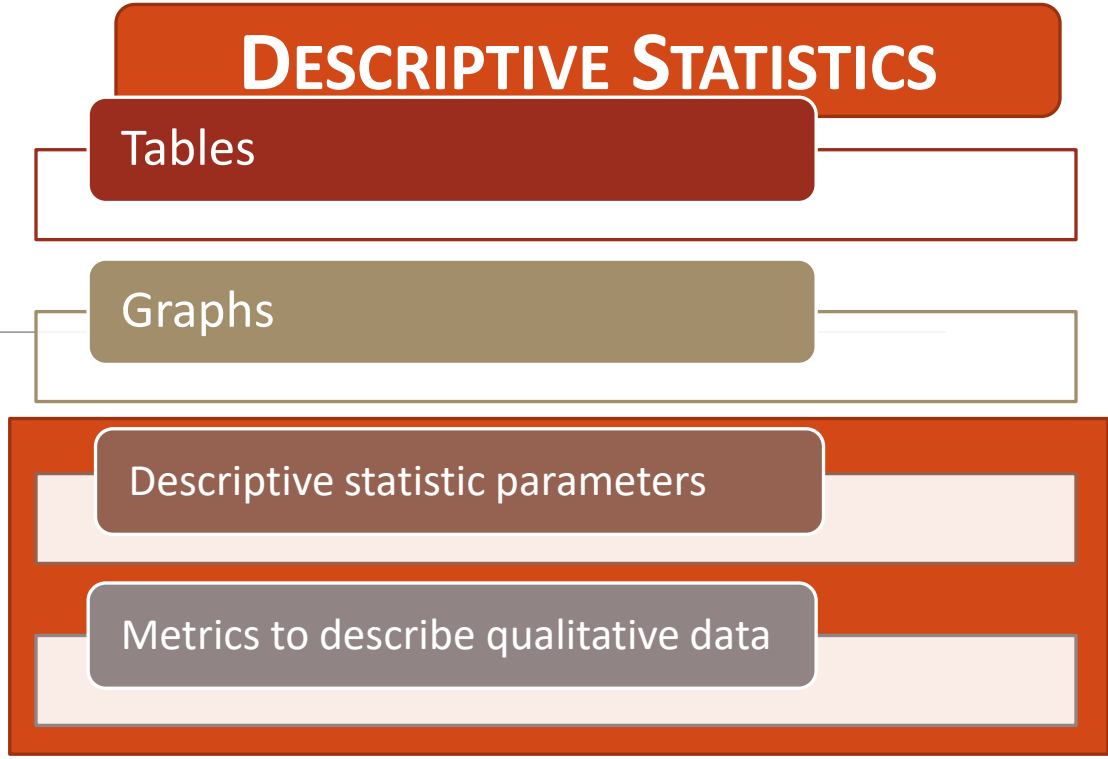
DESCRIPTIVE STATISTICS II

OUTLINE



Metrics to describe **qualitative data**

- Report
- Proportion
- Rate

Descriptive statistic parameters

- Measures of centrality
- Measures of spread
- Measures of symmetry
- Measures of localization

Applied on quantitative variables or qualitative ordinal variables

Metrics for qualitative variables

REPORT

PROPORTION

RATE

Report

Is calculated just for positive the number 'a' and 'b' when b≠0.

- Formula: a:b OR a/b
- Denominator ('b') does not necessarily include the numerator's subjects ('a').

2 out of the 10 persons consulted on a specific date by a family doctor had the value of systolic blood pressure higher than normal. Which is the report SBP normal / SBP not-normal?

The SBP normal / SBP not-normal = 8/2 = 4 → There is one subject with a not-normal value of SBP to every 4 subjects with normal SBP

Proportion

Is the report in which the numerator is part of the denominator.

- Formula: a:(a+b) OR a/(a+b)
- Take values between 0 and 1 (0 and 100 if it is expressed as %).

Prevalence is a proportion defined by the formula:

Prevalence = (number of cases of disease) / (volume of the population)

Incidence = the number of new cases of disease occurring in a population at risk over a period of time

Incidence = (the number of new cases of disease over a period of time) / (population at risk in the same period of time)

Proportion

At the emergency service of a county hospital, 1200 patients were presented for consultation. 420 were hospitalized (200 women and 220 men).

- Proportion of hospitalized subjects = 420/1200*100 = 35%
- Proportion of female subjects hospitalized = 200/420*100 = 48%
- Proportion of male subjects among hospitalized subjects = 220/420*100 = 52%

Rate

The rate reflects the risk of an event occurring over time (e.g., the number of subjects per unit of time - second / minute / hour / day / week / month / year).

- Has positive values in the range [0, ∞)
- Examples: morbidity rate, attack rate, mortality rate, natality rate

In a city with a population of 100,000, 200 live births were registered in 1999. What is the birth rate?
Birth rate = 200 / 100,000*1,000 = 2 ‰

Rate

Relative risk (RR symbol) is the ratio of the incidence rate to those exposed and the incidence rate in those unexposed.

- RR = 1 → the risk in the exposed group is equal to the risk in the non-exposed group.
- RR > 1 → the exposure is a risk factor for the pathology of interest.
- RR < 1 → the exposure is a protective factor for the pathology of interest.

Rate

The attributable risk (symbol AR) is given by the difference between the incidence rate in the exposed group (I_e) and the incidence rate in the unexposed group (I_n).

The percentage attributable risk (AR%) is the percentage of the exposure pathology of interest due to exposure:

$$RA\% = 100 * [(Ie-In) / Ie]$$

RA% is the percentage of the pathology of interest in the exposed group that will be removed if the exposure ceases.

Descriptive statistic parameters

How data resulted from medical studies are summarized is dictated by the type of variable and the scale of measurement. The same principle is applied also when statistics descriptive metrics are calculated.

Descriptive statistics parameters

Measures of Centrality	Measures of Spread
○ Mean ○ Mediana ○ Mode ○ Central value ○ …	○ Range (amplitude) ○ Variance ○ Standard deviation ○ Coefficient of variance ○ Standard error
Measures of Symmetry	**Measures of Localization**
○ Skewness ○ Kurtosis	○ Quartile ○ Percentiles

Measures of centrality

Simple values that give us information about the distribution of data

Parameters:

- Mode
- Median
- Mean (arithmetic mean)
- Geometric mean
- Harmonic mean
- Central value

Measures of centrality: mode

Called also Modal Value

- Is the most frequent value on the sample

There is no mathematical formula for calculus

Correspond the value of the highest pick on the graphic of frequency distribution

- Identify the mode for all previously graphical presentations

MODE(number*1*, number*2*, …, number*n*)

Measures of centrality: mode

Unimodal series:

2	1	2	1	1

- The age of patients hospitalized with the diarrheic syndrome at 1st Pediatric Clinic between 11.01 – 11.08.2008

Bimodal series:

2	1	2	1	1
2	2	1	3	3

Trimodal series (Multimodal):

2	1	2	1	1
2	3	3	3	4

Measures of centrality: mode

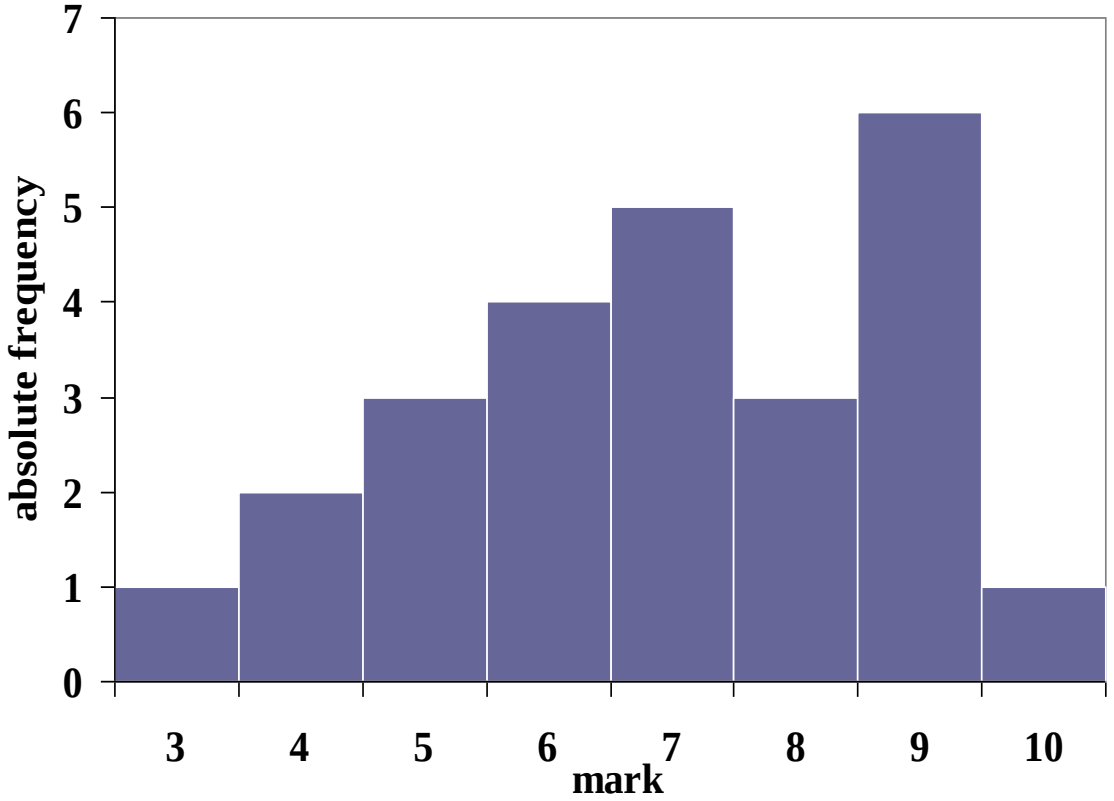
Extreme values do not influence it.

For a sample of

n = 25 students the marks of the practical exam at Informatics were:

3, 4, 9, 5, 4, 6, 7, 7, 8, 5, 9, 7, 9, 5, 6, 9, 10, 6, 7, 7, 8, 9, 8, 9, 6

Mode = 9



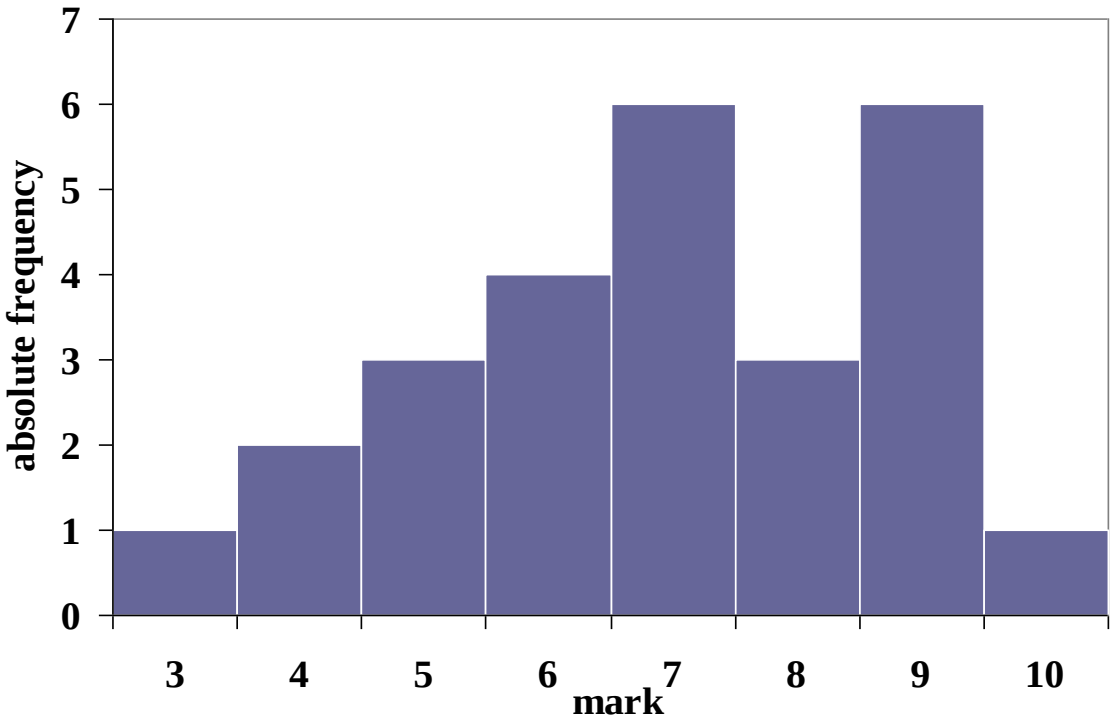
Measures of centrality: mode

Bi-modal series

For a sample of 26 students, the marks obtained at Informatics exam were:

3, 4, 9, 5, 4, 6, 7, 7, 8, 5, 9, 7, 9, 5, 7, 6, 9, 10, 6, 7, 7, 8, 9, 8, 9, 6

Mode = 7 & 9



Measures of centrality: median

Is the value that split the series of data into two half

Steps to finding the median:

- Sort the data ascending
- Locate the position of the median in the string and determine its value
- Its value is equal to the value of the 50th percentile

If the sample size is odd, we will use the following formula:

$$Me = X_{\frac{n+1}{2}}$$

If the sample is even, we will use the following formula:

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

Measures of centrality: median

1. It is not affected by extreme values of data series.
2. The median value could be not representative for the data on the series if individual data did not group in the neighbour of the central value (median).
3. Median is a measure of central tendency that minimizes the sum of absolute values of deviations from a value X on the line of the real numbers.

Measures of centrality: median

3, 4, 9, 5, 4, 6, 7, 7, 8, 5, 9, 7, 9, 5, 7, 6, 9, 10, 6, 7, 7, 8, 9, 8, 9, 6

Numbers are ordered ascending:

3	4	4	5	5	5	6	6	6	6	7	7	7	7	7	7	8	8	8	9	9	9	9	9	9	10
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}

- n = 26 (even number)
- $Me = (X_{13}+X_{14})/2 = (7+7)/2 = 7$

  = MEDIAN(number*1*, number*2*, …, number*26*)

Measures of centrality: mean

The sum of all data series divided by the sample size

Changing a single data series does not affect modal or median values but will affect the arithmetic mean

Population (the mean of a variable in a population is known):

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

Sample (is necessary to be calculated):

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

Measures of centrality: mean

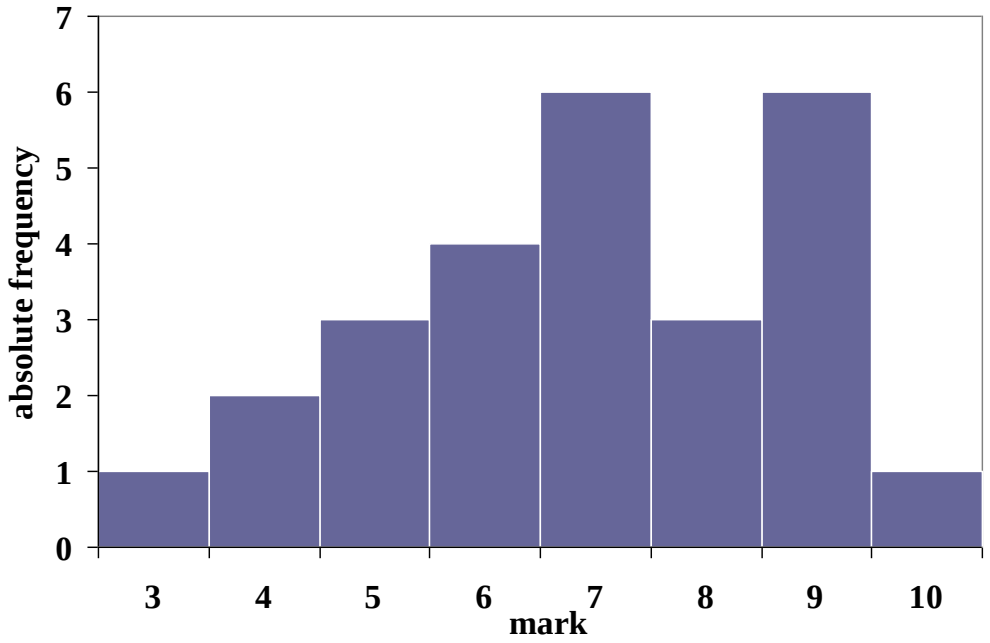
3	4	4	5	5	5	6	6	6	6	7	7	7	7	7	7	8	8	8	9	9	9	9	9	9	10
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}

Arithmetic mean:

= (3+4+…+9+10)/26

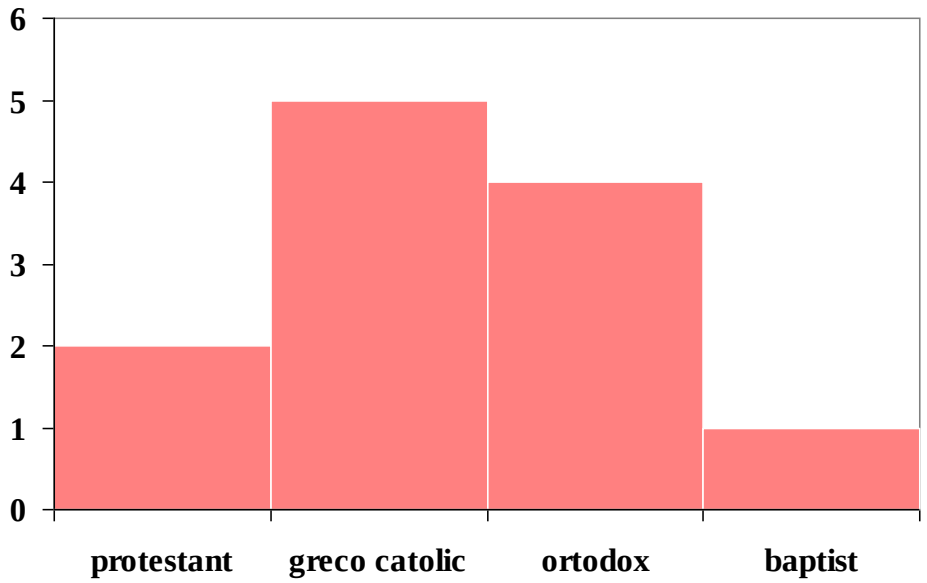
= 6.92

=AVERAGE (number*1*,…, number*26*)



Measures of centrality: mean

- Is the preferred measure of centrality both as a parameter (population) or as a statistic (sample).
- It has significance just IF the variable of interest is quantitative.



Measures of centrality: mean

Properties:

1. Any value of the series is taken into account in calculating the mean.
2. Outliers may influence the arithmetic mean by destroying its representativeness.
3. The value of the arithmetic mean is among the data series.
4. The sum of the differences between individual values and mean is zero:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

Measures of centrality: mean

Properties:

5. Changing the origin of the measurement scale of X-variable will influence the mean, Let X" = X + C (where C is a constant).
6. Transformation of the measurement scale of X-variable will influence the mean, Let X" = h*X (where h is a constant).
7. The sum of squares of deviations from the arithmetic mean is the minimum sum of squares of deviations from X of the values of series.

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \min_{X \in R} \sum_{i=1}^{n}(X_i - X)^2$$

Measures of centrality: weighted mean

Every X_i value is multiplied with a non-negative weight W_i, which indicate the importance of the value reported to all other values:

$$m_x = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

If the weights W_i are chosen to be equal and positive, we will obtain the arithmetic mean.

Measures of centrality: geometric mean

Used to describe the proportional growth (including exponential growth)

$$G = \sqrt{X_1 \cdot X_2}$$

Medical application:

- reporting experimental IgE results [Olivier J, Johnson WD, Marshall GD, The logarithmic transformation and the geometric mean in reporting experimental IgE results: what are they and when and why to use them? Ann Allergy Asthma Immunol. 2008;100(4):333-7.]
- The intravaginal ejaculation latency time (IELT) [Waldinger MD, Zwinderman AH, Olivier B, Schweitzer DH, Geometric mean IELT and premature ejaculation: appropriate statistics to avoid overestimation of treatment efficacy. J Sex Med. 2008;5(2):492-9.]

Measures of centrality: harmonic mean

Used to average the rates

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Example:

- A blood donor fills a 250 mL blood bag at 70 mL on the first visit and 90 mL on the second visit, What is the average rate at which the donor fills the bag?
- Harmonic mean = 2/(1/70+1/90)= 78.75 mL
- Arithmetic mean = 80 mL

Central value

$$\text{Central value} = \frac{X_{min} + X_{max}}{2}$$

Advantages and disadvantages

Average	Advantages	Disadvantages
Mean	Use all data Mathematically manageable	Influenced by outliers Distorted by skewed data
Median	Not influenced by the outliers Not distorted by skewed data	Ignore most of the data
Mode	Easily determined for qualitative data	Ignore most of the data
Geometric mean	Appropriate for right-skewed data	Appropriate if the log transformation produces a symmetrical distribution
Weighted mean	Count relative importance of each observation	Weights must be known or estimated

Measures of spread

Spread related to the central value

The data are more spread as their values are more different by each other

---

Parameters:

1. Range
2. Variance (VAR)
3. Standard deviation (STDEV)
4. Coefficient of variation
5. Standard Error

Measures of spread: range

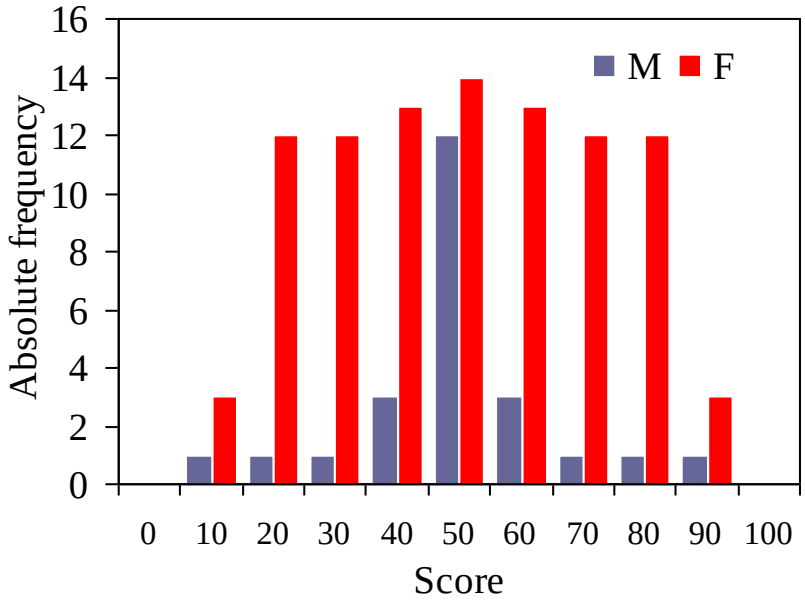
$R = X_{max} - X_{min}$

It tells us nothing about how the data vary around the central value

Outliers significantly affect the value of the range

RANGE (Descriptive Statistics)

- R_M = 90-10 = 80
- R_F = 90-10 = 80
  - Equal values BUT different spreads



Measures of spread: mean of deviation

From the mean:

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

From the Median:

$$R_{Me} = \frac{\sum_{i=1}^{n} |X_i - Me|}{n}$$

StdID	Mark	R_{Mean}	R_{Median}
34501	8	1.20	0.00
27896	3	-3.80	-5.00
32102	4	-2.80	-4.00
32654	8	1.20	0.00
32014	9	2.20	1.00
31023	9	2.20	1.00
30126	5	-1.80	-3.00
34021	9	2.20	1.00
33214	9	2.20	1.00
32016	4	-2.80	-4.00
Mean	6.80		
Median	8.00		

Measures of spread: mean of deviation

We analyze how different are the marks from the mean of ten students by using distances

The deviation is higher as the mark is far away form the mean

To quantify how the distribution is diverted to other distribution we calculate the sum of deviations

The difference from the mean is very close to zero

StdID	Note	R_{Mean}	R_{Median}
34501	8	1.20	0.00
27896	3	-3.80	-5.00
32102	4	-2.80	-4.00
32654	8	1.20	0.00
32014	9	2.20	1.00
31023	9	2.20	1.00
30126	5	-1.80	-3.00
34021	9	2.20	1.00
33214	9	2.20	1.00
32016	4	-2.80	-4.00
Sum		0.00	-12.00

Measures of spread: squared deviation from the mean

The squared deviation from the mean

Thus, the sum of squared deviation from the mean it will be obtained:

$$SS = \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

StdID	Note	R_{Mean}	R_{Mean}^2
34501	8	1.20	1.39
27896	3	-3.80	14.59
32102	4	-2.80	7.95
32654	8	1.20	1.39
32014	9	2.20	4.75
31023	9	2.20	4.75
30126	5	-1.80	3.31
34021	9	2.20	4.75
33214	9	2.20	4.75
32016	4	-2.80	7.95
Sum		**0.00**	**55.60**

Measures of spread: variance

The mean of the sum of squared deviation from the mean is called variance (it is expressed as squared units of measurements of observed data)

Population variance:

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

Sample variance (the sample variance tend to sub estimate the population variance):

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

Measures of spread: variance

Steps:

1. Calculate the mean,
2. Find the difference between data and mean for each subject,
3. Calculate the squared deviation from the mean,
4. Sum the squared deviation from the mean,
5. Divide the sum to n if you work with the entire population or at (n-1) if you work with a sample,
6. $s^2 = 55.60/9 = 6.18$

StdID	Mark	R_{Mean}	R_{Mean}^2
34501	8	1.20	1.39
27896	3	-3.80	14.59
32102	4	-2.80	7.95
32654	8	1.20	1.39
32014	9	2.20	4.75
31023	9	2.20	4.75
30126	5	-1.80	3.31
34021	9	2.20	4.75
33214	9	2.20	4.75
32016	4	-2.80	7.95
Sum		**0.00**	**55.60**

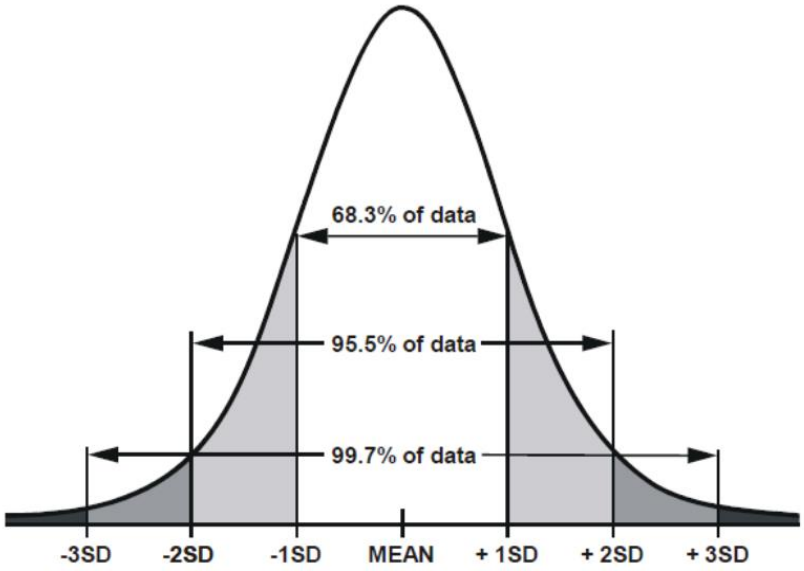
Measures of spread: standard deviation

Has the same unit of measurement as mean and data of the series

It is used in descriptive and inferential statistics

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Measures of spread: standard deviation



if the frequency distribution is symmetrical (bell shape) we have the properties:

mean=median=mode

Interval	% of contained observation
$\overline{X} \pm 1 \cdot s$	68.3
$\overline{X} \pm 2 \cdot s$	95.5
$\overline{X} \pm 3 \cdot s$	99.7

Measures of spread: coefficient of variation (CV)

A relative measure of dispersion.

Independent by the units of measurements.

Formula: $$CV = \frac{s}{\bar{X}} * 100$$

Interpretation:

CV < 10%	Homogenous
10% ≤ CV < 20%	Relative homogenous
20% ≤ CV < 30%	Relative heterogeneous
≥30%	Heterogeneous

Measures of spread: standard error

It is used as a measure of spread usually associated to mean (arithmetic mean).

It is used in computing the confidence levels.

$$ES = \frac{s}{\sqrt{n}}$$

Measures of symmetry

Measures of symmetry: skewness

Indicate for a series of data:

- Deviation from the symmetry
- Direction of the deviation from symmetry (positive / negative)

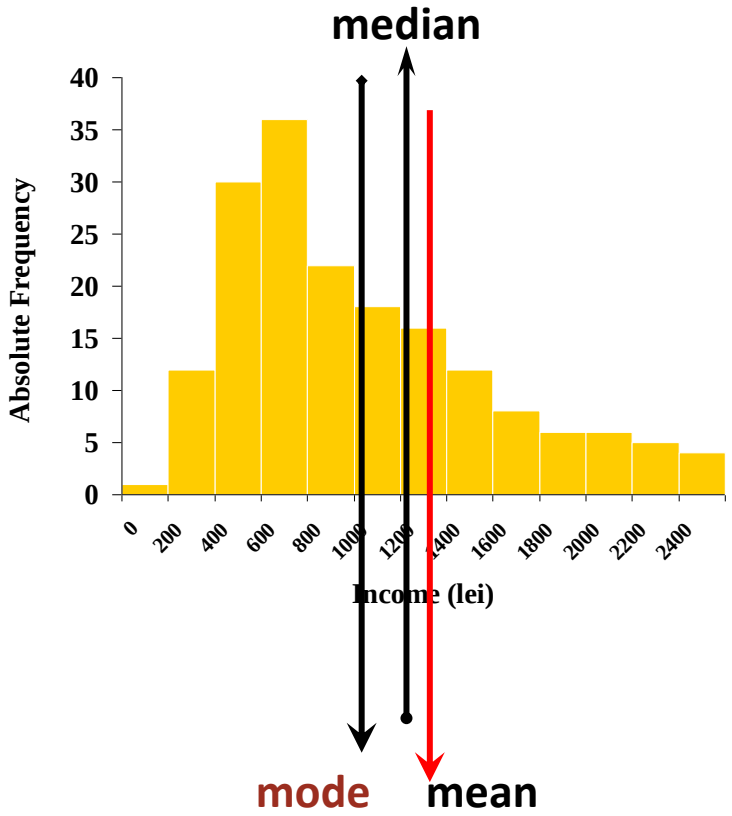
The formula for calculus:

$$M_3 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^3}{n}$$

Measures of symmetry: skewness

Positively skewed:

- **Mode** = 7000 Ron
- **Median** = 8870 Ron
- **Mean** = 9360 Ron

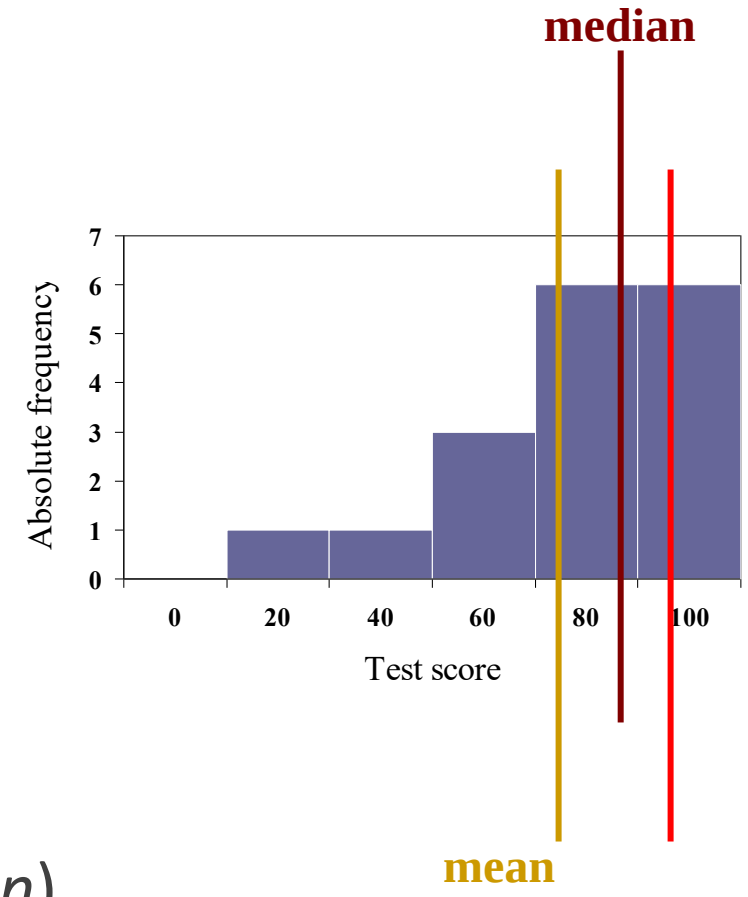


Mode < Median < Mean

Measures of symmetry: skewness

Negatively skewed:

Mode > Median > Mean



= SKEW(number*1*, …, number*n*)

Measures of symmetry: skewness

Interpretation [Bulmer MG, Principles of Statistics, Dover, 1979,] – applied to the population

- If skewness is less than −1 or greater than +1, the distribution is **highly skewed**.
- If skewness is between −1 and −½ or between +½ and +1, the distribution is **moderately skewed**.
- If skewness is between −½ and +½, the distribution is **approximately symmetric**.

Can you conclude anything about the population skewness looking to the skewness of the sample? → Inferential statistics

Measures of symmetry: kurtosis

A measure of the shape of a series relative to the Gaussian shape

$$\alpha_4 = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^4}{S^4} - 3$$

= KURT(number*1*, … , number*n*)

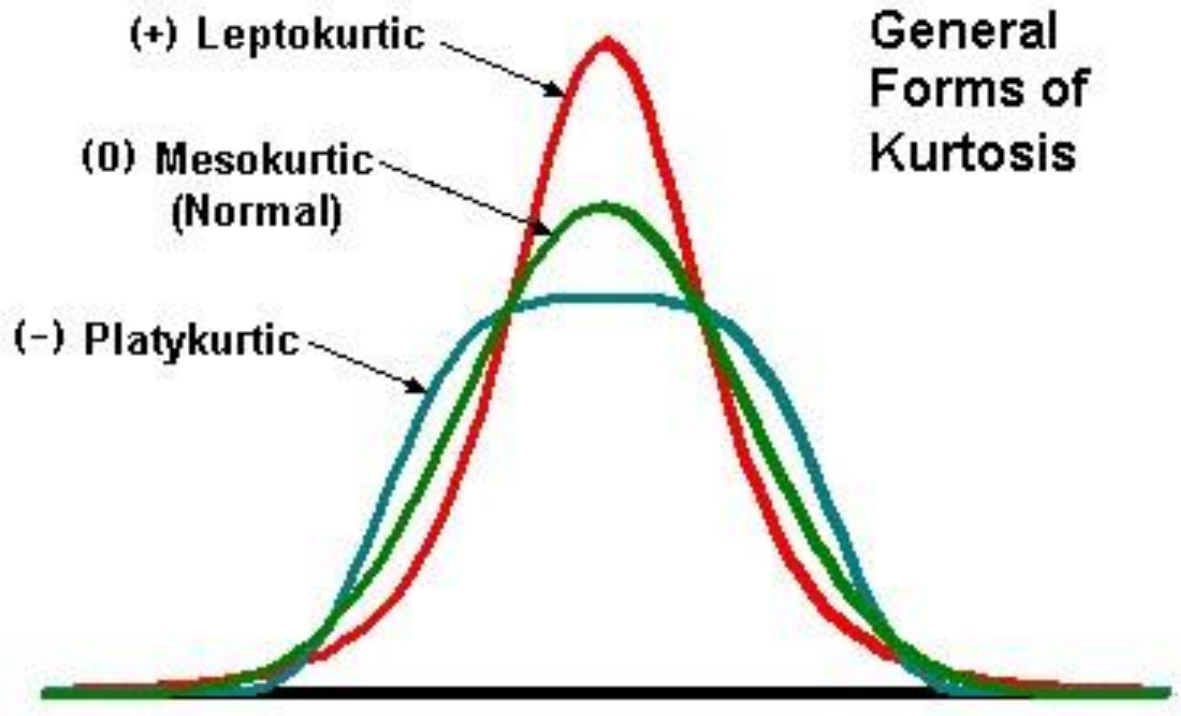
Measures of symmetry: kurtosis

The reference standard is a normal distribution, which has a kurtosis of 3.

Excess kurtosis (kurtosis in Excel) = kurtosis – 3

- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0), Any distribution with kurtosis ≅3 (excess ≅0) is called **mesokurtic**.
- A distribution with kurtosis <3 (excess kurtosis <0) is called **platykurtic**, Compared to a normal distribution, its central peak is lower and broader, and its tails are shorter and thinner.
- A distribution with kurtosis >3 (excess kurtosis >0) is called **leptokurtic**, Compared to a normal distribution, its central peak is higher and sharper, and its tails are longer and fatter.

Measures of symmetry: kurtosis



Excel function for quartile:
QUARTILE

Measures of localization

QUARTILE

PERCENTILE

DECILES

Measures of localization: quartiles – deciles

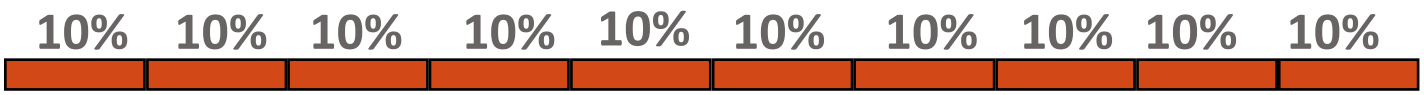
Quartiles:

- Split the series into 4 equal parts:

(minimum) **(median)** **(maximum)**

Decile:

- Split the series into 10 equal parts:



Percentile:

- Split the series into 100 equal parts

Quartiles & symmetry of a distribution

The symmetry of distribution could be analyzed using quartiles.

Let Q_1, Q_2, and Q_3 be 1^{st} (1/3), 2^{nd} (1/2) and 3^{rd} (3/4) quartiles:

- $Q_2-Q_1 \approx Q_3-Q_2$ (≈ almost equal) → the distribution is almost symmetrical
- $Q_2-Q_1 \neq Q_3-Q_2$ → the distribution is asymmetrical (through left or right)

Measures of localization: quartiles

2.80	2.97	3.05	3.25	3.40	3.45	3.80	4.10	4.30	4.40
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}

$Q_1 = 3.03$

$Q_2 = 3.43$

$Q_3 = 4.15$

$Q_2 - Q_1 = 3.43 - 3.03 = 0.40$

$Q_3 - Q_2 = 4.15 - 3.43 = 0.72$

How do you interpret this result???

Measures of centrality: type of variables

	Qualitative nominal	Qualitative ordinal	Quantitative
Mode	Yes	Yes (NOT recommended)	Yes (NOT recommended at all)
Median	No	Yes	Yes
Mean	No	No	Yes (if data is symmetric and unimodal)

Measures of centrality: type of variables

	Qualitative nominal	Qualitative ordinal	Quantitative
Range	No	Yes (NOT the best method)	Yes (NOT the best method)
Standard deviation	No	No	Yes (if data is symmetric and unimodal)

Units of measurements: importance

If to each data from a series add or subtract a constant:

- The mean will increase or decrease with the value of the added constant
- The standard deviation will NOT be changed

If each data from a series is multiplied or divided with a constant:

- The mean will be multiplied or divided with the value of the constant
- The standard deviation will be multiplied or dividid with the value of the constant

RECALL!

The units of measurements have an influence on statistical parameters.

Statistical parameters should be applied according to the type of data.

Mean, Standard deviation, and Range are sensitive to outliers.

When we use a summary statistic to describe a data set, we lose a lot of the information contained in the data set.

HOMEWORK (optional) for the theoretical

The total number of points that can be obtained is 15.

It is necessary to obtain at least 12 points in order for the topic to be considered.

October 30 - Novembre 12

The results will be communicated to you after 20 Nov (in Teams).

The results will be (yes / no). Those who will have "yes" to the final grade (which will be calculated after the practical exam (30%) and the theoretical exam (70%)) will receive extra 0.2 points for HOMEWORK 1.

For those who have "no", the final grade will be **unchanged**


Thank You
shift