DESCRIPTIVE STATISTICS II

Sorana D. Bolboacă

OBJECTIVES

- Measures of spread: variance, standard deviation, coefficient of variatio, standard error, amplitude
- Measures of symmetry: skewness and kurtosis

DESCRIPTIVE STATISTICS PARAMETERS

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

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

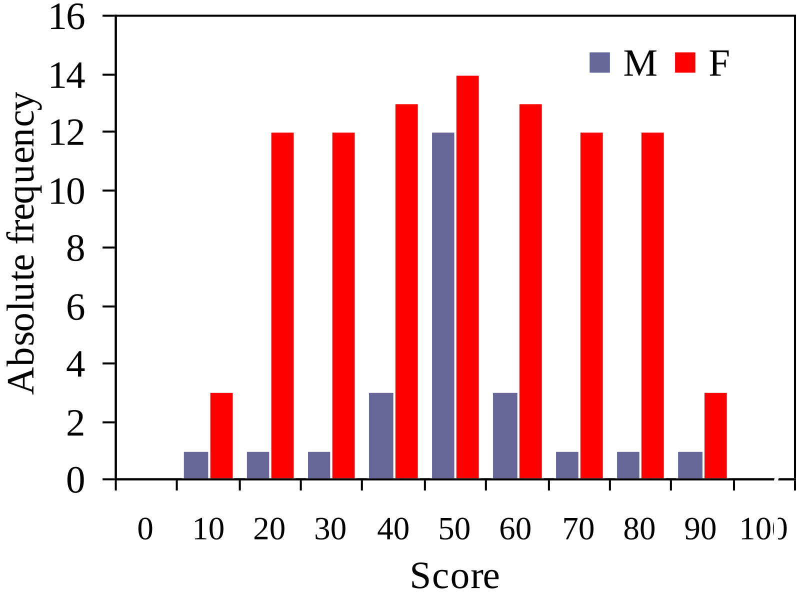
MEASURES OF SPREAD

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

- 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 = 90-10 = 80
- R_F = 90-10 = 80
  - Equal values – different spreads



MEASURES OF SPREAD: MEAN OF DEVIATION

- From the mean:

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

- 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:

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

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

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

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
$\overline{X} \pm 1 \cdot s$	68.3
$\overline{X} \pm 2 \cdot s$	95.5
$\overline{X} \pm 3 \cdot s$	99.7

Measures of Spread: Coeficient of Variation

- Relative measure of dispersion
- Calculus formula: $$CV = \frac{s}{\overline{X}}$$
- 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

- It is used in computing the confidence levels

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

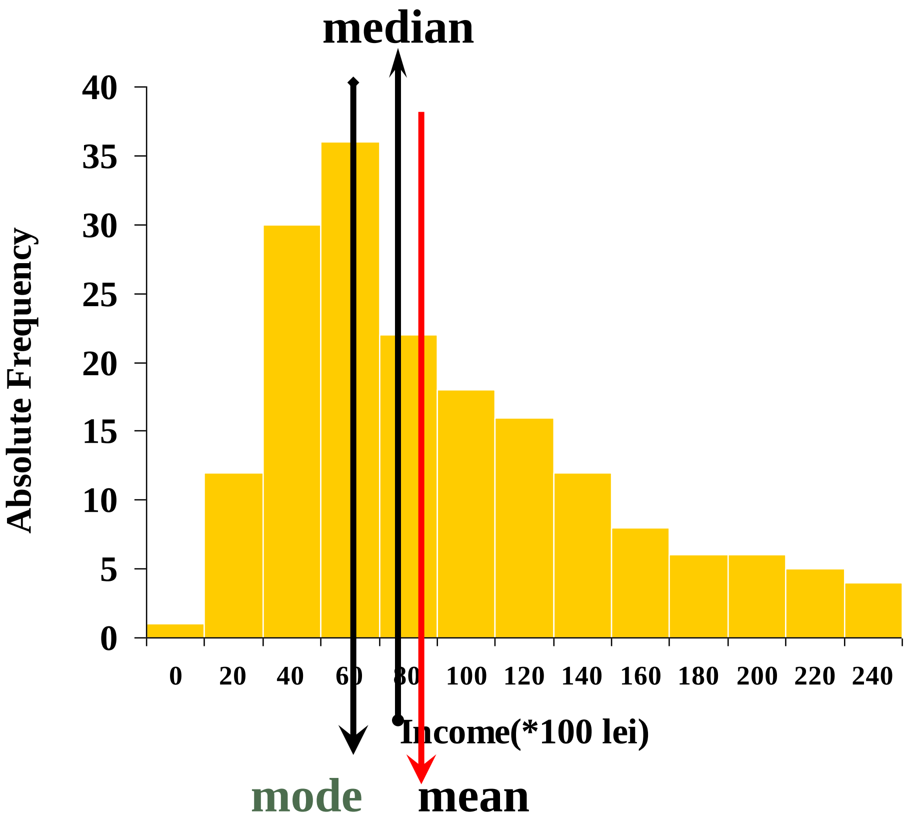
MEASURES OF SYMMETRY: SKEWNESS

- 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

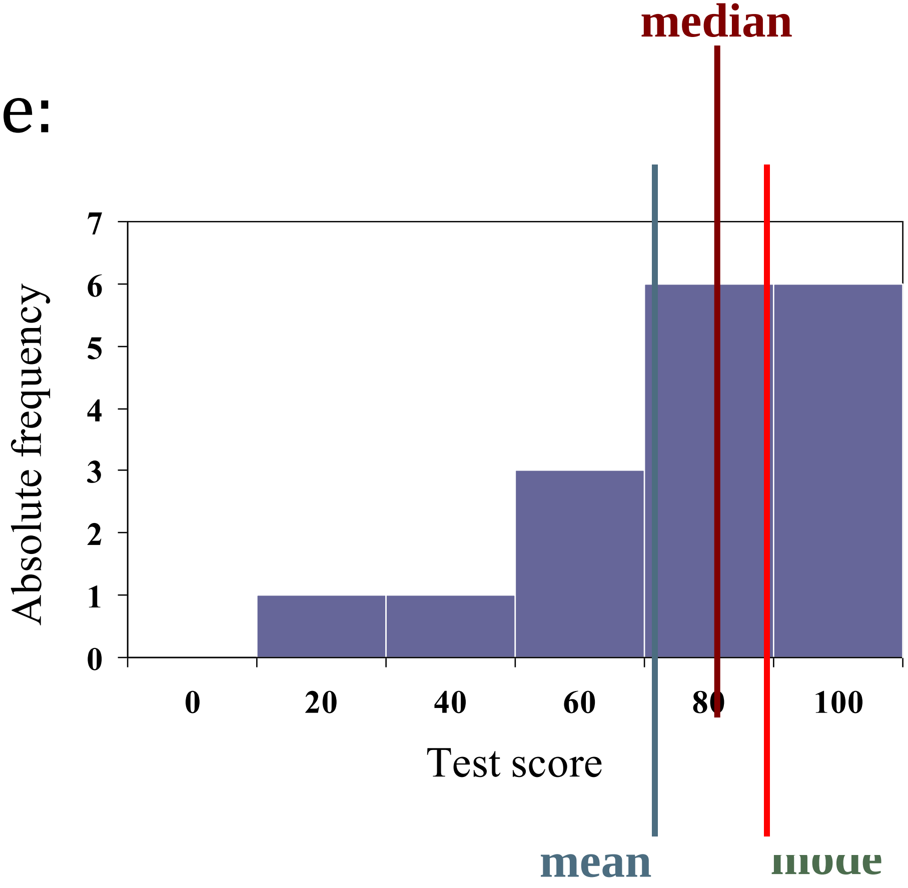
- Left asymmetry / positive:
  - □ **Mode** = 7000 Ron
  - □ **Median** = 8870 Ron
  - □ **Mean** = 9360 Ron



- **Mode < Median < Mean**

MEASURES OF SYMMETRY: SKEWNESS

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



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.