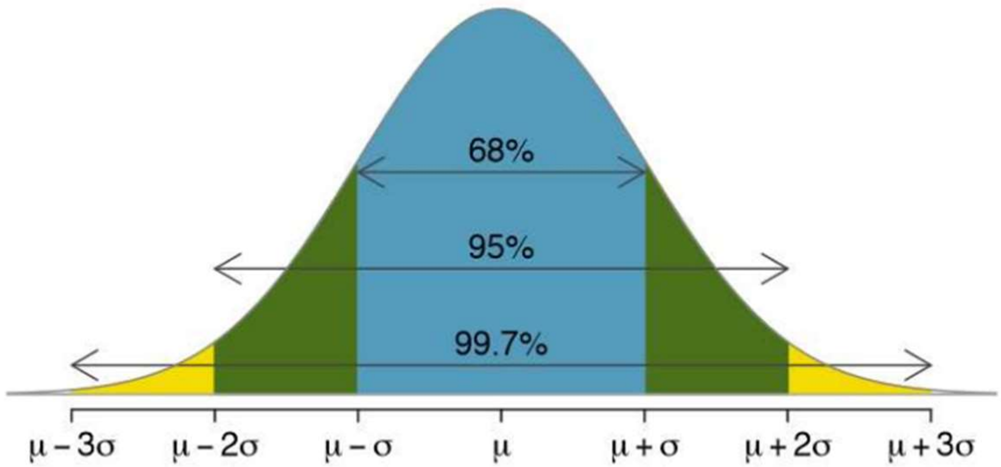


A random sample is collected from the population, and the resulting

Sample statistics are used to determine the **lower limit (LL)** and **upper limit (UL)** of an interval.

We need to have some confidence in our hypothesis, and that confidence is 1- α. As we have an interval, we need to divide α on 2 equal parts. α is the significance level of this study.

We usually assign 90% to 99% of confidence to our estimation.



Case 1:
Confidence intervals when σ is known

Confidence Intervals

CASE 1: CI to find μ (SD of the population is known)

An agency is looking to outsource some of its activities and will base the decision in the average salary of the employees. One company was chosen to be analysed and a sample of 500 employees drew an average salary of € 2000 monthly. According to previous studies, the salaries follow a normal distribution with a standard deviation of €300 monthly. What is the average salary of all the employees of this company at a 95% confidence?

X = salary of the employees (€)

Population	Sample	CI

We need to define the variable of study and display the data we have.

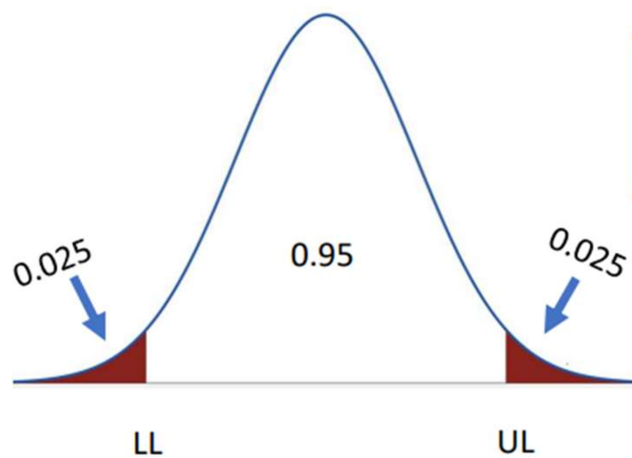
Confidence Intervals

CASE 1: CI to find μ (SD of the population is known)

An agency is looking to outsource some of its activities and will base the decision in the average salary of the employees. One company was chosen to be analysed and a sample of 500 employees drew an average salary of € 2000 monthly. According to previous studies, the salaries follow a normal distribution with a standard deviation of €300 monthly. What is the average salary of all the employees of this company at a 95% confidence?

X = salary of the employees (€)

Population	Sample	CI
$\sigma = 300$	$n = 500$	$1 - \alpha = 0.95$
	$\bar{x} = 2000$	$\alpha = 0.05$
		$\alpha/2 = 0.025$



Once we identify our goal, we need to find a formula that bring us to the solution

Confidence Intervals

CASE 1: CI to find μ (SD of the population is known)

An agency is looking to outsource some of its activities and will base the decision in the average salary of the employees. One company was chosen to be analysed and a sample of **500** employees drew an average salary of € **2000** monthly. According to previous studies, the salaries follow a normal distribution with a standard deviation of €**300** monthly. What is the average salary of all the employees of this company at a **95**% confidence?

X = salary of the employees (€)

The formula we will use in this case belongs to the Normal distribution. See the formula below:

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

Or clearing it

$$\boxed{L = \bar{x} \pm Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}}$$

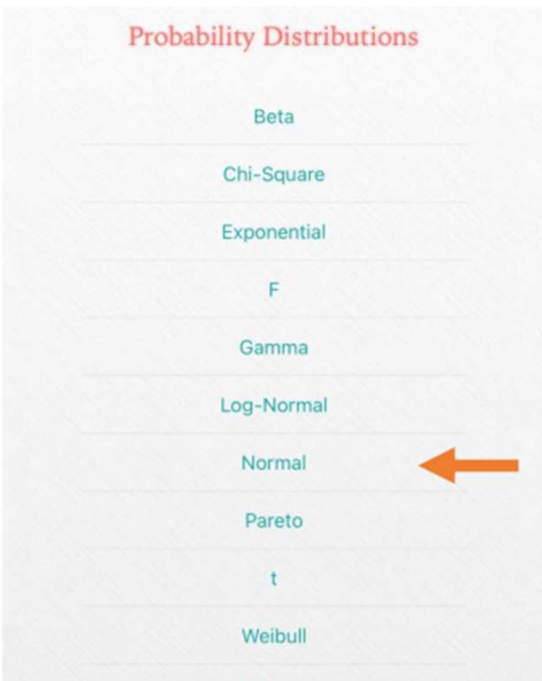
Confidence Intervals

X = salary of the employees (€)

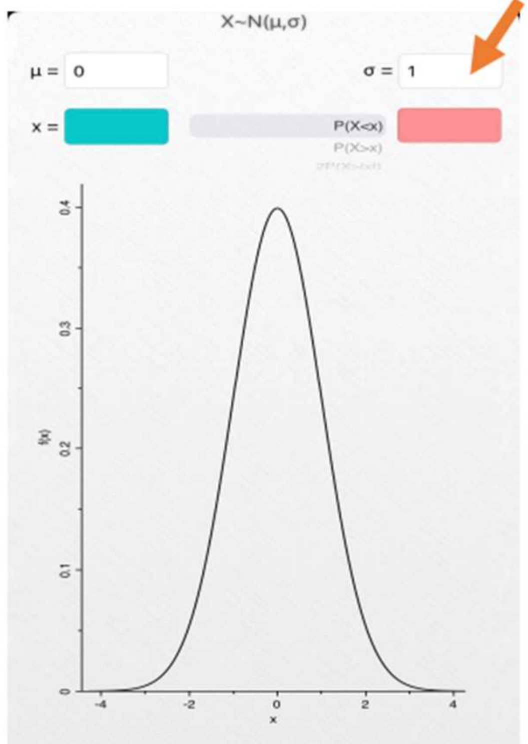
Once we have the formula, we know that we will work with Normal Distribution, therefore, we will need this table to find the LL and the UL in Z. We have used it before and for this exercises we will use the "**Probabilities Distribution**" app.

Confidence Intervals

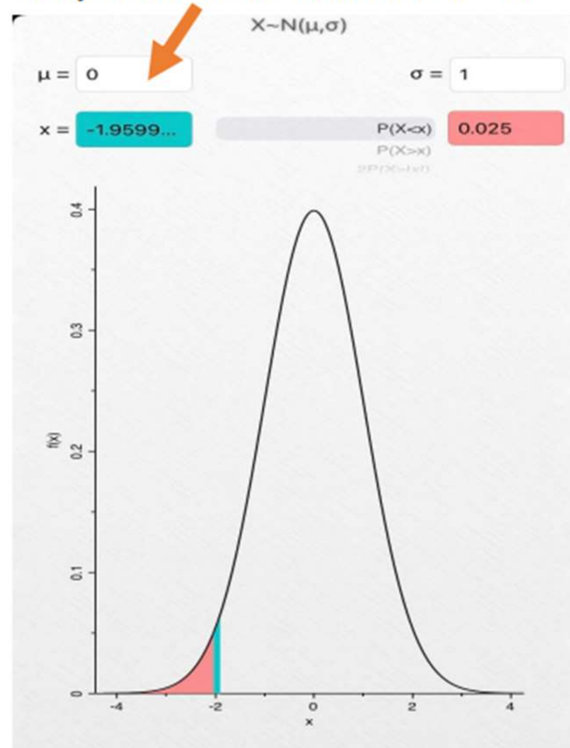
Step 1: Choose the distribution (Normal)



Step 2: Write in P(X>x)= α/2



Step 3: See the Z number in x = ...



Confidence Intervals

Let's get back to the exercise!

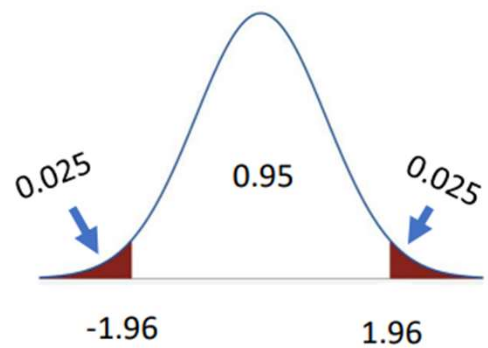
X = salary of the employees (€)

Population	Sample	CI
$\sigma = 300$	$n = 500$	$1 - \alpha = 0.95$
	$\bar{x} = 2000$	$\alpha = 0.05$
		$\alpha/2 = 0.025$

$$L = \bar{x} \pm Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

$$LL = 2000 - 1.96 * 300/\sqrt{500} = \mathbf{1973.70}$$

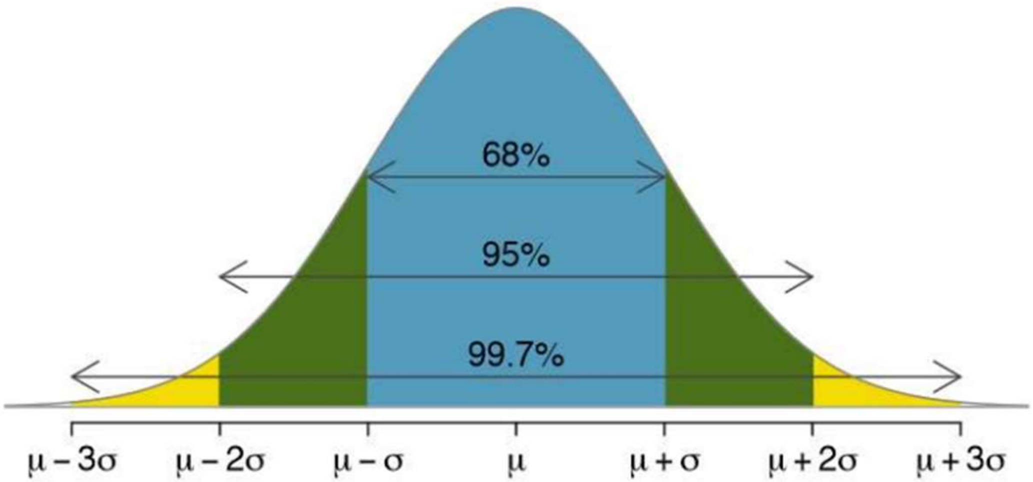
$$UL = 2000 + 1.96 * 300/\sqrt{500} = \mathbf{2026.29}$$



Our z values will be those that we found on the app. We just replace our data in the formula

$$1973.70 \leq \mu \leq 2026.29$$

Answer: The average salary of the employees that work at this place is between € 1973.70 and €2026.29.



Case 2

Confidence intervals when σ is unknown

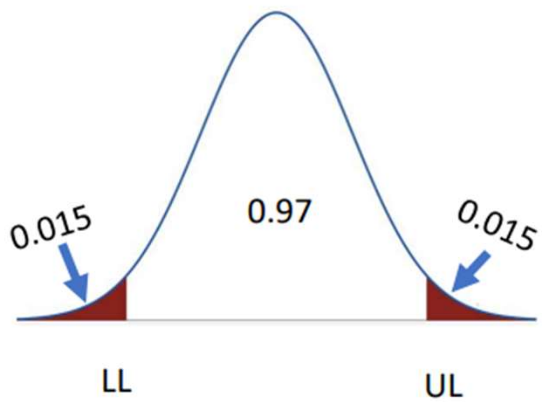
Confidence Intervals

CASE 2: CI to find μ (SD of the population is unknown)

An inspector goes to a biscuit factory to do an auditory to find out the average weight of the biscuit packages. In order to do that, he randomly takes 300 packages that draw an average weight of 250 grams with a standard deviation of 25 grams. At a 97% of confidence, what is the average weight of the biscuit packages of the factory?

X = weight of the packages (in grams)

Population	Sample	CI



We need to define the variable of study and display the data we have as we did in the previous example

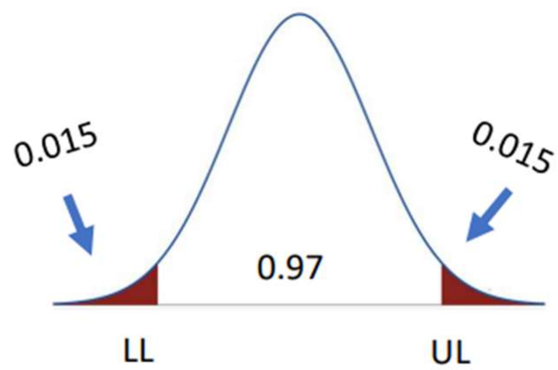
Confidence Intervals

CASE 2: CI to find µ (SD of the population is unknown)

An inspector goes to a biscuit factory to do an auditory to find out the average weight of the biscuits packages. In order to do that, he randomly takes 300 packages that draw an average weight of 250 grams with a standard deviation of 25 grams. At a 97% of confidence, what is the **average** weight of the biscuit packages of the factory?

X = weight of the packages (in grams)

Population	Sample	CI
	$n = 300$	$1 - \alpha = 0.97$
	$\bar{x} = 250$	$\alpha = 0.03$
	$S = 25$	$\alpha/2 = 0.015$



We don't have any data from the population, then we need to use a new formula

We will need a formula that includes only sample data and CI data

Confidence Intervals

CASE 2: CI to find μ (SD of the population is unknown)

An inspector goes to a biscuit factory to do an auditory to find out the average weight of the biscuits packages. In order to do that, he randomly takes 300 packages that draw an average weight of 250 grams with a standard deviation of 25 grams. At a 97% of confidence, what is the **average** weight of the biscuit packages of the factory?

X = weight of the packages (in grams)

When we need to find $\boldsymbol{\mu}$ and we have only sample data, we use the T- of student distribution. Find the formula below:

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

Or clearing it

$$\boxed{L = \bar{x} \pm t_{\alpha/2} * S/\sqrt{n-1}}$$

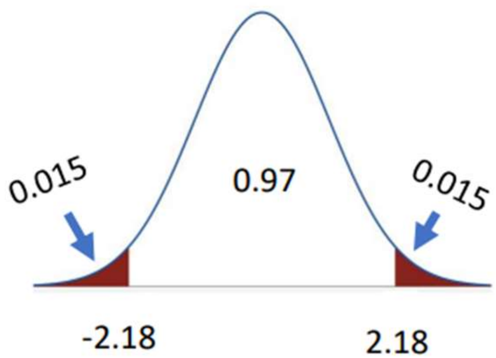
v: n-1 (degrees of freedom)

Confidence Intervals

Let's get back to the exercise!

X = weight of the packages (in grams)

Population	Sample	CI
	n = 300	$1 - \alpha = 0.97$
	$\bar{x} = 250$	$\alpha = 0.03$
	S = 25	$\alpha/2 = 0.015$



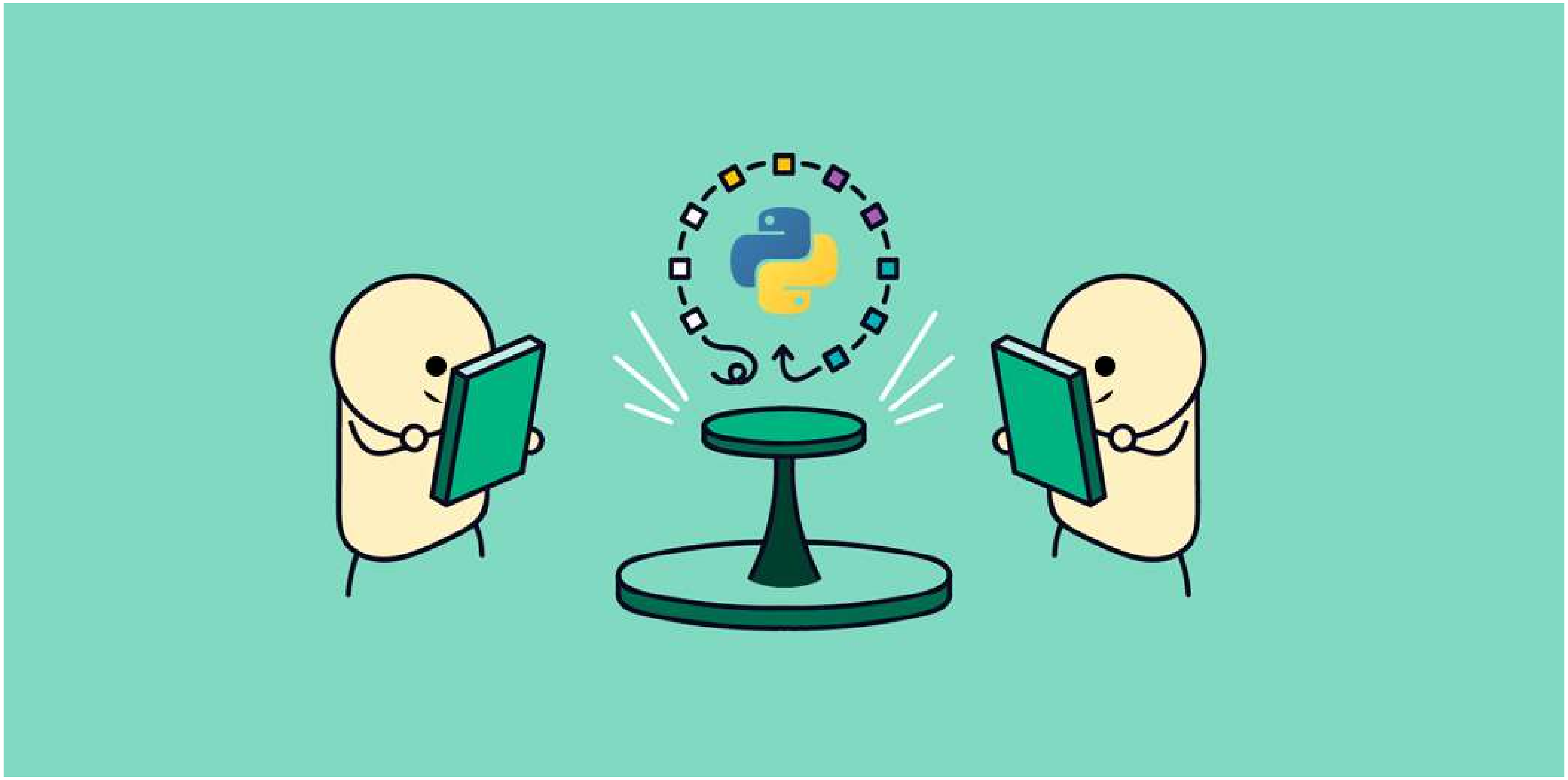
We check on the app the "t" distribution to see the t values. Remember to fill the degree of freedom: v = n-1 → v= 299

$$246.85 \leq \mu \leq 253.15$$

$$LL = 250 - 2.18 * 25/\sqrt{300-1} = \mathbf{246.85}$$

$$UL = 250 + 2.18 * 25/\sqrt{300-1} = \mathbf{253.15}$$

Answer: The average weight of the packages at this factory is between 246.85 grams and 253.15 grams



Let’s do some practice with Python

THAT’S ALL FOR TODAY

THANK YOU