BIEM/F

A.Y. 2024/2025

BLAB

HANDOUTS

STATISTICS

-FIRST PARTIAL-

WRITTEN BY

MATILDE BALDINI

TEACHING DIVISION

This handout is written by students with no intention of replacing university materials.

It is a useful tool for studying the subject, but does not guarantee preparation as exhaustive and complete as the material recommended by the University.

STATISTICS

Decision-making processes under uncertainty require the use of the **statistical method**, a set of procedures necessary and functional for the analysis of data, aimed at extracting information of immediate practical value, potentially supporting various decision-making processes.

We use Statistics to evaluate scenarios, define company strategies or political maneuvers. Some examples of its application are:

- Identifying the purchasing habits of customers of a company.
- Establishing the most effective marketing strategy for a target audience.
- Forecasting the number of orders for the next year.

DATA COLLECTION AND PREPARATION

PRIMARY SOURCES

- Observations
- Surveys
- Experiments

SECONDARY SOURCES

- Processing of primary sources (paper or electronic format).

Data may refer to:

- A **population** (universe, N): The set of all units of interest in a study.
- A **sample** (subset, n): A fraction of elements selected within the population.

To ensure the representativeness of the samples, it is essential to draw them randomly. This way, the samples accurately reflect all units in the population and avoid biases towards specific groups of subjects.

In **Simple Random Sampling:**

- Each unit is randomly and independently selected as a sample unit.
- Each unit of the population has an equal chance of being chosen.
- All the possible samples of the same size **n** have the same chance of being chosen.

STATISTICAL UNIT AND DATA ORGANIZATION

- **Observation/Case/Statistical Unit**: The entity (population unit) for which information is collected.
- **Variable**: A characteristic of interest (of the cases).
- **Values/Levels**: Distinct values taken by the variable.
- **Data/Measurements**: Observations of the variable of interest measured on the cases considered.

Data is typically arranged into datasets:

- **Rows**: Represent cases.
- **Columns**: Represent variables.

CLASSIFICATION OF VARIABLES

Qualitative (categorical):

The levels are labels that indicate attributes, representing membership in groups or categories with specific characteristics (e.g., gender, region of birth, industry sector).

1. **Nominal**: Values cannot be ordered in any way ((e.g., eye colour).
2. **Ordinal**: Values can be ordered but without quantifying differences (e.g., level of education, level of satisfaction).

Quantitative (numerical):

The levels are numerical values (e.g., age, height, number of children, amount spent).

1. **Discrete**: Data generated by a counting process (integer numbers; e.g., number of children).
2. **Continuous**: Data generated by a measurement process (real numbers; e.g., amount spent).

PARAMETERS AND STATISTICS

In general, we are interested in measuring certain characteristics of a dataset; it is necessary to distinguish between:

- **Population Parameters** (its proportion is p): Measures that describe or summarize a characteristic of the population.
- **Sample Statistics** (its proportion is $\hat{p}$): Measures that describe or summarize a characteristic of a sample.

DESCRIPTIVE STATISTICS

Graphical and numerical methods for the synthesis and elaboration of data:

- Applicable to data relating to the entire population or a sample.
- Includes techniques for the preparation, synthesis, and presentation of data.
  - **Synthesis**: Mean, variance, correlation.
  - **Presentation**: Tables and graphs.

It allows us to obtain the Population Parameters and Sample Statistics.

INFERENTIAL STATISTICS

Methods that allow inference and predictions about population characteristics (Parameters) based on information extracted from sample data (Statistics).

The reliability of statistical analysis depends on the risk of using sample information to make inferences about the population.

It is necessary to consider the random mechanism (probability) and inherent randomness in the sampling process.

THE BASICS

The elementary commands are **expressions** or **assignments**.

R is **case-sensitive,** spaces are **NOT** allowed

`<-` is used for **assignments** (immediately saved but not displayed).

The **values** can be numerical, characters, logical (`TRUE`, `FALSE`).

Types of **errors**:

- `NaN` (Not a Number): Results from square root of negative numbers or 0/0.
- `Inf` or `-Inf`: The number is too large to be displayed.
- `NA` (Not Available): Missing value, in case of datasets.
- `NULL`: Empty set, without any content.

There are 3 classes of **operators**:

- **Arithmetic** (can only be applied to numerical values, the result is a numerical value):
  - `+` and `-`
  - `*` and `/`
  - `^` (raise to power)
- **Relational** (the result is a logical value):
  - `<` and `>`, `<=` and `>=` (applicable only to numerical values)
  - `==` (equal) and `!=` (different)
- **Logical** (the result is a logical value):
  - `!` (not)
  - `&` (and) and `|` (or)

The function `Args()` lists the arguments of a function.

When the arguments of a function are written in **named form** (each argument is specified with its name), their order is not important.

`?` followed by the name of a function provides help on the function.

DATA STRUCTURES

R's fundamental data structures are:

- **Vector**: A set of elements sharing a common data type.
  - Function `c()`: Concatenates an arbitrary number of elements (mostly vectors).
  - Function `rep()`: Generates a vector by replicating a value for a specified number of times.
  - Function `length()`: Returns the size of a vector (count of elements).
  - Function `names()`: Assigns or retrieves names to a vector's elements.
  - To **access** specific elements or **transform** selected elements it is possible to use the `[]` operator after the vector's name, specifying as arguments the positions of the elements to be selected, the names or eventual conditions.
- **Matrix**: A set of elements all of the same data type organized in rows and columns.
  - Function `matrix()`: Creates a matrix with elements of the same data type:
    - `x`: The vector containing the elements.
    - `nrow`: The number of rows.
      - Function `rownames()`: Gives access and possibility to modify the names of the rows
    - `ncol`: The number of columns.
      - Function `colnames()`: Gives access and possibility to modify the names of the columns
    - `byrow`: A logical value specifying whether the matrix should be filled by rows (`TRUE`) or columns (`FALSE`).
- **Factors**: Vectors that reassign the values of another vector, often of varying types, by associating each value with a specific level (or category).
  - Function `levels()`: Retrieves the levels associated with a factor and can be used to arrange the values of vectors.
  - Function `factors()`: Creates a factor starting from a vector (often associated with the `levels()` function to arrange the values). Used when the variable is ordinal.

```
name_variabile.F <- factor(name_variabile, levels = c("...", "...",
"..."))
```

- **Dataframe**: A set of columns (potentially) of different types (vectors of different types, and/or factors).
  - Function `dim()`: Tells the dimension of data frame (number of rows and columns).
  - `dataframe$column_name` is a way to access a single variable in a dataframe.
- **List**: A set of structures of any type, typically used to organize the outcomes of complex functions.

DESCRIPTIVE STATISTICS

UNIVARIATE STATISICS

Univariate statistics

1. **Frequency Distribution**: Tables and graphs.
2. **Summary Measures**: When appropriately selected, they allow to make comparisons easier and more immediate, and to communicate the largest amount of information in the simplest possible way.
   - **Central Tendency Measures**:
     - Mean.
     - Median.
     - Mode.
   - **Non-Central Tendency Measures**:
     - Quantiles.

- **Dispersion Measures**:
  - Range.
  - Interquartile range.
  - Variance.
  - Standard deviation.
  - Coefficient of variation.

3. **Shape**: Box plot.

FREQUENCY DISTRIBUTION

Data related to a variable can be collected on N cases of a population or on n cases of a sample.

Considering data collected on a sample of n units: $x_1, x_2, \ldots, x_n$.

Raw data are not easy to analyse and interpret, and it is necessary to organize them effectively to emphasize their salient information value. This can be done by using tables and charts, which are built taking into account:

- The type of data: qualitative (nominal or ordinal) or quantitative (discrete or continuous)
- The number K of distinct values (or categories, or levels) observed in data: $x_1^*, x_2^*, \ldots, x_n^*$.

QUALITATIVE VARIABLES

The data is organized into a **frequency distribution**, a table reporting:

1. **Levels**: Distinct categories for each variable.
2. **Absolute Frequency** (f_K, $total = n$): The number of cases for each category.
3. **Relative Frequency** (p_K, $total = 1$): The proportion of cases for each category relative to the total number of cases: $p_k = \frac{\text{Absolute Frequency}}{\text{Total Number of Cases}} = f_k/n$.

FREQUENCY TABLE

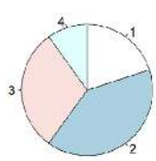
```
distr.table.x(x, freq = c("counts", "prop"), total = TRUE, data)
```

To construct a frequency table with the data:

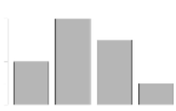
- `x`: Can be both a vector and a factor.
- `freq`: Specifies the type of frequency.
  - `"counts"`: Absolute frequencies.
  - `"prop"`: Relative frequencies.
  - `"perc"`: Percentages.
- `total`: If set to `TRUE`, includes the total count.
- `data`: The dataframe of reference.

The graphical representation of these variables can be through:

- **Pie Charts**: A circle divided into slices (categories) whose areas are proportional to the observed frequencies of the modalities. It provides information on the relative importance of each category (not about the order).
  - **Use**: Only for **nominal qualitative** variables.



- **Bar Charts**: A set of bars (categories) of equal width whose heights are proportional to the observed frequencies of the modalities.
  - **Use**: For nominal (categories don't have an order) and ordinal (categories have an order; factors) **qualitative** variables.



GRAPHICAL REPRESENTATION

```
distr.plot.x(x, freq = "counts", plot.type, bw = FALSE, data)
```

To create bar or pie charts:

- `x`: Can be both a vector and a factor.
- `plot.type`: Specifies the type of chart:
  - `"bars"`: Bar charts.
  - `"pie"`: Pie charts.
- `bw`: Sets graphs to grayscale (`TRUE`) or color (`FALSE`).
- `data`: The dataframe of reference.

QUANTITATIVE VARIABLES

For quantitative (numerical) variables, it is important to consider that:

- The **number of distinct values** taken by discrete variables can be small, as well as relatively large.
- **Continuous variables** typically take a different value for each observation (number of observed distinct values = total number of cases).

Frequency tables may face some challenges when it comes to quantitative variables:

1. **Discrete Variables**: Tables with many distinct values may lack clarity.
2. **Continuous Variables**: Each value may be unique. For continuous variables (or discrete variables taking a large number of distinct values), the frequency table becomes hard to interpret and ineffective (many rows/categories and many values are characterized by possibly low frequencies). In these cases it is convenient to classify data into **interval classes**:
   - Intervals should be contiguous and non-overlapping, and should cover the entire set of observed values.
   - The endpoints of the intervals should be clearly defined. By convention, the intervals' lower endpoint is included in the interval, while the upper endpoint is excluded (except for the final interval).
     1. **Lower and Upper Limits**: Define the boundaries of the interval.
     2. **Width** (w): The width of each interval class, can be the same or different. For intervals of equal width: $w = \frac{\text{Max} - \text{Min}}{\text{Number of Classes}}$. Unequal-width intervals are used when:
        - Data is concentrated in a narrow range.
        - Observations are sparse elsewhere.

     If the identified endpoints are consistent, the variable is correctly tabulated (with properly sorted intervals) and can be graphically represented using a histogram. However, if the identified endpoints are inconsistent, the table is still generated (with a warning message), but the histogram cannot be built due because of inconsistencies.

FREQUENCY TABLE

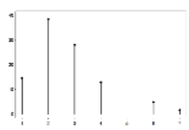
```
distr.table.x(x, freq = c("counts", "prop"), (interval OR breaks), f.digits, p.
digits, d.digits, total = TRUE, data)
```

To construct a frequency table with the data:

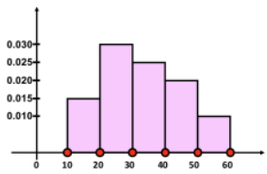
- `x`: Can be both a vector and a factor.
- `freq`: Specifies the type of frequency.
  - `"counts"`: Absolute frequencies.
  - `"prop"`: Relative frequencies.
  - `"perc"`: Percentages.
  - `"dens"`: Densities (histogram, if the classes have different width).
- `interval= T`: Indicates that the variable is measured in interval classes, because R cannot recognize the numerical ordering of the classes, and treats the intervals as characters otherwise.
- `breaks`: Specifies the number or size of the intervals (histogram).
  - If it is a single number, it indicates the number of intervals with equal width to use → the use of the frequency density, relative frequency, or absolute frequency doesn't change the way the graph looks.
  - If it is a vector of increasing values (`breaks=c(...,...,...)`), it specifies the intervals' endpoints (min and max value must be included between the first and last breaks).
- `f.digits`, `p.digits`, `d.digits`: Used to specify the number of decimals for relative frequencies, percentages, and densities.
- `total`: If set to `TRUE`, includes the total count.
- `data`: The dataframe of reference.

When it comes to **visually representing** quantitative variables, we use:

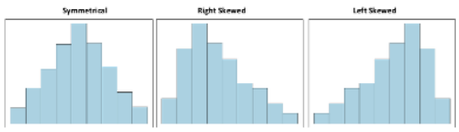
- **Spike (or Stick) Plots**: Takes into account both the observed values and the distances among them. Each observed value, on the horizontal axis, is associated a vertical stick whose height corresponds to its frequency.
  - **Use**: Only for **quantitative discrete** variables.



- **Histograms**: A diagram with adjacent rectangles, where each interval class is represented by a rectangle(base= w_k; area= p_k) that shows its frequency density (height= $c_k = p_k/w_k$) .
  - **Use**: For both discrete and continuous **quantitative** variables.



  - The choice of the number of intervals (5-30) depends both on the number of observations and on the specific features of the variable. There is no optimal criterion for determining the number of classes.
    - Highlight distinctive features and disparities within the data, avoiding an excessive number of intervals with low frequency or an insufficient number of classes with very high frequency.
    - Effectively depict the distribution's shape (symmetrical or asymmetrical), and the data's spread (presence of any "tails").
  - Shape of the distribution:
    - **Symmetrical Distribution**: Data is evenly distributed around the center.
    - **Asymmetrical Distribution**: Have a tail that extends primarily in one direction.
      - **Right-Skewed**: The tail extends toward higher values.
      - **Left-Skewed**: The tail extends toward lower values.



All distributions, both symmetrical and skewed, can have particularly long tails.

- **Width of the Intervals**: Frequently, most of the data tend to cluster within a relatively narrow range of values, whereas few values are spread out across a much larger range, resulting in long tails in the distribution. To simplify the representation of the frequency distribution in these situations, it is sensible to use intervals with different widths. Due to the different widths, the histogram must be built using the frequency densities.

GRAPHICAL REPRESENTATION

```
distr.plot.x(x, freq, plot.type, (interval OR breaks), data)
```

To create spike plots or histograms:

- `x`: Can be both a vector and a factor.
- `freq`: Specifies the type of frequency.
  - `"counts"`: Absolute frequencies.
  - `"prop"`: Relative frequencies.
  - `"perc"`: Percentages.
  - `"dens"`: Densities (histogram).
- `plot.type`: Specifies the type of chart:
  - `"spike"`: Spike charts.
  - `"hist"`: Pie charts.
- `interval= T`: Indicates that the variable is measured in interval classes, because R cannot recognize the numerical ordering of the classes, and treats the intervals as characters otherwise.
- `breaks`: Specifies the number or size of the intervals (histogram).
  - If it is a single number, it indicates the number of intervals with equal width to use → the use of the frequency density, relative frequency, or absolute frequency doesn't change the way the graph looks.
  - If it is a vector of increasing values (`breaks=c(...,...,...)`), it specifies the intervals' endpoints (min and max value must be included between the first and last breaks)
- `data`: The dataframe of reference.

CUMULATIVE FREQUENCIES

A cumulative frequency distribution summarizes the running total of frequencies up to each modality or interval. The cumulative frequency corresponding to a value (or to an interval for variables measured in classes) is the sum of all the frequencies associated to values (or classes) lower than or equal to the value (or class) itself.

We usually refer to the distribution of cumulative relative frequencies: $F_k = p_1 + p_2 + ... + p_k$.

However, also absolute frequencies or percentage can be cumulated. In fact, for a numerical variable it is possible to calculate the frequency cumulated at each real number.

The cumulative frequency function or distribution function, associates to each real number x the frequency of observed values lower than or equal to it: $F(x) = Freq(X \leq x)$.

FREQUENCY TABLE

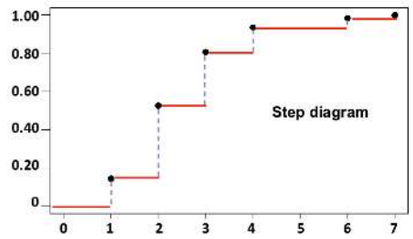
```
distr.table.x(x, freq = c("counts", "prop","cum"), total = TRUE, data)
```

To construct a frequency table with the data:

- `x`: Can be both a vector and a factor.
- `freq`: Specifies the type of frequency.
  - `"counts"`: Absolute frequencies.
  - `"prop"`: Relative frequencies.
  - `"perc"`: Percentages.
  - `"dens"`: Densities.
  - `"cum"`: Cumulative frequencies.
- `total`: If set to `TRUE`, includes the total count.
- `data`: The dataframe of reference.

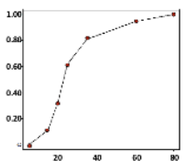
Some graphical representations for this kind of frequency measure are:

- **Step Diagrams**: The values between two units are not observed and, for this reason, they do not contribute to the cumulative frequency.
  - **Use**: ordinal qualitative variables (if the levels are not recognised as ordinal, a factor needs to be used) and discrete quantitative variables.
  - The values smaller than the minimum have $F_k = 0$, while the values greater than the maximum have $F_k = 1$.



- **Ogives/Cumulative Frequency Curves**: A graph based on a data classified into intervals. The ogive is a line connecting the frequencies (or percentages) cumulated at the upper endpoints of the interval classes. The ogive is built assuming that the cumulative frequency increases at a constant rate within each interval.

  It is obtained by smoothing out the step diagram, starting from raw data. If raw data is not available, the function can only be approximated by assuming that values are uniformly distributed within the classes (considering $x \in [a,b) x \in [a,b) F(x) = Freq(X \leq x) = F_k + (x-a)c_k$).
  - **Use**: quantitative variables (usually continuous).



GRAPHICAL REPRESENTATION

```
distr.plot.x(x, freq="prop", plot.type="cum", (interval=T or breaks), data)
```

To create step diagrams or ogives:

- `x`: Can be both a vector and a factor.
- `freq`: Specifies the type of frequency.
  - `"prop"`: Relative frequencies.
- `plot.type`: Specifies the type of chart.
- `interval= T or breaks`: Indicates that the variable is measured in interval classes, results in an ogive, rather than a step diagram.
- `data`: The dataframe of reference.

MEASURES OF CENTRAL TENDENCY

A central tendency measure is a single value that summarises all the observed data:

- Aims at describing the “centre” of the data.
- There are different measures, depending on how we can or how we decide to define the “centre” of the data:
  - Mode.
  - Median.
  - Mean.

MODE

The Mode is the value (or category) most frequently observed in a set of data.

The “centre” of data is taken as the value describing the most typical “behaviour” of cases with respect to a variable.

The mode can be calculated for all the types of data, both qualitative and quantitative. However, it might be not particularly effective when a variable takes a very large number of distinct values (usually continuous quantitative variables).

The mode may not be unique (two or more variables with the same frequency), it can be weak (rather small relative frequency), and in some cases it might not exist (same relative frequency for every variable). For continuous variables the modal class is used: data is classified into intervals, and it is the interval with the highest frequency density (concentration of relative frequency, depends on the classification used).

MEDIAN

The median is the value with the middle position in the ordered sequence of data.

The “centre” is taken as the value ideally dividing the data into two blocks of the same size. It therefore separates the higher half of values from the lower half of values.

Being based only upon the position of the order data and not on the specifically observed values, the median can be calculated both for ordinal qualitative and for quantitative data. However, it cannot be calculated for nominal variables.

The median divides ordered data into two halves:

1. If n is odd, the median is the middle value.
2. If n is even, the median is the average of the two middle values.

When starting from raw data, the median corresponds to the first value with a cumulative frequency higher than or equal to 0.5 (if equal, mean between that value and the next one).

For grouped data, the median can only be approximated under the assumption that the frequency associated with each interval is uniformly distributed within the interval, starting from the median class. The formula to be used is $F_{k-1} + (Me - x_k)c_k = 0.5$, where

- F_{k-1} is the value o the cumulative frequency of the class before.
- Me is the value of the median to be found.
- x_k is the value of the lower boundary of the median class.
- c_k is the frequency density of the median class.

MEDIAN

```
median(x, na.rm=T)
```

To get the central tendency measures for a variable in a dataset:

- `x`: Can be both a vector and a factor.
- `na.rm`: Stands for “NA remove”, it is a logical value that makes R ignore missing values.

MEAN

The (arithmetic) mean is defined as the sum of all data on a variable divided by the total number of cases: $\bar{x} = \sum \frac{x_i}{n}$. We use:

- $\bar{x}$ when the mean is computed on a sample (statistic).
- μ when the mean is computed on the population (parameter).

The mean can only be calculated for numerical variables.

The “centre” is taken as the value such that the sum of the negative deviations from the mean coincides with the sum of the positive deviations, resulting in a total deviation sum of 0 (centre of gravity of the observed data): $\sum(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = \sum x_i - \sum x_i = 0$.

The mean is a non-robust statistic, which means that it is particularly sensitive to extreme values/outliers (unlike the median). The mean, since it is more **sensible,** is useful to **compare** different years and understand **trends.**

For discrete variables, the mean can be precisely calculated based on the observed distinct values and their relative or absolute frequencies: $\bar{x} = \sum \frac{x_i}{n} = \sum \frac{x_k^* f_k}{n} = \sum x_k^* \frac{f_k}{n} = \sum x_k^* p_k$.

However, when a variable is measured in classes and only the frequency table of the classified variable is available, the mean cannot be calculated precisely, because the original raw numeric data is not available. In this situation, the mean can only be approximated, under the assumption that the frequency associated with each interval is uniformly distributed over the interval: $\bar{x} = \sum \frac{m_k f_k}{n}$, where:

- m_k is the midpoint of each class, used to approximate the mean.
- f_k is the relative frequency of each class.

MEAN

```
mean(x, na.rm=T)
```

To get the central tendency measures for a variable in a dataset:

- `x`: Can be both a vector and a factor.
- `na.rm`: Stands for “NA remove”, it is a logical value that makes R ignore missing values.

SUMMARY STATISTICS

```
distr.summary.x(x, stats="central", digits=2, f.digits=4, data)
```

To get the central tendency measures for a variable in a dataset:

- `x`: Can be both a vector and a factor.
- `stats`: Specifies the type of statistic.
  - `"central"`: Central tendency measures.
  - `c("mode", "median", "mean")`: Central tendency measures.
- `digits`: Number of decimals for the statistics.
- `f.digits`: Number of decimals for the frequency.
- `data`: The dataframe of reference.

This function treats the variables divided into classes like qualitative variables, only computing the mode (by using the maximum relative frequency, rather than to the frequency density).

SYMMETRICAL AND ASYMMETRICAL DISTRIBUTIONS

The shape of the distribution, the existence of outliers and possible asymmetries, lead to differences between mean and median that are due to their different robustness:

- If the distribution is symmetrical, mean and median are almost equal.
- If the distribution is asymmetrical, we have two possible cases:
  - **Positively Skewed**: The mean is greater than the median.
  - **Negatively Skewed**: The mean is less than the median.

The main consequence of this is that, the more the values are concentrated around the centre, the more the central tendency measures are representative.

NON-CENTRAL TENDENCY MEASURES

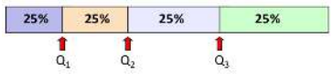
If a distribution is strongly skewed and/or has long tails, limiting attention to central tendency measures might lead to loss of information and to a partial or inaccurate description of the distribution's characteristics. In this case, it is important to provide summaries describing the behaviour of the distribution around or far from its centre and with this purpose we use the non-central tendency measures (quantiles):

- Quartiles.
- Quintiles.
- Deciles.
- Percentiles.

QUARTILES

Quartiles divide the ordered sequence of data into 4 blocks with (possibly) the same number of cases (25% each):

- The first quartile, Q_1 (or P25) separates the smallest 25% of the data from the remaining 75%, of values greater than Q_1.
- The second quartile, Q_2 (or P50) coincides with the median and divides the data into two blocks containing 50% of the values each.
- The third quartile, Q_3 (or P75) separates the highest 25% of data from the remaining 75% of values smaller than Q_3.



In some cases it is not possible to find values that divide the distribution exactly, and quartiles may lie between two values.

In general, in the case of discrete or ordinal variables (i.e. with few distinct values), as well as in the case of manual calculation, we will identify the three quartiles as the values at which the cumulative frequency equals or exceeds 0.25, 0.5 and 0.75 for the first time. In the case of numerical values, this approach returns quartiles possibly different from those obtained using R, which approximates the quartile values through appropriate interpolation.

Just like for the median, Q_1 and Q_3 can be approximated by considering cumulated frequencies in interval classes:

- Q_1: $F_{k-1} + (Q_1 - x_k)c_k = 0.25.$
- Median= Q_2: $F_{k-1} + (Q_2 - x_k)c_k = 0.5$
- Q_3: $F_{k-1} + (Q_3 - x_k)c_k = 0.75$

QUARTILES

```
distr.summary.x(x, stats = "quartiles", digits=2, f.digits=4, data)
```

To get the quartiles for a variable in a dataset:

- `x`: Can be both a vector and a factor.
- `stats`: Specifies the type of statistic.
  - `"quartiles"`: Quartile measures.
- `digits`: Number of decimals for the statistics.
- `f.digits`: Number of decimals for the frequency.
- `data`: The dataframe of reference.

PERCENTILES

Quartiles and central tendency measures describe **50% of data concentrated around the center,** to get info on the **tails** we need **percentiles**. Percentiles are useful when we have **long tails** and are calculated on **raw data,** independently from the analyst's choice.

Percentiles divide data into 100 groups with, possibly, the same number of data.

We denote by P_q the q^{th} percentile, which separates the $q\%$ of the smallest data from the remaining $(100 - q)\%$ (minimum value of the top $(100 - q)\%$).

PERCENTILES

```
distr.summary.x(x, stats = "percentiles", digits=2, f.digits=4, data)
```

To get the percentiles for a variable in a dataset:

- `x`: Can be both a vector and a factor.
- `stats`: Specifies the type of statistic.
  - `"percentiles"`: Percentile measures.
  - `c( "p1" , "…" , "p50" , "…" , "p99" )`: Percentile measures.
- `digits`: Number of decimals for the statistics.
- `f.digits`: Number of decimals for the frequency.
- `data`: The dataframe of reference.

BOX PLOT

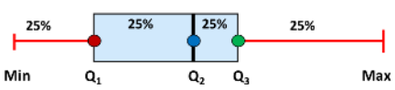
The box plot (or box-and-whiskers plot) provides a univocal, effective and schematic representation of the distribution of a numerical variable, without resorting to a histogram.

The simplest box plot is based on the five summary numbers:

- The minimum observed value in the data.
- The three quartiles.
- The maximum observed value.

The plot is based on a box and two whiskers.

- The box extends from the first to the third quartile (includes 50% of data), and is divided by the median.
- The whiskers connect the box to the lowest and highest observed value respectively.



The box plot summarises the salient features of the distribution: the centre and the local and global dispersion around it.

Being based on summary measures (quartiles, minimum and maximum values), the box plot does not change as the specific interval classes used to build the histogram vary. It is impossible to obtain the boxplot for variables measured in classes.

BOXPLOT

```
distr.plot.x(x, plot.type="boxplot", data)
```

To create a boxplot graph:

- `x`: Can be both a vector and a factor.
- `plot.type`: Specifies the type of chart.
- `data`: The dataframe of reference.

A more detailed version of the box plot also allows to identify and visualize extreme values. The differences here can be found in the whiskers:

- Values deviating from the box more than 1.5 times its length (interquartile range) are flagged as outliers and are identified by a specific symbol in the plot.
  - $\text{IQR} = Q_3 - Q_1$
- Whiskers connect the box to minimum and maximum regular values, that do not deviate from the box more than 1.5 times its length:
  - Lower tail: $Q_1 - 1.5 \cdot \text{IQR}$.

MINIMUM REGULAR VALUE

```
min(data$variable [data$variable <= lower_tail])
```

OUTLIERS (LOWER TAIL)

```
data$variable [data$variable < lower_tail]
```

- Upper tail: $Q_3 + 1.5 \cdot \mathrm{IQR}$

MAXIMUM REGULAR VALUE

```
max(data$variable [data$variable >= upper_tail])
```

OUTLIERS (UPPER TAIL)

```
data$variable [data$variable > upper_tail]
```

DISPERSION MEASURES

Measures of dispersion (or variability) quantify and summarise the amount of dispersion in a dataset. They complement measures of central tendency to provide a fuller understanding of data distribution. These measures are:

1. **Range**
2. **Interquartile Range (IQR)**
3. **Standard Deviation**
4. **Variance**
5. **Coefficient of Variation**

RANGE

The range (/range of variation) is the difference between the highest and lowest observed values in a dataset. This measure is not robust (highly sensitive to outliers).

$$\mathrm{Range} = \mathrm{Max}(x) - \mathrm{Min}(x)$$

The range is rather simple to compute but does not provide insight into the distribution of values.

INTERQUARTILE RANGE (IQR)

The interquartile range is the difference between the third quartile and the first quartile, containing the central 50% of the data: $\mathrm{IQR} = Q_3 - Q_1$

VARIANCE

This measure excludes the smallest 25% and the largest 25% of the data, focusing on the central portion, and is robust.

The variance is a measure of the dispersion of data around their mean.

The deviation of x_i (value for the i^{th} observation) from the sample mean ($\bar{x}$) is $(x_i - \bar{x})$, and it also quantifies the error that would occur if x_i was replaced (or estimated) by the sample mean, considered as "representative" of the observed data.

The **sum of all deviations is 0** and we are interested in the **magnitude** rather than in the direction. The dispersion around the mean can be summarized by the mean of the squared deviations from the mean, called variance:

- **Population Variance**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$.

- N: Population size.
- x_i: Observed value for the i^{th} case.
- μ: Population mean.

- **Sample Variance**: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
  - n: Sample size.
  - x_i: Observed value for the i^{th} case.
  - $\bar{x}$: Sample mean.

  It is not exactly the mean of the squared deviations, as the sum is divided by $(n-1)$ instead of n. This compensates for the bias introduced when using the sample mean instead of the population mean.

The variance considers all observed values and, as it is the mean of the squared deviations from the mean (which is non robust and sensitive to extreme values), it is also **sensitive to extreme values** and **non robust**.

This provides a measure of reliability for the mean as a summary of data, since it can be interpreted as the "average" squared error incurred when replacing the raw data with their mean.

The sample variance can be calculated with an indirect formula from the **mean** of the **squared data and the square of the mean.** This is especially useful when the variance is calculated based on grouped data.

Sample Variance (Indirect Formula)

$$s^2 = \frac{n}{n-1}[\sum_{i=1}^{n}(\frac{x_i^2}{n}) - \bar{x}^2]$$

Population Variance (Indirect Formula)

$$\sigma^2 = \sum_{i=1}^{N}(\frac{x_i^2}{N}) - \mu^2$$

STANDARD DEVIATION

The standard deviation can be interpreted as a measure of the **average** (standard) **distance** (absolute deviation) **of the data from the mean.** It is the **square root of the variance.** When variance increases, the standard deviation and the dispersion of data increase as well.

- **Sample Standard Deviation**: $s = \sqrt{s^2}$.
- **Population Standard Deviation**: $\sigma = \sqrt{\sigma^2}$

It is possible to compute both the **variance** and the **standard deviation** from the frequency table: since the **mean** can be calculated as the sum of the observed distinct values wighted by their relative frequency ($\bar{x} = \sum x_k \cdot p_k$), and the **mean of squared values** can be obtained easily ($\bar{x}^2 = \sum x_k^2 \cdot p_k$), the **indirect formula** can be used to compute both s^2 and s.

In case of data grouped in classes, we **discretize** it using **midpoints** (m_k) and we obtain an **approximated** value: $\bar{x} = \sum m_k \cdot p_k$, and $\bar{x}^2 = \sum m_k^2 \cdot p_k$.

Variance and standard deviation both have a unit of measurement, therefore they are **absolute** measures of dispersion.

Standard deviation is easier to interpret than variance due to its units and, like variance, it is not robust to extreme values. Variance and standard deviation measure the spread of data around the mean but may overemphasize large deviations due to squaring.

COEFFICIENT OF VARIATION (CV)

The coefficient of variation measures the **amount of dispersion in the data relative to their mean.** It is **a-dimensional** and should be used to compare the **dispersion** of data with **different units of measurements.**

It does not have a well-defined range, so it is not possible to draw conclusion on dispersion based solely on this measure (it does not provide an absolute measure of dispersion for a single distribution).

$$\mathrm{CV} = \frac{s}{|\bar{x}|}, \quad \bar{x} \neq 0$$

DISPERSION MEASURES

```
distr.summary.x(x, stats = "dispersion", data)
```

To get the dispersion measures for a variable in a dataset:

- `x`: Can be both a vector and a factor.
- `stats`: Specifies the type of statistic.
  - `"dispersion"`: Dispersion measures.
  - `c( "range" , "IQrange" , "sd" , "var" , "cv" )`: Dispersion measures.
  - `"range"`: Range.
  - `"IQrange"`: Interquartile range.
  - `"sd"`: Standard deviation.
  - `"var"`: Variance.
  - `"cv"`: Coefficient of variation.
- `data`: The dataframe of reference.

BIVARIATE STATISTICS

In a variety of applications, there is interest in the study of two variables and their relationships. The methods for organizing and representing the data depend on the types of variables being examined:

- **Both qualitative:** Generally with a limited number of categories.
- **Qualitative-quantitative:** Typically one with few categories and the other continuous.
- **Both quantitative:** Typically both continuous.

QUALITATIVE VARIABLES

Qualitative (or discrete) variables generally have a limited number of categories. To organise the data in such cases, the joint frequency distribution is considered, which describes:

- The (distinct) pairs of categories observed in the two variables.
- The relevance (frequency or percentage) of each pair of categories.
  - **Absolute joint frequencies:** Number of cases presenting each pair of categories.
  - **Relative joint frequencies:** Proportion of cases for each pair of categories relative to the total (absolute joint frequencies/total number of cases), with the corresponding percentages.

The distribution of joint frequencies is organized into a two-way contingency (or cross) table where the cells report the absolute or relative joint frequencies for each pair of categories (including potential totals).

Distinct values taken by Y

Distinct values taken by X	$X \backslash Y$	y_1^*	y_2^*	...	y_J^*	Total
	x_1^*	f_{11}	f_{12}	...	f_{1J}	R_1
	x_2^*	f_{21}	f_{22}	...	f_{2J}	R_2
	...	...	...	...	...	...
	x_K^*	f_{K1}	f_{K2}	...	f_{KJ}	R_K
	Total	C_1	C_2	...	C_J	n

- X: Variable represented in rows.
- Y: Variable represented in columns.
- x/y: Categories of variables X/Y.
- f_{kj}: Absolute/relative frequencies of each pair.
- R_k/C_j: Marginal frequency distribution of X/Y. If there are no missing data, these coincide with the univariate distribution of the variables.

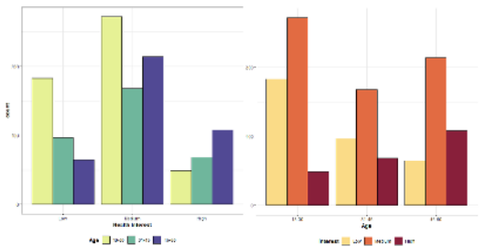
For graphical representation of a two-way table, bar charts are generally used:

- **Side-by-side bar diagrams**.
- **Stacked bar diagrams**.

SIDE-BY-SIDE BAR DIAGRAM

For each category of one variable, a set of bars of equal width is displayed, one for each category of the second variable.

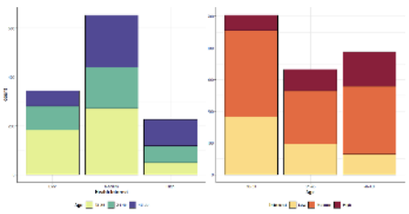
- The heights of the bars represent the joint frequencies.
- The appearance does not change whether absolute or relative joint frequencies are used.



STACKED BAR DIAGRAM

For each category of one variable, a single bar is constructed by stacking multiple segments, one for each category of the second variable.

- The heights of the segments represent the joint frequencies.
- The appearance does not change whether absolute or relative joint frequencies are used.



CONDITIONAL FREQUENCY DISTRIBUTION

Graphs can be built based on either variable, albeit differently. When two variables are crossed, one is always prioritized over the other. Such tools may not always effectively highlight the relevance of one variable's values given the other.

For more accurate analysis, comparing **conditional frequency distributions** is useful, i.e., distributions of one variable within subsets of cases defined by the different categories of the other variable.

Conditional distribution of Y given $X = x_k$:

$Y \mid X = x_k$

$\text{Freq}(Y = y_j^* \mid X = x_k^*) = f_{kj}/R_k, \quad \text{for } j = 1, 2, \ldots, J.$

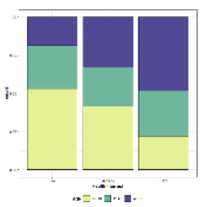
Conditional distribution of X given $Y = y_j$:

$X \mid Y = y_j$

$\text{Freq}(X = x_k^* \mid Y = y_j^*) = f_{kj}/C_j, \quad \text{for } k = 1, 2, \ldots, K$

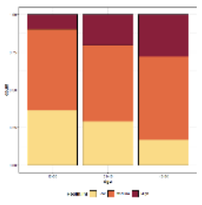
Distinct values taken by Y

Distinct values taken by X	$X \backslash Y$	y_1^*	y_2^*	...	y_J^*	Total
	x_1^*	f_{11}	f_{12}	...	f_{1J}	R_1
	x_2^*	f_{21}	f_{22}	...	f_{2J}	R_2
	...	...	...	...	...	...
	x_K^*	f_{K1}	f_{K2}	...	f_{KJ}	R_K



Distinct values taken by Y

Distinct values taken by X	$X \backslash Y$	y_1^*	y_2^*	...	y_J^*
	x_1^*	f_{11}	f_{12}	...	f_{1J}
	x_2^*	f_{21}	f_{22}	...	f_{2J}
	...	...	...	...	...
	x_K^*	f_{K1}	f_{K2}	...	f_{KJ}
	Total	C_1	C_2	...	C_J



FREQUENCY TABLE

```
distr.table.xy(x, y, freq=c("counts"), freq.type=c("joint"), total=TRUE, data)
```

To calculate joint and/or conditional distributions:

- `x`: Variable for rows.
- `y`: Variable for columns.
- `freq`: Specifies the type of frequencies in the table:
  - `"counts"`: Absolute frequencies.
  - `"prop"`: Proportional frequencies.
  - `"perc"`: Percentage frequencies.
- `freq.type`: Specifies the type of frequency distribution:
  - `"joint"`: Joint frequencies.
  - `"row"`: Conditional frequencies by row ($y \mid x$).
  - `"column"`: Conditional frequencies by column ($x \mid y$).
- `total`: Indicates whether totals should be included in the table (`TRUE` or `FALSE`).
- `data`: The name of the dataframe containing the variables.

For **numeric variables** to be classified into intervals: `breaks.x` and/or `breaks.y`.

For **interval class variables**: `interval.x = TRUE` and/or `interval.y = TRUE`.

GRAPHICAL REPRESENTATION

```
distr.plot.xy(x, y, freq="counts", freq.type="joint", plot.type="bars", bar.typ
e="stacked", data)
```

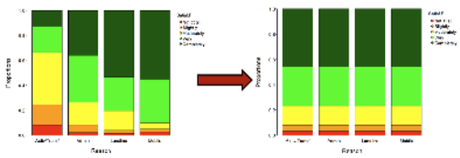
To create graphical representations (bar charts) for joint and/or conditional distributions:

- `x`: Variable for the horizontal axis.
- `y`: Variable for the vertical axis.
- `freq`: Specifies the frequencies displayed:
  - `"counts"`: Absolute frequencies.
  - `"prop"`: Proportional frequencies.
  - `"perc"`: Percentage frequencies.
- `freq.type`: Specifies the frequency type:
  - `"joint"`: Joint frequencies.
  - `"row"`: Conditional frequencies by row ($y \mid x$).
  - `"column"`: Conditional frequencies by column ($x \mid y$).
- `plot.type`: Type of chart to create.
- `bar.type`: Type of bar chart:
  - `"stacked"`: Stacked bar chart.
  - `"beside"`: Side-by-side bar chart.
- `data`: The name of the dataframe containing the variables.

For **numeric variables** to be classified into intervals: `breaks.x` and/or `breaks.y`.

For **interval class variables**: `interval.x = TRUE` and/or `interval.y = TRUE`.

If the two variables are independent, **conditional distributions** should match the **marginal distributions**. This means that the distribution of one variable should not change regardless of the categories of the other variable.



QUALITATIVE-QUANTITATIVE VARIABLES

Continuous variables typically have unique values for each observation, while discrete variables take on a limited or high number of values (e.g., age in years, number of transactions).

When analyzing two variables where at least one is continuous or discrete with many categories we can use:

1. **Two-way Tables**: May not facilitate analysis due to the large number of categories.
2. **Histograms**: A first approach is to classify the variable with many categories into **intervals** (two-way table) and compare **histograms** representing the distribution for each category of the other variable. This method becomes problematic if:
   - Differences between conditional distributions are not pronounced.
   - The conditioning variable has many categories.

SIDE-BY-SIDE BOXPLOTS

The most effective representation for comparing conditional distributions is **side-by-side boxplots**.

For each category of the qualitative variable, a boxplot is displayed to highlight the key features of the distribution of the numerical variable.

To compare distribution characteristics, summary measures of the numerical variable (y) conditioned on the qualitative variable (x) can be used.

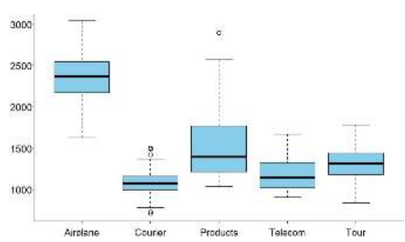
GRAPHICAL REPRESENTATION

```
distr.plot.xy(x, y, plot.type="boxplot", data)
```

To generate side-by-side boxplots:

- `x`: Qualitative variable.
- `y`: Numerical variable.
- `plot.type`: Set to `"boxplot"`.
- `data`: The name of the dataframe.

This function only works for numerical variables and does not support interval classification.



SUMMARY STATISTICS

SUMMARY STATISTICS

```
distr.summary.x(x, by1, by2, stats, digits=2, f.digits=4, data)
```

To calculates summary measures for a numerical variable conditioned on up to two variables:

- `x`: Numerical variable for summary statistics.
- `by1` / `by2`: Conditioning variables (up to 2).
- `stats`: Specifies the summary statistics to calculate. `"summary"` is for all:
  - Minimum ($\min$)
  - First Quartile (q_1)
  - Median (median or q_2)
  - Mean (mean)
  - Third Quartile (q_3)
  - Maximum ($\max$)
  - Standard Deviation (sd)
  - Variance (var)
- `digits`: Number of decimal places for output.
- `f.digits`: Decimal places for frequency outputs.
- `data`: The name of the dataframe.

To classify **continuous conditioning variables** into intervals, use: `breaks.by1` and/or `breaks.by2`.

For variables already classified into **interval classes**, set: `interval.by1 = TRUE` and/or `interval.by2 = TRUE`.

QUANTITATIVE VARIABLES

When analyzing two continuous numerical variables, **classifying them into intervals** in a two-way table excessively compresses the data. Instead, a **scatterplot** is typically used.

SCATTERPLOT

A scatterplot is a graph where each observation is represented by a point in a plane, with coordinates corresponding to the values observed for the two variables, plotted on the horizontal and vertical axes.

The scatterplot allows you to:

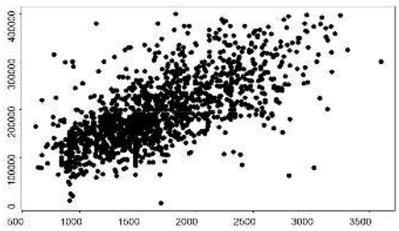
1. Visualize the **joint distribution** of the variables.
2. Identify possible **relationships** between the variables.
3. Detect **outliers** or anomalous values.

GRAPHICAL REPRESENTATION

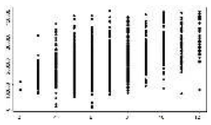
```
distr.plot.xy(x, y, plot.type="scatter", data)
```

To generate a scatterplot:

- `x`: Variable for the horizontal axis.
- `y`: Variable for the vertical axis.
- `plot.type`: Set to `"scatter"` to produce a scatterplot.
- `data`: The dataframe containing the variables.



Even if one of the two variables is **discrete**, a scatterplot remains useful because each observation is represented as a unique pair of values.

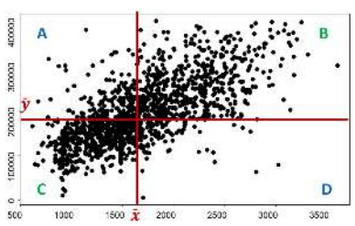


COVARIANCE

The strength of the relationship between variables decreases as the data becomes more **dispersed** along the y-axis. From observing the scatterplot configurations, it becomes clear that criteria are needed to **quantify the intensity** of the relationship between the two variables.

Scatterplots can be analyzed based on the **means** of the two variables, dividing the plot into four quadrants:

- **Quadrants A and D (Discordant Pairs)**: Observations where one variable has values **greater** than the mean while the other has values **less** than the mean, and vice versa.
- **Quadrants B and C (Concordant Pairs)**: Observations where both variables have values either **greater** or **less** than their respective means.



There ere two different types of **association**:

1. **Positive/Direct Association**:
   - A prevalence of **concordant pairs** (positive cross product, $(x_i - \bar{x})(y_i - \bar{y}) > 0$): positive product of deviations, both values are either greater or smaller than their respective means.
   - As one variable increases, the other tends to increase.
2. **Negative/Inverse Association**:
   - A prevalence of **discordant pairs** (negative cross product, $(x_i - \bar{x})(y_i - \bar{y}) < 0$): negative product of deviations, one value is greater while the other is smaller than its mean.
   - As one variable increases, the other tends to decrease.

To assess whether concordant or discordant pairs prevail, i.e. if between two variables a direct or inverse (or no) relation exists, the average of the cross-products is considered, the so-called **covariance**.

Covariance measures the relationship between two variables, considering both the direction and magnitude of deviations from their means.

Population Covariance:

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

- μ_X and μ_Y are the means of X and Y in the population.
- N is the population size.

Sample Covariance:

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- $\bar{x}$ and $\bar{y}$ are the sample means of X and Y.
- n is the sample size.

The sample covariance is slightly biased as it divides by $n-1$ rather than n.

Covariance can also be calculated with an **indirect calculation formula** using the means of the product of the data and the product of the means:

Population Covariance (Indirect Formula):

$$\sigma_{XY} = \frac{1}{N}\left(\sum_{i=1}^{N} x_i y_i\right) - \mu_x \mu_y$$

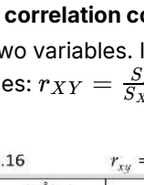
Sample Covariance (Indirect Formula):

$$s_{XY} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

This formula is particularly useful when working with **aggregated data**, such as variables recorded in interval classes.

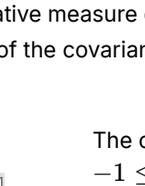
Positive Covariance:

- Concordant pairs dominate.
- As one variable increases, the other tends to increase (positive association).



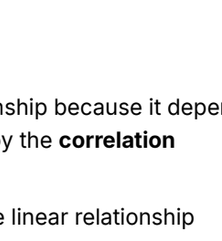
Negative Covariance:

- Discordant pairs dominate.
- As one variable increases, the other tends to decrease (negative association).



Covariance Near Zero:

- No clear association.



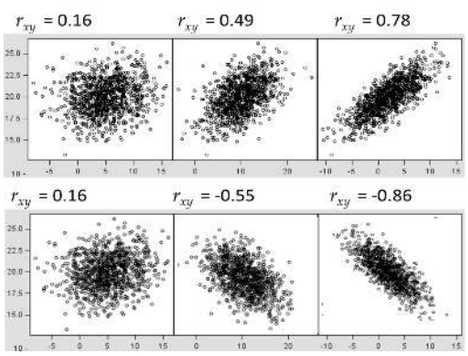
Covariance indicates the direction of the relationship but not its strength. It is an absolute measure, dependent on the units of the variables. For example:

1. **Straight Line (Positive or Negative Slope)**: Strong, monotonic relationship.
2. **Parabola**: Strong relationship, but non-monotonic.

LINEAR CORRELATION COEFFICIENT

Covariance does not provide a standardized measure of the strength of the relationship because it depends on the units of the variables. For this reason, it is often supplemented or replaced by the **correlation coefficient**, which standardizes the covariance.

The **linear correlation coefficient** (r_{XY}) is a relative measure of the strength of the linear relationship between two variables. It is defined as the ratio of the covariance to the product of the standard deviations of the variables: $r_{XY} = \frac{S_{XY}}{S_X S_Y}$.

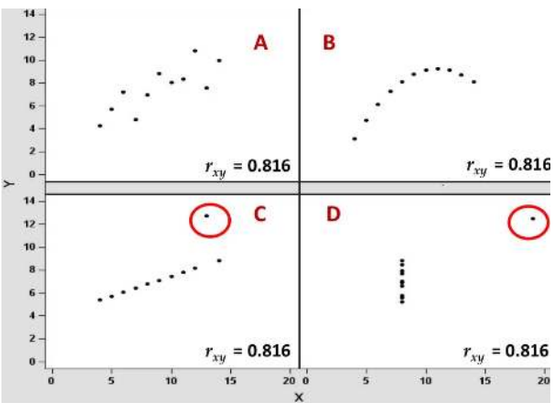


The coefficient takes values between -1 and $+1$ ($-1 \leq r_{XY} \leq +1$):

- $r_{XY} = +1$: Indicates a perfectly **linear, positive** (direct) relationship, all points lie on a straight line with positive slope.
- $r_{XY} = -1$: Indicates a perfectly **linear, negative** (inverse) relationship, all points lie on a straight line with negative slope.
- $r_{XY} = 0$: Indicates no **linear relationship** between the variables, the variables may still have a nonlinear relationship.

In some cases, the correlation coefficient may not be linear:

- Outliers may distort the correlation.
- The same mean, variance, covariance, and correlation may represent different types of relationships.
  - **Case A:** Strong alignment between the relationship and correlation.
  - **Case B:** Perfect nonlinear relationship (e.g., parabola).
  - **Cases C/D:** Outliers significantly alter r_{XY}.



A low correlation coefficient does not imply that the variables are unrelated, it only means there is no strong linear relationship. Correlation only measures the **strength of linear relationships**. Nonlinear relationships and the presence of outliers can make the coefficient unreliable.

COVARIANCE AND CORRELATION

```
cov(x, y)
```

```
cor(x, y)
```

To compute covariance and correlation for numerical variables:

- `x` and `y`: Vectors representing the two variables (e.g., `dataframe$variable_name`).
- `use = "complete"`: Excludes missing values from the calculation.

CONFOUNDING VARIABLES

In data analysis, confounding variables are variables that can distort or obscure the true relationships between variables of interest (i.e., two variables depend on a variable that has a causal relationship with both, called confounding variable). This happens when raw data analysis provides information that is significantly different from the analysis of the same data aggregated by other variables.

We talk about **Simpson's Paradox** when two variables show a certain association when the confounding variable is not taken into account, but when such confounder is eventually taken into account, the relationship **disappears** or is **reversed** (when groups are aggregated).

Conditional distributions should only be compared if they are **homogeneous** concerning confounding factors. Identifying and controlling these confounding factors is the responsibility of the researcher (should use summary measures and plot, but also use statistical evidence to support intuitions).

The existence of a linear relationship between two variables does not imply causality (**Correlation ≠ Causation**). The causal relationship may even be reversed based on context or additional factors.

INFERENTIAL STATISTICS

The statistical **inference problem** arises when one is interested in evaluating measures that describe (or summarize) the characteristics of an entire population, called parameters, but collecting data on all units of the population is prohibitively expensive (cost or time), difficult or even impossible.

In such cases, it may be necessary or convenient to collect data only on a random sample of units drawn from the population and to infer about population parameters based on the sample summary measures calculated on observed data, called statistics.

The inferential process consists of:

1. **Sampling**: Data is collected from a random sample of size n from the population.
2. **Statistics**: Measures summarizing the sample (e.g., mean, variance) are used to make inferences about population parameters.
3. **Uncertainty Assessment**: Assume a known distribution for the population variable:
   - Evaluate the **reliability** of inferences (generalisation) and quantify associated risks.
   - Evaluate the **relationship** between the **parameter** and the **distribution**, as well as the distribution of the considered **statistics** over **all possible samples** of size n that can be drawn from the population.

We consider **random variables** to describe the outcome of **drawing a sample** from a population. A variable is **random** because it is **not possible** to know **a priori** which value it will take among the possible ones, we only know the **probability** of observing **each possible value.**

Inferential statistics involves the procedures used to draw conclusions about population parameters (θ) based on statistics computed from a random sample ($X_1, X_2, \ldots, X_n$).

There are two types of estimation:

1. **Point Estimation**: Provides a single value as the estimate of the population parameter (e.g., sample mean).
2. **Interval Estimation**: Provides a range of values within which the parameter is likely to lie, with a specified level of confidence.

Hypothesis testing on a parameter involves evaluating which of two competing hypotheses about a parameter is better supported by sample data.

RANDOM VARIABLES

Random variables (RVs) describe the outcome of extracting a sample from the population. They can be:

- Discrete.
- Continuous.

DISCRETE RANDOM VARIABLES

A **discrete random variable** X is a variable whose value is unknown but can only take a countable number of distinct values. While the actual value of the RV is unknown prior to sampling, the probability of each outcome can be described using a **probability function**.

The **probability function** $P_X(x)$ assigns a probability p_x to each value x:

$$P_X(x) = \begin{cases} p_1 & \text{if } x = x_1, \\ p_2 & \text{if } x = x_2, \\ \vdots & \\ p_n & \text{if } x = x_n, \\ 0 & \text{otherwise} \end{cases}, \text{ where } P_X(x) = \text{Prob}(X = x).$$

The properties of probability functions:

1. $0 \leq P_X(x) \leq 1$, for all x.
2. $\sum_x P_X(x) = 1$.

The **cumulative distribution function (CDF)** $F_X(x)$ gives the probability that X is at most x: $F_X(x) = \text{Prob}(X \leq x)$

EXPECTED VALUE (MEAN)

The expected value ($E(X)$) is the weighted average of all possible values of X, weighted by their probabilities: $E(X) = \mu = \sum_x x \cdot P_X(x)$

VARIANCE

The variance measures the expected quadratic deviation of the values of X from the expected value: $\text{Var}(X) = \sigma^2 = E((X-\mu)^2) = E(X^2) - \mu^2 = \sum_x (x-\mu)^2 \cdot P_X(x)$

If the probability function exactly reflects the composition of the population, the expected value and the variance will coincide with the population mean and variance.

BERNOULLI DISTRIBUTION

A **Bernoulli distribution** describes whether an event with two possible results occurs:

- **Success** ($X = 1$).
- **Failure** ($X = 0$).

The probability function is defined as:

$$P_X(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} p^x(1-p)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$, where the parameter p indicates the probability of success.

EXPECTED VALUE

$E(X) = (1-p) \cdot 0 + p \cdot 1 = p$

VARIANCE

$\text{Var}(X) = (0-p)^2(1-p) + (1-p)^2 p = p - p^2 = p(1-p)$

CONTINUOUS RANDOM VARIABLES

A **continuous random variable** can take any value within an interval, its value is unknown, and it is derived from a **measurement process**.

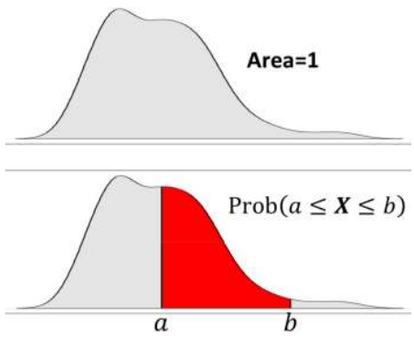
The probability that a continuous r.v. X takes a specific value x must be 0, whatever x is.

The probability of a continuous random variable X is described by a **probability density function (PDF)**, $f_X(x)$, which satisfies the following properties:

1. $f_X(x) \geq 0$ for all x.
2. The total area under the curve of $f_X(x)$ is 1:
   $$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

The probability that X falls within an interval $[a, b]$ is given by:

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$



For continuous (not discrete) random variables we have that $P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$. This equivalence arises because the probability of X taking any specific value is 0.

The CDF for a continuous random variable X is: $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt.$

It is important to note that $\text{Prob}(a \leq X \leq b) = F(b) - F(a)$ for all $a < b$.

For **continuous random variables**, the expected value and variance are calculated similarly to the discrete case, but sums are replaced by integrals.

EXPECTED VALUE

$$E(X) = \mu = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx$$

VARIANCE

$$\text{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_X(x)\,dx$$

NOTABLE DISTRIBUTIONS

Defining a density function that adequately describes the population (based on past experience or on the researcher's assumptions) can be complicated. Theoretical models have been developed to describe some typical situations.

Notable distributions depend on parameters that act on their shape so as to fit the assumed characteristics of the population's distribution.

1. **UNIFORM DISTRIBUTION**

   A uniform random variable has a constant probability density over a given interval.

2. **CHI SQUARE DISTRIBUTION**

2. **NORMAL DISTRIBUTION**

   The **normal distribution** is the most important probability distribution.

   A random variable X follows a normal distribution with parameters μ (mean) and σ^2 (variance), denoted as: $X \sim N(\mu, \sigma^2)$.

   The PDF is: $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ with $-\infty < x < \infty$.

   - **Symmetry**: Bell-shaped and symmetric around μ (controls location of the peak).
     - Larger μ: Distribution shifts to the right.
     - Smaller μ: Distribution shifts to the left.
   - **Centrality**: Mean, median, and mode are equal to μ.
   - **Dispersion**: Spread depends on σ:
     - Larger σ: Flatter and more dispersed (wider curve).
     - Smaller σ: Taller and more concentrated (narrower curve).

CUMULATIVE DISTRIBUTION FUNCTION (CDF)

```
pnorm(q, mean=0, sd=1)
```

- `q`: Value at which to calculate the CDF given the parameters → $F(q) = P(X \leq q)$
- `mean`: Mean of the distribution (default is 0).
- `sd`: Standard deviation (default is 1).

The result is a probability.

$F(q) = P(X \geq q)$ → `1-pnorm(q)`

$F(q) = P(q_1 \leq X \leq q_2)$ → `pnorm(q2)-pnorm(q1)`

PERCENTILES

```
qnorm(p, mean=0, sd=1)
```

- `p`: Order of the percentile at which the CDF equals p → $F(x_{1-p}) = P(X \leq x_{1-p}) = p$
- `mean`: Mean of the distribution (default is 0).
- `sd`: Standard deviation (default is 1).

The result is a percentile.

`c(qnorm(p1),qnorm(p2))` → interval between p_1 and p_2.

LINEAR TRANSFORMATION AND STANDARDIZATION

A linear transformation of a random variable X is expressed as:

$Y = a + bX$, where a and b are constants.

The **expected value** and **variance** of Y are related to those of X as follows:

EXPECTED VALUE

$E(Y) = E(a + bX) = a + bE(X) = a + b\mu$

VARIANCE → STANDARD DEVIATION

$\mathrm{Var}(Y) = \mathrm{Var}(a + bX) = b^2\mathrm{Var}(X) = b^2\sigma^2$ →
$\mathrm{Sd}(Y) = |b| \cdot \mathrm{Sd}(X) = |b| \cdot \sigma$

it is not always possible to easily determine the probability or density distribution of Y based on the distribution of X. Even so, in the particular case of a normally distributed r.v. X, with expected value μ and variance σ^2, any linear transformation of X has a normal distribution.

If $X \sim N(\mu, \sigma^2)$, then $Y = a + bX \sim N(a + b\mu, b^2\sigma^2)$.

This property ensures that linear transformations of normally distributed random variables remain normally distributed.

STANDARDIZATION

Standardization is a specific linear transformation. Given any random variable X, the standardized variable is $Z = \frac{X-\mu}{\sigma}$. Z is defined based on the expected value and variance of X.

EXPECTED VALUE

$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X)-\mu}{\sigma} = 0$

VARIANCE

$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{Var}(X) = 1$

By standardizing a normal random variable $X \sim N(\mu, \sigma^2)$, we obtain the **standard normal distribution**: $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

This distribution has some properties:

1. Symmetric around 0.
2. $\mathrm{Prob}(Z \leq 0) = 0.5 = \mathrm{Prob}(Z \geq 0)$
3. $\mathrm{Prob}(Z \leq z) = p$ implies:
   - $\mathrm{Prob}(Z \geq z) = 1 - p$.
   - $\mathrm{Prob}(Z \geq -z) = p$.

It is also important to note that, since $z = \frac{x-\mu}{\sigma}$, we have that $x_{1-p} = \mu + z_{1-p}\sigma = \mu - z_p\sigma$.

LINEAR COMBINATIONS OF RANDOM VARIABLES

JOINT DISTRIBUTION OF TWO RANDOM VARIABLES

To study linear combinations of random variables, we rely on the concept of **joint distributions**, which describe the probabilities or densities associated with pairs (or intervals) of values for two random variables X and Y.

- For discrete random variables: $P_{X,Y}(x,y) = \text{Prob}(X = x, Y = y)$
- For continuous random variables: $\text{Prob}(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y)\,dx\,dy$

COVARIANCE

$\text{Cov}(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$

CORRELATION

$\text{Corr}(X,Y) = \rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$

A particularly important joint density distribution is the **bivariate normal distribution**. If X and Y have a joint normal distribution, then they singularly have a normal distribution (while the opposite is not true).

For a linear combination Z = aX + bY, we have:

EXPECTED VALUE

$E(Z) = E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y$

VARIANCE → COVARIANCE

$\text{Var}(Z) = \text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y)$

→ $\text{Cov}(X,Y) = \sigma_{XY} = \rho_{XY}\sigma_X\sigma_Y$

The distribution of Z depends on the joint distribution of the two r.v.

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ have a **joint normal distribution**, then any linear combination $Z = aX + bY$ is also normally distributed: $Z \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY})$.

INDEPENDENT RANDOM VARIABLES

If X and Y are **independent**, then the probability of observing certain values for one r.v. doe not depend in any way on the values taken by the other r.v., so that:

$P_{X,Y}(x,y) = \text{Prob}(X = x, Y = y) = \text{Prob}(X = x)\text{Prob}(Y = y) = P_X(x) \cdot P_Y(y)$

$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$

This means that the probability of jointly observing values of X and Y can be determined from the (marginal) distributions of the two r.v.

When two r.v. are independent they are also **linearly independent**, that is: $\text{Cov}(X,Y) = \text{Corr}(X,Y) = 0$

SUM AND MEAN OF IID (INDEPENDENT AND IDENTICALLY DISTRIBUTED) RANDOM VARIABLES

Given X with expected value μ and variance equal to σ^2, we consider n r.v. $X_1, X_2, \ldots, X_n$ with the following properties:

- They are independent, therefore all the pairs have covariance equal to 0.
- They all have the same distribution as X.

This is the case when we consider n units randomly selected from the same population and each of them describes the random result of the selection.

If we take n independent (covariance is 0) and identically distributed (i.i.d.) random variables, we can define the **sum** and **mean.** These are random variables themselves, with expected values and variances derived from μ and σ^2 (same of each).

- **Sum**: $S = X_1 + X_2 + \cdots + X_n$.
  - **EXPECTED VALUE**: $E(S) = E(X_1) + \cdots + E(X_n) = n\mu$ → $E(X_i) = \mu$
  - **VARIANCE**: $\text{Var}(S) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = n\sigma^2$ → $Var(X_i) = \sigma^2$
- **Mean**: $\bar{X} = \frac{S}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$
  - **EXPECTED VALUE**: $E(\bar{X}) = E(\frac{X_1}{n}) + \cdots + E(\frac{X_n}{n}) = \frac{n \cdot \mu}{n} = \mu$ → $E(\frac{X_i}{n}) = \frac{\mu}{n}$
  - **VARIANCE**: $\text{Var}(\bar{X}) = \text{Var}(\frac{X_1}{n}) + \cdots + \text{Var}(\frac{X_n}{n}) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$ → $Var(\frac{X_i}{n}) = \frac{\sigma^2}{n^2}$

If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, the **sum** and **mean** are also normally distributed:

$S \sim N(n\mu, n\sigma^2)$

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

CENTRAL LIMIT THEOREM (CLT)

The **Central Limit Theorem** states that for n sufficiently large (typically >30), the distribution of the **sum** and **mean** of i.i.d. random variables can be approximated by the normal distribution, regardless of the original distribution of X.

If n is not sufficiently large, nothing can be said about expected value and variance, unless we assume that the random variables are normally distributed.

BERNOULLI DISTRIBUTION

For n i.i.d. random variables $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$, where p is the probability of success:

EXPECTED VALUE	VARIANCE
$E(X) = p$	$\text{Var}(X) = p(1-p)$

The **sum** becomes the number of successes: $S = X_1 + X_2 + \cdots + X_n \rightarrow S \sim N(np, np(1-p))$

The mean becomes the proportion of successes: $\hat{P} = \frac{S}{n} \rightarrow \hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right)$

The normal approximation holds for n>30.

POINT ESTIMATION

With reference to point estimation, it is necessary to distinguish between:

- **Parameter (θ)**: Measurable characteristic of the population with reference to a random variable X. Examples:
  - Population mean (μ).
  - Population standard deviation (σ).
  - Proportion (p).
- **Estimator ($\hat{\theta}$)**: Statistic used to estimate the parameter. It is a function of the outcome of a random sample of n independent and identically distributed draws $X_1, X_2, \ldots, X_n$ distributed as X. Example:
  - Sample mean ($\bar{X}$)
- **Estimate ($\hat{\theta}_n$)**: Sample realisation of an estimator calculated from the specific sample actually drawn. Estimate is the result of that process.

In inference, the **characteristics** of the population are **not known.**

It will necessarily have to be based on the specific estimate corresponding to a specific sample extracted from the population.

We reflect upon the reliability of the procedure by focusing on the expected value and the dispersion of the estimates around the parameter of interest. With what probability can we expect realizations of the estimator far away from the parameter of interest?

We will NOT be able to make any assessment, especially with respect to how close this estimates is to the value of the parameter (which is unknown).

We evaluate estimators based on two properties:

1. **Unbiasedness**: A point estimator is **unbiased** if its expected value equals the true parameter value for all sample sizes and parameters: $E(\hat{\theta}_n) = \theta$.

   The **bias** of an estimator is: $D(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta$

   - If $D(\hat{\theta}_n) = 0$, the estimator is **unbiased**.
   - If the bias decreases as $n \rightarrow \infty$, the estimator is **asymptotically unbiased**: $\lim_{n\to\infty} D(\hat{\theta}_n) = 0 \iff \lim_{n\to\infty} E(\hat{\theta}_n) = \theta$

   This does not mean that a particular value of the estimator must be exactly the correct parameter, but that the estimator has the capability of estimating a parameter correctly on the average.

   Sample mean, variance and proportion are unbiased estimators of the corresponding parameters.

2. **Efficiency**: The **variance** (or **standard error**)of an unbiased estimator is the expected squared deviation of a generic realisation of the estimator from the parameter of interest: $E[(\hat{\theta}_n - \theta)^2] = \text{Var}(\hat{\theta}_n)$.

   This is important because it expresses how dispersed each single estimation will be around the actual parameter. The unbiased estimator with the smallest variance will be the most efficient and most reliable.

ESTIMATOR FOR THE POPULATION MEAN

The sample mean $\bar{X}$ is the most natural estimator for the population mean $E(X) = \mu$, based on an independent identically distributed random sample.

Whatever the distribution of X:

- The value of the mean estimator will be equal to the mean itself ($E(\bar{X}) = \mu$), making it therefore a **unbiased estimator** for μ.

  SAMPLE MEAN

  ```
  Mean <- mean(dataframe$column_name)
  ```

- The **dispersion** decreases as the size of the sample increases:
  - $\sigma^2 = \mathrm{Var}(X)$
  - $SE_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}}$

  STANDARD ERROR

  ```
  Standard_Error <- sigma / sqrt(n)
  ```

  The SE is the expected deviation from the population mean of a generic and not of a specific estimate.

  The lower the standard error of the estimator the higher the probability of estimates close to the population mean.

Additionally, we know that this unbiased estimator is the **most efficient** (minimum variance) and that, if the r.v. is normally distributed, the distribution of the estimator will be so as well. If the size of X is big enough, we can state this independently from its distribution (CLT).

ESTIMATOR FOR THE POPULATION VARIANCE

The variance of X depends on the variance of the population σ^2, information typically unknown. To estimate it we use the sample variance $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$.

It can be shown that $E(S^2) = \sigma^2$. The sample variance is an unbiased estimator for the population variance, whatever the distribution of X:

The **unadjusted sample variance** $\tilde{S}^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}$ is characterised by the expected value $E(\tilde{S}^2) = \left(\frac{n-1}{n}\right) \cdot \sigma^2$. This makes it a biased asymptotically unbiased estimator for the population's variance.

SAMPLE VARIANCE

```
Variance_S2 <- var(dataframe$column_name)
```

STANDARD ERROR

```
Standard_Error <- Variance_S2 / sqrt(n)
```

ESTIMATOR FOR THE POPULATION PROPORTION

Another parameter one is typically interested in is the proportion p of cases in a population that exhibit a characteristic (coded as 'success'), estimated on the basis of a simple random sample of n units, by measuring the r.v., indicating for each unit whether a success was observed (1) or not (0).

To estimate p we use the sample proportion $\hat{P}$. The r.v. are i.i.d. according to a Bernoulli distribution with parameter p, and $\hat{P}$ is therefore their sample mean. From the properties of the sample mean we have:

- $\hat{P}$ is an unbiased estimator for p: $E(\hat{P}) = E(X) = p$.
- The variance of the estimator (never known but estimated by substituting p with its estimate $\hat{p}$) is smaller the larger the sample: $\mathrm{Var}(\hat{p}) = \frac{\mathrm{Var}(X)}{n} = \frac{p(1-p)}{n}$

If the size of the sample is large enough, the distribution of $\hat{P}$ can be approximated by a normal distribution: $P \approx N(p, \frac{p(1-p)}{n})$

PROPORTION p ESTIMATE, STANDARD ERROR OF ESTIMATOR

DISTRIBUTION OF CATEGORICAL VARIABLE

```
distribution <- table(dataframe$column_name)
```

SAMPLE PROPORTION

```
proportion <- Y / n  # Y is the number of favorable cases
```

STANDARD ERROR

```
Standard_Error <- sqrt(proportion * (1 - proportion) / n)
```

When asked if the evaluated expected deviation is the exact difference between estimate and the parameter we cannot answer. The probability to draw a sample leading to an estimate close to μ is higher than the probability to draw a sample leading to an estimate far from μ.

Additionally, to reach a lower SE, we need to increase the sample size: $\mathrm{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

However, we cannot state that a higher sample size leads to a more reliable estimate. It might be that the larger sample contains some extreme cases and the smaller contains very central observations.

The same rationale applies for estimating a proportion: in the case of an estimator $\bar{X}$ (sample mean) of which the standard deviation σ is known, if we want to keep the standard error below a certain value z, the condition that will need to be satisfied is: $\mathrm{SE}_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \leq z \rightarrow \sqrt{n} \geq \frac{\sigma}{z} \rightarrow n \geq \frac{\sigma^2}{z^2}$ (with the possibility of using an estimate S^2 for the variance, so the sample variance, if its value in the population is unknown).

SAMPLE MEAN AND SAMPLE VARIANCE

```
mean()
var()
```

In the case we need to evaluate mean and/or variance with data sets including some missing values (NA), we can use some other functions on R.

MISSING VALUES

```
na.rm=TRUE
complete.cases
```

- `na.rm = TRUE`: Excludes the missing values.
- `complete.cases`: If used with the `sum` function as well, it will determine the number of rows with real values (not NA).

ESTIMATOR FOR THE DIFFERENCE BETWEEN MEANS

In some cases the parameter we might be interested in evaluating may be the difference between the means of two populations $\mu_x - \mu_y$.

To estimate the difference between the means of two populations described by two r.v. X and Y on the basis of two samples $X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_n$, it is natural to use the estimator $\bar{X} - \bar{Y}$. Whatever the distribution of the random variables is:

- $(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$, so the estimator is unbiased.
- The variance, and thus the standard error, of the estimator depends on the relationship between the two populations and on their joint distribution of X and Y.
- The distribution of $\bar{X}$ and $\bar{Y}$ depends on X and Y and their joint distribution, but as always remains the possibility to approximate with a normal if the sample is large enough.

About this, we have to make two distinctions:

1. **Independent sample**: Samples are possibly of different sizes, drawn independently, and made up of different statistical units. In this case the sample averages X and Y are independent and thus **uncorrelated**.
   - The **variance** of the estimator is: $\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$.
   - The **standard error** of the estimator is: $\mathrm{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$
2. **Paired sample**:
   - If the **variances** are **unknown and different**, we can substitute the parameter with the estimates: $\mathrm{Var}(\bar{X} - \bar{Y}) = \frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}$
     - **Standard error** ($\sigma_X^2 \neq \sigma_Y^2$): $\mathrm{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$
   - If the **variances** are **unknown and equal**, the common variance is the pooled sample variance (weighted average of sample variances): $S_{\mathrm{Pool}}^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$
     - **Standard error** ($\sigma_X^2 = \sigma_Y^2$): $\mathrm{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{S_{pool}^2}{n_X} + \frac{S_{pool}^2}{n_Y}}$

   In the case of paired samples of size n, two measurements are available for each sample unit: X_i and Y_i. The differences $D_i = X_i - Y_i$ can therefore be considered as n measurements from the r.v. $\bar{D} = \bar{X} - \bar{Y}$

   The **expected value** of all the differences $D_i = X_i - Y_i$ is equal to the estimate of the sample, so $E(\bar{D}) = \mu_D = E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y$.

   The **variance**, instead, is: $\mathrm{Var}(\bar{D}) = \frac{\mathrm{Var}(D)}{n} = \frac{\sigma_D^2}{n}$, where $\mathrm{Var}(D) = \mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$.

   However, it is hard to conclude that the variance of D is known. To **estimate** its variance, we can proceed as the **sample variance** of the differences, related to the sample variances and covariance:

$$S_D^2 = \sum_{i=1}^{n} \frac{(D_i - \bar{D})^2}{n-1} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} + \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1} - 2\sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = S_X^2 + S_Y^2 - 2S_{XY}$$

ESTIMATORS FOR THE DIFFERENCE BETWEEN PROPORTIONS

When we are interested in comparing the proportions of a certain phenomenon in two different populations, the parameter of interest is the difference between the proportions in the two populations: $p_X - p_Y$.

To estimate the difference between the proportions of 'successes' in two populations, it is natural to use the estimator $\hat{P}_X - \hat{P}_Y$ where $\hat{P}_X$ and $\hat{P}_Y$ are the **sample means** of two samples whose units have Bernoulli distributions.

For the properties of the sample means, we have that $E(\hat{p}_X - \hat{p}_Y) = E(\hat{p}_X) - E(\hat{p}_Y) = p_X - p_Y$, so $\hat{P}_X - \hat{P}_Y$ is an unbiased estimator for $p_X - p_Y$.

The **variance** of the estimator depends on the relationships between X and Y. In the case of **independent** samples, we have:

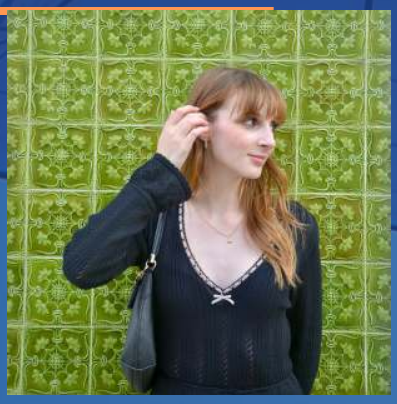
$$\mathrm{Var}(\hat{P}_X - \hat{P}_Y) = \mathrm{Var}(\hat{P}_X) + \mathrm{Var}(\hat{P}_Y) = \frac{\mathrm{Var}(X)}{n_X} + \frac{\mathrm{Var}(Y)}{n_Y} = \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}$$

The variance is unknown (it depends on p_X and p_Y), but it can be estimated by replacing p_X and p_Y with their estimates (observed sample proportions).

IN SUMMARY

Parameter	Estimator	Estimate	Samples	Standard error, *SE*	Standard error estimate, *se*
μ	$\overline{X}$	$\overline{x}$		$\sigma/\sqrt{n}$	$s/\sqrt{n}$
p	$\hat{P}$	$\hat{p}$		$\sqrt{p(1-p)/n}$	$\sqrt{\hat{p}(1-\hat{p})/n}$
$\mu_X - \mu_Y$	$\overline{X} - \overline{Y}$	$\overline{x} - \overline{y}$	Independent	$\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$	$\sigma_X^2 = \sigma_Y^2 \rightarrow \sqrt{\frac{s_{Pool}^2}{n_X} + \frac{s_{Pool}^2}{n_Y}}$ $\sigma_X^2 \neq \sigma_Y^2 \rightarrow \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$
			Paired	$\sqrt{\frac{\sigma_D^2}{n}} = \sqrt{\frac{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}{n}}$	$\sqrt{\frac{s_D^2}{n}} = \sqrt{\frac{s_X^2 + s_Y^2 - 2s_{XY}}{n}}$
$p_X - p_Y$	$\hat{P}_X - \hat{P}_Y$	$\hat{p}_X - \hat{p}_Y$	Independent	$\sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}$	$\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n_X} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n_Y}}$

FOR DOUBTS OR SUGGESTIONS ON THE HANDOUTS



MATILDE BALDINI

matilde.baldini@studbocconi.it
@_matildebaldini_
+39 3470273884

FOR INFO ON THE TEACHING DIVISION

VITTORIA NASONTE

vittoria.nasonte@studbocconi.it
@_vittorian_
+39 3274441476

ELENA CACIOLI

elena.cacioli@studbocconi.it
@elenacaciolii_
+39 3928931605

TEACHING DIVISION

OUR PARTNERS

700+ CLUB

ETHAN SUSTAINABILITY

DELIVERY VALLEY
NO GENDER KITCHEN

LA PIADINERIA