DESCRIPTIVE STATISTICS

Sorana D. BOLBOACĂ

OUTLINE

- Measures of Centrality
- Measures of Spread
- Measures of Localization
- Measures of Symmetry

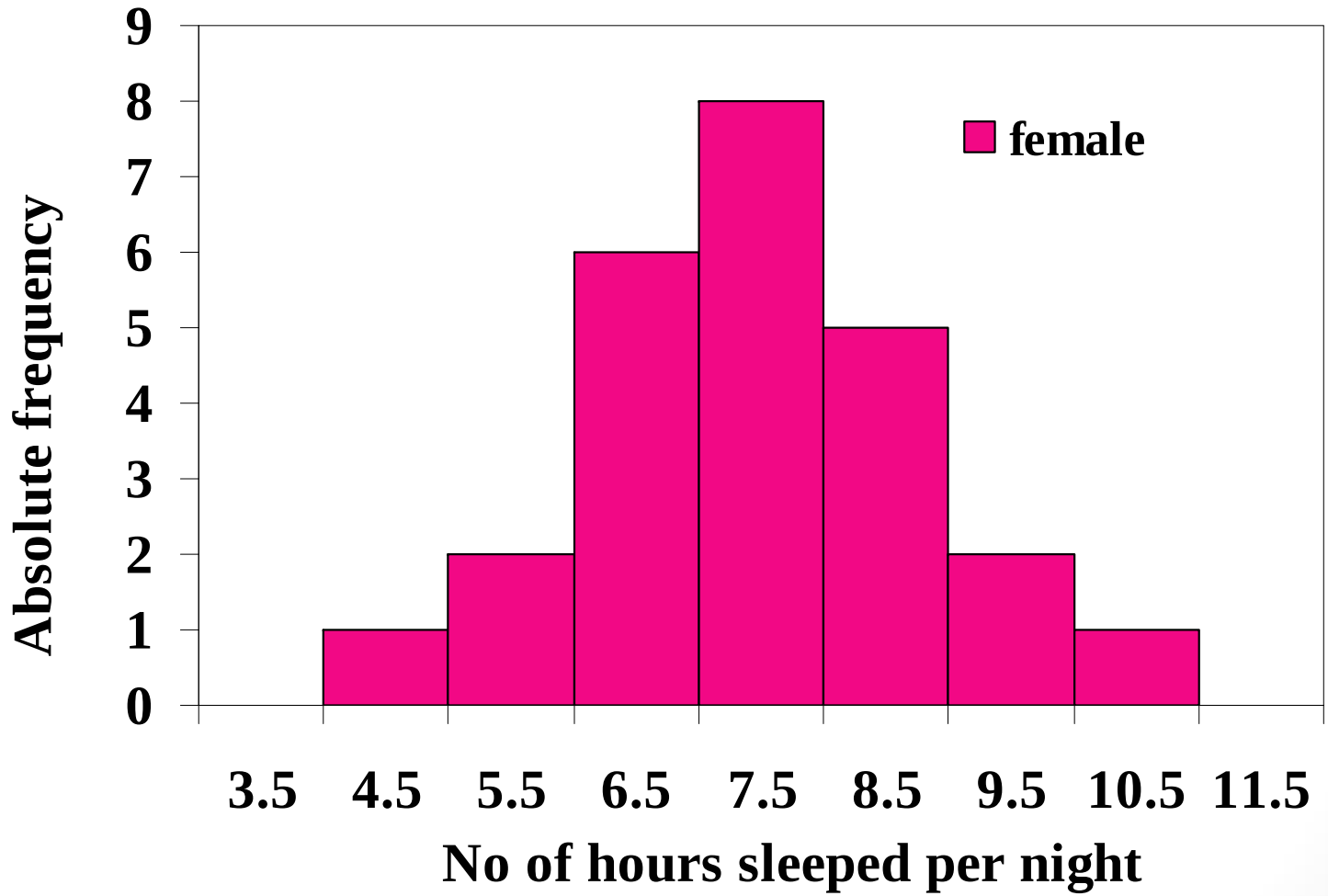
DESCRIPTIVE STATISTICS PARAMETERS

Measures of Centrality	Measures of Spread
✓ Mean ✓ Mediana ✓ Mode	✓ Range ✓ Variance ✓ Standard deviation ✓ Coefficient of variance ✓ Standard error
Measures of Symmetry	**Measures of Localization**
✓ Skewness ✓ Kurtosis	✓ Quartile ✓ Percentiles

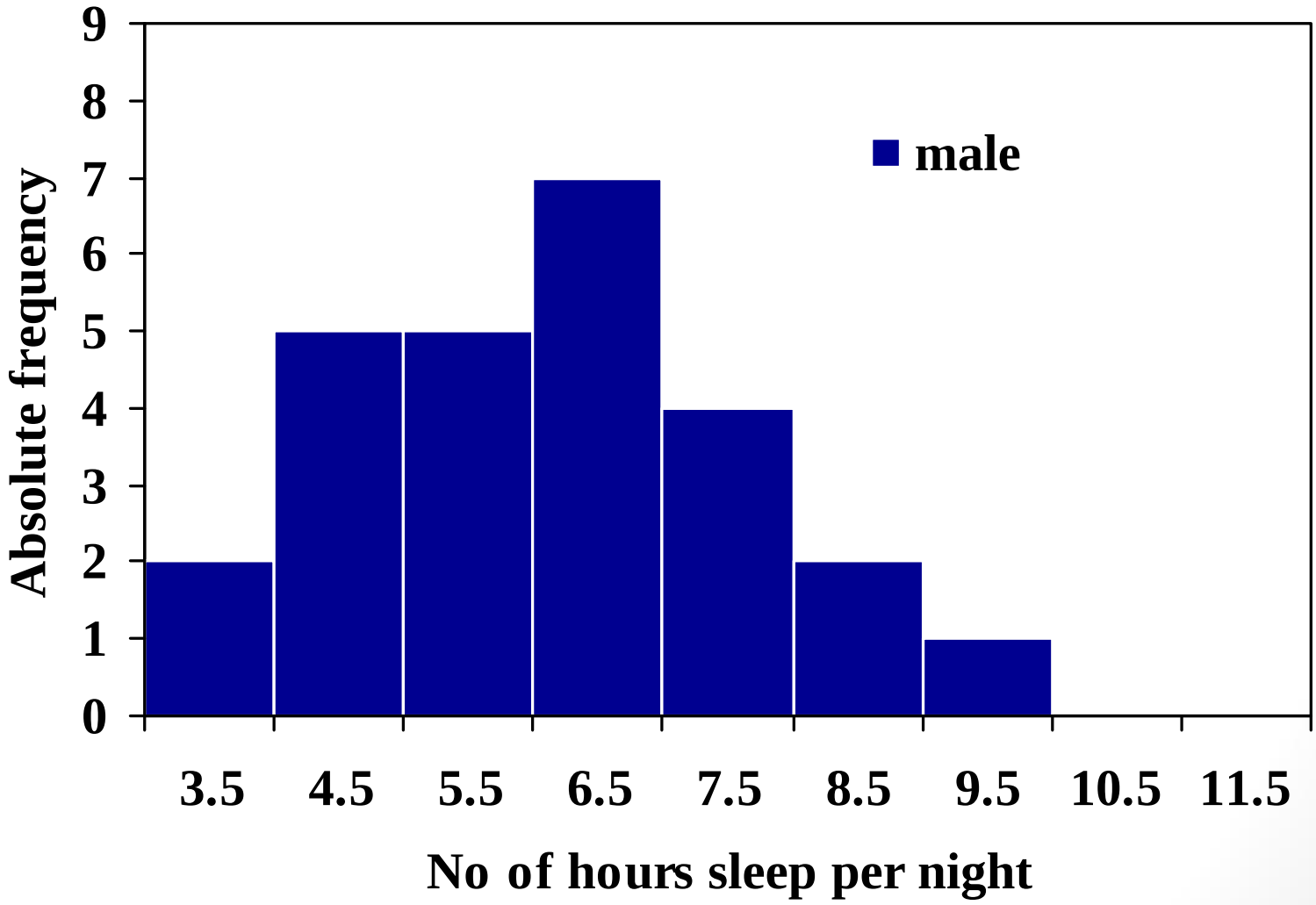
MEASURES OF CENTRALITY

- Simple values that give us information about the distribution of data
- Parameters:
  - Mode
  - Median
  - Mean

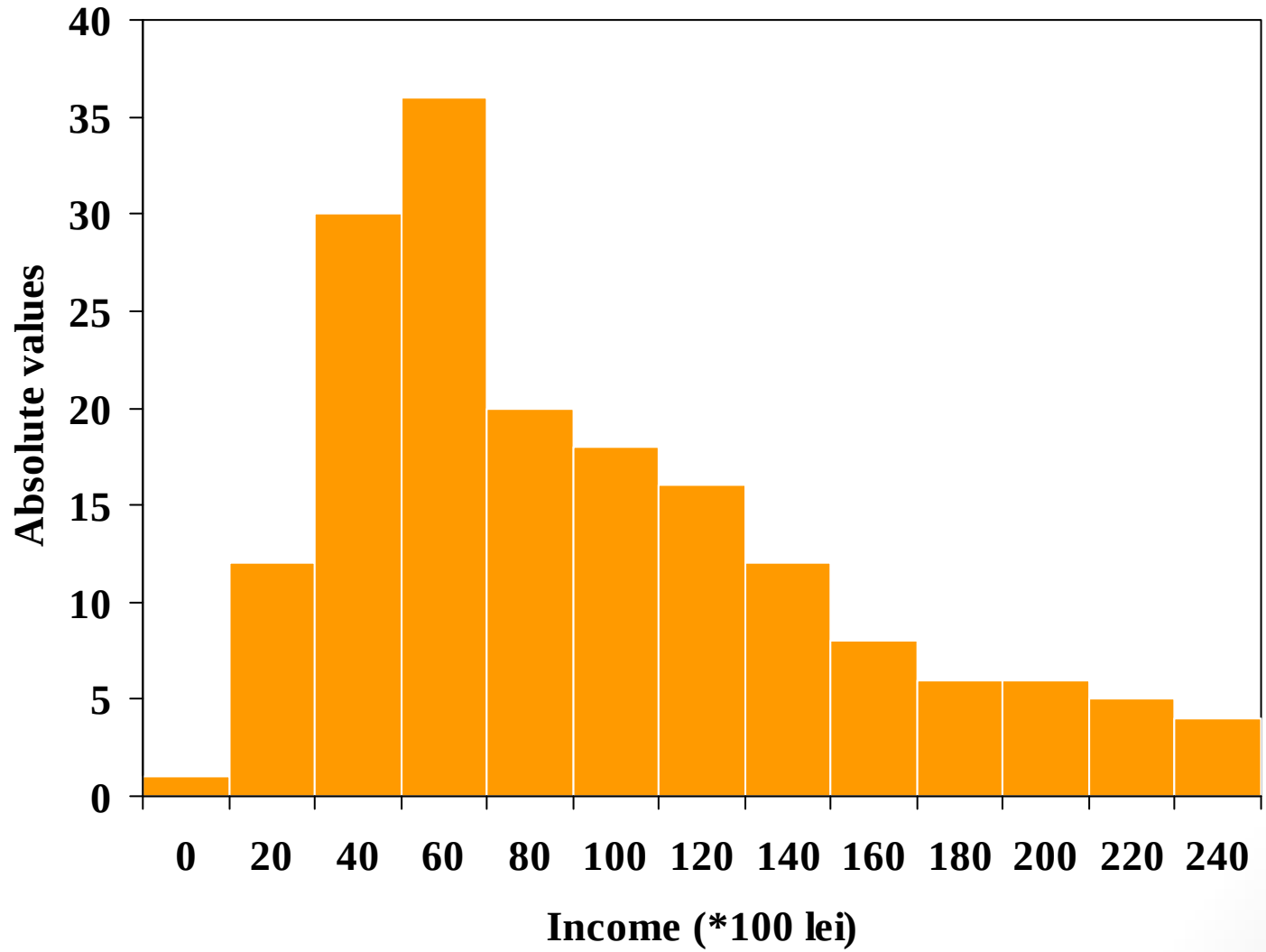
MEASURES OF CENTRALITY



MEASURES OF CENTRALITY



MEASURES OF CENTRALITY



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
- Excel: MODE(number1.number2. …. number)

MEASURES OF CENTRALITY: MODE

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

2	1	2	1	1

- 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

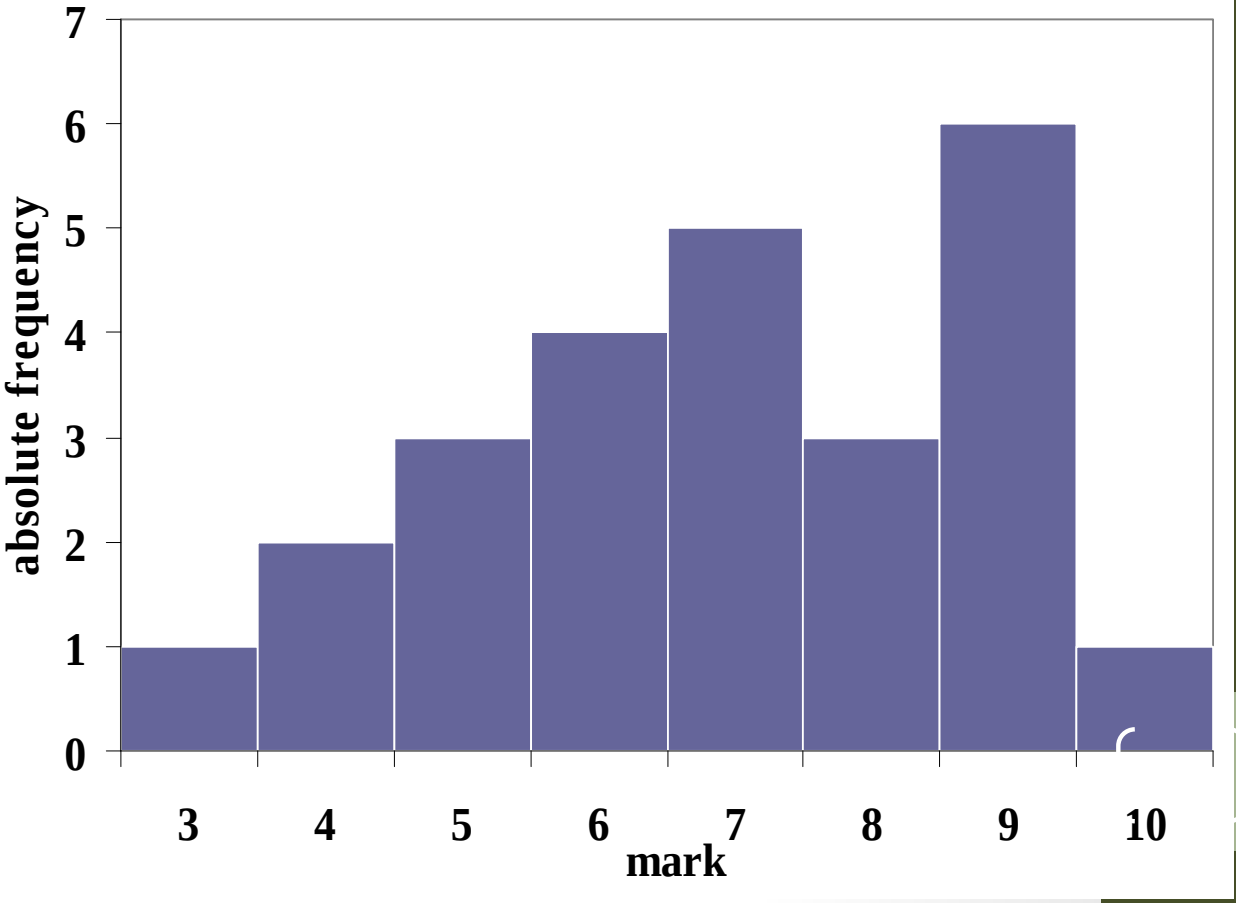
- It is NOT influenced by extreme values

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



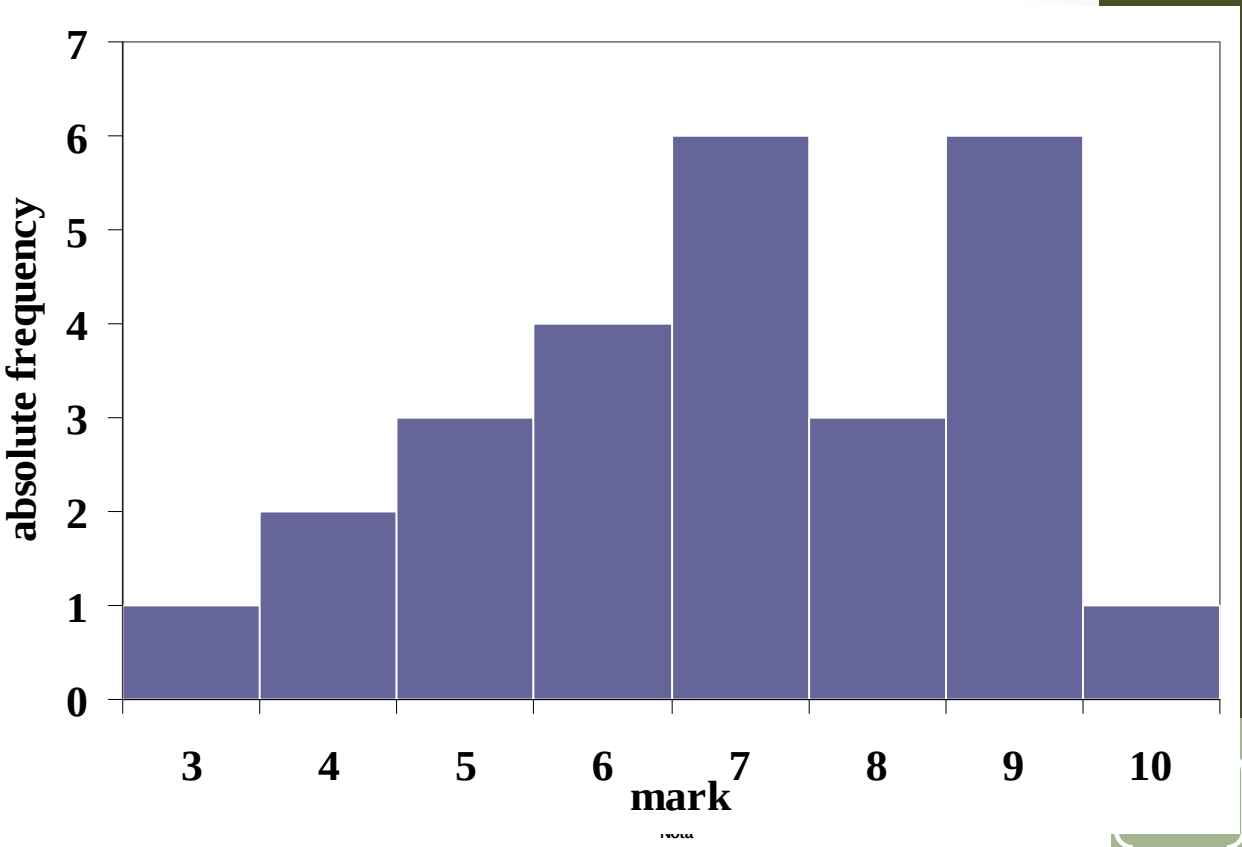
MEASURES OF CENTRALITY: MODE

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

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

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

- If 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 grouped 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)
- $Me = (X_{13}+X_{14})/2 = (7+7)/2 = 7$
- **Excel**: = MEDIAN(number1.number2.....number26)

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):

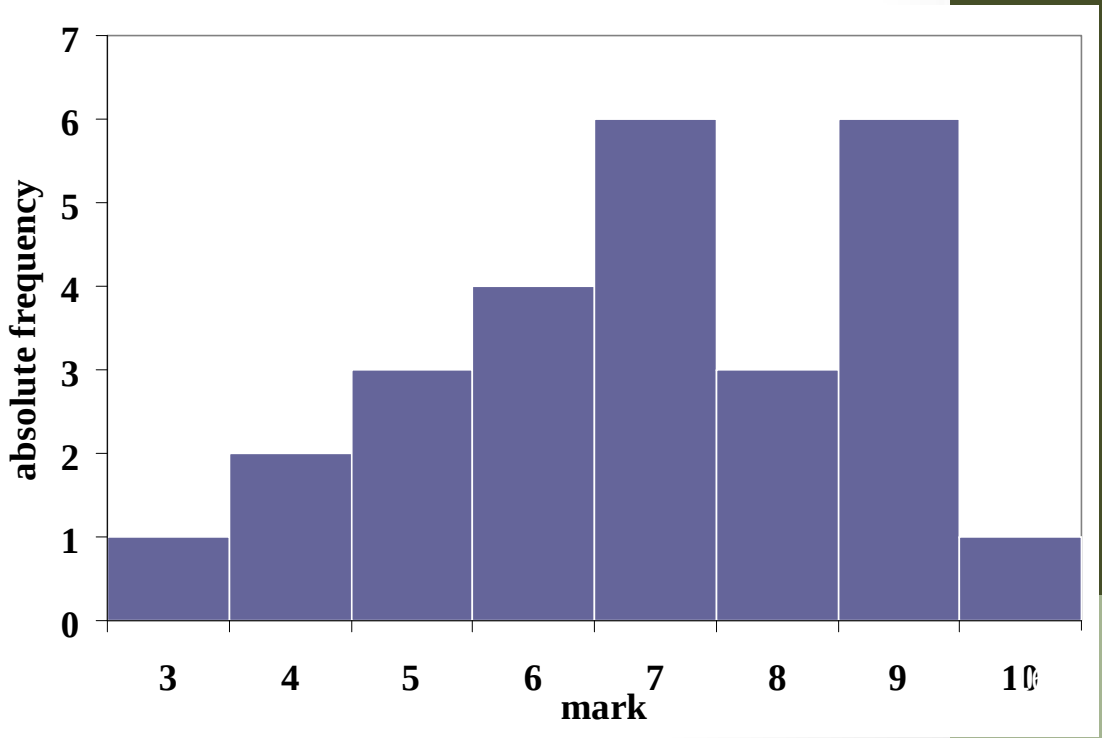
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

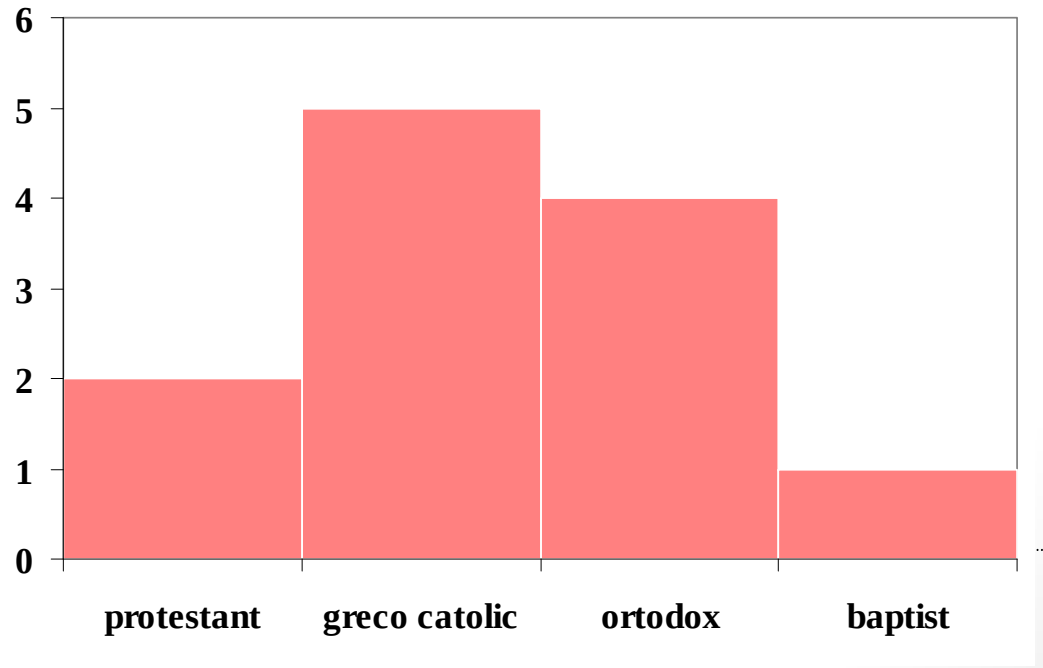
- **Excel**:
- =AVERAGE (number1..... number26)



MEASURES OF CENTRALITY: MEAN

- Is the preferred measure of centrality both as a parameter for describing data and as estimator.

- It has significance just IF the variable of interest is on interval scale.



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. 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 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. 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 id multiply 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 choose to be equal and positive we will obtain the arithmetic mean.

MEASURES OF CENTRALITY: OTHER

- Quadratic mean (root mean square. abbreviated RMS): measure the magnitude of a varying quantity

$$RMN = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

- Central value:

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

MEASURES OF CENTRALITY: TYPE OF VARIABLES

	Nominal	Ordinal	Metric
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 SPREAD

- Spread related to the central value
- The data are more spread as their values are more different by each other

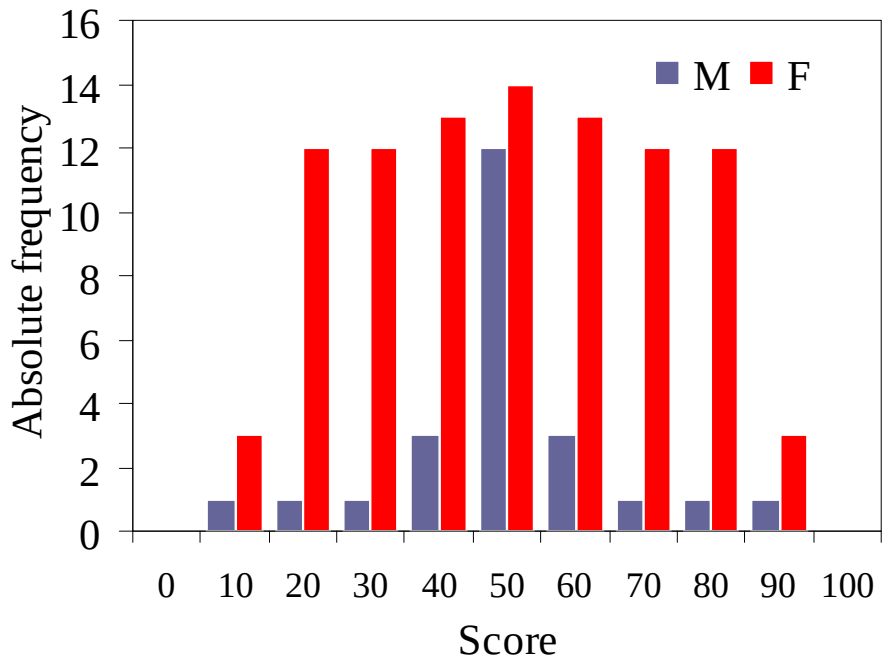
Parameters:

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

MEASURES OF SPREAD

- $\mathbf{R = X_{max} - X_{min}}$
- It tells us nothing about how the data vary around the central value
- Outliers significantly affect the value of range
- Excel: RANGE (Descriptive Statistics)

- $R_M = 90\text{-}10 = 80$
- $R_F = 90\text{-}10 = 80$
  - Equal values – 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 analyse how different are the marks from the mean of ten students by using distances
- The deviation is greater as the mark is further 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 obtain:

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

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 sum of squared deviation form the mean is called variance (it is expressed as squared units of measurements of observed data)
- Population variance:

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

- 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

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

MEASURES OF SPREAD: STANDARD DEVIATION

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

MEASURES OF SPREAD: COEFICIENT OF VARIATION

- Relative measure of dispersion
- Calculus formula: $CV = \frac{s}{\bar{X}}$
- Evaluation of standard deviation reported to mean
- Has the advantage of being a parameter independent by the units of measurements

MEASURES OF SPREAD: COEFICIENT OF VARIATION

- Interpretation of Homogeneity:

Coefficient of Variation (CV)	Interpretation: The population could be considered
CV < 10%	Homogenous
10% ≤ CV < 20%	Relative homogenous
20% ≤ CV < 30%	Relative heterogeneous
> 30%	Heterogeneous

MEASURES OF SPREAD: STANDARD ERROR

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

- It is used in computing the confidence levels

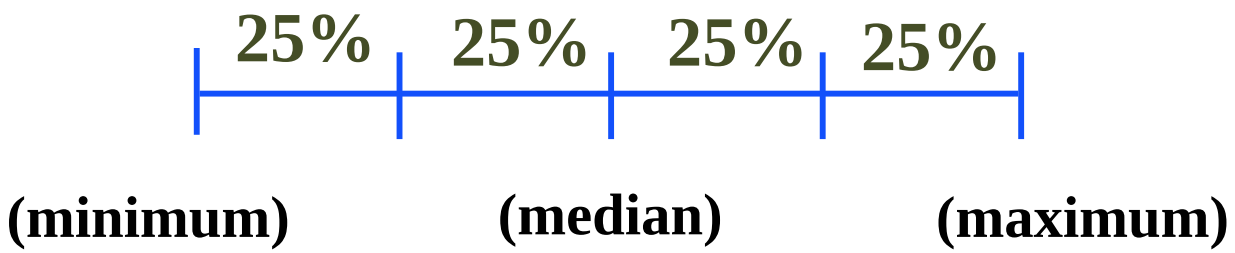
MEASURES OF LOCALIZATION

- Quartile
- Percentile
- Deciles

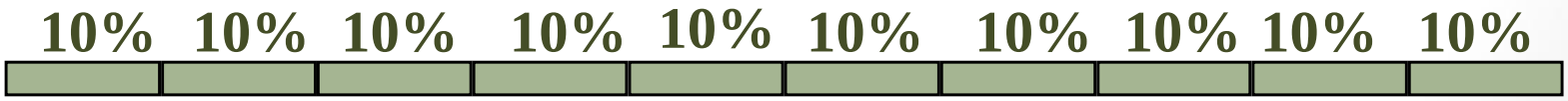
- Excel function for quartile:
- **QUARTILE**

MEASURES OF LOCALIZATION: QUARTILES – DECILES

- **Quatiles**:
  - Split the series in 4 equal parts:



- **Decile**:
  - Split the series in 10 equal parts:



MEASURES OF LOCALIZATION: PERCENTILE

- Percentile: Split the series in 100 equal parts
- The symmetry of a 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:
  - $Q_2-Q_1 \approx Q_3-Q_2$ ($\approx$ almost equal) $\rightarrow$ the distribution is almost symmetrical
  - $Q_2-Q_1 \neq Q_3-Q_2 \rightarrow$ 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 SYMMETRY: SKEWNESS

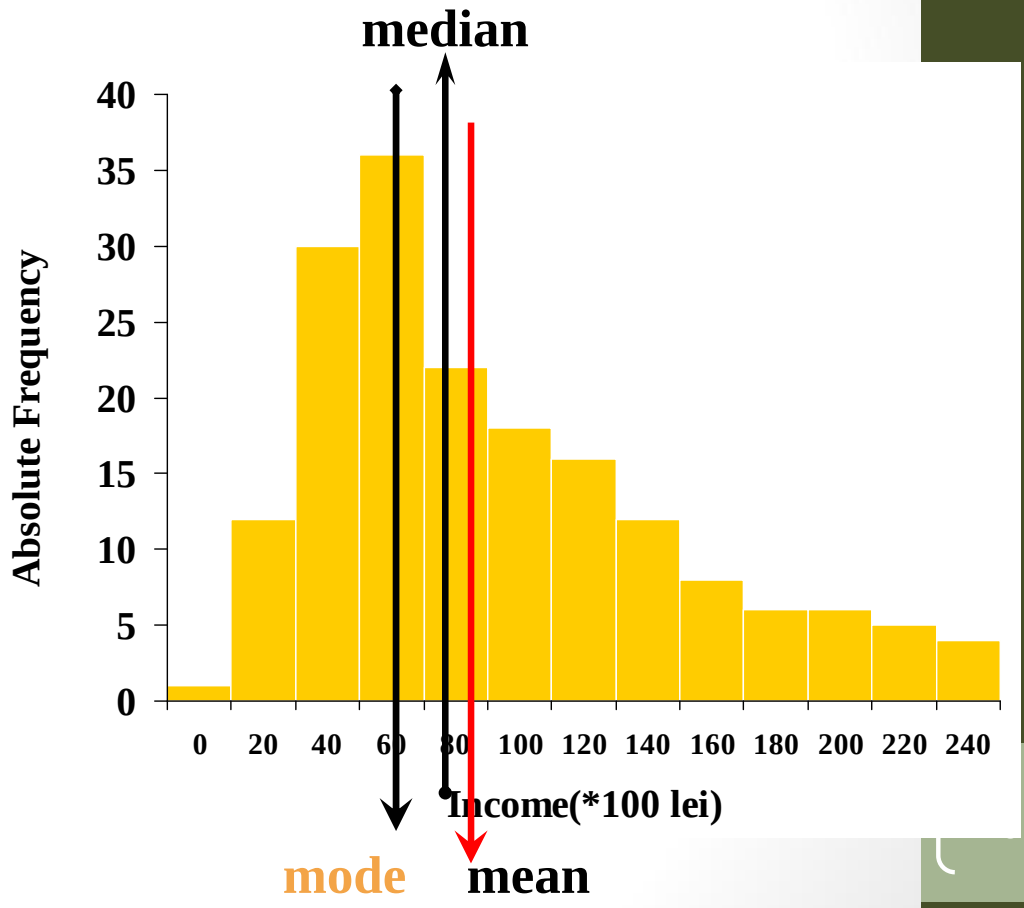
- Indicate for a series of data:
  - Deviation from the symmetry
  - Direction of the deviation from symmetry (positive / negative)
- Formula for calculus:

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

MEASURES OF SYMMETRY: SKEWNESS

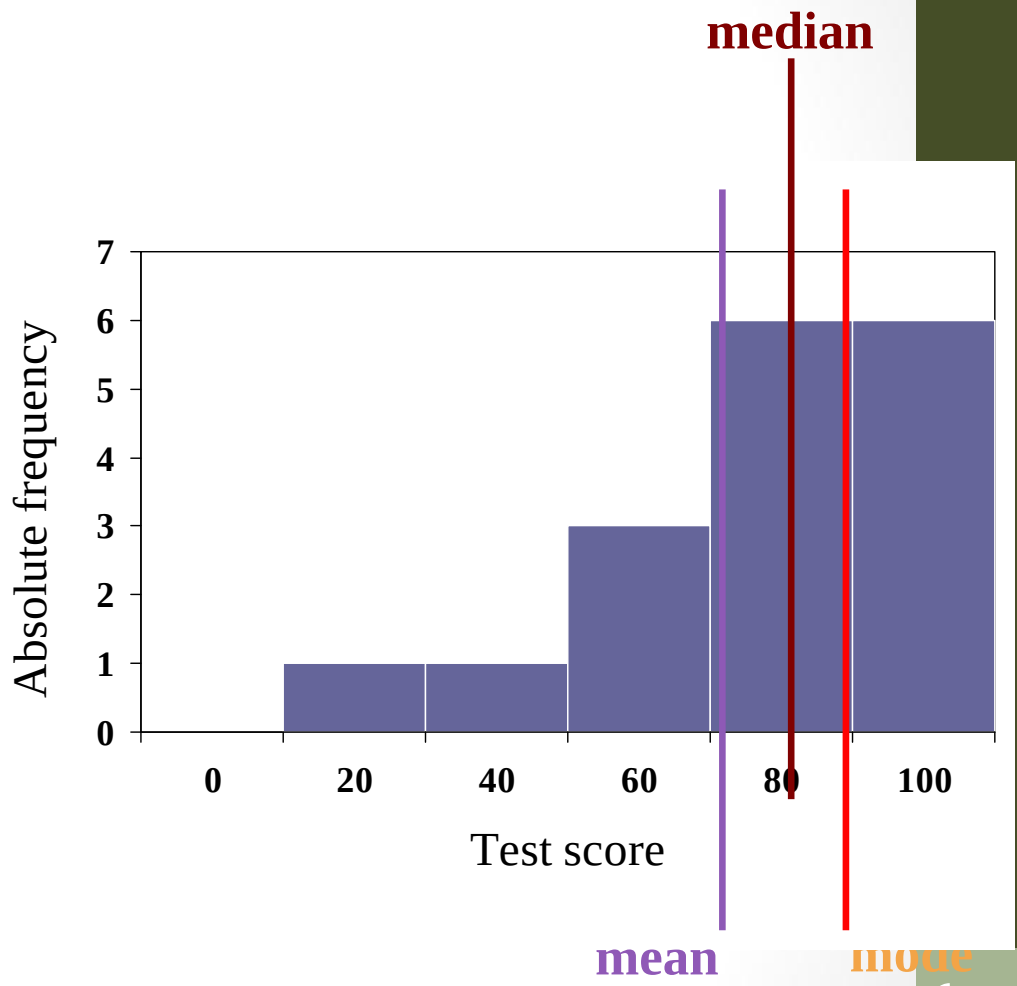
- Left asymmetry / positive:
  - **Mode** = 7000 Ron
  - **Median** = 8870 Ron
  - **Mean** = 9360 Ron

- **Mode < Median < Mean**



MEASURES OF SYMMETRY: SKEWNESS

- Right asymmetry / negative:
- **Mode > Median > Mean**
- **Excel:**
- = SKEW(number*1*. .... number*n*)



MEASURES OF SYMMETRY: SKEWNESS

- Interpretation [Bulmer MG. Principles of Statistics. Dover, 1979.] – applied to 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 Gaussian shape

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

- Excel:

= 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 SPREAD

	Range	Standard deviation
Nominal	No	No
Ordinal	Yes (NOT the best method)	No
Metric	Yes (NOT the best method)	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 multiply or divide with a constant:
  - The mean will be multiply or divide with the value of the constant
  - The standard deviation will be multiply or divide with the value of the constant

REMEMBER!

- The units of measurements have influence on statistical parameters.
- Statistical parameters should be applied according to the type of data.
- Sensitive to outliers: Mean. Standard deviation. Range.
- When we use a summary statistic to describe a data set we lose a lot of the information contained in the data set.
- It is important that we do not use summary measures to obscure vital characteristics of a data set.

TASK

- Using reading material about how journals recommend to report statistical results summarize which type of measures of centrality, spread, localization and symmetry fit according with type of variables.