MBA USP ESALQ

# Introduction to the Statistics

Prof. Wilson Tarantin Jr.

# Types of variables

- Variables can be classified into:
  - **Qualitative**: they are non-metric variables, and attribute categories or classifications
    - They can attribute two or more categories
    - The descriptive analysis of qualitative variables is done through frequency tables and charts, as these variables do not allow the calculation of position and dispersion measures
  - **Quantitative**: they are metric variables, and attribute count or measurement
    - They can be discrete or continuous
    - The descriptive analysis of quantitative variables can be done by several statistical tools, including the position and dispersion measures

# Types of variables

- **Qualitative**: examples
  - Likert scale
  - Income bracket
  - Nationality
  - Marital status
  - Education level
  - Soil type
  - Vaccinated or not
  - Comorbidities types

# Types of variables

- **Quantitative**: examples
  - Lifetime
  - Income (in R$)
  - Quantity of children
  - Height of the person (in cm)
  - Weight (in kg)
  - Return of shares on the stock market
  - Environment temperature
  - Profit/loss of the company

# Descriptive statistics

# 1. Frequency table

- Quantity of occurrences per category
  - **Qualitative**
    - Directly, it presents the amount of occurrences for each category
  - **Quantitative**
    - Discrete variable: the analysis is similar to the case of the qualitative variable, that is, it shows the amount of occurrences for each discrete value of the variable
    - Continuous variable: an initial categorization by classes or ranges is necessary in order to present the amount of occurrences in each category generated

# 1. Frequency table

- Elaborating a frequency table
  - The types of frequencies reported are:
    - **Absolute frequency**: occurrence count in each category
    - **Relative frequency**: percentage of each category compared to the total of observations
    - **Cumulative frequency**: sum of the absolute frequency for each new category
    - **Cumulative relative frequency**: sum of the frequency relative to each category
  - **Example:** Data of the country of origin of 300 people who were in a lecture: The frequency table for the variable "country of origin" is available in the support spreadsheet in the **tab Frequency Tables**.

# 2. Position measures

- **Mean**
- **Median**
- **Mode**
- **Percentiles**
- **Quartiles**
- **Deciles**

# 2. Position measures

- **Mean**
  - It is a simple arithmetic mean for the variable, that is, the sum of values ($X_i$) contained in the variable divided by the total amount of observations (n) ($n$)
  - $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

# 2. Position measures

- **Median**
  - It is the central element of the variable distribution, considering that the variable is with its *n* values organized in ascending order
  - Half of the values of the variable are higher or equal to the value of the median, and half of the values are smaller or equal to the value of the median
  - $$Md(X) = \begin{cases} \dfrac{X_{\frac{n}{2}} + X_{\left(\frac{n}{2}\right)+1}}{2} & , \ \textbf{\textit{If n is even}} \\ X_{\left(\frac{n+1}{2}\right)} & , \ \textbf{\textit{If n is odd}} \end{cases}$$

# 2. Position measures

- **Mode**
  - It is the value that appears most frequently in the observations of a variable
  - Mode can also be calculated for qualitative data
  - It is possible that there is not the mode of a variable (especially if it is a continuous variable)
    - It occurs when no value is repeated

# 2. Position measures

- **Percentiles**
  - It's the elements of the distribution of the variable that divide the observations into **hundred equal parts**, considering that the variable is with its values organized in ascending order
    - 14th Percentile
    - 42th Percentile
    - 60th Percentile ...
  - $\boldsymbol{Pos(P_i) = \left[(n-1).\left(\frac{P_i}{100}\right)\right] + 1}$

# 2. Position measures

- **Quartiles**
  - They are the elements of the distribution of the variable that divide the observations into **four equal parts**, considering that the variable is with its values organized in ascending order
    - 1st Quartile: 25% of the observations are smaller than the 1st quartile
    - 2nd Quartile: it is the median
    - 3rd Quartile: 25% of the observations are higher than the 3rd quartile
    - 1st Quartile = 25th Percentile
    - 2nd Quartile = 50th Percentile
    - 3rd Quartile = 75th Percentile

# 2. Position measures

- **Deciles**
  - It is the elements of the distribution of the variable that divide the observations into **ten equal parts**, considering that the variable is with its values organized in ascending order
    - 1st Decile
    - 3th Decile
    - 8th Decile ...
    - 1st Decile = 10th Percentile
    - 3rd Decile = 30th Percentile
    - 8th Decile = 80th Percentile

# 3. Dispersion measures

- **Range**
- **Variance**
- **Standard deviation**
- **Standard error**
- **Coefficient of variation**

# 3. Dispersion measures

- **Range**
  - It presents the difference between the maximum value and the minimum value of a variable
  - $\boldsymbol{A = X_{máx} - X_{mín}}$
    - Maximum value: greater value of the variable
    - Minimum value: lowest value of the variable

# 3. Dispersion measures

- **Variance**
  - It shows the dispersion of the observations of a variable around its mean
  - $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$
    - In this case, it is the sample variance

# 3. Dispersion measures

- **Standard deviation**
  - It is a measure derived from variance, making the interpretation of the dispersion more simple around the mean
    - The variance is defined in square terms, which makes it difficult for interpretation
  - The standard deviation is the square root of variance
  - $S = \sqrt{S^2}$

# 3. Dispersion measures

- **Standard error**
  - It is the standard deviation of the mean of the variable
  - $S_{\bar{X}} = \frac{S}{\sqrt{n}}$
    - *S* is the standard deviation of the variable and *n* the sample size
  - The higher the sample size, the less the standard error in the estimate of the mean of the variable more precise is the estimated mean

# 3. Dispersion measures

- **Coefficient of variation (CV)**
  - It is a measure of relative dispersion, as it relates the standard deviation and the mean of a variable
  - It can be used to perform comparisons between samples, for example.
  - The smaller the CV, the more homogeneous the values of the variable and more concentrated are the values in relation to the mean
  - $\boldsymbol{CV = \frac{S}{\bar{X}} . 100}$

# 4. Form measures

- **Skewness and Kurtosis**
  - Skewness: place of concentration of the distribution
    - Symmetric Curve: **Mean = Median = Mode**
    - Asymmetric Curves – Right: it has the longer tail to the right Mean **> Median**
    - Asymmetric Curves – Left: it has the longer tail to the left Mean **< Median**
  - Fisher's Coefficient of Skewness:
  - $g_1 = \frac{n^2 \,.\, M_3}{(n-1).(n-2).S^3}$ in which $M_3 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n}$

# 4. Form measures

- **Skewness and Kurtosis**
  - Kurtosis: flatness of the distribution curve
    - The normal curve as reference, it is possible to observe if the curves have smaller dispersion or greater dispersion around the mean
  - Fisher's Coefficient of Kurtosis
  - $g_2 = \frac{n^2 . (n+1) . M_4}{(n-1).(n-2).(n-3).S^4} - 3 . \frac{(n-1)^2}{(n-2).(n-3)}$ in which $M_4 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{n}$

# 4. Form measures

- **Skewness and Kurtosis**
  - **Skewness**
    - $g_1 = 0$ it indicates symmetric curve
    - $g_1 > 0$ it indicates positive asymmetric curve (right)
    - $g_1 < 0$ it indicates negative asymmetric curve (left)
  - **Kurtosis**
    - $g_2 = 0$ it indicates curve with normal distribution
    - $g_2 > 0$ it indicates curve with elongated distribution
    - $g_2 < 0$ it indicates curve with flat distribution

# Position, dispersion and form

- **Joint application of measures**

  - **Example:** A consumer is analyzing the price of a product that you want to buy. To generate more information for their decision, they collect 100 prices for the product. Since the "price" is a quantitative variable, the analyses will be carried out through the measures of position, dispersion and form approached previously.

    - The database with the 100 price observations is in the support spreadsheet **in the Descriptive – Quantitative tab**.

# 5. Relation between variables

- **Bivariate analysis measures**
  - Often, the interest can be in the relation between <u>two</u> variables. In these cases, before it is important to know the type of variable:
    - **Qualitative variables**: analysis of **the association** by the chi-squared test ($\chi^2$)
    - **Quantitative variables**: **correlation** analysis through covariance and Pearson correlation coefficient

# 5. Relation between variables

- **Chi-squared test: qualitative variables**
  - It begins through a joint distribution table of frequencies (or cross-classification table) that presents the **absolute frequencies observed** for each pair of categories of variables

| | | Variable B | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Category 1** | **Category 2** | **Category 3** | **…** | **Category J** | **Total** |
| **Variable A** | **Category 1** | $n_{11}$ | $n_{12}$ | $n_{13}$ | … | $n_{1J}$ | $\Sigma_{L1}$ |
| | **Category 2** | $n_{21}$ | $n_{22}$ | $n_{23}$ | … | $n_{2J}$ | $\Sigma_{L2}$ |
| | **Category 3** | $n_{31}$ | $n_{32}$ | $n_{33}$ | … | $n_{3J}$ | $\Sigma_{L3}$ |
| | **…** | … | … | … | … | … | … |
| | **Category I** | $n_{I1}$ | $n_{I2}$ | $n_{I3}$ | … | $n_{IJ}$ | $\Sigma_{LI}$ |
| | **Total** | $\Sigma_{C1}$ | $\Sigma_{C2}$ | $\Sigma_{C3}$ | … | $\Sigma_{CJ}$ | $N$ |

# 5. Relation between variables

- **Chi-squared test: qualitative variables**
  - Next, the **absolute frequencies expected** for each pair of categories of variables are identified
  - ***absolute frequency expected*** $_{11} = \frac{(\sum \boldsymbol{L1} \,.\, \sum \boldsymbol{C1})}{\boldsymbol{N}}$
    - The same calculation is performed for each pair of categories of the contingency table, keeping the respective line and column correspondences in the numerator

# 5. Relation between variables

- **Chi-squared test: qualitative variables**
  - Subsequently, the **residues** for each pair of categories of variables are identified
  - ***residues*** $_{11}$ = ***absolute frequency observed*** $_{11}$ - ***absolute frequency expected*** $_{11}$
    - Residues are identified for each pair of categories of the contingency table

# 5. Relation between variables

- **Chi-squared test: qualitative variables**
  - Finally, the $\chi^2$ values of each pair of categories are calculated
  - $$\chi^2 = \frac{(residue)^2}{(absolute\ frequency\ expected)}$$
  - And the $\chi^2$ individual values are added to obtain the value *of* χ2 total of the analysis

# 5. Relation between variables

- **Chi-squared test: qualitative variables**
  - Based on the total value $\chi^2$, the following test is performed:
    - $H_0$: the variables are associated in a random way
    - $H_1$: the association between the variables is not done in a random way
  - Considering the level of significance and the degrees of freedom, if the value of $\chi^2$ statistics is greater than its critical value, there is a significant association between the two variables ($H_1$)
    - Critical value of the $\chi^2$ distribution with (I – 1) (J – 1) degrees of freedom

# 5. Relation between variables

- **Chi-squared test: qualitative variables**

  - **Example:** A study was done with 200 people to analyze the joint behavior of the "health plan operator" with the variable "level of satisfaction" of the consumer. The objective is to analyze if there is the statistically significant association between these variables. (Source: Fávero and Belfiore, 2017, Cap. 8)

    - The data of the contingency table obtained from the sample is in the support spreadsheet in **the Association – Qui² tab**.

# 5. Relation between variables

- **Pearson correlation coefficient**
  - It is used to identify the correlation between two quantitative variables
  - It begins by the calculation of the covariance between the two variables and subsequently it is obtained the Pearson correlation coefficient
  - $\boldsymbol{cov(X,Y)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \,.\, (Y_i - \bar{Y})}{n-1}$
  - $\boldsymbol{\rho_{XY}} = \frac{cov(X,Y)}{S_X \,.\, S_Y}$, since it $\boldsymbol{\rho_{XY}}$ varies between -1 and 1

$\rho_{XY}$ = -1 (perfect negative)
$\rho_{XY}$ = 0 (without correlation)
$\rho_{XY}$ = 1 (perfect positive)

# 5. Relation between variables

- **Pearson correlation coefficient**

  - **Example:** the course coordinator wants to analyze if there is a correlation between the students grades in different subjects. To do so, they set a database with the grades of 30 students for the mathematics, physics and literature subjects. Next, they want to calculate the pairs of correlations between the grades of mathematics – physics, mathematics – literature and physics – literature. Which are the correlations of Pearson obtained?

    - The data are in the support spreadsheet in the **Pearson's Correlation tab**.

# Probability Distributions

# Variables and distributions

- **Primordial characteristics**
  - **Discrete random variable**: it is the one that does not assume decimal values, that is, it is composed of integer values
    - Examples: number of children; amount of patients per day; amount of seeds per hectare; the amount of cases of diseases per day...
  - **Continuous random variable**: it is the one that can assume any values contained in the real numbers
    - Examples: distance between cities; monthly salary; a person height; return of the share...

# Discrete variables

- **Probability distributions:**
  - **Uniform**
  - **Bernoulli**
  - **Binomial**
  - **Negative binomial**
  - **Poisson**

# Discrete variables

- **Discrete uniform distribution**
  - All possible values have the same probability of occurrence
  - $\boldsymbol{P(X = x_i) = \frac{1}{n}}$
  - The parameter *n* represents the amount of possible values

# Discrete variables

- **Discrete uniform distribution**
  - **Example:** The probabilities of the possible results when throwing a die are:
    - P(X=*1*) = 1/6
    - P(X=*2*) = 1/6
    - P(X=*3*) = 1/6
    - P(X=*4*) = 1/6
    - P(X=*5*) = 1/6
    - P(X=*6*) = 1/6

# Discrete variables

- **Bernoulli distribution**
  - The values of the variable can assume only two possible results, which are called success ($x$=1) or failure ($x$=0)
  - The Bernoulli distribution presents the probability of success ($p$) or of failure ($1 - p$) when only one experiment occurs
  - $\boldsymbol{P(X = x) = p^x \,.\, (1 - p)^{1-x}}$

**Binary logistic!**

# Discrete variables

- **Bernoulli distribution**

  - **Example:** The probability ($p$) of a candidate be approved (x=1) in an test for a class board is 48%. If each candidate can only perform the test once, analyze the possible probabilities through the Bernoulli distribution.

    - ***P(X=1)*** = $(0.48)^1 . (1 – 0.48)^0 = 0.48$ or 48%
    - ***P(X=0)*** = $(0.48)^0 . (1 – 0.48)^1 = 0.52$ or 52%

    - X=1 is approval and X=0 it is failure in the test.

# Discrete variables

- **Binomial distribution**
  - The binomial distribution occurs when there are (*n*) independent repetitions of the Bernoulli experiment and the probability of success (*p*) is constant in all repetitions
  - The variable in the binomial model indicates the amount of successes (*k*) in the (*n*) repetitions of the experiment
  - $\boldsymbol{P(X=k) = \binom{n}{k} . p^k . (1-p)^{n-k}}$ in which $\boldsymbol{\binom{n}{k} = \frac{n!}{k!(n-k)!}}$

**Multinomial logistic!**

# Discrete variables

- **Binomial distribution**

  - **Example:** it is known in an industry that the probability (*p*) of finding defective parts in each batch produced is 6.50%. It is produced 12 batches (*n*) of the piece per month. Analyze the following probabilities (*k*):

    a) What is the probability of finding defective parts in 2 batches in the month?
    b) What is the probability of finding defective parts in 4 batches in the month?
    c) What is the probability of finding defective parts in maximum 2 batches?

    - The spreadsheet for help is in the support spreadsheet in the **Binomial Distribution tab**.

# Discrete variables

- **Negative binomial distribution**
  - In the binomial distribution, independent trials of Bernoulli are performed (*x*) until it is obtained (*k*) successes
  - The probability of success (*p*) is constant in all trials performed
  - The variable in the binomial model indicates the amount of tests (*x*)
  - $\boldsymbol{P(X=x) = \binom{x-1}{k-1} . p^k . (1-p)^{x-k}}$ in which $\boldsymbol{\binom{x-1}{k-1} = \frac{(x-1)!}{(k-1)![(x-1)-(k-1)]!}}$

# Discrete variables

- **Negative binomial distribution**

  - **Example:** in a amusement park, there is a machine in which the player must capture any item using the commands of a mechanical arm. Consider that the probability (p) of the player capturing some item in each move is 11%. Identify the following probabilities:

    a) The player needs 10 moves to capture 3 items.
    b) The player needs 20 moves to capture 3 items.
    c) The player needs 5 moves to capture 1 item.

  - The spreadsheet for help is in the support spreadsheet in **the Negative Binomial Distribution tab**.

# Discrete variables

- **Poisson Distribution**
  - The Poisson distribution indicates the probability of the number of successes ($k$) in a certain continuous exposure
  - Examples of exposure: time and area
  - $$\boldsymbol{P(X = k) = \frac{e^{-\lambda} \cdot \lambda^{k}}{k!}}$$
    - In which $\boldsymbol{\lambda}$ is the average rate estimated of occurrence of the event (success) in a exposure

# Discrete variables

- **Poisson Distribution**

  - **Example:** A doctor noticed that the average rate of occurrence ($\lambda$) of patients with a rare disease in their clinic is 2 per year. Accepting that this variable is in Poisson distribution, estimate:

    a) The probability of the doctor attending 1 patient with this disease in one year.
    b) The probability of the doctor attending 3 patients with this disease in one year.
    c) The probability of the doctor not attending patients with this disease in one year.
    d) The probability of the doctor attending 10 patients with this disease in the next two years.

  - The spreadsheet for help is in the support spreadsheet in **the Poisson Distribution tab**.

# Continuous variables

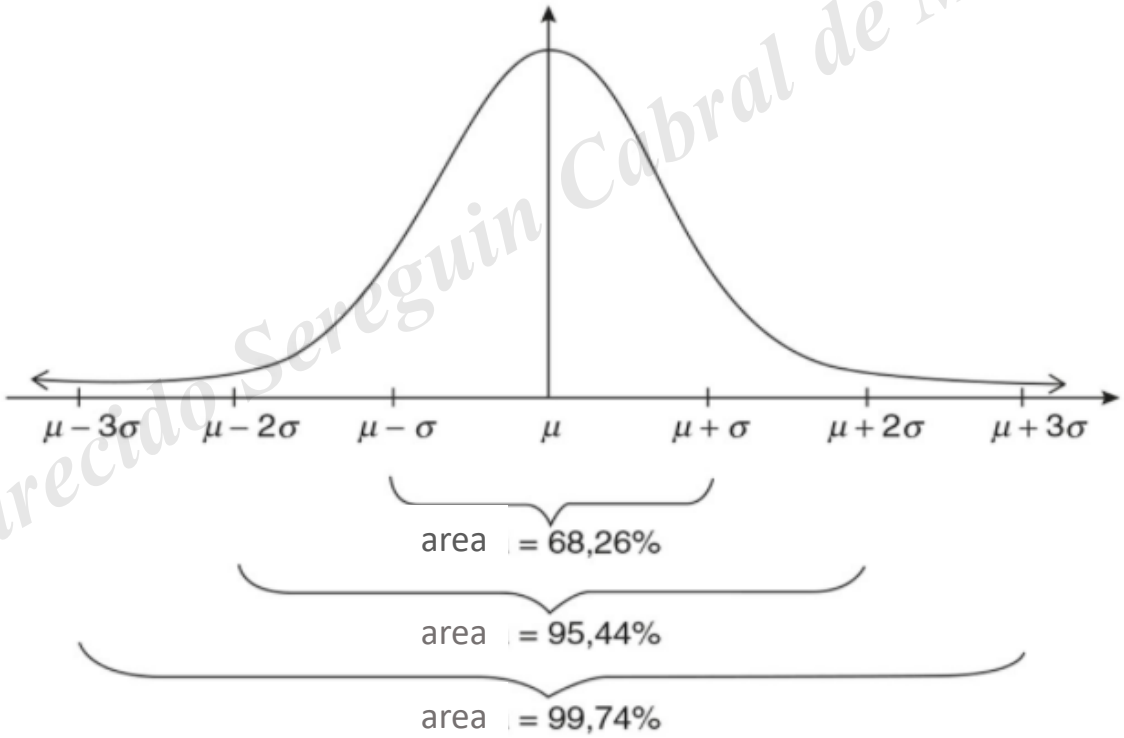
- **Probability distributions:**
  - **Normal (Normal Standard)**
  - **Chi-squared**
  - ***Student's t***
  - **Snedecor's F**

# Continuous variables

- **Normal distribution**
  - It is the Gaussian distribution, with a curve in a shape of bell
  - The relevant parameters of the normal distribution are the mean (μ) and the standard deviation (σ) of the variable
  - $$f(x) = \frac{1}{\sigma . \sqrt{2\pi}} . e^{-\frac{(x-\mu)^2}{2 . \sigma^2}}$$

# Continuous variables

- **Normal distribution**





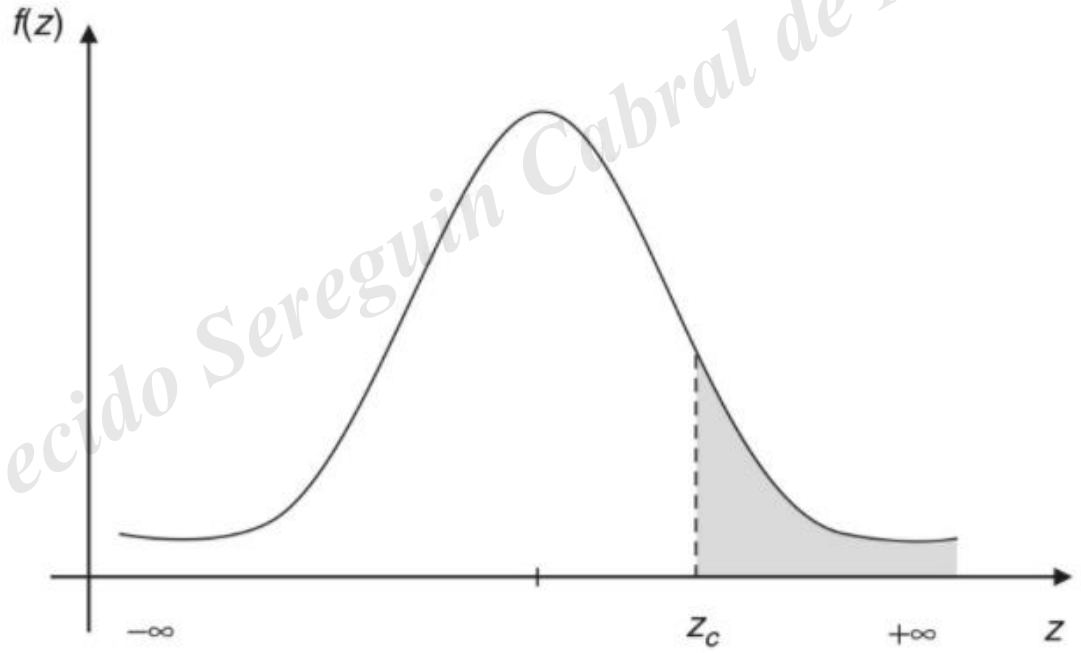

Source: Fávero e Belfiore (2017, Cap. 5)

# Continuous variables

- **Standard normal distribution**
  - To obtain the standard normal, the variable is transformed through Z-score
  - The transformation by Z-score makes the variable mean = 0 and standard deviation = 1, and it does not change the original distribution
  - $Z = \frac{(X - \mu)}{\sigma}$
  - **Example:** critical values in a standard normal table and in Excel

# Continuous variables

- **Standard normal distribution**




Source: Fávero e Belfiore (2017, Cap. 5)

# Continuous variables

- **Standard normal distribution**

  - **Example:** An investor calculated that the monthly average return of a share on the stock market was 2.80%. In the same period, the standard deviation of the share was 1.20%. Based on the normal distribution, calculate:

    a) The probability of the return of the share is greater than 4% per month.
    b) The probability of the return of the share is lower than 3% per month.
    c) The probability of the return of the share is negative.
    d) The probability of the return of the share is greater than 1% and lower than 5% per month.

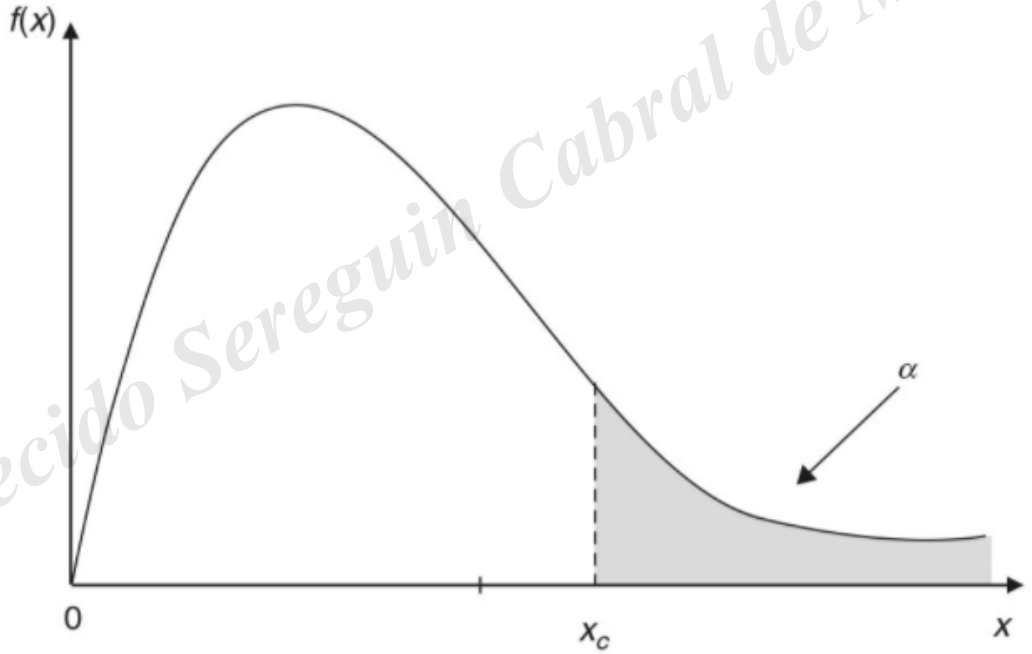
    - The spreadsheet for help is in the support spreadsheet in **the Normal Distribution tab**.

# Continuous variables

- **Chi-squared distribution ($\chi^2$)**
  - The distribution $\chi^2$ presents asymmetric and positive curve
  - It is formed from the sum of squares of *v* independent random variables with standard normal distribution
  - **Example:** critical values in a chi-squared table and in Excel

# Continuous variables

- **Chi-squared distribution ($\chi^2$)**




Source: Fávero e Belfiore (2017, Cap. 5)

# Continuous variables

- **Chi-squared distribution ($\chi^2$)**

  - **Example:** A researcher in botany identified that a variable of their study follows a chi-squared distribution and it has 7 degrees of freedom. Based on this information, the researcher calculated:

    a) The probability of it finding a value X>6.
    b) The probability of it finding a value X<8.
    c) The value of X that makes the P(X>x) 5%.
    d) The value of X that makes the P(X<x) 90%.

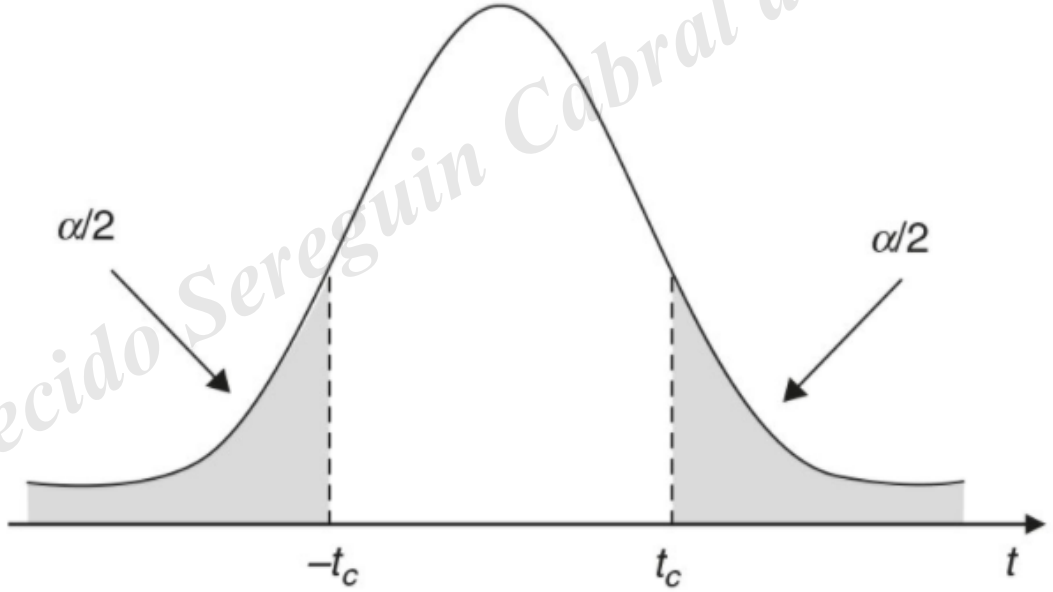
    - The spreadsheet for help is in the support spreadsheet in **the Chi-squared Distribution tab**.

# Continuous variables

- ***Student's T* Distribution**
  - The *Student's T* Distribution is similar to the standard normal distribution, that is, it it is bell-shaped and it is symmetric around the mean
  - On the other hand, the *Student's T* Distribution has longer tails and this allows more extreme values
  - An application of the *Student's T* Distribution is the hypothesis test for means
  - **Example:** critical values in a *Student's* T table and in Excel

# Continuous variables

- ***Student's T* Distribution**




Source: Fávero e Belfiore (2017, Cap. 5)

# Continuous variables

- ***Student's T* Distribution**

  - **Example:** The manager of a company's quality control identified that a relevant variable for its control presents Student's T distribution and it has 7 degrees of freedom. What are your analyses in these situations:

    a) The probability of it finding T > 2.5.
    b) The probability of it finding T < - 2.5.
    c) The probability of it finding T>-1 and T<2.
    d) The T value so that P(T>t) = 5%.

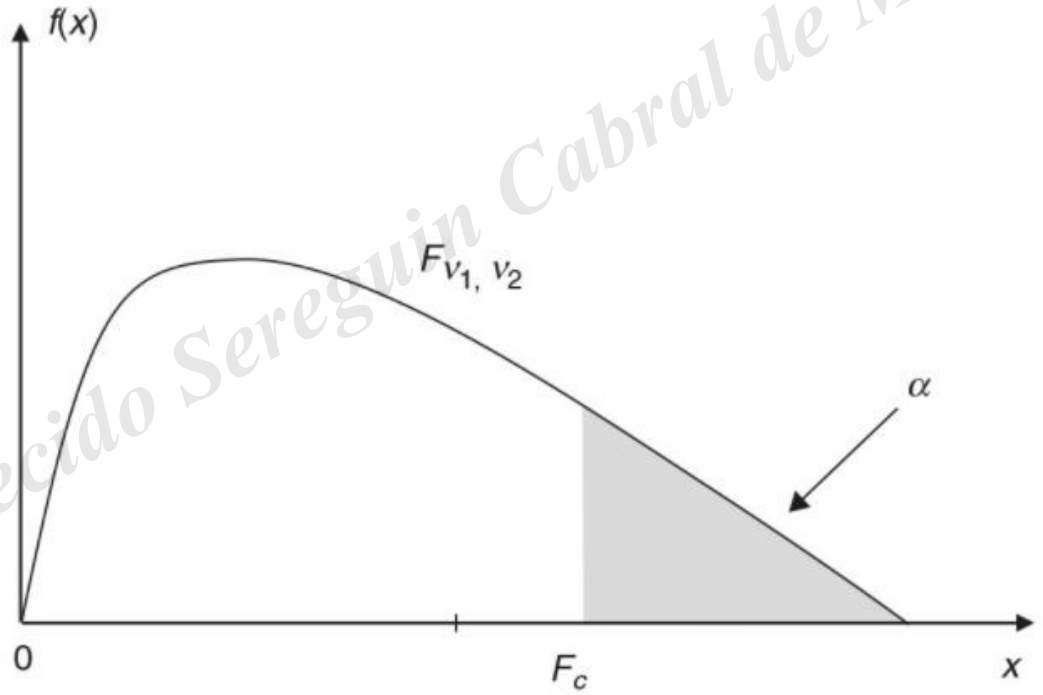
  - The spreadsheet for help is in the support spreadsheet in the Student's **T Distribution tab**.

# Continuous variables

- **Snedecor's F-Distribution**
  - A common application of the Snedecor's F-Distribution (Fischer) is the analysis of variances
  - It has distribution of asymmetric and positive probabilities when the degrees of freedom $v_1$ and $v_2$ are small
  - **Example:** critical values in a Snedecor's F table and in Excel

# Continuous variables

- **Snedecor's F-Distribution**




Source: Fávero e Belfiore (2017, Cap. 5)

# Continuous variables

- **Snedecor's F-Distribution**

  - **Example:** A data scientist is evaluating a variable with Snedecor's F-distribution. This variable presents 17 degrees of freedom in the numerator and 28 degrees of freedom in the denominator. Determine:

    a) The probability of it finding $X > 1.5$.
    b) The probability of it finding $X < 1.0$.
    c) The probability of it finding $2<X<3$.
    d) The F-value so that $P(X>x) = 5\%$.

  - The spreadsheet for help is in the support spreadsheet in the **Snedecor's F-Distribution tab**.

# Hypothesis Tests

# Hypothesis Tests

- **Purpose**
  - In some cases, we can be interested in testing characteristics about population parameters, such as mean and standard deviation (variance)
  - Given the impossibility of obtaining the population data, we use the population samples
  - To test the interest parameter through samples, we use the statistical hypothesis tests

# Tests types

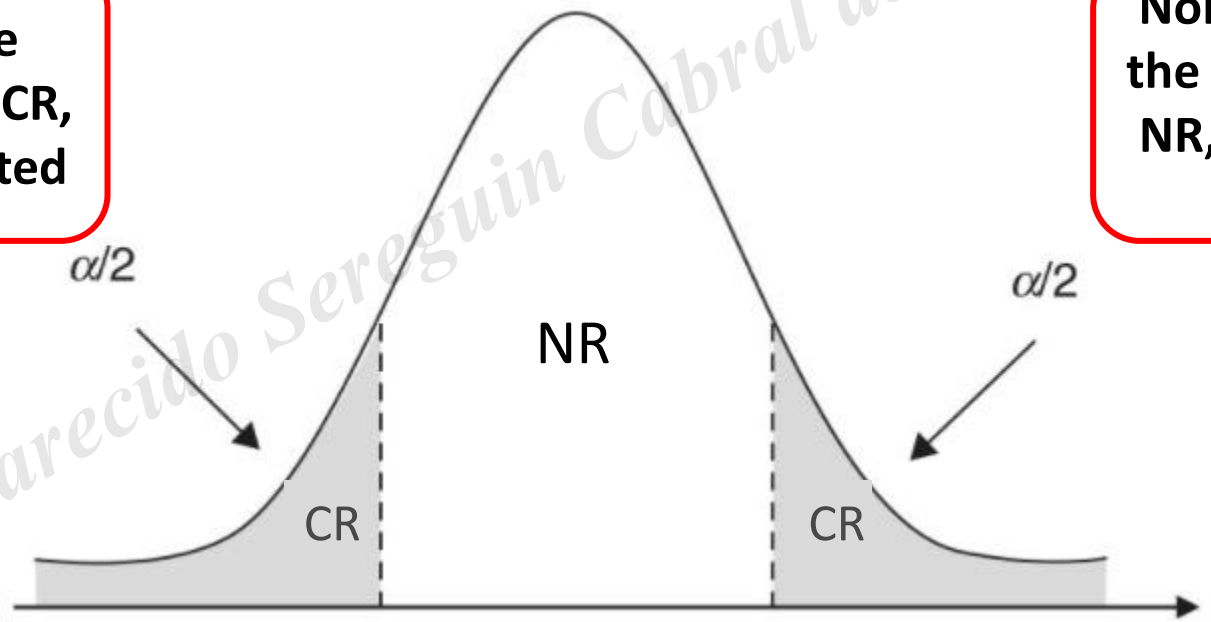
- **Two-tailed test**
  - In the two-tailed test, for a θ parameter, the interest is to test:
    - $H_0$: θ = $θ_0$ (null hypothesis)
    - **$H_1$: θ θ $θ_0$ (alternative hypothesis)**
  - The objective is to verify if the parameter is **statistically different** of a certain value of interest
  - It is necessary to define the **significance level (α)** desired for the analysis

# Tests types

- **Two-tailed test (bilateral)**

**Critical Region (CR): if the statistics of the test falls in CR, the null hypothesis is rejected**

**Non-Rejection Region (NR): if the statistics of the test falls in NR, the null hypothesis is not rejected**



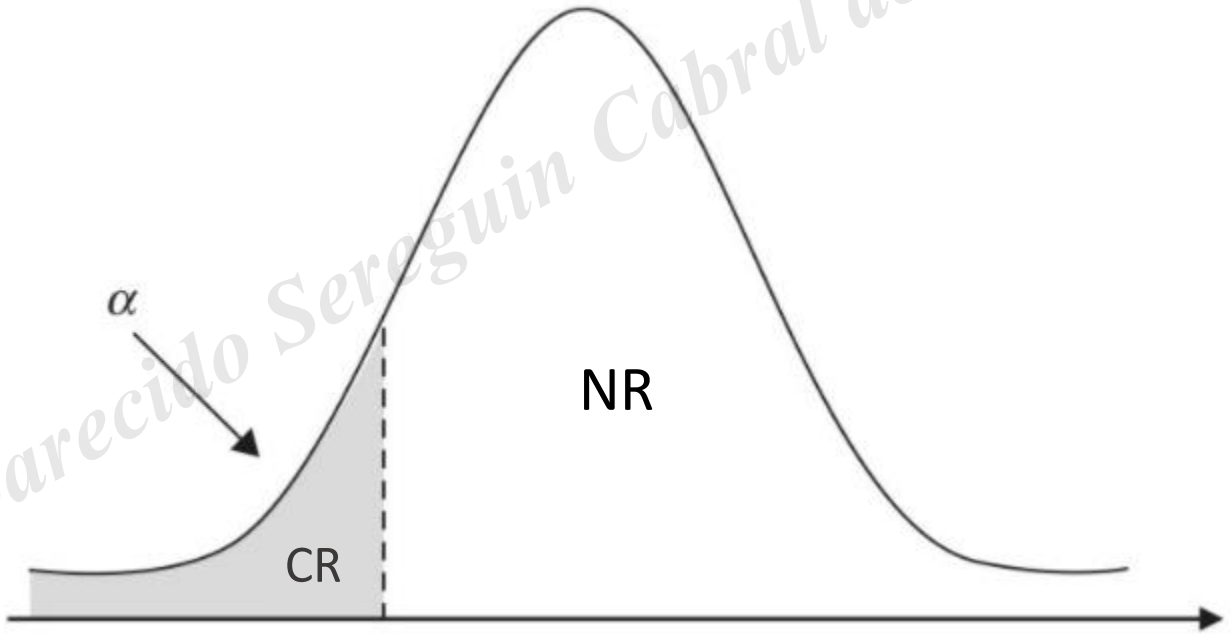

Source: Fávero e Belfiore (2017, Cap. 7)

# Tests types

- **Left tailed test (unilateral)**
  - In the left tailed test, for a θ parameter, the interest is to test:
    - $H_0$: $\theta = \theta_0$ (null hypothesis)
    - **$H_1$: $\theta < \theta_0$ (alternative hypothesis)**
  - The objective is to analyze if the parameter is **statistically lower** than a certain value of interest

# Tests types

- **Left tailed test (unilateral)**



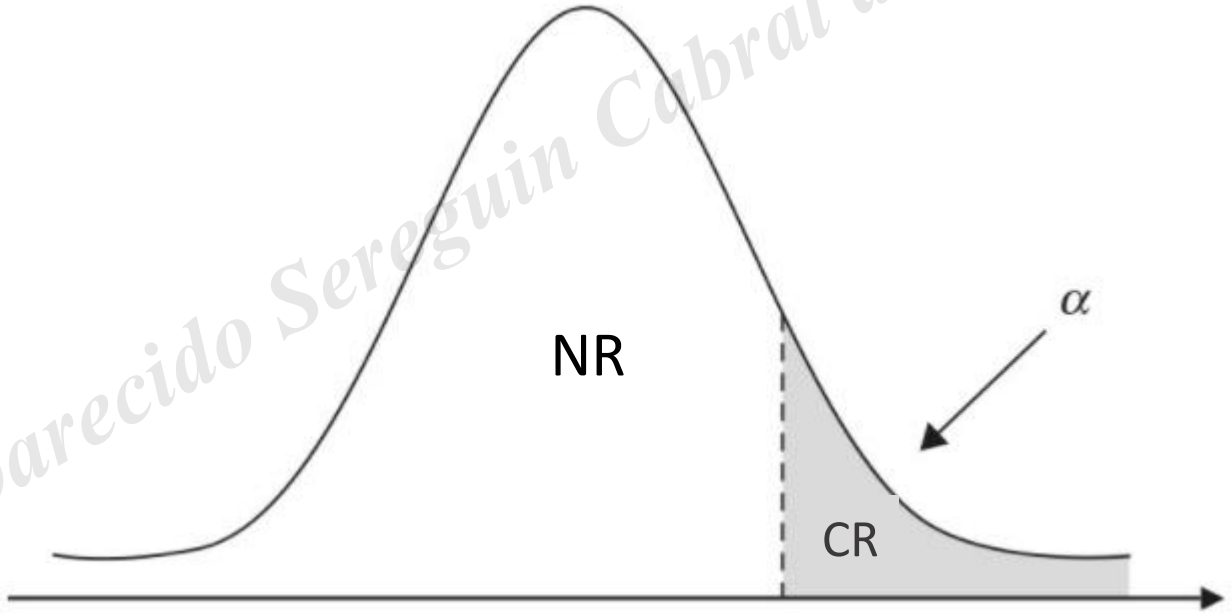

Source: Fávero e Belfiore (2017, Cap. 7)

# Tests types

- **Right tailed test**
  - In the right tailed test, for a θ parameter, the interest is to test:
    - $H_0$: $\theta = \theta_0$ (null hypothesis)
    - **$H_1$: $\theta > \theta_0$ (alternative hypothesis)**
  - The objective is to verify if the parameter is **statistically greater** than a certain value of interest

# Tests types

- **Right tailed test**




Source: Fávero e Belfiore (2017, Cap. 7)

# Types of errors

- **Possible errors in decision-making**
  - **Error type I:** rejecting null hypothesis ($H_0$) when it is true
  - **Error type II:** not rejecting null hypothesis ($H_0$) when it is false
- The correct decisions are:
  - Reject $H_0$ when it is false
  - Do not reject $H_0$ when it is true

# Test Significance

- **Test significance level**
  - The level of significance (α) indicates the probability of rejecting $H_0$ when it is true, that is, the probability of commiting the error type I
  - Some significance levels recurrently used:
    - α = 1%
    - α = 5%
    - α = 10%
  - The confidence level of the test is defined as 1 – α

# P-value and hypothesis test

- **P-value and significance level**

  - One way of testing statistically the hypothesis is comparing the value of calculated statistics in the data with the critical value for the significance level

  - It is also possible to obtain the p-value for the calculated statistics, then compare it to the significance level chosen

    - If p-value < significance level (α) it **is rejected $H_0$**
    - If p-value > significance level (α) it **is not rejected $H_0$**

  - **The p-value is the probability associated with the value of the calculated statistics**

# Hypothesis Tests

- **Z-Test for means of a sample**
  - It is applied when the population standard deviation is known and the distribution of the variable is normal (or using large samples)
  - The test statistics is:
  - $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
  - The relevant distribution for critical values is the standard normal

# Hypothesis Tests

- **Z-Test for means of a sample**

  - **Example:** A manufacturer of cardboard boxes wants to verify if the amount of cardboard that is used in each box of type 1 is according to its historical standard, since there are evidences that consumption has increased. Traditionally, it is used on average 100 g of cardboard in each box and the standard deviation is 12g. A sample was collected to verify if the current average is greater than the traditional average.

    - The sample collected is in the support spreadsheet in **the Z-test means tab**.

# Hypothesis Tests

- ***Z*-Test for means of a sample**
  - It is applied when the population standard deviation is not known, then it is used the sample standard deviation
  - Test statistics is similar to Z, but with a sample standard deviation:
  - $\boldsymbol{T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}}$
  - The relevant distribution is the *Student's t* with *n-1* degrees of freedom

# Hypothesis Tests

- ***Z*-Test for means of a sample**

  - **Example:** The average time of processing a certain task on a machine has been 18 minutes. New concepts were introduced to reduce the average time of processing. Therefore, after a certain period, a sample of 25 elements was collected, obtaining the mean time of 16.808 minutes with a standard deviation of 2.733 minutes. Check if this result highlights an improvement in the mean time of processing. Consider α = 1%. (Source: Fávero and Belfiore, 2017, Cap. 7)

    - The data are in the support spreadsheet in the T-**test means tab**.

# Hypothesis Tests

- **T-Test for correlations**
  - After the correlation coefficient (*r*) is estimated between quantitative variables, it is possible to test the significance of the estimated parameter
  - Test statistics:
  - $$t = \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}}$$
  - The relevant distribution is the *Student's t* with *n-2* degrees of freedom

# Hypothesis Tests

- ***T*-Test for correlations**

  - **Example:** Back to the example of the correlation between students's grades, now the objective is to assess if the correlations obtained for the grades samples are significant. The course coordinator used the significance level of 5% for their analyses.

    - The data are in the support spreadsheet in the **Pearson's Correlation tab**.

# Hypothesis Tests

- **Chi-squared test for a sample**
  - It is applied when the variable assumes two or more categories (K) and the objective is to verify if there are differences between the frequencies observed (O) and expected (E)
  - The test statistics is:
  - $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
  - The relevant distribution is the chi-squared with *k-1* degrees of freedom

# Hypothesis Tests

- **Chi-squared test for a sample**

  - **Example:** A store wants to verify if the amount sold in each day of the week varies according to the week day. The data for sales in each day of a chosen week randomly were tabulated. In this case, the objective is to test if the observed and expected frequency are equal or if they are different. (Source: Fávero and Belfiore, 2017, Cap. 8)

    - The tabulated data are in the support spreadsheet in the **Chi-squared Test A Sample tab**.

# Hypothesis Tests

- **F-Test for comparison of variances**
  - It is applied to compare the variances of two independent samples
  - The test statistics is:
  - $$F = \frac{S^2_{greater}}{S^2_{lower}}$$
  - The relevant distribution is the Snedecor's F, with *n-1* degrees of freedom in the numerator and *n-1* degrees of freedom in the denominator

# Hypothesis Tests

- **F-Test for comparison of variances**
  - **Example:** A logistic company is analyzing which routes are better predicting the delivery schedule for its largest customer. Delivery time data were collected during 35 days for each route. The logistic director wants to test the hypothesis that the route B has greater variability in the delivery time compared to the route A.
    - The data are in the support spreadsheet in the **F-Test Variances tab**.

# Confidence interval

- **Confidence interval for the mean**
  - When we obtain the estimate for the population mean from a sample, we can also build its confidence interval, that is, an interval of possible values for the population parameter
  - It is necessary to establish the confidence level of the analysis (for example, 95%)
  - $IC = (\overline{x} - Z \cdot \frac{s}{\sqrt{n}} \ , \ \overline{x} + Z \cdot \frac{s}{\sqrt{n}})$ **or** $IC = \left(\overline{x} - t \cdot \frac{s}{\sqrt{n}} \ , \ \overline{x} + t \cdot \frac{s}{\sqrt{n}}\right)$

**Large samples / Known variance**

**Small samples / Unknown Variance**

- **Z** and **t** are two-tailed values; in the **T** distribution it is used *n-1* degrees of freedom

# Confidence interval

- **Confidence interval for the mean**
  - **Example:** An engineer collected a sample of 25 parts from the assembly line and found that the mean size was 47cm and the standard deviation was 1cm. What is the confidence interval with 95% for this estimated mean?
    - The data are in the support spreadsheet in the **Confidence Interval – Mean tab**.

# Hypothesis Tests

- **T-Test for comparison of means in two independent samples**
  - In order to compare the means of two independent samples of the same population through t-test, it is necessary to compare population variances of the two groups
    - For example, before a F-test can be done for comparison of variances
  - The calculation of t-statistic and degrees of freedom depends on: if the population variances are different or if they are homogeneous

# Hypothesis Tests

- **T-Test for comparison of means in two independent samples**
  - T-statistics for different population variances
  - $T = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{S^2 1}{n_1} + \frac{S^2 2}{n_2}}}$
  - The degrees of freedom are $v = \frac{\left(\frac{S^2 1}{n_1} + \frac{S^2 2}{n_2}\right)^2}{\frac{\left(S^2 1 / n_1\right)^2}{(n_1 - 1)} + \frac{\left(S^2 2 / n_2\right)^2}{(n_2 - 1)}}$

# Hypothesis Tests

- **T-Test for comparison of means in two independent samples**
  - T-statistics for homogeneous population variances
  - $T = \frac{(\overline{X_1} - \overline{X_2})}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
  - In which $S_p = \sqrt{\frac{(n_1 - 1).S^2{}_1 + (n_2 - 1).S^2{}_2}{n_1 + n_2 - 2}}$ and the degrees of freedom are $n_1+n_2-2$

# Hypothesis Tests

- **T-Test for comparison of means in two independent samples**

  - **Example:** in an industry, the production manager did a survey with 30 measurements of temperature (in °C) of the two main furnaces of the production line, where the products of the same type are produced. Between these, 15 measurements were from the furnace A and 15 measurements were from the B one. The aim is to verify if the mean temperature is considerably different between the furnaces.

    - The data are in the support spreadsheet in the **T-Test Two Independent Samples tab**.

# Reference

Fávero, Luiz Paulo; Belfiore, Patrícia. (2017).

Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Rio de Janeiro: Elsevier