

STATISTICS

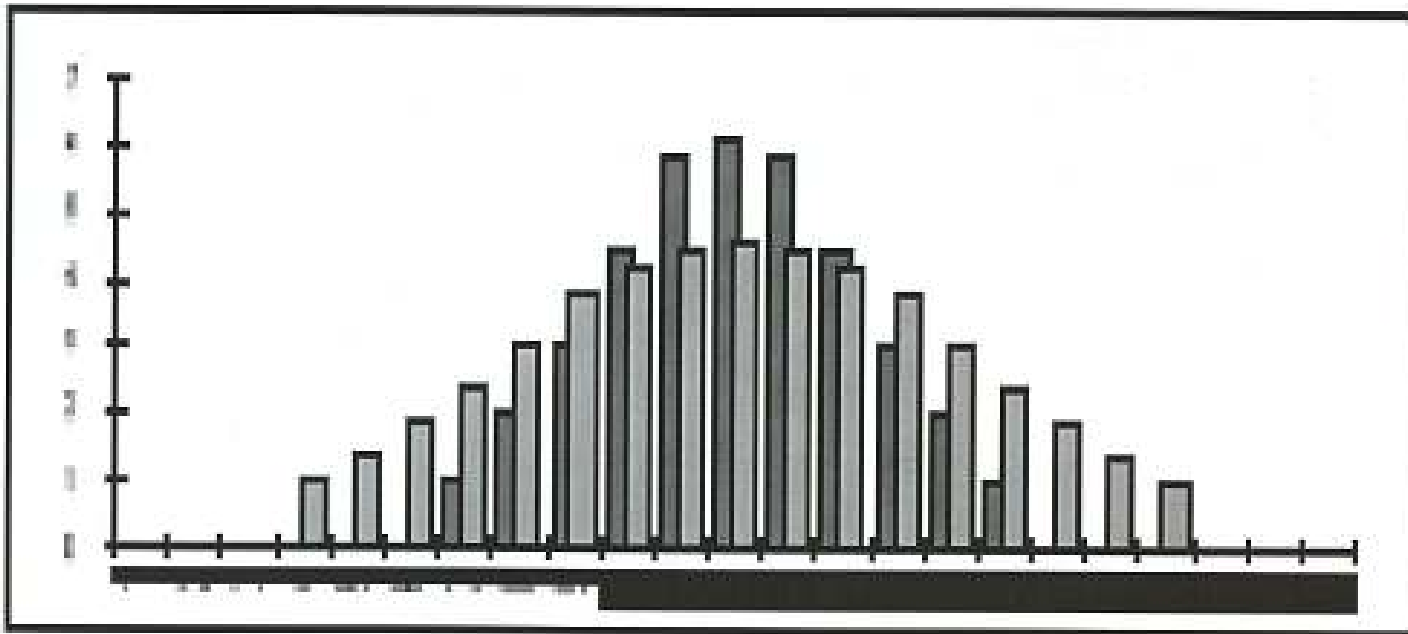
Analysis on quantitative criteria (2)

Dispersion

2) Dispersion

Diversity and variability of the values is called dispersion.

Diversity of the values can be expressed on the one hand by the differences in the values from each other and by a mean value of the deviations, on the other.



Metrics on dispersion

The most important metrics on dispersion:

1. Range (R (or IQR))
2. Mean deviation, δ
3. Standard deviation, σ
4. Relative standard deviation, V
5. Mean difference, G

1. Extremes and range-type inequality indicators

- Maximum
 - Maximum value of the data set (x_{max})
- Minimum
 - Minimum value of the data set (x_{min})
- Range-type inequality indicators are based on them
 - Range (range of standard deviation) $P = x_{max} - x_{min}$
 - Range-ratio (ratio of the range of the data set) $K = \frac{x_{max}}{x_{min}}$
 - Relative range $Q = \frac{x_{max} - x_{min}}{\bar{x}}$

1. Extremes and range-type inequality indicators

- **Interquartile range** **$IQR = Q_3 - Q_1$**

It involves the middle 50% of the range of interpretation.

- **Interdecile range** **$IDR = D_9 - D_1$**

It involves the middle 80% of the range of interpretation.

Metrics of dispersion

2. Mean deviation: arithmetic mean of the differences from average.

It shows that on average how much the criteria differ from the arithmetic mean.

The unit equals to the units of the original data.

Unweighted:

$$\delta = \frac{\sum_{i=1}^n |d_i|}{n}$$

$$d_i = x_i - \bar{x}$$

Weighted:

$$\delta = \frac{\sum_{i=1}^k f_i \cdot |d_i|}{\sum_{i=1}^k f_i}$$

Standard deviation type inequality indicators

- **Non-specific (absolute) indicator (x_i):** unweighted standard deviation;
- **Specific indicator (y_i):** weighted standard deviation;
- Real inequalities can be measured by the **relative standard deviation**;
 - **Non-specific indicator: unweighted** relative standard deviation (unweighted standard deviation in percentage of the mean);
 - **Specific indicator: weighted** relative standard deviation (weighted standard deviation in percentage of the mean);

Unweighted absolute standard deviation: for non-specific indicators

- Square root of the square of the difference of the elements of the data set (x_i) from their mean;

- **Its formula:**

x_i = absolute indicator in region i ;

n = number of elements;

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$$

- **Set of values:** $0 \leq \sigma \leq \infty$

The higher the value σ , the better x_i is dispersed around the mean;

- **Unit:** as the unit of the original values (x_i);

Unweighted relative standard deviation: for non-specific indicators

- Real inequalities can be measured by the relative standard deviation;
- **Relative standard deviation:** for absolute indicators;

- **Its formula:**

σ = standard deviation of the data set x_i
 \bar{x} = mean of the data set x_i

$$v = \frac{\sigma}{\bar{x}} * 100$$

- **Its calculation:**

the standard deviation is divided by the mean and then multiplied by 100 (standard deviation is expressed in percentage of the mean)

- **Set of values:** $0 \leq v \leq \infty$;

The higher the value v , the better x_i is dispersed around the mean;

- **Unit:** % ;

Weighted standard deviation: for specific indicators

- For specific indicators (y_i);
- Square root of the square of the difference of the elements of the data set (y_i) from their weighted mean;

- **Its formula:**

y_i = specific indicator in region i ;

f_i = weight;

$$\sigma = \sqrt{\frac{\sum_i (y_i - \bar{y})^2 f_i}{\sum_i f_i}}$$

- **Set of values:** $0 \leq \sigma \leq \infty$;

The higher the value σ , the better y_i is dispersed around the mean;

- **Unit:** as the unit of the original values (y_i);

Weighted relative standard deviation: for specific indicators

- Real inequalities can be measured by the relative standard deviation;
For specific indicators: with weighted relative standard deviation;

- **Its formula:**

σ = **weighted** standard deviation of the data set y_i ;

\bar{y} = **weighted** mean of the data set y_i ;

$$v = \frac{\sigma}{\bar{y}} * 100$$

- **Its calculation:**

the weighted standard deviation is divided by the weighted mean and then multiplied by 100 (weighted standard deviation is expressed in percentage of the weighted mean);

- **Set of values:** $0 \leq v \leq \infty$;

The higher the value v , the better y_i is dispersed around the mean;

- **Unit:** % ;

Example: Characteristics of the dispersion

- **Data:** 1 2 4 1; arranged in increasing order: 1 1 2 4
- **Range:** max-min=4-1=3
- **Quartiles:**
- **Standard deviation:**

Percentiles

	Percentiles		
	25	50	75
Weighted Average(Definition 1)	1.0000	1.5000	3.5000
Tukey's Hinges	1.0000	1.5000	3.0000

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1-2=-1	1
1	1-2=-1	1
2	2-2=0	0
4	4-2=2	4
Total	0	6

$$SD = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{6}{3}} = \sqrt{2} = 1.414$$

The error of mean (standard error, SE);

- It expresses how reliable is the average calculated from the sample;
- If we could repeat the experiment several times (if it would be an infinite number of measurement series), then it had shown the dispersion of the average of each series around the real mean of the population;
- **Its calculation:** Standard error (SE): $SE = SD / \sqrt{n}$;

;

Standard deviation, or standard error?

- **Standard deviation (SD):** standard deviation of the sample, namely dispersion of the data around the mean;

In case of normal distribution, the interval of mean $\pm 2SD$ comprises 95% of the data;

- **Standard error ($SE=SD/\sqrt{n}$):** reliability of the mean, dispersion of the mean around the (unknown) real mean of the population;

In case of normal distribution, the interval of mean $\pm 2SE$ comprises the real mean with around 95% probability;

Characteristics of the standard deviation (1)

- If the same constant value (a) is added to all x value ($x + a$), the standard deviation remains unchanged;
- If all the x value is multiplied with the same constant (k) number ($k \cdot x$), the standard deviation is changing k -fold;
- The **sum of squares of differences** is the **lowest** when calculating it **with deviations from the average**;
- The **variance** can be written as the **difference between the square of the quadratic mean and the the square of the arithmetic mean**:

$$\sigma^2 = \overline{X}_q^2 - \overline{X}^2$$

Characteristics of the standard deviation (1)

- **Variance** equals to the sum of the **inner variance** (σ_B^2) and **outer variance** (σ_K^2) of the sub-populations;

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

- Its value 0, if $x = \text{constant}$;
- Its threshold

$$0 \leq \sigma \leq \bar{x} \sqrt{N-1}$$

An example for the characteristics of the standard deviation

x_i	$d_i = x_i - \bar{x}$	d_i^2	y_i	$d_i = y_i - \bar{y}$
100	-100	10000	150	-100
150	-50	2500	200	-50
210	+10	100	260	+10
240	+40	1600	290	+40
300	+100	10000	350	+100
Σ 1000	0	24200	1250	0
\bar{x} 200			\bar{y} 250	
		$\sigma^2=4840$		$\sigma^2=4840$
		$\sigma=69,6$		$\sigma=69,6$

An example for the characteristics of the standard deviation

x_i	$d_i = x_i - \bar{x}$	d_i^2	y_i	$d_i = y_i - \bar{y}$	d_i^2
100	-100	10000	110	-110	12100
150	-50	2500	165	-55	3025
210	+10	100	231	+11	121
240	+40	1600	264	+44	1936
300	+100	10000	330	+110	12100
Σ 1000	0	24200	1100		29282
\bar{x} 200			\bar{y} 220		
		$\sigma^2=4840$			$\sigma^2=5856,4$
		$\sigma=69,6$			$\sigma=76,52$

An example for the dispersion and the standard deviation (1)

Water consumption (m ³) (medium value of the interval: x _i)	Number of flats (f _i)	f'
– 15	5	5
15 – 25	17	22
25 – 35	15	37
35 – 45	8	45
45 –	5	50
Total	50	-

$$\sigma = \sqrt{\frac{5(10 - 28,2)^2 + 17(20 - 28,2)^2 + \dots + 5(50 - 28,2)^2}{50}}$$

$$\sigma = 11,26m^3$$

$$V = \frac{\sigma}{x} = \frac{11,26}{28,2} = 0,4$$

An example for the dispersion and the standard deviation (2)

10 14 15 23 58

Fortune of 5 people in mHUF is above.

Calculate the following parameters:

- a) Mean
- b) Mean absolute deviation
- c) Variance, standard deviation, relative standard deviation

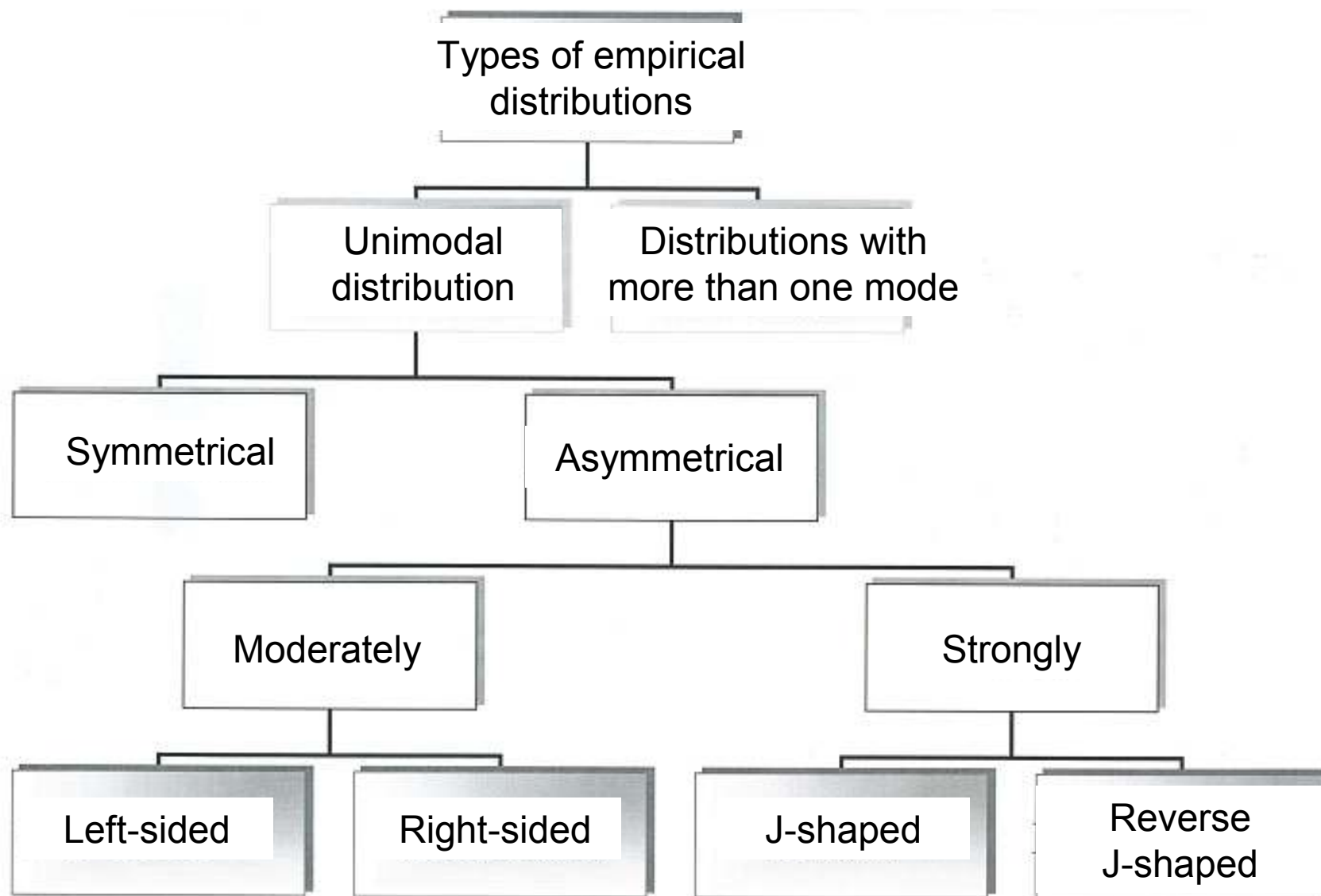
Interprete the results!

Metrics for dispersion

5) Average difference, G (Gini indicator):
Arithmetic average of the absolute deviations of the variants from each other. (It is used mostly when analyzing concentration.)

$$G = \frac{1}{n^2} \cdot \sum_{j=1}^n \sum_{i=1}^n |x_i - x_j|$$

$$G = \frac{1}{n^2} \cdot \sum_{j=1}^k \sum_{i=1}^k f_i \cdot f_j \cdot |x_i - x_j|$$

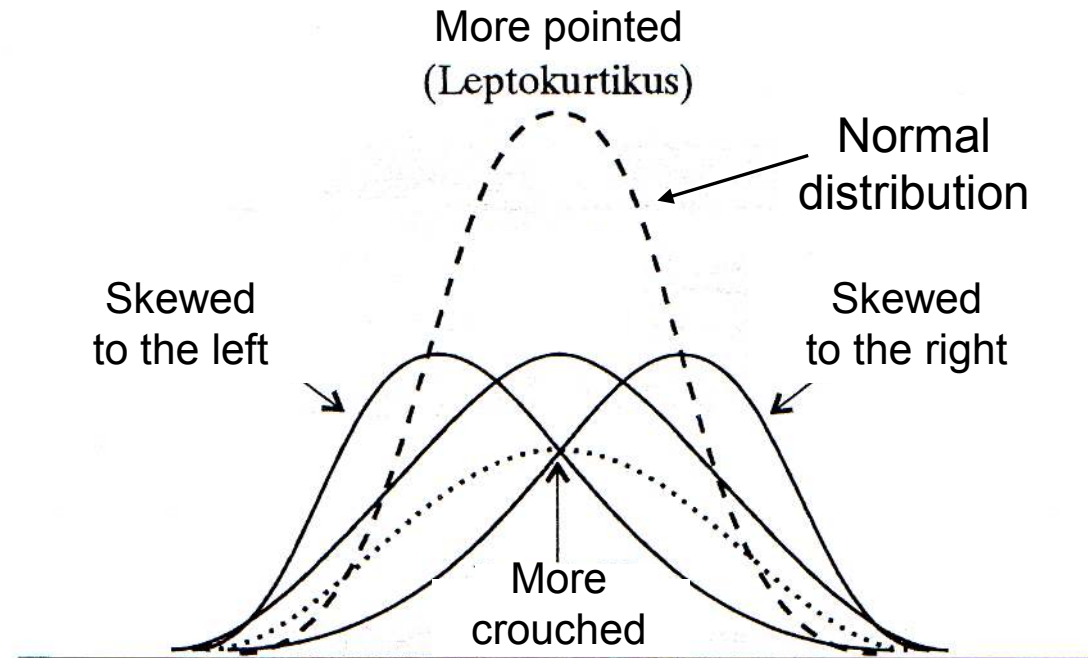


3) Shape indicators

They are used to describe in a solid numerical form that to what context and in what extent a given distribution differs from the frequency curve of the normal distribution.

They are dimensionless indicators.

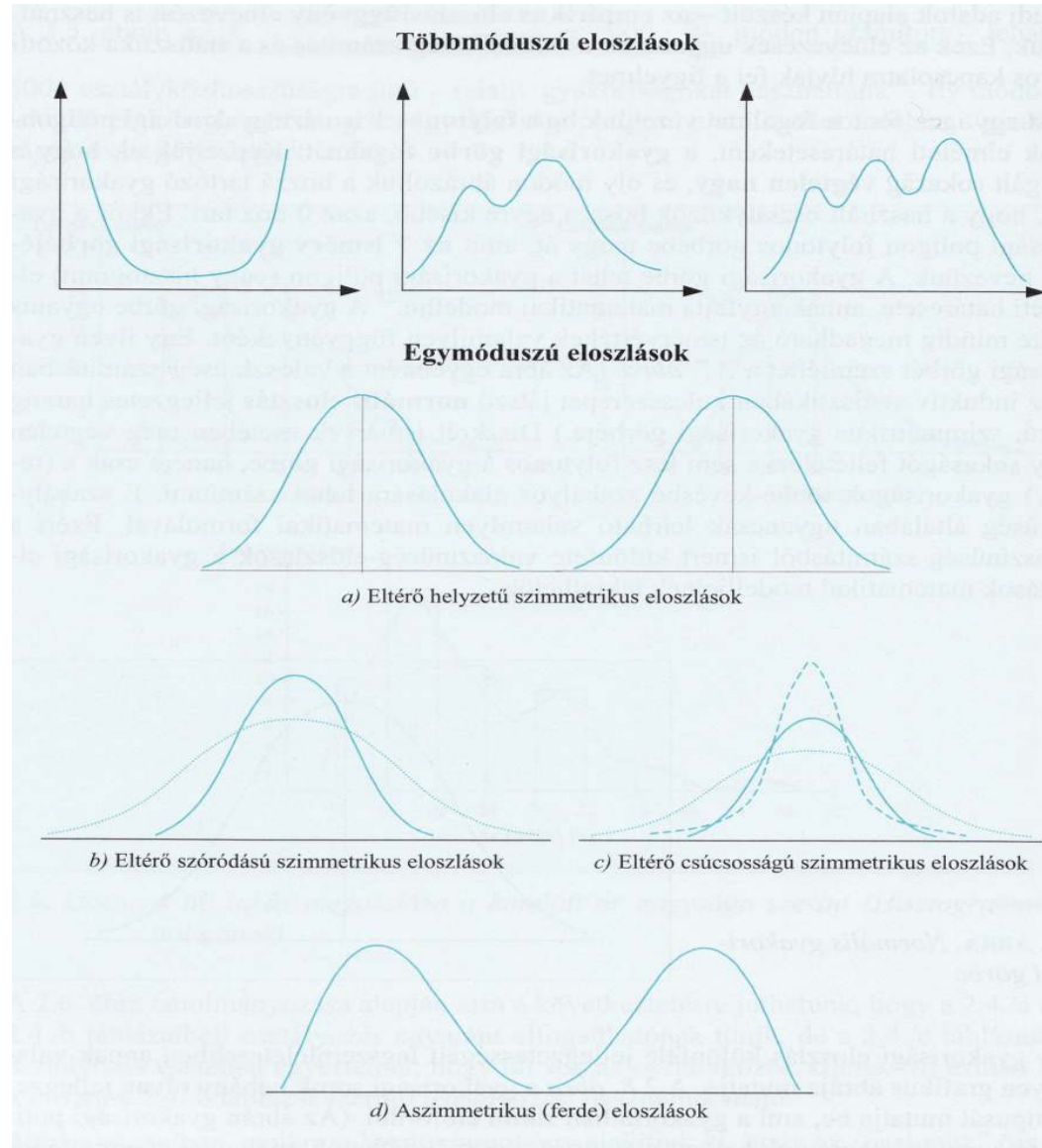
Shape indicators and position indicators



Possible differences of the unimodal frequency distribution from the normal frequency curve

Frequency distributions with different characteristics

Distributions with more than one mode →



Symmetric distributions with →
b) different dispersion
c) different kurtosis

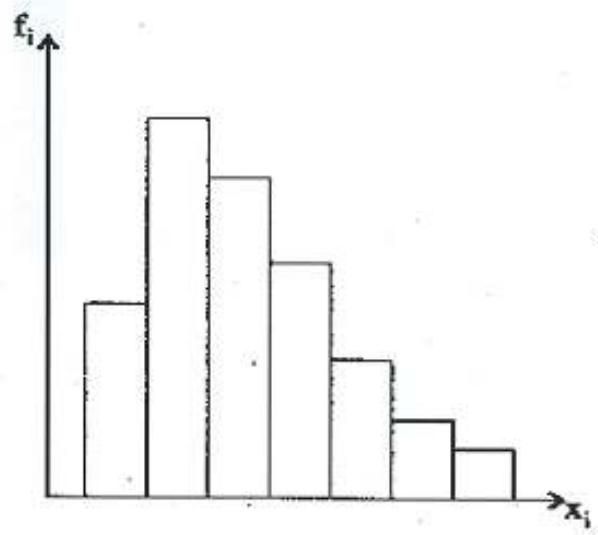
← Unimodal distributions

a) Symmetric distributions with different positions

← Asymmetric distributions

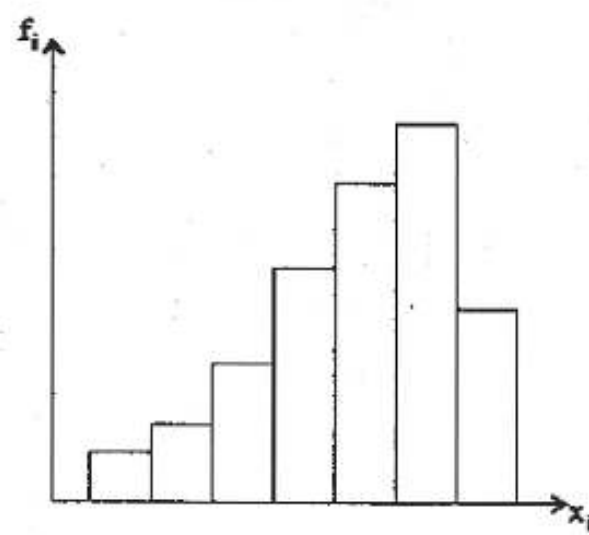
Asymmetric distributions

Left asymmetry



$$M_o < M_e < \bar{x}$$

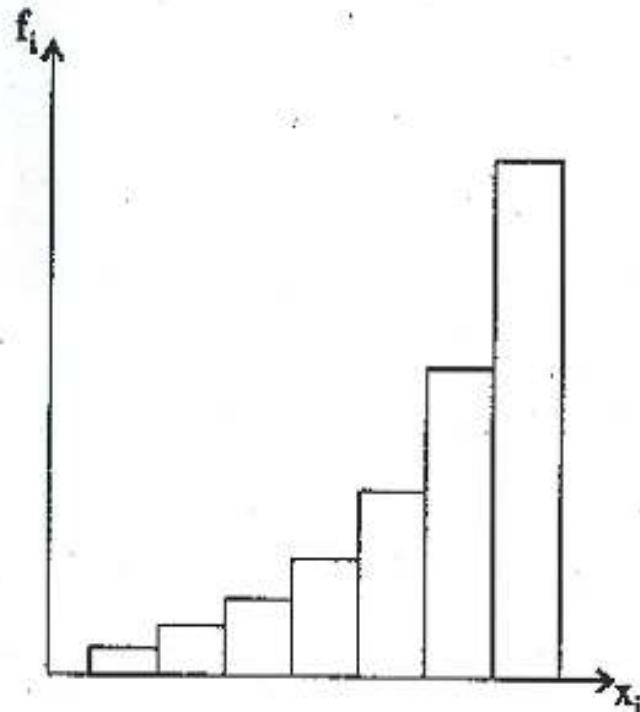
Right asymmetry



$$\bar{x} < M_e < M_o$$

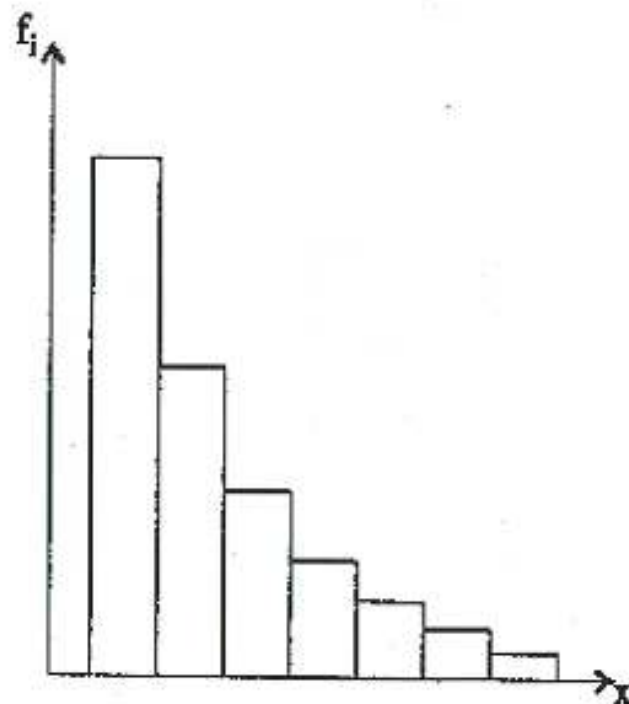
Strongly asymmetric distributions

J-shaped



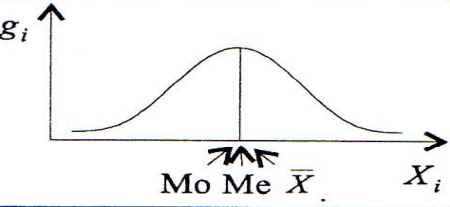
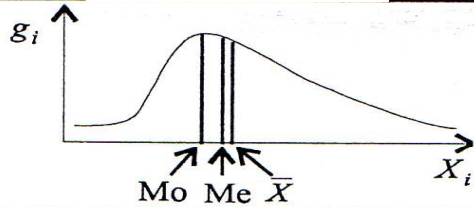
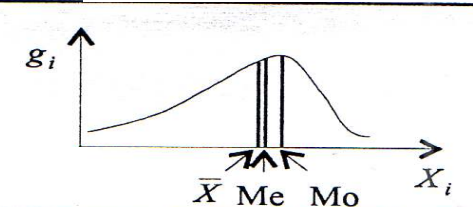
J

Reverse J-shaped (l)



l

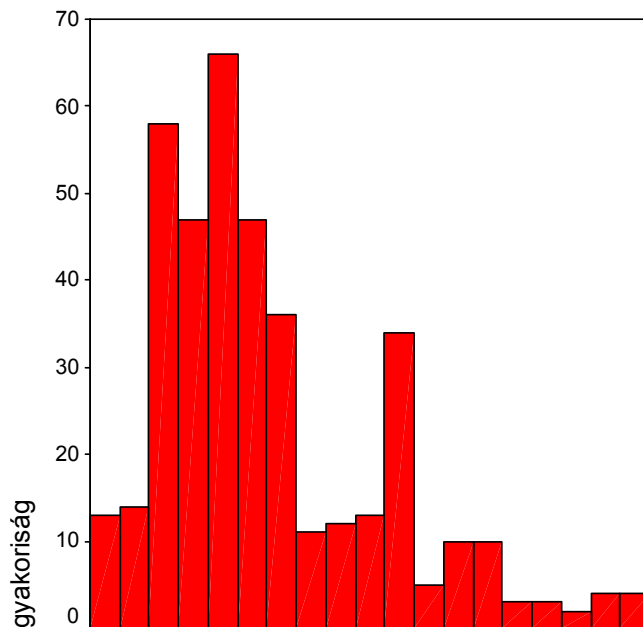
Location indicators for symmetric and asymmetric distributions

symmetric distribution	asymmetric distribution	
	Left asymmetry	Right asymmetry
 <p>A graph showing a symmetric bell-shaped distribution. The vertical axis is labeled g_i and the horizontal axis is labeled X_i. A vertical line marks the center, with arrows pointing to it from the labels Mo, Me, and \bar{X} on the x-axis.</p>	 <p>A graph showing a left-skewed distribution. The vertical axis is labeled g_i and the horizontal axis is labeled X_i. Three vertical lines mark Mo, Me, and \bar{X} on the x-axis, with arrows pointing to them. The order from left to right is Mo, Me, and \bar{X}.</p>	 <p>A graph showing a right-skewed distribution. The vertical axis is labeled g_i and the horizontal axis is labeled X_i. Three vertical lines mark \bar{X}, Me, and Mo on the x-axis, with arrows pointing to them. The order from left to right is \bar{X}, Me, and Mo.</p>
$Mo = Me = \bar{X}$	$Mo < Me < \bar{X}$	$Mo > Me > \bar{X}$
$(Q_3 - Me) = (Me - Q_1)$	$(Q_3 - Me) > (Me - Q_1)$	$(Q_3 - Me) < (Me - Q_1)$

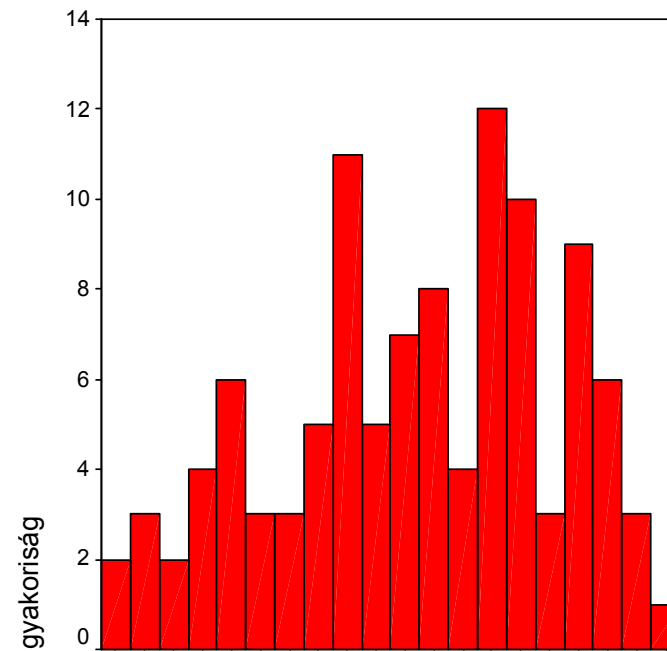
Statistics characterizing the distributions

SKEWNESS

$$s = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^3}$$



Jobbra ferdül, $s > 0$

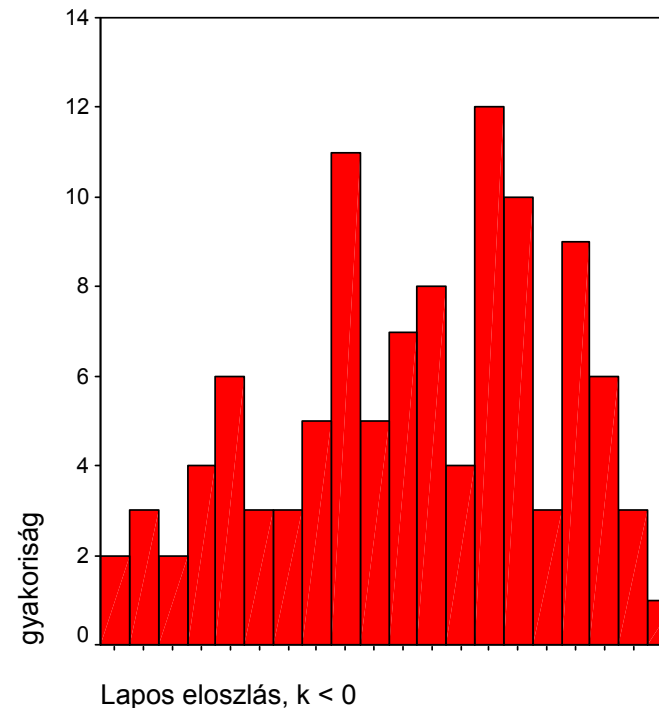
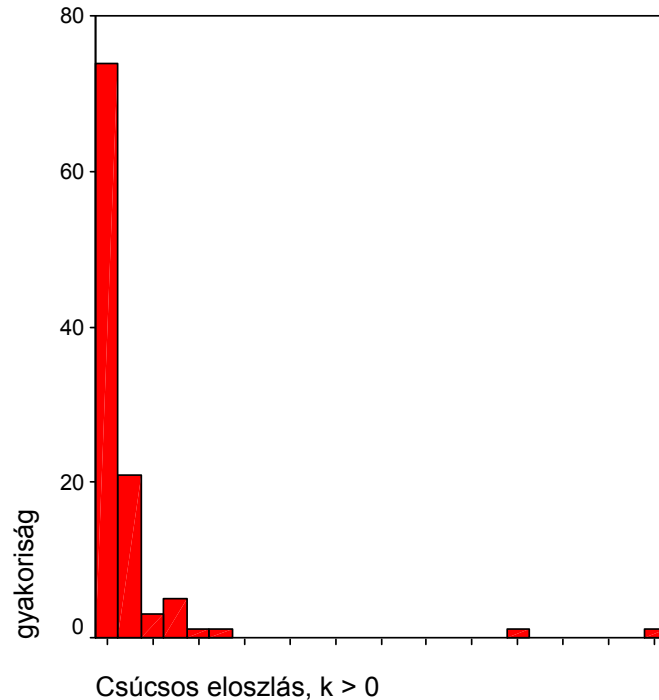


Balra ferdül, $s < 0$

Statistics characterizing the distributions

$$k = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \right)^4} - 3$$

CURTOSIS



Asymmetry indicators (1)

Pearson's **A**-indicator

Its sign shows the direction of asymmetry.

$A > 0$: left asymmetry stretching to right

$A < 0$: right asymmetry stretching to left

$A = 0$: symmetric distribution.

Its absolute value has no upper limit.

- $A > 1$: quite strong asymmetry

$$A = \frac{\bar{x} - Mo}{\sigma}$$

Asymmetry indicators (2)

F-indicator of asymmetry

It is based on quartiles:

$$F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}$$

In case of symmetric distribution: **F = 0**

In case of left asymmetry: **F > 0**

In case of right asymmetry : **F < 0**

$$-1 \leq F \leq 1$$

4) Further methods of analysis

- Concentration
- Time series analysis with means

Concentration

- ❑ **Economic life:** conglomeration, concentration of resources
- ❑ **Statistical criterion:** analysis of a population according to quantity
- ❑ **Concentration:** a significant part of the total value (value amounts) is concentrated on a few elements of the population

Concentration

Concentration can be analyzed through comparing relative frequencies (g_i) and relative value amounts (z_i). If g_i and z_i values belonging to the individual class intervals are identical, this can be interpreted as the lack of concentration, while their difference indicates concentration.

Value amount: $s_i = f_i \cdot x_i$

characteristic to the given group

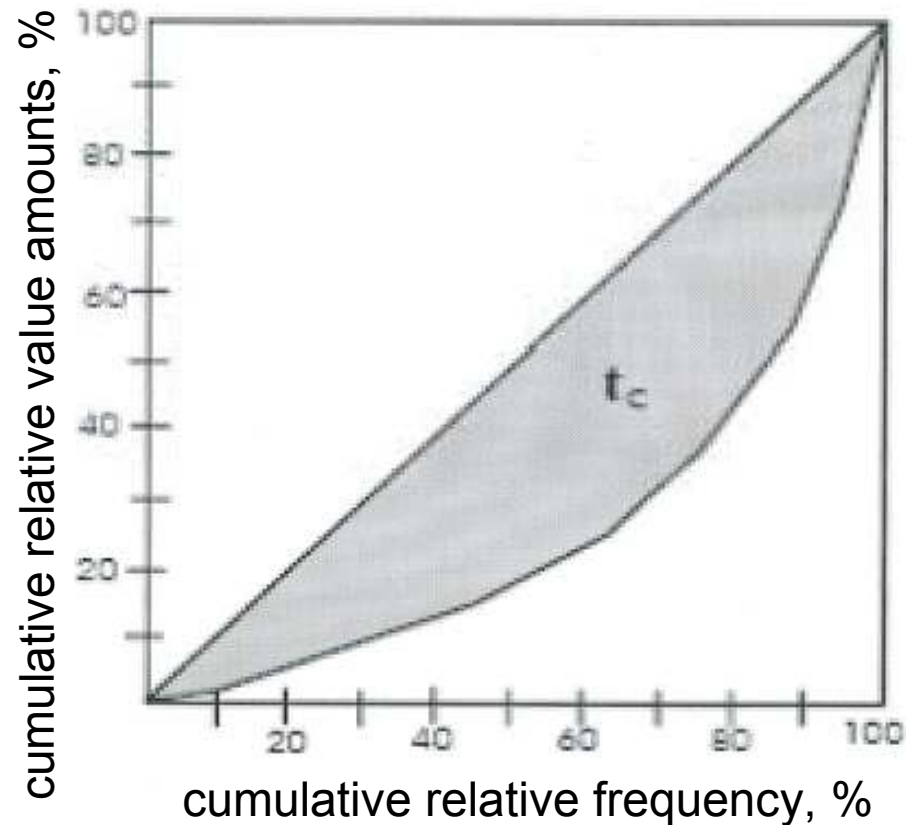
multiplication of x_i value (for class interval frequency series *the middle of the class*) and frequency f_i

Relative value amount: $z_i = \frac{s_i}{\sum s_i}$

Relative frequency: $g_i = \frac{f_i}{\sum f_i}$

Lorenz curve

A figure placed into an unit-sided box, which represents the cumulative relative value amounts as a function of the cumulative relative frequency values.



- If every individual (or, in other words, all subsets of the population) equally share from the value amount, then Lorenz curve will be the diagonal of the square. However, if inequalities are in the population, then the Lorenz curve is running below the diagonal. The more the curve differs downward from the diagonal, the greater the concentration of the given quantity (income, wealth, etc.).

Lorenz curve

- in the absence of concentration the curve coincides with the diagonal
- the farther the curve is from the diagonal, the greater the degree of concentration

Application

- illustrate of relative concentration
- interpolation
- comparison of the concentration of several criteria
- temporal or spatial comparison of the concentration of a given criterion

Concentration coefficient

The area enclosed by the Lorenz curve and the diagonal is called concentration area.

If the concentration area is related to the area of the triangle then, based on the quotient, we can conclude to the degree of concentration. The ratio of the concentration area is measured with the concentration coefficient.

$$K = \frac{G}{2\bar{x}}$$

where G is the mean difference (Gini indicator)

Range of interpretation for K is the interval [0; 1]. In the lack of concentration K=0, and the closer K is to 1, the stronger the concentration.

Concentration

- ❑ Absolute concentration: concentrates only to a few units of the value of amounts (e.g. energy industry, automobile manufacturing)
- ❑ Relative concentration: the value of amounts, in relative sense, concentrates only to a few units (e.g. personal income)

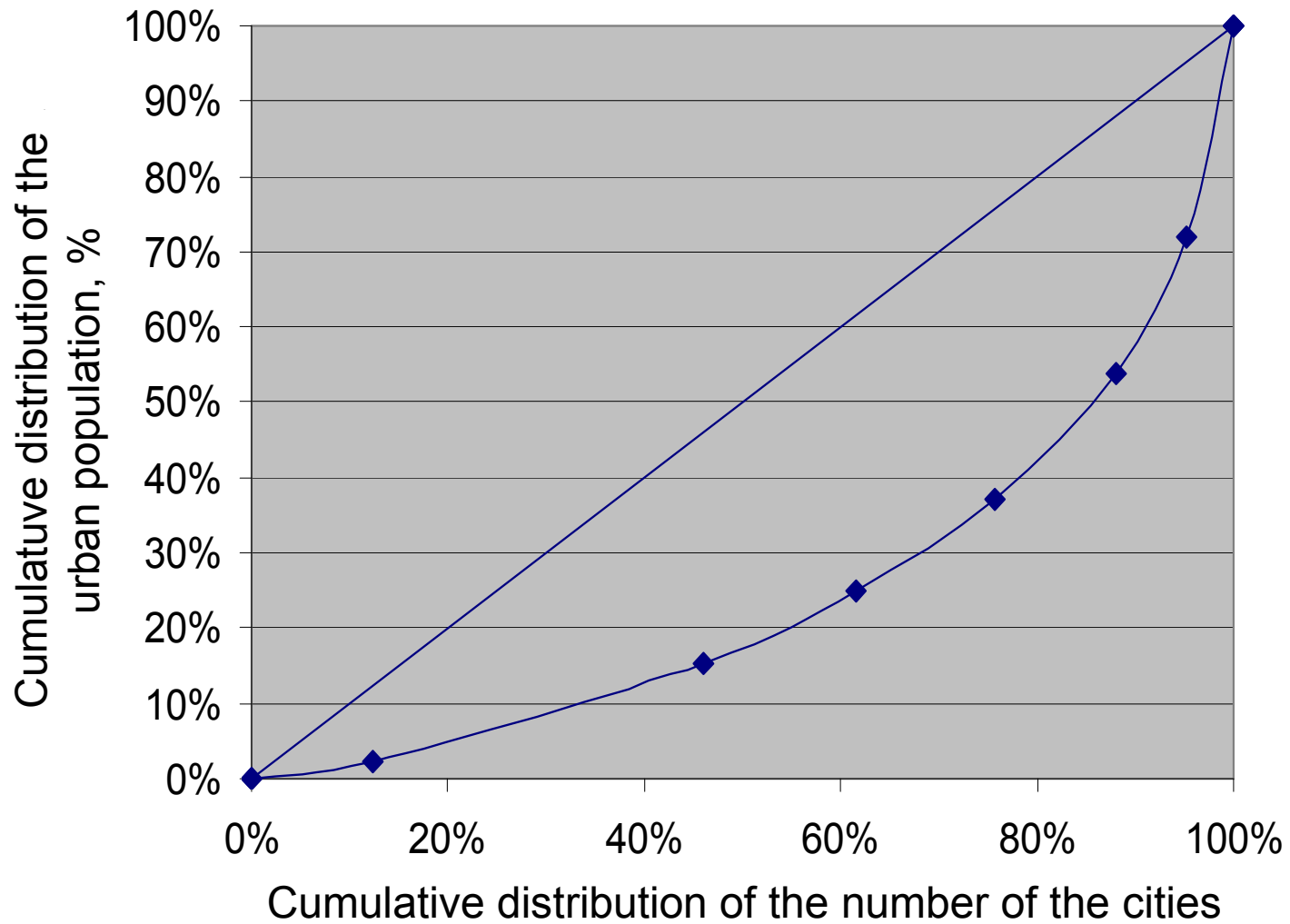
Concentration

VAUE OF AMOUNT (s)	POPULATION (n)
capital, property, production, sales, earnings	business organizations
export, import	countries, products, business organizations
agricultural land, assets, livestock	business organizations, owners
household income, property	inhabitants, residential households

Example:

Number of population	Length of class interval	Cities			
		no.	cumulative distribution, number, %	population	cumulative distribution, population, %
- 5999	?	24	12.40	109 216	2.40
6000-11999	6000	65	46.10	580 133	15.30
12000-17999	6000	30	61.60	430 350	24.90
18000-23999	6000	27	75.60	546 159	37.00
24000-39999	16000	24	88.00	751 171	53.70
40000-79999	40000	14	95.30	821 709	72.00
80000-	?	9	100.00	1 261 569	100.00
Total:		193		4 500 307	

Lorenz curve of the urban population in Hungary



Example:

Is there concentration in the distribution of the population in cities in Hungary excluding Budapest?

Distribution of the population - cities (population at the end of 1997)

Population	Number of cities
2 000–4 999	20
5 000–9 999	61
10 000–49 999	105
50 000–99 999	11
100 000–	8
Total:	205

Source: Statistical Yearbook of Hungary, 1997

Calculate data of the relative frequency and relative value of amount series!

relative frequencies:

Population	No. of cities (f_i)	Relative frequency, % (g_i)
2 000–4 999	20	9.8
5 000–9 999	61	29.8
10 000–49 999	105	51.1
50 000–99 999	11	5.4
100 000–	8	3.9
Total:	205	100.0

Source: Statistical Yearbook of Hungary, 1997

Work table of the relative values of amounts

Population	No. of cities (f_i)	Middle of the class interval (x_i)	Value of amounts, population ($f_i x_i$)	Relative value of amounts % (z_i)
2 000–4 999	20	3 500	70 000	1.3
5 000–9 999	61	7 500	457 500	8.3
10 000–49 999	105	30 000	3 150 000	57.2
50 000–99 999	11	75 000	825 000	15.0
100 000–	8	125 000	1 000 000	18.2
Összesen:	205	–	5 502 500	100.0

Cumulative relative frequencies and values of amounts

Population	Cumulative relative	
	Frequency, % (g'_i)	Value of amounts, % (z'_i)
2 000–4 999	9.8	1.3
5 000–9 999	39.6	9.6
10 000–49 999	90.7	66.8
50 000–99 999	96.1	81.8
100 000–	100.0	100.0

The distribution of savings deposits

Funds deposited (thousand HUF)	The number of deposit books
- 100	18
101 - 500	15
501 - 1000	9
1 001 - 3 000	10
3 001 - 5 000	5
5 001 -	3
Total	60

Illustrate the concentration of deposits!

Determine the concentration coefficients!

Deposit (thousand HUF)	The number of deposit books	X_i	g_i (%)	S_i	Z_i (%)	g_i^l	Z_i^l
0 - 100	18	50	30,0	900	1,3	30,0	1,3
101 - 500	15	300	25,0	4 500	6,4	55,0	7,7
501 - 1000	9	750	15,0	6 750	9,6	70,0	17,3
1 001 - 3 000	10	2 000	16,7	20 000	28,5	86,7	45,8
3 001 - 5 000	5	4 000	8,3	20 000	28,5	95,0	74,3
5 001 - 7 000	3	6 000	5,0	18 000	25,7	100,0	100,0
Total	60		100,0	70 150	100,0		

If the cumulative relative frequency values are significantly higher than the cumulative relative values of amounts \Rightarrow the concentration of the urban population can be observed.

The concentration is strong when a large proportion of the population belongs to a small proportion of the total amount of values, and vice versa.

If the share of the units from the values of amounts is the same \Rightarrow cumulative relative frequencies and cumulative relative values of amounts are equal ($g_i = z_i$). This indicates the lack of concentration.

\Rightarrow The curve coincides with the diagonal of the square.

In case of total concentration the curve coincides with the coordinate axes.

The area enclosed by the curve and the diagonal characterizes the relative size of concentration.

Example:


Gasoline consumption (l/100 km)	number of cars
3.0 – 6.0	8
6.1 – 10.0	9
10.1 – 20.0	3
Total:	20

Is there concentration in gasoline consumption of the cars belonging to each category?



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!