#### Medical Statistics II. Sample Characteristics

Assoc. Prof. Katarína Kozlíková, RN., PhD. IMPhBPhITM FM CU in Bratislava katarina.kozlikova@fmed.uniba.sk

#### Content

- Introduction
- Sample Characteristics
- Measures of Location
  - Averages
  - Structural Characteristics
- Measures of Variability
  - Connected with Averages
  - Connected with Structural Characteristics
- Measures of Shape
  - Skewness
    - Excess Kurtosis
  - Literature

líková, 2013

2013/14 Medical Statistics II - Sample characteristcs

#### Introduction

- The aim of this lecture is to familiarize students with the sample characteristics that are most commonly used in the evaluation of medical experiments and occur in medical scientific and professional literature
- It follows up the lecture Medical Statistics I Basic Terminology and uses the terminology explained there
- To practice the lectured topics and for further study is recommended to use the literature from the list of literature

#### **Sample Characteristics**

- A descriptive characteristics
  - A number calculated from a statistical set in a well defined way (using a formula)
- The aim of descriptive characteristics
  - Numbers obtained in this way are used to characterise a set consisting of many numbers by only a few numbers
- Descriptive statistics
  - A part of statistics that deals with descriptive characteristics
- Sample characteristics
  - Descriptive characteristics applied to a sample
- There exist three groups of descriptive statistics
   Measures of location
   Measures of variability
   Measures of shape

## Measures of Location

#### Measures of Location -Introduction

- In the most common case, the obtained data are organised as a set of numbers
- Quantitative data are usually clustered around a central tendency
  - They are organised according some (statistically defined) distribution
- A central tendency
  - A single number intended to typify the numbers in the set
- Measures of location are used for estimation of location of this distribution, or its "middle", that is, where on the (numerical) axis – on the scale is the position of the central tendency

#### Measures of Location – Central Tendency (1)

- A central tendency can be estimated in following way
  - If all the numbers in the list are the same, then this number should be used
  - If the numbers are not the same, the average is calculated by combining the numbers from the list in a specific way and computing a single number as being the average of the list
    - The method of calculation depends on the distribution of data in the set
    - In this lecture, we assume a normal distribution of data generally
- A single number is then used for consecutive calculations

#### Measures of Location – Central Tendency (2)

- A central tendency is a measure of the "middle" or "typical" value of a data set
- Several types of averages are used according to what kind of data are represented by the numbers
  - Means (averages)
  - Median
  - Mode

Kozlíková, 2013

Quantiles

#### Measures of Location - Averages

- Take into account the values of all measurements of the quantity
- Used for normally distributed quantitative data
- The most used means

Kozlíková, 2013

- Arithmetic mean (weighted)
- Geometric mean (weighted)
- Harmonic mean (weighted)
  - Weighted means are used for very large samples, for sorted data (data in classes), for data of different "weights"

## Arithmetic Mean (Mean)

- The best estimate of the measured quantity *X* from *n* measurements
  - Calculated as the sum of all measured values and then divided by the number of the measurements

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n} \quad \text{or another notation} \quad \overline{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$$

Used nomenclature:

Kolíková, 2013

- $\overline{X}$ : average (arithmetic mean)
- $x_i$ :  $i^{th}$  measurement of the quantity X
- i : measurement number (i = 1, 2, ..., n)

n : sample size  $\Sigma$  : summation sign

#### Arithmetic Mean - Properties

- The sum of differences of all values from the arithmetic mean equals zero
  - Suitable to check the calculation

líková, 2013

- The sum of squared differences of all values from the arithmetic mean is less then the sum of squared differences of all values from any other value
  - Used for calculations of variance and standard deviation
- An arithmetic mean of arithmetic means of subsets of a set is not the arithmetic mean of the whole set

In such case, the <u>weighted</u> arithmetic mean has to be used

## Weighted Arithmetic Mean (Weighted Mean)

 If different "weights" (frequencies) of input data have to be considered

$$\overline{X} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + \dots + n_k \cdot x_k}{n_1 + n_2 + \dots + n_k}$$

where

$$n_1 + n_2 + \ldots + n_k = n$$

or another notation

or another notation

$$\overline{\mathbf{x}} = \frac{\sum_{j=1}^{k} (n_j \cdot \mathbf{x}_j)}{\sum_{j=1}^{k} n_j}$$
$$\sum_{j=1}^{k} n_j = n$$

Used nomenclature:

Kozlíková, 2013

- $\overline{x}$ : mean (arithmetic, weighted)
- $x_j$ : mid-point of the  $j^{\text{th}}$  class
- $n_j$ : frequency ("weight") of the j<sup>th</sup> class
- j : class number (j = 1, 2, ..., k)

- *n* : sample size
- k : number of classes
- $\boldsymbol{\Sigma}$  : summation sign

C

#### **Geometric Mean**

- Suitable for data that can be expressed using coefficients (for example, mean increase, percentage, area, volume – positive values of the measured quantity having an absolute zero)
  - Obtained by multiplying *n* positive numbers and then taking the *n*<sup>th</sup> root

$$\overline{X}_G = \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdot \ldots \cdot X_n}$$

or another not

or another notation 
$$\overline{X}_{G} = \left(\prod_{i=1}^{n} X_{i}\right)^{\frac{1}{n}}$$
  
 $\overline{X}_{G} = Z^{\frac{1}{n} \sum_{i=1}^{n} \log_{Z}(x_{i})}$ 

or a notation using logarithms

#### Used nomenclature:

- $\overline{x}_{c}$ : geometric mean
- $X_i$ : i<sup>th</sup> measurement of the quantity X
  - *i* : measurement number (i = 1, 2, ..., n)
  - n : sample size

- $\Pi$  : multiplication sign
- $\Sigma$  : summation sign
- z : basis of logarithm

#### **Comparison of the Arithmetic Mean and the Geometric Mean**



Figure 1: An example of an arithmetic mean (left) and a geometric mean (right) calculated from the same data (4; 18; 3).

Kozlíková, 2013

2013/14 Medical Statistics II - Sample characteristcs

#### Weighted Geometric Mean

If different "weights" (frequencies) of input data have to be considered

$$\overline{X}_{G} = \sqrt[n]{X_{1}^{n_{1}} \cdot X_{2}^{n_{2}} \cdot \ldots \cdot X_{k}^{n_{k}}} \text{ or another notation} } \overline{X}_{G} = \sqrt[n]{\prod_{j=1}^{k} X_{j}^{n_{j}}}$$
where
$$n_{1} + n_{2} + \ldots + n_{k} = n \qquad \text{or another notation} \qquad \sum_{j=1}^{k} n_{j} = n$$

$$\underbrace{\text{Used nomenclature:}}_{\overline{X}_{G}: \text{ geometric mean}} \qquad n: \text{ sample size} \\ x_{j}: \text{ mid-point of the } j^{\text{th}} \text{ class} \\ n_{j}: \text{ frequency } (,, \text{weight"}) \text{ of the } i^{\text{th}} \text{ class} \\ j: \text{ class number } (j = 1, 2, ..., k)$$

$$\underbrace{\text{We fixed} 2013} \qquad 2013/14 \text{ Medical Statistics II - Sample characteristes} \qquad 1$$

#### Harmonic Mean

- Appropriate for situations when the average of rates or ratios is desired (for example, given effort at open time consumption, parallel connection of electrical resistances)
  - Calculated as the reciprocal of the arithmetic mean of the reciprocals of all measured values



### Weighted Harmonic Mean

 If different "weights" (frequencies) of input data have to be considered

$$\overline{X}_{H} = \frac{n_{1} + n_{2} + n_{3} + \ldots + n_{k}}{\frac{n_{1}}{x_{1}} + \frac{n_{2}}{x_{2}} + \frac{n_{3}}{x_{3}} + \ldots + \frac{n_{k}}{x_{k}}} \quad \text{or another notation} \quad \overline{X}_{H} = \frac{\sum_{j=1}^{k} n_{j}}{\sum_{j=1}^{k} \frac{n_{j}}{x_{j}}}$$
where
$$n_{1} + n_{2} + \ldots + n_{k} = n \quad \text{or another notation} \quad \sum_{j=1}^{k} n_{j} = n$$

$$\frac{\text{Used nomenclature:}}{\overline{x}_{H}: \text{ harmonic mean}} \quad n: \text{ sample size}$$

 $x_j$  : mid-point of the *j*<sup>th</sup> class

Kozlíková, 2013

- $n_j$ : frequency ("weight") of the  $j^{th}$  class
- j : class number (j = 1, 2, ..., k)

- *k* : number of classes
- $\boldsymbol{\Sigma}$  : summation sign

## Measures of Location – Structural Characteristics

- They are based on distribution of elements in an ordered set, that is, they are based on the ranks of the measurements, and not on the values of individual measurements
- The most used structural characteristics:
  - Median
  - Mode

Kozlíková, 2013

- Quantiles:
  - Quartiles
  - Deciles
    - Percentiles

#### Median

- Divides an ordered sample into two equally sized parts (with the same probability 0.50)
  - Is calculated according to the below mentioned formulas depending on whether the sample size is odd or even

Odd *n*: 
$$\widetilde{X} = X_{\frac{n+1}{2}}$$
 Even *n*:  $\widetilde{X} = \frac{X_n + X_n}{\frac{2}{2} + \frac{2}{2} + 1}}{2}$ 

#### Used nomenclature:

 $\widetilde{x}$  : median

líková, 2013

- *n* : sample size
- $x_n$  :  $n^{\text{th}}$  measurement of the quantity X in the sorted sample

#### Median - Examples



#### Figure 2: Examples of medians.

Kozlíková, 2013

In the case of odd number of elements in an ordered sample (the graph on the left), the median equals the "middle" value (value of the item F).

In the case of even number of elements in an ordered sample (the graph on the right), the median is calculated from the two "middle" values (average of values of the items E and F).

#### Median - Sorted Samples (can be omitted without loss of continuity )

 In a sorted and ordered sample, the median interval (the median class, in which the median is) is established, from which the median is interpolated

$$\widetilde{X} = \widetilde{X}_{L} + \frac{\widetilde{h}}{\widetilde{n}} \cdot \left(\frac{n}{2} - n_{(-1)}^{cum}\right)$$

Used nomenclature:

 $\widetilde{X}$  : median

Kozlíková, 2013

 $\widetilde{h}$  : width of the median interval

- $\widetilde{X}_{X}$ : lower end of the median interval
- $\widetilde{n}$ : frequency of the median interval
- *n* : sample size

 $n_{(-1)}^{cum}$ : cumulative frequency of the interval before the median interval



#### • The most frequent value in the sample

Â



Figure 3: An example of a mode.

The mode is represented by the item G.

Ka erína Kozlíková, 2013

2013/14 Medical Statistics II - Sample characteristcs

#### Mode – Sorted Sample (can be omitted without loss of continuity )

 In an ordered sample, the modal interval (the class with the highest frequency) is found, from which the mode can be interpolated in any of the two ways

$$\hat{x} = \hat{x}_{L} + \hat{h} \cdot \frac{\hat{n}_{(+1)}}{\hat{n}_{(-1)} + \hat{n}_{(+1)}} \quad \text{or} \quad \hat{x} = \hat{x}_{L} + \hat{h} \cdot \frac{\hat{n} - \hat{n}_{(-1)}}{2\hat{n} - \hat{n}_{(-1)} - \hat{n}_{(+1)}}$$

Used nomenclature:

- $\hat{x}$  : mode  $\hat{h}$  : length of the modal interval
- $\hat{x}_L$ : lower end of the modal interval
- $\hat{n}$  : frequency of the modal interval
- $\hat{n}_{(-1)}$  frequency of the interval before the modal interval
- $\hat{n}_{(+1)}$ : frequency of the interval after the modal interval

#### Symmetric Distribution

The mean, the median, and the mode have the same value



Figure 4: A histogram and a polygon of a symmetric distribution.

2013/14 Medical Statistics II - Sample characteristcs

Kozlíková, 2013

#### Asymmetric Distribution Left-Sided

 At the numerical axis, the mode is at most to the left (has the highest count, but a low value), the mean is at most to the right (has a high value) and the median is in the middle



Figure 5: A histogram and a polygon of an asymmetric left-sided distribution.

Ka erína Kozlíková, 2013

 $\hat{X} < \hat{X} < \overline{X}$ 

2013/14 Medical Statistics II - Sample characteristcs

#### Asymmetric Distribution Right-Sided

 At the numerical axis, the mean is at most to the left (has a low value), the mode is at most to the right (has the highest count and a high value) and the median is in the middle

 $\overline{X} < \overset{\sim}{X} < \hat{X}$ 

Kozlíková, 2013



Figure 6: A histogram and a polygon of an asymmetric right-sided distribution.

2013/14 Medical Statistics II - Sample characteristcs

#### Approximate Relation Between Mode, Mean, and Median

• The distance between the mean and the mode is <u>approximately</u> three times larger than is the distance between the median and the mean

$$\hat{x} - \overline{x} \cong 3 \cdot (\widetilde{x} - \overline{x})$$

• The relation can be used to estimate the mode or to check the correctness of its calculation

líková, 2013

• The mean and the median can usually be calculated more easily

#### Quantile

- Quantile  $K_j^{(m)}$  divides the probability distribution of a random variable into *m* parts
  - *m* denotes the number of parts of the set
  - *j* is the order number of the quantile
    - *j* = 0, 1, 2, ...., m-1, m
- Each part has the probability 1/m
- Quantile value indicates how much of the set has a value less than or equal to the quantile value
- Quantile special names:
  - Lower j = 1:  $K_1^{(m)}$ 
    - Upper j = m 1:  $K_{m-1}^{(m)}$

## Calculation of the Quantile (can be omitted without loss of continuity )



## The Mostly Used Quantiles

- Quartile  $Q_i$ 
  - Divides an ordered sample into m = 4 parts, all with probability 0.25
- Sextile *S<sub>i</sub>* 
  - Divides an ordered sample into m = 6 parts, all with probability 1/6
- Decile D<sub>i</sub>
  - Divides an ordered sample into m = 10 parts, all with probability 0.10
- Percentile  $P_i$ 
  - Divides an ordered sample into m = 100 parts, all with probability 0.01

Examples
$$Q_1 = P_{25}$$
 $Q_2 = D_5 = P_{50} = \widetilde{X}$  $Q_3 = P_{75}$ líková, 20132013/14 Medical Statistics II - Sample characteristics

## Measures of Variability

#### Measures of Variability – Introduction (1)

- Statistical data as numeric variables are always various (they vary, have different values)
- A small degree of variability means a small mutual diversity of values of a given variable (the values are very similar)
  - In this case, the central tendencies such as mean, median or mode, are good characteristics of the general size of the value of the variable in the sample
- High variability means a large mutual difference among values of a given variable
  - In this case, the calculated parameters are good characteristics of the general size of the value of the variable in the sample

#### Measures of Variability – Introduction (2)

- Characteristics of the middle of the set (the central tendencies)
  - Only <u>indicate</u> the location information of a statistical sample on numerical axis
  - They do not inform
    - How are the values scattered around the centre in the set
    - Whether there are any outliers in the sample
- Measures of variability (variability characteristics)

Reflect the distribution of values of the variable around the mean value of the set

## Measures of Variability Connected with Averages

- Take into account the values of the measurements of the quantity *X*
- The mostly used measures:
  - Standard deviation
  - Mean deviation (Average error)
  - Standard error of the mean
  - Coefficient of variation
  - Dispersion coefficient

líková, 2013

#### Variance (Sample Variance)

- Characterizes the scatter of measurements around the sample mean
  - Calculated as sum of squared deviations of random variables in a random sample from their sample mean divided by the number of terms in the sum minus one

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}$$

#### Used nomenclature:

- $s_x^2$ : sample variance
- $\overline{x}$ : average (arithmetic mean)
- $x_i$ : *i*<sup>th</sup> measurement of the quantity X
- i : measurement number (i = 1, 2, ..., n)

*n* : sample size

- n-1: degrees of freedom
- $\boldsymbol{\Sigma}$  : summation sign

#### **Do Not Confuse the Sample Variance** with the Population Variance!

Population variance







Used nomenclature:

Kozlíková, 2013

- $\sigma_x^2$  : population variance
- $\mu$  : true value of the mean
- $x_i$ : *i*<sup>th</sup> measurement of the quantity X
- i : measurement number (i = 1, 2, ..., n)
- N : population size
  - $\boldsymbol{\Sigma}$  : summation sign

#### Weighted Variance (Sample Variance)

- Characterizes the dispersion of all measurements around the weighted arithmetic mean of the sample
  - It is calculated analogously to the "normal" mean, but takes into account the frequency of each class

$$S_x^2 = \frac{\sum_{k=1}^{\kappa} (x_j - \overline{x})^2 \cdot n_j}{n-1}$$

where

 $n_1 + n_2 + \ldots + n_k = n$ 

Used nomenclature:

Kozlíková, 2013

- $s_x^2$  : sample variance
- $x_i$ : mid-point of the *j*<sup>th</sup> class
- $n_i$ : frequency ("weight") of the j<sup>th</sup> class
  - : class number (j = 1, 2, ..., k)

- $\overline{x}$  :mean (weighted)
- *n* : sample size
- k : number of classes
- $\boldsymbol{\Sigma}$  : summation sign

#### Standard Deviation (Sample Standard Deviation)

- Characterizes the scatter of measurements around the sample mean
  - Calculated as the square root of the sample variance

$$S_{x} = \sqrt{\frac{\sum_{i=1}^{n} (x_{i} - \overline{x})^{2}}{n-1}}$$

or another expression

 $S_x = \sqrt{S_x^2}$ 

# Used nomenclature: $S_x$ : sample standard deviation $S_x^2$ :sample variance $\overline{X}$ : average (arithmetic mean)n: sample size $x_i$ : $i^{th}$ measurement of the quantity Xn-1: degrees of freedomi: measurement number (i = 1, 2, ..., n) $\Sigma$ : summation sign

#### **Do Not Confuse** The Sample Standard Deviation With The Population Standard Deviation!

The square root of the variance

 $\sigma_{x} = \sqrt{\frac{\sum_{i=1}^{N} (x_{i} - \mu)^{2}}{N}}$  $\mu = \frac{\sum_{i=1}^{N} X_i}{N'}$ where

Used nomenclature:

Kozlíková, 2013

- $\sigma_x$ : population standard deviation N : population size
- $\mu$  : true value of the mean
- $x_i$ : *i*<sup>th</sup> measurement of the quantity X
- i: measurement number (i = 1, 2, ..., n)
- $\Sigma$  : summation sign

#### Weighted Standard Deviation (Sample Standard Deviation )

- Characterizes the dispersion of all measurements around the weighted arithmetic mean of the sample
  - It is calculated analogously to the "normal" mean, but takes into account the frequency of each class

$$S_{x} = \sqrt{\frac{\sum_{k=1}^{k} (x_{j} - \overline{x})^{2} \cdot n_{j}}{n-1}}$$

where

 $n_1 + n_2 + \ldots + n_k = n$ 

Used nomenclature:

Kozlíková, 2013

- $s_x$ : sample standard deviation
- ix; mid-point of the j<sup>th</sup> class
  - : frequency ("weight") of the j<sup>th</sup> class

: class number 
$$(j = 1, 2, ..., k)$$

 $\overline{x}$  :mean (weighted)

- *n* : sample size
- *k* : number of classes
- $\boldsymbol{\Sigma}$  : summation sign

#### Mean Deviation (Average Error)

Mean absolute deviation





#### Standard Error of the Mean (Error in the Mean)

 Used for comparison of sample means (from the same population)

or

$$se_{x} = \sqrt{\frac{\sum_{i=1}^{n} (x_{i} - \overline{x})^{2}}{(n-1) \cdot n}}$$

#### Used nomenclature:

 $s_x$ : sample standard deviation

 $Se_x = \frac{S_x}{\sqrt{n}}$ 

- $\overline{x}$ : average (arithmetic mean)
- $x_i$ : *i*<sup>th</sup> measurement of the quantity *X*
- i : measurement number (i = 1, 2, ..., n)

- $se_x$ : sample standard error
  - *n* : sample size
  - $\boldsymbol{\Sigma}$  : summation sign

#### Weighted Error in the Mean

 Used for comparison of sample means (from the same population)

$$se_{x} = \sqrt{\frac{\sum_{j=1}^{k} (x_{j} - \overline{x})^{2} \cdot n_{j}}{(n-1) \cdot n}}$$

where

$$\sum_{j=1}^{k} n_j = n$$

#### Used nomenclature:

- $se_{\star}$  : standard error of the mean
  - $x_{j}$ : mid-point of the  $j^{\text{th}}$  class
  - $n_j$ : frequency ("weight") of the  $j^{\text{th}}$  class
- *j* : class number (j = 1, 2, ..., k)

- $\overline{x}$  :mean (weighted)
- *n* : sample size
- k : number of classes
- $\boldsymbol{\Sigma}$  : summation sign

#### Coefficient of Variation (1) (Sample Coefficient of Variation)

- It is a characteristic of variability (probability distribution) of a random variable
  - In general, it is defined as the ratio of the standard deviation and the (absolute) value of the mean (average)

$$V_{x} = \frac{S_{x}}{\left|\overline{x}\right|} \cdot 100\%$$

Used nomenclature:

Kozlíková, 2013

- $V_x$ : variability coefficient
- $\overline{x}$ : average (arithmetic mean)

 $s_x$ : sample standard deviation

*n* : sample size

#### **Coefficient of Variation (2)**

- Is a relative measure of variability
- Is not affected by the absolute values of the character
- Is used
  - For a mutual comparison of variability of two or more sets with significantly different levels of values (for example, mass in grams and kilograms)
  - For statistical control (laboratory results)



#### **Coefficient of Dispersion** (can be omitted without loss of continuity )

- Is a relative measure of variability (probability distribution) of a random variable
  - In general, it is defined as the mean deviation calculated from the median and the (absolute) value of the median

$$d_{x} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} |x_{i} - \widetilde{x}|}{|\widetilde{x}|}$$

#### Used nomenclature:

- $d_x$ : coefficient of dispersion
- $x_i$ : value of the *i*<sup>th</sup> measurement
- $\widetilde{x}$  : median
- *n* : sample size

serína Kozlíková, 2013

2013/14 Medical Statistics II - Sample characteristcs

#### Measures of Variability Connected with Structural Characteristics

- Take into account the <u>ranks</u> of the measurements of the quantity X in an <u>ordered</u> sample
- The mostly used measures
  - Ranges

Kozlíková, 2013

- Range (Sample range)
- Inter-quantile range
- Inter-quartile range
- Inter-quartile deviation

#### Range (Sample Range)

- The difference between the highest value (maximum) and the lowest value (minimum) of the sample
  - The largest ordered statistics minus the smallest ordered statistics

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

#### Used nomenclature:

- R : range
- $x_{max}$ : the maximal value of the measured quantity X
- $x_{min}$ : the minimal value of the measured quantity X

#### Inter-Quantile Range

• The difference between the upper and the lower quantile

$$R_{K} = K_{m-1}^{(m)} - K_{1}^{(m)}$$

- A nonparametric measure of variability
  - It does not take into account the size of all values file
  - Excludes the impact of outliers

Kozlíková, 2013

- In the case of deciles, in this interval would lie 80% of values  $(D_9 D_1)$
- In the case of percentiles, in this interval would lie 98% of values  $(P_{99} P_1)$

#### Inter-Quartile Range

• The difference between the upper quartile and the lower quartile

$$R_Q = Q_3 - Q_1$$

- A nonparametric measure of variability
  - A special case of the inter-quantile range
  - The interval, in which lies the half of the sample set
  - It is often used with the median

líková, 2013

- Is often used to find outliers in the sample
  - Outliers are observations that fall below  $Q_1 1.5 \cdot R_Q$  or above  $Q_3 + 1.5 \cdot R_Q$

#### Inter-Quartile Range – Example

Figure 7: Box plot (plot at the top) with an inter-quartile range ( $R_Q = IQR = Q_3 - Q_1$ ) and a probability density function of a normal N(0, $\sigma^2$ ) population

Available at: http://en.wikipedia.org/wiki/Interquartile\_range [17. 9. 2013]



Kalerína Kozlíková, 2013

2013/14 Medical Statistics II - Sample characteristcs

### **Quantile Deviation**

- Determined as an average of positive deviations of neighbouring quantiles  $\mathbf{k}^{(m)} = \mathbf{k}^{(m)}$ 
  - After editing, the formula is

$$R_{\kappa^{(m)}} = rac{K_{m-1}^{(m)} - K_{1}^{(m)}}{m-2}$$

- Measure of variability, which is neither affected by extremely high nor by extremely low values
- It takes into account the variability of intermediate values
  - In the case of deciles, variability of 80% of values
    - In the case of percentiles, variability of 98% of values

$$R_{D/8} = \frac{D_9 - D_1}{8}$$

$$R_{P/98} = \frac{P_{99} - P_1}{98}$$

#### **Quartile Deviation**

 Determined as an average of positive deviations of neighbouring quartiles

$$R_{Q/2} = \frac{Q_3 - Q_1}{2}$$

- It is a measure of the spread through the middle half of a distribution
- It is neither influenced by extremely high nor extremely low scores

Kozlíková, 2013

 Quartile deviation is an ordinal statistic and is most often used in conjunction with the median

## **Measures of Shape**

#### Skewness (1)

- Skewness measures the direction and degree of asymmetry of the distribution of a random variable
- There are more measures of skewness, the most used is the coefficient of skewness
  - For the selection of the normal distribution, the skewness is calculated according to one of the equivalent formulas



It is defined only for  $s_x \neq 0$  and  $n \geq 3$ 

Ko líková, 2013 2013/14 Medical Statistics II - Sample characteristcs

#### Skewness (2)

- If the calculated skewness
  - Equals zero (As = 0)
    - It characterizes a symmetric distribution (fig. 2 and 7)
  - Is positive (As > 0)
    - It characterizes right sided skewness, that is, left sided asymmetric distribution, when the mean is greater than the median (fig. 3)
  - Is negative (As < 0)

Kozlíková, 2013

• It characterizes left sided skewness, that is, right sided asymmetric distribution, when the mean is smaller than median (fig. 4)

#### **Excess Kurtosis (1)**

- Kurtosis characterizes the "peakedness" (flatness or sharpness) of a distribution
  - There are more measures of kurtosis, the mostly used is the excess kurtosis
  - For the sample of the normal distribution, the excess kurtosis is calculated according to one of the equivalent formulas



I is defined only for  $s_x \neq 0$  and  $n \ge 4$ 

a arína Kozlíková, 2013

2013/14 Medical Statistics II - Sample characteristcs

#### **Excess Kurtosis (2)**

- If the calculated excess kurtosis
  - Equals zero (Exc = 0)
    - It characterizes a normal distribution (fig. 7)
  - Is positive (*Exc* > 0)
    - It characterizes a more peaked distribution (with a sharp peak) than the normal distribution, therefore, more values are concentrated around the central tendency, less values are at the tails of the distribution
  - Is negative (*Exc* < 0)</li>

líková, 2013

 It characterizes a less peaked distribution (with a rounded peak) than the normal distribution, therefore, less values are concentrated around the central tendency, more values are at the tails of the distribution

#### Literature

- HAWKINS, D. *Biomeasurement: Understanding, Analysing and Communicating Data in the Biosciences.* Oxford University Press : New York, 2007. 300 p. ISBN 978-0199265152.
- CHAJDIAK, J. *Štatistika jednoducho.* Statis : Bratislava, 2003. 88 s. ISBN 80-85659-28.
- KOZLÍKOVÁ, K. *Základy spracovania biomedicínskych meraní I.* Asklepios : Bratislava, 2003. 88 s. ISBN 80-7167-064-2.
- KOZLÍKOVÁ, K. Lekárska štatistika I. Základné pojmy. [Cit. 1. 10. 2012].
   Dostupné na internete: http://portal.fmed.uniba.sk/
- KOZLÍKOVÁ, K., MARTINKÁ, J. *Základy spracovania biomedicínskych meraní II.* Asklepios : Bratislava, 2009. 204 s. ISBN 978-80-7176-137-4.
- KOZLÍKOVÁ, K., MARTINKA, J. *Theory And Tasks For Practicals On Medical Biophysics.* Librix : Brno, 2010. 248 p. ISBN 978-80-7399-881-1.
- KOZLÍKOVÁ, K., MARTINKA, J., KNEZOVIĆ, R. Otázky na overenie znalostí z vybraných kapitol lekárskej fyziky a biofyziky. EQUILIBRIA : Košice, 2013. 174 p. ISBN 978-80-8143-105-0.
- STN ISO 3534-1. Štatistika. Slovník a značky. Časť 1: Všeobecné štatistické termíny a termíny používané v teórii pravdepodobnosti. Slovenský ústav technickej normalizácie : Bratislava, 2008. 90 s.

Comment:

líková, 2013

If not stated else, the author of the text is the author of the figures and graphs as well.