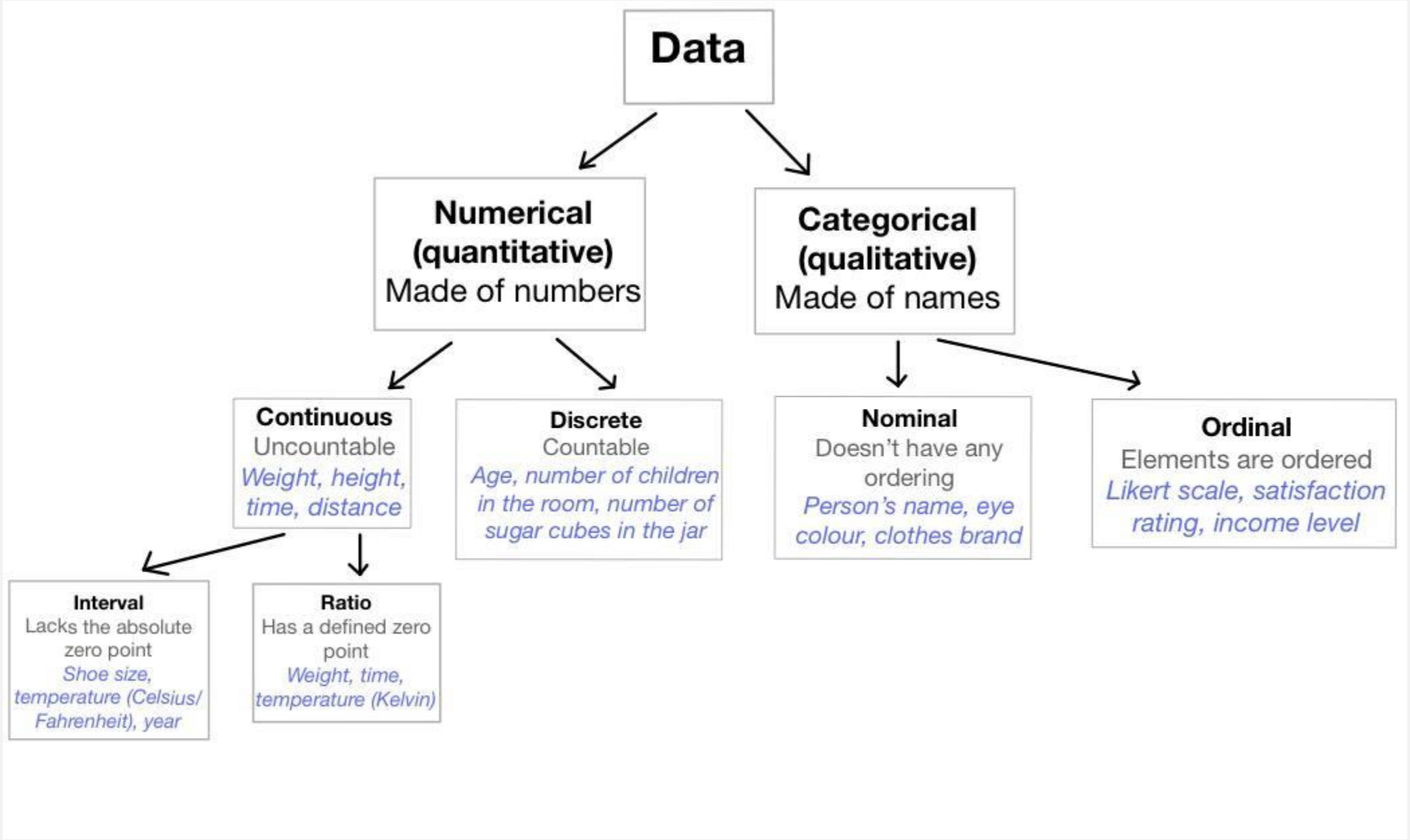
- **ELEMENTS OF STATISTICS**


Data
Numerical (quantitative)
Made of numbers
Categorical (qualitative)
Made of names
Continuous
Uncountable
Weight, height, time, distance
Discrete
Countable
Age, number of children in the room, number of sugar cubes in the jar
Nominal
Doesn't have any ordering
Person's name, eye colour, clothes brand
Ordinal
Elements are ordered
Likert scale, satisfaction rating, income level
Interval
Lacks the absolute zero point
Shoe size, temperature (Celsius/ Fahrenheit), year
Ratio
Has a defined zero point
Weight, time, temperature (Kelvin)

NUMERIC RECORD DATA

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an n-by-d data matrix, where there are n rows, one for each object, and d columns, one for each attribute

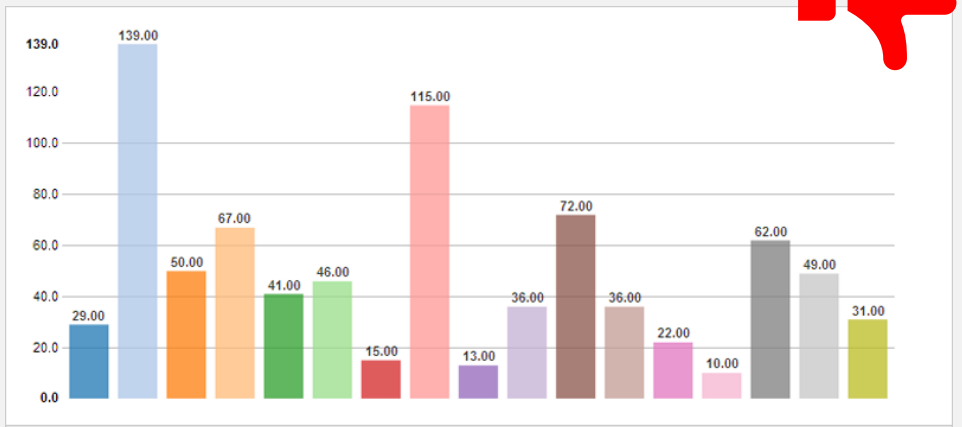
Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

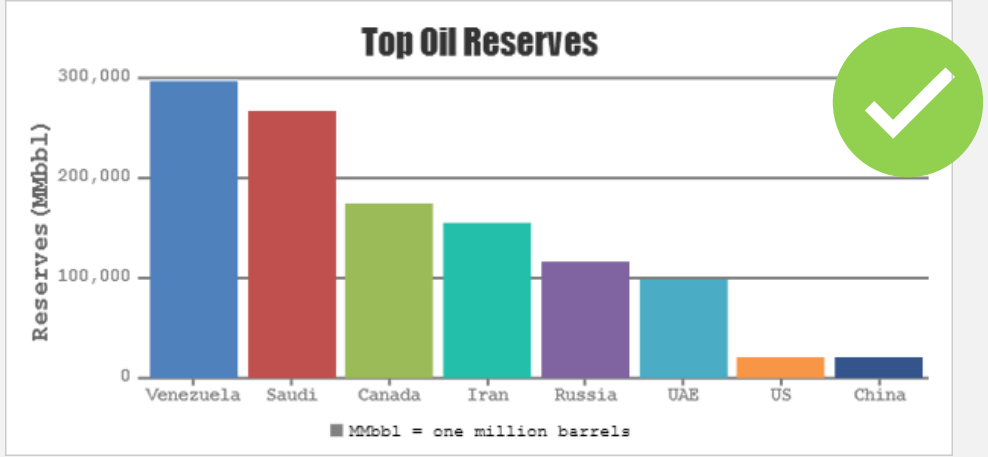
CATEGORICAL DATA

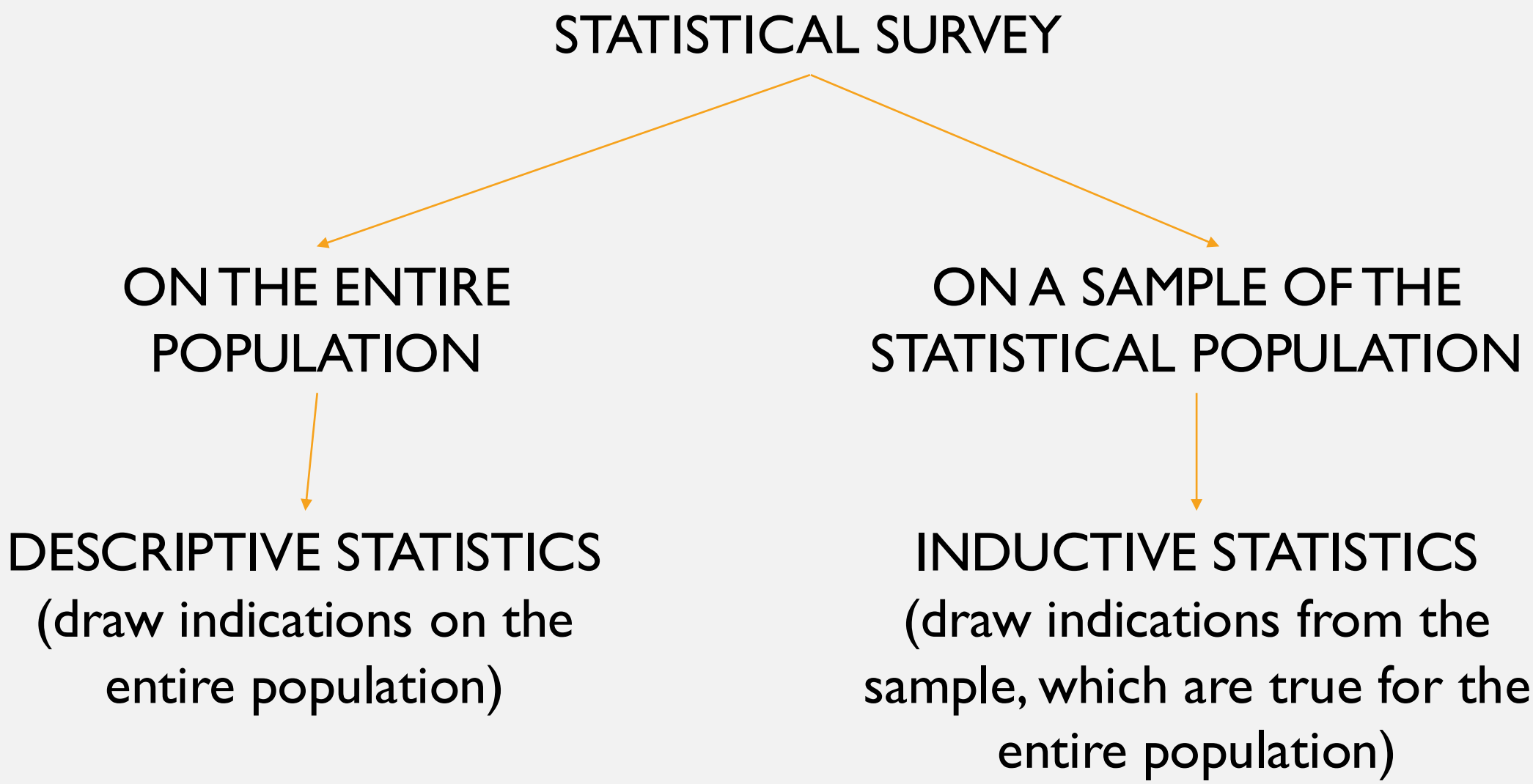
- Data that consists of a collection of records, each of which consists of a fixed set of categorical attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	Low	No
4	Yes	Married	High	No
5	No	Divorced	Medium	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Medium	Yes
9	No	Married	Medium	No
10	No	Single	Medium	Yes

GOOD AND BAD VISUALIZATION






STATISTICAL SURVEY
ON THE ENTIRE POPULATION
ON A SAMPLE OF THE STATISTICAL POPULATION
DESCRIPTIVE STATISTICS (draw indications on the entire population)
INDUCTIVE STATISTICS (draw indications from the sample, which are true for the entire population)

Population, statistical unit, statistical sample

- **Statistical population**: set of elements to which the statistical investigation refers:
  - opinions of Americans regarding a new presidential election: all USA citizens
  - genes overexpressed in individuals affected by obesity: all individuals with obesity
  - …

- **Statistical unit**: each element of the statistical population, the smallest unit from which data are collected:
  - A citizen, an individual with obesity….

- **Statistical sample (sample)**: any set of statistical units drawn from the entire population. A sample is therefore a subset of measurements selected from the population
  - 50 individuals with obesity-related conditions (randomly selected).

Random variable

- The *collective phenomenon* manifests according to different modalities across the various statistical units; therefore, we shall refer to it as a random variable.

- The value assumed by the random variable in a given statistical unit shall be referred to as an **observation**.

Example:

- **random variable**: expression level of gene AAA
- **observation**: person X's gene AAA has an expression level equal to 12.3, person Y's gene AAA has an expression level of 10.2, person Z's gene AAA….

Quantitative or qualitative variable

- **Quantitative variable**: when it takes numerical values:
  - **Continuous**: it assumes continuous values within an interval (a person's weight and height, intensity levels of samples on microarray, level of gene expression, etc.)
  - **Discrete**: it assumes discrete values such as number of samples, number of over-expressed genes, number of patients, etc.

- **Qualitative variable**: when it takes non-numerical values
  - **Ordinal**: the data are ordered (good–average–poor, cold–lukewarm–hot…)
  - **Categorical**: male/female, phenotype, groups of diseased/healthy patients, etc.

The data table

- The coded data from a statistical survey conducted on n **statistical units** with reference to p **variables**, are collected in a table that is referred to as the "**data matrix**"

N.	sex	Academic qualification	age	weight	N. hospitalizations
1	M	Lower secondary school	36	65	3
2	F	Degree	45	70	1
…	…	…	…	…	…
N	F	High school	60	55	6

Data analysis

- When the dataset is large, direct analysis of the matrix does not allow the salient aspects of the phenomenon to be grasped immediately. It is therefore necessary to obtain a synthesis through a **statistical processing of the data**

STATISTICAL INDECES

- to **synthesize** a certain characteristics
- to **compare** different situations

STATISTICAL INDECES

CENTRAL TENDENCY
- mean
- median
- mode
- quantiles
- percentiles

DISPERSION
- range of variation
- average absolute deviation
- variance
- standard deviation
- coefficient of variation

SHAPE
- coefficient of asymmetry
- coefficient of kurtosis

Absolute and percentage frequencies

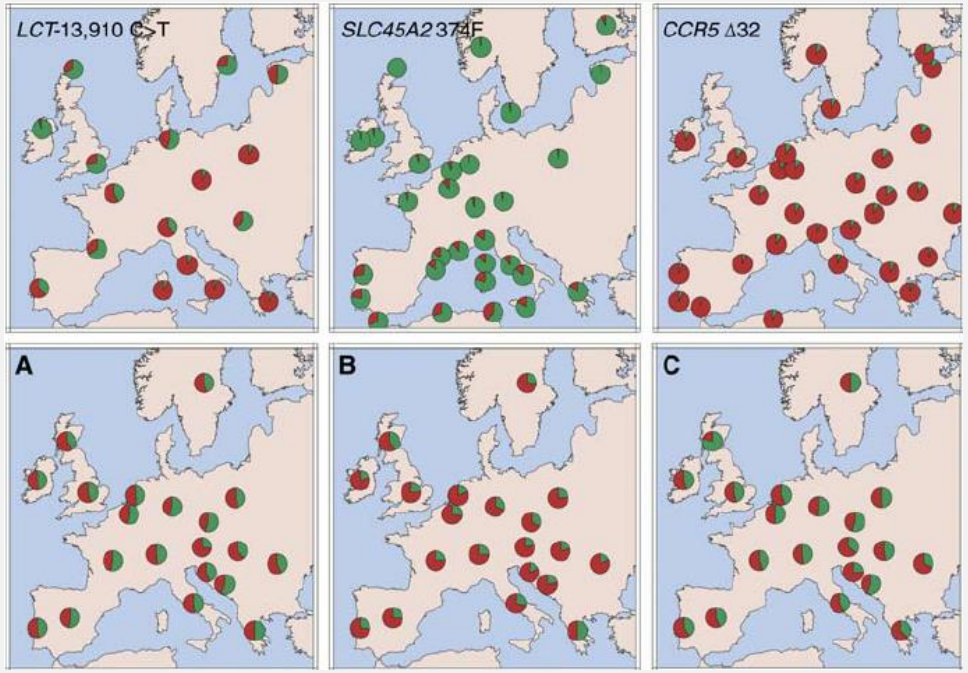
- When the sample for which we intend to describe the statistical variables is very large, rather than considering all values, one may report only distinct values and indicate, for each value, how many times it occurs.
- Let $Y = (y_1, y_2, \ldots, y_N)$ be a discrete statistical variable. We define as **modalities** the distinct values among $y_1, \ldots, y_N$ and as **absolute frequency** of a modality the number of times it is observed in the expression of the statistical variable.

Y=(60, 80, 92, 100, 83, 96, 74, 63, 80, 100, 90, 75, 74, 92)

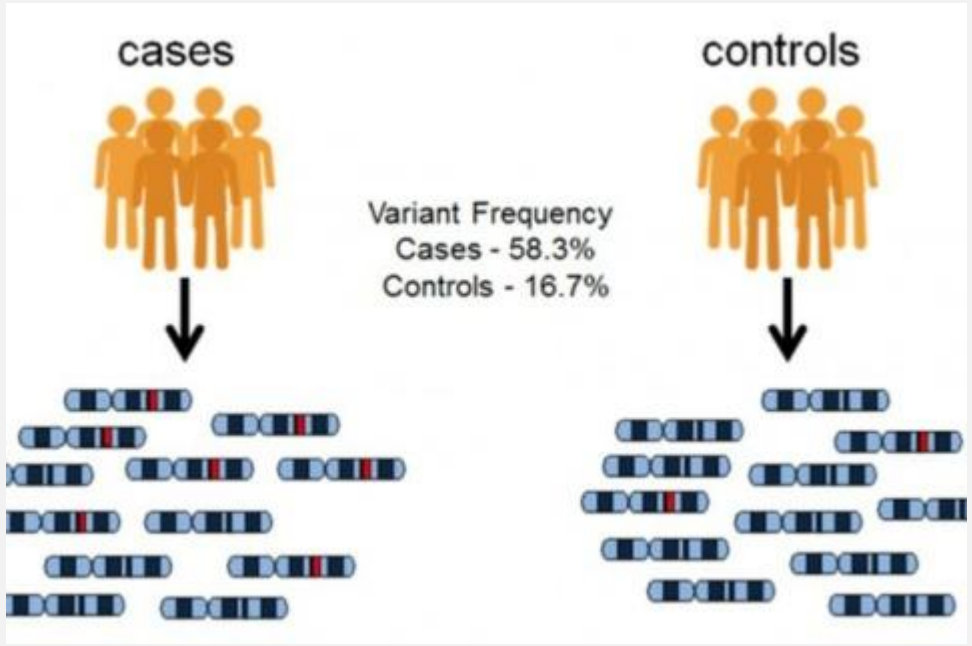
Modality	60	80	92	100	83	96	74	63	90
Absolute frequency	1	2	2	2	1	1	2	1	1

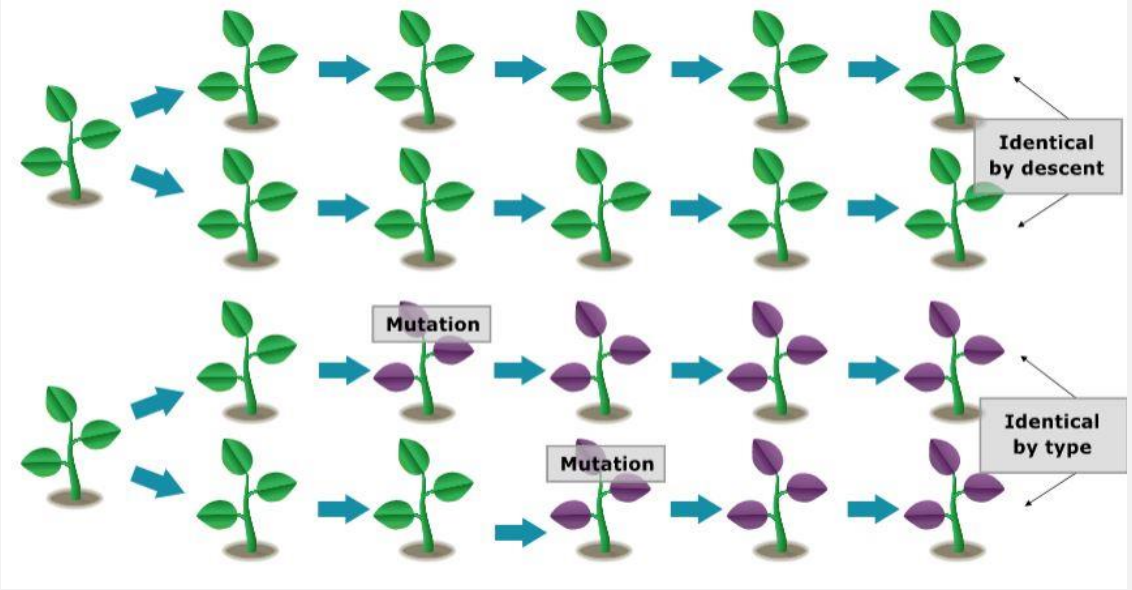
Absolute and percentage frequencies

Let Y be a statistical variable, and f the absolute frequency of modality z. We define the **relative frequency** of modality z as the ratio f/N, where N is the number of elements in the sample. The **percentage frequency** is obtained by multiplying the relative frequency by 100.



Maps showing the frequency distribution of individual genetic variants





Mean

- **Mean of a population**: sum of all values of the population variables divided by the number of units in the population (N)

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Where N is the number of elements in the population, X_i is the i-th observation of the variable X_i

- **Mean of a sample**: sum of all values of the variables in a subset of the population divided by the number of units in that sample (n)

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

Where n is the number of elements in the sample, X_i is the i-th observation of the variable X_i

Given the following set of measurements of gene expression levels:

55.20	18.06	28.16	44.14	61.61	4.88	180.29	399.11	97.47	56.89	271.95	365.29	807.80

mean of the population:

$$\mu = \frac{\sum_{i=1}^{13} 55.20 + 18.06 + 28.16 + 44.14 + 61.61 + \ldots + \ldots + 807.80}{13} = \frac{2390{,}85}{13} = 183.9115$$

mean of the sample: (55.20; 18.06; 28.16; 44.14):

$$\overline{X} = \frac{55.20 + 18.06 + 28.16 + 44.14}{4} = \frac{145.56}{4} = 36.39$$

The mean of any sample $\bar{X}$ may differ substantially from that of the entire population μ. The larger the sample, the closer the sample mean will be to the population mean

Mean

- **Weighted mean of a population**: a weight is assigned to each variable; all variable values are summed, each multiplied by its weight, and the resulting total is divided by the sum of the weights

$$\mu = \frac{\sum_{i=1}^{N} p_i X_i}{\sum_{i=1}^{N} p_i}$$

- The expected value of a variable X, denoted by $E[X]$, is defined as the mean of X computed over a large number of experiments

Exam	Grade	Credits (cfu)
Physics	21	7
Chemistry	25	10
Mathematics	26	6
Biology	24	5
…	…	…

$$\text{weighted mean} = \frac{(\text{grade x cfu}) + (\text{grade x cfu}) + (\text{grade x cfu}) + (\text{grade x cfu})}{(\text{cfu} + \text{cfu} + \text{cfu} + \text{cfu})}$$

$$\text{weighted mean} = \frac{(21 \text{ x } 7)+(25 \text{ x } 10)+(26 \text{ x } 6)+(24 \text{ x } 5)}{(7+10+6+5)} = 24.04$$

Mode

- The **mode** is the most frequent value of a distribution, or more precisely, the most recurrent category of the variable (i.e., the one associated with the highest frequency).

962	1005	1003	768	980	965	1030	1005	975	989	955	783	1005

- The mode of this sample is 1005, as it occurs 3 times.

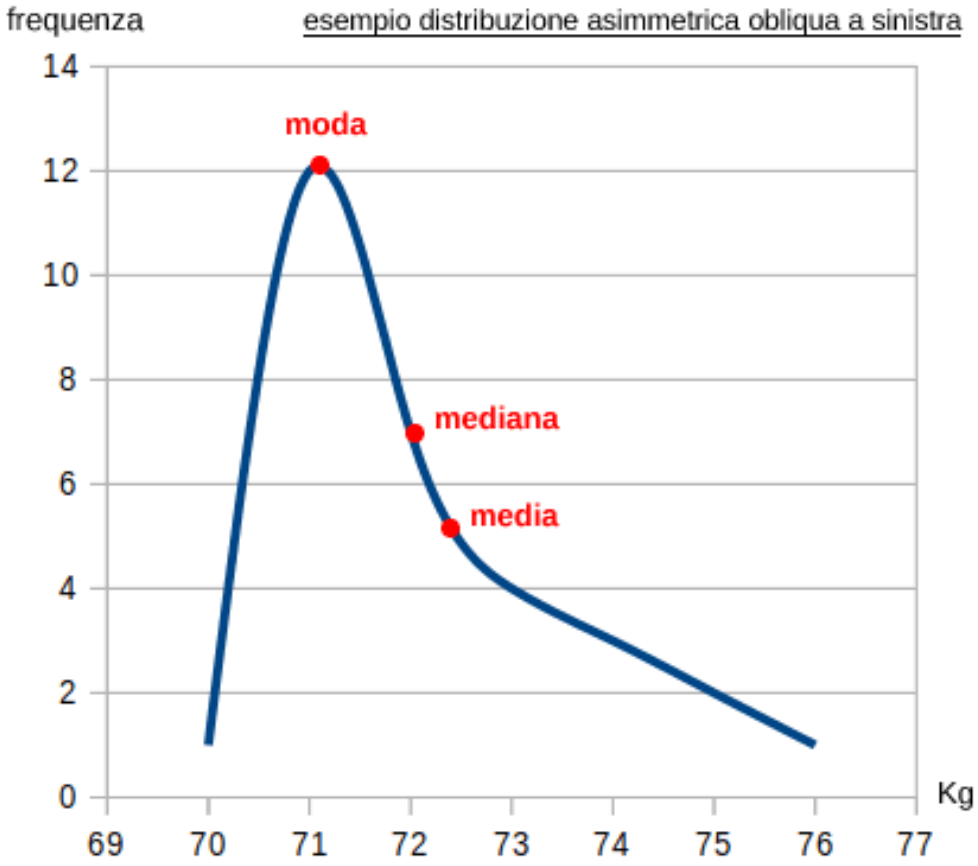
Characteristics of the mode:

- it is used solely for descriptive purposes, because it is less stable and less objective than other measures of central tendency
- to identify the mode of a distribution, graphical methods can be used, such as histograms
- it may differ within the same data series when distribution classes (intervals) of different widths are formed
- to identify the mode within a frequency class, without knowing how the data are distributed, the assumption of a uniform distribution is adopted.

kg	frequenza
70	1
71	12
72	7
73	4
74	3
75	2
76	1

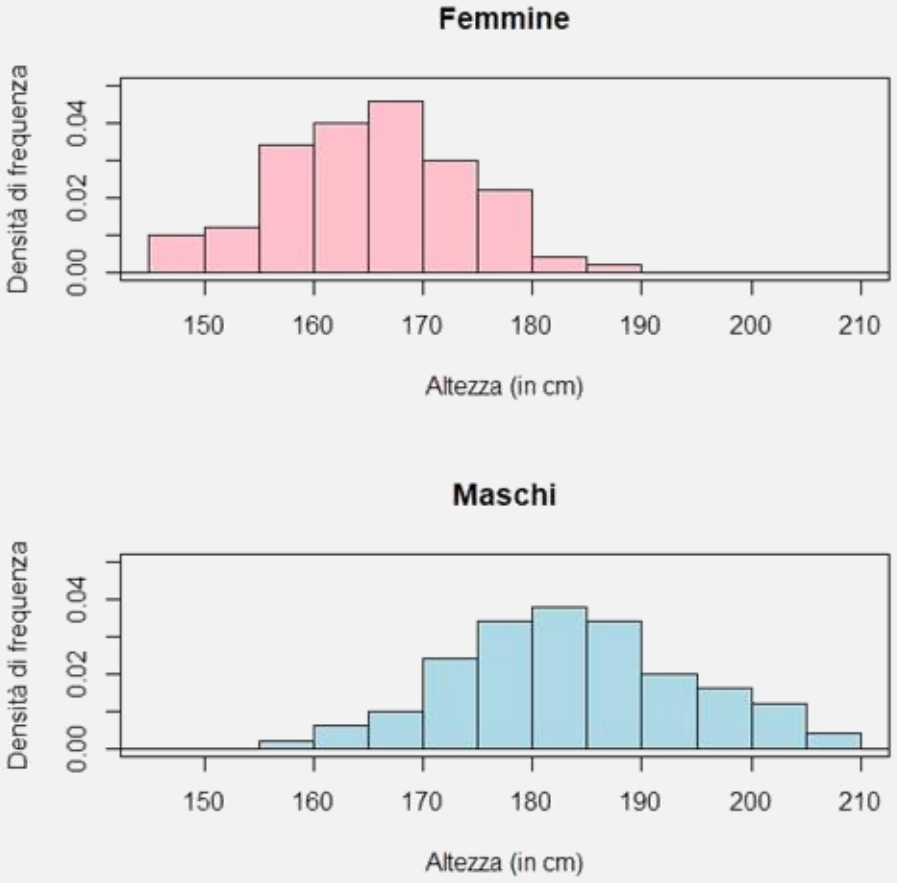
moda	71
mediana	72
media	72,2

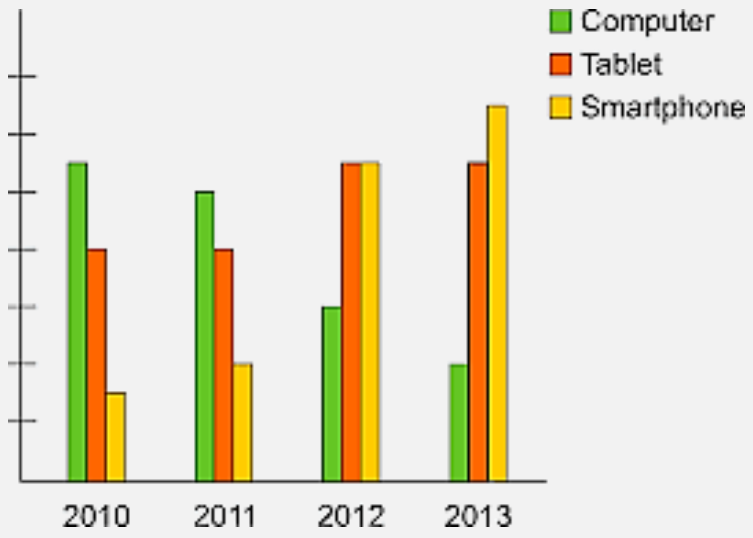
WWW.ANDREAMININI.ORG



Histograms

- A histogram describes the relative frequency of data within an interval (a, b) and is used to visualize the data distribution.





Unimodal/Bimodal Distributions

- A distribution may exhibit multiple modes:
  - **Unimodal distributions**: frequency distributions that have a single mode, that is, a single point of maximum (representing both the relative and the absolute maximum)
  - **Bimodal or k-modal distributions**: frequency distributions that exhibit two or more modes, that is, two (or k) relative maxima
    - Example: by measuring the heights of a group of adolescents in which the larger part is female and the smaller part male, one obtains a bimodal distribution, with a primary and a secondary mode.

Zero-modal Distribution

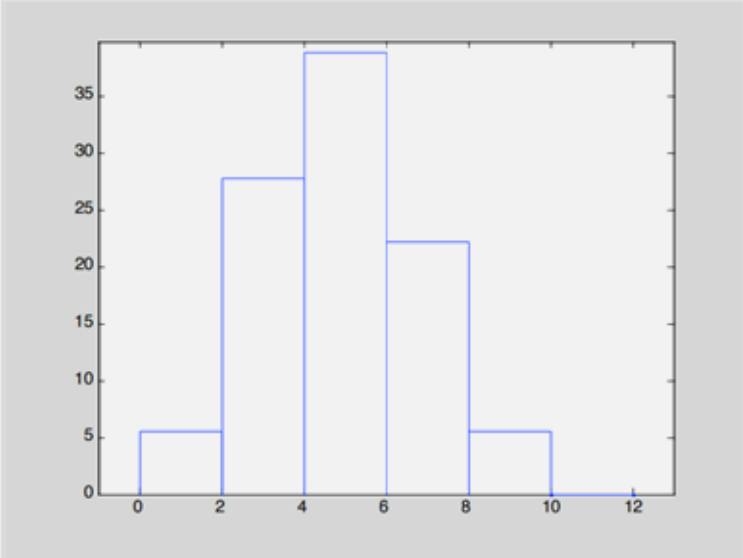
- No value has a higher frequency than the others:

$$A = \{1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6\}$$

Unimodal distribution

C'è un solo valore con una frequenza più elevata degli altri

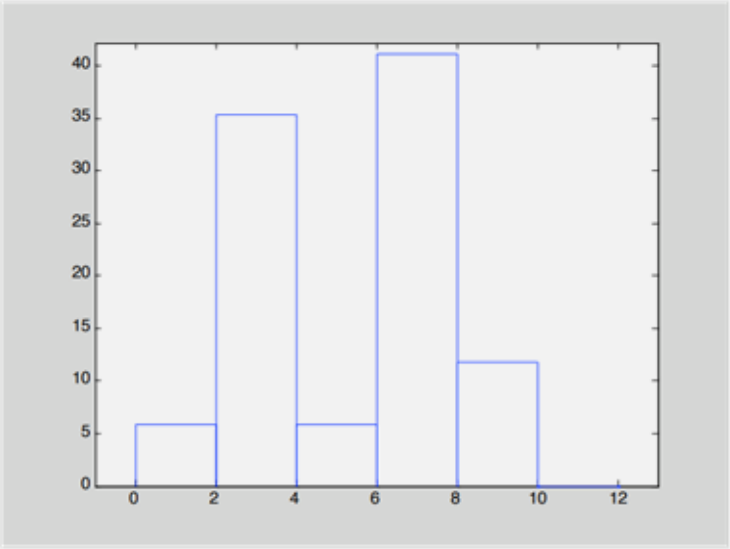
A = {1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8}

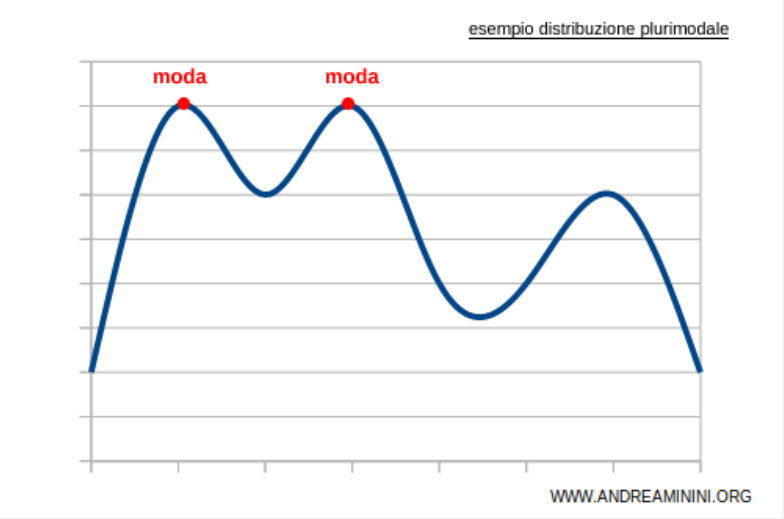


Bimodal distribution

Ci sono due valori con una frequenza più elevata degli altri.

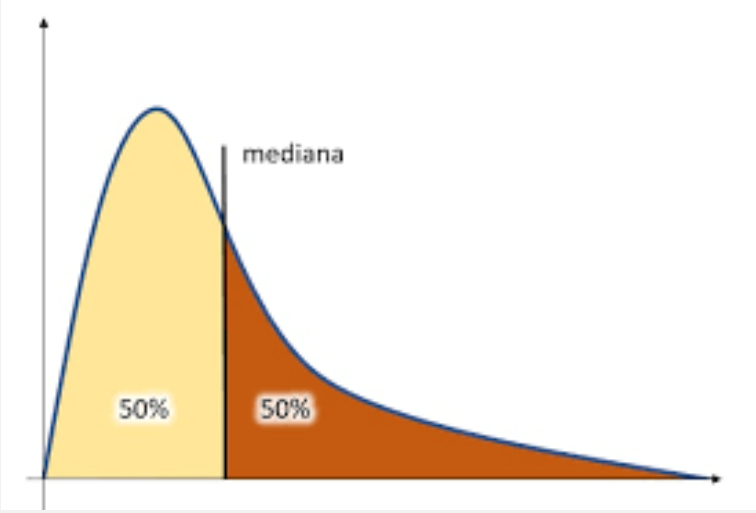
A = {1, 2, 2, 3, 3, 3, 3, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8}





Median

- The **median** is the value that occupies the central position in an ordered set of data.
- It is a robust measure, as it is minimally influenced by the presence of outliers.
- Characteristics:
  - it is used when one seeks to mitigate the effect of extreme values;
  - in a distribution or data series, each value drawn at random has the same probability of being below or above the median.



Calculation of the Median

To calculate the median of a dataset, one must:

1. arrange the values in ascending or descending order and count the total number n of data;
2. if the number (n) of data is odd, the median corresponds to the numerical value of the central datum, i.e., the one occupying position $(n+1)/2$;
3. if the number (n) of data is even, the median is estimated by using the two central values occupying positions $n/2$ and $n/2+1$:
    a. with a small number of observations, the median is taken to be the arithmetic mean of these two intermediate observations;
    b. with a large number of observations grouped into classes, one sometimes resorts to proportions.

Example.
Consider the following sample:

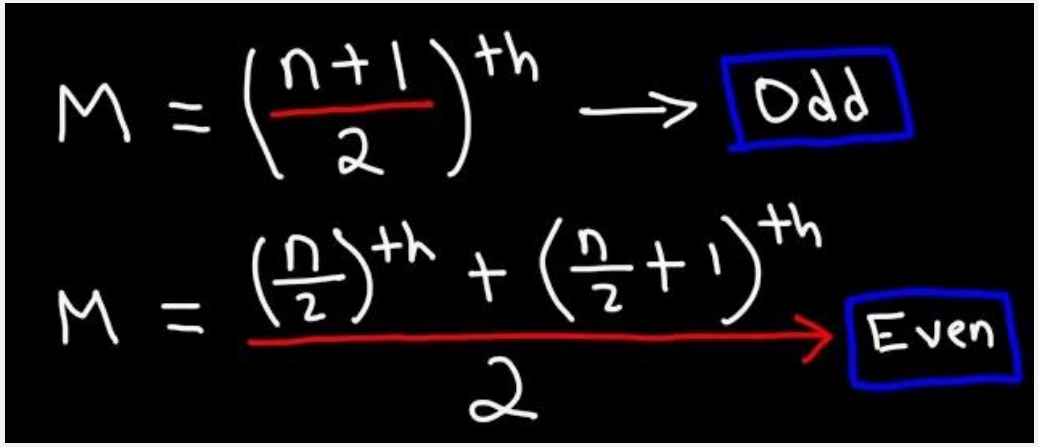
96	78	90	62	73	89	92	84	76	86

Order the samples in ascending order:

62	73	76	78	**84**	**86**	89	90	92	95

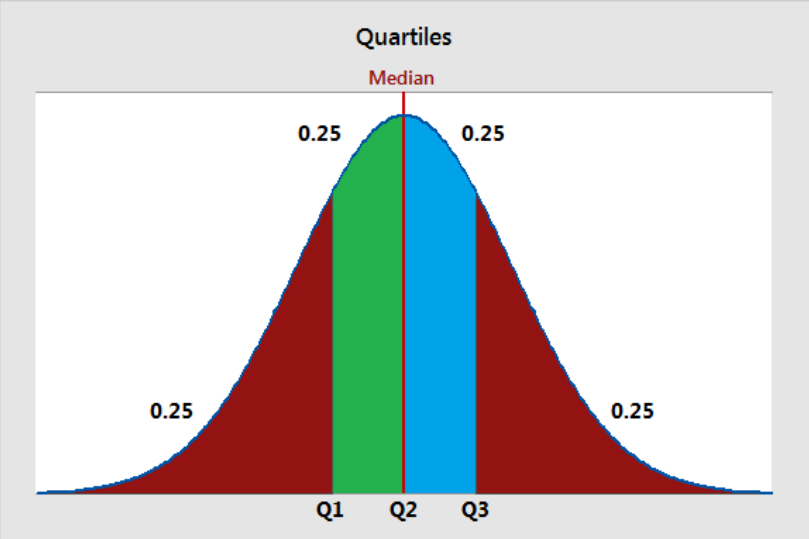
1. Since the number of samples is even ($n = 10$), the median is computed as the mean of the two central *elements*:

$$median = 84 + 862 = 85$$



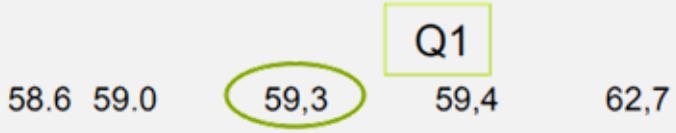
Quantiles

- The **quantiles** are a family of measures, to which the median also belongs, that are distinguished according to the number of equal parts into which they divide a distribution.
- The **quartiles** partition the distribution into 4 parts of equal frequency, where each part contains the same fraction of observations:
  - The **first quartile** is defined as the number $q1$ such that 25% of the statistical data are less than or equal to $q1$.
  - The **second quartile** is defined as the number $q2$ such that 50% of the statistical data are less than or equal to $q2$. The second quartile corresponds to the median.
  - The **third quartile** is defined as a number $q3$ such that 75% of the statistical data are less than or equal to $q3$.

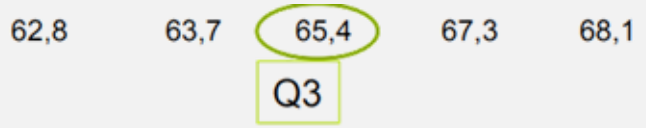


Example.
Let us consider a study that examines restaurant waiting times in a sample of 10 customers:

The median is equal to 62.75
One considers the lower half of the data, i.e., all values below the median, and on this subset of data the median is computed; the value obtained is Q1

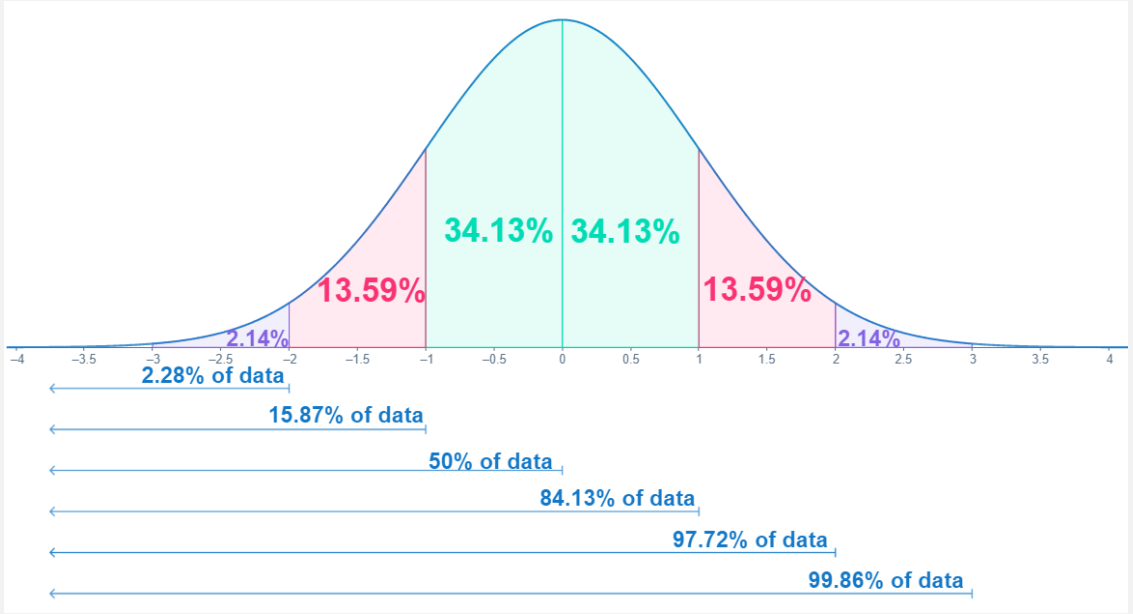


One considers the upper half of the data, i.e., all values above the median, and on this subset of data the median is computed; the value obtained is Q3



Deciles and percentiles

- Analogously, the following are defined:
  - **Deciles**: 9 points that divide the ordered distribution into 10 equal parts
  - **Percentiles**: 99 points that divide the ordered distribution into 100 equal parts



Range of variation

- The **range** of a distribution is the difference between the largest and the smallest value in the distribution:

$$C = x_{max} - x_{min}$$

This index is relatively coarse, as it provides no information regarding the variability of intermediate values.

- Example: the range of the following distribution:

$$25 - 26 - 28 - 29 - 30 - 32 \rightarrow C = 32 - 25 = 7$$

Deviation

- The **deviation** measures how far each datum x_i departs from the mean value, hence:

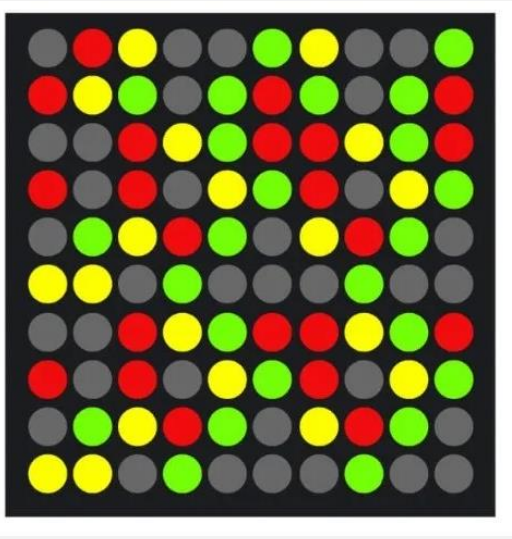
$$s = x_i - \bar{X}$$

- Example: Let us consider the following intensities recorded from the microarray spots:

$$435.02, 678.14, 235.35, 956.12, \ldots, 1127.82, 456.43$$

  - The mean of these values is: 515.13; their deviations are:

$$956.12 - 515.13 = 440.99 \quad 235.35 - 515.13 = -279.78 \quad 678.14 - 515.13 = 163.01 \quad 435.02 - 515.13 = -80.11$$



Absolute deviation

Using s, several other indices of variability can be derived:

- The **mean absolute deviation** is denoted by s_m and is the arithmetic mean of the absolute values of the deviations

$$s_m = \frac{\sum_{i=1}^{n} |x_i - \bar{X}|}{n}$$

Variance

- **Variance of the population**: optimal measure to characterize the variability of a population. An indicator of the dispersion of a variable or statistical distribution that one obtains by calculating the average of the squares of the deviations from the arithmetic mean (μ).

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

where N is the number of observations of the entire population; μ is the mean of the population; x_i is the i-th observation

- **Variance of a sample**:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

where n is the number of observations of the sample; $\bar{X}$ is the mean of the sample; x_i is the i-th observation

When n is high enough, the difference between the two formulas are tiny.

Example.

Let us consider the following sample of observations:

$$\{2,3,6,9,15\}$$

Computation of the mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{2+3+6+9+15}{5} = 7$$

Computation of the sample variance:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{(2-7)^2 + (3-7)^2 + (6-7)^2 + (9-7)^2 + (15-7)^2}{4} = 27.5$$

Deviance

In the calculation of some statistics, one uses the deviance, obtained by the numerator of the variance. It is a measure of dispersion. It is obtained by calculating the sum of the squares of the deviations of the data in a distribution from the mean.

$$Dev = \sum_{i=1}^{N} (X_i - \bar{X})^2$$

Variance is the mean of the deviance; that is, it indicates how much the values of the distribution vary with respect to the mean

Standard deviation, or root mean square deviation

- Variance has the disadvantage of being a quadratic quantity and therefore is not directly comparable with the mean or with the other values of the distribution.

- To obtain a measure expressed in the same unit of measurement as the original variable, it suffices to extract the square root of the variance.

- The **standard deviation** is a measure of distance from the mean and therefore always has a positive value.

- It is a measure of the dispersion of the random variable around the mean.

Standard deviation, or root mean square deviation

- **Population standard deviation**:

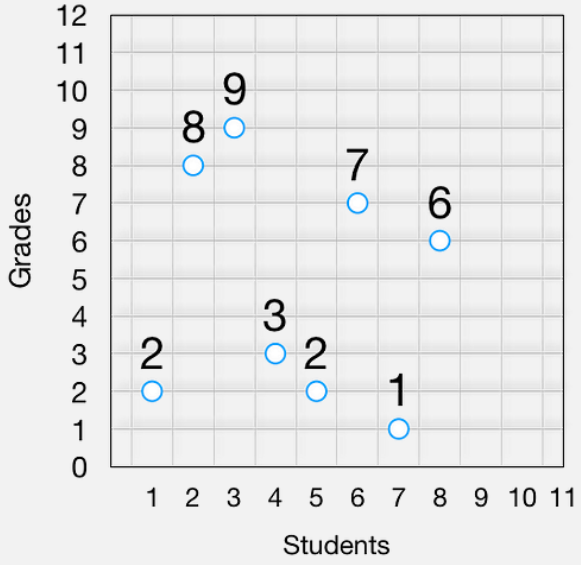
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

where N is the number of observations in the entire population; μ is the population mean; X_i is the i-th observation.

Sample standard deviation:

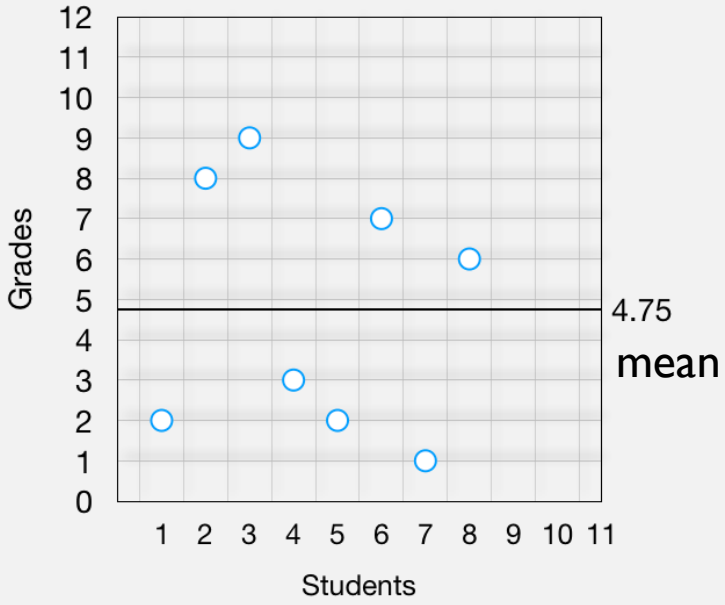
$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

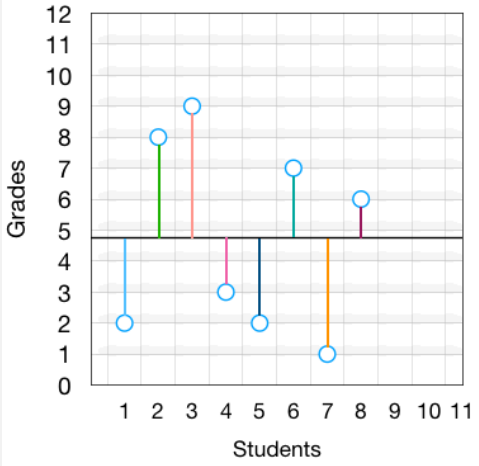
where n is the number of observations in the sample; X is the sample mean; x_i is the i-th observation.



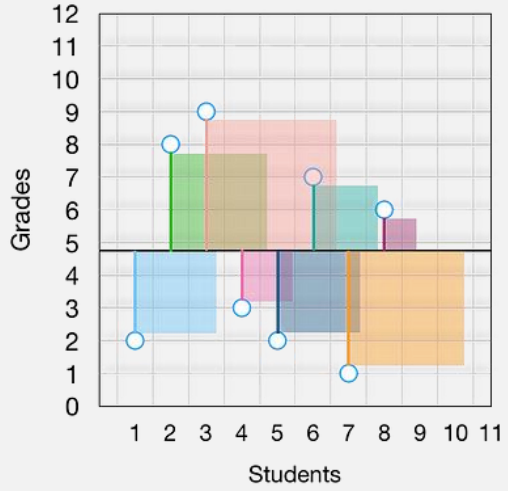
Student	Grade
1	2
2	8
3	9
4	3
5	2
6	7
7	1
8	6

$$\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N} = \frac{2+8+9+3+2+7+1+6}{8} = \frac{38}{8} = 4.75$$





$$x - \bar{x} =$$
$$(2-4.75) + (8-4.75)$$
$$+ (9-4.75) + (3-4.75)$$
$$+ (2-4.75) + (7-4.75)$$
$$+ (1-4.75) + (6-4.75)$$



$$\sum (x_n - \bar{x})^2 =$$

$$7.5625 + 10.5625 + 18.0625 + 3.0625 + 7.5625 + 5.0625 + 14.0625 + 1.5625 = 67.5$$

Variance

$$\frac{\sum (x_n - \bar{x})^2}{N} = \frac{67.5}{8} = 8.45 \text{ points}^2$$

Standard deviation

$$\frac{\sum (x_n - \bar{x})^2}{N}$$

Mean ( ) = 8.45

$$\sqrt{} \; 8.45 \; 2.91 = 2.91$$

Example.
Let us consider the grades of two students:

- Anna: *30, 30, 28, 27, 26*
- Stefano: *21, 30, 30, 30, 30*

Both have the same mean of the grades (*28.2*)

Let us compute the standard deviation:
- $\sigma_{Anna} = 1.78$
- $\sigma_{Stefano} = 4.02$

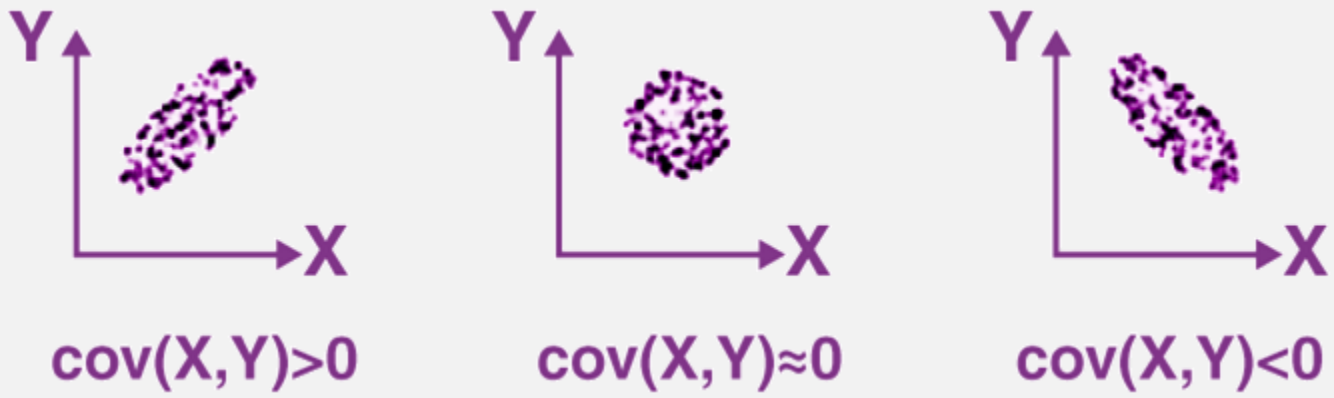
What does this mean? Anna's grades are more concentrated (closer) than those of Stefano

Covariance

- Index that makes it possible to verify whether between two statistical variables there exists a linear relationship.

- Considering two series $\{x_i\}$ and $\{y_i\}$, $i = 1,2, \dots, n$, it compares the pairs of deviations $(x_i - \bar{x})$ and $(y_i - \bar{y})$:

$$Cov(X, Y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$



Covariance can be:

- **POSITIVE**: when X and Y tend to vary in the same direction; that is, as X increases, Y also tends to increase; and as X decreases, Y also tends **to** decrease.

- **NEGATIVE**: when the two variables tend to vary in opposite directions; that is, when as one variable increases, the other variable tends to decrease (and vice versa)

- **ZERO**: when there is no tendency for the two variables to vary in the same direction or in opposite directions. When $Cov(X, Y) = 0$ it is also said that X and Y are uncorrelated or linearly independent.

Correlation

- It is a more rigorous approach that allows the study the **degree of intensity** of the linear relationship between pairs of variables

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(VarX)(VarY)}}$$

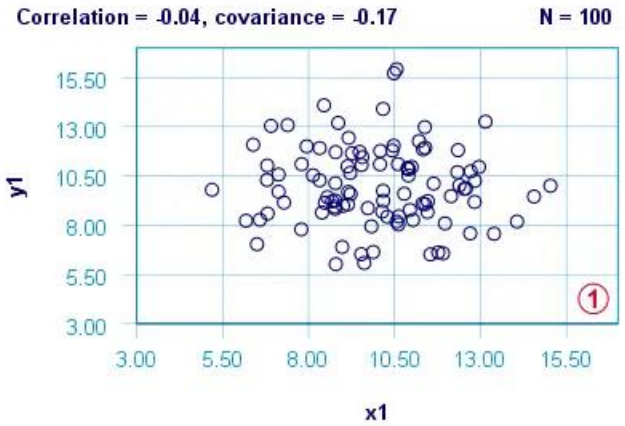
Pearson correlation coefficient

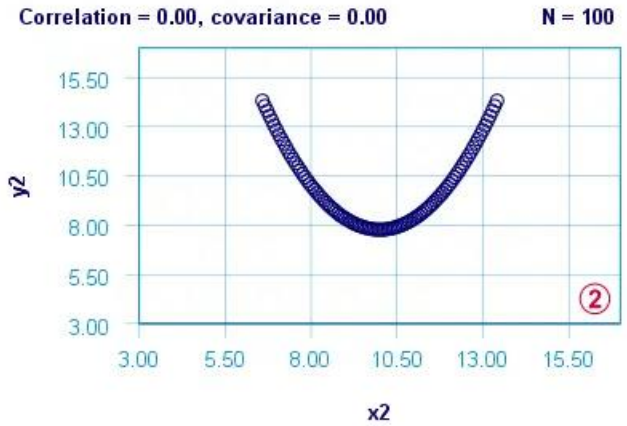
The correlation coefficient allows us to:

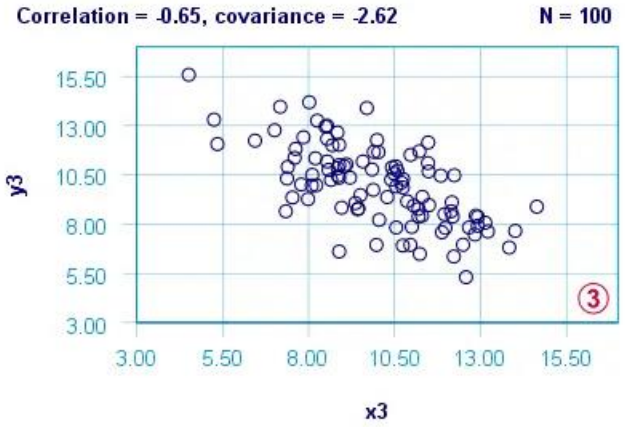
- summarize the strength of the **linear** relationship between the variables
- assess the apparent association between the variables

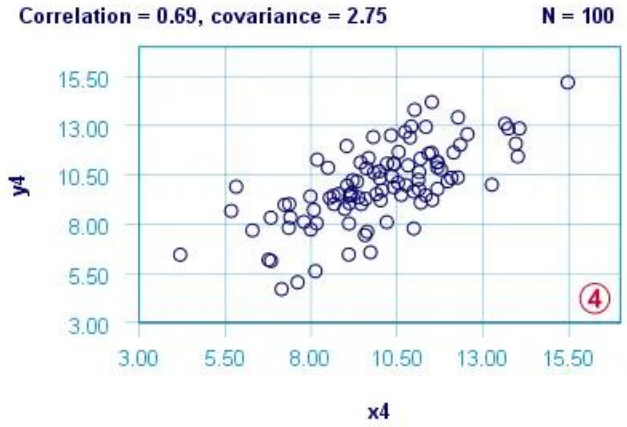
The correlation coefficient:

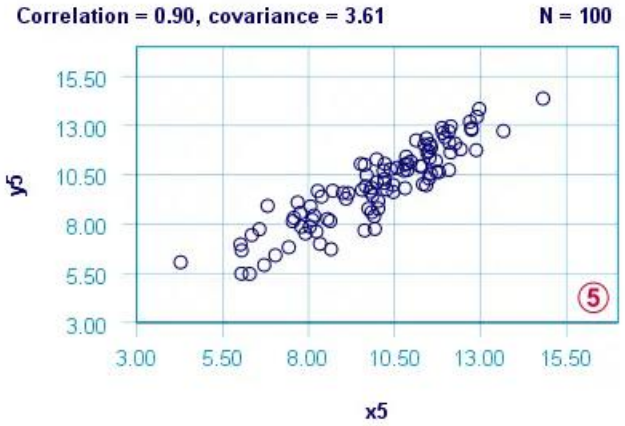
- ranges from -1 to 1 (if equal to 1 or -1: perfectly correlated)
- is positive when the values of the variables increase together
- is negative when the values of one variable increase as the values of the other decrease
- is not influenced by the units of measurement

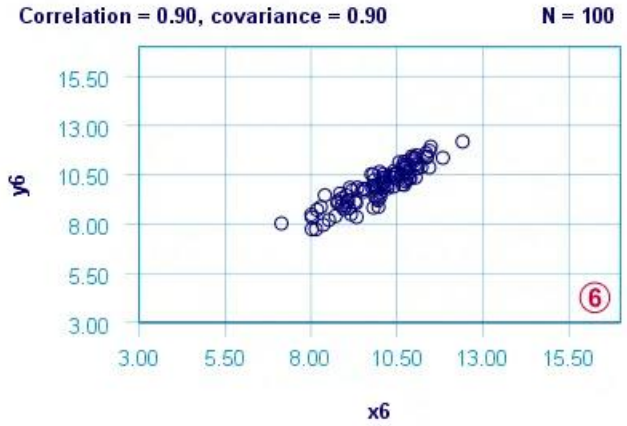

Correlation = -0.04, covariance = -0.17
N = 100
y1
x1
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
①

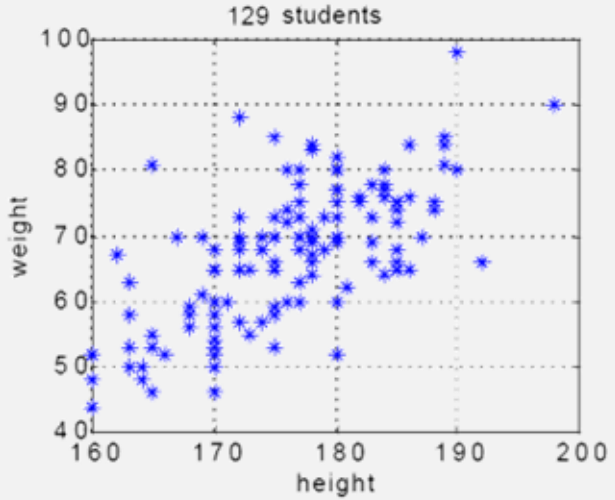

Correlation = 0.00, covariance = 0.00
N = 100
y2
x2
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
②

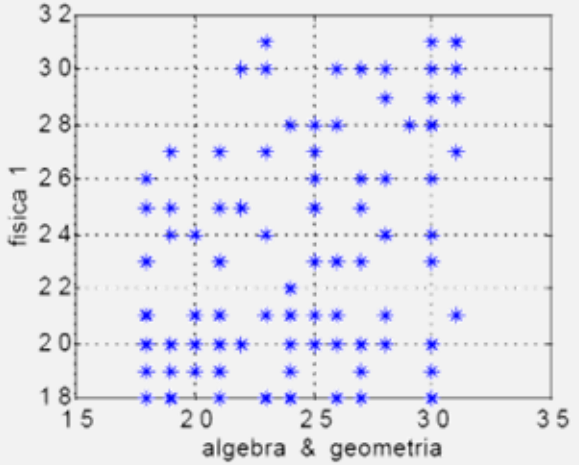

Correlation = -0.65, covariance = -2.62
N = 100
y3
x3
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
③

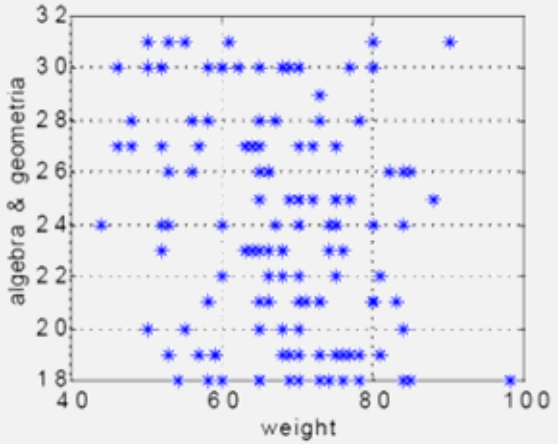

Correlation = 0.69, covariance = 2.75
N = 100
y4
x4
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
④

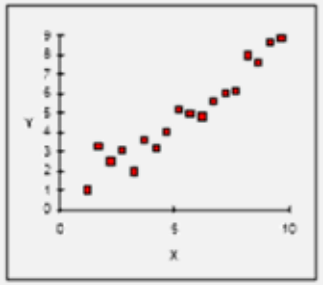

Correlation = 0.90, covariance = 3.61
N = 100
y5
x5
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
⑤


Correlation = 0.90, covariance = 0.90
N = 100
y6
x6
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
⑥



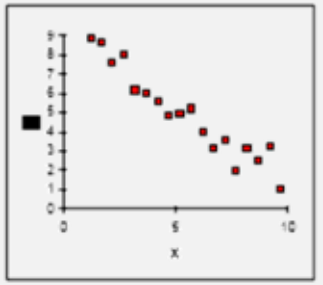


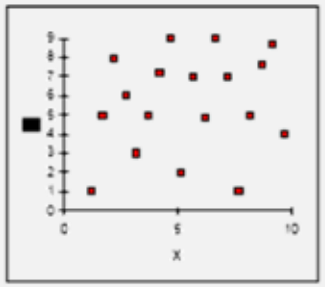




r=0,96

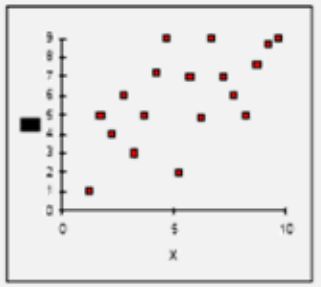
r=-0,96



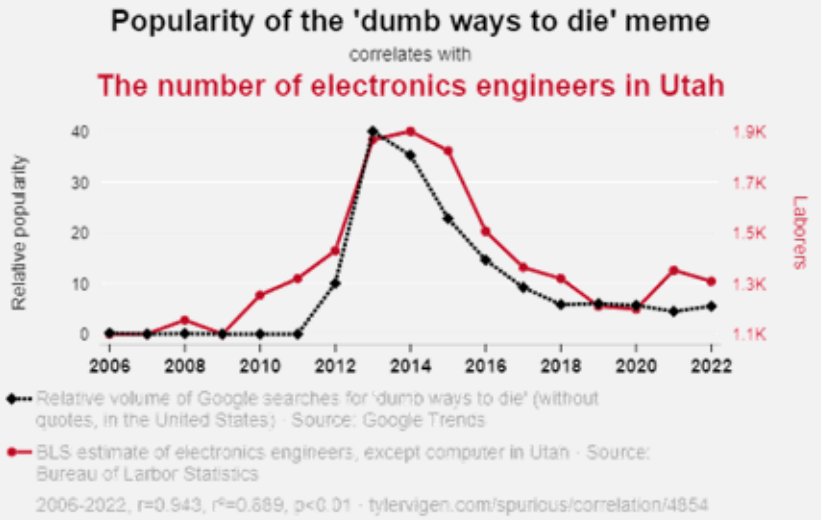


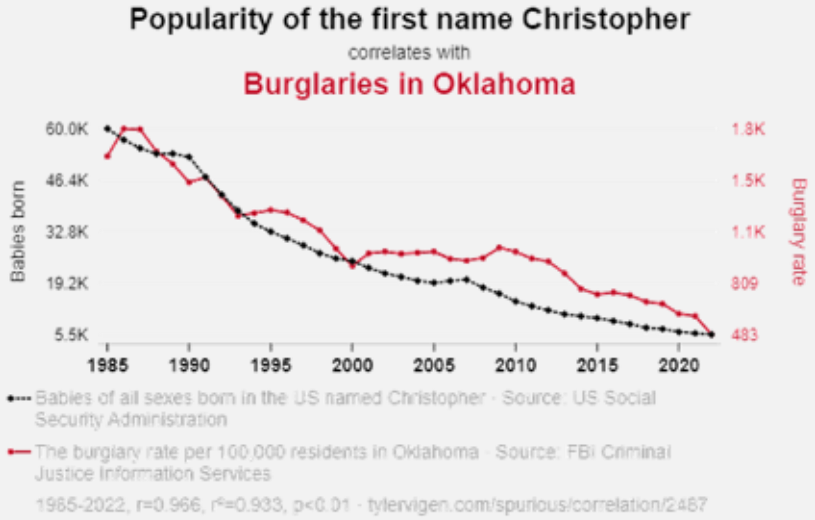
r=0,12

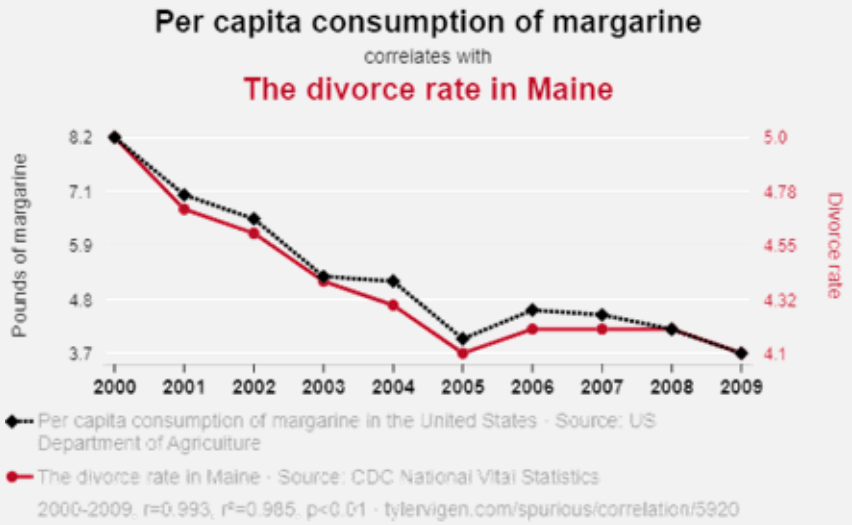
r=0,62



CORRELATION IS NOT CAUSATION





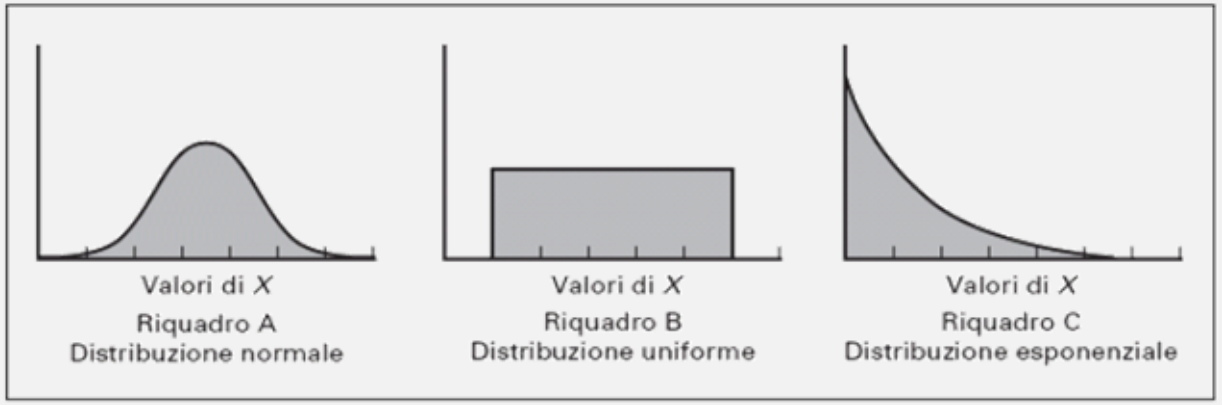


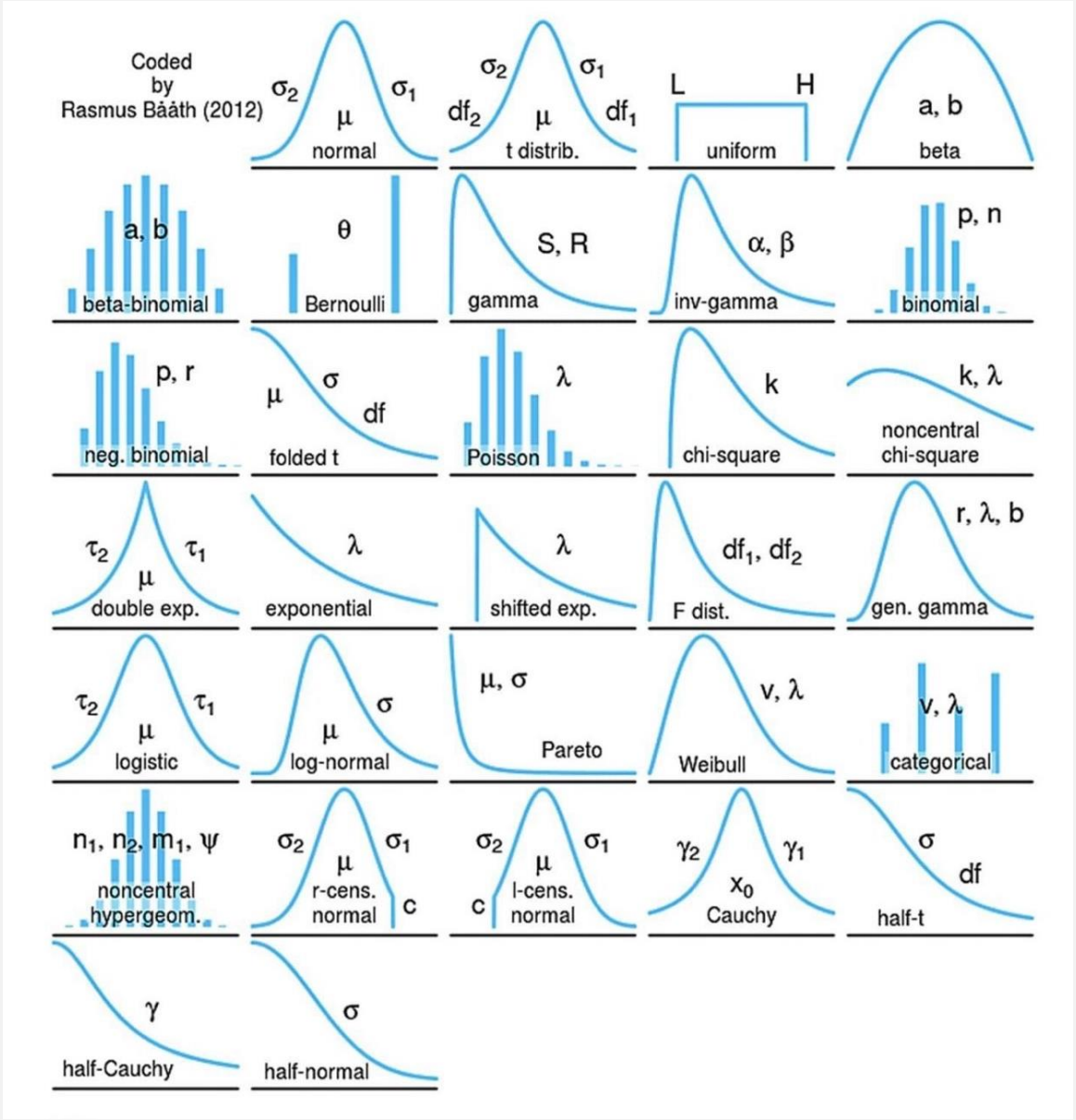
Continuous probability distributions

- A continuous probability density function is a model that analytically defines how the values assumed by a continuous random variable are distributed

- When one has a mathematical expression suitable for representing a continuous phenomenon, we are able to calculate the probability that the random variable takes values lying within intervals

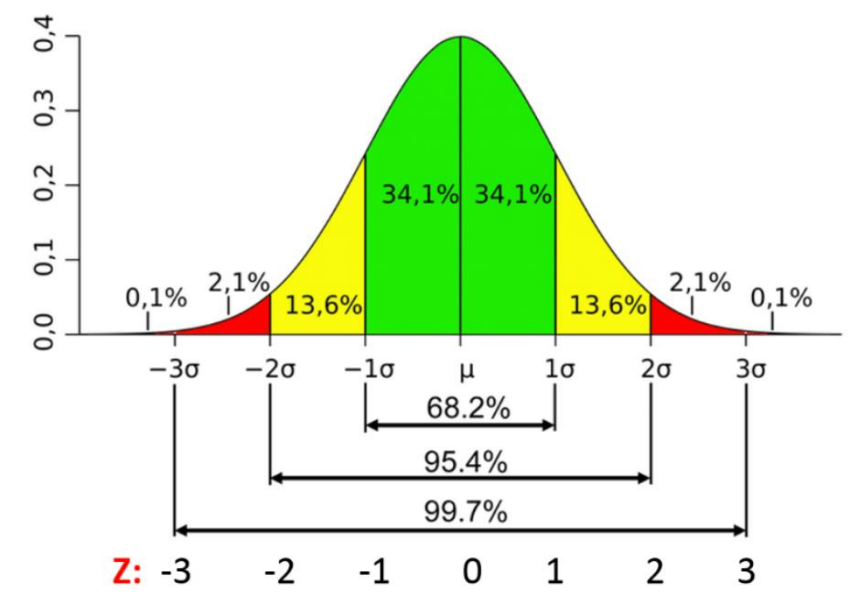
- The figure graphically depicts three probability density functions: normal, uniform, and exponential

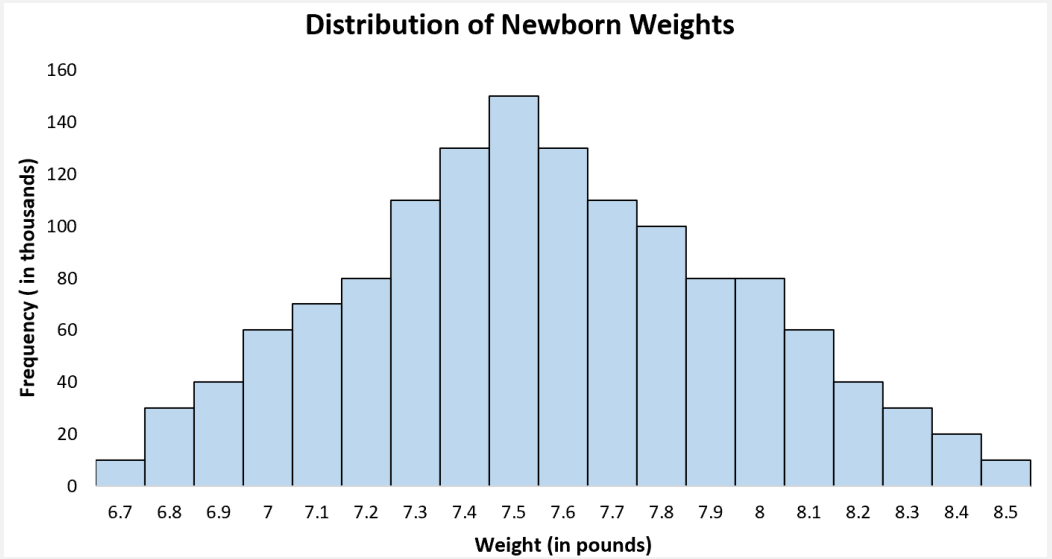


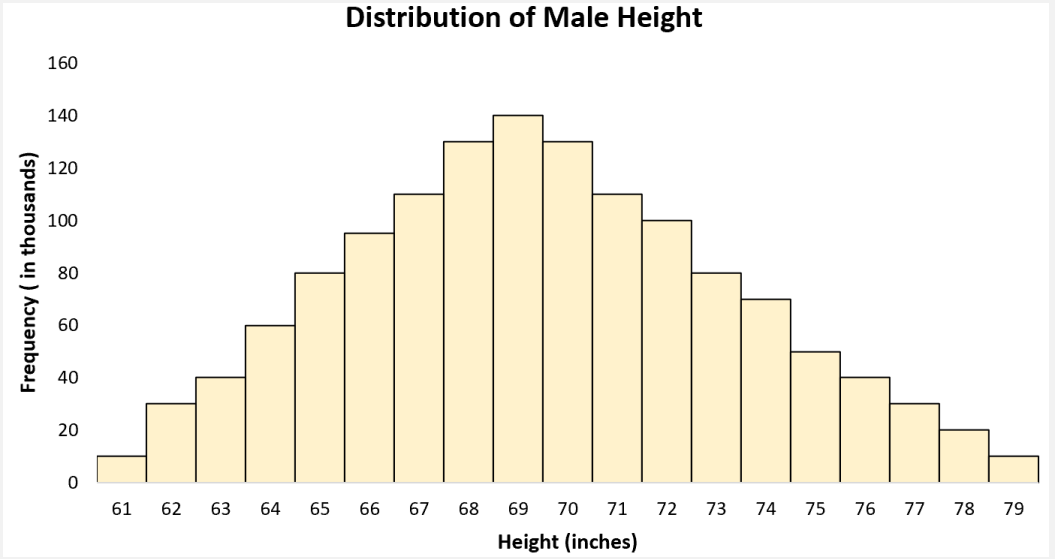

Coded
by
Rasmus Bååth (2012)
σ2
σ1
μ
normal
σ2
σ1
df2
μ
df1
t distrib.
L
H
uniform
a, b
beta
a, b
beta-binomial
θ
Bernoulli
S, R
gamma
α, β
inv-gamma
p, n
binomial
p, r
neg. binomial
μ
σ
df
folded t
λ
Poisson
k
chi-square
k, λ
noncentral
chi-square
τ2
τ1
μ
double exp.
λ
exponential
λ
shifted exp.
df1, df2
F dist.
r, λ, b
gen. gamma
τ2
τ1
μ
logistic
μ
σ
log-normal
μ, σ
Pareto
v, λ
Weibull
v, λ
categorical
n1, n2, m1, ψ
noncentral
hypergeom.
σ2
σ1
μ
r-cens.
normal
c
σ2
σ1
μ
c
l-cens.
normal
γ2
γ1
x0
Cauchy
σ
df
half-t
γ
half-Cauchy
σ
half-normal

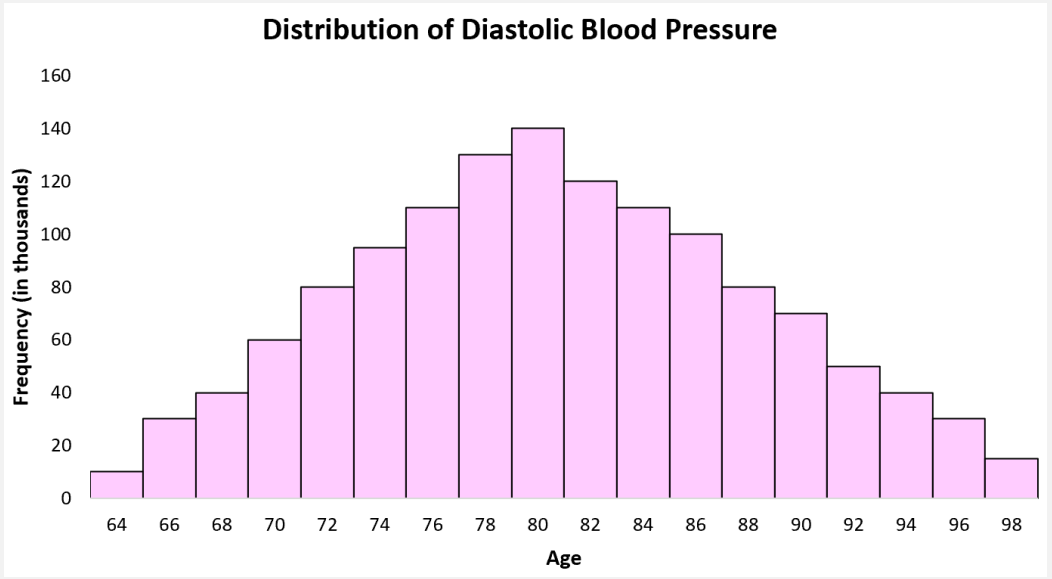
Normal distribution

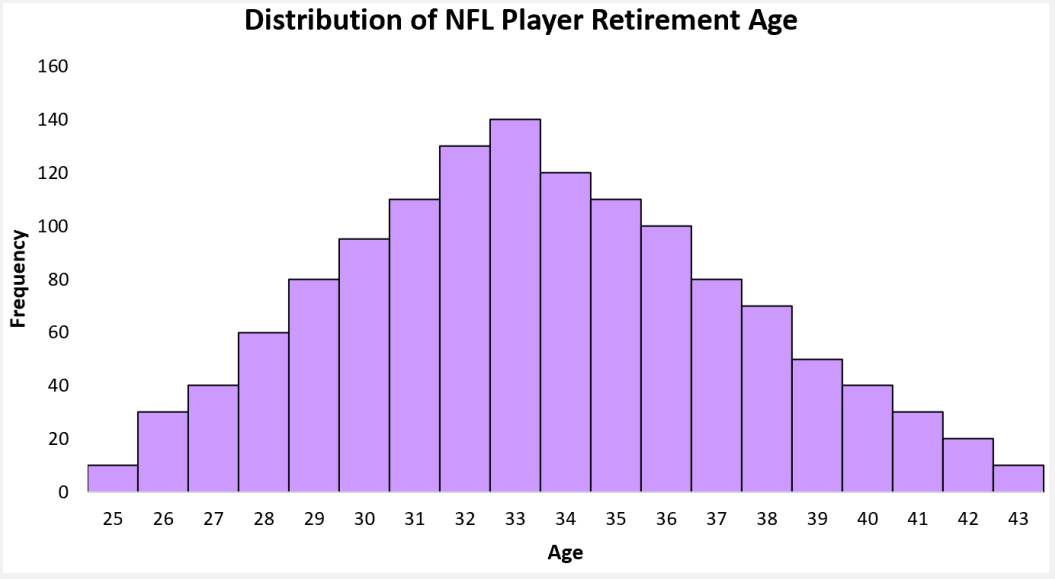
- The normal distribution (or Gaussian distribution) is the most widely used continuous distribution in statistics.
- The normal distribution is important in statistics for three fundamental reasons:
  1. Several continuous phenomena appear to follow, at least approximately, a normal distribution.
  2. The normal distribution can be used to approximate numerous discrete probability distributions.




Distribution of Newborn Weights
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
6.7
6.8
6.9
7
7.1
7.2
7.3
7.4
7.5
7.6
7.7
7.8
7.9
8
8.1
8.2
8.3
8.4
8.5
Weight (in pounds)


Distribution of Male Height
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
Height (inches)


Distribution of Diastolic Blood Pressure
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
64
66
68
70
72
74
76
78
80
82
84
86
88
90
92
94
96
98
Age


Distribution of NFL Player Retirement Age
Frequency
0
20
40
60
80
100
120
140
160
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
Age

Normal distribution

The normal distribution has several important characteristics:

- The normal distribution has a bell-shaped and symmetric form
- Its measures of central tendency (expected value, median) coincide
- Its interquartile range equals 1.33 times the standard deviation; that is, it covers an interval between $\mu - 2/3\,\sigma$ and $\mu + 2/3\,\sigma$
- The normally distributed random variable takes values between $-\infty$ and $+\infty$

Normal distribution

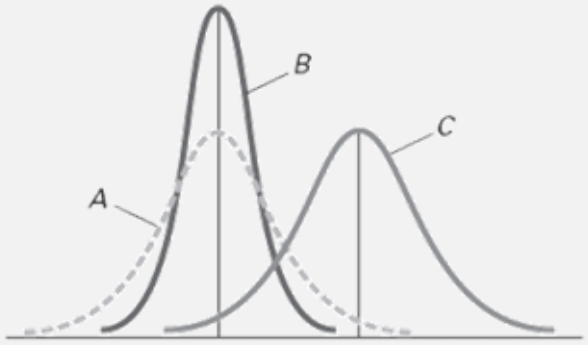
We will use the symbol $f(X)$ to denote the mathematical expression of a probability density function. In the case of the normal distribution, the normal probability density function is given by the following expression:

Normal probability density function

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(\frac{1}{2}\right)\left[\frac{X-\mu}{\sigma}\right]^2}$$

Where μ is the expected value of the population; σ is the standard deviation of the population; X represents the values assumed by the random variable, $-\infty < X < +\infty$

We note that, since e e π are mathematical constants, the probabilities of a normal distribution depend solely on the values assumed by the two parameters μ and σ. By specifying particular combinations of μ and σ, we obtain different normal probability distributions.



Exercise.

The following are the heights of ten dogs participating to a show:

$$Y = (40,42,38,41,40,45,46,42,42,41)$$

Compute median, mean, and variance.

To comupte the median, data must be ordered:

$$38, 40, 40, 41, 41, 42, 42, 42, 45, 46$$

Mean: sum of the data divided by 10 → 41.7 cm

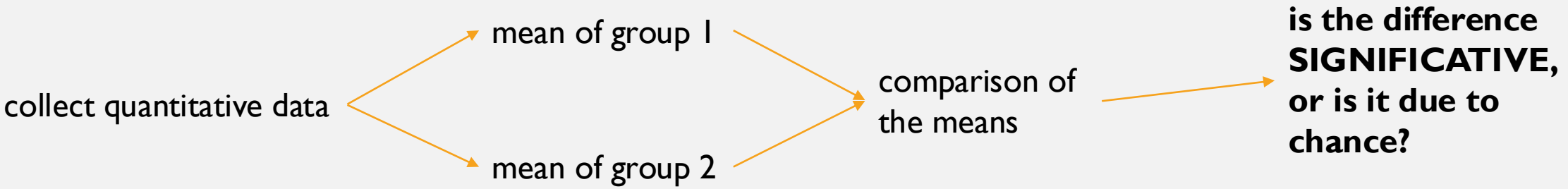
Median: if the number (n) of data is even, the median is estimated using the two central values that occupy positions $n/2$ and $n/2+1$→ mean of the fifth and sixth value → 41.5 cm

Variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$ → 5.01

Comparing two means: the Student *t* test

is the difference between the means of the two samples significant?

Can you state that the observed difference is not due to chance but that, instead, there is genuinely a difference between the means of the two populations?



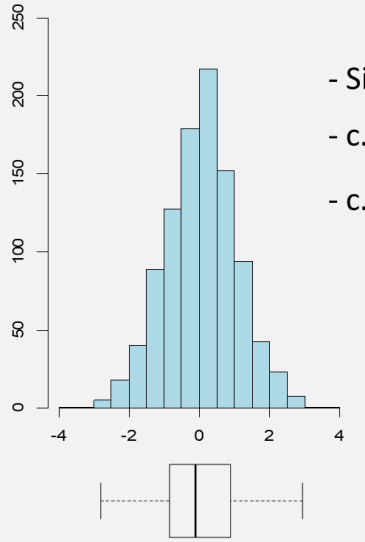
Comparing two means: the Student *t* test

comparison of the means → ZERO HYPOTHESIS: the difference is due to chance! → accept or refuse the zero hypothesis? → **Student's *t*-test**

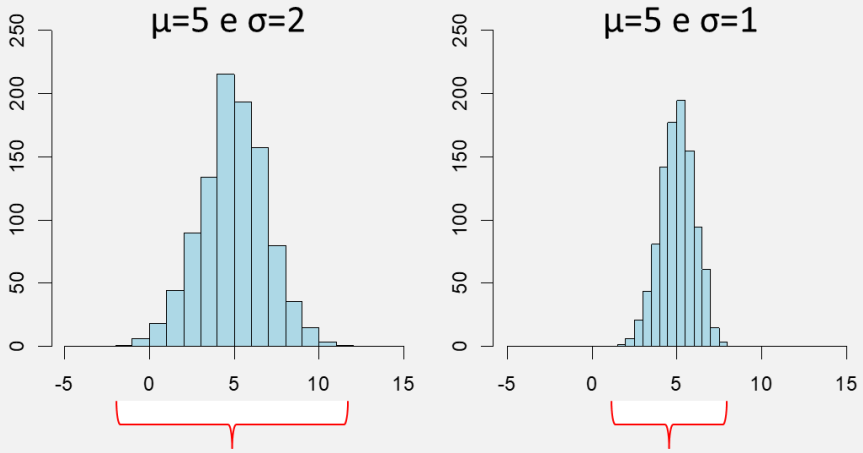
Conditions:

1. Independence of observations
2. Normality of the populations being compared
3. Homogeneity of variance

Analisi dell'istogramma



- Simmetria (media ≈mediana)
- c. 2/3 dei dati in un intervallo μ±σ
- c. 95% dei dati in un intervallo μ±2σ



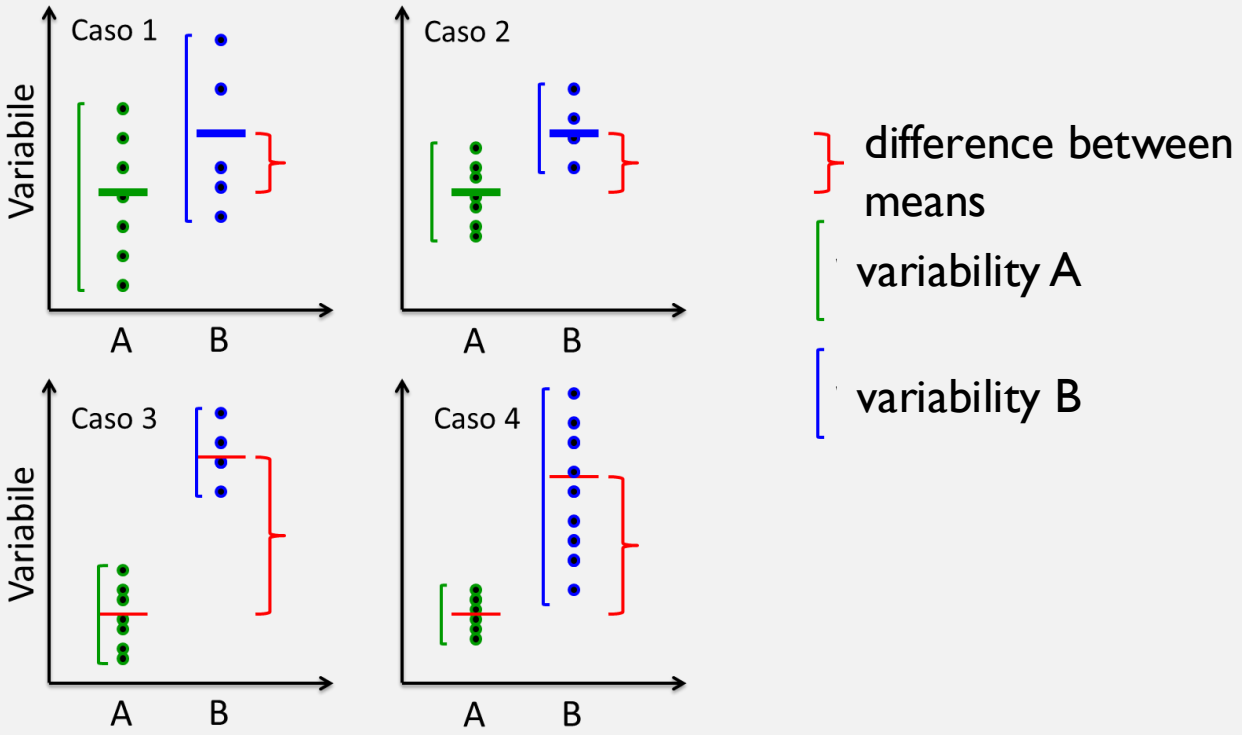
Comparing two means: the Student *t* test

difference between means

dimensional factor

$$t = \frac{m_a - m_b}{S} \sqrt{\frac{n_a\, n_b}{n_a + n_b}}$$

average standard deviation



difference between means

variability A

variability B