

# **STATISTICS**

## **Analysis on quantitative criteria (1)**

**Grouping, positional and calculated  
means, position indices (quantiles)**

# **ANALYSIS OF A POPULATION ACCORDING TO ONE CRITERION**

- **Individual and grouped data,**
- **Position indicators,**
- **means,**
- **dispersion,**

- **Grouping of a population**

Grouping is a division of a population according to a given criterion.

Grouping should be

- *free from overlapping*
- complete
- Namely: *all data* (each element of the population) should belong exactly to one and only one group of the population.

A grouping criterion may be

- **Qualitative criterion:** e.g. grouping of students according to subjects, grouping of inhabitants according to settlement type.
- **Quantitative criterion:**
  - A simpler case is when we speak about **recurring, discrete variants** (criterion values): e.g. grouping families according to the number of children, grouping students according to their mark on Statistics, which can be 1, 2, ... 5.
  - In case of **non-recurring variants**, the variants of the population ranked in increasing order can be classified into intervals. E.g. monthly payment of a company's employees can be grouped into intervals.

- **Characterization of a grouped population**

- The first important information on a grouped population is the number of data in each individual group. This information is called frequency.

- *Frequency, relative frequency*

- **Frequency ( $f_i$ )** specifies that how many units of the population are included in the  $i$ -th group.

- **Relative frequency ( $g_i$ )** specifies that how much proportion of the population (in percentage) is included in the  $i$ -th group:

$$g_i = \frac{f_i}{\sum_{i=1}^m f_i} = \frac{f_i}{n}$$

- where  $g_i$  is the relative frequency of the  $i$ -th group,  $m$  is the number of groups,  $n$  is the number of the population.

- **Cumulative frequency:** means an accumulated summation.
  - **Cumulative frequency:** means that how many units belongs to the first  $k$  groups of the population.

$$f_i' = \sum_{i=1}^k f_i$$

- **Cumulative relative frequency :** means that how much fraction belongs to the first  $k$  groups of the population.

$$g_k' = \sum_{i=1}^k g_i = \frac{\sum_{i=1}^k f_i}{\sum_{i=1}^m f_i} = \frac{\sum_{i=1}^k f_i}{n}$$

- For quantitative criteria – according to the values – *upward and downward cumulated frequency can be distinguished*. If higher variants are assigned to higher *i*-values (e.g. if workers with higher salary belong to intervals of higher serial number), then:
  - ***upward cumulated frequencies ( $f_i'$ ) / relative frequencies ( $g_i'$ )*** show that altogether how many data / how much fraction (%) can be found **in the first *i* intervals**.
  - ***downward cumulated frequencies ( $f_i''$ ) / relative frequencies ( $g_i''$ )*** show that altogether how many data / how much fraction (%) can be found **in the *i*-th and the following intervals**.

- **Summed up value:** In case of quantitative criteria, summed up value is the **sum of values being in a given group**. It is indicated by:  $s_i$ . If identical values are in a given group then summed up value can be calculated as follows:

$$s_i = f_i \cdot x_i$$

- **Relative summed up value:** it means the share of the summed up value of a given group *within the whole summed up value*:

$$z_i = \frac{f_i \cdot x_i}{\sum_{i=1}^m f_i x_i}$$

- **Cumulative rows (series)** can be calculated both from absolute and relative summed up values.



- ***Statistical row:*** listing statistical data *on given criteria*. (It is advisable to give them in a table.)
- ***Frequency row:*** listing frequency data *according to given groups*.

More:

- ***relative frequency row;***
- ***cumulative frequency row;***
- ***relative summed up value row, etc.***

- **Example.** In a warehouse 200 boxes of juice are stored with four kinds of capacity, in the table below. (The volume is in decilitre.)

	Values Volume (dl) $x_i$	Frequency $f_i$	Relative frequency $g_i (%)$	Summed up value $s_i$	Rel. summed up value $z_i$	Cumulative summed up value $s'$	Cumulative rel. summed up value $z'$
	20	15					
	10	25					
	3,3	50					
	2	110					
Total	35.3	200					

# Means and position indicators in case of individual and grouped data

Characterization of data sets and empirical distributions:

## ***Positional mean values:***

- *modus (value occurring most frequently, a „most fashionable / most trendiest – a la mode”- value);*
- *median;*

## ***Position indicators:***

- *quantiles;*

## ***Calculated means:***

- arithmetic average (simple and weighted);
- harmonic average (simple and weighted);
- geometric mean (simple and weighted);
- quadratic average (simple and weighted);
- chronological average (simple and weighted);

## ***Positional means:***

### ***Modus***

#### ***Concept of modus in case of recurrent data:***

- **For a discrete criterion:** the most frequently occurring element;
- **For a continuous criterion:** the maximum value of the frequency curve;

**Example:** If we take marks of Statistics essays and mark 4 is the most frequent, then modus of the marks is 4. The modus is not always clear, it may be two or more modes of a data set.

For non-repeating individual data, if each data differ from each other (frequency of all values is 1), it makes no sense to talk about modus.

# Characteristics of modus

## Advantages:

- typical values;
- it can be used for all measurement scales;
- not sensitive to extremes, outliers;

## Drawbacks:

- it may not exist or more than one modus may also occur;

## ***The concept of modus in case of interval-grouped data:***

**Modal interval: wherein the density of the data is the largest.**

**EXAMPLE:** If we analyze essay marks of a course, where 100 points was the maximum, and the interval between the points 60 and 80 includes the densest data, then this is the modal interval.

Since there may be more class intervals with the same data density, the definition of the class intervals is not always clear. **Modal class interval** can well be illustrated by a histogram: This is the class interval, which **includes the tallest column.**

## ***Assessment of the modus:***

The modal class interval is an interval, around which the data are most concentrated. In the lack of knowledge of basic data this is of course only an estimate.

**The simplest option for selecting mode is the middle of the modal class,** but based on practical experiences it seems a better estimate if density of the two adjacent intervals is also taken into account, believing that the mode is closer to that edge of the interval, density of which is closer to the modal density.

**The calculation would be: to determine that how much the data density of the two neighbouring classes is smaller than the data density of the modal class.** Let mark these differences  $k_1$  and  $k_2$ , and select the value dividing the interval in ratio  $k_1 : k_2$ . Based on this, the estimated value of the mode can be determined by the following formula:

$$Mo = x_{i0} + \frac{k_1}{k_1 + k_2} \cdot h_i$$

where  $x_{i0}$  is the lower threshold of the modal class, while  $h_i$  is the width of the modal class.

**EXAMPLE.** If value of  $k_1$  and  $k_2$  are 7 and 4, respectively, the lower threshold of the modal class is 50, and the width of the modal class is 10, then the estimated value of the mode:

$$Mo = 50 + \frac{7}{7 + 4} \cdot 10 \cong 56,4$$

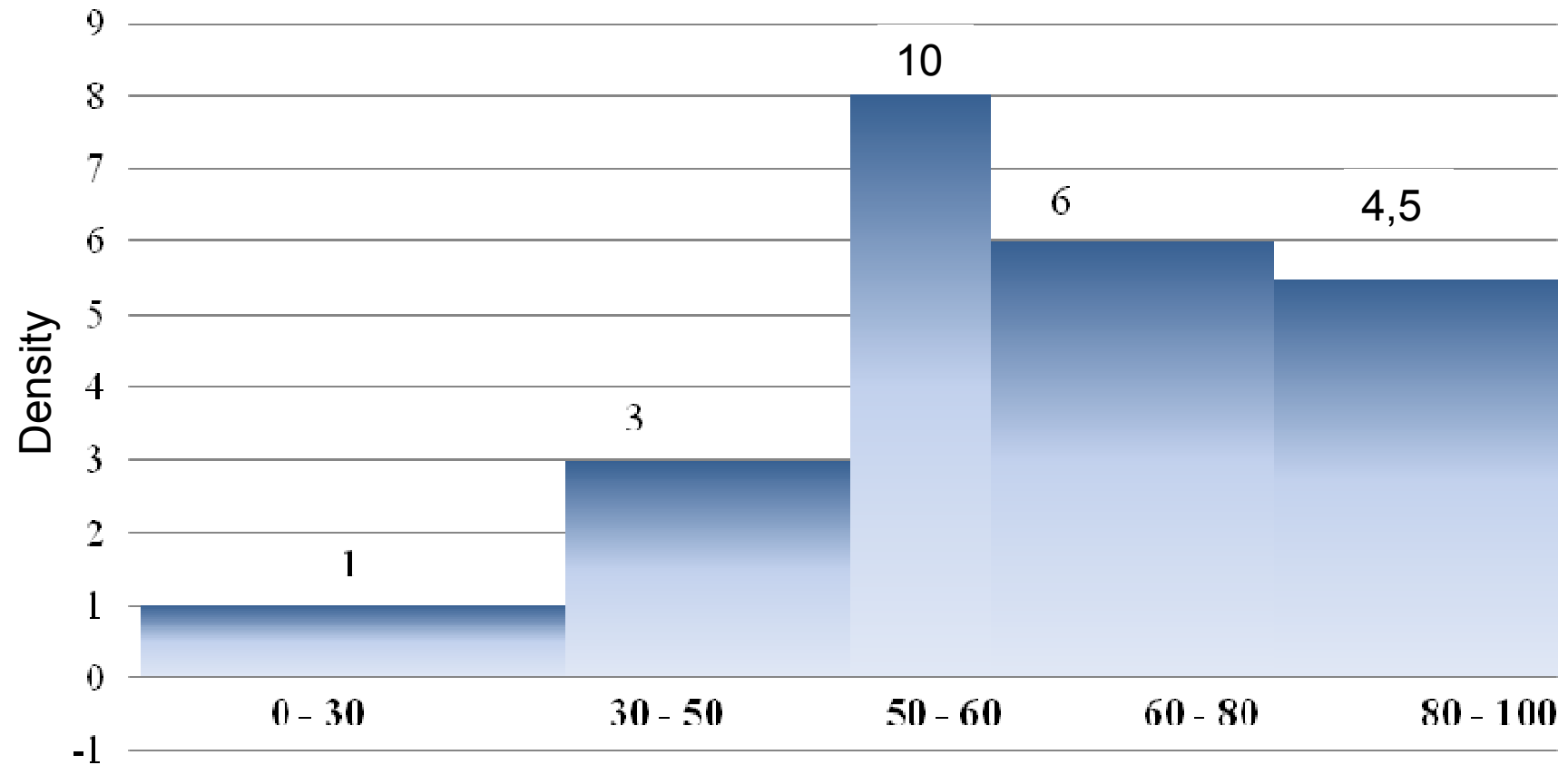
### EXAMPLE:

Distribution of the students according to score-intervals:

<b>Thresholds (class intervals)</b>	<b>Frequency <math>f_i</math></b>	<b>Width of class intervals <math>h_i</math></b>	<b>Density <math>f_i / h_i</math></b>
0 — 30	30	30	1
30 — 50	60	20	3
50 — 60	100	10	10
60 — 80	120	20	6
80 — 100	90	20	4,5
<b>Total</b>	<b>400</b>		

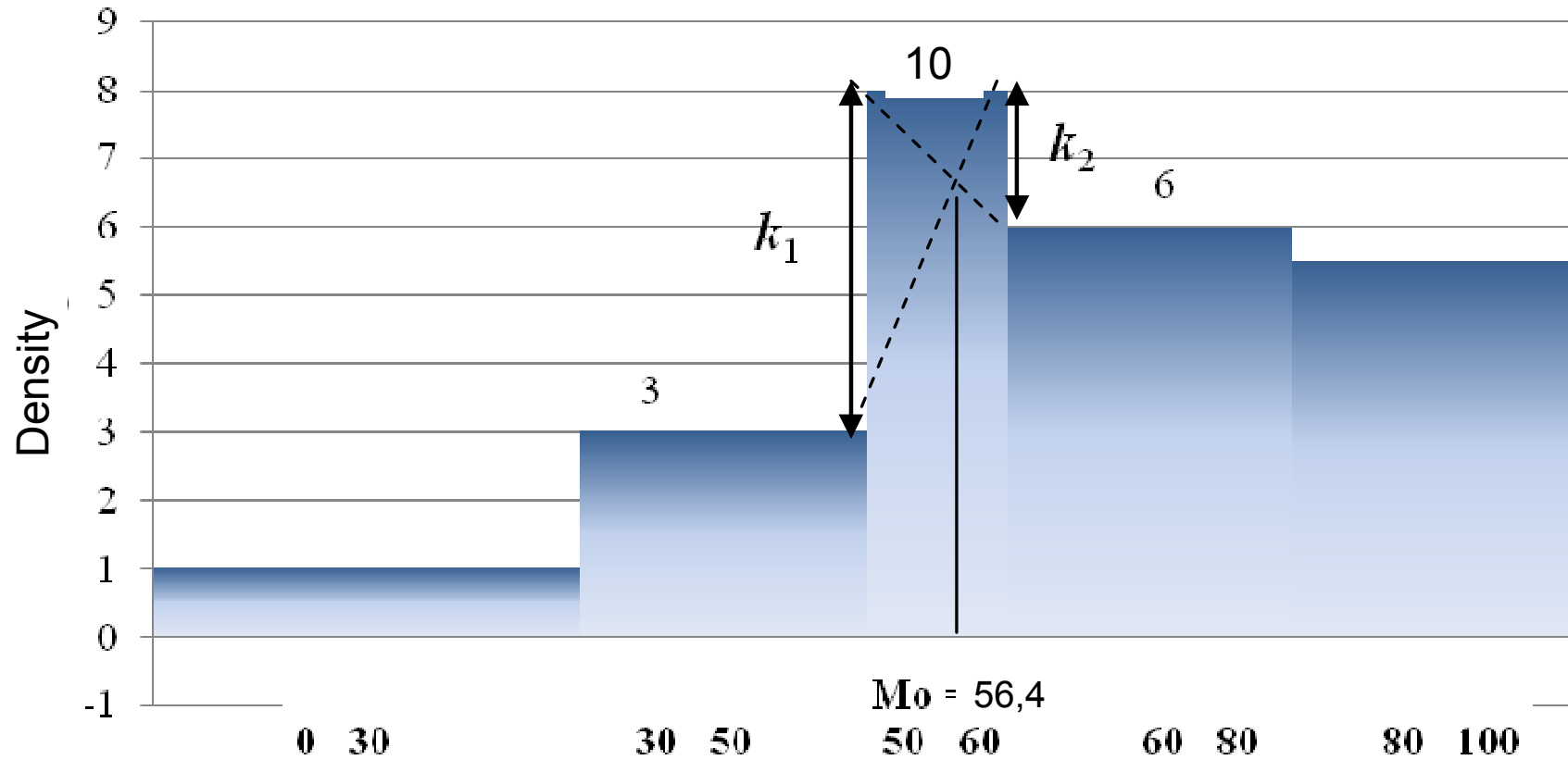


Distribution of the students according to scores  
(Histogram)



The most frequent class interval is the 4th, while the densest is the 3rd.

Distribution of the students according to scores  
(Histogram)



*Interpretation:* scores are condensed to the greatest extent around the value of 56.4

## Quantiles:

They give a concise description of the location of the criteria.

## They are named:

**Median ( $M_{e2i}$ )** if the data set is divided into **two** parts;

**Quartile ( $Q_{4i}$ )** if the data set is divided into **four** parts;

**Quintile ( $K_{5i}$ )** if the data set is divided into **five** parts;

**Decile ( $D_{10i}$ )** if the data set is divided into **ten** parts;

**Pentile ( $Pen_{20i}$ )** if the data set is divided into **twenty** parts;

**Percentile ( $Per_{100i}$ )** if the data set is divided by **one hundred** parts;

## Their definition (for discrete criteria):

- a) Prepare ranking (ascending)
- b) Determine the serial number of quantile values (division point)  
(i.e., what quantile I want to define?)

$$s_j = \frac{j}{k}(n+1)$$

### where:

n = number of data;

k = the number of equal parts (quantiles);

j = 1,2,...k-1 sequence number within the given quantile values (i.e. it specifies that quantile value of which serial number is searched);

**The quantile questioned is the data for the  $s_j$ -th serial no.**

# Example: quantiles of the monthly gross salaries based on a sample of n (=95) elements

**Median:**

$$s = \frac{1}{2}(95 + 1) = 48$$

$$\text{Me} = 201,000 \text{ HUF}$$

**Quartiles: (Q<sub>j</sub>)**

$$s_1 = \frac{1}{4}(95 + 1) = 24$$

$$Q_1 = 159,000 \text{ HUF}$$

$$s_2 = \frac{2}{4}(95 + 1) = 48$$

$$Q_2 = 201,000 \text{ HUF}$$

$$s_3 = \frac{3}{4}(95 + 1) = 72$$

$$Q_3 = 269,000 \text{ HUF}$$

## ***Median:***

Among the ordered data, median is the middle value; or in other words: the median is the value which divides the collated data into two equal parts.

*In case of individual data,*

a) **if the data are of odd number**, then the last-mentioned definition is clear, because there is a middle data, before which is the same amount of data as after;

b) **if the data are of even number** then there are two middle data; in this case any of them can be considered median. In practice, arithmetic mean of the two values are usually given.

⇒ ***The median is the value at which up to 50% of the data are smaller and up to 50% of the data are higher.***

# Advantages of median

- clearly defined;
- it can be determined not only in the case of quantitative characteristics, but it can be ranked in case of quality criteria, as well;
- it is independent of extreme values;

# Disadvantages of median

- it can be calculated only from elements arranged in rank;
- if a significant proportion of the individuals have the same variant, then it is not practical to use;

## ***Definition of median in case of class interval frequency series:***

I look for a value that divides the data set arranged in rank into two equal parts. This is only an estimate because we do not know the basic data, only the frequency series.

**First, we determine the class interval containing the median.** It is easy to do when you consider that both the pre-class intervals and the subsequent class intervals containing the median comprise less than half of all data. **Let the class index containing the median is  $i$ .**

To get half to the data, additional data of the  $i$ -th class still need to be taken, that is, some share of the  $i$ -th class. Then an equal proportion of the  $i$ -th interval should also be taken, and thus the median has already been determined.

**The calculation is as follows**, if you already know which class contains the median (let the serial number of it is  $i$ ):



- Let all classes preceding the  $i$ -th class comprising the median include  $f'_{i-1}$  number of data. We need to take yet  $(n/2 - f'_{i-1})$  number of data from the  $i$ -th class so that to get half of all data.
- Let's see how much ratio is the number of the data taken from the  $i$ -th class compared to the frequency of the  $i$ -th class. The value that splits the  $i$ -th class with the same percentage split between, will be the median.

Namely:

$$\text{Me} = x_{i0} + \frac{\frac{n}{2} - f'_{i-1}}{f_i} \cdot h_i$$

The same with relative frequencies:

$$\text{Me} = x_{i0} + \frac{0,5 - g'_{i-1}}{g_i} \cdot h_i$$

Here,  $x_{i0}$  is the lower threshold comprising the median, while  $h_i$  is the width of this class.

## ***Relative terms of median - quantiles:***

### **quartiles, deciles, pentiles, percentiles**

**Median:** the "middle value": 50% of the data arranged in rank are smaller, while the other 50% are greater than median.

Similarly, other position indicators can also be defined:

what value can be found at

- ✓ a quarter,
- ✓ three-quarter,
- ✓ one-third,
- ✓ some proportion of  $p$  of the data.

**The general formula for determining the a  $p$ -quantile:**

$$K_{\text{vant}}(p) = x_{i_0} + \frac{p - g'_{i-1}}{g_i} \cdot h_i$$

Frequently used ratios, quantiles were given special names, some of which are:

**Lower quartile:**  $\frac{1}{4}$ -th of the data are smaller than this value, while  $\frac{3}{4}$ -th of the data are higher than this value:

$$Q_1 = x_{i_0} + \frac{0,25 - g'_{i-1}}{g_i} \cdot h_i$$

**Middle quartile:** the same as the median.

**Upper quartile:**  $\frac{3}{4}$ -th of the data are smaller than this value, while  $\frac{1}{4}$ -th of the data are higher than this value:

$$Q_3 = x_{i_0} + \frac{0,75 - g'_{i-1}}{g_i} \cdot h_i$$

**Upper decile:**  $\frac{9}{10}$ -th of the data are smaller than this value, while  $\frac{1}{10}$ -th of the data are higher than this value.

**Lower pentile:** 95% of the data are smaller than this value, while 5% of the data are higher than this value.

**Upper percentile:** 99% of the data are smaller than this value, while 1% of the data are higher than this value.

# **AVERAGES**

# Characterization of the distribution of a population / sample

- determination of typical values;
- examination of the difference of the data;
- analysis of the distribution curve of the population / sample;

Means	Dispersal measures
<ul style="list-style-type: none"> <li>• positional (median, modus)</li> <li>• calculated</li> </ul> $\bar{X}, \bar{X}_h, \bar{X}_q, \bar{X}_g, \bar{X}_k$	<ul style="list-style-type: none"> <li>• standard deviation (<math>\sigma</math>)</li> <li>• relative standard deviation (V)</li> <li>• range (R)</li> <li>• interquartile range (IQR)</li> </ul>
Measurement of asimmetry	Other characteristics
<ul style="list-style-type: none"> <li>• Pearson indicator (A)</li> <li>• F indicator</li> <li>• <math>\beta_1</math> indicator</li> <li>• Graphical representation</li> </ul>	<ul style="list-style-type: none"> <li>• concentration</li> <li>• quantiles</li> <li>• momentums</li> <li>• graphic images</li> </ul>

# Requirements for means

- clear calculation;
- typical values;
- illustrative, good legibility;
- medium position:  $X_{\min} \leq K \leq X_{\max}$  ;

# Characteristics of means

- the quantitative criterion is characterized by a single figure;
- dimension: unit of the criterion;



# Means:

## Calculated means

**Arithmetic**  $\bar{x}$

**Harmonic**  $\bar{x}_h$

**Geometric**  $\bar{x}_g$

**Quadratic**  $\bar{x}_q$

- **Chronological**  $\bar{x}_k$

## Positional means

**Modus (Mo)**

**Median (Me)**

# Arithmetic mean

is the value, with which substituting the original values to be averaged, their sum remains unchanged.

For individual values:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Weighted form:

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

# Simple mean

The values occur only once

Mark (x)	Number of students / mark (f)
5	1
4	1
3	1
2	1
1	1
<b>Total</b>	<b>5</b>

The values occur repeatedly but at the same number of occasions

Mark (x)	Number of students / mark (f)
5	2
4	2
3	2
2	2
1	2
<b>Total</b>	<b>10</b>

# Weighted mean

The values occur repeatedly but at different number of occasions

Mark (x)	Number of students / mark (f)
5	3
4	8
3	6
2	2
1	1
<b>Total</b>	<b>20</b>

# Mathematical properties of the arithmetical average

- The amount of the deviations of each element from the average is 0:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

- When an "a" constant value is added to each element, the arithmetic average of the items deviates with "a" from the original average.
- If each item is multiplied with a "b" constant value, then the average of the items is "b" times higher than the original average of the elements.

# Mathematical properties of the arithmetical average

- If  $\bar{x}$  is the average of the elements  $x_1, x_2, \dots, x_n$  and
- $\bar{y}$  is the average of the elements  $y_1, y_2, \dots, y_n$
- then the average of  $x_1 + y_1; x_2 + y_2; \dots; x_n + y_n$  is  $\bar{x} + \bar{y}$ .
- Subtracting an arbitrary "a" constant from each element, the sum of the squares of these differences will be minimal if "a" is just the constant, i.e.,  $\bar{x}$

$$\sum_{i=1}^n (x_i - a)^2 \text{ minimal, if } a = \bar{x}$$

# An example to the properties of the arithmetical average

$x_i$	$x_i+50$	$x_i \cdot 1,1$	$Z=x_i+x_i \cdot 1,1$
100	150	110	210
150	200	165	315
210	260	231	441
240	290	264	504
300	350	330	630
$\Sigma$	1000	1250	2100
$\bar{x}$	<b>200</b>	<b>250</b>	<b>420</b>

# Advantages of the arithmetical average

- It is clear and understandable, its calculation is simple.
- Each data set has one and only one arithmetic average.
- Each item will be taken into account when calculating it.
- When calculating it, the knowledge of the individual values is not necessary, it is enough to know their number and amount.



# Drawbacks of the arithmetical average

- It is sensitive to outliers (trimmed mean).
- We can not take into account the unique values when using frequencies of classes.

# Geometric mean

Geometric mean is the number written in the place of the individual values, with which their multiplication is unchanged.

In case of unique values:

$$\bar{X}_g = \sqrt[n]{\prod_{i=1}^n X_i}$$

Weighted average form:

$$\bar{X}_g = \sqrt[n]{\prod_{i=1}^n X_i^{f_i}}$$

# The volume index of GDP in Hungary

Period	Previous quarter = 100%
2008. 1st quarter	100,9
2008. 2nd quarter	99,8
2008. 3rd quarter	99,0
2008. 4th quarter	98,1

Source: KSH Flash Report

Average rate of change:

$$\overline{x}_g = \sqrt[4]{1,009 \cdot 0,998 \cdot 0,99 \cdot 0,981} = \sqrt[4]{0,978} = 0,994 = 99,4\%$$

## Harmonic average:

**Simple harmonic average:** reciprocal of the average of the reciprocal to be averaged.

It can be used for averaging reverse intensity ratios.

$$\bar{X}_h = \frac{1}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

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

Shift hour needs of producing 100 workpieces for 4 machine types

Shift hour / 100 work pieces

Type	Performance
I.	45
II.	55
III.	40
IV.	43

$$\bar{x} = \frac{1}{\sum_{i=1}^n \frac{1}{x_i}} = \frac{1}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{1}{\frac{1}{45} + \frac{1}{55} + \frac{1}{40} + \frac{1}{43}} =$$

$$\bar{x} = \frac{1}{\frac{0,088}{4}} = \frac{1}{0,022} = 45.45 \text{ shifts / 100 workpieces}$$

- Weighted harmonic average

**When ratios are averaged and the numerator of the ratios is given as weight.**

$$\bar{X}_h = \frac{1}{\frac{f_1 \frac{1}{x_1} + f_2 \frac{1}{x_2} + \dots + f_n \frac{1}{x_n}}{f_1 + f_2 + \dots + f_n}} \quad \bar{X}_h = \frac{\sum_{i=1}^n f_i}{\sum_{i=1}^n f_i \frac{1}{x_i}}$$

Average amount per capita of a product  
manufactured by region, Northern Hungary, 2008

Areal unit	Produced quantity ( $f_i$ ) thousand t	Avarage per capita ( $x_i$ ) t/capita	$\frac{f_i}{x_i}$
Borsod-Abaúj- Zemplén	114.04	39.38	2.896
Heves	90.64	33.57	2.700
Nógrád	14.10	19.29	0.731
Northern-Hungary	218.78	-	6.327

$$\bar{x} = \frac{\sum_{i=1}^n f_i}{\sum_{i=1}^n \frac{f_i}{x_i}} = \frac{114,04 + 90,64 + 14,10}{\frac{114,04}{39,38} + \frac{90,64}{33,57} + \frac{14,10}{19,29}} = \frac{218,78}{6,327} = 34.58 \text{ t/capita}$$



## Harmonic average

- **It can only be applied if reciprocals of the values to be averaged have tangible sense.**
- In practice, the weighted form occurs frequently:
  - calculation of the average from data of value amounts;
  - Calculation of complex ratios;

## Geometric mean:

**It shows the average rate of development.**

The change can be expressed in absolute (sum) and relative (multiplicative) degree.

Its size is decided by the two extremes.

**It can be used for a value series of one-way tendency.**

If the change is not one-way, the statistical series should be split into sections.

It serves typical values on temporal or intensity range of the statistical series.

## Quadratic average:

- ✓ It is sensitive to outliers;
- ✓ When replacing the values to be averaged by quadratic average, their sum of squares are unchanged;

$$\bar{X}_q = \sqrt{\frac{\sum x_i^2}{n}}$$

$$\bar{X}_q = \sqrt{\frac{\sum f_i x_i^2}{\sum f_i}}$$

## ***Chronological average***

The chronological average is a type of mathematical average.

**Applicable for averaging of state time series, where data are available in equidistant intervals.**

**E.g. we can count average stock and average headcount using chronologic average.**

Chronological average is calculated as follows: the first data (NYK) is divided by 2, then add the rest of the data, as well as half of the last data, finally they are divided by one less than the total number of elements (n).

$$\bar{X}_k = \frac{\frac{x_1}{2} + x_2 + \dots + x_{n-1} + \frac{x_n}{2}}{n-1}$$

### **Example:**

On January 1, the set is 40 thousand HUF = NyK (opening stock)

On January 31, the set is 48 thousand HUF

On February 28, the stock is 46 thousand HUF

On March 31, the stock is 44 thousand HUF = ZK (ending stocks)

**Determine the chronological average of the stock!**

# Means

	Unweighted	Weighted
Arithmetic $\bar{x}$	$\bar{x} = \frac{\sum x_i}{n}$	$\bar{x} = \frac{\sum f_i x_i}{n}$
Harmonic $\bar{x}_h$	$\bar{x} = \frac{n}{\sum \frac{1}{x_i}}$	$\bar{x} = \frac{\sum f_i}{\sum \frac{f_i}{x_i}}$
Geometric $\bar{x}_g$	$\bar{x} = \sqrt[n]{\prod x_i}$	$\bar{x} = \sqrt[n]{\prod x_i^{f_i}}$
Quadratic $\bar{x}_q$	$\bar{x} = \sqrt{\frac{\sum x_i^2}{n}}$	$\bar{x} = \sqrt{\frac{\sum f_i x_i^2}{\sum f_i}}$

Ordered magnitude of averages,  
calculated from the same positive values

$$x_{\min} \leq \bar{x}_h \leq \bar{x}_g \leq \bar{x} \leq \bar{x}_q \leq x_{\max}$$

$\bar{x}_h$  and  $\bar{x}_g$  sensitive to extreme low outliers

$\bar{x}$  and  $\bar{x}_q$  sensitive to extreme high outliers

**Example 1:** simple/unweighted averages – the values occur only once each (unique values) or the same number of times

The values to be averaged: 3, 4, 5, 8 – the values occur only once each

(or: 3, 3, 4, 4, 5, 5, 8, 8 – the values occur several times but always the same number of times)

**Example:**

- a) Calculate arithmetic, harmonic, geometric and quadratic averages!
- b) Compare the results received!
- c) Determine the order of averages calculated from the same positive numbers!
- d) If there is an extreme low outlier among the values to be averaged (e.g. 1), which averages are sensitive to it?
- e) Which averages are influenced mostly, if there is an extreme high outlier (e.g. 32) among the values to be averaged?

# Solution

Arithmetic mean

$$\bar{x} = \frac{3 + 4 + 5 + 8}{4} = 5$$

Harmonic average

$$\bar{x}_h = \frac{4}{\frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{8}} = 4.404$$

Geometric average

$$\bar{x}_g = \sqrt[4]{3 \cdot 4 \cdot 5 \cdot 8} = 4.681$$

Quadratic average

$$\bar{x}_q = \sqrt{\frac{3^2 + 4^2 + 5^2 + 8^2}{4}} = \sqrt{\frac{114}{4}} = \sqrt{28,5} = 5.339$$



**Example 2:** (weighted average – the values occur several times but not always the same number of times)

The values to be averaged and the weights belonging to them:

$(x_i)$	Data:	k:	3,	4,	5,	8
$(f_i)$	Frequency:		4,	4,	1,	1

Example:

a) Calculate arithmetic, harmonic, geometric and quadratic averages!

# Solution

Arithmetic mean

$$\bar{x} = \frac{4 \cdot 3 + 4 \cdot 4 + 1 \cdot 5 + 1 \cdot 8}{10} = 4.1$$

Geometric average

$$\bar{x}_g = \sqrt[10]{3^4 \cdot 4^4 \cdot 5^1 \cdot 8^1} = 3.907$$

Harmonic average

$$\bar{x}_h = \frac{10}{\frac{4}{3} + \frac{4}{4} + \frac{1}{5} + \frac{1}{8}} = \frac{10}{2.658} = 3.762$$

Quadratic average

$$\bar{x}_q = \sqrt{\frac{4 \cdot 3^2 + 4 \cdot 4^2 + 1 \cdot 5^2 + 1 \cdot 8^2}{10}} = 4.347$$

## **Example 1.** (unique values)

In a residential area of Budapest the owners of three-room apartments were asked, how much their previous monthly overhead costs were. The following data were generated in thousand HUF.

75, 64, 69, 80, 76, 77, 86, 79, 65, 72, 73, 75, 75, 70

### Example:

Characterize the previous month overhead costs of the owners of three-room apartments with averages that can be used (mean, modus, median)!

# Solution

Arithmetic mean:

$$\bar{X} = \frac{75 + \dots + 70}{14} = 74$$

Average overhead costs of the owners of apartments in the previous month was 74 thousand HUF.

Preparing an order:

64, 65, 69, 70, 72, 73, **75, 75, 75**, 76, 77, 79, 80, 86

Median:

$$\frac{n+1}{2} = \frac{15}{2} = 7,5$$

Me = 75 thousand HUF → previous month overhead costs of half of the owners of apartments were less than 75 thousand HUF, while the other half of the owners of apartments had higher than 75 thousand HUF overhead costs.

Modus:

Mo=75 ezer Ft

Previous month overhead costs of the most apartment owners is 75 thousand HUF.

## Example 2. (equal class intervals)

At a gas station, the daily amount of petrol sold, according to passenger cars was as follows

petrol sold (litre)	number of cars
10 – 19	10
20 – 29	28
30 – 39	42
40 – 49	15
50 – 59	5
Total	100

Task:

Calculate and interpret the average!

Estimate median and modus and describe their meaning!

# Solution

petrol sold (litre)	number of cars	middle of class	cumulative frequency
10 – 19	10	15	10
20 – 29	28	25	38
30 – 39	42	35	80
40 – 49	15	45	95
50 – 59	5	55	100
Összesen	100	---	---

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{10 \cdot 15 + 28 \cdot 25 + \dots + 5 \cdot 55}{100} = 32,7 \text{ litres}$$

The cars were refueled of 32.7 litres on average at the gas station on the given day.

# Solution

Median:

$$S_{me} = \frac{n}{2} = \frac{100}{2} = 50 \quad \text{and} \quad f' \geq 50 \rightarrow Me \text{ is in the 3rd class interval}$$

$$Me = x_{me,0} + \frac{\frac{n}{2} - f'_{me-1}}{f_{me}} h_{me} = 30 + \frac{50 - 38}{42} \cdot 10 = 32,86 \text{ litres}$$

Half of the cars were refueled less than 32.86 liters of petrol, while the other half were refueled more than this amount on the given day.

Modus: is in the 3rd class interval

$$Mo = x_{mo,0} + \frac{k_1}{k_1 + k_2} \cdot h_{mo} = 30 + \frac{(42 - 28)}{(42 - 28) + (42 - 15)} \cdot 10 = 33.41 \text{ litres}$$

The most cars were refueled around 33.41 litres on the given day.

### Example 3. (non-equidistant class intervals)

Mean salaries at a company in 1999

Salaries (thousand HUF)	Headcount
40 – 50	12
50 – 60	20
60 – 80	34
80 – 100	32
100 – 150	14
150 – 200	3
Total	115

Example:

Calculate and interpret the mean!

Estimate median, modus and quartiles and describe their meaning!



# Solution

Only for  
MODUS!

Salaries (thousand HUF) (x)	Headcount (f)	Middle of class (x)	Cumulative frequency (F)	f* New class interval = 20 thousand HUF
40 – 50	12	45	12	24
50 – 60 (Q1),(Mo)	20	55	32	40
60 – 80 (Me)	34	70	66	34
80 – 100 (Q3)	32	90	98	32
100 – 150	14	125	112	5,6
150 – 200	3	175	115	1,2
Összesen	115	---	---	---

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{45 \cdot 12 + 55 \cdot 20 + \dots + 175 \cdot 3}{115} = 75.1 \text{ HUF}$$

$$f^* = \frac{\text{frequency}}{\text{orig. cl. intv. length}} \cdot \text{new cl. intv. length}$$

Company workers earn an average of 75.1 HUF.

# Solution

Median:

$$S_{me} = \frac{n}{2} = \frac{115}{2} = 57,5 \quad (\text{Me is in the 3rd class interval.})$$

$$Me = x_{me,0} + \frac{\frac{n}{2} - f'_{me-1}}{f_{me}} \cdot h_{me} = 60 + \frac{57,5 - 32}{34} \cdot 20 = 75 \text{ thousand HUF}$$

Half of the workers earned less than 75 thousand HUF (while the other half more than this amount) at the given year.

Lower quartile:  $\frac{n}{4} = \frac{115}{4} = 28,7$  (Q1 is found in the 2nd class interval.)

$$Q_1 = x_{q1,0} + \frac{\frac{n}{4} - f'_{q1-1}}{f_{q1}} \cdot h_{q1} = 50 + \frac{28,75 - 12}{20} \cdot 10 = 58.375 \text{ thousand HUF}$$

A quarter of the workers earned less than 58.4 thousand HUF, while three-quarter of them earned more than this amount at the given year.

# Solution

Upper quartile:

$$S_{q3} = \frac{3n}{4} = \frac{3 \cdot 115}{4} = 86,25 \quad (\text{Q3 is in the 4th class interval.})$$

$$Q_3 = x_{q3,0} + \frac{\frac{3 \cdot n}{4} - f'_{q3-1}}{f_{q3}} \cdot h_{q3} = 80 + \frac{86,25 - 66}{32} \cdot 20 = 92,65 \text{eFt}$$

A quarter of the workers earned more than 92.65 thousand HUF (while three-quarter of them earned less than this amount) at the given year.

Modus:

(Mo is in the 2nd class interval.)


$$MO = x_{mo,0} + \frac{k_1}{k_1 + k_2} \cdot h_{mo} = 50 + \frac{(40-24)}{(40-24) + (40-34)} \cdot 10 = 57,270 \text{ HUF}$$

Most of the workers earned 57.27 thousand HUF at the given year.



Always look on the bright side  
of things!

**We finished for today, goodbye!**



ямарваа нэг зүйлийн гэгээлэг  
талыг нь үргэлж олж харцгаая  
өнөөдөртөө ингээд дуусгацгаая, баяртай

让我们总是从光明的一面来看待事物吧！

今天的课程到此结束，谢谢！

دعونا ننظر دائما إلى الجانب المشرق من  
الأشياء!

انتهينا لهذا اليوم، وداعا!