Université
de Technologie
Tarbes
Occitanie Pyrénées

THE **COMPLETE GUIDE** TO

INFERENTIAL STATISTICS

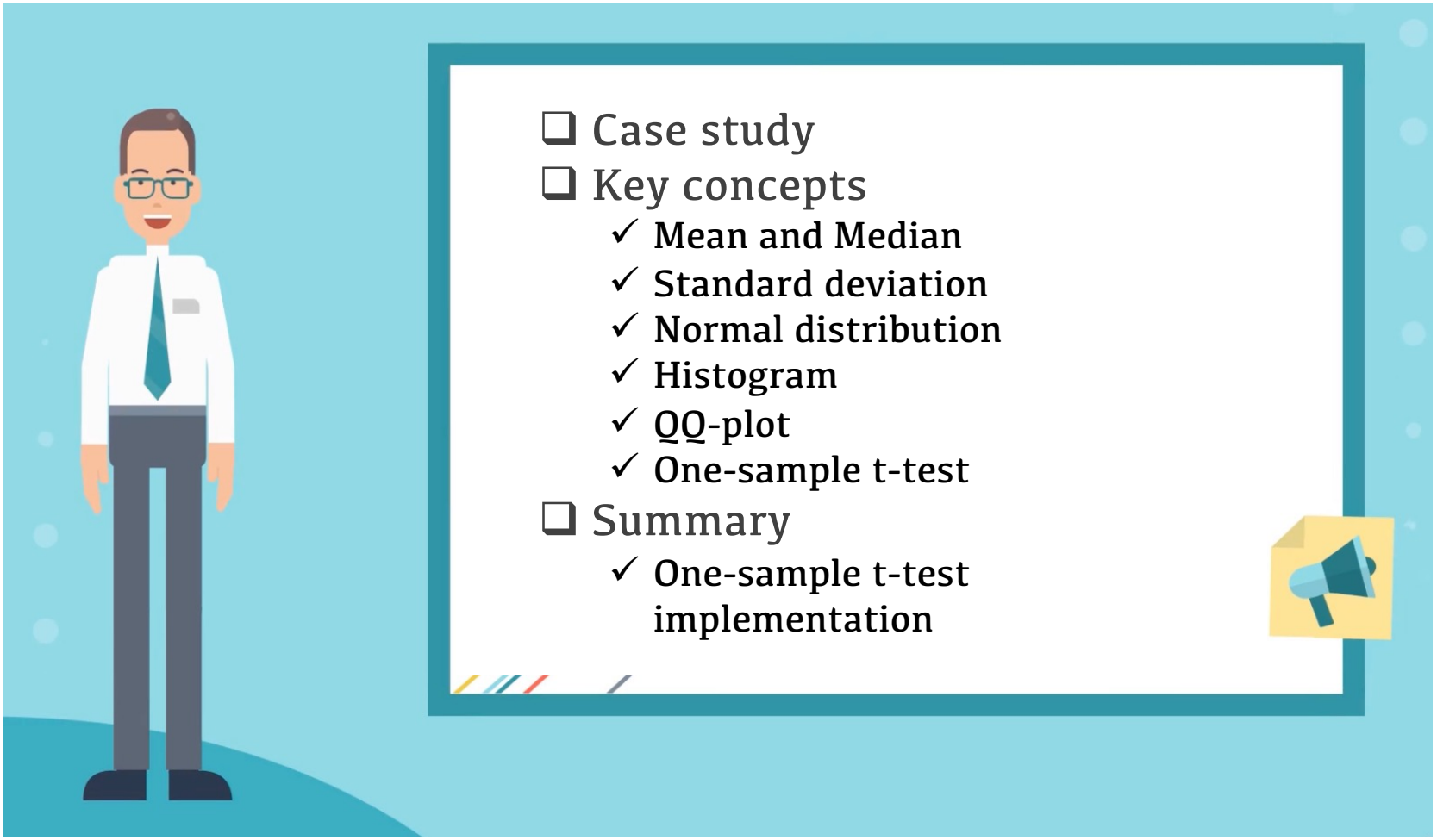
< By Raymond HOUE NGOUNA />

Université de Technologie Tarbes Occitanie Pyrénées

</SECTION 3>

ONE-SAMPLE T-TEST

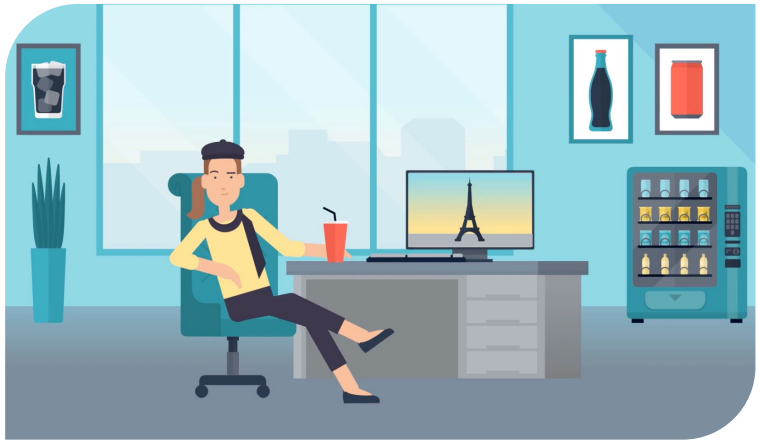
SECTION OUTLINE



CASE STUDY

CASE DESCRIPTION

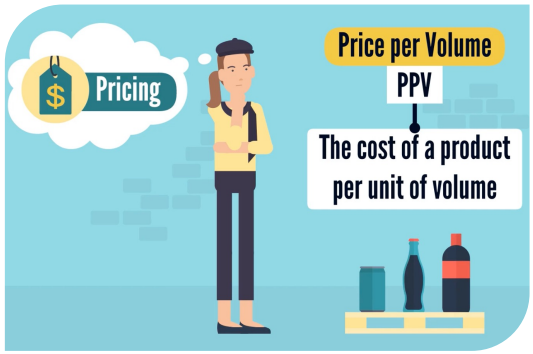
Sophie, the Marketing Director of a multinational soft drink company, supervises the European market.



- ✓ Her primary goal is to align brand pricing with top management's recommended positioning.
- ✓ This is vital because other marketing elements like targeting, packaging, and distribution aim to position these brands as premium in a competitive market.

SOPHIE'S APPROACH

- ✓ Sophie assesses her brand's pricing alignment with its high-end image by examining the price per volume, a standardized method to compare prices across various package sizes.



- ➔ She randomly selects 100 stock-keeping units (SKUs) among all sold in Europe,
- ➔ And checks the price in the database of a market research company based on the data.

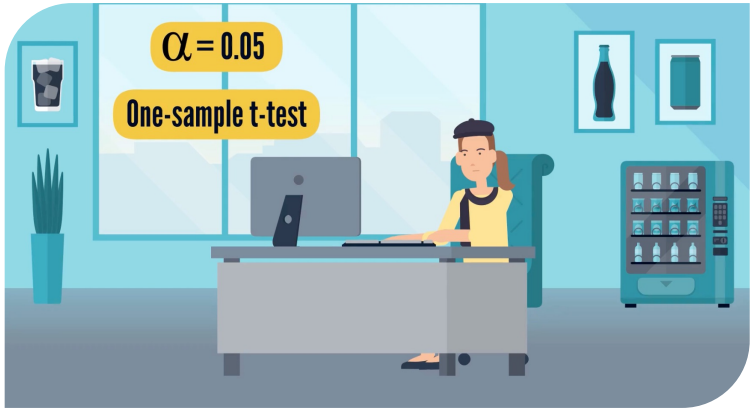
SOPHIE'S RESEARCH

- ✓ Sophie found that the average PPV for the 100 SKUs in the sample is € 6.3 per liter.
  - She aims to verify if this average price per volume significantly exceeds the broader European markets.
- ✓ The known average is € 6.21 per liter.
  - Sophie knows that such a step involves statistical testing.

- The test aims to determine if the observed difference in average pricing between her sample and the overall market is not just a result of random variation, but indicates a real substantial difference.
- A significant difference implies that her brand's pricing strategy differs from the market average, aligning with the company's high-end positioning goals.

THE RESEARCH'S OUTCOMES

- ✓ Fully aware of her plan, Sophie sets the significance level at 0.05 and conducts a one-sample t-test.
  - Sophie's test results indicate that the mean price per volume of €6.43 is not significantly different from €6.21 at the 0.05 significance level.



- *This outcome is disappointing for Sophie, as she had hoped her product's price per volume would surpass the market average.*
- *A positive aspect of this analysis is that it highlighted a potential issue with the company's pricing policy.*
- *Sophie can now address this issue with the company's top management and suggest raising the prices of some products and brands in the portfolio.*

MEAN AND MEDIAN

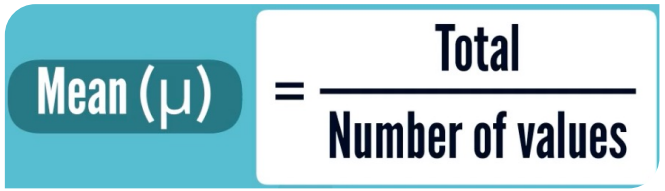
Université de Technologie Tarbes Occitanie Pyrénées

MEAN & MEDIAN

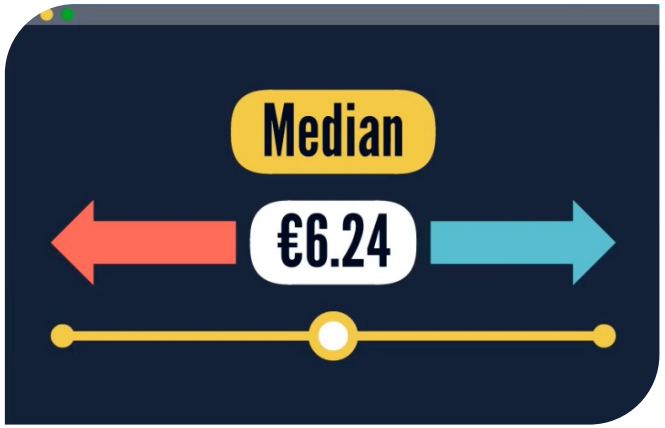
- ✓ Mean and median are both measures of central tendency.
- ✓ They represent the typical or central value of a data set, offering unique summaries and insights into its central characteristics.

EXAMPLES

- ✓ The mean represents the average of a dataset.
  - Sophie's average transaction value across 100 SKUs, with a total price per volume (PPV) of €642.71, is €6.4271.

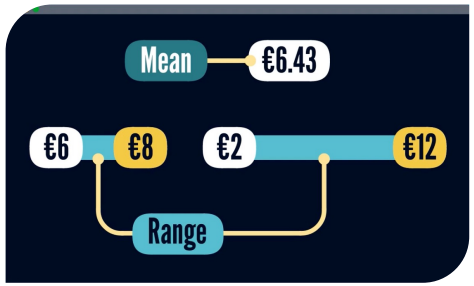


- ✓ The median, on the other hand, is the middle value of a dataset when the values are arranged in ascending order.
  - The median PPV of the 100 SKUs is €6.24, meaning half of the SKUs have a PPV below €6.24, while the other half have a PPV above it.



STANDARD DEVIATION

DATA RANGE



- ✓ Knowing Sophie's SKU portfolio's mean price per volume at €6.43 is helpful.
- ✓ Understanding the full PPV range is crucial to interpreting this value.

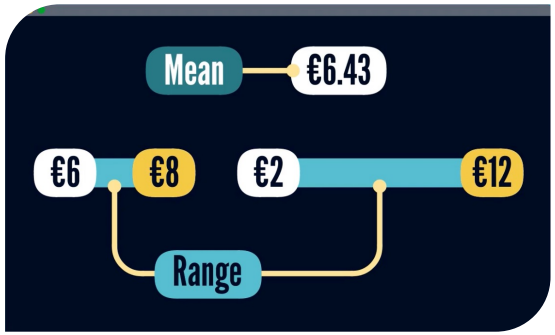
- ✓ A narrow price range of 6 to €8 per liter differs significantly from a broader range of 2 to €12,
- ✓ Impacting the brand positioning and pricing strategy.
- ✓ The PPV in Sophie's case ranges from 3.02 to €9.50 ➔ €6.48.

RANGE INTERPRETATION

- ✓ It's important to recognize that, while range is a useful metric, it can be misleading for general data interpretation as a single outlier can greatly affect the results.
  - If Sophie raises the PPV to €13.02, this would significantly increase the price range from €3.02 to €13.02,
  - This will result in a jump of €10, a substantial increase from the previous range of €6.48.

- ✓ Is there a better metric that more accurately reflects the data?

STANDARD DEVIATION

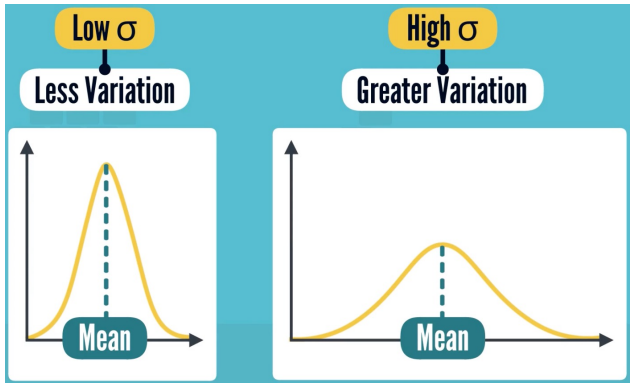


- ✓ Knowing Sophie's SKU portfolio's mean price per volume at €6.43 is helpful.
- ✓ Understanding the full PPV range is crucial to interpreting this value.

- ✓ A narrow price range of 6 to €8 per liter differs significantly from a broader range of 2 to €12,
- ✓ Impacting the brand positioning and pricing strategy.
- ✓ The PPV in Sophie's case ranges from 3.02 to €9.50 ➔ €6.48.

INTERPRETATION OF STANDARD DEVIATION

- ✓ The standard deviation measures the average deviation of data points from the mean.
- ✓ It reflects how spread out the data is.
  - A low standard deviation indicates that the data points are closely clustered around the mean, showing minimal variability.
  - In contrast, a high standard deviation suggests a wider spread of values, indicating greater variability.
- ✓ Standard deviation is a valuable metric because it effectively assesses the consistency or variability of the data.



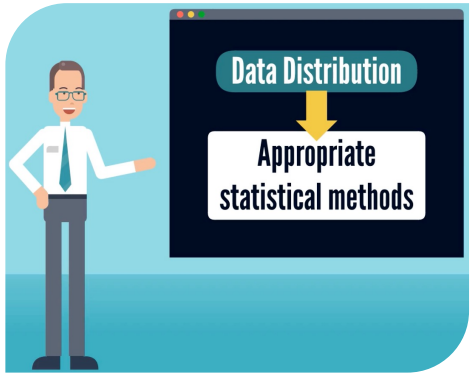
The standard deviation of the PPV in Sophie's product sample is €1.37, with an average of €6.43:

- the prices fluctuate around this mean by €1.37,
- hence a uniform pricing strategy with a moderate variation around the average.

NORMAL DISTRIBUTION

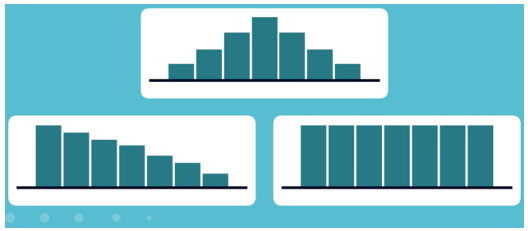
Université de Technologie Tarbes Occitanie Pyrénées

DATA DISTRIBUTION



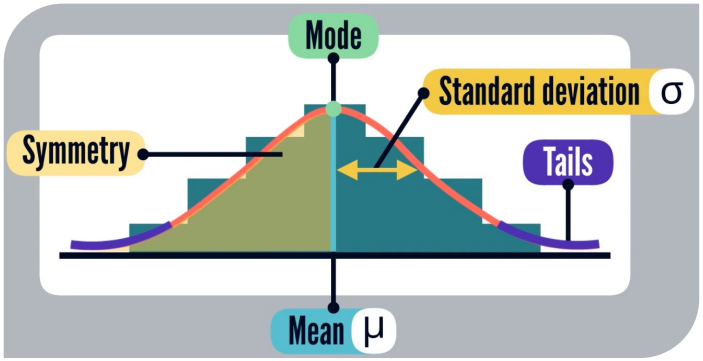
- ✓ In statistical testing, a distribution refers to the pattern of data values and their frequency of occurrence in a data set or population.
  - It provides information about how the data are spread or distributed among different values.

- A distribution represents how often each value occurs in the data set, effectively illustrating the shape of the data spread such as whether they are:
  - clustered around a central value,
  - skewed to one side,
  - or evenly distributed across a range.



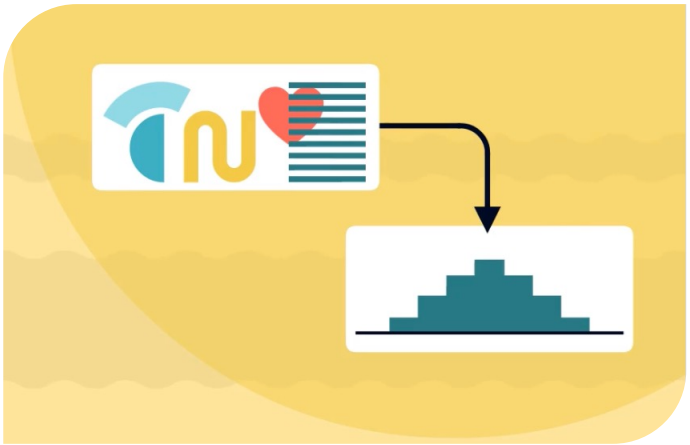
GAUSSIAN DISTRIBUTION

- ✓ Many statistical tests require data to be normally distributed.
- ✓ The Gaussian distribution *(by Carl Friedrich Gauss),* commonly known as the normal distribution, is certainly the most famous.



- It features a symmetric bell-shaped curve.
- Most data points are clustered near the average, with fewer as the distance from the mean increases.
- Symmetry means the data are equally likely to occur on either side of the mean.
- The mode coincides with the mean marking the distribution's peak, and the tails extend indefinitely, suggesting rare but possible extreme values.

WHY USE NORMAL DISTRIBUTION?

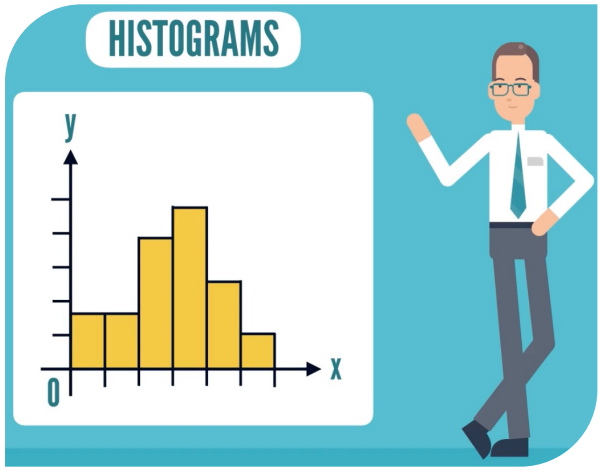


- ✓ Normal distribution is widely used in statistical testing due to its theoretical properties and relevance to many phenomena.
  - Many statistical tests assume that the data follow a normal distribution because it simplifies the mathematical calculations and makes certain statistical inferences more confident.

- ✓ We explore various tests, including examining normality tests to determine precisely if a data set adheres to a normal distribution.

HISTOGRAM

HISTROGRAM



- ✓ A histogram is a graphical representation of the distribution of a data set.
  - It provides a visual summary of the frequency distribution,
  - By grouping data into intervals or bins, and displaying the number of data points falling within each bin.

USEFULNESS OF HISTOGRAMS

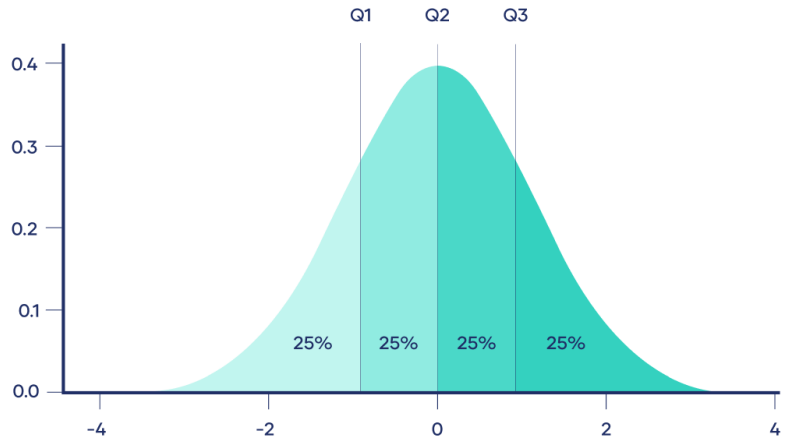
- ✓ Histograms are valuable (visual) tools in exploratory data analysis.
  - They are useful for understanding the data's shape, central tendency, and spread.
  - Essential for conducting statistical tests and drawing conclusions.

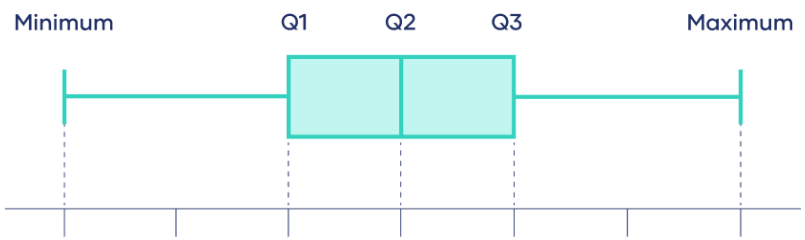
- ✓ Before performing statistical tests, histograms help us understand underlying patterns in data and gain insights into their characteristics.

Q-Q PLOT

QUANTILE

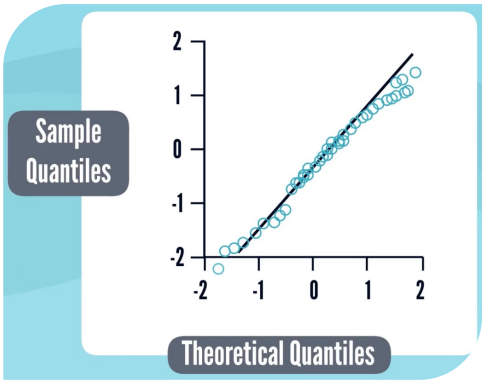
- ✓ In statistics, a quantile is a value that divides a dataset into intervals with a specific proportion of the data falling below it.
  - Quantiles are particularly useful for summarizing the distribution of data and identifying patterns, such as skewness, outliers, or central tendency.





  - The first quartile (Q1) is the 25th percentile ➔ 25% of the data is below this value.
  - The second quartile (Q2) is the 50th percentile, which is also the median.
  - The third quartile (Q3) is the 75th percentile ➔ 75% of the data below this value.

Q-Q PLOT



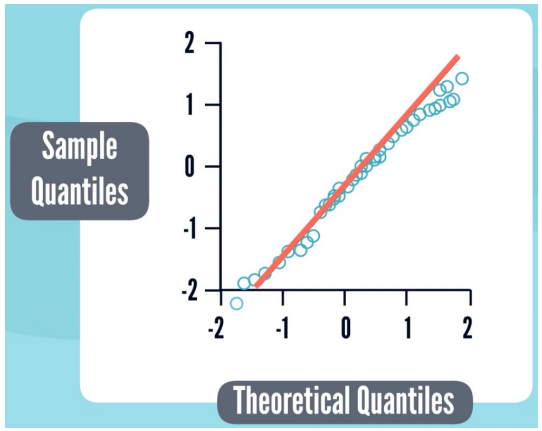
- ✓ A quantile-quantile plot, or q-q plot, graphically assesses whether a data set aligns with a theoretical distribution, such as the normal distribution.
- ✓ This tool enables analysts to visually assess the degree to which the data match the expected distribution.

- ✓ Q-Q plots are therefore crucial in determining whether the assumptions underlying statistical tests are met,
  - which is essential for conducting valid and reliable statistical analyses.
- ✓ By using Q-Q plots, analysts can ensure the data appropriately fit the required distributional criteria,
  - thereby strengthening the credibility and accuracy of their statistical inferences.

HOW TO INTERPRET A Q-Q PLOT

The process:

- First, calculate the quantiles for your sample data.
- Then, calculate the corresponding quantiles of the theoretical distribution for comparison.
- The theoretical quantiles are plotted on the x-axis.
- The sample quantiles are plotted on the y-axis.

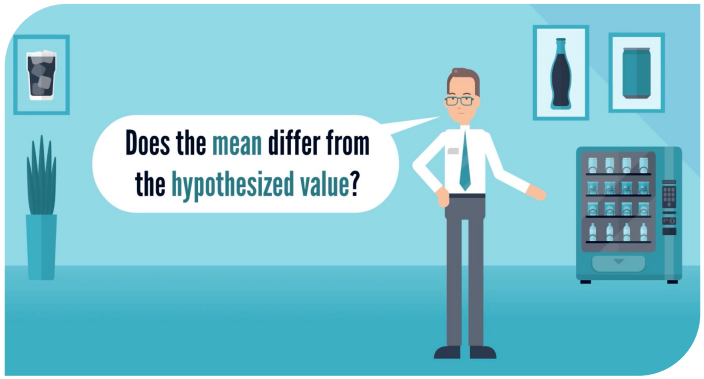


✓If the points on the Q-Q plot lie approximately along a straight line, it suggests that the sample distribution is similar to the theoretical distribution.

ONE-SAMPLE T-TEST

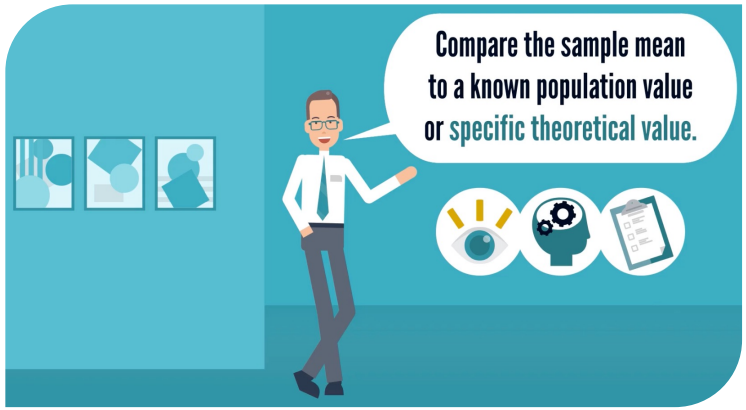
Université de Technologie Tarbes Occitanie Pyrénées

RESEARCH QUESTION



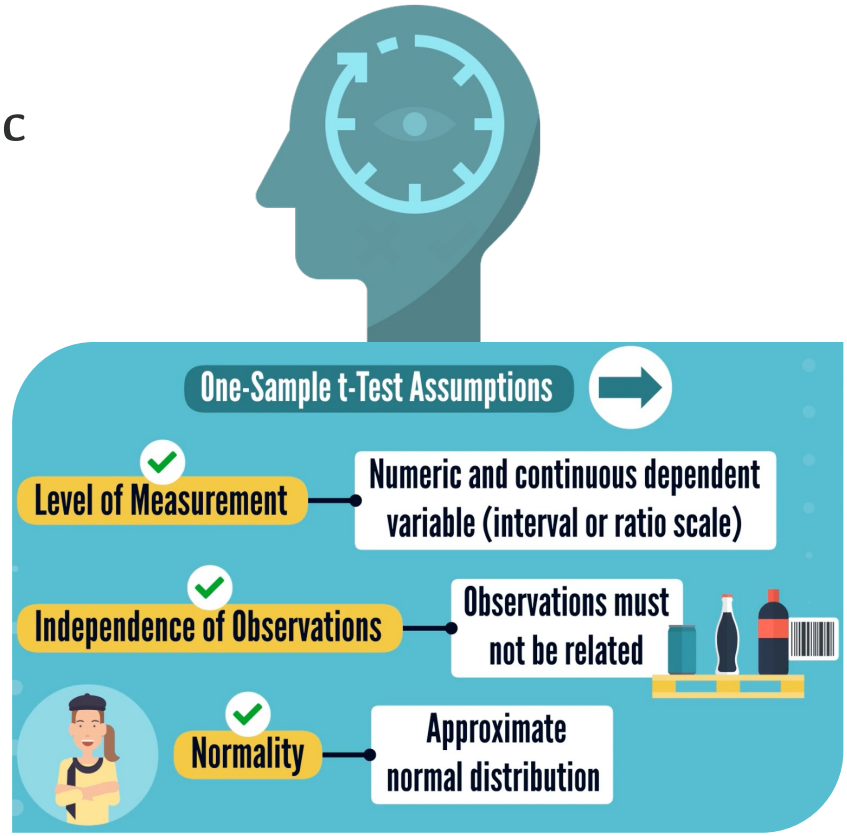
- ✓ The one-sample t-test is a statistical hypothesis test used to determine whether the mean of a single sample significantly differs from the hypothesized value.

- It's beneficial when you compare the sample mean to a known population value or a specific theoretical value.
- The hypothesized value can be chosen based on the researcher's or analyst's expert intuition, prior knowledge, or existing reports.

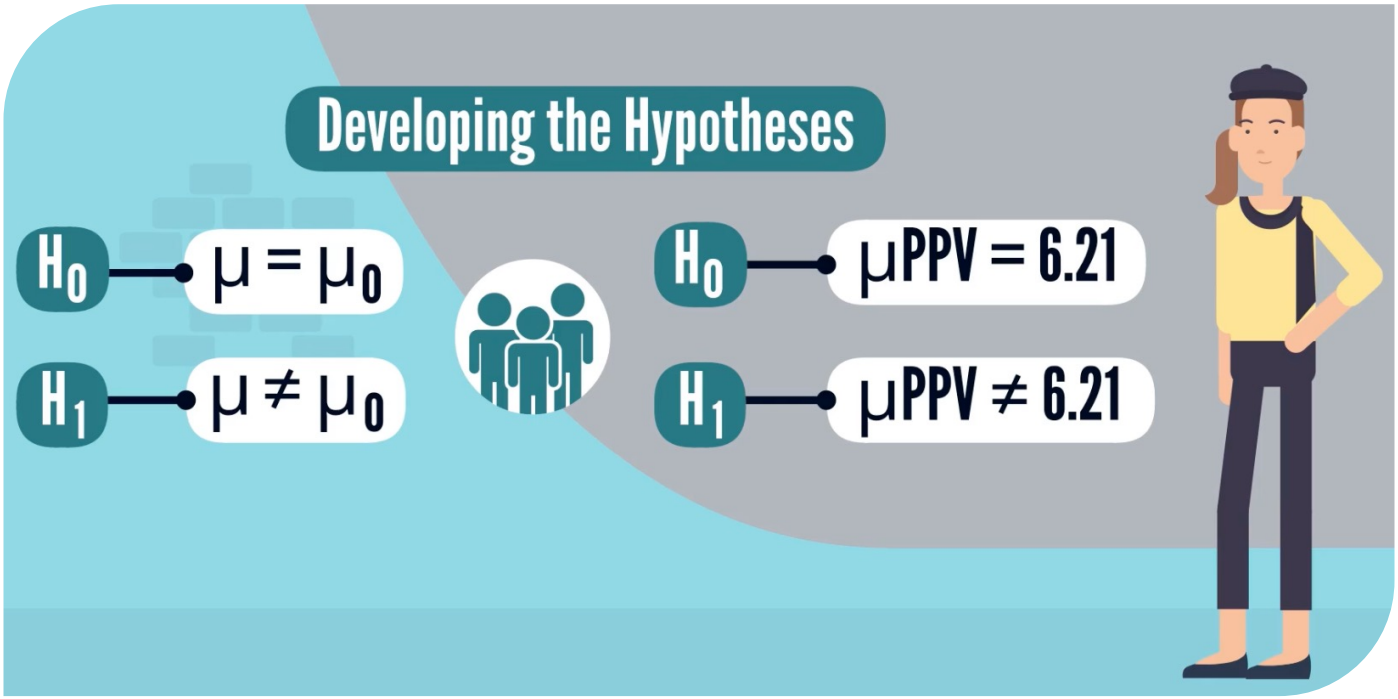


ONE-SAMPLE T-TEST ASSUMPTIONS

- ✓ Level of measurement
  - The dependent variable should be numeric and continuous,
  - Measured on an interval or ratio scale.
- ✓ Independence of observations
  - Observations must not be related.
  - This is often assumed when data is randomly and independently collected.
- ✓ Normality
  - Both the sample and the population for the test variable should exhibit an approximate normal distribution.

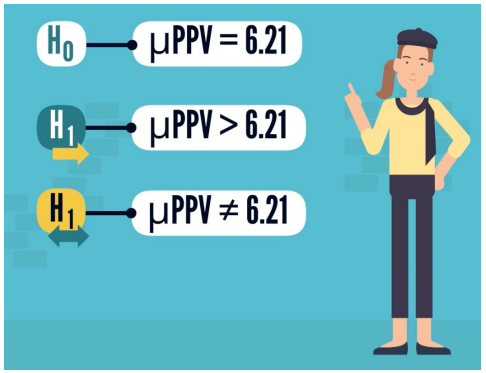


EXAMPLES T-TEST HYPOTHESES



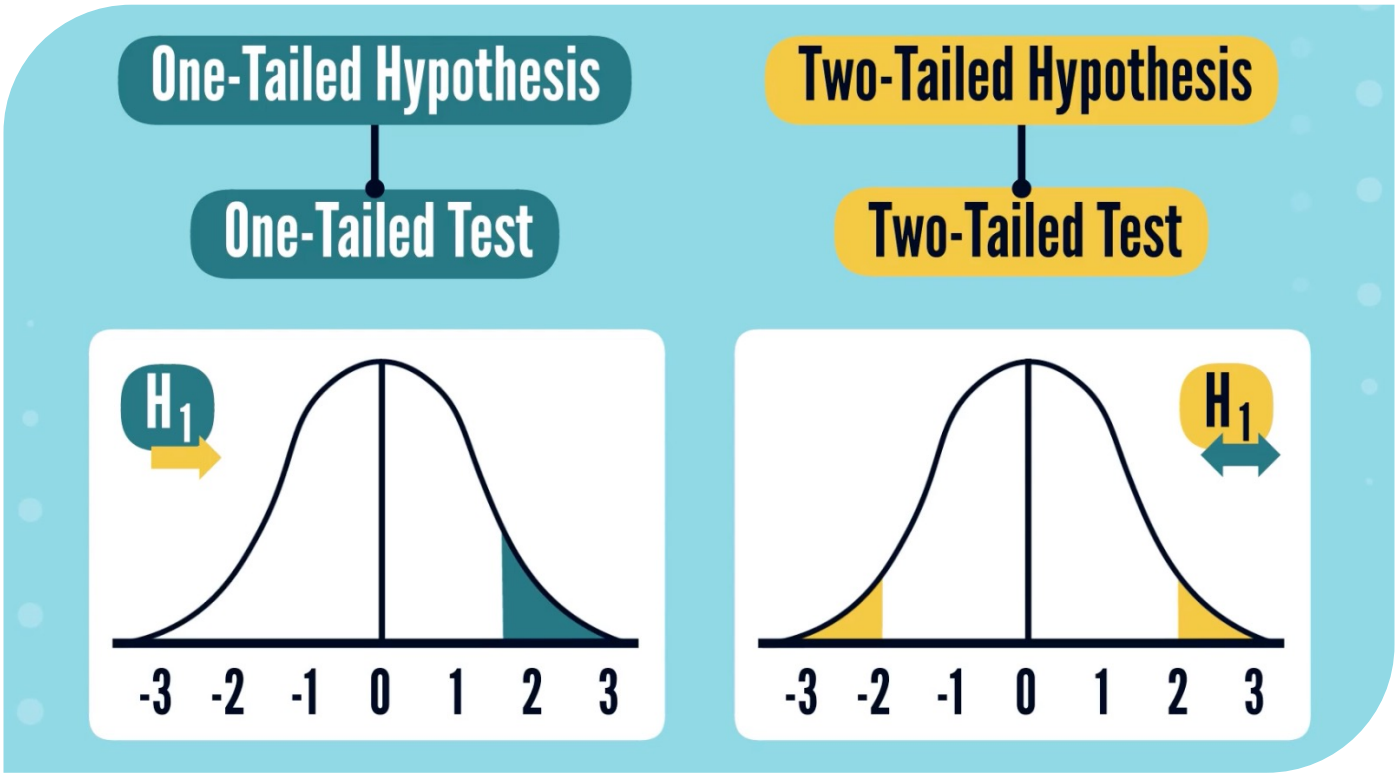
ONE-TAILED VS TWO-TAILED TEST

- ✓ Both alternative hypotheses evaluate evidence for a particular effect or difference.
- ✓ They vary in the direction of their claims.

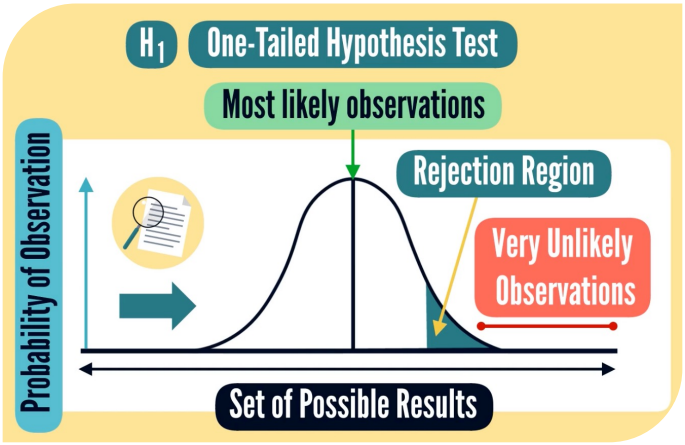


- In a <u>one-tailed</u> or one-sided alternative hypothesis, the population parameter is distinctly specified as exceeding or being less than the hypothesized value ➔ a specific or directional claim.
- In a <u>two-tailed</u> or two-sided alternative hypothesis, the claim regarding the population parameter is non-directional.

DIFFERENCES BETWEEN ONE & TWO-TAILED



ONE-TAILED TEST

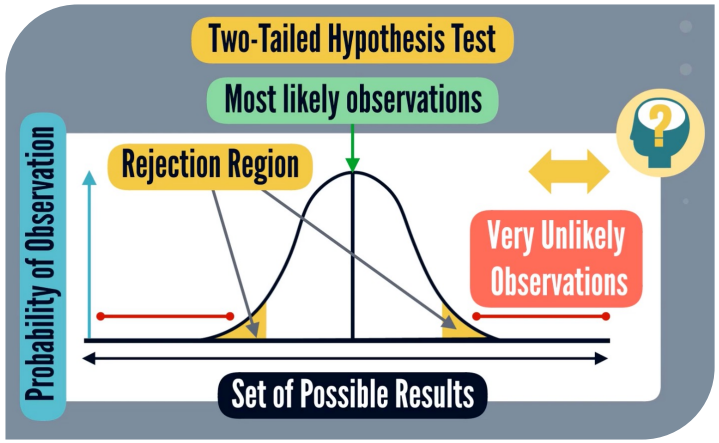


- ✓ In a one-tailed hypothesis test, the critical region is located on only one side of the distribution:
  - either on the right or left, depending on the direction specified by the alternative hypothesis.

- ✓ When to use this test?
  - There is strong theoretical or practical evidence supporting a specific direction of effect, with no interest in effects in the opposite direction.
  - The decision to reject or not reject the null hypothesis is based on whether the observed values fall within this one-sided critical region.

TWO-TAILED TEST

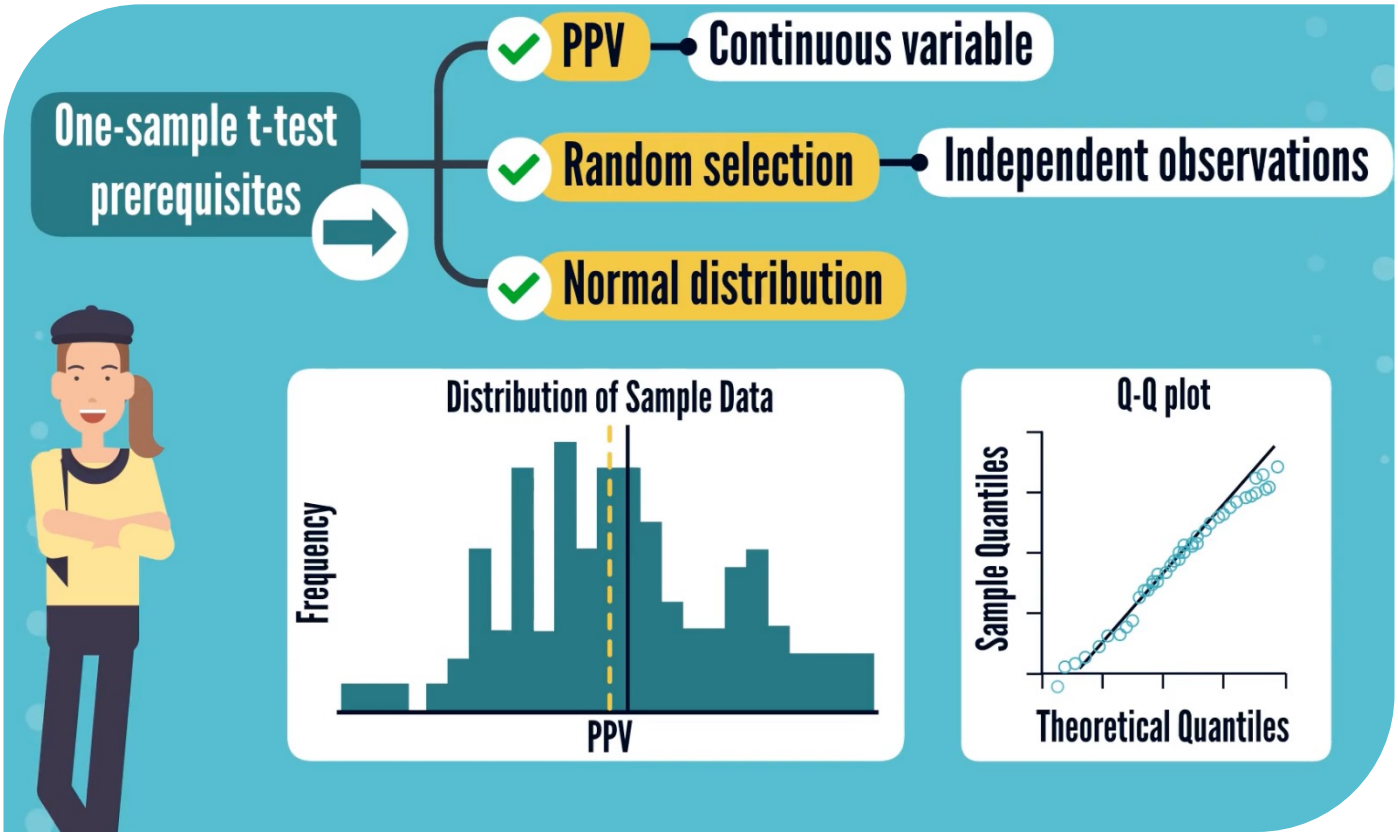
- ✓ In a two-tailed hypothesis test, the critical regions are divided between both tails of the distribution to account for extreme values on either side of the hypothesized value.



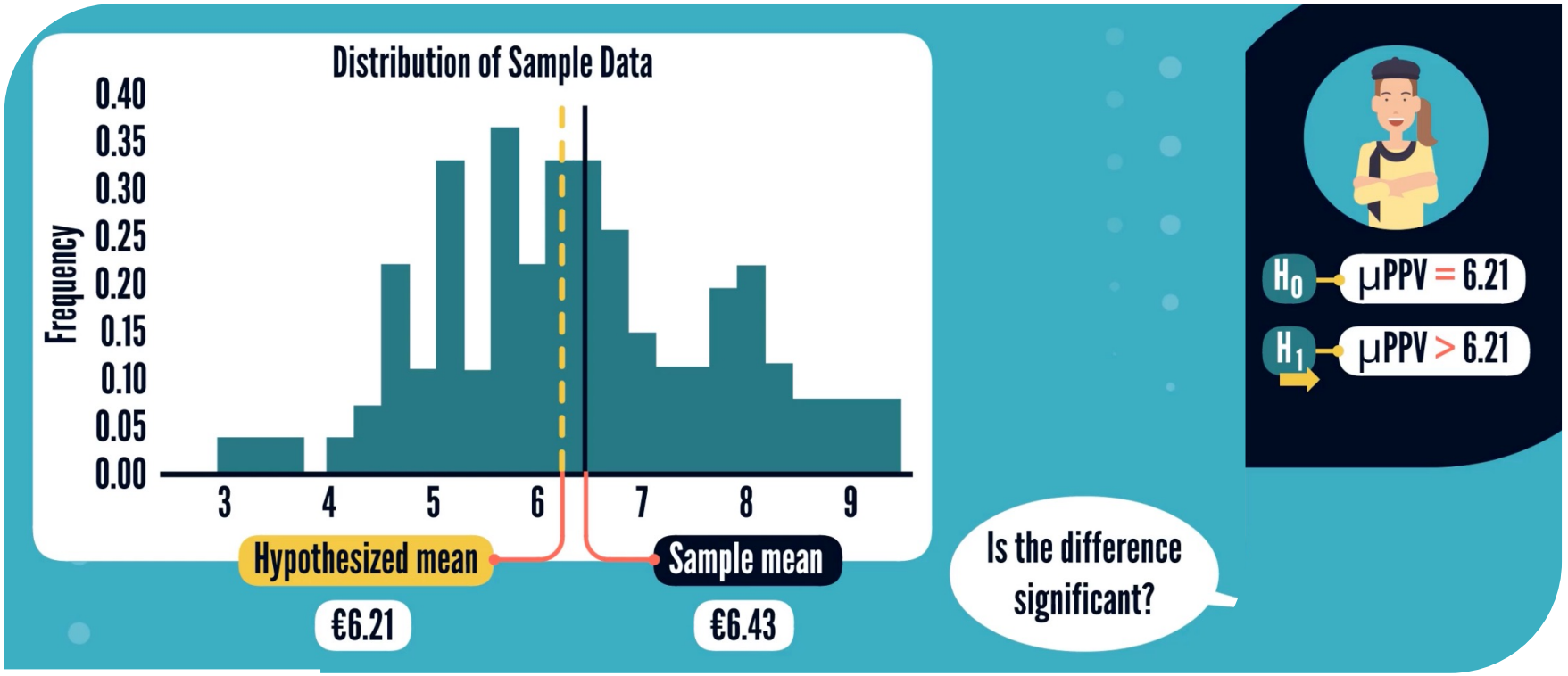
- ✓ When to use this test?
  - ○There is no strong prior assumption or theory predicting the direction of the effect, and the goal is to determine if there is a significant deviation from the hypothesized value in either direction.
  - ○The decision to reject or not reject the null hypothesis is based on whether extreme values are found in either tail of the critical region.

SUMMARY

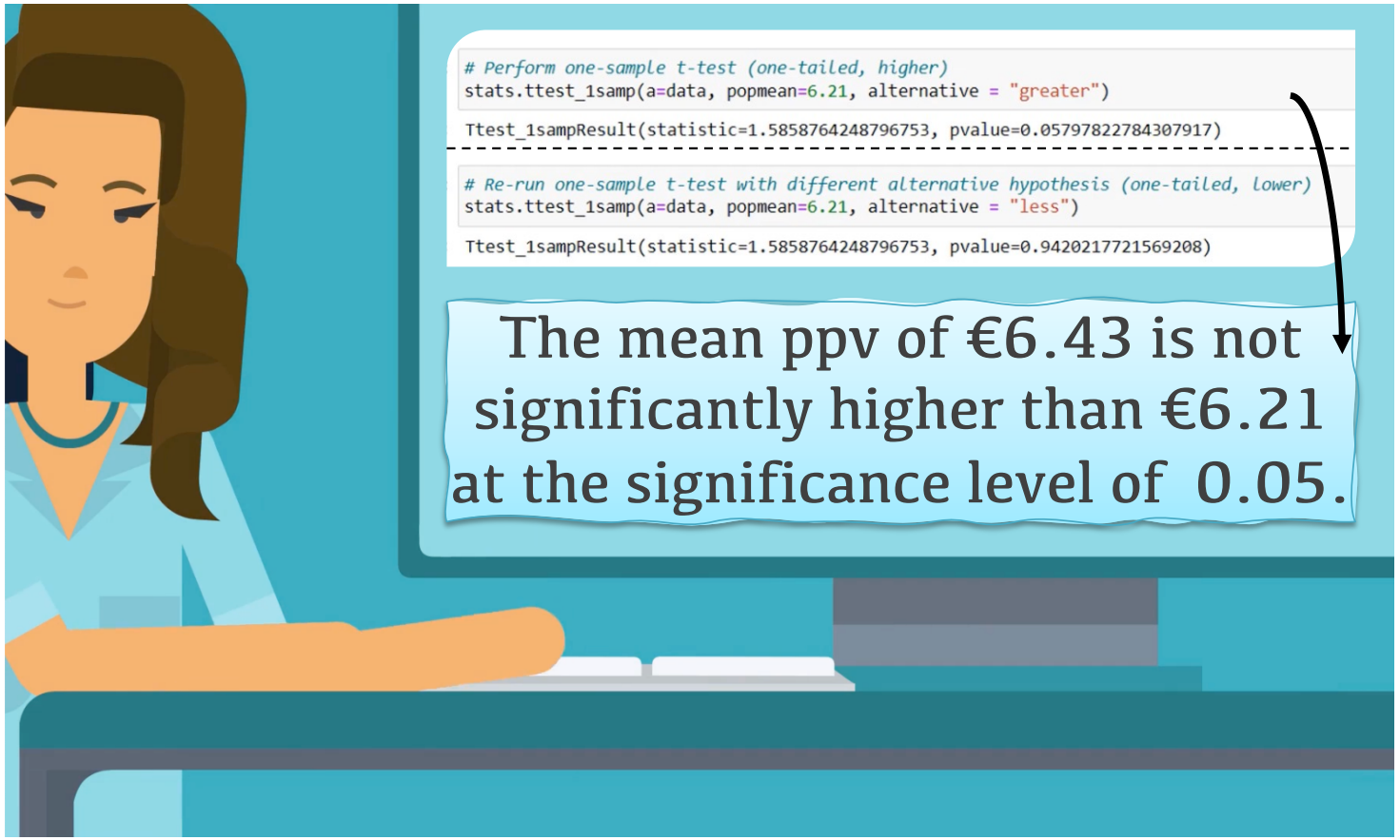
ONE-SAMPLE T-TEST ASSUMPTIONS

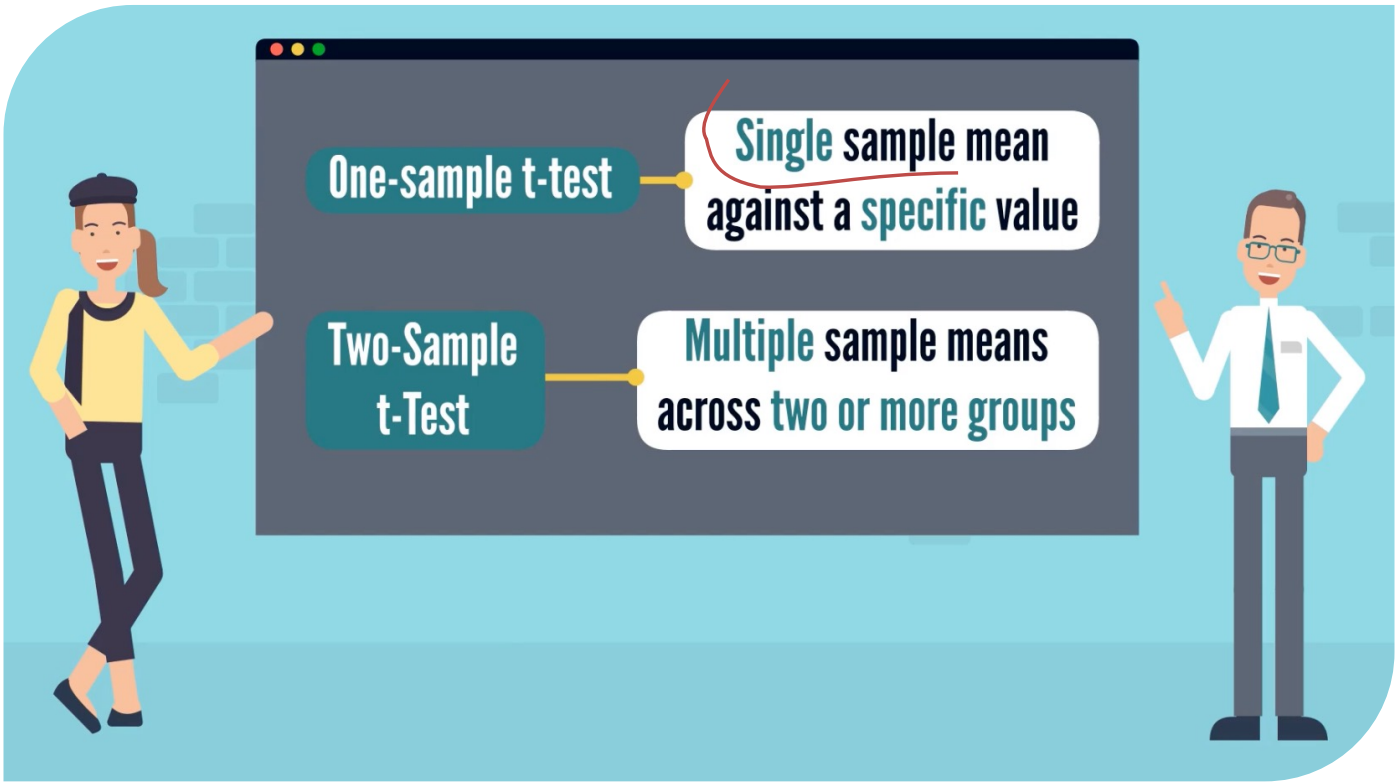


RELATIVE DIFFERENCE: OBSERVED & HYPOTHESIZED MEANS



RESULTS


# Perform one-sample t-test (one-tailed, higher)
stats.ttest_1samp(a=data, popmean=6.21, alternative = "greater")
Ttest_1sampResult(statistic=1.5858764248796753, pvalue=0.05797822784307917)
# Re-run one-sample t-test with different alternative hypothesis (one-tailed, lower)
stats.ttest_1samp(a=data, popmean=6.21, alternative = "less")
Ttest_1sampResult(statistic=1.5858764248796753, pvalue=0.9420217721569208)
The mean ppv of €6.43 is not significantly higher than €6.21 at the significance level of 0.05.


One-sample t-test
Single sample mean against a specific value
Two-Sample t-Test
Multiple sample means across two or more groups