Simple Statistics

Statistics can be divided in three areas:

I DESCRIPTIVE STATISTICS

a) Presentation of data in tables and graphs.
b) Synthetic indices that describe the distribution of data.

II MATHEMATICAL STATISTICS

Probability calculus and theoretical distributions.
Theoretical probability distributions: Binomial, Poisson, Hypergeometric, Normal, ...

III INFERENTIAL STATISTICS

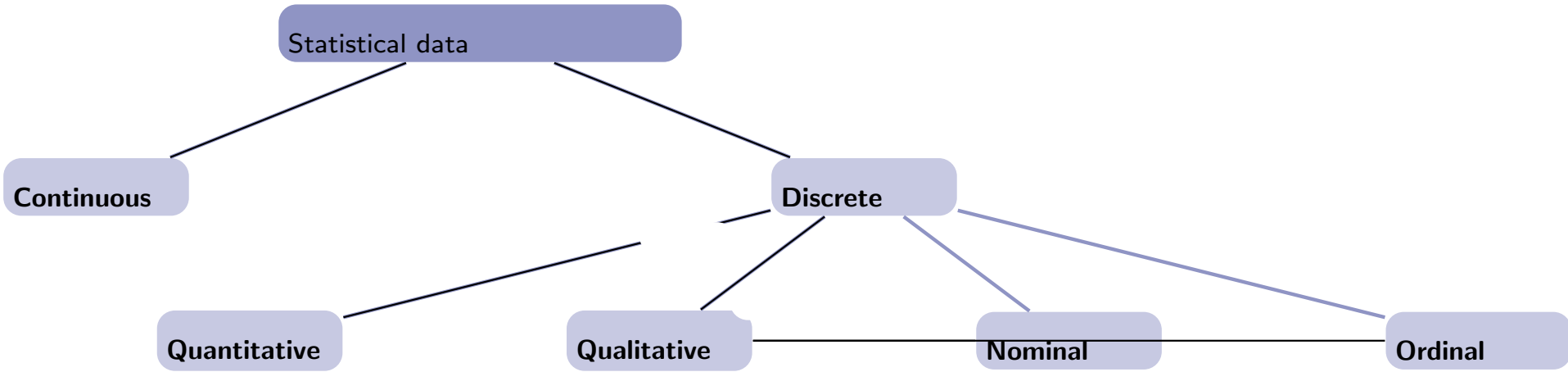
Deduction of general laws from a random sample of data.

a) Parametric hypothesis (about mean, variance, . . .) and functional hypothesis (about the distribution).
b) Univariate, bivariate, multivariate distributions.
c) Parametric and non-parametric statistics.
d) One-sample, two-sample, n-sample tests.

During the statistical analysis, data features have to be considered carefully.

The experimental phase in which information are collected is a key point: both description methods and applicable tests rely on experimental phase.

Statistical universe



Then, four types of scale exist:

- ⋆ Continuous
- ⋆ Discrete Qualitative Nominal
- ⋆ Discrete Quantitative Ordinal
- ⋆ Discrete Quantitative

Depending on the kind of data, different statistical options are available.

In general, on Qualitative data fewer statistical options are available than on quantitative data.

For example, on Quantitative data one might calculate the average, whereas, on qualitative nominal data this couldn't be done.

Again, on Nominal data fewer statistical options are available than on Ordinal data.

Ordinal data can be treated as Nominal; the converse is not true.

Quantitative data can be always considered as Ordinal; the converse is not true.

A **random variable** (or **stochastic variable**) is a variable whose values are associated to a probability law.

When a variable is not associated to a probability law it is called deterministic variable.

The **random variables** are split into two main categories that are treated in different ways:

1. **Discrete random variables**
   that are usually used to model qualitative (nominal or ordinal) data frequencies.

2. **Continuous random variables**
   expressed on a continuous scale of values. Values are often associated to a proper **unit of measurement**.

The difference between **discrete** and **continuous** random variables is often not clearly defined, but it can depend from the number of **categories** associated to the variable.

For example, the measurement of an height might be modeled by a discrete variable if it is measured in inches. If more decimal digits are considered then the number of categories increase and the same variable can be considered as a continuous variable.

Discrete variables have less explicative power than continuous ones:

- ⋆ statistics computed on continuous variables contains more information;
- ⋆ to get good quality statistics using discrete variables a greater **number of observations** is required.

Data should be collected using the more accurate scale. When possible, data should be collected on a continuos scale. During the analysis data can be split into classes. This produce more readable results at the cost of information loss.

A set of observed data on a variable is called Sample.

A first and basic data series processing is the **data ordering**, in ascending or descending way.

The difference between the highest value (maximum) and the lowest value (minimum) is called **range**.

Data can then be categorized into classes. The number of values (or **statistical units**) for each class can so be counted.

Values of frequencies of each category produce a **frequency distribution**, often called simply **distribution**.

The table below reports the number of events counted on 45 time frames of the same length.

5	6	3	4	7	2	3	2	3	2	6	4	3	9	3
2	0	3	3	4	6	5	4	2	3	6	7	3	4	2
5	1	3	4	3	7	0	2	1	3	1	5	0	4	5

Tabella 1: Number of events.

The first step is the **definition of classes**:

- ⋆ the **minimum** (**0**, in the table) and the maximum (**9**) must be identified;
- ⋆ the number of data within each class (how many zeros, how many ones, . . .) must be counted.

Class	h	0	1	2	3	4	5	6	7	8	9
Absolute Freq	n	3	3	7	12	7	5	4	3	0	1
Relative Freq	f	0.07	0.07	0.15	0.27	0.15	0.11	0.09	0.07	0.00	0.02
Cumulative Abs Freq		3	6	13	25	32	37	41	44	44	45
Cumulative Rel Freq	F	0.07	0.14	0.29	0.56	0.71	0.82	0.91	0.98	0.98	1.00

Table 2: Distribution of absolute and relative frequencies of the number of events.

The **class** row represents a category and, in this case, it is a count value.

The **absolute frequency** of a class is the count of data values of that class.

The **relative frequency** of a class is its absolute frequency divided by the total count of observations (in this example: 45).

The **cumulative (absolute or relative) frequency** of a class is the (absolute or relative) frequency of this class added to the sum of frequencies of all the lower classes. It requires data that can be ordered.

The use of the relative frequencies instead of absolute frequencies is especially useful when two or more distributions with different total counts of observation must be compared.

The cumulative frequency provides useful information to estimate the total count of observations less (or higher) than a set value.

The frequency distribution provides an easy reading of most important characteristics of data: in the example, the value 3 represents the **central tendency**.

The minimum and the maximum are, respectively, 0 and 9 and they provide the range, a measure of the **variability**.

In this example, frequencies tend to decrease in a similar way in both directions starting from 3. As the number of classes with non null frequency above the center is greater than the number of classes below the center, the **shape** of the distribution is not symmetric but asymmetric.

Group	**h**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Absolute Freq	n	3	3	7	12	7	5	4	3	0	1
Relative Freq	f	0.07	0.07	0.15	0.27	0.15	0.11	0.09	0.07	0.00	0.02
Cumulative Abs Freq		3	6	13	25	32	37	41	44	44	45
Cumulative Rel Freq	F	0.07	0.14	0.29	0.56	0.71	0.82	0.91	0.98	0.98	1.00

Table 3: Distribution of absolute and relative frequencies of the number of events.

How many classes should be created when building a frequency table? This is the first question when data grouping is required. The choice depends a lot on the total number **N** of observations and, to a lesser degree, data variability.

Frequency distributions tend to be representative of the phenomenon only when enough observations are available.

An empirical rule suggests to use 4-5 classes when observations are 10-15 until a maximum of 15-20 classes when observations are more than a hundred.

Too few classes cause a loss of information about distribution features.

On the other side, **too many classes** return useless information because the distribution shape is not clearly visible.

Groups with equal widths are not mandatory, but they are suggested for an easier reading.

Plots and parameter computation require some non intuitive cares if classes have different sizes.

Many methods have been proposed to compute the number of classes, which include:

Sturges' formula: $C = 1 + \frac{10}{3}\log_{10}(N)$

Scott's formula: $h = \frac{3.5 \cdot s}{\sqrt{N}}$

when:

C is the number of classes;

N is the total count of observations;

s is the standard deviation;

h is the optimal size of categories.

These methods are implemented on the most common statistical software.

Data grouping requires some care when the variable is **continuos**.

The table shows 40 measurements, in cm:

107	83	100	128	143	127	117	125	64	119
98	111	119	130	170	143	156	126	113	127
130	120	108	95	192	124	129	143	198	131
163	152	104	119	161	178	135	146	158	176

Table 4: 40 measurements

Making a class for each centimeter is useless.

$\Downarrow$

For this reason data grouping is required. Each class includes more than one category.

$\Downarrow$

The total range can be defined knowing the minimum and the maximum. During the data grouping the lower limit of the first class and the upper limit of the last class may not correspond to observed values. Of course, observed values must be included in the global range.

$\Downarrow$

In this example, a total range of 140 cm can be created, from 60 cm to 199 cm, limits included.

$\Downarrow$

The number of classes is chosen based on the number of observations (40). In this example, total number of classes can intuitively be 7, each one with a size of 20 cm.

The lower and the upper limit for each class must be clearly defined, to avoid uncertainty in the assignment of a value to two contiguos classes.

In the example, the following classes can be built: 60-79 the first one, 80-99 the second one, 100-119 the third one and so on until 180-199 for the last one.
Data grouping such as 60-80, 80-100, 100-120, ... must be avoided.

If data are collected with two decimal digits then the class 60-79 must be read as 60.00-79.99 cm. In the same way the class 180-199 must be read as 180.00-199.99 cm.

Alternatively, classes built on continuos scales can be shown in the following way: $[60; 80)$, $[80; 100)$, ... where "[" stands for "limit included" and ")" stands for "limit not included".

Classification tables are obtained from data split into classes. Different classification tables can so produce different data descriptions.

For small samples, differences due to data categorization can be appreciable. Increasing the sample size, subjective choices have smaller effect to distribution shape.

Remember that the first and the last class should not have open bounds on extremes (for example "< 80" the first one and "> 180" the last one).

The information about the mimimum or the maximum will be lost using open bound extreme classes. For this reason, it is impossible to know the central value of an open bound class.

Graphics of open bound extreme classes are impossible to draw or they must be drawn in a subjective way.

Class	x	60-79	80-99	100-119	120-139	140-159	160-179	180-199
Abs Freq	n	1	3	10	12	7	5	2
Rel Freq %	f%	2.5	7.5	25.0	30.0	17.5	12.5	5.0
Cum Rel Freq %	F%	2.5	10.0	35.0	65.0	82.5	95.0	100.0

Tabella 5: Absolute and relative frequencies of 40 categorized measurements

The number of classes computed with the Sturges' and Scott's formulas is 6 and 9, respectively. This can be a confirm to the previous empirical choice.

The frequency distribution table clearly provides the main characteristics of the data series:

- ⋆ **Position** (central tendency);
- ⋆ **Dispersion or variability**.

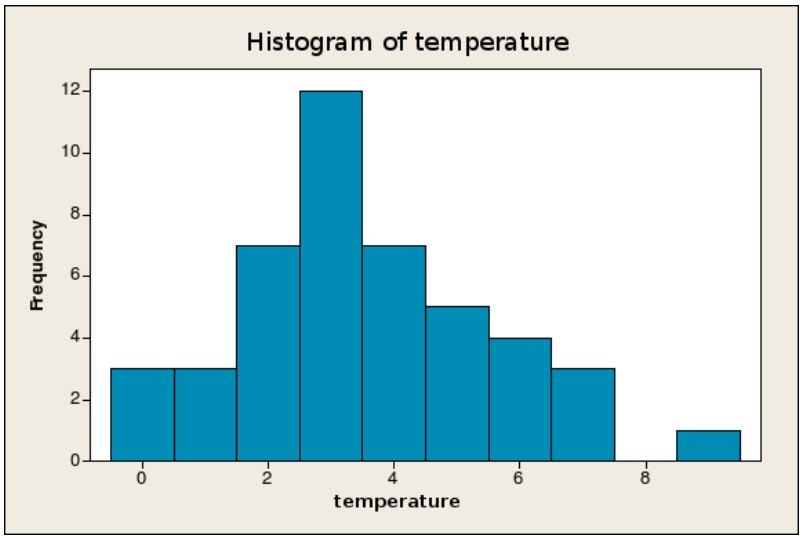
Main drawback: **the distribution within each class is not known**.

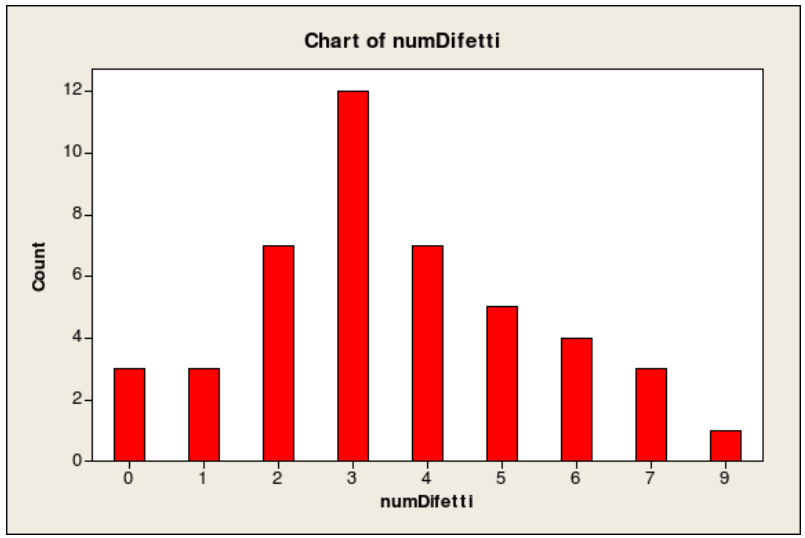
Central values of each class are used to estimate distribution parameters (mean, variance, symmetry, kurtosis). **The main assumption is that, inside each class, data are uniformly distributed**.

This procedure involves an approximation, with respect to the use of individual observations, because the **implicit main assumption is usually not true**.

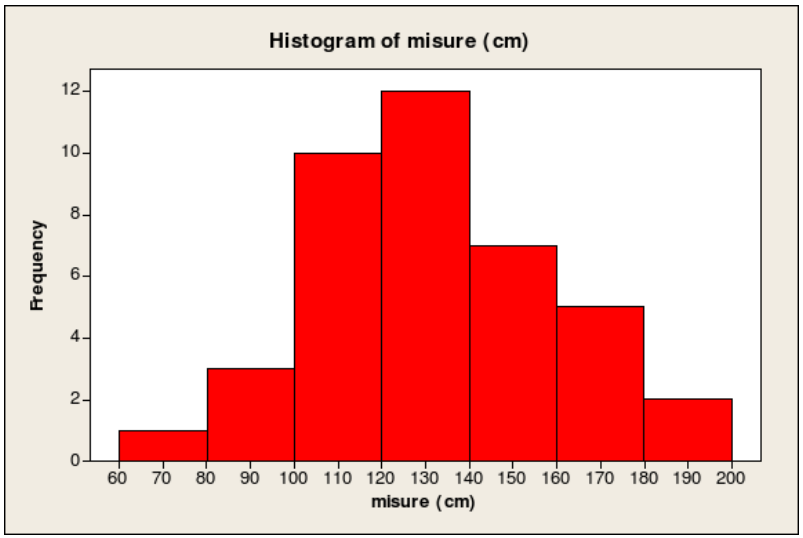
The **plot type** must be chosen in relation to the data type and then the data scale.

Histograms or **bar chart** can be used for **individual data series**. These two graphics are often mistaken for each other.





Measurements values are reported in the horizontal axis, typically grouped in classes. The vertical axis shows the absolute, relative or percentage frequency for each class.



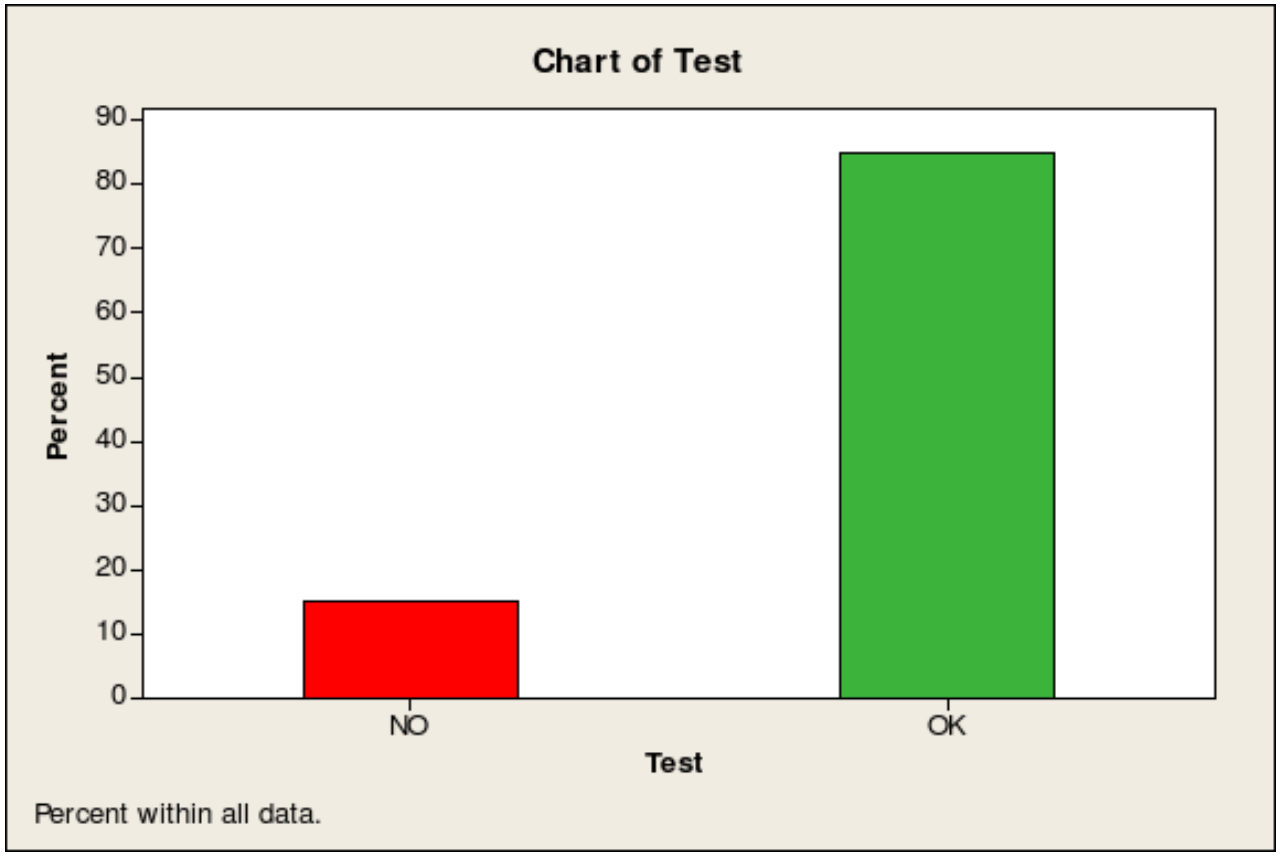
In an histogram **the rectangle areas must be proportional** to the corresponding frequencies.

The **bar chart** or bar graph is a chart with rectangular bars. Widths of bars are equal while heights are proportional to values that each bar represents.

Unlike histograms, bar charts don't have contiguos rectangles. In fact, the bar are separated. Bar charts are used for plotting discrete (or "discontinuous" or categorical) data. Names, labels or symbols are reported in the x-axis instead of numeric measurements.

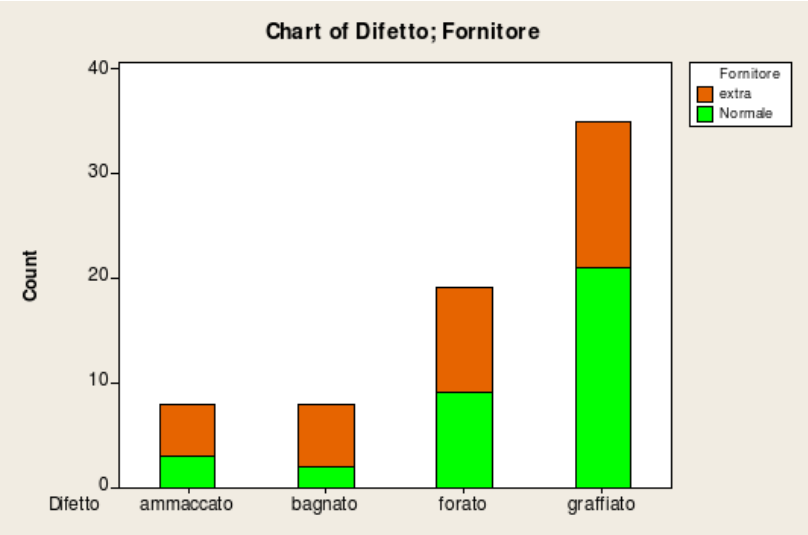
The bar widths are all equal, because they have only a symbolic meaning with qualitative data.

Bar chart examples:

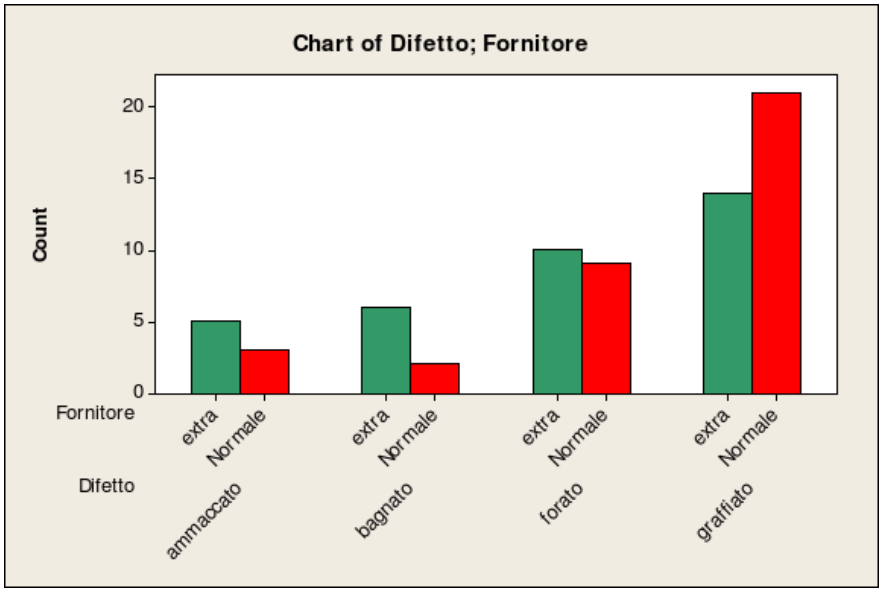


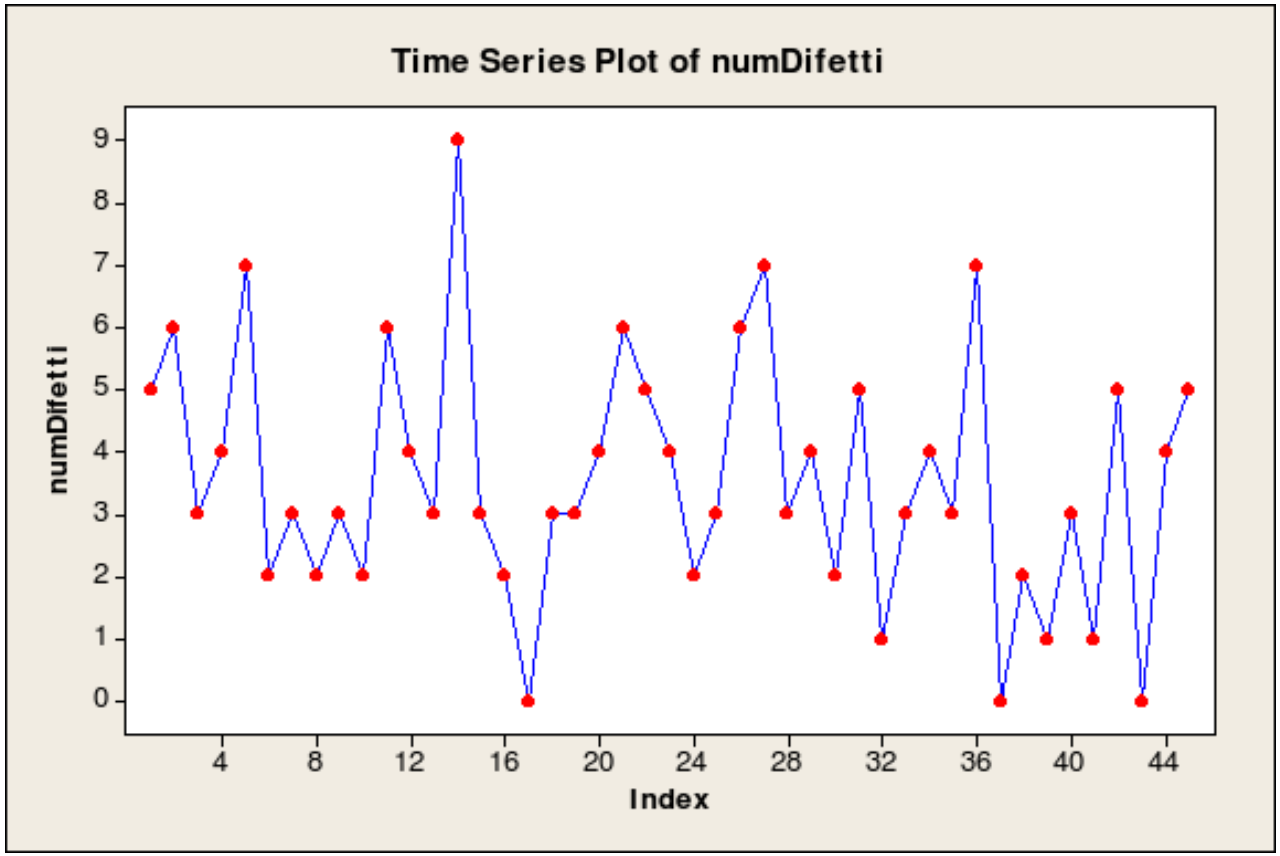
The bar chart represents discrete data.

Stacked Bar



Grouped Bars





Time Series Plot Examples:

Plots saw until now provide a visual summary of main characteristics of frequency distributions.

Graphs are more intuitive than numbers. However they are less detailed than numeric basic statistics.

Almost all descriptive information of **qualitative data** is provided by frequency tables and plots.

Objective (i.e. numeric) summaries that can be processed mathematically are required for quantitative data. An objective analysis should give same results, with same data, when done by different researchers.

Types of statistical indices which can be calculated for data description are three:

1. centrality indices;
2. variability indices;
3. shape indices.

1. Centrality indices

- ⋆ Mean
- ⋆ Median
- ⋆ Mode
- ⋆ Quartiles
- ⋆ Percentiles

2. Variability indices

- ⋆ Range
- ⋆ Interquartile range
- ⋆ Mean absolute deviation
- ⋆ Deviance
- ⋆ Variance
- ⋆ Standard deviation
- ⋆ Standard error

3. Shape indices

- ⋆ Skewness
- ⋆ Kurtosis

Measures of central tendency or measures of position provide the value around which data are distributed. The central tendency measurement is one of the most appropriate indices to summarize a data series.

The measure of central tendency is the first and more intuitive index about data distribution.

The suggested indices are mainly three: **mean, median** and **mode**.

The choice of the measure of central tendency depends from distribution features and from scale type.

The (simple) **arithmetic mean** is the most commonly used measure of central tendency.

It is called simply **mean** or average when the context is clear.

It is defined as the sum of values of all observations divided by the total count of observations.

Given a sample of n units $\underline{X} = X_1, X_2, \ldots, X_i, \ldots, X_{n-1}, X_n$ the (sample) arithmetic mean is defined via the equation:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

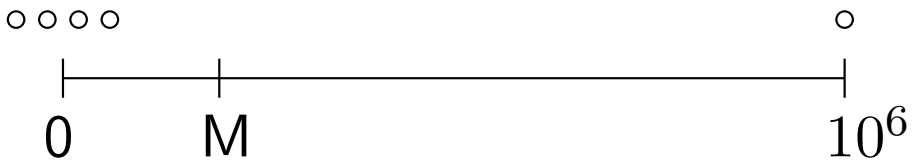
The arithmetic mean is the equilibrium point of a data series.

The arithmetic mean is the center of mass of a data series. The arithmetic mean may not accord with the intuitive notion of "middle" and it is particularly influenced by **outliers** (or extreme values). If a single value is far from the main group of values (outlier) then this value can weight considerably in the mean calculation.

Example: Calculation of average income between the following values (expressed in euros):

$$1000 \quad 1200 \quad 1500 \quad 2000 \quad 10^6$$

The arithmetic mean is 201,140 euro: it is influenced by the only high value of the series. The arithmetic mean is far from the most large group of values:



When data are grouped in classes, the representative value for each class is the central value.

The **arithmetic mean for frequency distributions** grouped in classes, is called **weighted arithmetic mean**.

Given a frequency table of n classes with:

- ⋆ central values: $X_1, X_2, \ldots, X_i, \ldots, X_{n-1}, X_n$
- ⋆ absolute (or relative) frequencies: $f_1, f_2, \ldots, f_i, \ldots, f_{n-1}, f_n$

then the (weighted) arithmetic mean is

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i}$$

Simple and weighted arithmetic mean can have several employments. This is due to some properties of the mean:

- ⋆ **the additive quantities are the most common quantities**;
- ⋆ **the arithmetic mean adjusts random errors** in the observations, it is the more accurate estimate on repeated measurements;
- ⋆ **the arithmetic mean is the easier** of algebraic means.

When quantities are not additive or when data are obtained from ratios, then other types of mean are used: the **geometric mean**, the **harmonic mean** and the **quadratic mean**.

Simple and weighted arithmetic mean can have several employments. This is due to some properties of the mean:

- ⋆ **the additive quantities are the most common quantities**;
- ⋆ **the arithmetic mean adjusts random errors** in the observations, it is the more accurate estimate on repeated measurements;
- ⋆ **the arithmetic mean is the easier** of algebraic means.

When quantities are not additive or when data are obtained from ratios, then other types of mean are used: the **geometric mean**, the **harmonic mean** and the **quadratic mean**.

Simple and weighted arithmetic mean can have several employments. This is due to some properties of the mean:

- A **trimmed mean** (similar to an adjusted mean) is a method of averaging that removes a small designated percentage of the largest and smallest values before calculating the mean.
- After removing the specified outlier observations, the trimmed mean is found using a standard arithmetic averaging formula.
- The use of a trimmed mean helps eliminate the influence of outliers or data points on the tails that may unfairly affect the traditional or arithmetic mean.

Example of trimmed mean A skating competition produces the following scores: 6.0, 8.1, 8.3, 9.1, and 9.9.

The mean for the scores would equal:

$$\frac{6.0 + 8.1 + 8.3 + 9.1 + 9.9}{5} = 8.28$$

To trim the mean by a total of 40%, we remove the lowest 20% and the highest 20% of values, eliminating the scores of 6.0 and 9.9.

Next, we calculate the mean based on the calculation:

$$\frac{8.1 + 8.3 + 9.1}{3} = 8.5$$

The **median** is the numeric value separating the higher half of a sample from the lower half. It can be found by arranging all the observations from lowest value to highest value and picking the middle one.

It is a **robust measure**, because it is **little influenced by outliers**.

Its use is **required for ordinal data or ranks**.

It is computed on the total count of observations. It is used to mitigate the effect of outliers or to consider only information provided by ranks.

Each value randomly extracted from a distribution or from a data series has the same probability to be lower or higher than the median.

The mean is the measure of central tendency used in parametric statistics. The median is the measure of central tendency used in almost all non parametric statistical tests.

To **calculate the median of a data series** it needs:

- ⋆ arranging all the observations from lowest value to highest value (or from the highest value to lowest value) and counting the total number n of observations;
- ⋆ if n is odd then the median is the middle observation, i.e. the observation in the position $\frac{n+1}{2}$;
- ⋆ if n is even then the median can be calculated as the average between the observations in positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

The median splits the ordered data series in two equivalent parts.

Mean and median coincide in a **symmetric distribution**.

When the distribution is not symmetric the median is more appropriate then the mean to describe the "middle" of data, i.e. the "typical" value of a data series.

The mean is more influenced by outliers. Extreme values shift the mean from the group of most common values making it different.

If extreme values are closer (or farther) to central values of the data series, the mean could change while the median would still remain the same.

The mean should be analyzed together with the median, which is not affected by outliers.

Example: In the previous example, about the calculation of average income in a data series, the median is much less than mean and closer to the most common values because it is not influenced by the extreme value:

$$1000 \qquad 1200 \qquad \boxed{1500} \qquad 2000 \qquad 10^6$$

The **quartiles** of a data series can be defined as follow:

Q1 = the median of the first half of the ordered series, i.e. it cuts off lowest 25% of data, or highest 75%.

Q2 = the median.

Q3 = the median of the second half of the ordered series, i.e. it cuts off lowest 75% of data, or highest 25%.

More generally, the **percentiles or quantiles** of a data series can be defined.

For example, the 37-th percentiles (P_{37}) is the value that cuts off lowest 37% of data.

The P_{100} is the value that cuts off the 100% (more or less!!) of data.

Then:

- $\boldsymbol{P_0 = Min}$
- $\boldsymbol{P_{25} = Q1}$
- $\boldsymbol{P_{50} = Median}$
- $\boldsymbol{P_{75} = Q3}$
- $\boldsymbol{P_{100} = Max}$

Along with the arithmetic mean and the median, also **quartiles** and **percentiles** can be used as centrality indices, altough they concern data variability.

These indices are important to find anomalous cases. Moreover, the calculation of P_{90}, P_{95} or P_{99} values provides more useful information than just the maximum of the data series.

The **mode** is the value that occurs most frequently in a data set.

It is not influenced by outliers; however it is used only for descriptive aims because **it is less steady and less objective than other measures of central tendency**.

Indeed, when the same continuos data are grouped, the mode can be different for different class widths.

When the variable is discrete, and then classes are unambiguous, the value of the mode is more objective; when the variable is continuos, the choice of classes makes the mode a subjective index, depending from the number of classes.

When data are continuos, the **uniform distribution** within classes is hypothesized to find the mode of data distribution.

Frequency distributions with one mode only are called **unimodal distributions** while distributions with two or more modes are called **bimodal or multimodal distributions**.

Multimodal distributions can be the result of few observations or data rounding. They are usually due to the overlap of more distributions with different central tendencies.

When data distribution shows two or more modes, the researcher should suspect that data are not homogeneous but that they come from different set with different central tendencies.

When data distribution shows two or more modes, **the calculation of the general data mean might be incorrect because the fundamental assumption that all data come from the same universe or population, with only one central tendency, may not be true**.

There are other measures of central tendency, rarely used in statistics:

- ⋆ mean interval;
- ⋆ interquartile mean;
- ⋆ midhinge;
- ⋆ Tukey's trimean.

The **second important feature** of a data distribution is the **dispersion or variability**.

The dispersion defines if values are more or less concentrated around the central tendency.

The first and the simplest measure introduced to describe the data variability is the **range**. It is defined as the **difference between the maximum and the minimum** in a data series.

Range = maximum - minimum

It takes only positive values and it is an intuitive and simple statistical tool, especially when the data are ordered.

The drawbacks of this measure are:

- ⋆ **it doesn't show the variation inside the data set**, notably to detect outliers;
- ⋆ **it is dependent from the number of observations**. When the number of data increases the range tends to increase too.

The range is an inefficient misure of data dispersion.

To compare more distributions, it requires samples with equal sizes, a limiting requirement for research and data analysis.

To illustrate one of drawbacks of the range as variability index, the following example is given.

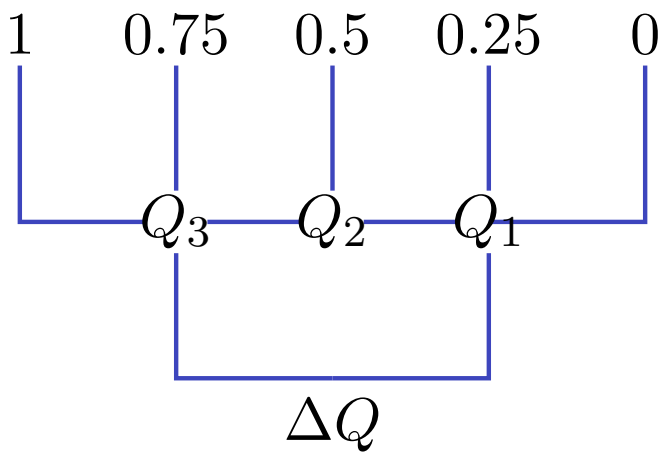
Example:

The figure below shows 4 data series with the same range:

The 4 data series have the same **range**, but this value is due only to the extreme values. The range doesn't provide information about distribution "within" samples.

The **interquartile range** is the difference between the third and first quartile. It doesn't use extreme values.

The interquartile range is more stable compared to the range.



The **quantiles** are often used as measures of non-central tendency because a given quantile always detect the same fraction of observations. Quantiles are mainly used for descriptive purposes.

Anomalous values (lower than P_5 or higher than P_{95}) **may depend by special causes**. Some times can be spent to investigate this special anomalous values.

Anomalous values very often depend from special causes, that require more care.

When the distribution shape is unknown or the distribution is strongly asymmetric, quantiles provide easy and robust indications to discriminate the **most frequent** values, that can be considered **"common"**, from the **less frequent** (or **"anomalous"**) ones.

Mean deviations are the most appropriate measure of variability for a data series. The sum of mean deviations is always zero, by definition, because the mean is the centre of mass of the distribution. For this reason a transformation is required. It can be implemented in two ways:

- ⋆ **the mean absolute deviation**;
- ⋆ **the sum of squares (of mean deviations) or deviance**.

The **mean absolute deviation** (S_m) for non categorized data is defined by:

$$S_m = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

where:

x_i represents the i-th observed value;

$\overline{x}$ represents the sample mean;

n represents the total count of observations.

For categorized data it is defined by:

$$S_m = \frac{\sum_{i=1}^{n} |x_i - \overline{x}| n_i}{n}$$

where:

x_i represents the i-th class;

n_i represents the count of observations of i-th class;

Some **non parametric statistical tests** use the **median absolute deviation**. It is the mean of the absolute differences between the single values and their median. Formulas are identical to the previous ones, replacing the mean with the median.

The median is the value that minimize te absolute differences. For this reason, the median absolute deviation is always less than the mean absolute deviation. The two values are equal if the distribution is symmetric and so mean and median coincide.

The **deviance** or **Sum of Squares (SS)** of mean deviations is the **basic measure of data variability**, used in all parametric statistics.

$$SS = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

or

$$SS = \sum_{i=1}^{n} (x_i)^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i \right)^2$$

For categorized data it is defined by: $SS = \sum_{i=1}^{n} \left(\overline{x_i} - \overline{\overline{x}}\right)^2 n_i$

where:

$\overline{x_i}$ represents the center value of i-th class;

$\overline{\overline{x}}$ represents the sample mean.

n_i represents the count of observations of i-th class.

The deviance value depends from two distribution features: mean deviations and total count of observations.

The first one is a measurement of data variability and it is the effect that should be estimated.

The second one is a limiting factor for the use of the deviance.

The variance or mean of squares (MS) is an "average deviance", i.e. a deviance divided by the total count of observations.

The **variance of a population**, which is indicated with $\boldsymbol{\sigma^2}$, is obtained dividing the deviance by N, **the count of all population units (it can be infinity)**:

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

where μ is the "true" population mean.

The **sample variance**, which is indicated with s^2, is obtained dividing the deviance by $n-1$, the so called number of **degrees of freedom (df)**:

$$\hat{V}_x = s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

Of course, when n (the sample size) is large, differences between sample variance and population variance are small; when n is small, differences are significant.

The sample variance is always used when **sample data are used to know the population characteristics**, i.e. when **inferential statistics** is used.

s^2 is usually used as an estimate of σ^2.

It is not easy to explain why **the deviance must be divided by $n-1$**.

The simplest answer is based on the fact that **$n-1$ is the number of indipendent observations**.

The sum of mean deviations is 0, consequently one value of a series of indipendent values is constrained: it is not free to take any values, when the other $n-1$ values are known.

The number of degrees of freedom is given by the number of observations minus the number of costants that are already calculated or parameters that are already estimated from data.

In the case of the variance, the sample mean is the constant used to compute mean deviation: degrees of freedom are so $n-1$.

Dividing by $n - 1$ affects the calculated variance, particularly when the sample size is small: in this case the variance value is amplified.

While the **mean is a linear value**, the **variance is a squared value**. To compare the variance with the mean, the variance should be brought back to a linear value. The unit of measurement of the variance is the square of the original unit of measurement.

Unlike range, the variance considers also the **dispersion** of observations within the **range**.

The measurement of variance is amplified by observations far from mean.

The **standard deviation is the square root of the variance**. It is indicated with σ when considering the population standard deviation and with s when considering the sample standard deviation. It is a good measure of variability and it allows to bring back the variability index to the original unit of measurement.

The standard deviation is a measure of distance from mean and so it is always positive. It is an average measure of the dispersion around the mean of observed values of random variable.

The standard deviation for a sample can be calculated, from the data series, as:

$$s_x = \sqrt{\hat{V}_x} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

The deviance is the best variability index from a strictly **mathematical** point of view, followed by the variance and then by the standard deviation.

Given two samples, x e y, indipendent each other, a sample z can be defined as: $z = x + y$.
In this case, it can be shown that the variance of z is the sum of variances of x and y:

$$V_z = V_x + V_y$$

This relationship is not true for the standard deviation:

$$S_z \neq S_x + S_y$$

$$S_z = (V_x + V_y)^{0.5}$$

The variance is the true measure of variability, as in a complex system each variation imposed to the system generates a degree of disorder (entropy) proportional to the square of the variation.

Example.

A firm has two shops, in two different cities: Padova and Milano. The series of daily sales of each of two shops are avalaible. The two series can be hypothesized as indipendents.

The series of daily total sales of the firm can be easily calculated as the sum of sales of Padova and Milano.

Day	1	2	3	4	5	6	7
Milano	1010.46	997.01	1012.31	1004.02	992.07	1000.30	1000.21
Padova	795.03	796.74	793.34	818.11	794.01	793.27	807.63
Firm total	1805.49	1793.75	1805.65	1822.13	1786.08	1793.57	1807.84

$V_{Milano} = 51.8 \qquad V_{Padova} = 91.1 \qquad V_{Milano} + V_{Padova} = 142.9$

$V_{Total} = 142.9$

Another variability index often used in statistical computation is the **standard error** (se).

$$se_x = \overline{s}_x = \frac{s_x}{\sqrt{n}}$$

The standard error is a measure of the dispersion of the sample mean around the true mean.

The standard error is a measure of distance from the mean and therefore it has always a positive (or null) value.

Shape indices refer typically to two distribution features: skewness and kurtosis.

The analysis of distribution shape still use rudimentary*.5cm The analysis of distribution shape still use rudimentary indices and definition itself is often ambigous.

Shape indices are usually not object of inference, they are used as a description of the distribution shape.

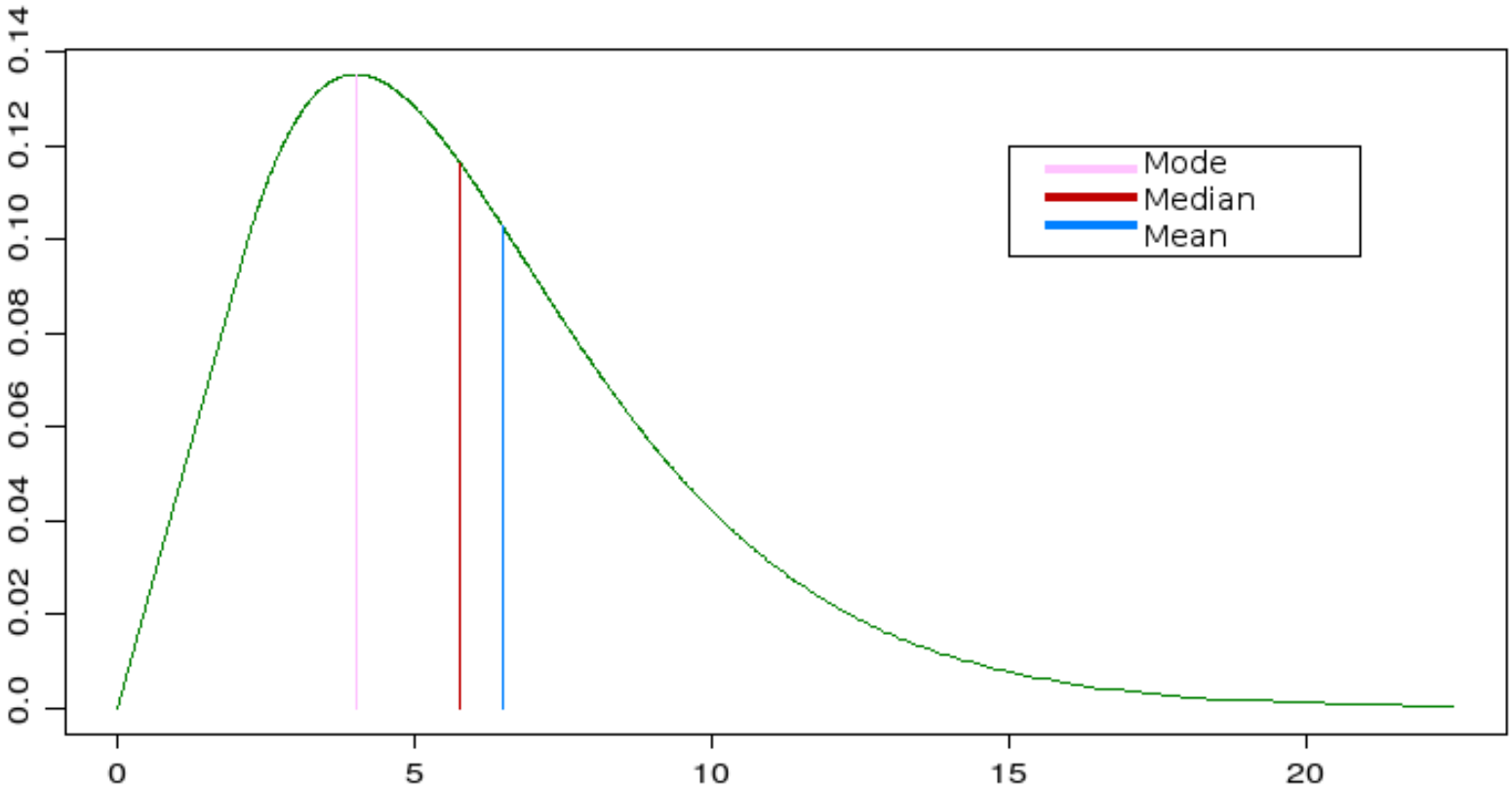
If an unimodal distribution is symmetric, then mean, median and mode coincide.

When farthest values from mean are the largest values, then there is a **right asymmetry**. In this case, the **mode** is smaller than the **median** that is smaller than the **mean**.

When farthest values from mean are the smallest values, then there is a **left asymmetry**. In this case, the **mean** is smaller than the **median** that is smaller than the **mode**.

The coincidence of three measurements of central tendency is a necessary but not sufficient condition to have a symmetric distribution.

Right asymmetry



A small number of observations or an inadequate categorization obtained from grouping data may produce a form of asymmetry. This is usually due to too few classes; in this case the asymmetry is false and must be distinguished from the **true asymmetry that can exist only in the real distribution of the population**.

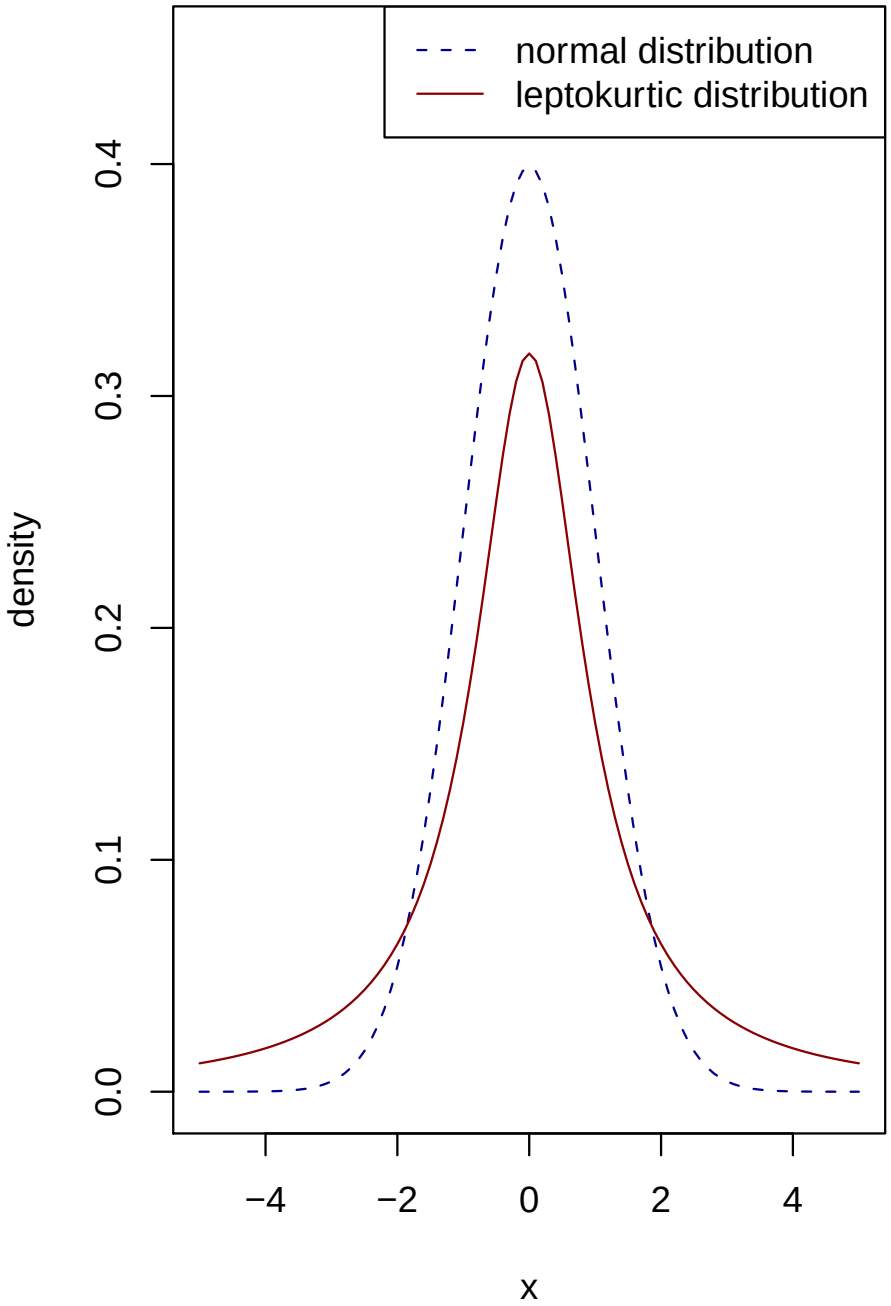
A basic property of skewness indices of a distribution is that they should be null if, and only if, the distribution is symmetric.

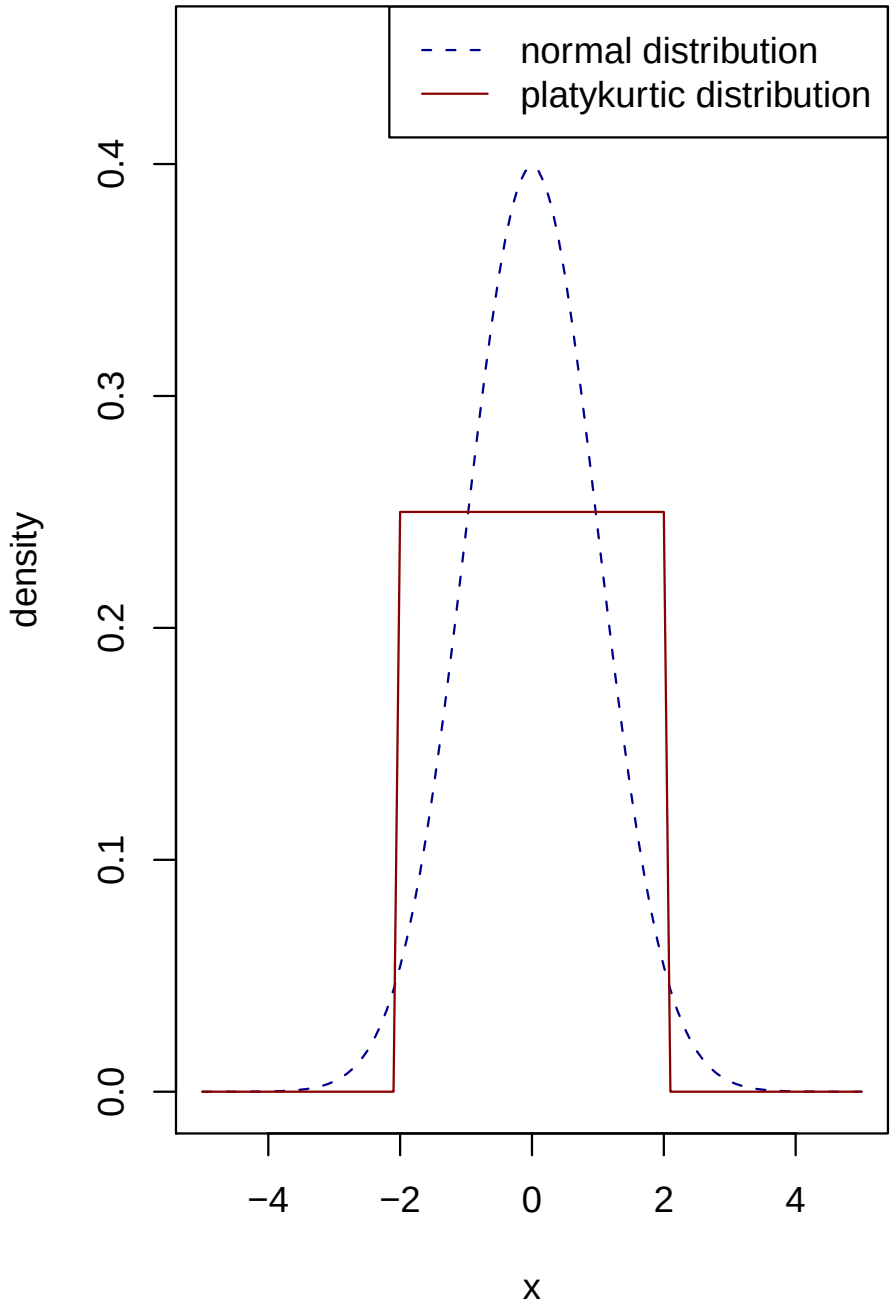
However, the proposed index doesn't have this property: if the distribution is symmetric its value is zero, but it can be null also for asymmetric distributions.

The **kurtosis** is a measure of the heaviness of the tails of a distribution. It measures differences of a symmetric distribution from a gaussian curve.

With respect to kurtosis, an **unimodal symmetric distribution** can be:

- **mesokurtic**, if its shape is like the normal distribution;
- **leptokurtic**, if it has a more acute peak around the mean and fatter tails;
- **platykurtic**, if it has lower, wider peak around the mean (that is, a higher probability than a normally distributed variable of values near the mean) and thinner tails.

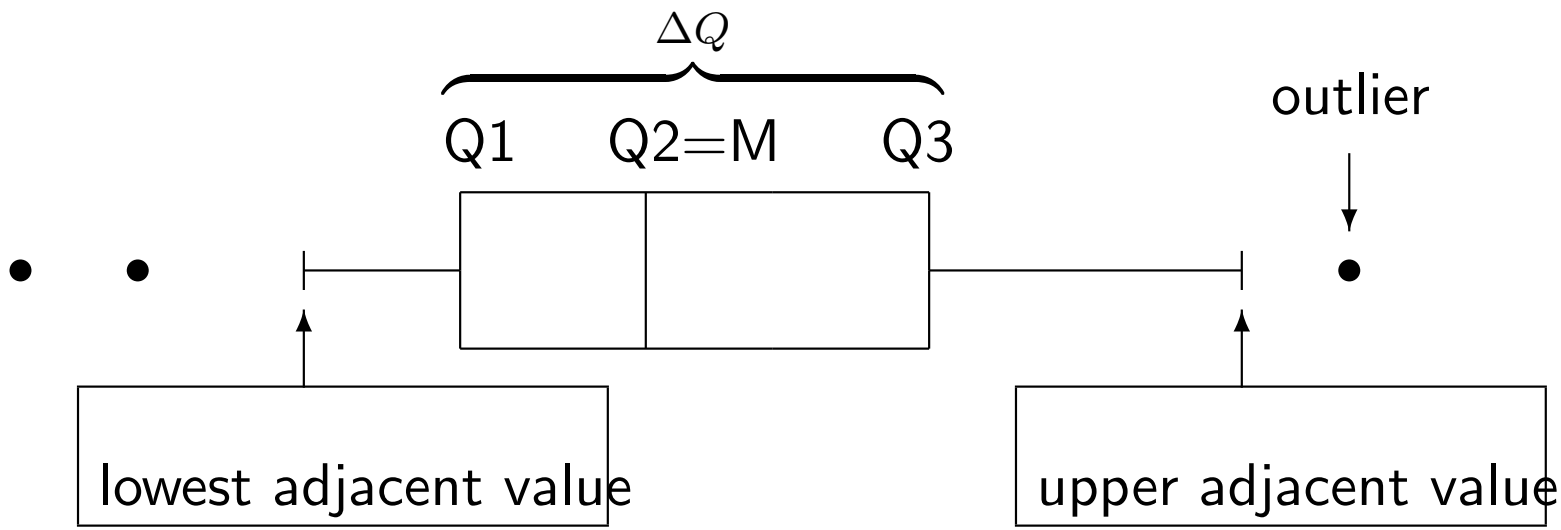

normal distribution
leptokurtic distribution
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x


normal distribution
platykurtic distribution
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x

The box-and-whisker diagram or box-and-whisker plot or, simply, the **box plot** is a convenient way of graphically depicting a distribution of numerical data. The popularity of the box plot (or boxplot) has grown with the implementation in statistical software. The plot is easy to draw and read.

A box plot shows three fundamental characteristics of a statistical distribution:

- ⋆ **the data spread or variability**, with respect of mean and/or median;
- ⋆ **the skewness**;
- ⋆ **the presence of outliers**.



The **vertical line** within the box is the **median**.

The **two external vertical lines** that delimit the box are the first (Q_1) and the third (Q_3) quartiles.

The difference between the third and the first quartiles is called **interquartile range**. It provides a measure of dispersion of the distribution. By definition, the 50% of observations are included in the interquartile range.

A small interquartile range means that the 50% of observed values is close to the median. The size of the interquartile range increases when data dispersion (variability) increases.

The boxplot provides also information about the shape of the distribution (skewness): if the distance of Q1 from the median differs from the distance of Q3 from the median then the distribution is asymmetric.

The lines that start from the box borders and stop with a short vertical line are the so called "whiskers". The limits given by two whiskers define the "non-outlier" range. Any data not included between the whiskers shall be marked as an outlier with a dot.

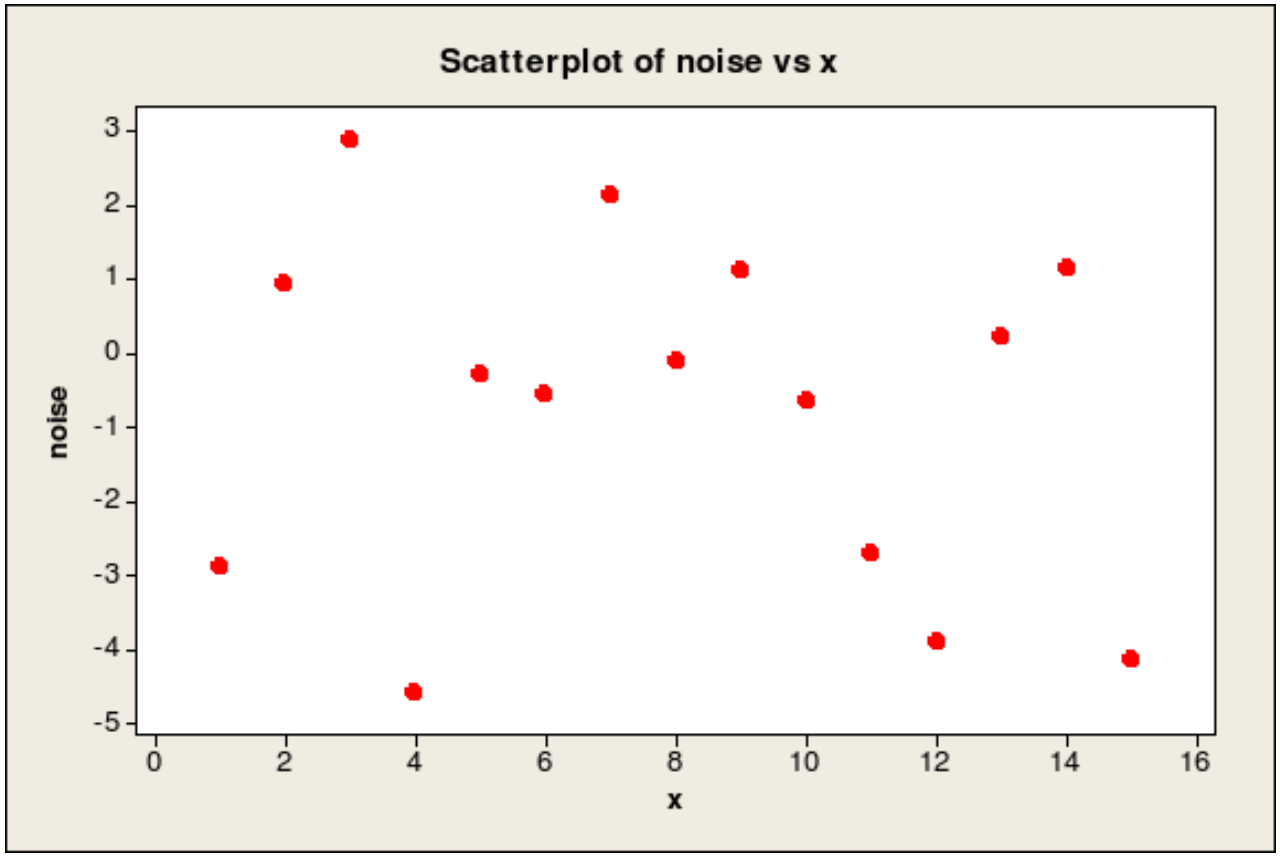
Statistical indices proposed until now work all on a single variable.

Given two or more variables, the type and the intensity of the relationship between variables can be analyzed.

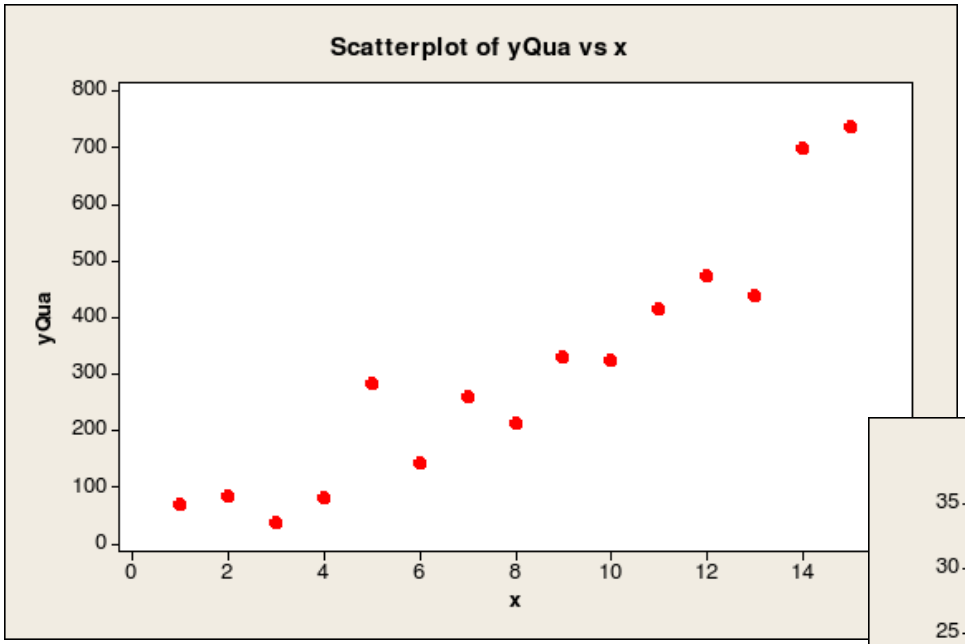
The classical scatterplot (or xyplot) is the most common plot type to show the relationship between two variables.

Like all graphic tools, it doesn't provide an exhaustive description but it gives hints about the analyzed phenomenon.

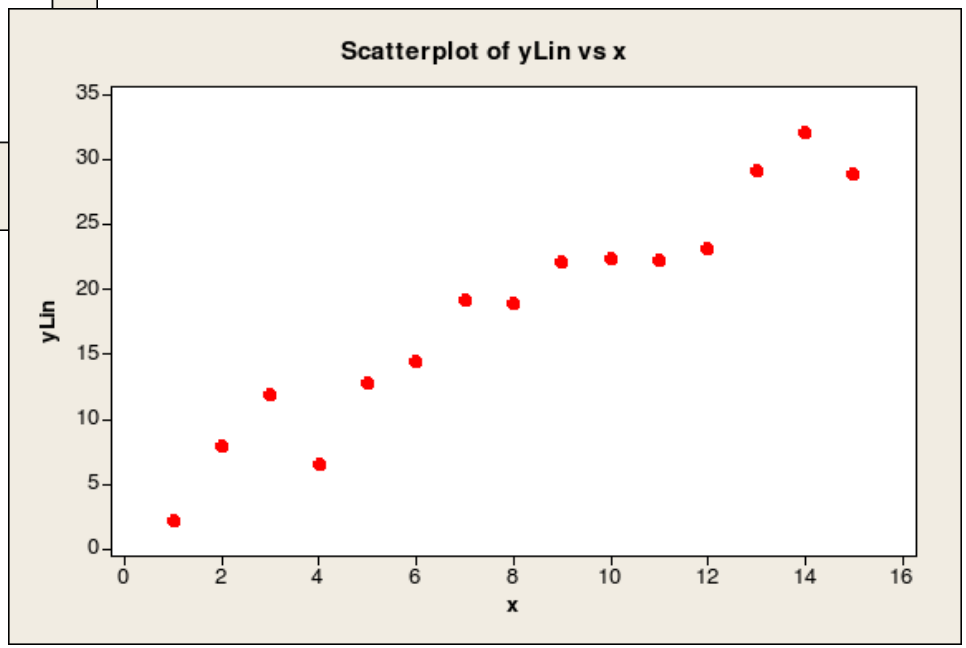
If a relationship between variables exists then the scatterplot must show a tendency of the values in variable in the y-axis to vary along with the values in variable in the x-axis.



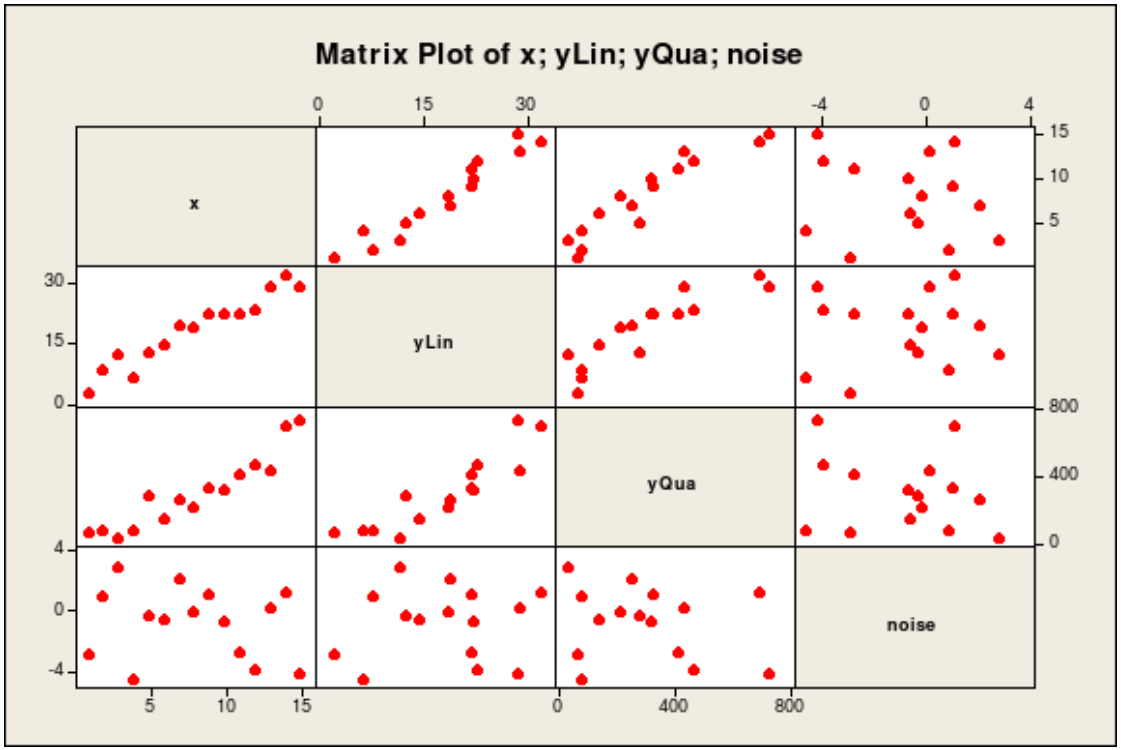
The plot shows that there is not a relationship between variables or, in other words, it exists a random relationship.



Plots show a relationship between variables.



In both cases the relationship is positive. When x increases also y increases, on average.



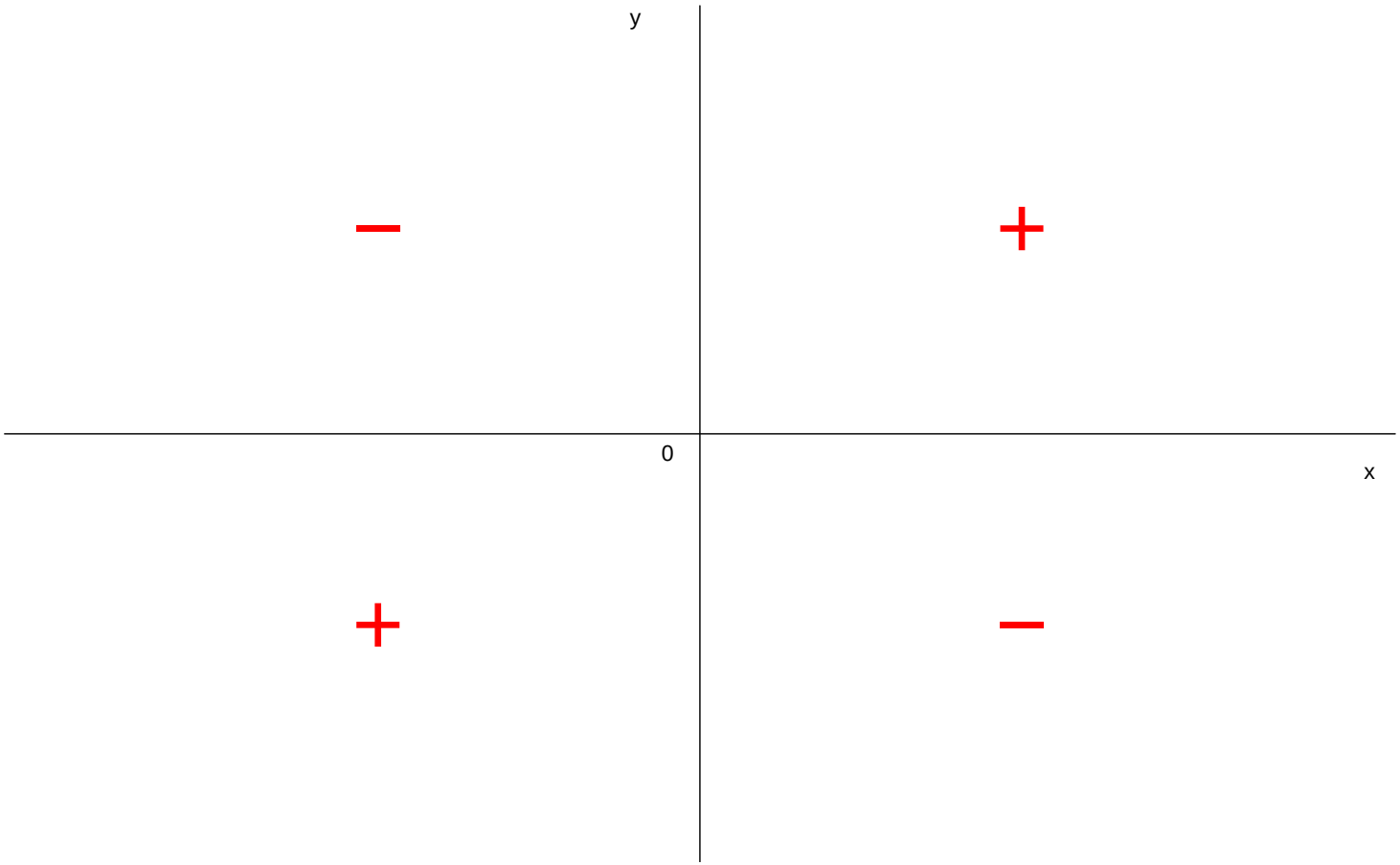
A **matrix plot** draws more scatterplots in only one display. It is useful to search for possible relationships between a **set of variables**.

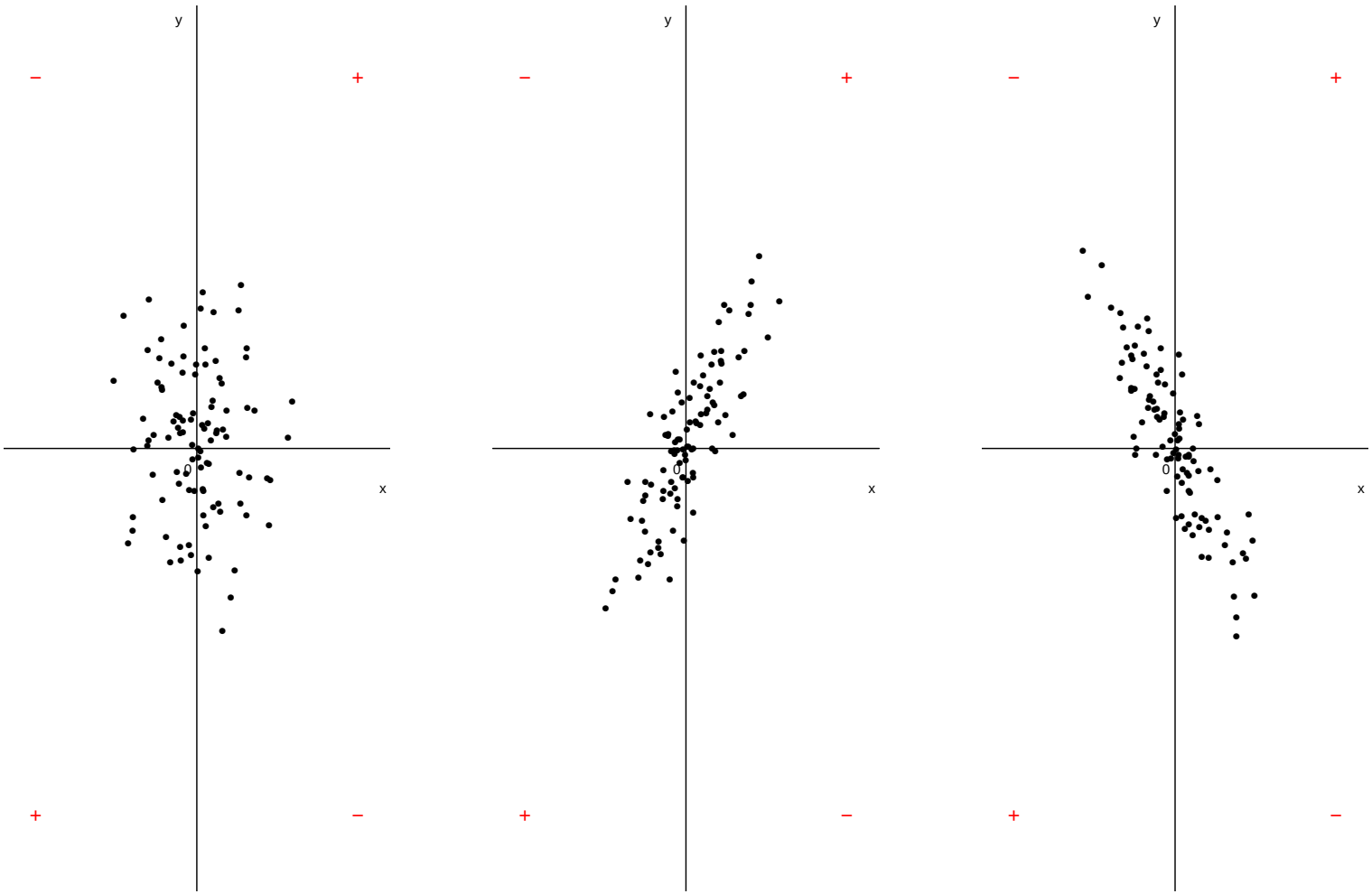
When a **linear relationship** between two variables exists, then the index that highlights this relationship is the **covariance** index.

$$Cov_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

The covariance can assume:

- ⋆ **positive** values, when an increasing linear relationship exists between the two variables;
- ⋆ **negative** values, when a decreasing linear relationship exists between the two variables;
- ⋆ close to **zero** values, when a linear relationship doesn't exist between the two variables.


y
0
x
−
+
+
−


y
x
0
−
+
+
−
y
x
0
−
+
+
−
y
x
0
−
+
+
−

- The graph of Page 87 recalls that the product between x and y is:
  - a positive number if x and y are both positive (first quarter, top right);
  - a negative number if x is negative and y is positive (second quarter, top left);
  - a positive number if x and y are both negative (third quartrer, bottom left);
  - a negative number if x is positive and y is negative (quarto quadrante, bottom right).

- The graphs of Page 88 show a state in which:
  - there is no correlation between x and y;
  - there is positive correlation between x and y;
  - there is negative correlation between x and y.

The following inequality can be demonstrated about the covariance:

$$-s_x s_y \leq \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1} \leq s_x s_y$$

As a consequence, the **correlation index** may be used as a standardized covariance:

$$Cor_{xy} = \frac{Cov_{xy}}{s_x s_y}$$

The correlation index may only assume values between -1 and 1:

$$-1 \leq Cor_{xy} \leq 1$$

Notes.

- ⋆ Correlation and covariance are indices that measure only linear relationships. Quadratic, cubic, sinusoidal, ... relationships may also exist between two variables.
- ⋆ The correlation index can only highlight the linear component of a relationship.
- ⋆ The presence of correlation may not be due to a cause-and-effect relationship between two variables.