PROF. CRESCENZIO GALLO

STATISTICAL DATA ANALYSIS IN HEALTHCARE RESEARCH

Università di Foggia

University of Foggia

Medical Area Departments

STATISTICS

How to Present Results Clearly?

- In this handout, we will attempt to delve into the most common statistical techniques for analyzing data collected in healthcare research.
- The basic concepts for properly interpreting the results of a study will also be provided.

DATA COLLECTION AND PRESENTATION

STATISTICAL ASPECTS
Data Collection

The information collected during the course of a controlled clinical trial (CCT), and usually reported on a special data collection form, generally concerns the socio-demographic and clinical characteristics of enrolled patients.

Each of these characteristics (e.g., age, sex, stage of disease, etc.) is called a "**variable**." The variables used within any study can be of two distinct types:

- *Numerical (continuous* or *discrete)*
- *Ordinal*
- *Nominal* (or *categorical*)

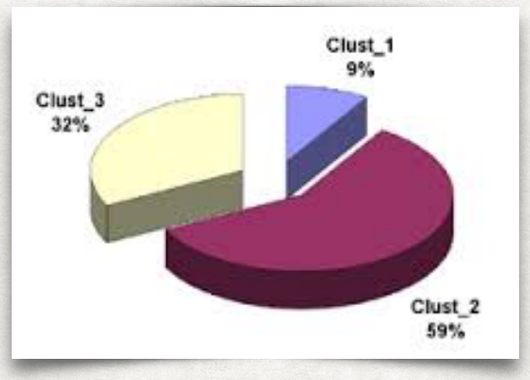
This distinction is critical because, as we will see later, it will guide us in choosing which summary measures and statistical tests to use.

STATISTICAL ASPECTS
Data Collection

- **Continuous** numerical variables can assume an infinite number of values within a certain range.
- Moreover, the distance between 3 and 4 is the same as between 20 and 21.
- This means that if we consider for example the weight, typical continuous variable, a subject who weighs 80 kg will have a weight that is actually double than a subject who weighs 40 kg.
- Age, blood pressure, glycemia, are all examples of continuous variables.

STATISTICAL ASPECTS
DATA COLLECTION

- **Discrete** numeric variables differ from continuous variables in that they can only take on a finite number of values within a specific range (e.g., the number of children, in the range 2-5, can only take on the value of 2, 3, 4, or 5).
- In **ordinal** variables, although values are placed in a predetermined order (e.g., a class IV heart failure is more severe than a class III heart failure, which in turn is more severe than a class II heart failure), there is no equidistance between values (i.e., we cannot say that a class IV heart failure is twice as severe as a class II heart failure or four times as severe as a class I heart failure).
- Stages of disease, or measures of quality of life, are typically ordinal variables.



- Finally, **nominal** variables express an "all or nothing" type of quality, with no set order.
- Examples are gender, race, presence/absence of a complication, etc.

STATISTICAL ASPECTS

DEFINITIONS OF NUMERICAL AND NOMINAL VARIABLE

- **CONTINUOUS (NUMERICAL) VARIABLE**

  Characterized by an infinite number of possible values between any two values. It therefore refers to a continuous set of values (e.g., blood pressure, blood glucose, etc.).

- **DISCRETE (NUMERICAL) VARIABLE**

  It can only take a finite number of values within a specific range of integers.

- **ORDINAL VARIABLE**

  Characterized by a predetermined order for ranking responses (e.g., disease stage, pain scale, etc.).

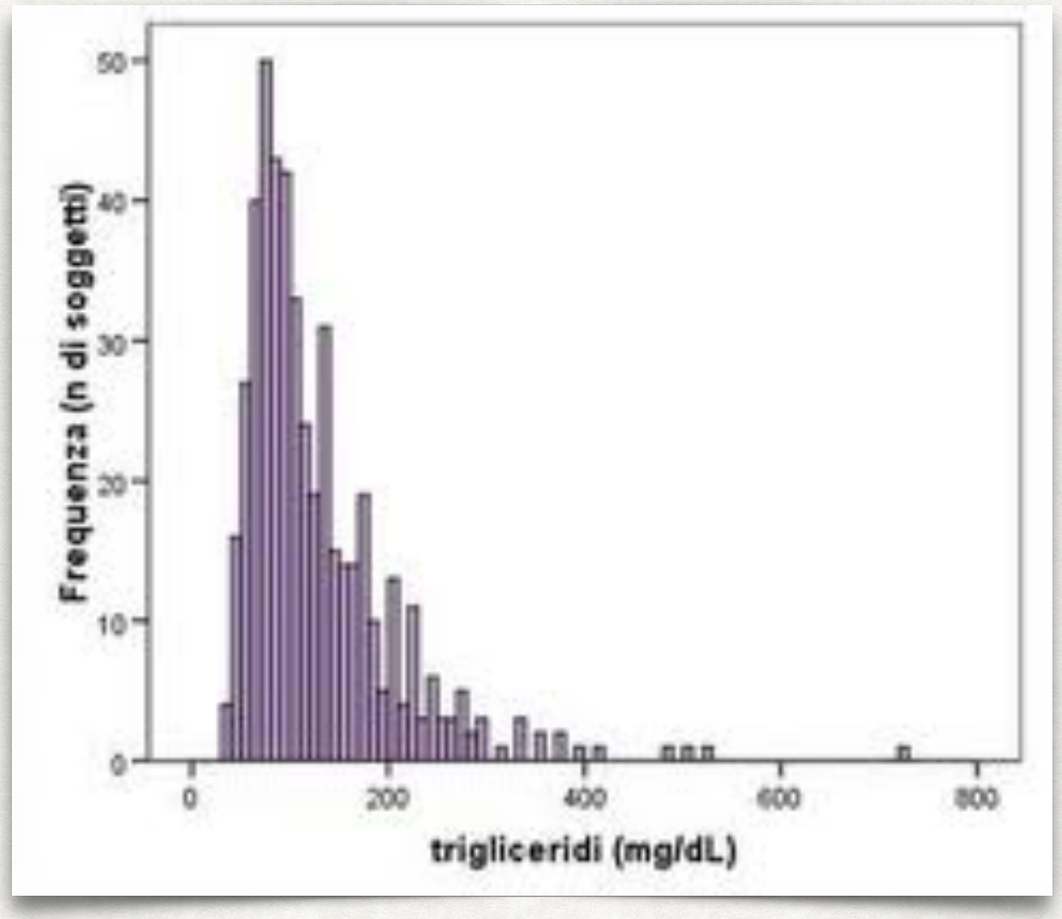
- **NOMINAL VARIABLE**

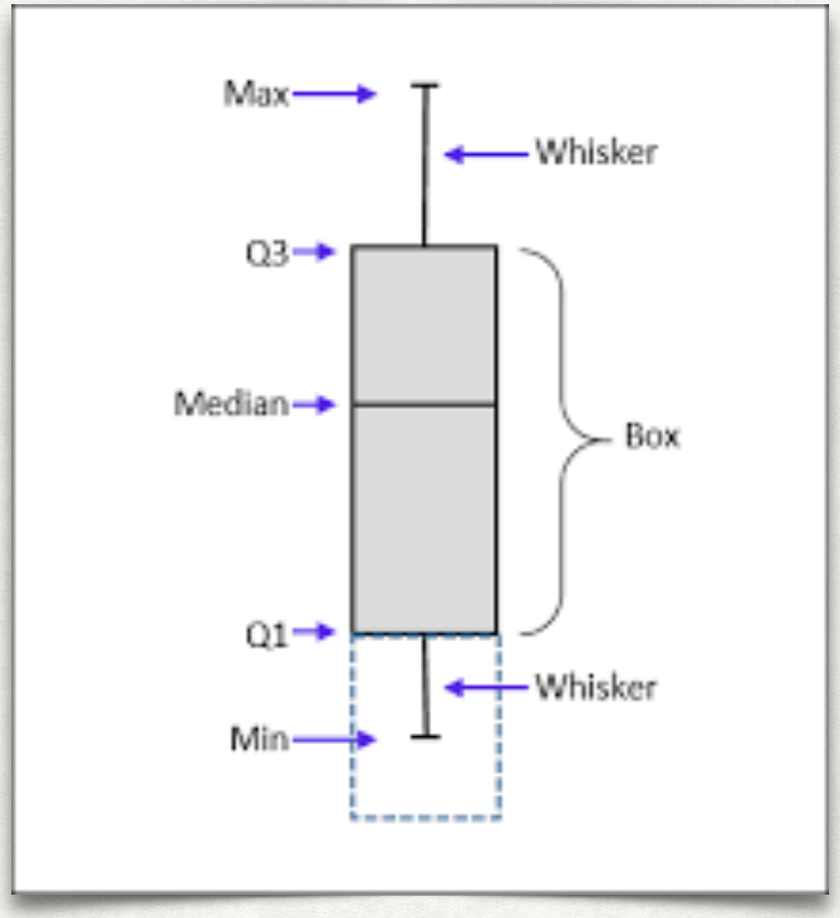
  Expresses a quality (e.g. sex, race, vital status). It does not imply any sorting, nor intermediate values between two values of the variable (e.g., the variable "sex" contemplates only the values M or F, without any sorting or intermediate values).

GRAPHICAL REPRESENTATION OF DATA

CONTINUOUS VARIABLE

A numerical variable can be represented graphically using a **histogram** or **box-plot**.

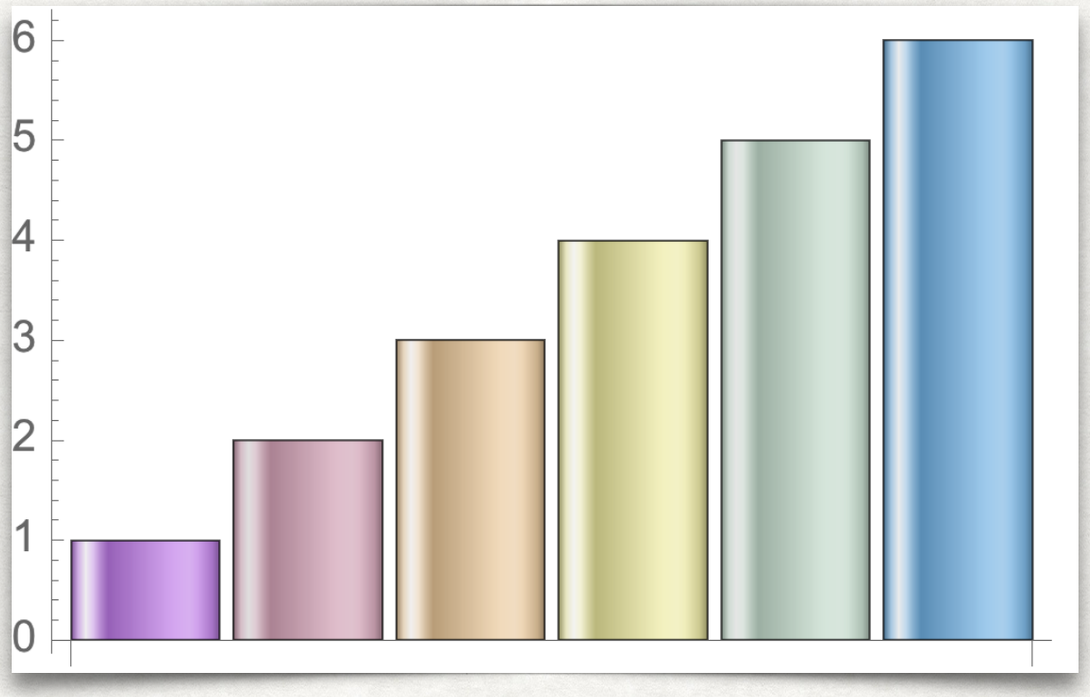


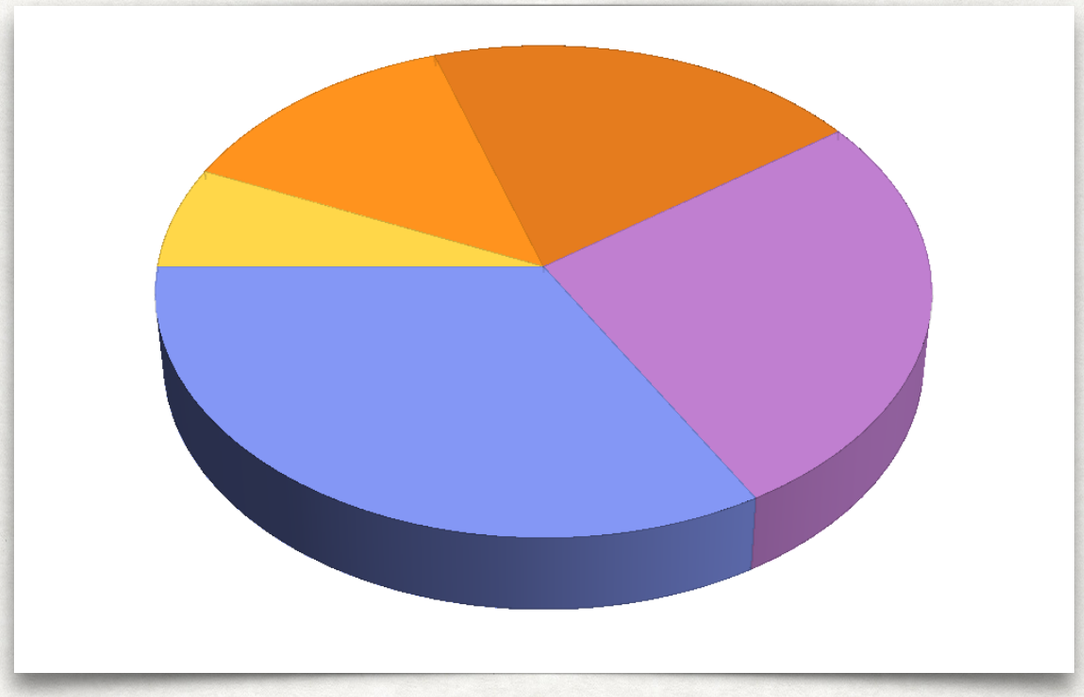


GRAPHICAL REPRESENTATION OF DATA

ORDINAL / NOMINAL VARIABLE

An ordinal or nominal variable is represented with a **bar** or **pie** chart.





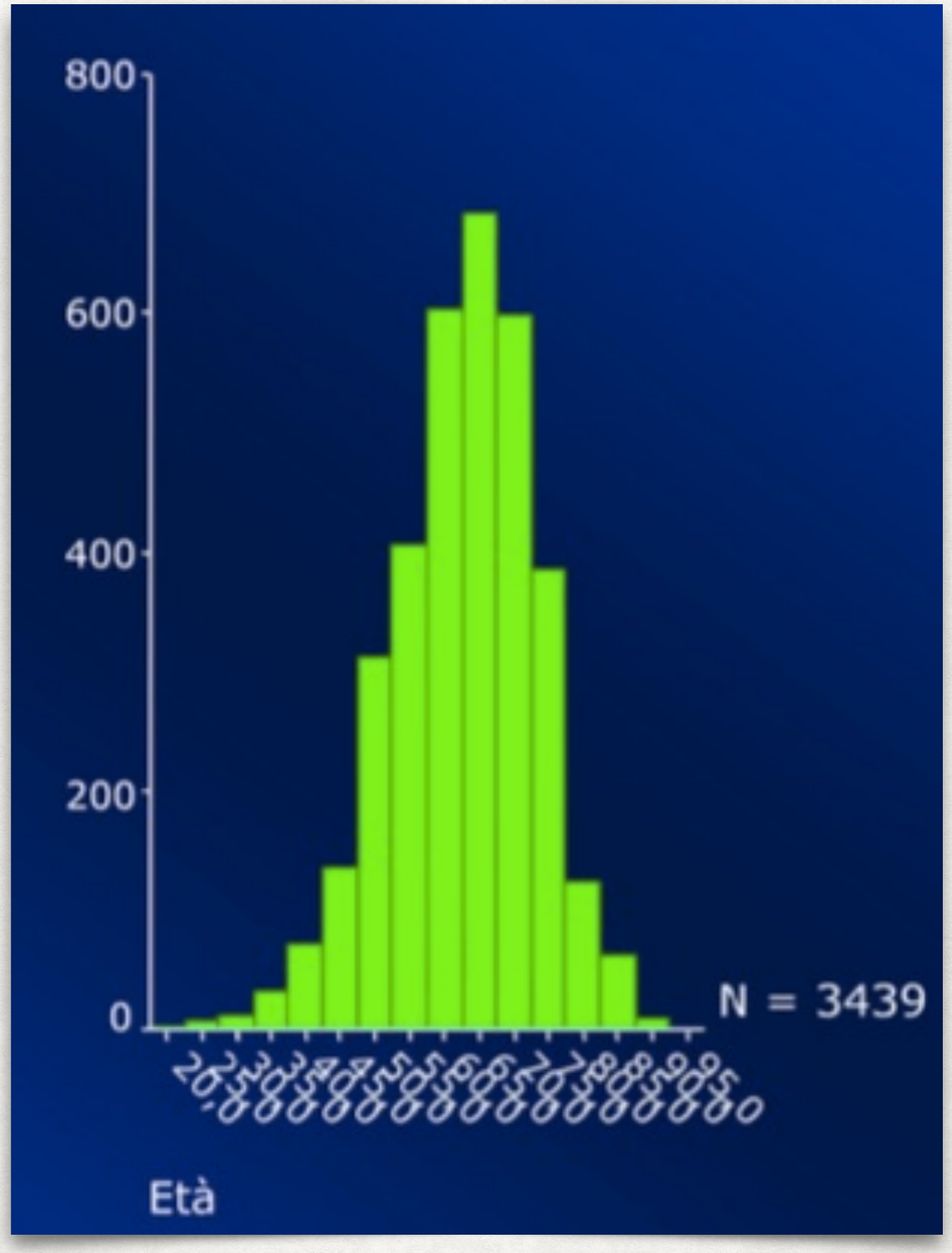
A discrete/ordinal variable can also be represented with a box-plot.

GRAPHICAL REPRESENTATION OF DATA

Histogram for Numerical Variables

The histogram shows the distribution of values by dividing the values into equally spaced intervals and plotting as a bar the count of cases in each interval.

- *A histogram is constructed dividing the values of the variable in object in intervals equally spaced, and bringing back under form of column the number of subjects that introduce a value inside the interval.*
- *The height of the column is proportional to the number of subjects in that range.*
- *In the example in the figure, the age histogram was plotted in a sample of 3439 subjects.*
- *Age was divided into 5-year intervals, and a count was made of how many subjects had an age that fell within each of the intervals.*
- *From the graph, it can be seen that most of the subjects were between the ages of 50 and 70, while only a few individuals were under 30 or over 80.*

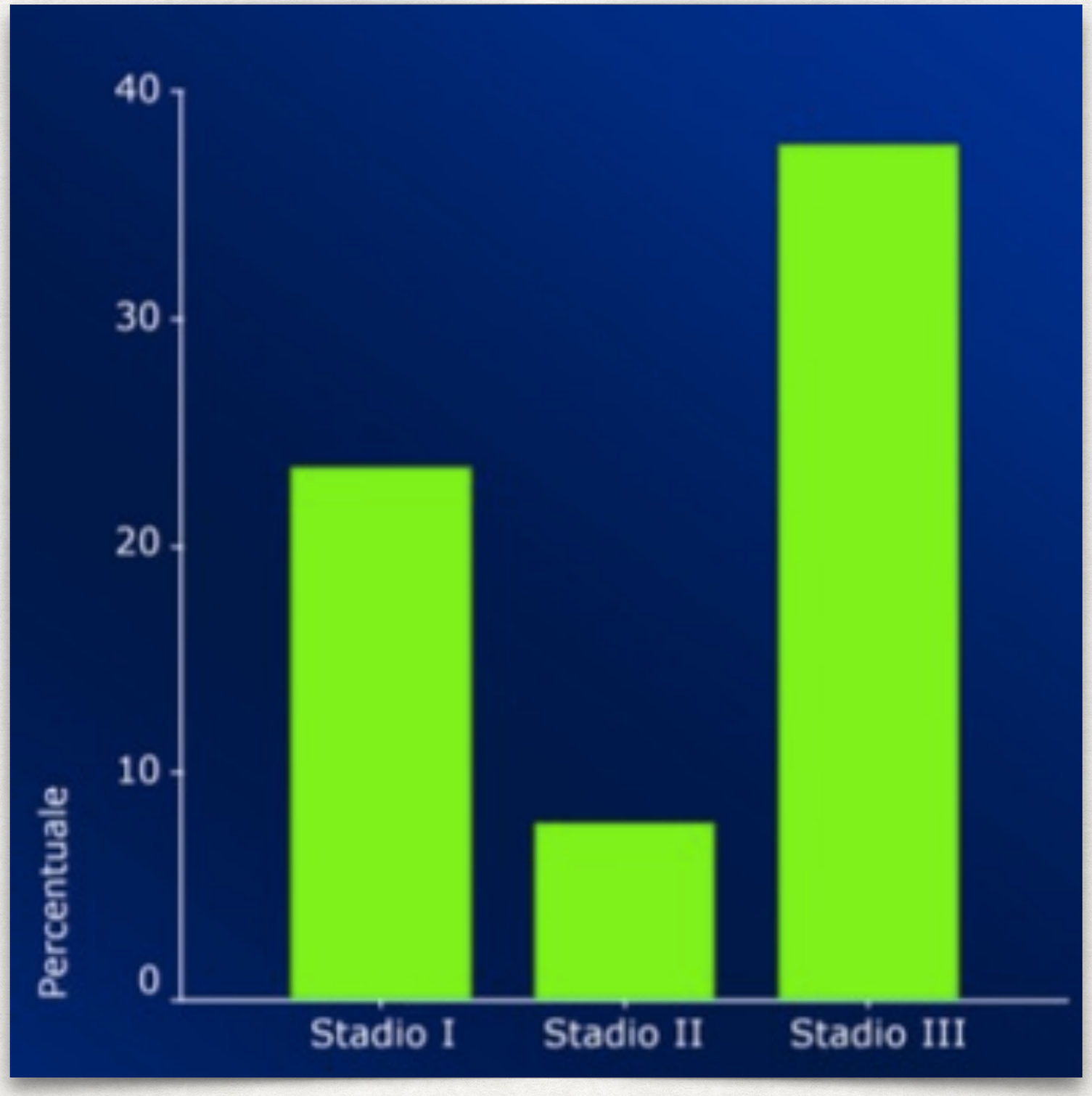


GRAPHICAL REPRESENTATION OF DATA

Bar Graph for Ordinal or Nominal Variables

The bar graph makes it easy to visually compare the categories of an ordinal variable.

- *Bar graphs can be used in the case of ordinal or nominal variables.*
- *Similar to the histogram, the height of each bar will depend on the number of subjects in that class.*
- *However, it is usually preferred to report data as percentages rather than absolute numbers.*
- *In the example shown in the figure, we can see what the percentage of subjects with stage I, II, and III disease is.*

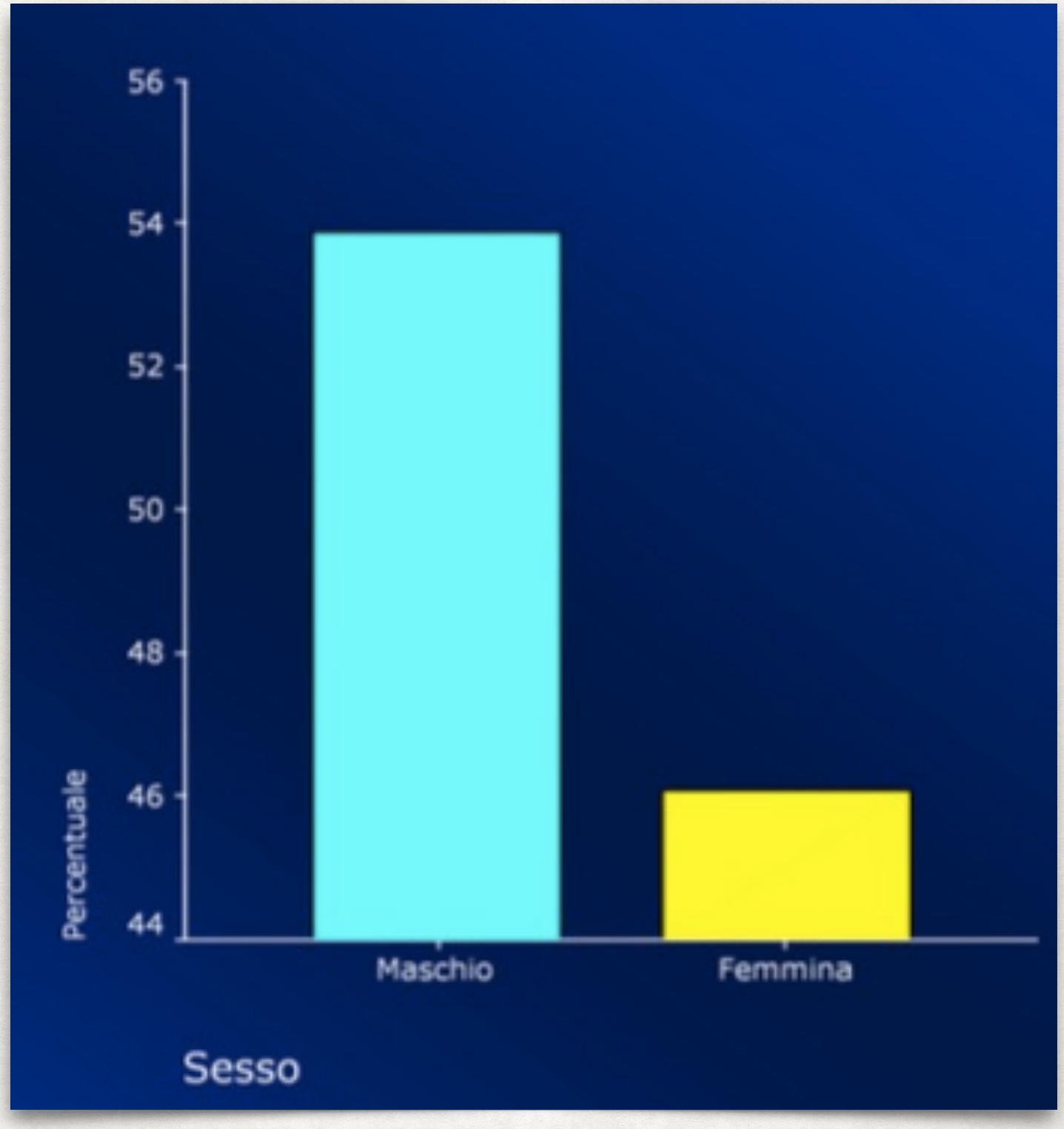


GRAPHICAL REPRESENTATION OF DATA

Bar Graph for Ordinal or Nominal Variables

The bar graph allows you to present the summary characteristics of a nominal variable.

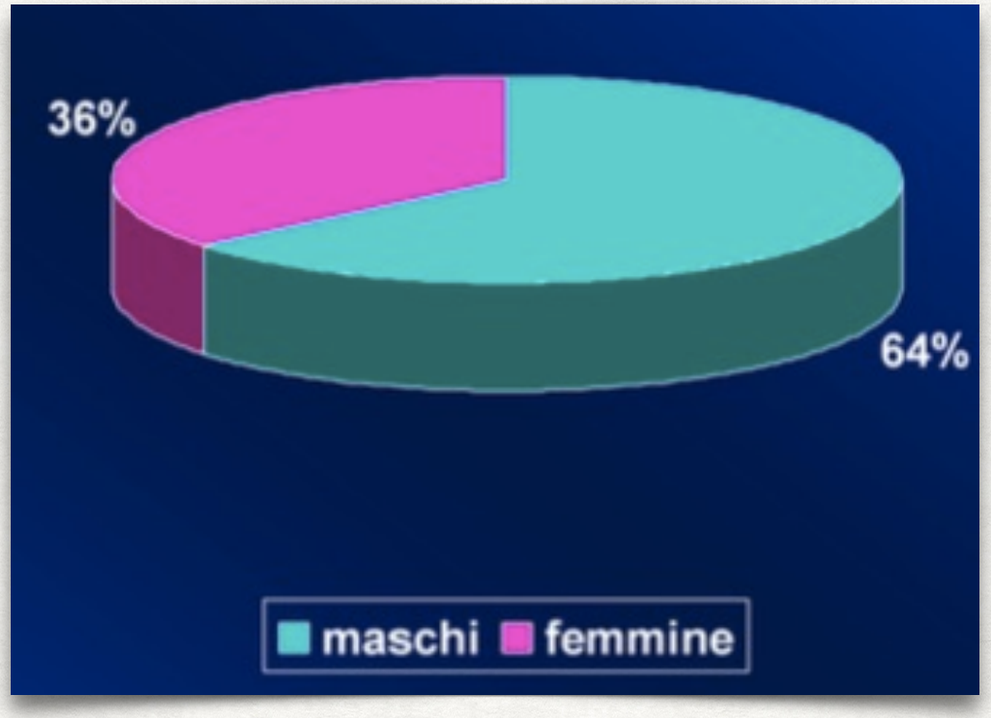
- *Similarly, for a nominal variable (in this case, sex), the height of the bars depends on the number (or percentage) of subjects in each of the classes.*



GRAPHICAL REPRESENTATION OF DATA

PIE CHART

Like the bar graph, the pie chart allows the percentage of subjects in each class (category) of the variable to be presented graphically.



- *The pie chart is an alternative way of representing ordinal or nominal variables.*
- *However, the graph becomes difficult to read if the variable is divided into many classes.*

DATA SUMMARY MEASURES

DATA SUMMARY MEASURES

MEAN, MEDIAN, MODE, RANGE, STANDARD DEVIATION

In addition to graphically, it is necessary to summarize the available information in numerical form.

In this regard, for an adequate description of the distribution of values of a variable in the population under study we need two **measures**, one of *central tendency*, the other of *dispersion*.

- *The measure of central tendency tells us around which value the observations tend to cluster, while the measure of dispersion will tell us by how much the individual observations deviate from the central value.*
- *For a continuous variable, the measure of central tendency is the **mean**, and the measure of dispersion is the **standard deviation**.*
- *For discrete variables, the **median** and **range** are used instead.*

DATA SUMMARY MEASURES

MEAN, MEDIAN, MODE, RANGE, STANDARD DEVIATION

Numerical variables (continuous and discrete) can then be summarized by a measure of **central tendency** and a measure of **dispersion**.

Variable	Measure of central tendency	Measure of dispersion
Continuous	Mean (**μ**)	Standard deviation (**σ**)
Discrete	Median	Range

- For any type of variable, it is also possible to calculate the **mode**, that is, the most frequent value in the distribution.
- For ordinal and nominal variables it is not possible to use summary measures (the values of the variable are categories, ordered or not), but only the **percentage of cases** in each category.

ARITHMETIC MEAN μ

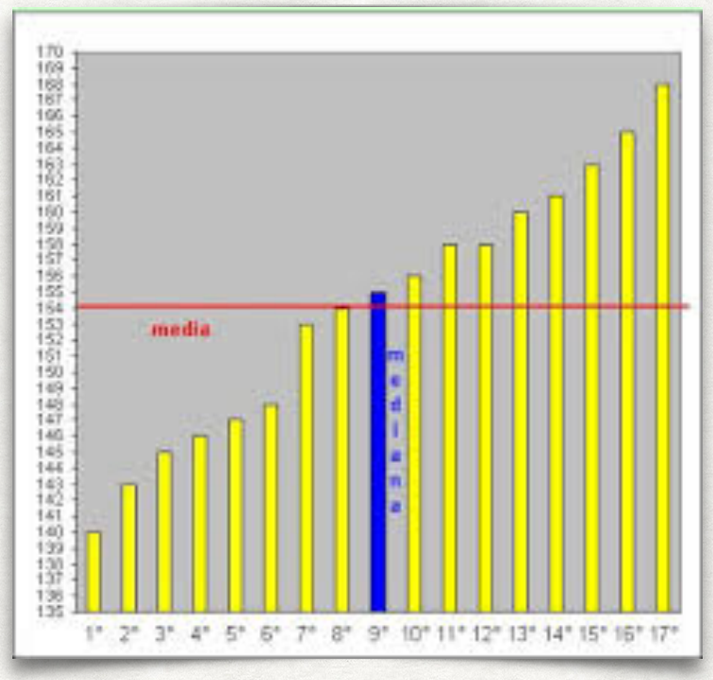
Sum (Σ) of the values of a variable (X) for each observation divided by the number of observations (N):

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

- If for example we have the five observations 2, 12, 6, 3, 7 their average will be: (2 + 12 + 6 + 3 + 7) / 5 = 6
- If instead we have the six measures 1, 2, 4, 5, 10, 20 the mean will be: (1 + 2 + 4 + 5 + 10 + 20) / 6 = 7

MEDIAN

Value that divides the population in half.



- It is the central value for an odd number of observations, and the average of the two central values for an even number of observations.
- To determine it, you must first arrange the values in ascending order.
- If for example we have the 5 observations: 2, 3, 6, 7, 12 the median will be: 6.
- If we have the 6 observations: 2, 3, 6, 7, 12, 15 the median will be: (6+7)/2 = 6.5
- In other words, it represents that value that divides the population in half, so that 50% have a value equal to or higher than the median, and 50% have a value equal to or lower than the median.

STANDARD DEVIATION σ (and VARIANCE σ^2)

Measures how far the values of a distribution deviate from the mean.

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

- If to es. we have the 5 observations: 2, 3, 6, 7, 12 the average is: 6
- The variance σ^2 is given by: $[(2-6)^2+(3-6)^2+(6-6)^2+(7-6)^2+(12-6)^2] / 5$
  $= [16+9+0+1+36] / 5 = 62 / 5 = 12.4$
- So the standard deviation σ is given by the square root of 12.4 i.e. 3.52

RANGE

Refers to the minimum and maximum value found in the study sample.

- Arranging the values in ascending order, the range is given by the first and the last value in the sequence.
- If for example we have the 5 observations: 2, 3, 6, 7, 12 the range is: 2 — 12
- In the case of an ordinal variable, the range (i.e. the minimum and maximum values found) measures the dispersion.

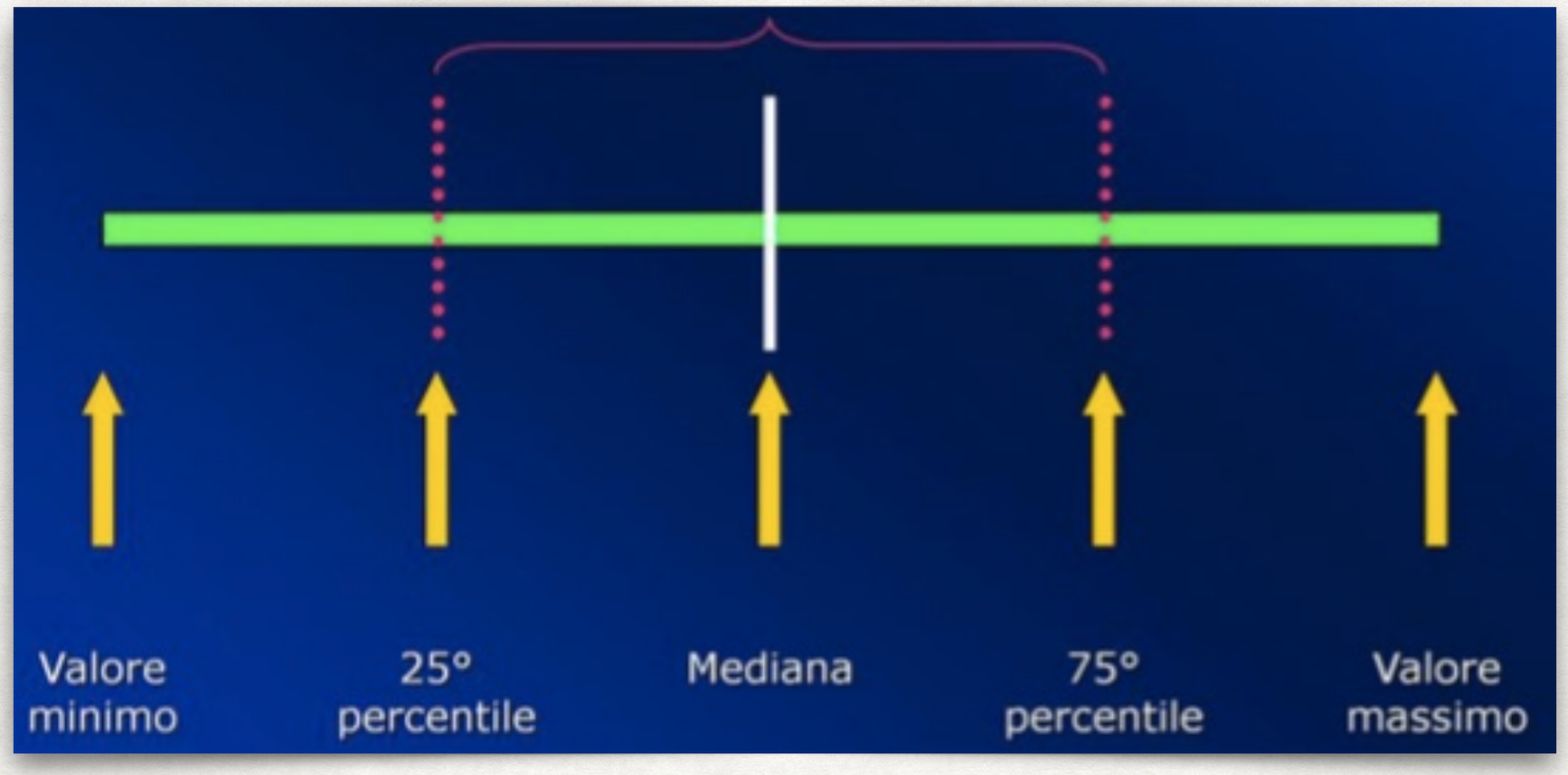
INTERQUARTILE RANGE

Refers to values for the first and third quartiles (25th and 75th percentiles).

- The range could give an inaccurate information; in fact, the highest and the lowest value could represent atypical extreme values, and not reflect the real variability of the measure in the population under examination.
- If, for example, we had 100 subjects, all aged 50 except for two subjects, one aged 10 and one aged 90, the range (10–90) might lead us to think that we are dealing with a population of much more variable age than it actually is.
- For this reason, the interquartile range is usually used.
- For example, quartiles are the values that divide the population into four groups, each of which contains 25% of the sample.

INTERQUARTILE RANGE

Refers to values for the first and third quartiles (25th and 75th percentiles).

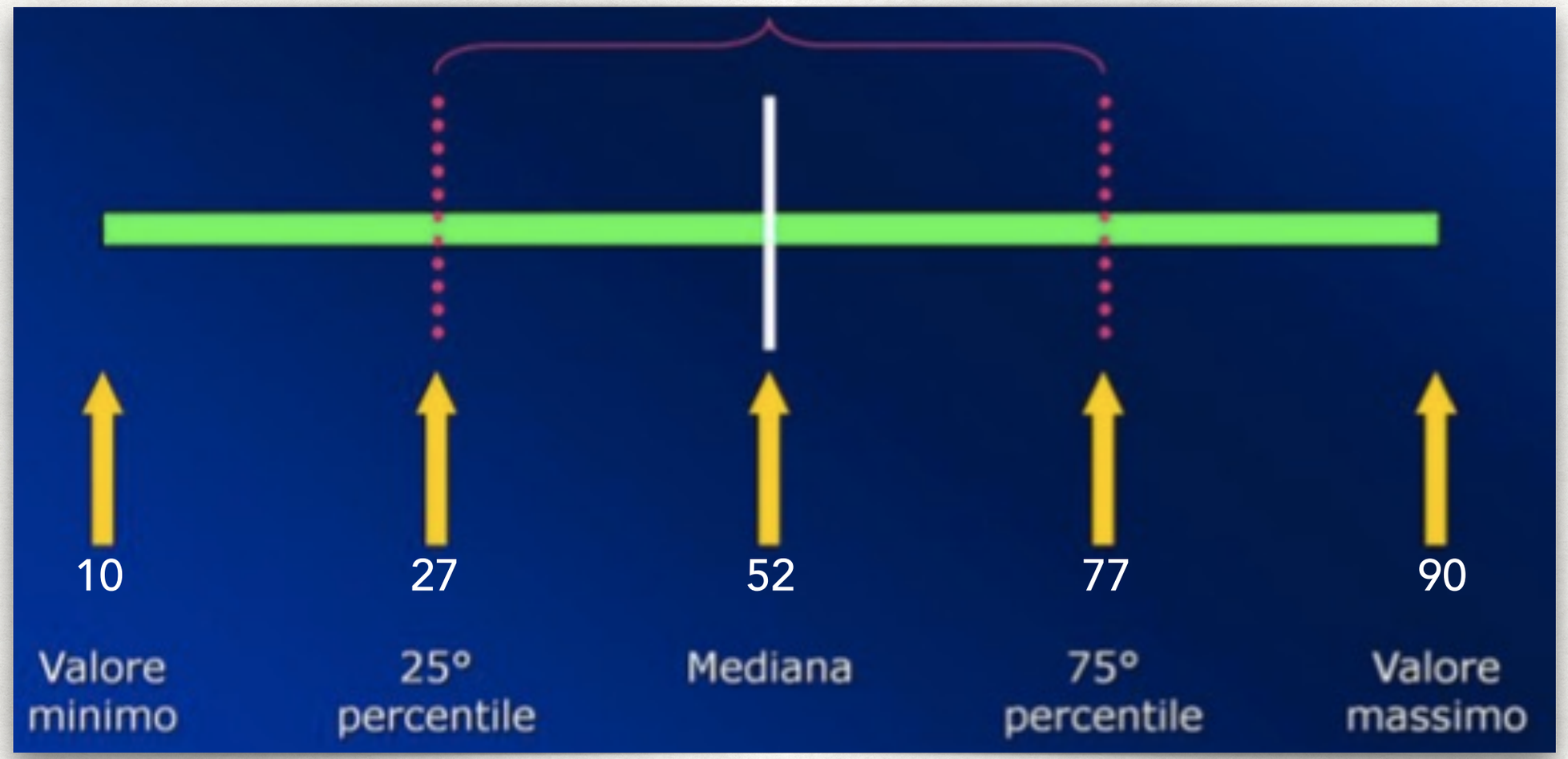


- More generally, *percentiles* are the values of a variable that divide the population under study into groups of equal numerosity.
- As is easy to guess, the median corresponds to the 50th percentile.
- The interquartile range is then defined by the 25th and 75th percentile values.

DATA SUMMARY MEASURES

EXAMPLE

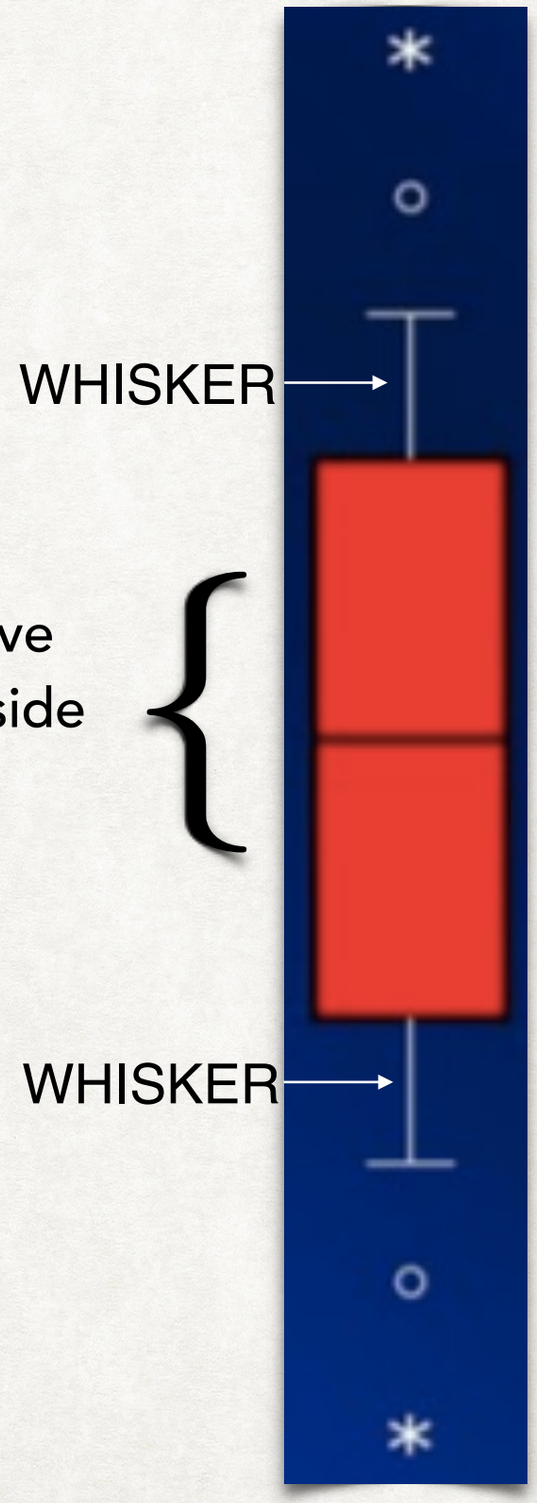
Median age of patient sample = 52 years.



- Range: 10 — 90 years.
- Interquartile range: 27 — 77 years.
- 25% are under 27 years old, 50% are between 27 and 77 years old, and 25% are over 77 years old.

DATA REPRESENTATION

BOX-PLOT



Far outlier: value > Q3 + 3×(Q3–Q1)

Outlier: value > Q3+1.5×(Q3–Q1)

Max: highest non-outlier value

Q3 (75th PERCENTILE)

MEDIAN

Q1 (25th PERCENTILE)

Min: lowest non-outlier value

Outlier: value < Q1 – 1.5×(Q3–Q1)

Far outlier: value < Q1 – 3×(Q3–Q1)

DATA REPRESENTATION

BOX-PLOT

- The definition of percentiles and interquartile range allows us to describe another graphical mode of data representation, represented by **box plots**.
- The box plot is a "box" shaped graph that contains a lot of information about our variable.
- The central line that cuts through the box represents the median, while the extremes of the box represent the interquartile range; in other words, 50% of the observations have a value within the box.
- The graph also shows us the so-called "outliers," i.e., those values that are lower than the 25th percentile by more than 1.5 times the length of the box or that are higher than the 75th percentile by more than 1.5 times the length of the box.

DATA REPRESENTATION

BOX-PLOT

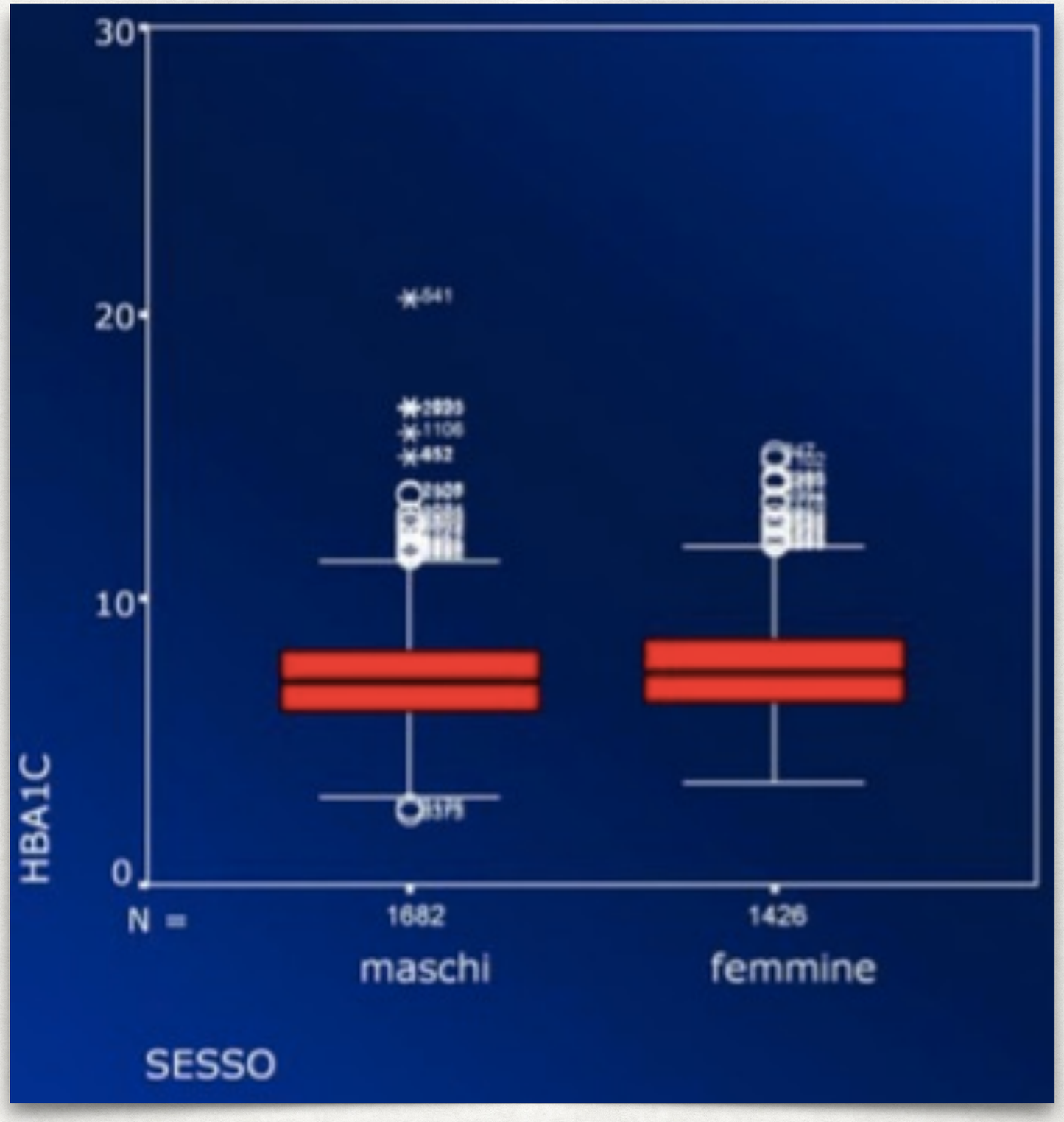
- Finally, extreme values are represented, i.e., those values that deviate from the 25th and 75th percentiles by more than 3 times the length of the box.
- This information is important because it allows us to identify those values that need to be verified, because they could be errors in data entry.
- If they are not typing errors, it is still necessary to check whether they are biologically plausible values or measurement errors.
- If the values are plausible, they should obviously not be removed.

DATA REPRESENTATION

BOX-PLOT

- In this example, box plots were used to graphically summarize the values of glycosylated hemoglobin (HbA_{1c}) in a population of individuals with diabetes, summarizing the data separately for males and females.

- The graph shows that, while in females there are only some outliers, marked by the small circles, among males there are also some extreme values, represented by the asterisks.



DATA REPRESENTATION

TABULAR FORM

- Data, in addition to graphical form, are generally summarized in **tabular form**.
- **Numeric** variables are expressed as mean and standard deviation, or as median and range (or interquartile range).
- **Categorical** variables are expressed as percentages.

This slide shows an example of a tabular summary of the characteristics of a population, divided by study branch.

Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial

	Diastolic blood pressure target group		
	≤90 mm Hg (n=6264)	≤85 mm Hg (n=6264)	≤80 mm Hg (n=6262)
Age (years)	61·5 (7·5)	61·5 (7·5)	61·5 (7·5)
Body-mass index (kg/m²)	28·4 (4·7)	28·5 (4·7)	28·4 (4·6)
Diastolic blood pressure (mm Hg)	105 (3·4)	105 (3·4)	105 (3·4)
Systolic blood pressure (mm Hg)	170 (14·4)	170 (14·0)	170 (14·1)
Serum creatinine (μmol/L)	89 (26)	89 (23)	89 (23)
Serum cholesterol (mmol/L)	6·0 (1·1)	6·1 (1·1)	6·1 (1·2)
Men/women (%)	53/47	53/47	53/47
Previous treatment (%)	52·3	52·7	52·6
Smokers (%)	15·9	15·8	15·9
Previous MI (%)	1·6	1·5	1·5
Other previous CHD (%)	5·9	6·0	5·9
Previous stroke (%)	1·2	1·2	1·2
Diabetes mellitus (%)	8·0	8·0	8·0

Data are mean (SD) or % of group. MI=myocardial infarction; CHD=coronary heart disease.

Table 1: **Characteristics at randomisation**

Lancet1998; 351: 1755–62

NORMAL (GAUSSIAN) AND ASYMMETRIC DISTRIBUTION

NORMAL (GAUSSIAN) DISTRIBUTION

DEFINITIONS

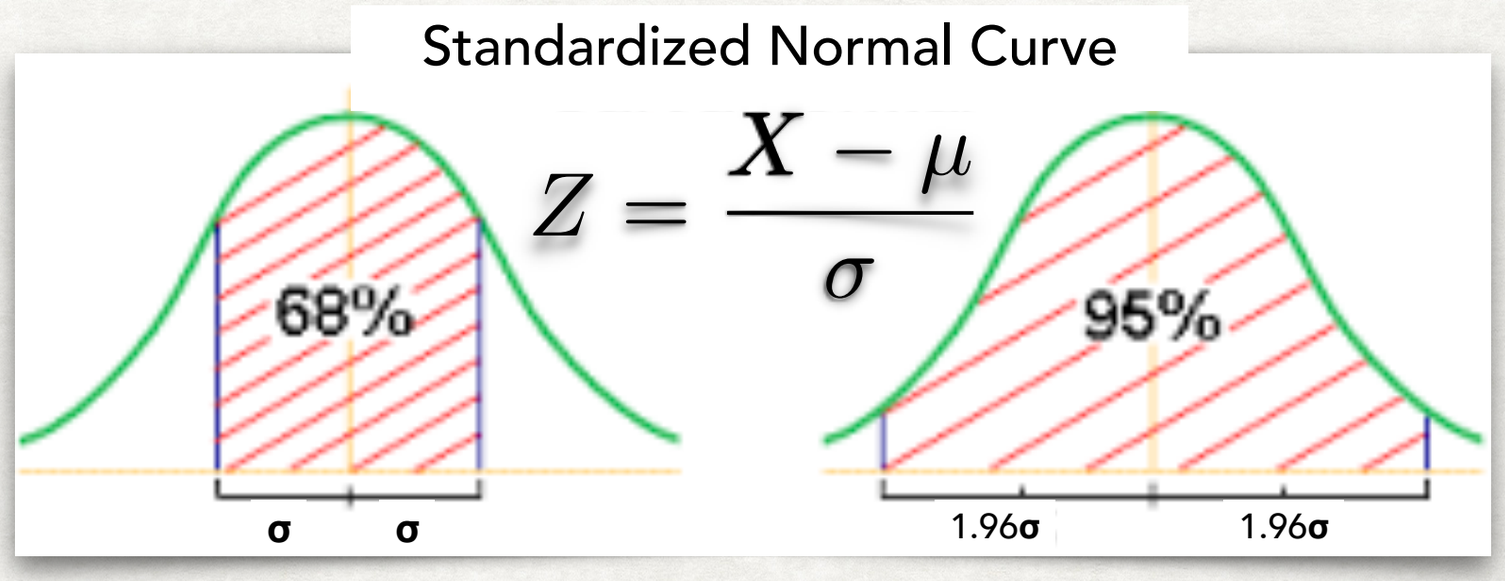
A continuous variable X is **normally distributed** when the mean, mode, and median coincide.

Also, a normal distribution Z is **standardized** if it has zero mean and unit standard deviation (obtained by subtracting the mean from each observation and dividing by the standard deviation). In a normal distribution:

- about 2/3 of the observations (68%) are within 2 standard deviations (μ±1σ);
- about 95% of the observations are within almost 4 standard deviations (μ±1.96σ);
- about 99% of observations are within almost 6 standard deviations (μ±2.58σ).

Furthermore, in a normal distribution:

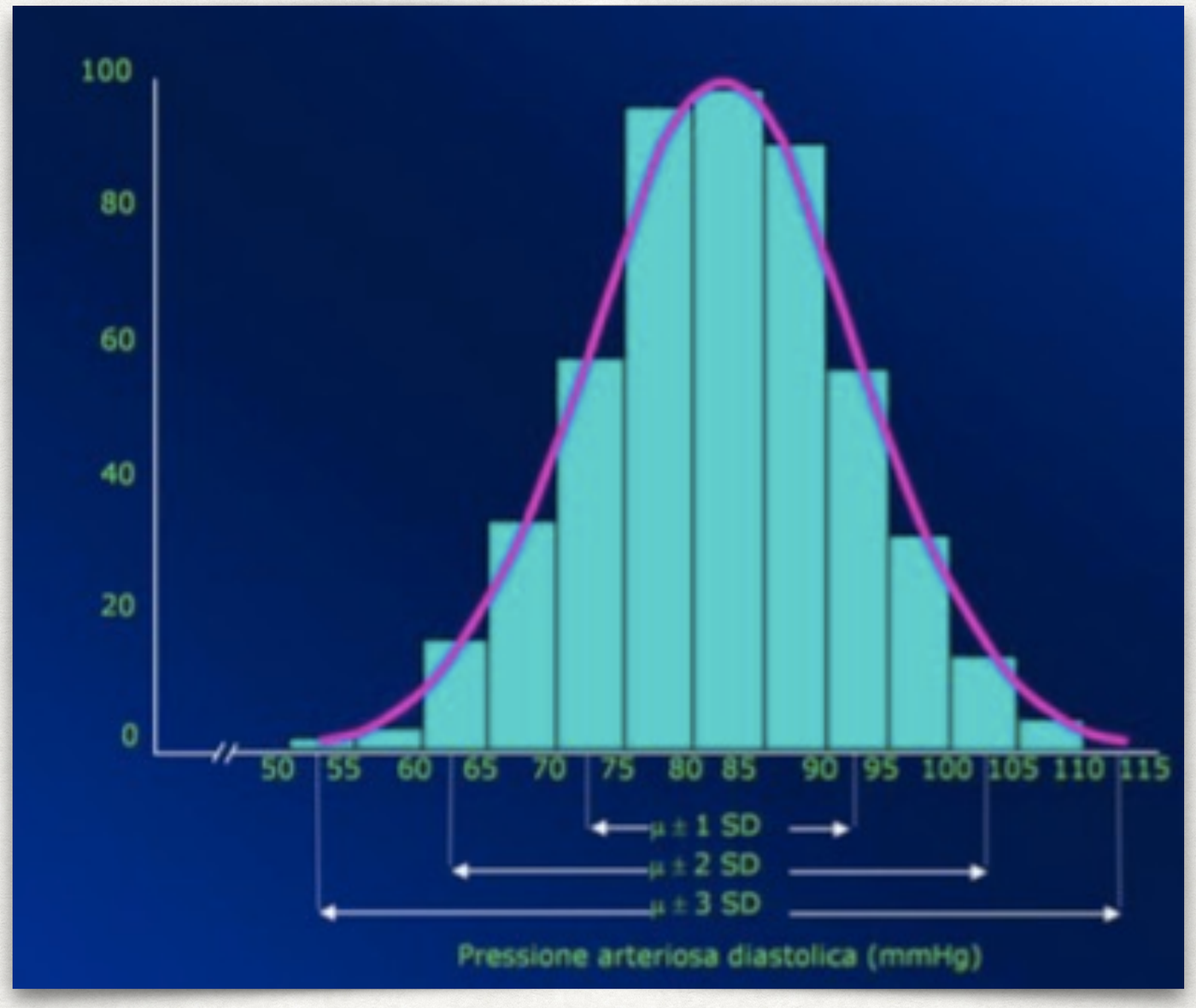
- 2.5th percentile ≃ μ-2σ
- 16th percentile = μ-1σ
- 50th percentile = μ
- 84th percentile = μ+1σ
- 97.5th percentile ≃ μ+2σ



NORMAL (GAUSSIAN) DISTRIBUTION

DEFINITIONS

- The graph shows a classic normal distribution, referring to the diastolic pressure values in a population.
- From the image it can be seen that, in the presence of a normal distribution, many subjects tend to have a value close to the average, while, as we move away from the average value, we find fewer and fewer subjects.
- In particular, only 2.5% of subjects will have a value higher than μ+1.96σ, and similarly only 2.5% of subjects will have a value lower than μ-1.96σ.

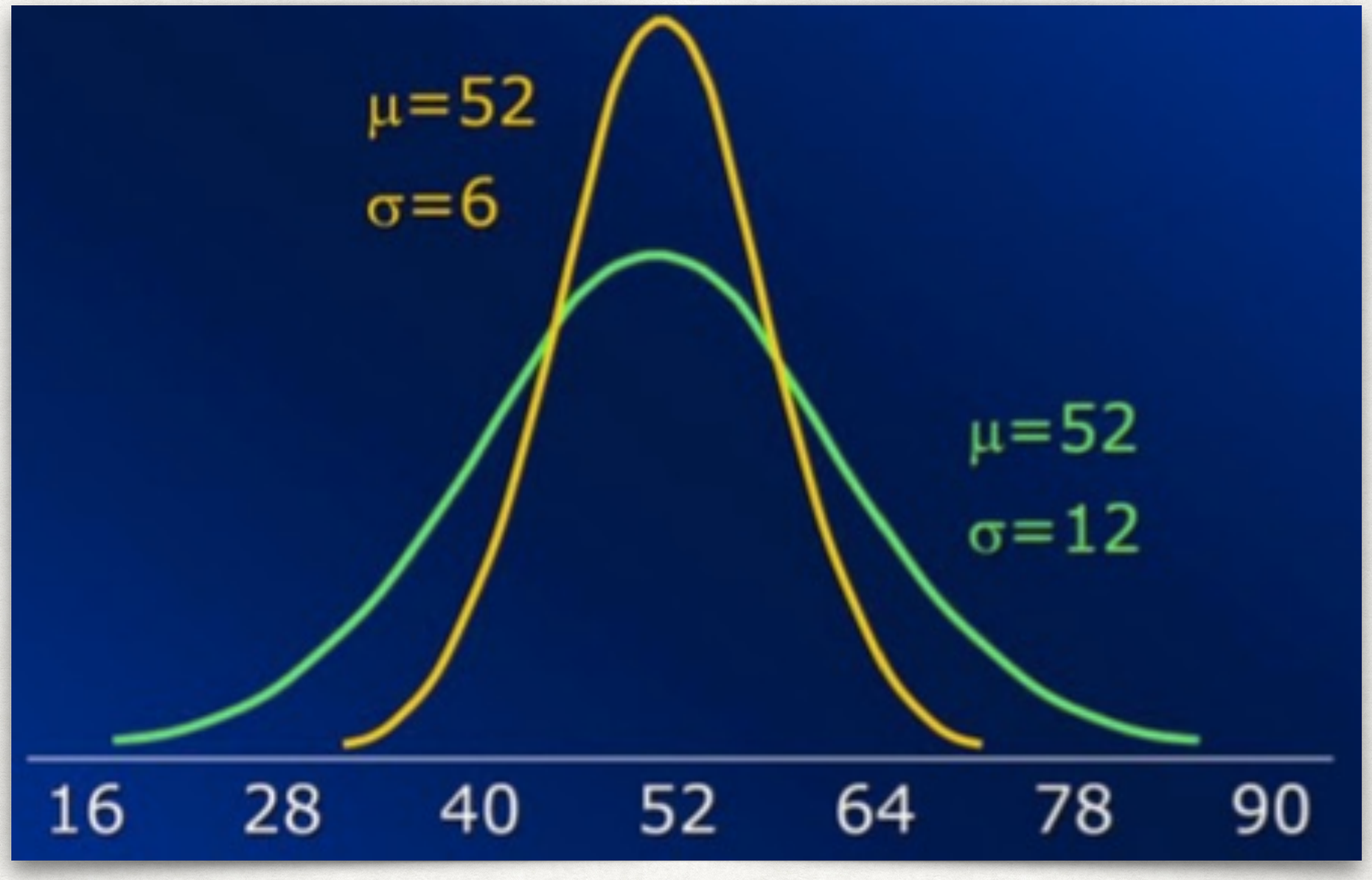


- In addition, subjects with values greater than μ+2.58σ and those with values less than μ-2.58σ will be only 0.5% in both cases.

NORMAL (GAUSSIAN) DISTRIBUTION

DEFINITIONS

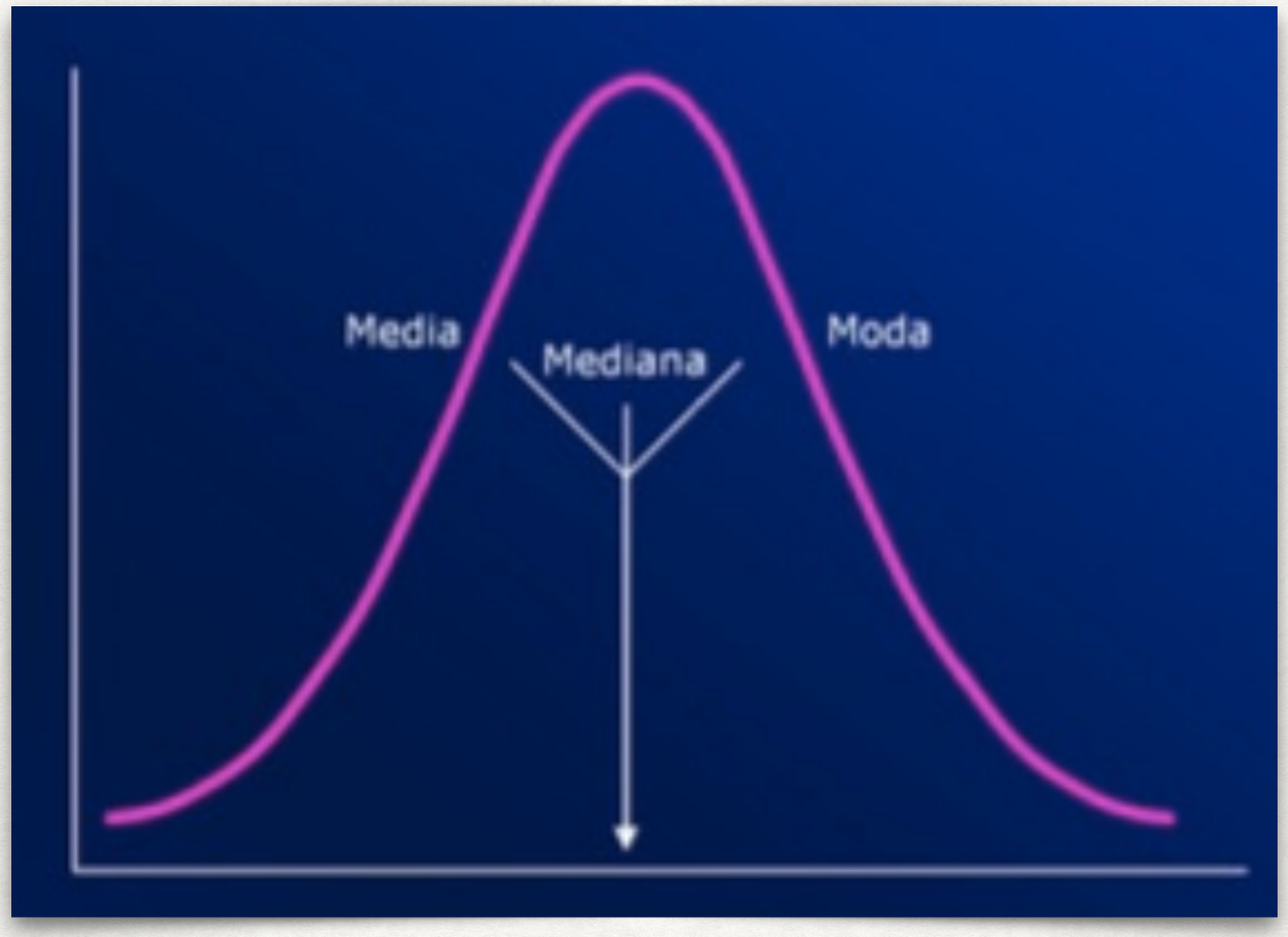
- The two curves represent the age distribution in two samples.
- As can be seen, the mean value is the same, but, in the case of the green curve, the standard deviation is double that of the yellow curve.



- This means that the subjects belonging to the population represented in yellow have a more homogeneous and less variable age than those belonging to the population in green.

NORMAL (GAUSSIAN) DISTRIBUTION

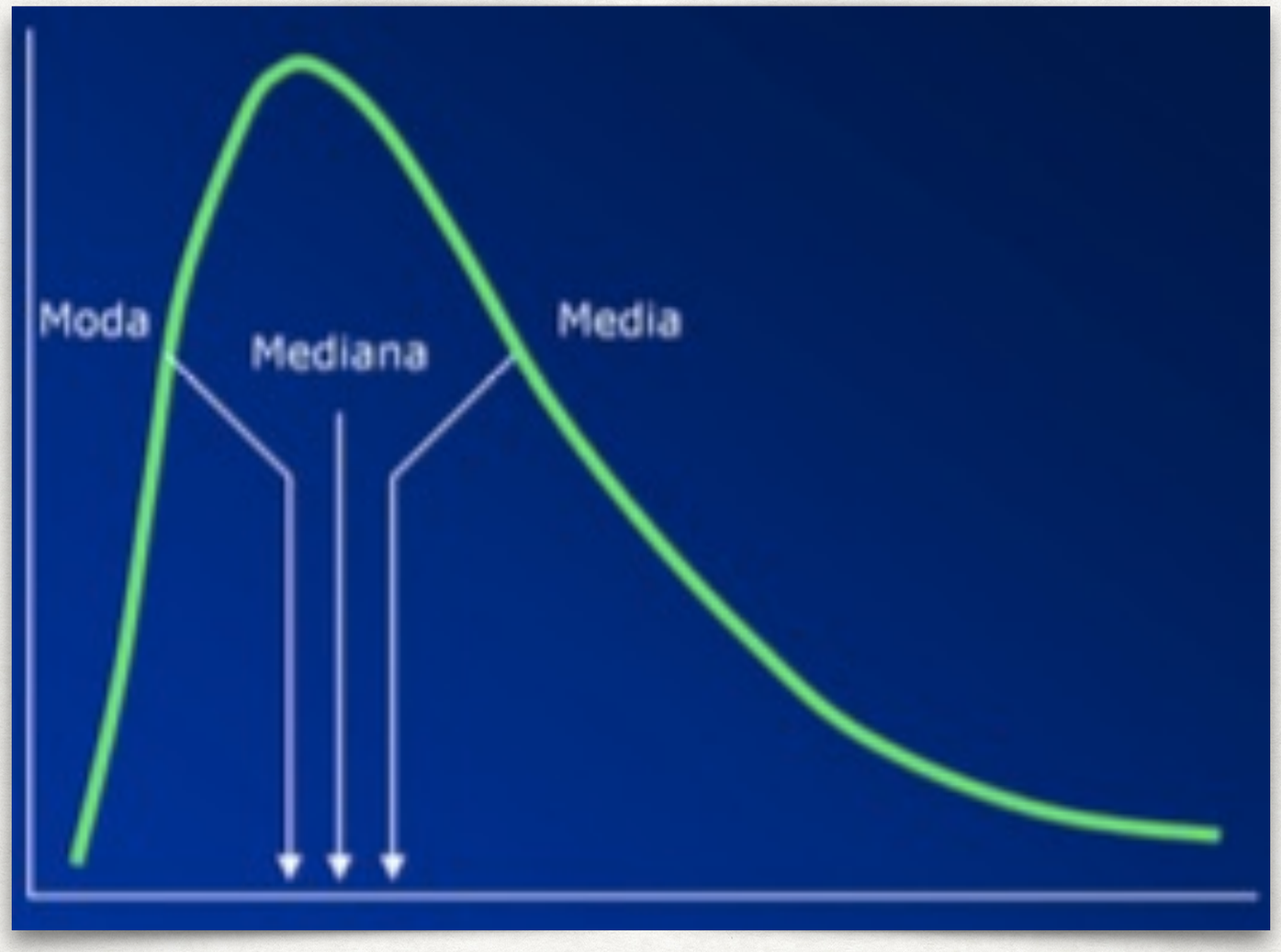
DEFINITIONS



- As already mentioned, one of the conditions for a continuous variable to be considered normally distributed is the coincidence of mean, mode and median.
- In other words, we can say that the curve is symmetrical with respect to the central value.

ASYMMETRIC DISTRIBUTION

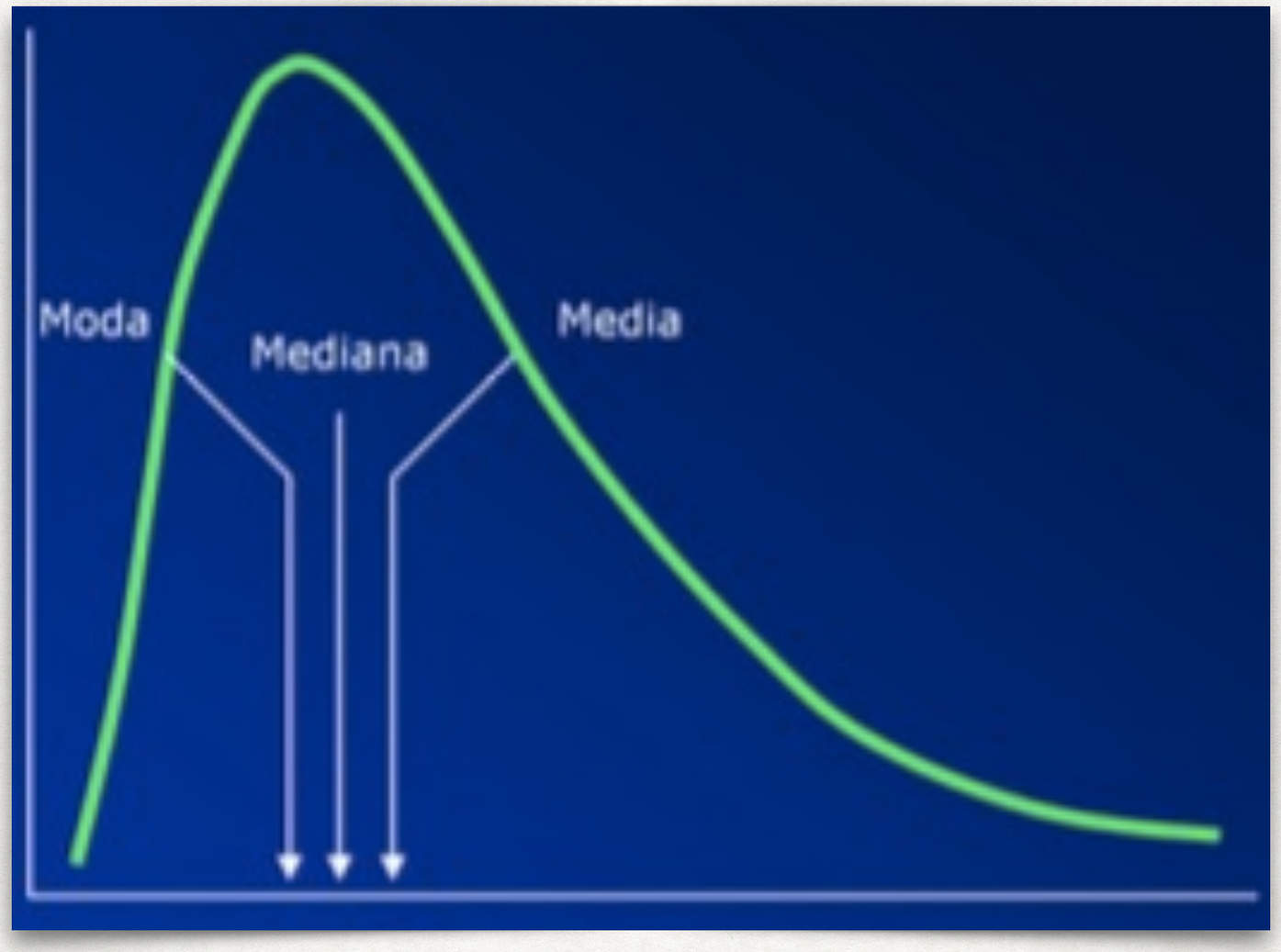
RIGHT SKEWNESS



- However, it may happen that a continuous variable takes on values that are not symmetrically distributed.
- For example, in some cases there may be a certain percentage of subjects who have a higher value than the rest of the sample, giving the curve an asymmetrical appearance with a "tail" to the right.

ASYMMETRIC DISTRIBUTION

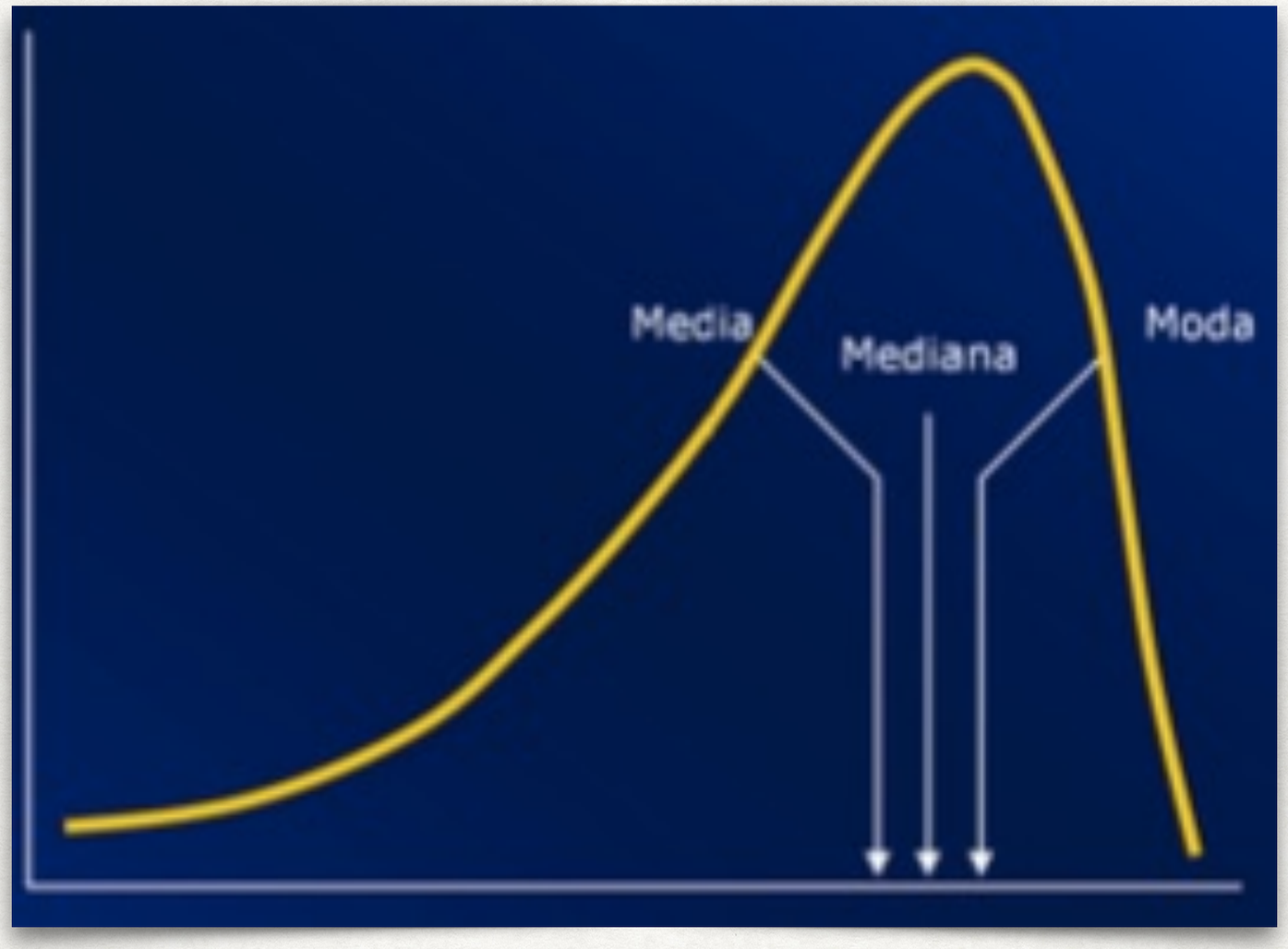
RIGHT SKEWNESS



- This could occur if, for example, blood glucose levels were measured in a population.
- Most subjects will have values in the normal range, but a small percentage may have consistently higher values because they have diabetes without knowing it.
- In these cases, the value of the mean will be higher than the median.

ASYMMETRIC DISTRIBUTION

Left Skewness

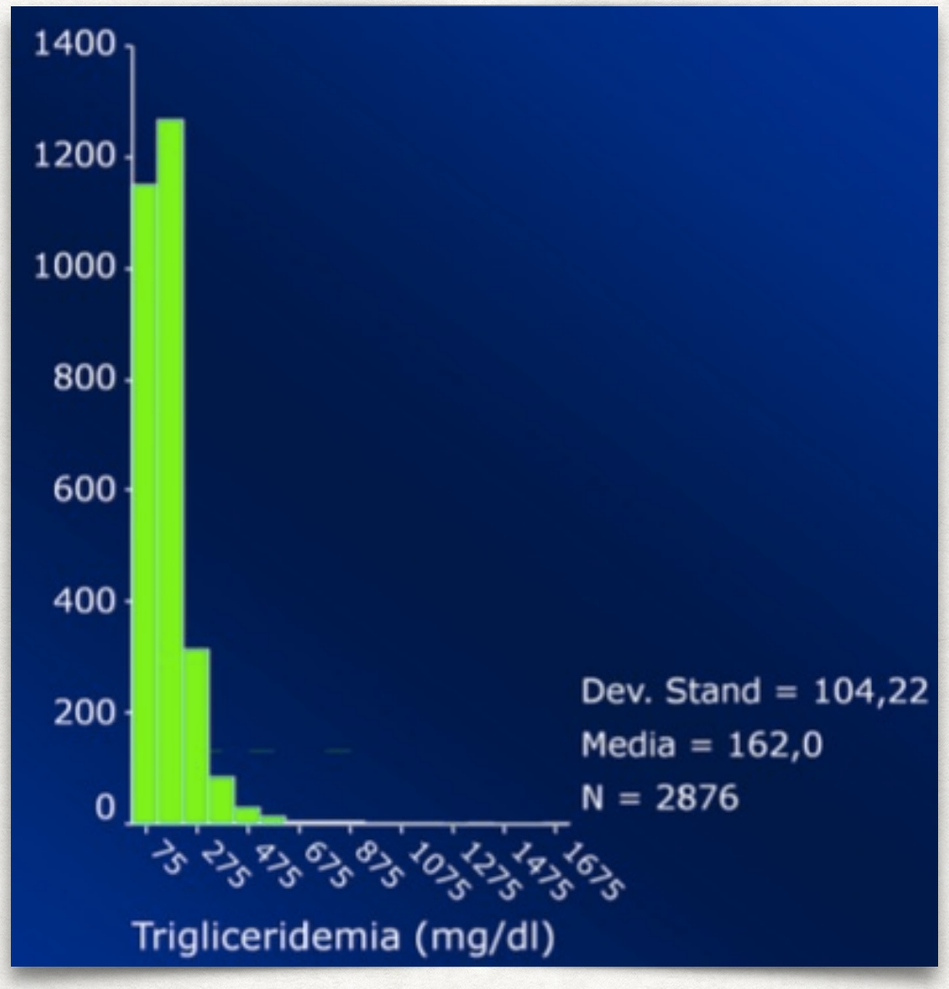


- In other cases, the opposite situation could occur, with a certain percentage of subjects presenting lower values than the rest of the population.
- In this case the curve will present a "tail" to the left and the value of the mean will be lower than that of the median.

DATA TRANSFORMATION

From Asymmetrical to Normal Distribution

- Sometimes a variable with a clearly asymmetrical distribution can be mathematically transformed so as to make its distribution normal.
- In the example shown in the graph, concerning the values of triglycerides in blood in a sample of subjects with diabetes, you can clearly see how the distribution is asymmetric, with a "tail" on the right.
- This distribution is determined by the presence of a certain proportion of patients with values of triglyceridemia markedly higher than the rest of the sample.

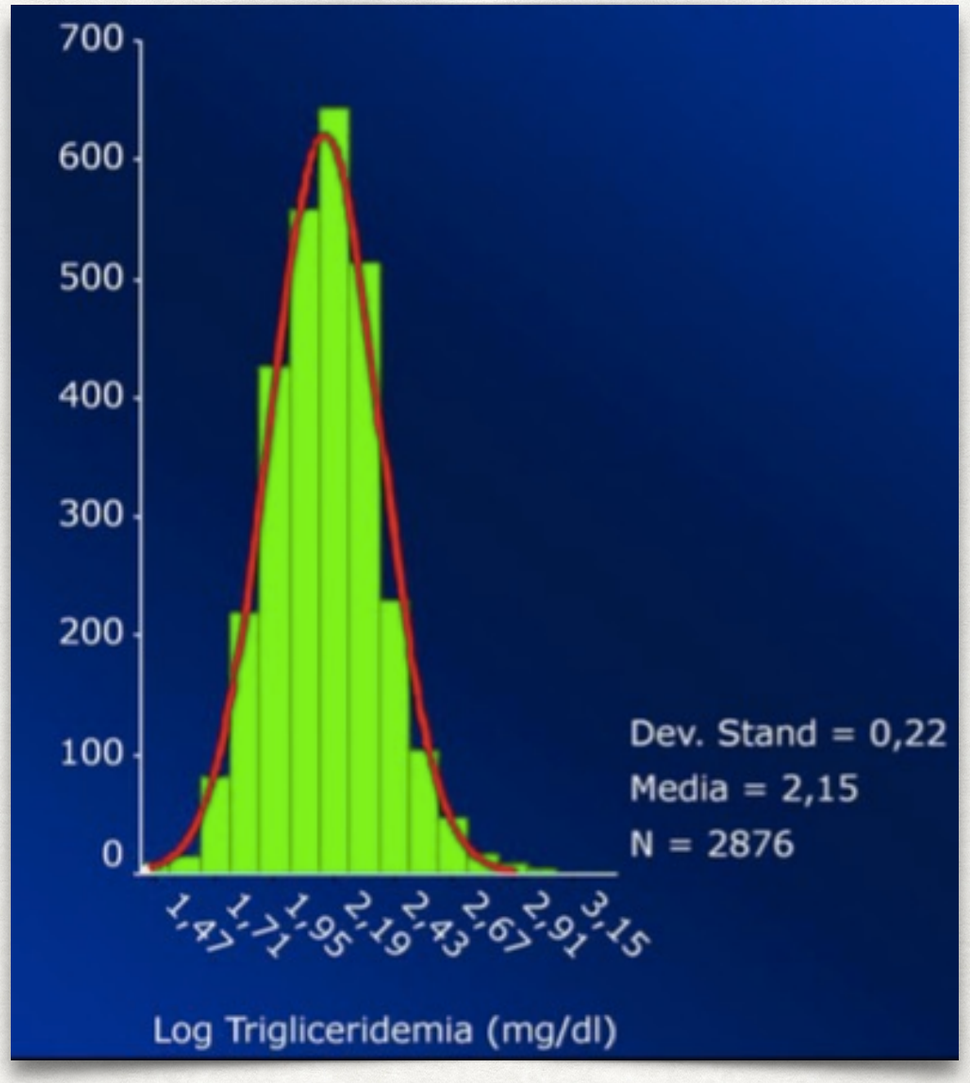


- That the distribution is not normal can also be seen by looking at the mean and standard deviation.
- As already pointed out, if the distribution were normal, 95% of the values should fall within the range $\mu \pm 1.96\sigma$.
- However, if we subtract from the mean value, equal to 162 mg/dl, the double of σ (ie, 208.44) we will obtain a negative number!
- In these cases a mathematical transformation of the variable can sometimes allow us to bring the distribution back to normality.

DATA TRANSFORMATION

From Asymmetrical to Normal Distribution

- This graph shows the distribution of the logarithm of triglyceridemia.
- Comparing this histogram with the previous one, we can clearly see how the distribution is now symmetrical and no longer has the long tail on the right, regaining the classic inverted-bell appearance.
- The use of logarithm, square root, or other mathematical functions can therefore sometimes be useful to obtain normality (we will see later that the requirement of normality of a continuous variable is essential for the use of some statistical tests, called *parametric*).



HYPOTHESIS TESTING

HYPOTHESIS TESTING

Once the data have been summarized in graphical or tabular form, the next step is to explore whether particular associations exist (e.g., "*Is there a relationship between obesity and metabolic control?*"; "*Are HbA_{1c} values higher in subjects with low levels of schooling?*").

STATISTICAL TESTS

Statistical tests are used to accept or reject (refute) a hypothesis

HYPOTHESIS TESTING

- Each statistical test starts with the **null hypothesis** of no differences.
- In the example above, the null hypothesis is represented by no differences in mean HbA_{1c} values based on BMI (no association).
- The purpose of the statistical test is to accept the null hypothesis or to refute it, thereby accepting the **alternative hypothesis** that significant differences exist.

HYPOTHESIS TESTING

All statistical tests start with the **null hypothesis** that there is no relationship between the variables under study.

If, for example, we wanted to test the hypothesis that, among subjects with diabetes, the level of metabolic control depends on the level of obesity, expressed by body mass index (or BMI), the starting null hypothesis would be represented by a complete absence of correlation between metabolic control and BMI.

The purpose of the statistical test will be to suggest whether to accept this null hypothesis, or to reject it, thereby accepting the **alternative hypothesis** that there is a significant relationship between metabolic control and level of obesity.

HYPOTHESIS TESTING

- Because the subjects studied, however numerous they may be, always represent a small sample of all those potentially enrollable and because of biological variability, **the results of a statistical test will always be in terms of probability**.
- This is because **the sample we study may not be representative** of the universe of patients and the conclusions we reach may therefore be erroneous.
- In accepting or refuting the null hypothesis it is therefore **always possible to make a mistake**; if, however, the probability of making such a mistake is very low, then we will accept with "sufficient" confidence the conclusions we have arrived at.

By what rule do we accept or reject the null hypothesis?

***p*-value**

Indicates the probability of making a mistake by rejecting the null hypothesis, that is, the probability of making a mistake by asserting that there is a difference between the groups being compared.

$p \leq 0.05$ → Statistically significant result → We reject the null hypothesis

HYPOTHESIS TESTING

EXAMPLE

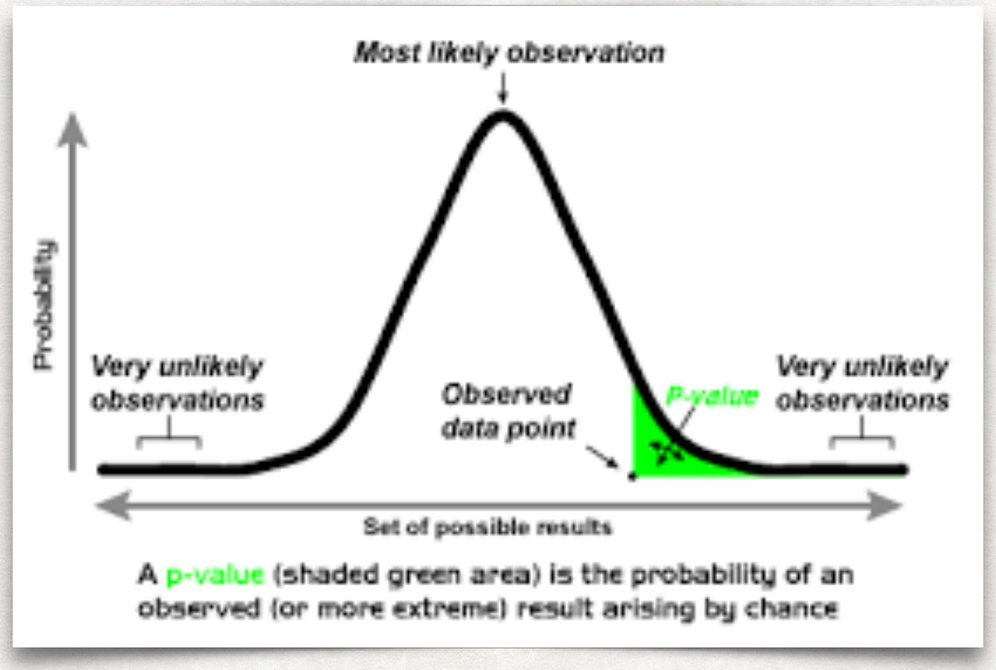
- Subjects with BMI between 25 and 27 have an HbA_{1c} of 7.28±1.7
- Subjects with BMI>30 have an HbA_{1c} of 7.51±1.6
- **Null hypothesis**: mean HbA_{1c} levels do not differ by BMI.
- **Alternative hypothesis**: mean HbA_{1c} levels are significantly higher in subjects with BMI>30 than in those with BMI between 25 and 27.

Using the appropriate statistical test [Student's *t*-test for unpaired data (*)] the value of *p* turns out to be 0.009

(*) *Later we will outline the criteria for choosing the appropriate statistical test in relation to the distribution.*

$p = 0.009$

- By rejecting the null hypothesis and thus stating that there is a difference in HbA_{1c} based on BMI levels, we will have a less than 1% chance of being wrong (9 in a thousand).
- In other words, the probability that there is no association is less than 1%.



HYPOTHESIS TESTING

EXAMPLE

A <u>statistically significant</u> result is not the same as saying it is <u>clinically relevant</u>.

- Suppose that treating 10,000 diabetics with hypoglycemic A yields HbA_{1c} values of 7.1±1.3 and that treating as many subjects with hypoglycemic B yields HbA_{1c} values of 7.3±1.4.
- Applying the appropriate statistical test, a $p<0.001$ value is obtained.
- The difference is statistically significant and is highly unlikely to be due to chance.
- But is it also clinically relevant?

- Warning. Very often **statistical significance** is confused with **clinical relevance** of the result obtained.
- The fact that the result obtained is statistically significant only implies that it is very likely that this result is true, and not due to chance.
- **However, the *p*-value cannot tell us whether this result is also clinically important, because this is a judgment that rests only with those who are evaluating the results, and is unrelated to statistics.**

HYPOTHESIS TESING

- If, for example, in a sample of 20,000 patients with diabetes we compare the efficacy of two hypoglycemic drugs and obtain average HbA_{1c} values of 7.1±1.3 with drug A and average values of 7.3±1.4 with drug B, this difference in average values will be highly significant.
- Is this data sufficient to state that drug A should be preferred? Probably not.
- In fact, from the epidemiological data, it is difficult to imagine that such a small difference in mean HbA_{1c} values would translate into an important difference in the risk of developing the complications of the disease.

HYPOTHESIS TESING

- Certainly, if drugs A and B had the same tolerability profile and cost, we should always prefer the one that has been shown to be more effective, even if only slightly.
- If, on the contrary, drug B was better tolerated or less expensive, then it might in specific circumstances be preferred, despite slightly (though statistically significant) lower efficacy.

P-VALUES AND CONFIDENCE INTERVALS

p<0.05

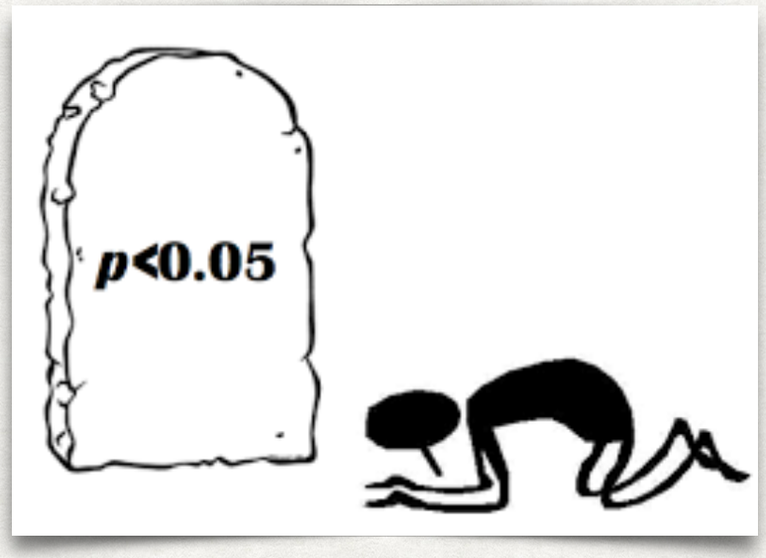
- If we reject the null hypothesis (i.e., if we state that one treatment is more effective than the other) we have a less than 5% chance of stating the falsehood.

Limits

- Arbitrary threshold value (e.g., why not 0.03 or 0.06?).
- The probability of finding a $p<0.05$ increases with the number of tests performed.
- The *p* value does not provide any indication of the clinical significance of the difference found.

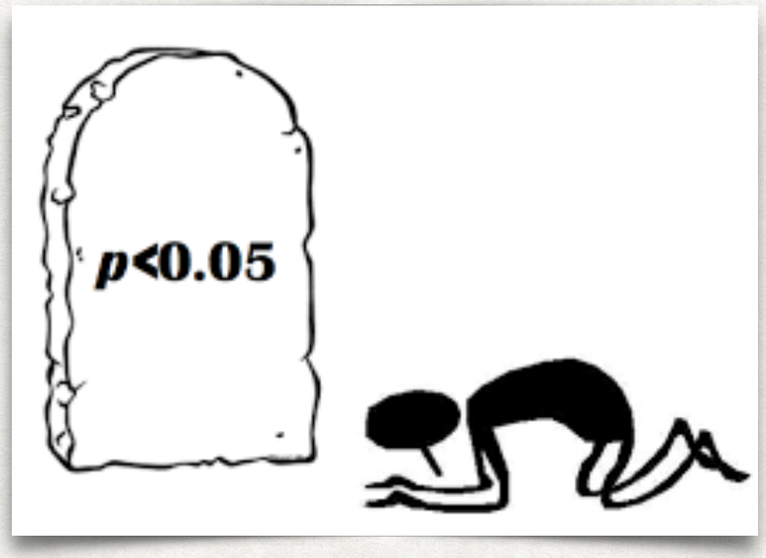
P-VALUES AND CONFIDENCE INTERVALS

- From the above, it is clear that the value of *p* has limitations.
- First of all, the threshold value to define a result as statistically significant is arbitrarily placed at 0.05.
- But can we say with confidence that a *p*=0.06 or 0.07 is not?
- Indeed, the risk of incorrectly accepting the alternative hypothesis would increase from 5% to 6-7%, but it would still be very low!



P-VALUES AND CONFIDENCE INTERVALS

- Moreover, when we apply a statistical test, the *p*-value obtained will refer to the probability of error for that specific test.
- If within a study — as always happens — we do many statistical analyses on our data, the probability of finding a *p*-value<0.05 by pure chance increases substantially as the number of tests performed increases.
- Finally, we have already pointed out that the *p*-value does not provide any indication regarding the clinical relevance of the results obtained.



CHOICE OF STATISTICAL TEST

CHOICE OF STATISTICAL TEST

To decide on the most appropriate type of test for a given question, one must ask the questions:

1. **What kind of variables should I study?**
   - Numerical (continuous or discrete)
   - Categorical (ordinal or nominal)
2. **If the variable is numeric, is it normally distributed?**
3. **How many groups do I need to compare?**
   - 2 groups;
   - 3 or more groups.

CHOICE OF STATISTICAL TEST

Parametric methods

- Based on mean μ and standard deviation σ (or variance σ2).
- They are used in the case of normally distributed variables.

Non-parametric methods

- Based on frequency and percentiles.
- They are used in the case of variables whose distribution is not normal.

The former are called "parametric" because they are based on the *mean* and *standard deviation* parameters. These tests can therefore be used only for normally distributed variables. In all other cases, non-parametric methods are used, which are independent of the distribution of the data.

PARAMETRIC STATISTICAL TESTS

- **Parametric** statistical methods are statistical tests based on the parameters mean and standard deviation.
- For this reason they are only usable for normally distributed numerical variables.

How do you evaluate the normality of a distribution?

Graphical approach

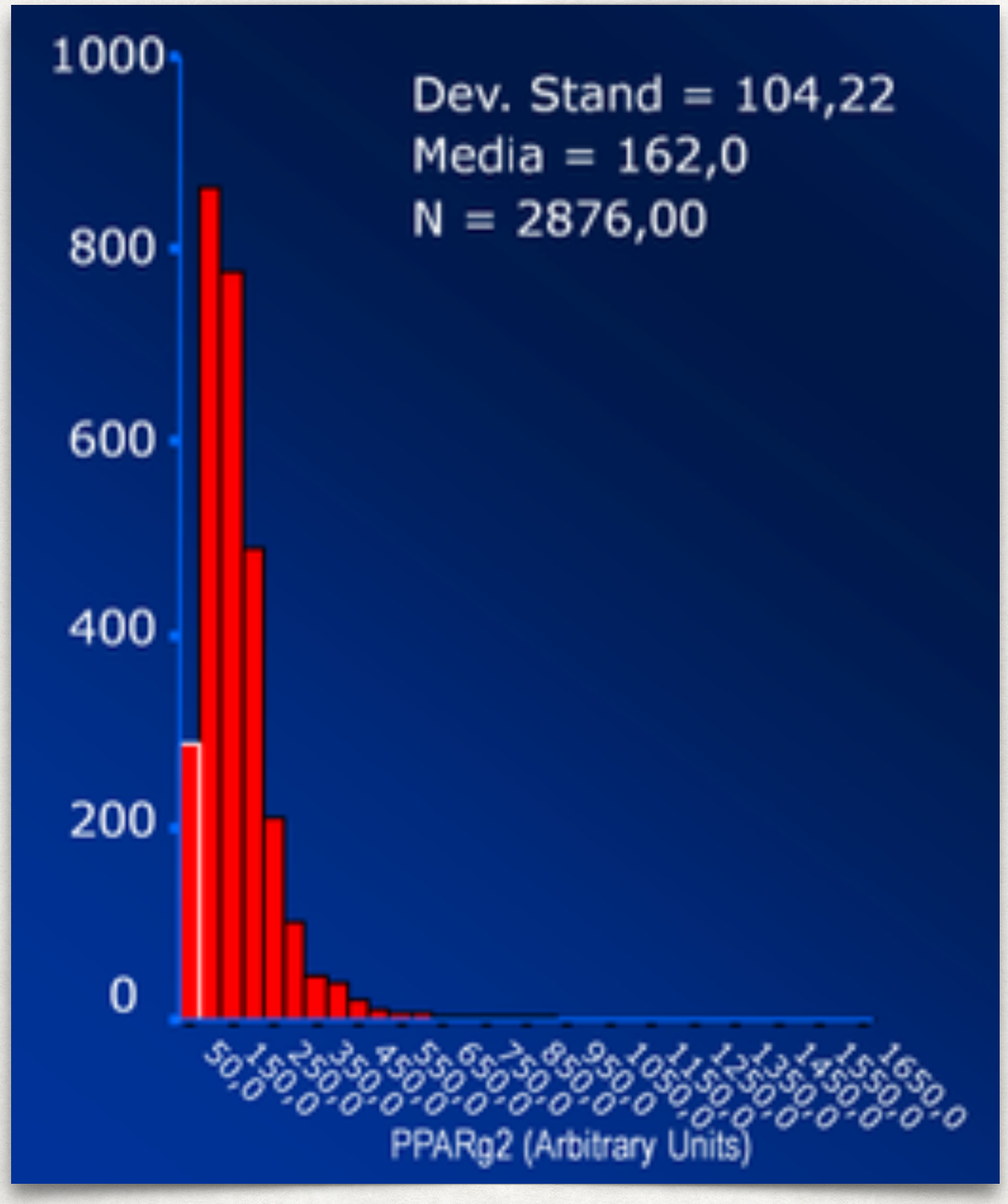
Skewness and Kurtosis

Statistical tests (Kolmogorov-Smirnov, χ^2, Shapiro-Wilk, …)

It is possible to use different approaches, more or less "exact", based on a graphical evaluation, which however could be imperfect, or on numerical parameters.

PARAMETRIC STATISTICAL TESTS

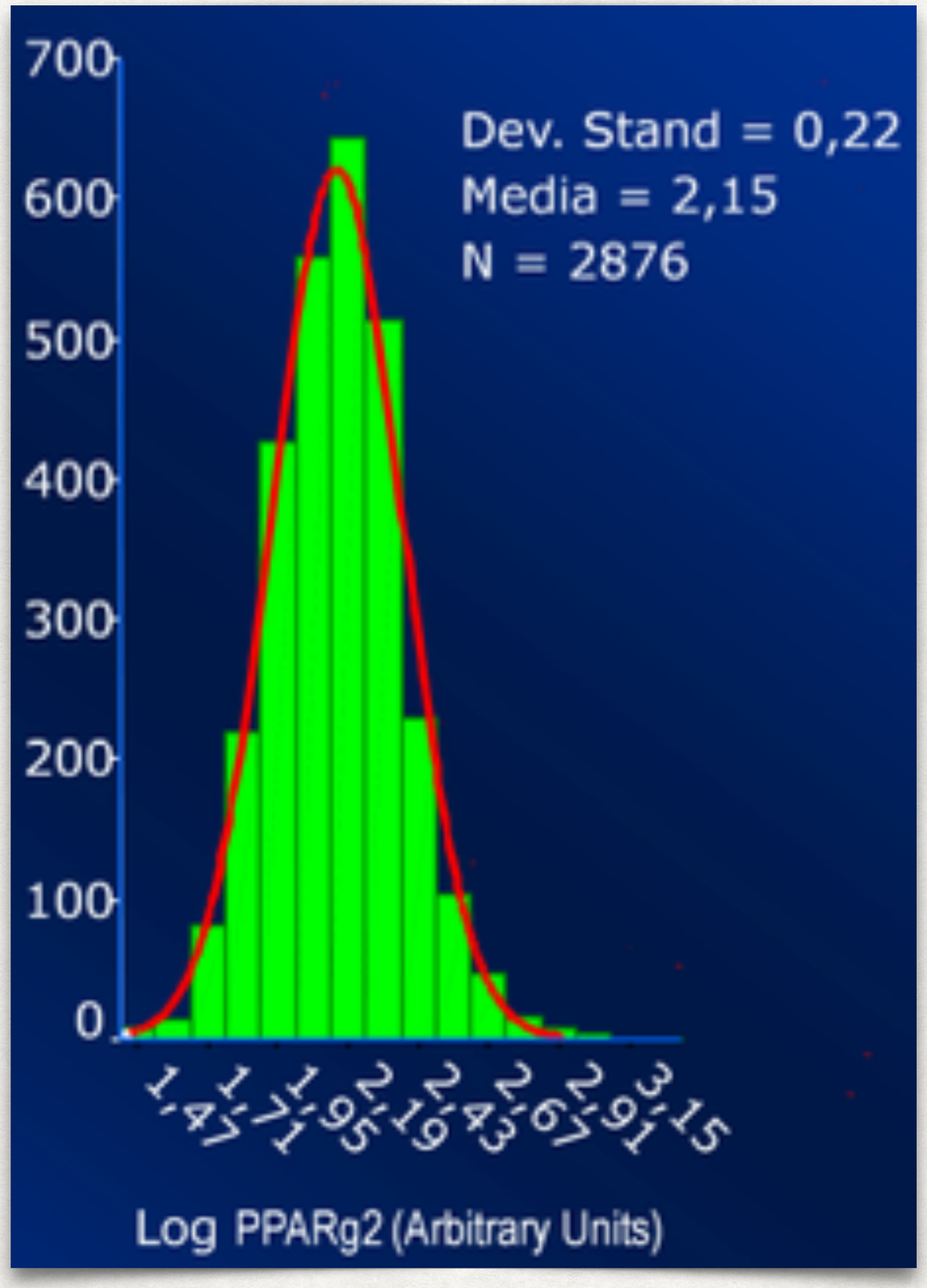
Graphical Approach



- The graphical approach, as already mentioned, consists in drawing the histogram of the variable in question, to verify whether or not it presents a "bell-shaped", symmetrical distribution.
- The example shows how the distribution of PPARg2 gene expression levels is clearly asymmetric (left graph).

PARAMETRIC STATISTICAL TESTS

Graphical Approach

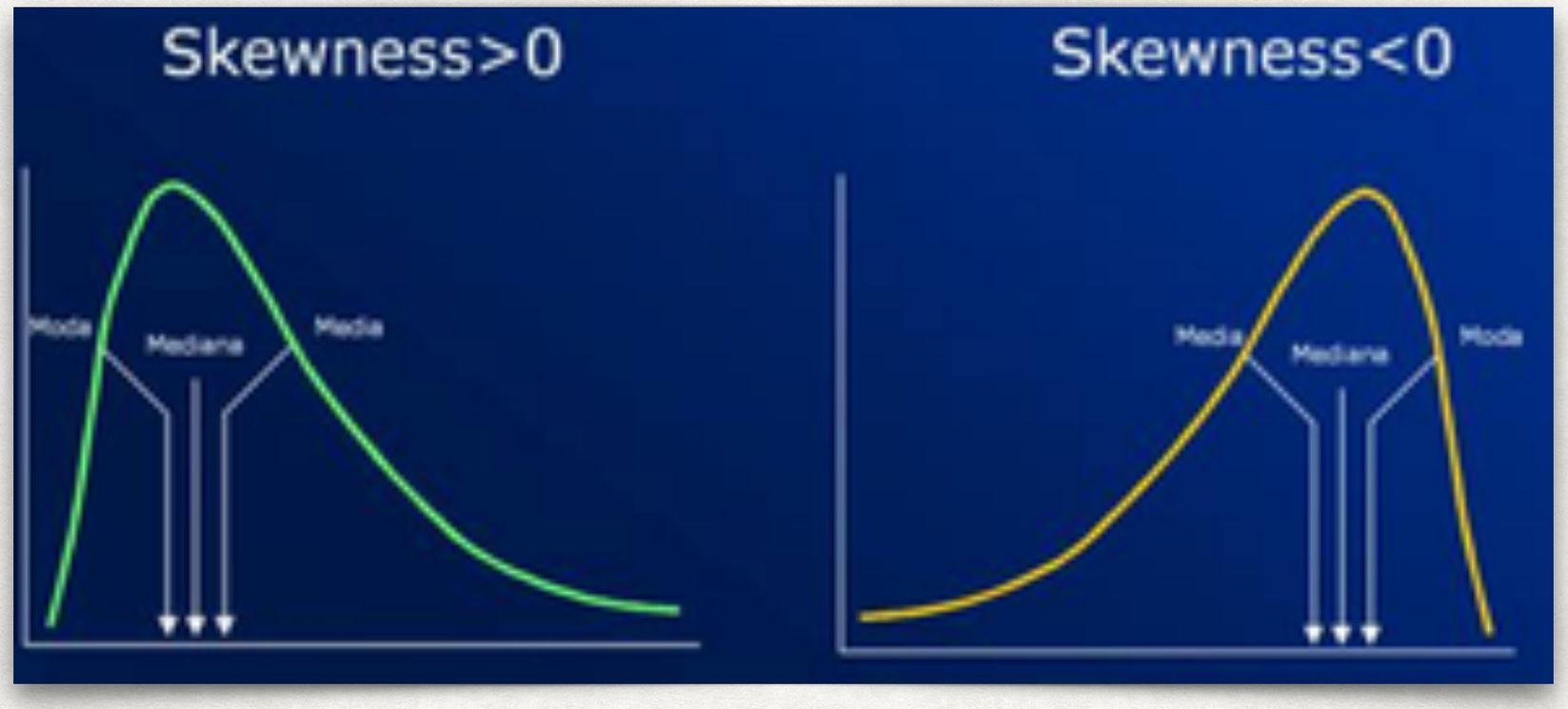


- However, when we draw the graph of the logarithm of PPARg2 mRNA levels, the distribution of the new variable is certainly closer to normal.
- However, the graphical impression is not sufficient.
- In fact, recall that, in addition to the bell-shaped appearance, for a variable to be normally distributed it is necessary that 95% of the observations fall within the interval defined by mean ± (approximately) 2 standard deviations (in the example 2.15±0.44).

PARAMETRIC STATISTICAL TESTS

Skewness and Kurtosis

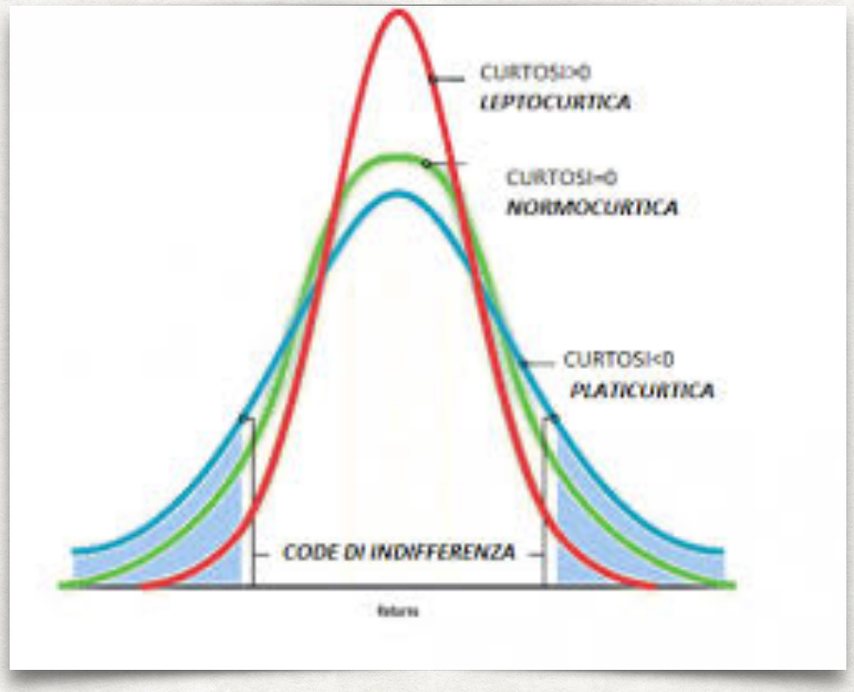
- Two numerical indices usually reported by all statistical software efficiently summarize the information concerning the distribution of a continuous variable: *skewness* and *kurtosis*.
- The **skewness** indicates the level of asymmetry of the distribution. In the case of perfect symmetry, the value will be zero. If the value is positive, then the distribution will be skewed to the right, while in the case of a negative value it will be skewed to the left.



PARAMETRIC STATISTICAL TESTS

Skewness and Kurtosis

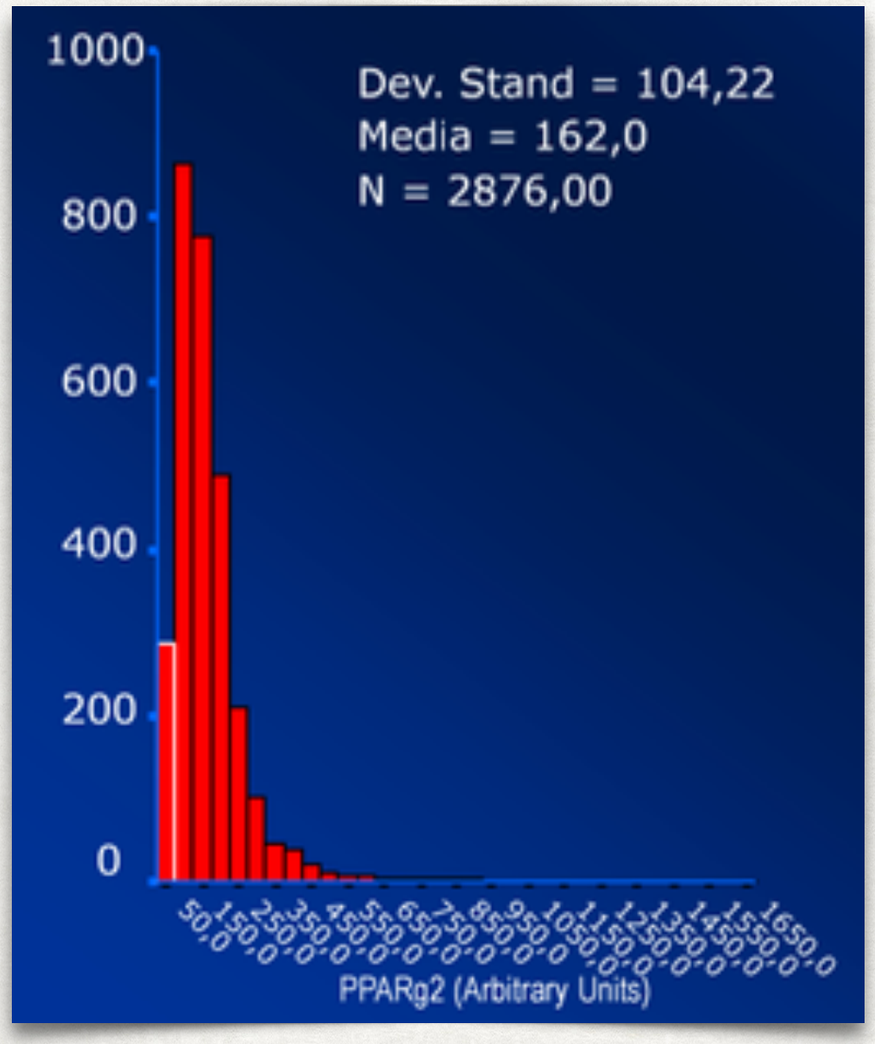
- In contrast, **kurtosis** indicates whether the bell-shaped distribution is very "narrow" or skewed.
- If the distribution is normal, the kurtosis will be zero (95% of observations contained within $\mu \pm 1.96\sigma$).
- If the kurtosis is negative, then the distribution is platycurtic, i.e., slack, which is equivalent to saying that less than 95% of the observations fall within $\mu \pm 1.96\sigma$
- Conversely, if the kurtosis is positive, then the distribution is leptokurtic, i.e., "narrow," meaning that more than 95% of observations fall within $\mu \pm 1.96\sigma$



PARAMETRIC STATISTICAL TESTS

Skewness and Kurtosis

						Asymmetry			Kurtosis		
	N	Min	Max	μ	σ	Value	SE	Value/SE	Value	SE	Value/SE
PPARg2	2876	28	1631	162.03	104.219	4.11	0.046	89.3	33.842	0.091	371.9



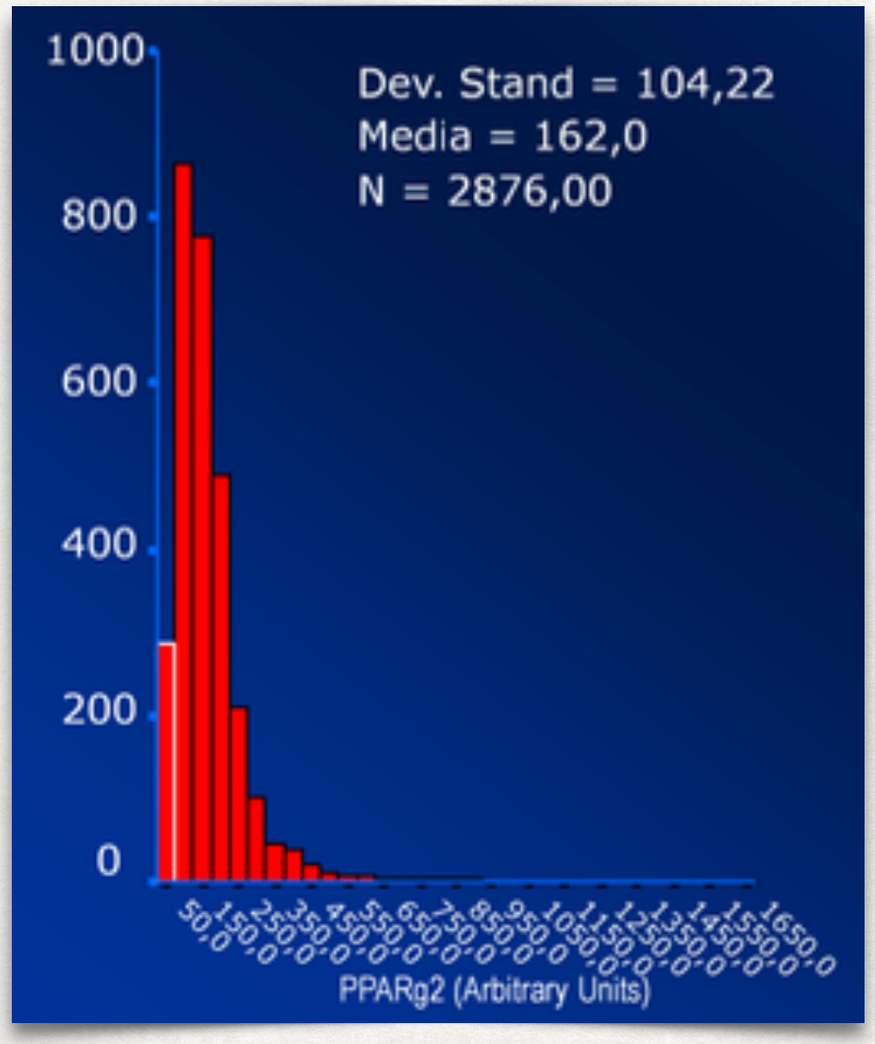
- If the measure of skewness or kurtosis divided by its standard error SE(*) gives a value greater than 2 or less than -2 then the distribution can not be considered normal.
- In the example (gene expression level of PPARg2) we can see that the value of skewness is positive, equal to 4.11: this is equivalent to say that the distribution is skewed to the right, as clearly evident from the histogram.
- This value, divided by its standard error (equal to 0.046), gives a result of 89.3, well greater than 2!

(*) The SE for skewness is approximately given by $(6/N)^{1/2}$

PARAMETRIC STATISTICAL TESTS

Skewness and Kurtosis

						Asymmetry			Kurtosis		
	N	Min	Max	μ	σ	Value	SE	Value/SE	Value	SE	Value/SE
PPARg2	2876	28	1631	162.03	104.219	4.11	0.046	89.3	33.842	0.091	371.9

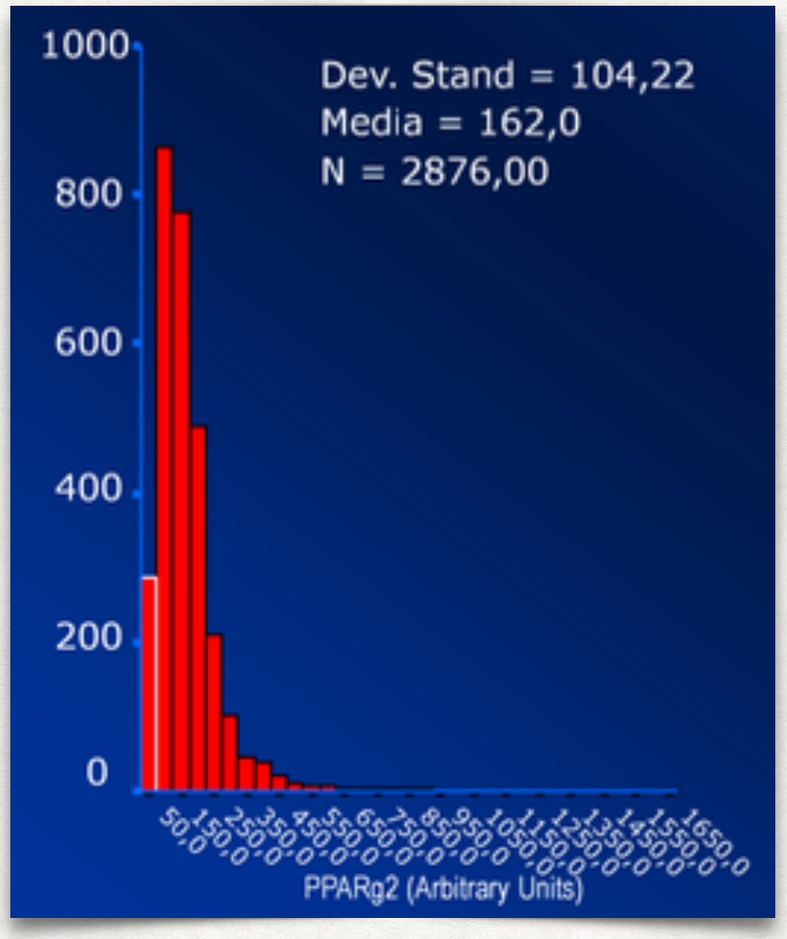


- Similarly, the kurtosis is positive and very high (33.842), indicating that the distribution is very "narrow". The value obtained by dividing the kurtosis by its SE(*) is 371.9
- Therefore, the distribution in question deviates significantly from a normal distribution.

(*) The SE for kurtosis is approximately given by $(24/N)^{1/2}$

PARAMETRIC STATISTICAL TESTS

KOLMOGOROV-SMIRNOV TEST



- A final way to assess whether a distribution differs significantly from normal is by applying a statistical test, called the Kolmogorov-Smirnov test.
- This test starts with the null hypothesis that the distribution under consideration does not differ from a normal distribution.

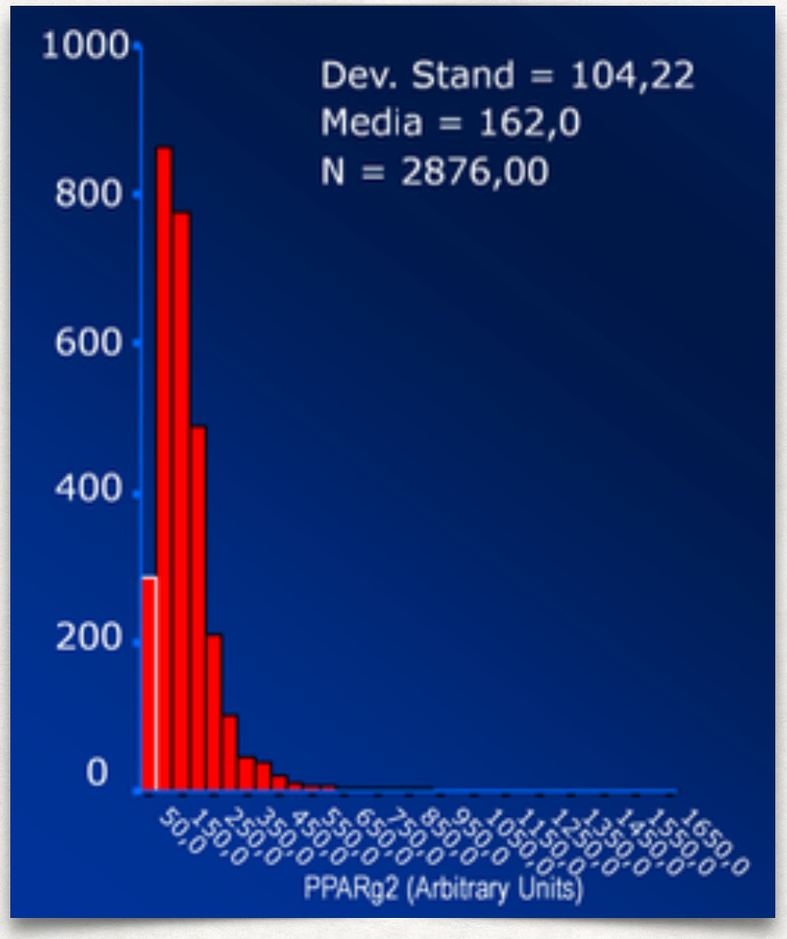
K-S test for one sample

PPARg2	
N	2876
μ	162.03
σ	104.22
Z of K-S	7.716
p-value (2 tails)	<0.0001

If $p < 0.05$, the distribution differs significantly from a normal distribution.

PARAMETRIC STATISTICAL TESTS

KOLMOGOROV-SMIRNOV TEST



K-S test for one sample

PPARg2	
N	2876
μ	162.03
σ	104.22
Z of K-S	7.716
p-value (2 tails)	<0.0001

If $p < 0.05$, the distribution differs significantly from a normal distribution.

- If the test results in a value of $p<0.05$, then we should reject the null hypothesis, and state that the distribution under consideration differs significantly from a normal distribution.
- Applying the K-S test to our triglyceride data, we will obtain a value of $p<0.0001$.
- Therefore, we will have to conclude that the distribution differs significantly from a normal distribution.

SOME USEFUL STATISTICAL INDICATORS

Standard Error (SE), Relative SE (RSE), Coefficient of Variation (CV)

N	μ	σ	**SE**	RSE	CV (RSD)
2876	162.03	104.219	**±1.943**	1.2%	89.3

- In addition to the normal statistical parameters Mean and Standard Deviation, we also use the **Standard Error** of the Mean (SE) which measures the "dispersion" of the sample distribution of the mean.
- The sample mean is an estimate of the population mean. However, different samples from the same population generally have different values of the sample mean, so there is a distribution of sample means (with its own mean and variance).
- The standard error of the mean is the standard deviation of the sample averages over all possible samples (of a given size) drawn from the population.
- It can be calculated as the ratio of the standard deviation to the square root of the sample size:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{N}}$$

In the example: SE = σ/√N = 1.943;

μ±1.96SE corresponds to the 95% confidence interval of the mean = [158.22, 165.84]

SOME USEFUL STATISTICAL INDICATORS

STANDARD ERROR (SE), RELATIVE SE (RSE), COEFFICIENT OF VARIATION (CV)

N	μ	σ	SE	**RSE**	CV (RSD)
2876	162.03	104.219	±1.943	**1.2%**	89.3

- The **relative standard error** of a sample mean (RSE) is the standard error divided by the mean and expressed as a percentage.
- It can only be calculated if the mean is a non-zero value.
- A variable with a lower relative standard error can be said to have a more accurate measure, since it has proportionally less sample variation around the mean.

$$\mathrm{RSE} = \frac{\mathrm{SE}}{\mu}$$

In the example: RSE = SE/μ = 1.943/162.03 = 0.012 = 1.2%

SOME USEFUL STATISTICAL INDICATORS

Standard Error (SE), Relative SE (RSE), Coefficient of Variation (CV)

N	μ	σ	SE	RSE	**CV (RSD)**
2876	162.03	104.219	±1.943	1.2%	**64.32%**

- **Coefficient of Variation** (CV), also known as Relative Standard Deviation (RSD), is a standardized measure of dispersion of a probability distribution or frequency distribution.
- The CV or RSD is widely used in laboratory analysis to express the precision and repeatability of a test.
- It shows the degree of variability from the population mean.
- It is often expressed as a percentage, and is defined as the ratio of the standard deviation to the mean (or its absolute value).

$$\mathrm{CV} = \mathrm{RSD} = \frac{\sigma}{\mu}$$

In the example: CV = σ/μ = 104.219/162.03 = 0.6432 = 64.32%

SOME USEFUL STATISTICAL INDICATORS

QUARTILE COEFFICIENT OF DISPERSION (QCD)

Dataset A = {2, 4, 6, 8, 10, 12, 14}						
N	range	μ	median	Q1	Q3	**QCD**
7	12	8	8	4	12	**50%**

Dataset B = {1.8, 2, 2.1, 2.4, 2.6, 2.9, 3}						
N	range	μ	median	Q1	Q3	**QCD**
7	1.2	2.4	2.4	2	2.9	**18%**

- A more robust possibility than the coefficient of variation is the **Quartile Coefficient of Dispersion** (QCD).
- The Quartile Coefficient of Dispersion is a descriptive statistic that measures dispersion and is used for within- and between-distribution comparisons.
- It is given by the interquartile range divided by the sum: (Q3-Q1) / (Q3+Q1).

In the example:

$$\text{QCD dataset A} = (Q3-Q1)/(Q3+Q1) = (12-4)/(12+4) = 8/16 = 0.5 = 50\%$$

$$\text{QCD dataset B} = (Q3-Q1)/(Q3+Q1) = (2.9-2)/(2.9+2) = 0.9/4.9 = 0.18 = 18\%$$

The quartile coefficient of dispersion of dataset A is 2.7 times larger (0.5/0.18) than that of dataset B

NON-PARAMETRIC STATISTICAL TESTS

- Nonparametric statistical methods are statistical tests based on the ranks of observations, i.e., their order number instead of the observations themselves.
- Thus, by transforming values into ranks, one loses information about the measure, retaining only information about the ordering.
- Whenever a numerical variable is not normally distributed, or we are faced with a categorical (ordinal or nominal) variable, it will be necessary to use **nonparametric tests**.
- Such tests are based on the ranks of the observations, not on their true value. In other words, the observations are put in ascending order, and each one is given a number corresponding to the position that observation occupies in the ranking (*rank*).
- Nonparametric statistical tests are then based on a comparison of the sums of the ranks.

NON-PARAMETRIC STATISTICAL TESTS

Example of Rank Calculation

Let the following values of a variable be given: 11, 25, 3, 26, 20

- To calculate the ranks, first order the values of the variable: 3, 11, 20, 25, 26
- Then we assign the ranks, which will take the following values: 1, 2, 3, 4, 5

Let the following values of a variable be given: 11, 25, 3, 25, 20

- To calculate the ranks, first order the values of the variable: 3, 11, 20, 25, 25
- Then we assign the ranks, which will take the following values: 1, 2, 3, **4.5, 4.5**

For identical data, the average of the ranks that would otherwise be attributable to the measures (had they been different) is considered.

Thus in the example, the value **4.5** is derived from the average of the ranks that would have been assigned to the two values (4 and 5).

NON-PARAMETRIC STATISTICAL TESTS

APPLICABILITY

Nonparametric statistical tests are warranted when:

- variables have **clear deviations** from normality, or are **highly skewed** or have **more than one peak** (multimodal distribution);
- the sample is **too small** to test whether the data distribution is normal;
- observations are represented by **categorical variables**.

Non-parametric tests are therefore used when a variable is not normally distributed, but also when there are few values available, and therefore it is not possible to understand what the distribution is.

If, for example, we measure the weight in 5 subjects, it will be very difficult to understand if these measures are normally distributed!

As a general rule, if the number of observations is <20, you should never use a parametric test.

CHOICE OF STATISTICAL TEST

Non-parametric methods are all the more "powerful" (in the statistical sense) the further away the distribution is from normal, while they are less powerful than the corresponding parametric methods when approaching normal.

- While there is no doubt about the necessity of having to use a nonparametric test for nominal and ordinal variables, one might wonder what error one would incur if, in the presence of a continuous variable, one always used a parametric test, regardless of the distribution of the variable or, conversely, always used a nonparametric test.
- In this regard, it is important to emphasize that nonparametric tests are more powerful the more the distribution deviates from normality, while, in the presence of a normal distribution, they tend to be less powerful than parametric tests.
- As a consequence, if we use a parametric test when the distribution is clearly non-normal, or conversely use a non-parametric test in the presence of a normal distribution, we may obtain results that are not statistically significant, when they would be if the more appropriate test had been used.

CHOICE OF STATISTICAL TEST

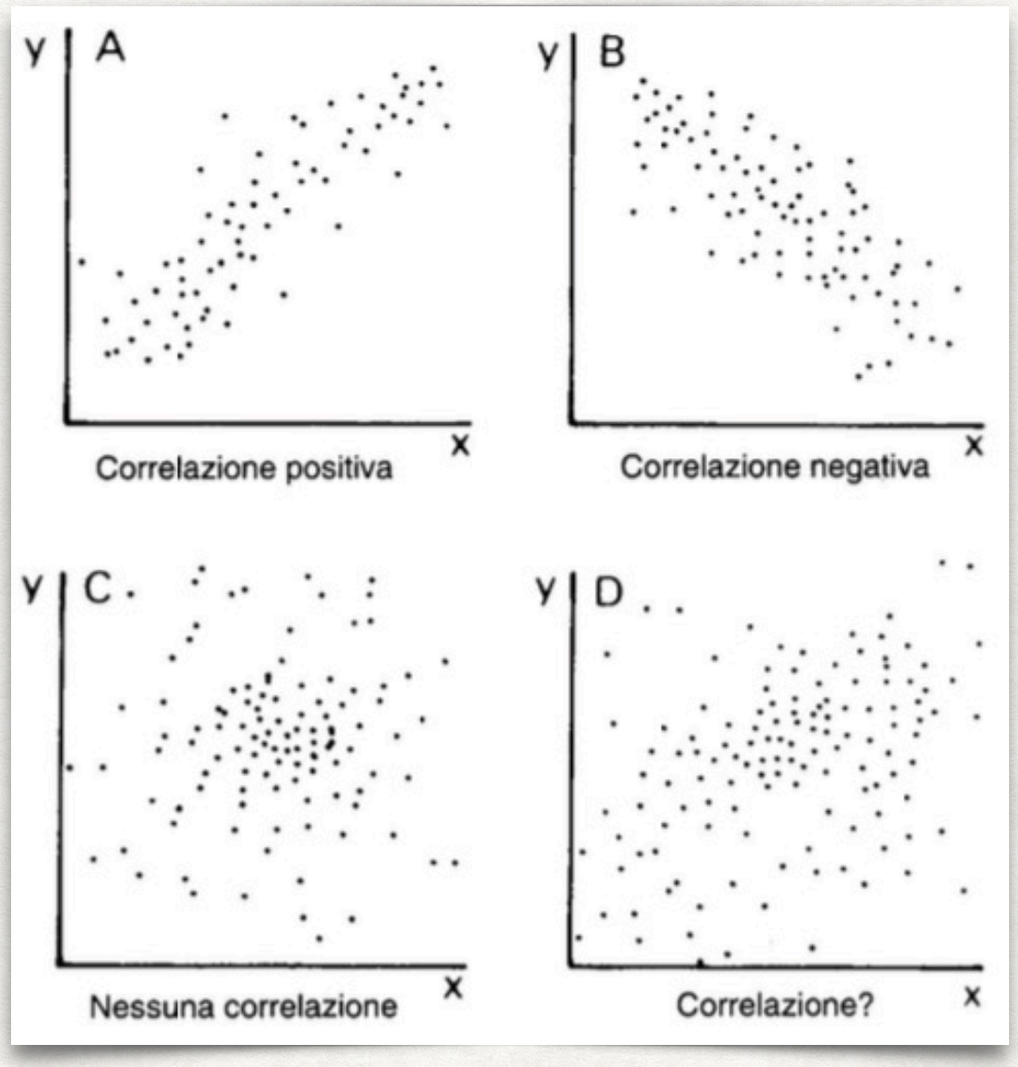
	Parametric tests	Non-parametric tests
Summary measures	Mean (μ) and standard deviation (σ)	Median and range
Comparison of 2 variables	Linear regression	Chi-square (χ^2)
Correlation measure between 2 variables	Pearson's R	Spearman's ρ Kendall's τ-b
Comparison of values between 2 different groups	Student's "unpaired" t-test	Mann-Whitney u-test
Before-after comparison	Student's "paired" t-test	Wilcoxon sign-rank test
Comparison of values among multiple groups	Analysis of variance (ANOVA)	Kruskal-Wallis test
Multivariate analysis	Multiple regression	Logistic regression

This table summarizes the most commonly used statistical tests.
It is clear that, for each situation, there is a parametric test and the corresponding non-parametric test.
There is therefore no reason to use a parametric test at all costs, even when it is not appropriate.

CORRELATION AND REGRESSION

CORRELATION AND REGRESSION

- **Correlation** looks at <u>whether there is a relationship</u> between two variables (how and how much two variables vary together).
- **Regression** analyzes the <u>form of the functional relationship</u> between variables (cause-effect relationship).
- **Simple regression** (linear or nonlinear): determine the form of the relationship between 2 variables (one independent and one dependent).



- **Multiple regression**: determine the form of the relationship between multiple variables (multiple independents and one dependent).
- Conoscendo la forma della relazione funzionale tra variabile indipendente e dipendente è possibile stimare il valore della variabile dipendente conoscendo quello della variabile indipendente (solo interpolazione).

CORRELATION BETWEEN NORMAL VARIABLES

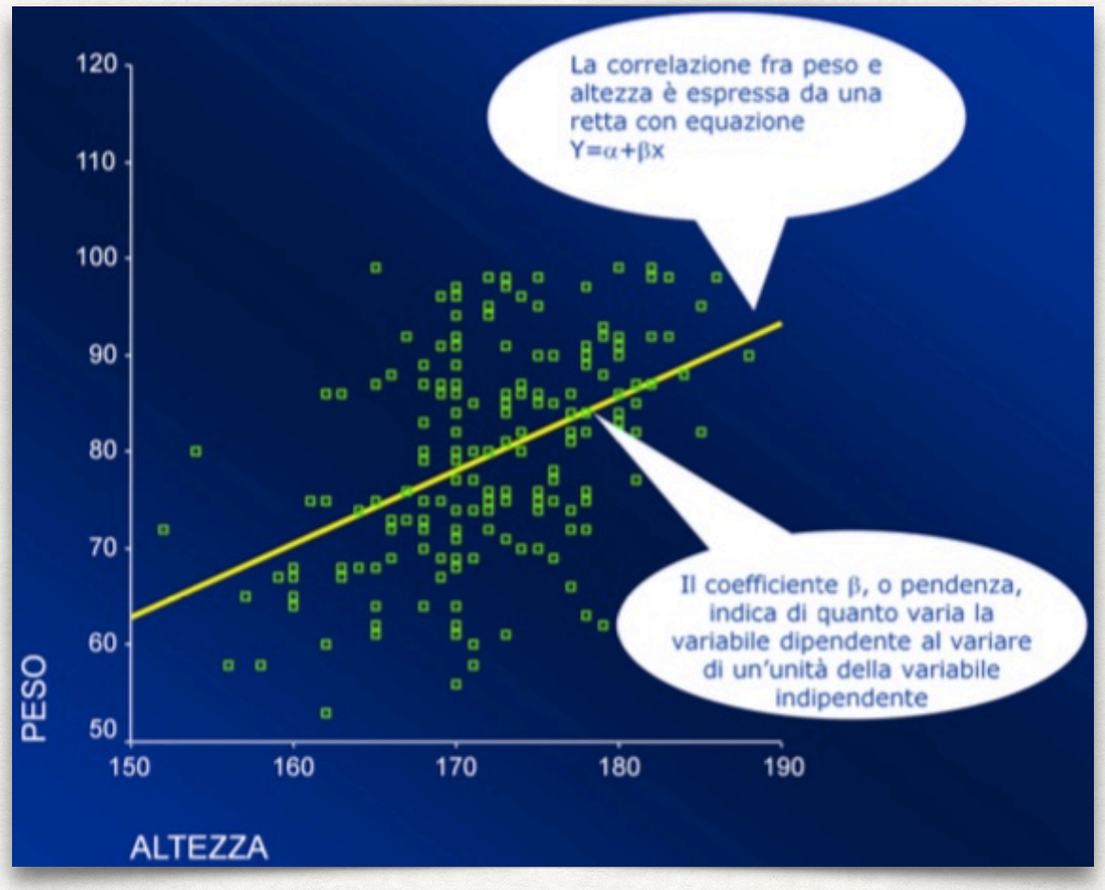
What is the relationship between weight and height?
Is systolic blood pressure related to age?

To these questions (search for a correlation between two normally distributed variables) it is possible to give an answer with **linear regression**.

With this technique it is possible to evaluate if and to what extent the values of a variable (**dependent variable**) vary (i.e. are correlated) with the variation of a second variable (independent variable).

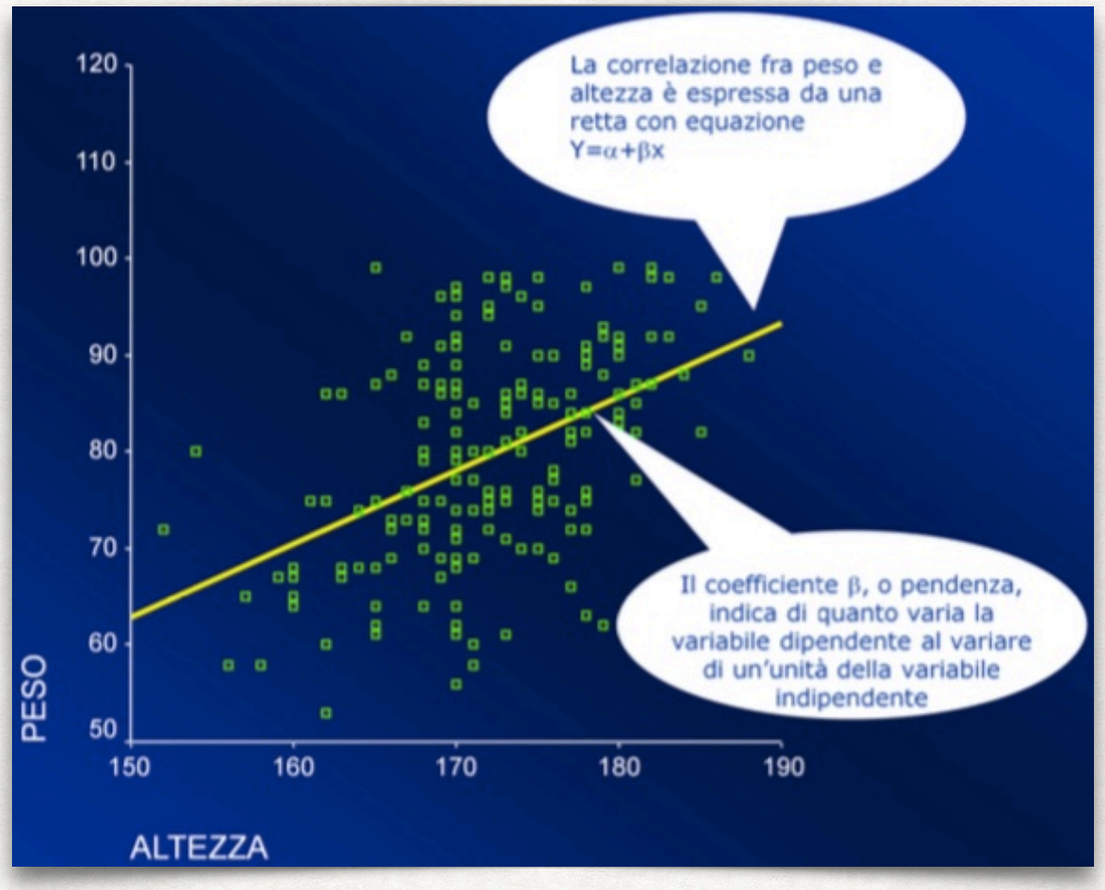
For example, in the relationship between weight and height the weight will be the dependent variable (i.e. the one that varies as height varies), while in the relationship between systolic pressure and age, the first will vary as a function of the second, so our dependent variable will be the systolic pressure, the independent one the age.

CORRELATION BETWEEN NORMAL VARIABLES



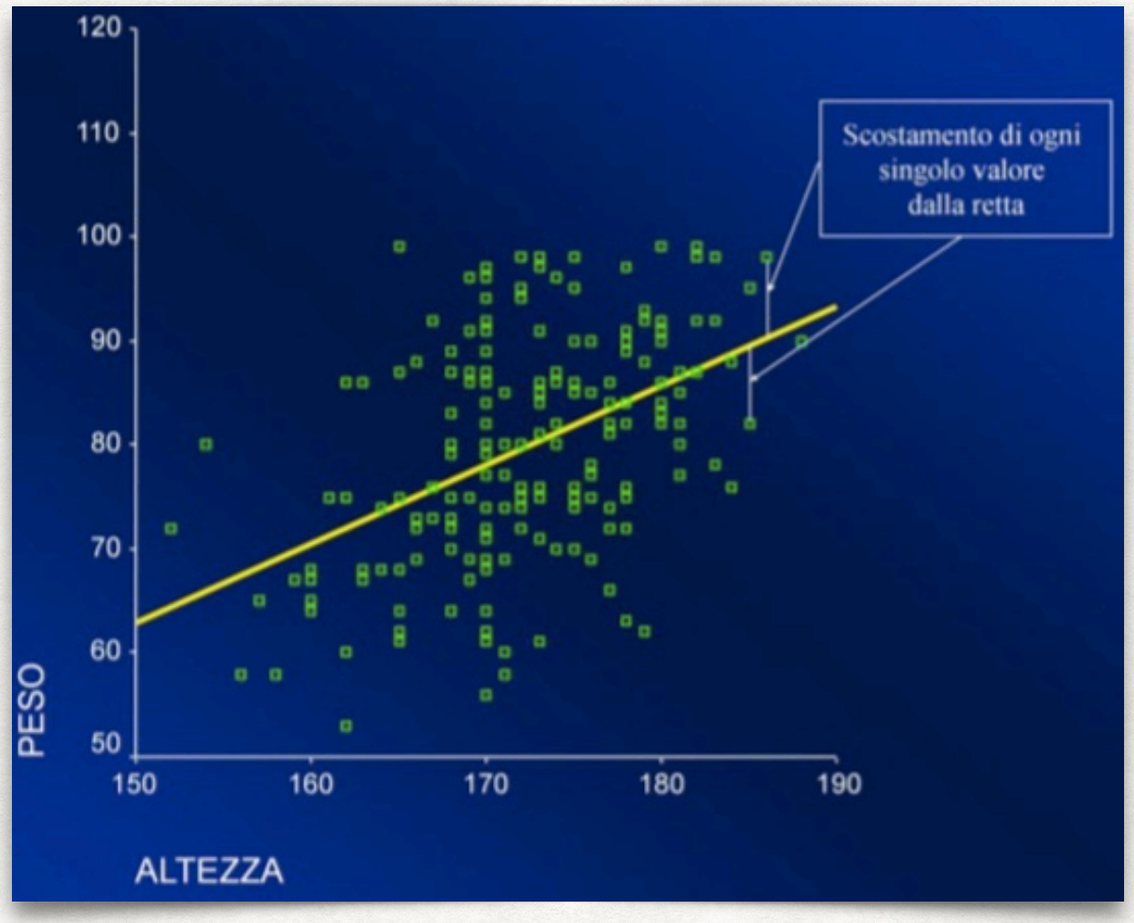
- In linear regression, the relationship between the dependent and independent variables can be represented in a graph with Cartesian axes.
- On the abscissas are reported the values of the independent variable, and on the ordinates the values of the dependent variable.
- The closer the relationship between the two variables, the more the points on the graph will tend to lie along a straight line.

CORRELATION BETWEEN NORMAL VARIABLES



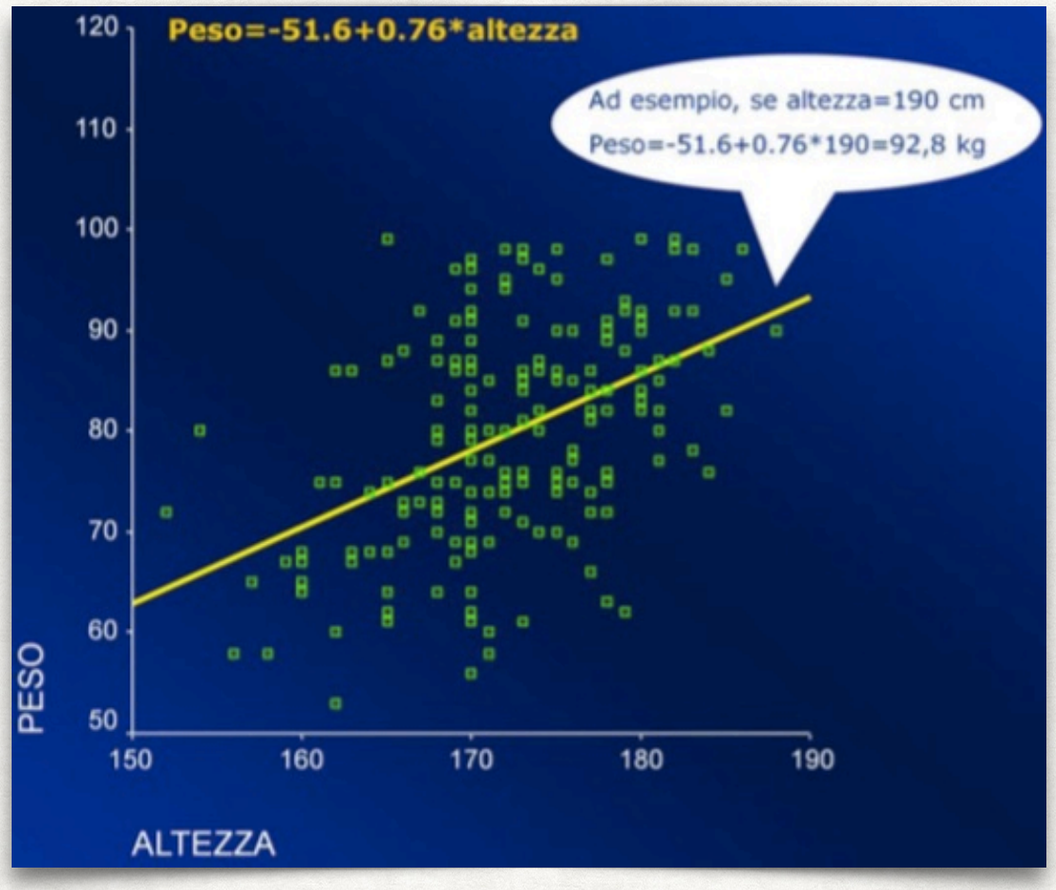
- From a mathematical point of view, the relationship between the two variables is expressed precisely by the equation of a straight line, where y represents the value of the dependent variable and x represents the value of the independent variable.
- The β coefficient, also called slope, indicates how much y varies for each variation of one unit of x.
- The value of α, or intercept, represents the value that y assumes when x is equal to zero.

CORRELATION BETWEEN NORMAL VARIABLES



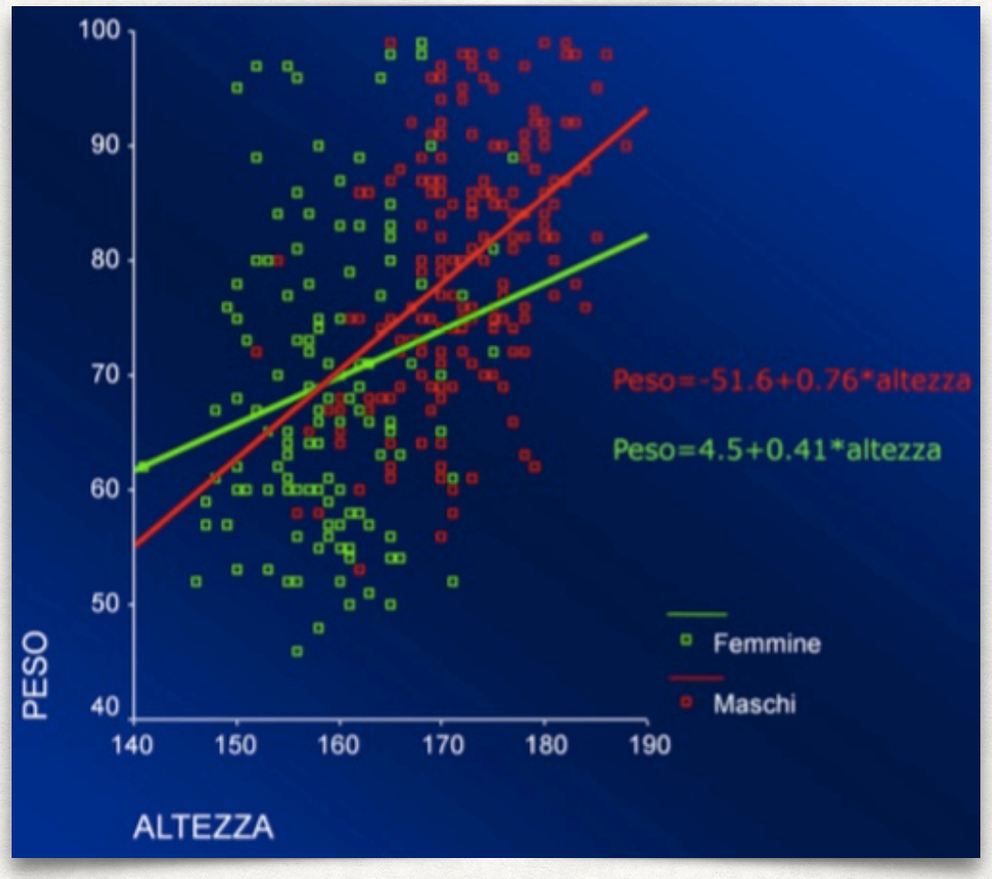
- Note that of all the possible lines describing the relationship between the two variables, the one for which the sum of the squared value of all the deviations of each point from the line is the least will be chosen (method of least squares).
- *Notice. The values of the deviations are squared because, compared to the value estimated by the equation, the values would be sometimes positive and sometimes negative, so their sum would be equal to zero.*

CORRELATION BETWEEN NORMAL VARIABLES



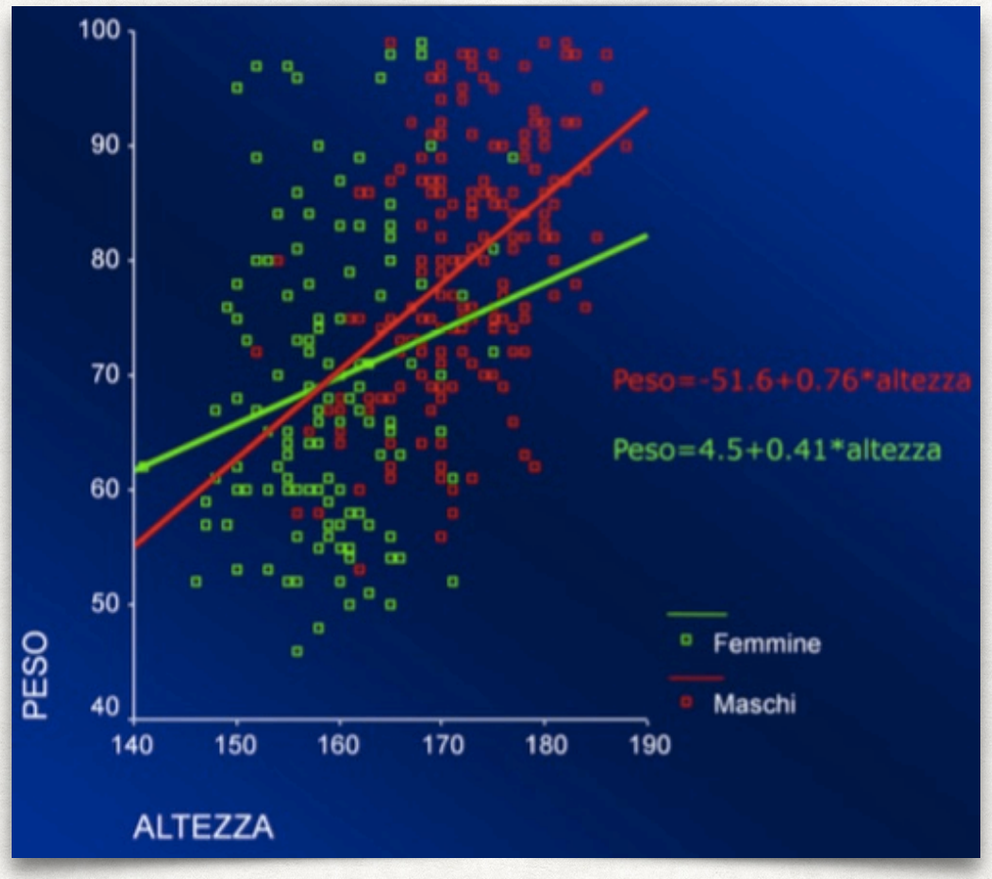
- In the case of the relationship between weight and height shown in the figure, the equation tells us that, on average, the weight tends to increase by 0.76 kg for each additional cm of height.
- Note that the value of the intercept, negative, is purely theoretical, as it would represent the weight of a subject 0 cm tall!
- From the equation of the line we can estimate that, for example, a subject 1.90 cm tall and belonging to the population under study (male subjects with diabetes), would have an average weight of 92.8 kg.

CORRELATION BETWEEN NORMAL VARIABLES



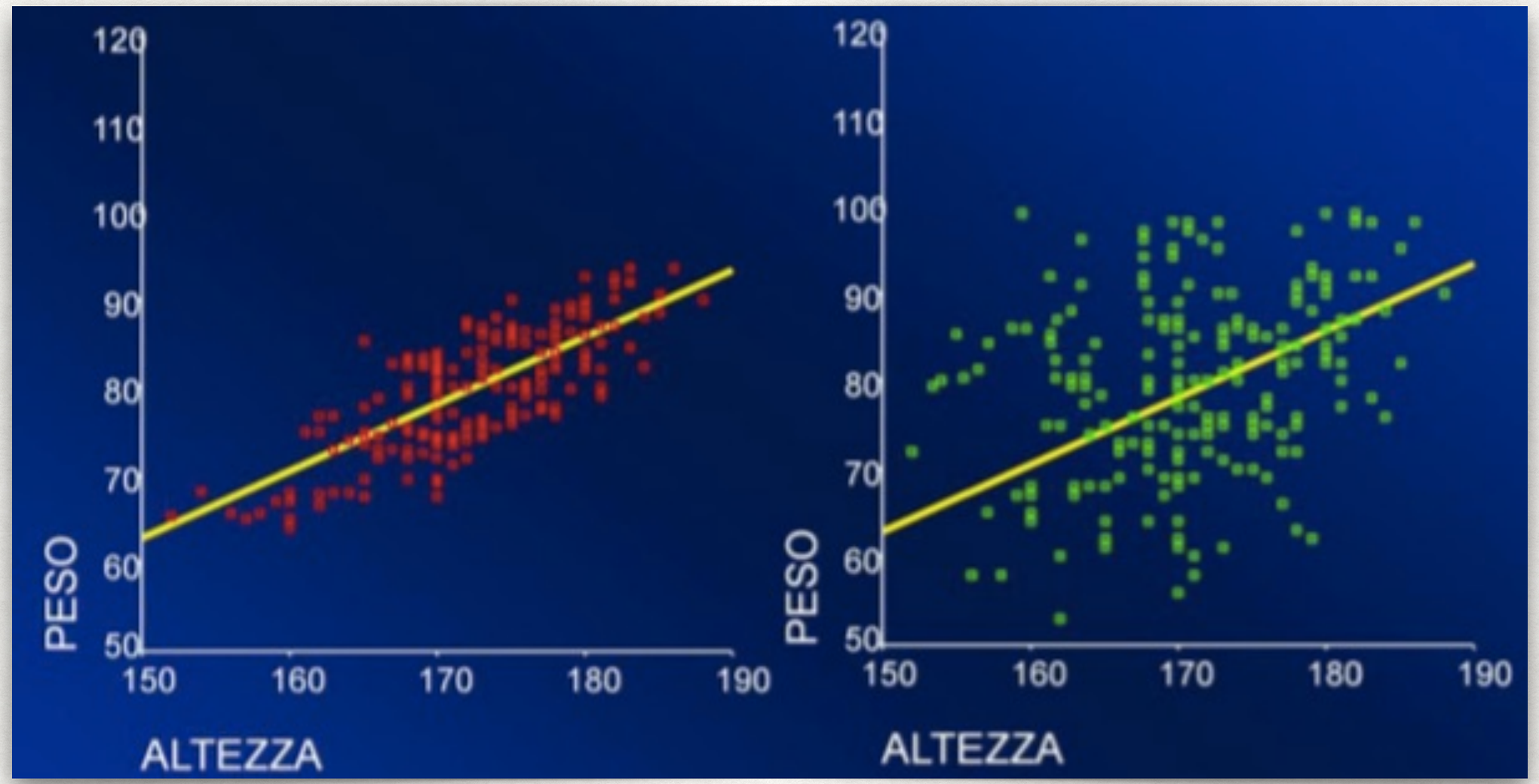
- In this graph, the relationship between weight and height has been estimated separately for males and females.
- As can be seen, the line describing the ratio between weight and height in males (red) is "steeper" than that of females (green).
- The slope (or beta coefficient) for males is in fact higher (0.76 for males, 0.41 for females).
- In other words, weight will tend to increase by an average of 0.76 kg for every cm more in height among males, while it will tend to increase by an average of 0.41 kg for every cm among females.

CORRELATION BETWEEN NORMAL VARIABLES



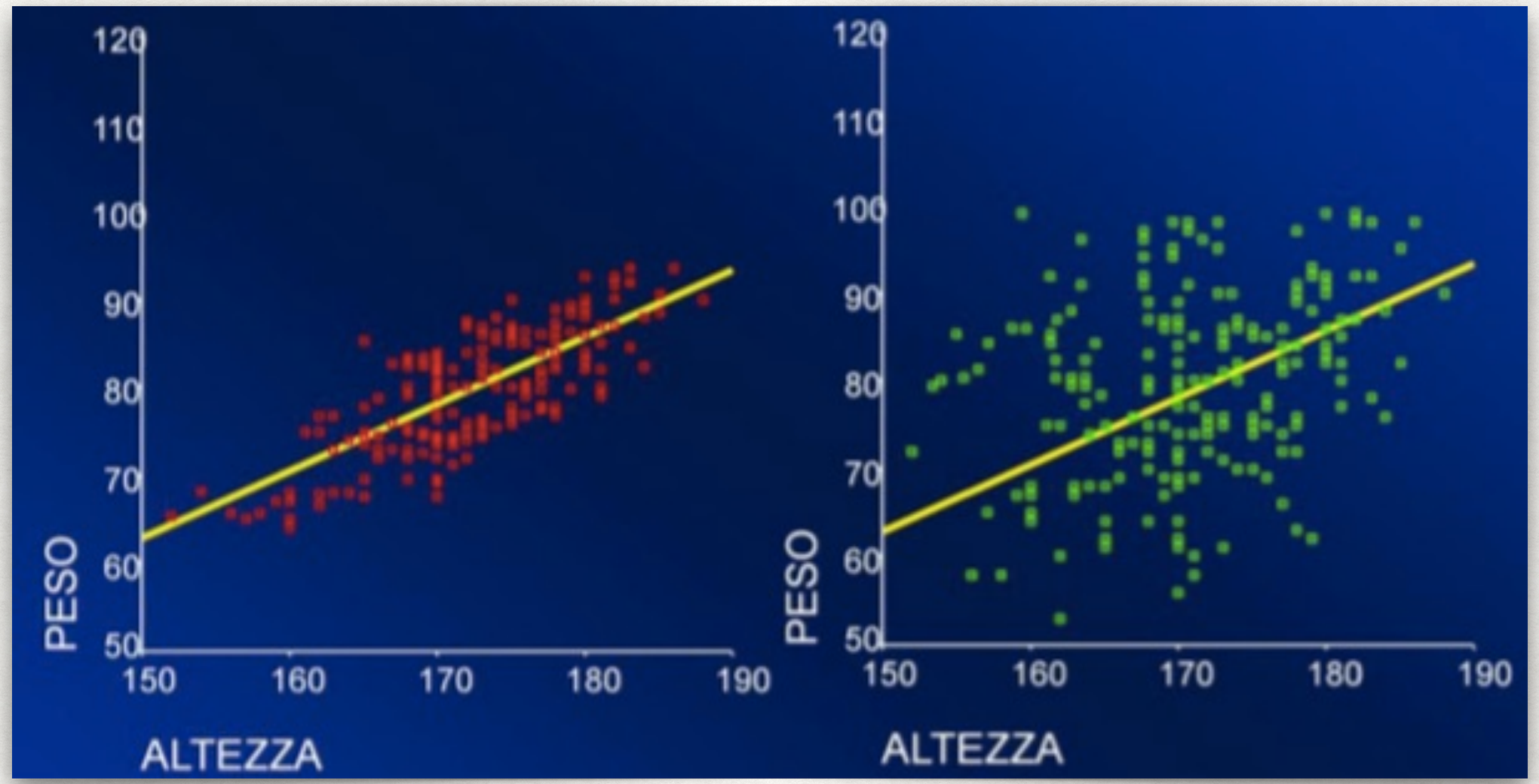
- When the β coefficient is positive, then the two variables are positively related (the dependent variable tends to increase as the independent variable increases, and the line tends to "rise" from left to right).
- When instead the beta is negative, then between the two variables there will be an inverse relationship (the dependent variable tends to decrease as the independent variable increases, and the line tends to "fall" from left to right).
- In the absence of correlation between the two variables, the β coefficient will have a value close to zero; graphically, the line will be roughly parallel to the x-axis.

CORRELATION BETWEEN NORMAL VARIABLES



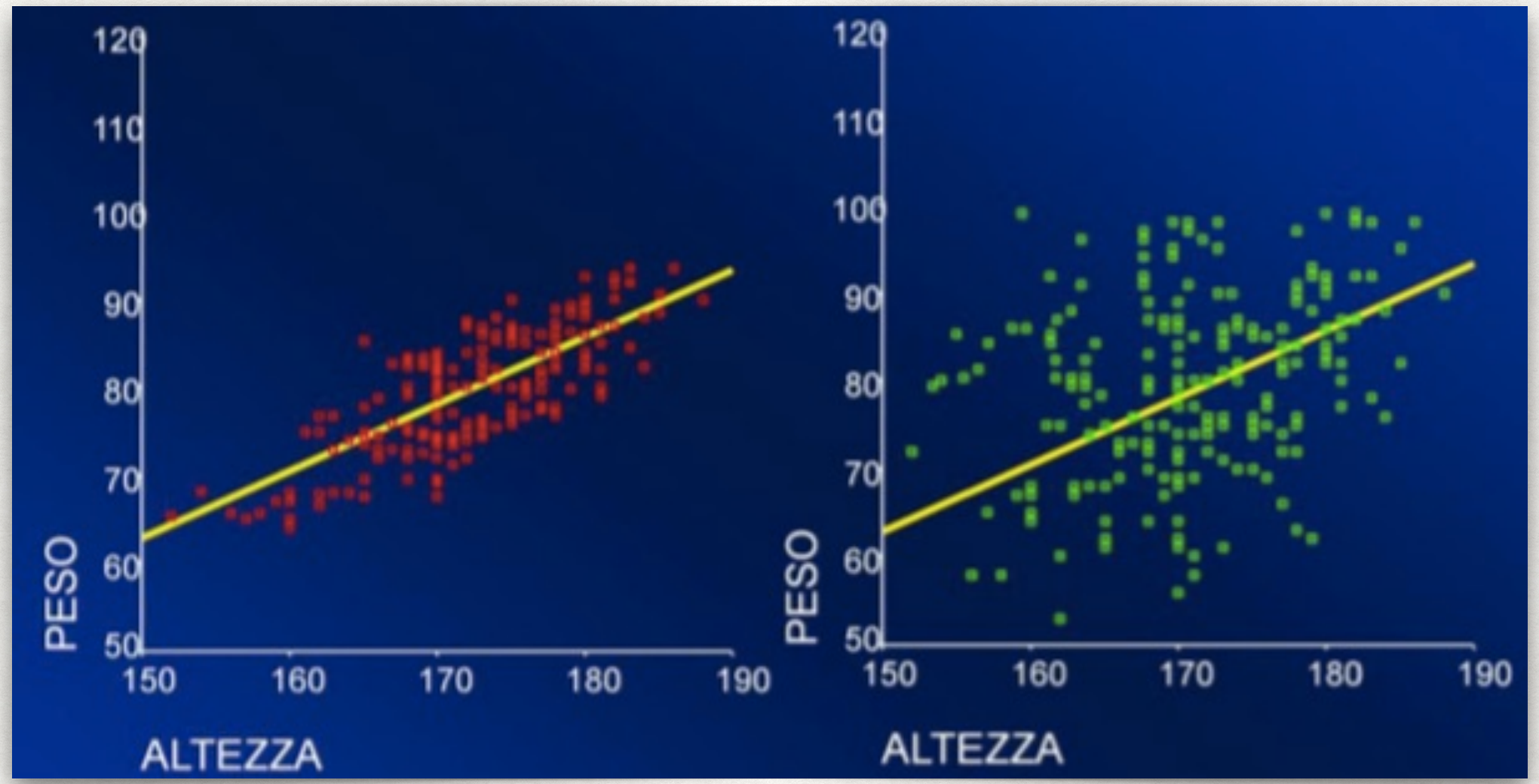
- Although linear regression provides an indication of how much the dependent variable tends to vary as a function of the independent variable, it gives us no indication of how strong that association is.
- Graphically, an initial indication will come from the scatter of the points around the line.

CORRELATION BETWEEN NORMAL VARIABLES



- If the points tend to be very close to the line, assuming a classic linear arrangement (as in the graph on the left) then we can argue that the linear relationship between the two variables is very close.
- If on the contrary the points are very dispersed, forming a cloud (as in the graph on the right) then the association will be less strong.

CORRELATION BETWEEN NORMAL VARIABLES



However, it is possible to go beyond the graphic impression, and express the strength of the association in quantitative terms, using Pearson's correlation coefficient $\boldsymbol{R}$.

CORRELATION BETWEEN NORMAL VARIABLES

PEARSON'S CORRELATION COEFFICIENT R

- The existence of a linear relationship between two normally distributed continuous variables *X* and *Y* can also be expressed by Pearson's correlation coefficient *R*.
- The coefficient *R* varies between -1 and +1, and is defined as their covariance divided by the product of the standard deviations of the two variables:

$$R = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- Covariance measures how much the two variables *X* and *Y* vary together.

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

CORRELATION BETWEEN NORMAL VARIABLES

PEARSON'S CORRELATION COEFFICIENT R

- If $0 < R \leq 0.3$ there is weak correlation; if $0.3 < R \leq 0.7$ there is moderate correlation; if $R > 0.7$ there is strong correlation.
- The sign indicates the direction of correlation:
  - if R is negative, as the independent variable increases, the dependent variable decreases;
  - if R is positive, as the independent variable increases, the dependent variable increases;
  - if $R = 0$ there is no correlation between the two variables.

R^2 indicates the percentage of variability in the dependent variable that is "explained" by the independent variable (also known as the *coefficient of determination*).

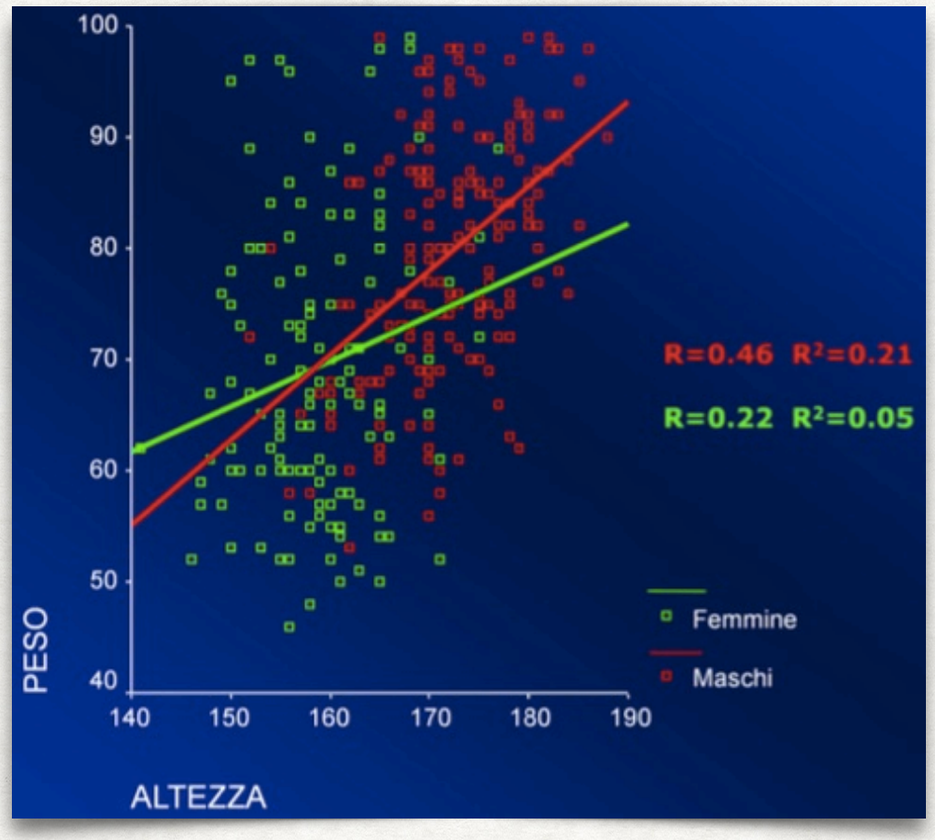
CORRELATION BETWEEN NORMAL VARIABLES

PEARSON'S CORRELATION COEFFICIENT R

- Pearson's correlation coefficient *R* then expresses the strength of the linear association between two normally distributed continuous variables.
- If *R* = 1 or -1, then the correlation is perfect (graphically, all points would fall exactly on the line).
- Therefore, the closer *R* is to 1 or -1, the greater the strength of the association between the two variables.
- The squared value of *R* provides additional information: it tells what percentage of the variability of the dependent variable is explained by the independent variable.
- If *R* = 1 or -1, R^2 would be equal to 1; in other words, 100% of the variability of our dependent variable would be explained by the independent variable.

CORRELATION BETWEEN NORMAL VARIABLES

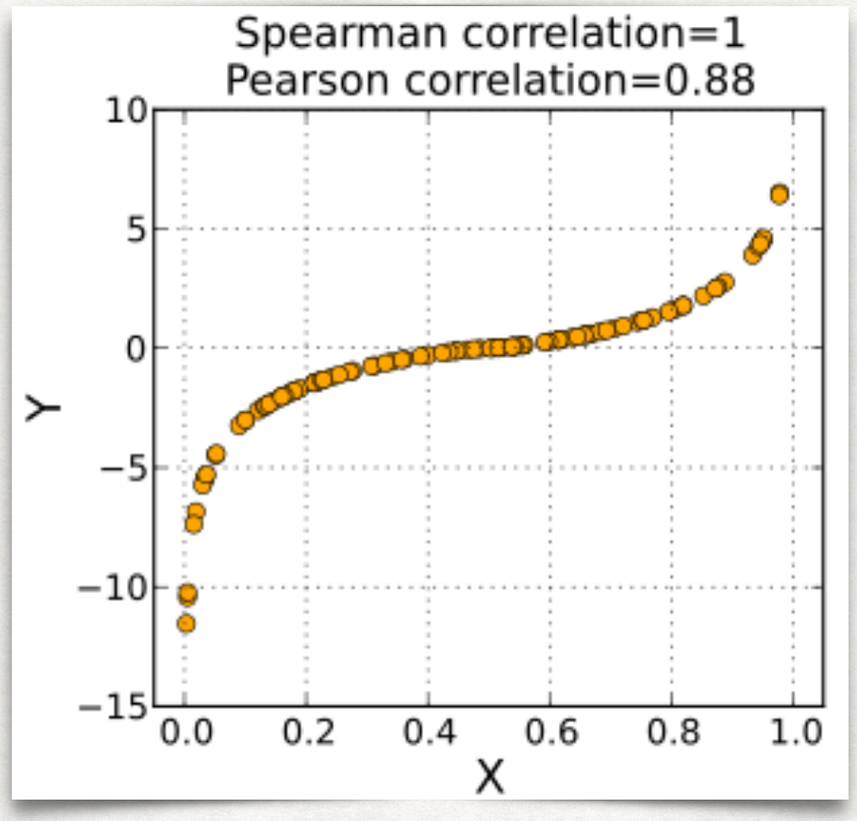
PEARSON'S CORRELATION COEFFICIENT R



- In the case of the weight-height ratio in males and females, it can be seen that the value of *R* is much higher in males.
- This means that, while in males the variability in weight is linked in an important way to the variability in height, in females this correlation is less strong. In other words, among women more than among men, we can find people with very different weights with the same height.
- The R^2 value shows us that among males the variability in height is responsible for 21% of the variability in weight, while among females height explains only 5% of the variability in weight.

CORRELATION BETWEEN NON-NORMAL VARIABLES

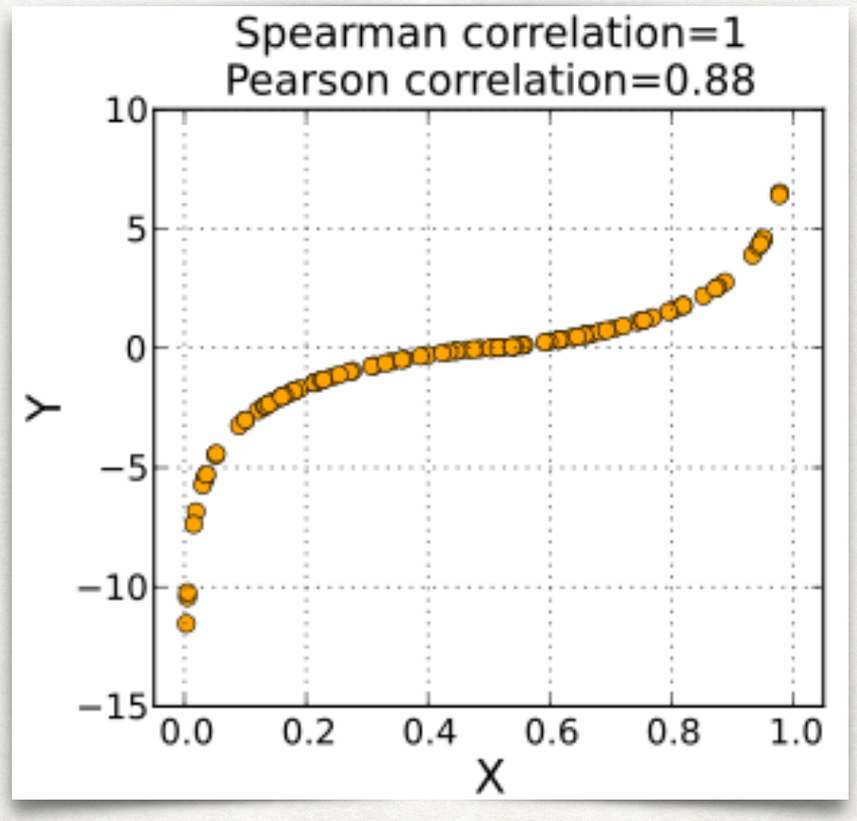
SPEARMAN'S RANK CORRELATION COEFFICIENT



- The **Spearman's rank correlation index** ρ is a nonparametric statistical measure of correlation.
- It measures the degree of relationship between two variables for which no other assumption is made than the ordinal, but possibly continuous, measure.
- Unlike Pearson's linear correlation coefficient, Spearman's coefficient does not measure a linear relationship even if interval measures are used.
- On a practical level, the ρ coefficient is simply a special case of Pearson's correlation coefficient where values are converted to ranks before calculating the coefficient.

CORRELATION BETWEEN NON-NORMAL VARIABLES

SPEARMAN'S RANK CORRELATION COEFFICIENT



- The **Spearman's rank correlation index ρ** is given by the formula:

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} D_i^2}{N(N^2 - 1)}$$

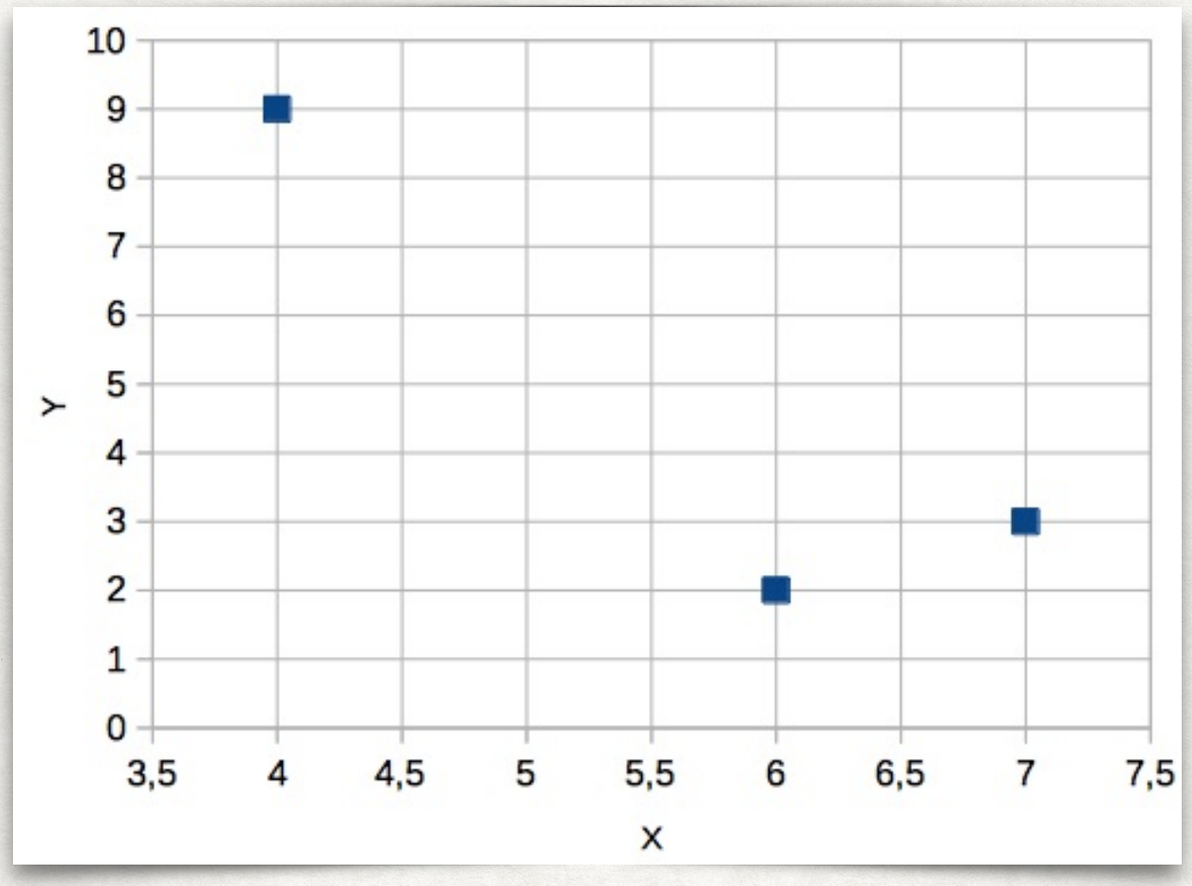
where $D_i = r_i - s_i$ is the difference of ranks (being r_i and s_i respectively the rank of the first variable and the second variable of the i-th observation).

- In the example, Spearman's coefficient is equal to 1. In contrast, Pearson's coefficient is suboptimal.

CORRELATION BETWEEN NON-NORMAL VARIABLES

SPEARMAN'S RANK CORRELATION COEFFICIENT

Data 1	Data 2	Rank 1	Rank 2	d	d^2
6	2	2	1	1	1
4	9	1	3	2	4
7	3	3	2	1	1



- Calculation example.

$$1-\left(\frac{6\Sigma d^2}{n(n^2-1)}\right)=1-\left(\frac{6\times 6}{n(n^2-1)}\right)$$

$$1-\left(\frac{6\times 6}{3(3^2-1)}\right)=-0.5$$

In a spreadsheet, you can use the RANK and CORREL functions.

COMPARISON

BETWEEN GROUPS

COMPARISON OF MEAN VALUES OF TWO (UNPAIRED) GROUPS

What is the difference in mean blood pressure levels between control group and treated patients?

To answer this question, we use **Student's *t*-test**:

$$t = \frac{\text{Difference between sample means}}{\text{Standard error of the difference between sample means}}$$

- Another situation we often run into when analyzing study data is comparing the mean values of a variable between two groups of individuals.
- For example, we might ask whether drug A had a greater effect on blood pressure than placebo.
- If the variable in question is normally distributed, this type of question can be answered using Student's *t*-test for unpaired data.

COMPARISON OF MEAN VALUES OF TWO (UNPAIRED) GROUPS

What is the difference in mean blood pressure levels between control group and treated patients?

To answer this question, we use **Student's *t*-test**:

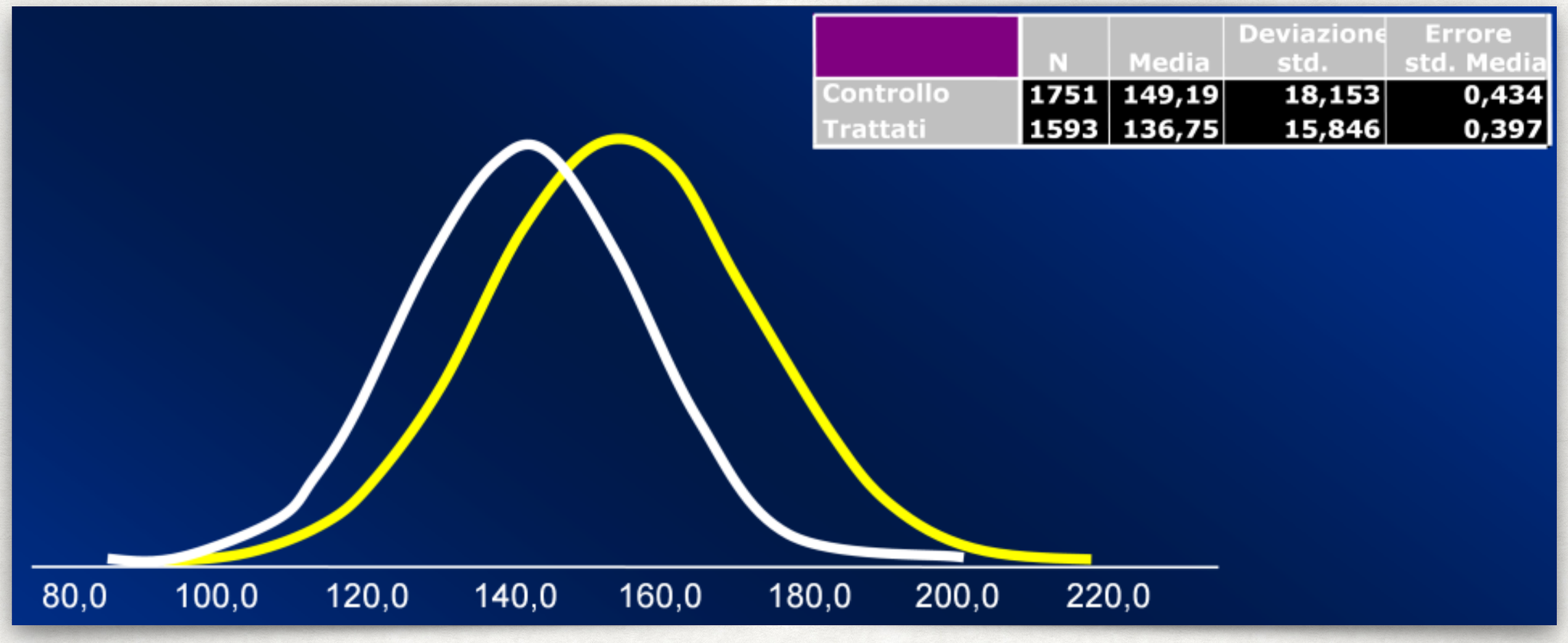
$$t = \frac{\text{Difference between sample means}}{\text{Standard error of the difference between sample means}}$$

- This test is based on the difference between the mean values in the two groups being compared, divided by the standard error of that difference.
- The standard error provides a measure of how accurately this difference between the mean values has been estimated (the more individuals we study, the more accurate the estimate will be).
- The larger the value of *t*, the more confident we will be in rejecting the null hypothesis of no difference between the two mean values.
- In particular, the result will be considered statistically significant ($p<0.05$) if the value of *t* will be > 1.96

COMPARISON OF MEAN VALUES OF TWO (UNPAIRED) GROUPS

STUDENT'S T TEST

	N	Media	Deviazione std.	Errore std. Media
Controllo	1751	149,19	18,153	0,434
Trattati	1593	136,75	15,846	0,397

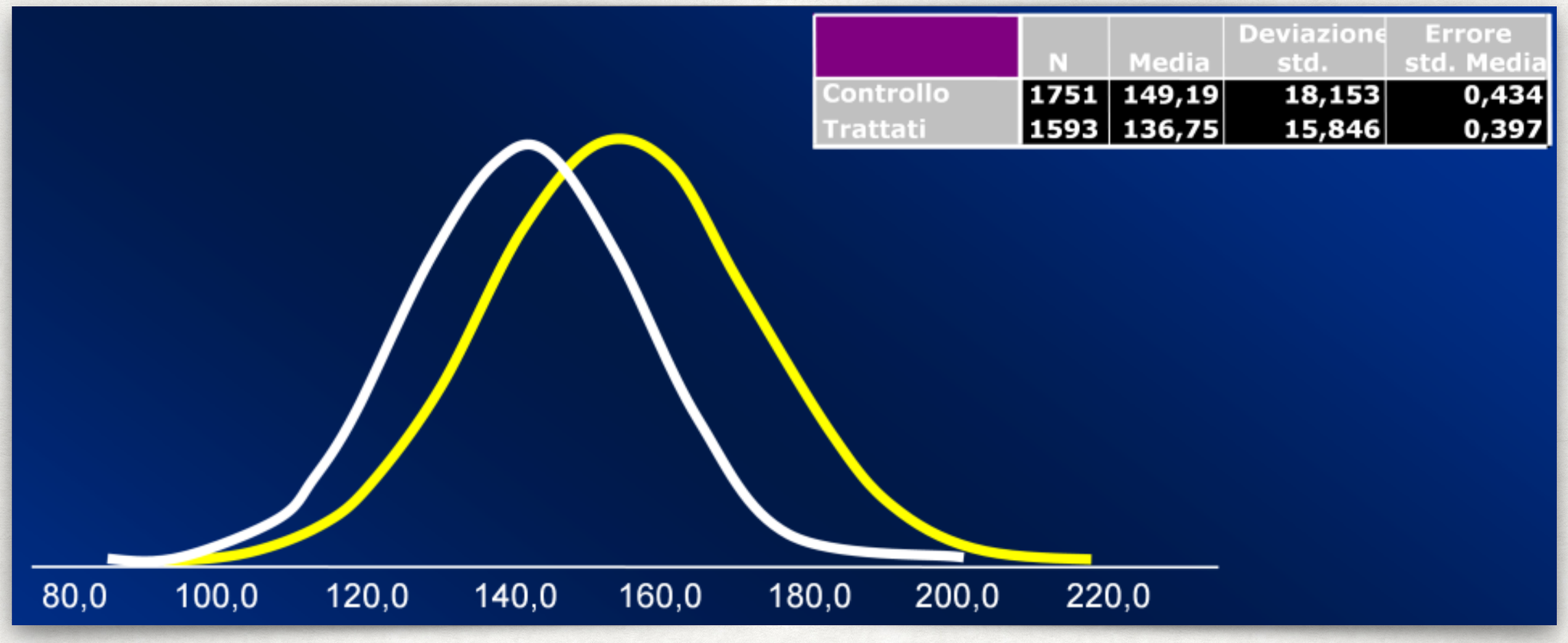


Notice. *The standard error of the difference between the means is calculated as the square root of the sum of their standard errors.*

- Difference between means: 149.19 − 136.75 = 12.44
- Standard error of the difference = $\sqrt{(0{,}434^2+0{,}397^2)} = \sqrt{0.346} = 0.588$
- $t = 12.44 / 0.588 = 21.156 \Longrightarrow p < 0.0001$

COMPARISON OF MEAN VALUES OF TWO (UNPAIRED) GROUPS

STUDENT'S T TEST

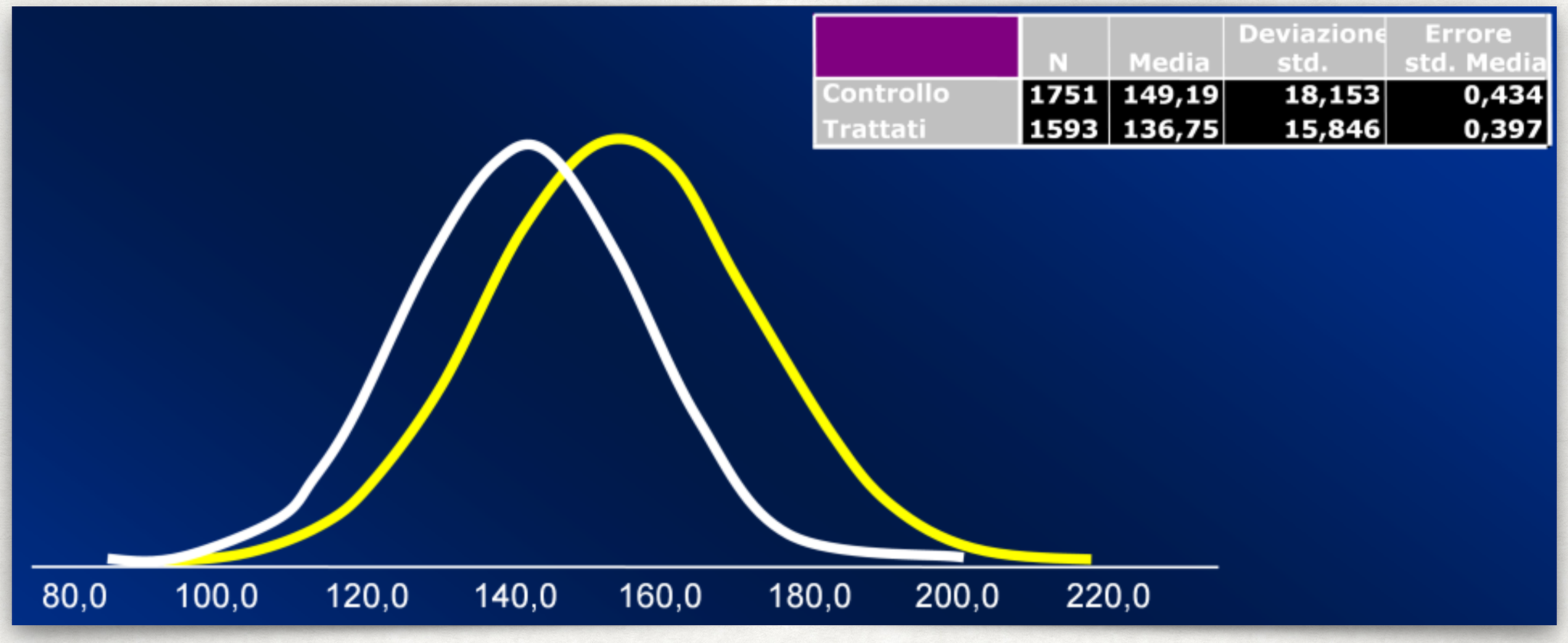


	N	Media	Deviazione std.	Errore std. Media
Controllo	1751	149,19	18,153	0,434
Trattati	1593	136,75	15,846	0,397

- This example represents a practical application of the Student's *t* test for comparison of mean systolic blood pressure values between a group of 1593 subjects treated with an antihypertensive drug and 1751 assigned to the control group, treated with a placebo.
- The mean blood pressure values were approximately 137 mmHg in the first group and 149 mmHg in the second. The graph shows the distribution of pressor values in the two groups.
- The null hypothesis is that the two distributions are not significantly different.

COMPARISON OF MEAN VALUES OF TWO (UNPAIRED) GROUPS

STUDENT'S T TEST



	N	Media	Deviazione std.	Errore std. Media
Controllo	1751	149,19	18,153	0,434
Trattati	1593	136,75	15,846	0,397

- Rejecting the null hypothesis, we state that the pressor values obtained with the study drug are instead significantly lower than those obtained with placebo.
- The application of the student's *t*-test (difference between the means divided by its standard error) will lead to a *t*-value of more than 21, which will correspond to a *p*-value of $p<0.0001$.
- We can therefore conclude that the blood pressure values obtained with the drug are significantly lower than those obtained with placebo.

COMPARISON BETWEEN MEAN VALUES OF A (PAIRED) SAMPLE

STUDENT'S T-TEST (FOR NORMAL DISTRIBUTION)

- In the paired Student's t-test (**paired *t*-test**), a sample is selected and two measurements are taken for each item in the sample (e.g., "before" and "after" a treatment).
- Each set of measurements is considered a sample. Unlike the "unpaired" hypothesis test, the two samples are not independent of each other.
- Paired samples are also called "paired samples" or "repeated measures."
- The distribution of the data must be normal.

COMPARISON BETWEEN MEAN VALUES OF A (PAIRED) SAMPLE

STUDENT'S T-TEST (FOR NORMAL DISTRIBUTION)

- *For example, if you want to determine whether drinking a glass of wine or drinking a glass of beer has the same or different impact on memory, one approach is to take a sample of — say — 40 people, having half of them drink a glass of wine and the other half drink a glass of beer.*
- *You then subject each of the 40 people to a memory test and compare the results.*
- *This is the "independent samples" approach.*

COMPARISON BETWEEN MEAN VALUES OF A (PAIRED) SAMPLE

STUDENT'S T-TEST (FOR NORMAL DISTRIBUTION)

- *Another approach could be to take a sample of 20 people, give each person a glass of wine and then subject them to a memory test.*
- *You then have the same people drink a glass of beer and perform the memory test again. Finally the results are compared.*
- *This is the approach used with **paired samples**.*

The advantage of this second approach is the sample can be smaller. Because the subjects sampled are the same for beer and wine, there is less chance that some external factor (*confounding* variable) will affect the result.

The problem with this approach is that it is possible for the results of the second memory test to be lower simply because the person absorbed more alcohol.

This can be corrected by separating the tests, for example by performing the test with beer one day after the test with wine.

COMPARISON BETWEEN MEAN VALUES OF A (PAIRED) SAMPLE

STUDENT'S T-TEST (FOR NORMAL DISTRIBUTION)

- In the paired Student's t-test (**paired *t*-test**) the purpose is then to compare the means of two samples that are not independent (the measures being taken for the same subjects).
- The analysis is performed on the N **differences** $d_i = (y_i - x_i)$ between individual pairs of observations (y_i=after; x_i=before).
- The null hypothesis is that the mean of the differences is equal to zero:
  $\boldsymbol{H_0: \mu_d = 0}$
- The alternative hypothesis is that the mean of the differences is <u>not</u> equal to zero:
  $\boldsymbol{H_A: \mu_d \neq 0}$
- The statistic that the test calculates is as follows:

$$t = \frac{\mu_d}{\mathrm{SE}(\mu_d)}$$

where $\mathrm{SE}(\mu_d) = \sigma_d / \sqrt{N}$

COMPARISON BETWEEN MEAN VALUES OF A (PAIRED) SAMPLE

STUDENT'S T-TEST (FOR NORMAL DISTRIBUTION)

- Under the null hypothesis, this statistic follows a Student's *t* distribution with *n*-1 degrees of freedom.
- We then use the appropriate table of Student's *t* distribution (with *n*-1 degrees of freedom) to compare the value of the statistic. This provides the *p*-value for the test.
- CAUTION: *The test is valid if the distribution of differences is approximately normal. Therefore, it is not advisable to use the paired t-test if there are extreme outliers or the distribution is strongly non-normal.*
- It would be useful to calculate a confidence interval for the mean difference to know within what limits the true population difference is likely to lie.
- A 95% confidence interval for the true mean difference is given by:

$$\mu_d \pm [t^* \times \mathrm{SE}(\mu_d)]$$

where t^* is point 2.5% of the t distribution with $n-1$ degrees of freedom.

- Notice. *The TTEST function is available in the spreadsheet to calculate both paired and unpaired t-tests.*

COMPARISON BETWEEN MEAN VALUES OF A (PAIRED) SAMPLE

STUDENT'S T-TEST (FOR NORMAL DISTRIBUTION)

Example. Suppose you have the following measures for a group of 20 people:

Person	Before	After	Difference
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1

- Calculating the mean and standard deviation of the differences we have: $\mu_d = 2.05$ and $\sigma_d = 2.837$; thereby $SE(\mu_d) = 2.837 / \sqrt{20} = 0.634$
- So $t = 2.05 / 0.634 = 3.231$ with 19 df
- From the table of Student's t distribution with 19 degrees of freedom we have $p = 0.004$
- Therefore, there is strong evidence that, on average, treatment leads to improvement
- Furthermore, since the point at 2.5% of the t distribution with 19 df is 2.093, the 95% CI will be: $2.05 \pm (2.093 \times 0.634) = [0.72, 3.38]$
- We can therefore be 95% confident that the improvement in the true mean difference lies between 0.72 and 3.38

COMPARISON OF MEAN VALUES OF SEVERAL GROUPS

ANOVA

In case we want to compare the average values of several groups (with normal distribution), we use the **Analysis of Variance** (ANOVA)

- If one is interested in comparing the mean values of a variable between more than two groups (e.g. systolic blood pressure comparing three different treatments), it will be necessary to use a statistical test that represents an extension of Student's *t*-test: the analysis of variance.
- To describe in detail this type of analysis is beyond our scope; it is sufficient to know that we start from the null hypothesis that the average values (three or more) compared are equal.
- The test produces a value (which is called *F*), to which corresponds a *p* value; if the latter is < 0.05, then we reject the null hypothesis and conclude that the compared mean values are not equal.

CORRELATION BETWEEN CATEGORICAL VARIABLES

What is the relationship between treatment and event?

	Event YES	Event NO	Total
Placebo	20	10	30
Active drug	8	22	30

- To answer this question (finding a correlation between two categorical variables), one must first know the number of patients in each response category (both in absolute numbers and as a percentage).
- Thus, we construct the **2x2 contingency table**, which directly compares the distribution of events according to the treatments administered.

Among the most widely used tests for statistical analysis of clinical data is the chi-square test (χ^2). This test is used to compare percentages between two or more classes of a variable. If, for example, we are comparing the effect of a drug versus placebo with respect to the development of a particular event (e.g., occurrence of myocardial infarction, mortality, development of complications, etc.), then the question we will ask is this: *is the percentage of events that occurred in the subjects treated with the drug lower, equal, or higher than that which occurred in the subjects who took the placebo?*

CORRELATION BETWEEN CATEGORICAL VARIABLES

What is the relationship between treatment and event?

	Event YES	Event NO	Total
Placebo	20	10	30
Active drug	8	22	30

- To answer this question, it is first necessary to construct a **contingency table**, in which we will report, in four separate cells, the number of subjects treated with the drug or placebo, which have developed or not the event of interest.
- In the example shown, a total of 60 patients were studied, 30 of whom were treated with the active drug and 30 with the placebo.
- Among the subjects treated with the active drug, 8 developed the event and 22 did not, whereas among those who took the placebo, 20 out of 30 subjects developed the event. Is this difference statistically significant?

CORRELATION BETWEEN CATEGORICAL VARIABLES

Drug/Observed event Contingency Table

	Event YES	Event NO	Total
Placebo n° (%)	20 (66.7%)	10 (33.3%)	30 (100%)
Drug n° (%)	8 (28.6%)	22 (73.3%)	30 (100%)
TOTAL n° (%)	28 (46.7%)	32 (53.3%)	60 (100%)

% of observed events

Drug/Expected event Contingency Table

	Event YES	Event NO	Total
Placebo n° (%)	14 (46.7%)	16 (53.3%)	30 (100%)
Drug n° (%)	14 (46.7%)	16 (53.3%)	30 (100%)
TOTAL n° (%)	28 (46.7%)	32 (53.3%)	60 (100%)

% of expected events assuming that drug or placebo have the same efficacy

- If in the contingency table we report the percentages of events/non-events in relation to treatment, we see that the percentage of patients who had an event is 66.7% among those who took placebo and 28.6% among those treated with the drug.
- Adding up the number of patients who had an event, regardless of the treatment received, we see that this is 28.

CORRELATION BETWEEN CATEGORICAL VARIABLES

Tabella di contingenza farmaco/evento <u>osservato</u>

	Event YES	Event NO	Total
Placebo n° (%)	20 (66.7%)	10 (33.3%)	30 (100%)
Drug n° (%)	8 (28.6%)	22 (73.3%)	30 (100%)
TOTAL n° (%)	28 (46.7%)	32 (53.3%)	60 (100%)

% of observed events

Tabella di contingenza farmaco/evento <u>atteso</u>

	Event YES	Event NO	Total
Placebo n° (%)	14 (46.7%)	16 (53.3%)	30 (100%)
Drug n° (%)	14 (46.7%)	16 (53.3%)	30 (100%)
TOTAL n° (%)	28 (46.7%)	32 (53.3%)	60 (100%)

% of expected events assuming that drug or placebo have the same efficacy

- At this point, the null hypothesis to be tested is that the percentage of events is the same in the two groups compared.
- Since the two groups are of equal size — the total number of events being 28 — we should expect, if the null hypothesis were true, that half of these events would have occurred among the patients who took the placebo and the other half among the patients who took the active treatment.

CORRELATION BETWEEN CATEGORICAL VARIABLES

	Event YES	Event NO	Total
PLACEBO observed	20	10	30
expected	14	16	
DRUG observed	8	22	30
expected	14	16	
TOTAL n°	28	32	60

Chi-square χ^2 test

$$\chi^2 = \sum_i \frac{(\text{Osservati}_i - \text{Attesi}_i)^2}{\text{Attesi}_i}$$

$\chi^2 = (20-14)^2 / 14 + (10-16)^2 / 16 + (8-14)^2 / 14$
$+ (22-16)^2 / 16 = 9.643$

- The chi square test is based on calculating the differences between observed and expected values within each of the 4 cells.
- The value of each difference (Observed-Expected) is squared and divided by the respective expected value. The values obtained are then summed, as in the example shown.
- A chi-square value of 3.84 corresponds to a value of p=0.05.
- If we obtain then, in a 2x2 table, a value $\chi^2 \geq 3.84$ (i.e. $p \leq 0.05$), we will reject the null hypothesis and conclude that the percentages are significantly different.
- In the case of the example, a chi square of 9.643 corresponds to a value of p=0.002.

FROM UNIVARIATE TO MULTIVARIATE ANALYSIS

FROM UNIVARIATE TO MULTIVARIATE ANALYSIS

- Whenever, in the context of a trial, there is an imbalance in one or more important prognostic factors or we want to estimate the role of a factor in the absence of confounding problems created by other variables, **it is necessary to use multivariate analysis**.
- In controlled clinical trials randomization is the most effective technique to ensure that the groups compared are as comparable as possible with regard to all clinical and socio-demographic characteristics.
- Only in this way do we have sufficient assurance that, should differences in efficacy be documented between the compared treatments, these differences can actually be attributable to the treatment and not to underlying differences in the compared groups.
- However, because randomization is by nature an entirely random assignment process, it may be the case that the study groups may differ with respect to one or more characteristics deemed important.
- It will therefore be necessary to account for these imbalances in the analysis of the study data, using appropriate multivariate analysis techniques.

FROM UNIVARIATE TO MULTIVARIATE ANALYSIS

	Drug A	Drug B	*p*
HbA_{1c}	7.1±1.6	7.5±1.7	0.0001
Age	62±10	64±11	0.0001
BMI	27.5±3.9	28.4±5.1	0.0001
Duration	11.2±8.6	11.7±8.5	0.1300

- Suppose we have two groups of patients with type 2 diabetes mellitus, one of whom has been randomized to be treated with hypoglycemic A and the other with hypoglycemic B.
- From the table, we infer that metabolic control, expressed as mean values of glycosylated hemoglobin (HbA_{1c}), is significantly(*) better with drug A than with drug B.
- However, analyzing the characteristics of patients in the two groups, we observe that patients assigned to drug B are significantly(*) older and have a significantly(*) higher body mass index (BMI).

Is drug A really more effective than drug B, or is the difference in metabolic control related to younger age and lower BMI?

What would be the true difference in HbA_{1c} values if the two groups were the same age and BMI?

(*) Statistically significant…

FROM UNIVARIATE TO MULTIVARIATE ANALYSIS

	Drug A	Drug B	*p*
HbA_{1c}	7.1±1.6	7.5±1.7	0.0001
Age	62±10	64±11	0.0001
BMI	27.5±3.9	28.4±5.1	0.0001
Duration	11.2±8.6	11.7±8.5	0.1300

- In contrast, the two groups did not differ (significantly) in the duration of diabetes.
- Questions arise at this point:
  - *Are the differences in mean HbA_{1c} values really related to treatment, or are they a consequence of differences in age and body weight?*
  - *What would be the true effect of treatment if the two groups were the same age and BMI?*
- These questions can be answered using multivariate analyses, which allow us to assess the independent role of each factor considered, all else being equal.

Is drug A really more effective than drug B, or is the difference in metabolic control related to younger age and lower BMI?

What would be the true difference in HbA_{1c} values if the two groups were the same age and BMI?

FROM UNIVARIATE TO MULTIVARIATE ANALYSIS

Multivariate analysis allows us to assess the independent role of each individual variable (factor), all else being equal.

$y = \alpha + \beta x \Leftarrow$ univariate analysis

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots \Leftarrow$ multivariate analysis

$$\mathrm{HbA}_{1c} = \alpha + \beta_1 \cdot \mathrm{drug} + \beta_2 \cdot \mathrm{BMI} + \beta_3 \cdot \mathrm{age} + \beta_4 \cdot \mathrm{duration}$$

- In the case of a continuous dependent variable, as in our example, the appropriate multivariate analysis is represented by **multiple linear regression**, which is the normal extension of simple linear regression, of which we have already spoken.
- Unlike simple linear regression, in which we had only one independent variable, in this case we are going to evaluate the extent to which the dependent variable varies as more than one independent variable varies. In this case, therefore, we will estimate more parameters, one for each independent variable included in the model.
- In our example, we are going to evaluate how the levels of HbA_{1c} vary with the variation of the drug used, the BMI, the age of the patients and the duration of the disease.

FROM UNIVARIATE TO MULTIVARIATE ANALYSIS

	β	*p*
HbA_{1c}	0.31	0.000
Age	-0.0053	0.08
BMI	18	0.007
Duration	14	0.001

Interpretation

- *With equal age, BMI and duration of diabetes, drug B is associated with a significantly higher HbA_{1c} of 0.31 compared to drug A;*
- *HbA_{1c} values also increased significantly with increasing BMI and duration of diabetes, whereas they tended to decrease with increasing age (although the latter result did not reach statistical significance).*

- The results of the multiple regression applied to our example are shown in the table. It shows us that the type of treatment is significantly related to HbA_{1c} values, as well as BMI and duration of diabetes.
- The interpretation of the results is therefore as follows: with the same age, BMI and duration of diabetes, HbA_{1c} values are significantly higher in patients treated with drug B than in those treated with drug A (the β value tells us that on average they are 0.31 higher).
- The regression also provides us with other insights, as we now know that, given the same treatment, metabolic control is still significantly worse as BMI and diabetes duration increase, whereas it does not change substantially as age changes.

LOGISTIC REGRESSION

- In the previous example, the **dependent variable** (HbA_{1c}) was a **continuous**, **normally distributed** variable. **Multiple regression** is used in these cases.
- In cases where the **dependent variable** is **dichotomous** (Yes/No), **binary logistic regression** is used.
- Similarly, for continuous variables with distributions far from normal, ordinal or nominal variables at multiple levels, logistic regression is used after dichotomizing the dependent variable.
- Thus, if the end-point of the study under consideration is represented by a categorical variable instead of a continuous one, the multivariate analysis model to be used is represented by binary logistic regression.
- It should also be remembered that parametric tests, to which multiple linear regression also belongs, require that the dependent variable be normally distributed. If it is not, it is preferable to dichotomize the variable with respect to a clinically relevant value, and use logistic regression.

LOGISTIC REGRESSION

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots$$

- Logistic regression is an extension of the χ^2 test, and the results are expressed as "**odds ratio**" (OR) with 95% confidence intervals (CIs):

$$\mathrm{OR} = e^{\beta_1}$$

- An OR > 1 indicates excess risk, whereas an OR value < 1 indicates lower risk than the reference category.
- If the 95% confidence interval includes the null value of 1, then the result is not statistically significant.

LOGISTIC REGRESSION

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots$$

- The equation describing logistic regression is not very different from that of multiple regression. In this case too, the linear relationship between the dependent variable and a series of independent variables is studied, each of which corresponds to a β coefficient.
- What distinguishes logistic regression from multiple regression is the quantity to the left of the equation.
- Here in fact we no longer have the value of y, i.e. the dependent variable, but the logarithm of the *odd*, i.e. the probability of having the event of interest divided by the probability of not having it.
- This characteristic means that the exponential of each β coefficient will represent the odds ratio, and therefore will express the risk of having the event for a specific category with respect to the reference category.

ODDS RATIO

The term **odds** refers to the ratio of the probability *p* of an event to the probability that this event does not happen (i.e., the 1-*p* probability of the complementary event):

$$\frac{p}{1-p}$$

Example (source: Wikipedia).

- In the U.S. town of Framingham, of a total of 656 subjects in the 50- to 60-year age group at the time of the first survey, 130 of them developed coronary artery disease during a 12-year follow-up.
- The probability *p* of the event "development of coronary artery disease" is given by the proportion 130/656 = 0.20 = 20%.
- The *odds* of developing coronary artery disease is: *p* / (1-*p*) = 0.20 / (1-0.20) = 0.20 / 0.80 = 0.25; that is, 1 to 4 (25%).

ODDS RATIO

- The natural logarithm of the *odds* is called the **logit**.
- The ratio between two *odds* is called odds ratio (OR) and measures the association between two factors, for example between a risk factor and a disease.
- The calculation of the odds ratio involves comparing the frequencies of occurrence of the event (e.g., disease) in subjects exposed and unexposed to the risk factor under study, respectively.
- If OR = 1, it means that the risk factor is irrelevant to the occurrence of disease.
- If OR > 1, the risk factor is or may be implicated in the occurrence of disease.
- If OR < 1, the risk factor is actually a defense against the disease.
- It is used in retrospective (case-control) studies, where data collection over time is not necessary; in fact, it does not calculate a trend and is, indeed, independent of the duration factor.

RELATIVE RISK

- In prospective studies, the calculation of relative risk is used for the same purpose.
- The **relative risk** (*risk rate*, **RR**) is the probability that a subject, belonging to a group exposed to certain factors, will develop the disease, compared to the probability that a subject belonging to a group not exposed will develop the same disease.
- This index is used in cohort studies where exposure is measured over time:

$$\mathrm{RR} = \frac{I(\text{exposed})}{I(\text{not exposed})}$$

where I = incidence, which is defined as:

$$\mathrm{I} = \frac{\text{number of new patients}}{\text{total number of people - number of patients}}$$

- If:
  - RR = 1 the risk factor has no influence on the occurrence of the disease;
  - RR > 1 the risk factor is involved in the occurrence of the disease;
  - RR < 1 the risk factor depends on the disease (defense factor).
- Examples of the application of this formula are studies concerning the correlation between smoking and the development of lung cancer, in which RR > 17 were found.

WEB RESOURCES FOR STATISTICAL CALCULATION

WEB RESOURCES FOR STATISTICAL CALCULATION

Course of **Data Analysis**

(teacher Prof. Barbaranelli, La Sapienza)

https://elearning2.uniroma1.it/course/view.php?id=1796

This course provides the necessary foundation for the study of the major techniques of multivariate data analysis.

Real Statistics Package

http://www.real-statistics.com

This site is a practical guide on how to do statistical analysis with a spreadsheet, without the need for sophisticated (and expensive) specialized software. It includes free software that extends the statistical capabilities of the spreadsheet, so that a wide variety of statistical analyses can be performed with ease.

WEB RESOURCES FOR STATISTICAL CALCULATION

Online text **Handbook of Biologica Statistics**

(J.H. McDonald)

http://www.biostathandbook.com

This is a complete and comprehensive textbook available online, showing how to make the choice of the appropriate statistical test for a particular experiment, and then apply it and interpret the results. It is accompanied by many examples and resources available on the web.

Statistical and Data Mining Software **Tanagra**

http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

TANAGRA is a data mining software for academic and research use. It offers a variety of data mining methods from exploratory data analysis, statistical learning, machine learning and databases.

WEB RESOURCES FOR STATISTICAL CALCULATION

Data Analysis and Data Mining Software **Orange**

https://orange.biolab.si/

ORANGE is an open source machine learning and data visualization software for professional and personal use. It has an interactive workflow-oriented application interface for data analysis, with a rich toolbox of methods and algorithms.

Statistical and exploratory data analysis software **Jamovi**

https://www.jamovi.org

JAMOVI is a free, open source statistical spreadsheet based on the R statistical language that is easy and straightforward to use. It is particularly suitable for rapid exploratory analysis of datasets, and contains effective tabular and graphical presentations of data.

Thanks for your attention