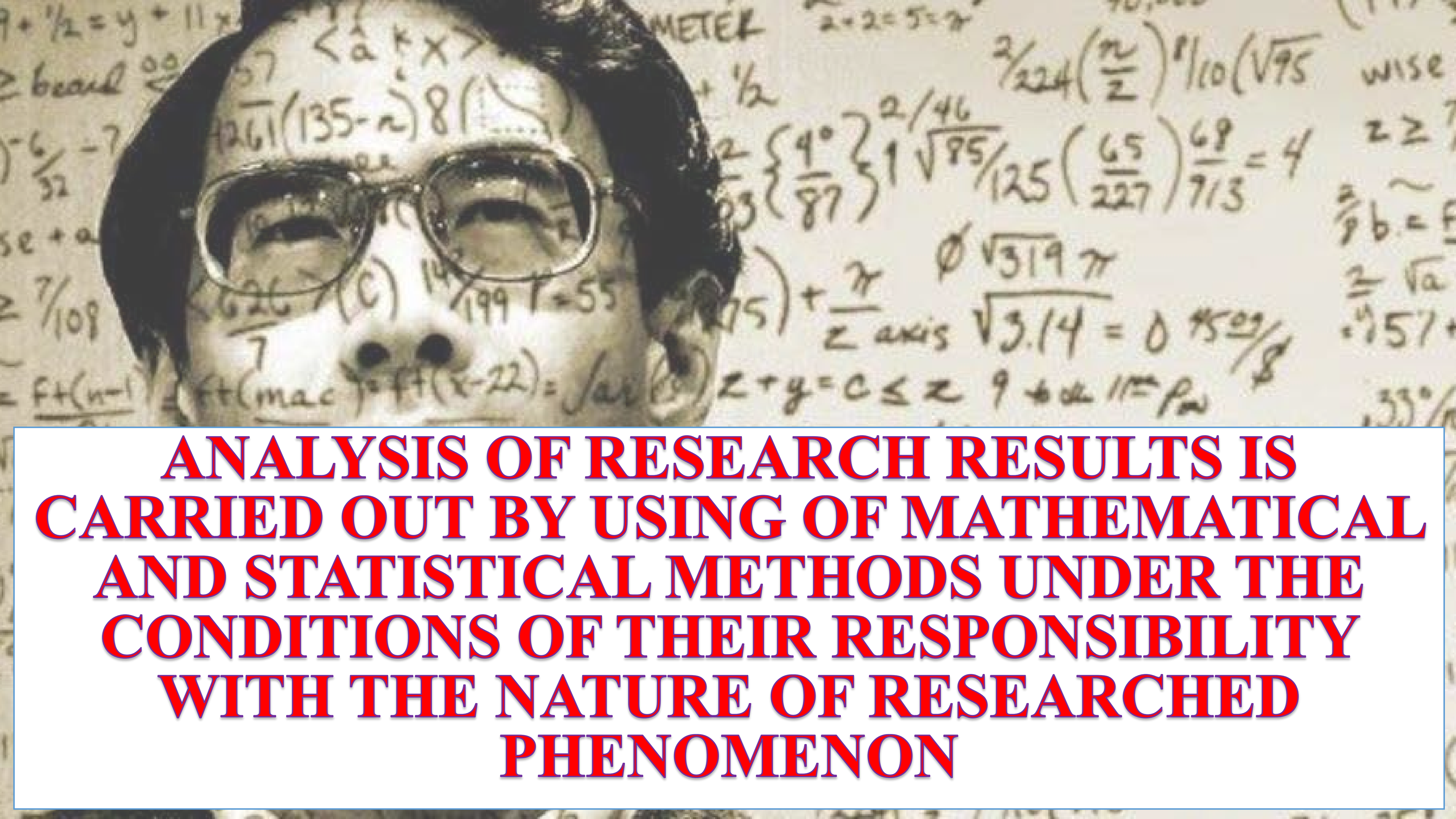
DATA ANALYSIS METHODS: AVERAGE VALUES EVALUATION OF THE RELIABILITY OF THE RESULTS OF THE STUDY. CHARACTERISTICS AND ANALYSIS OF STATISTICAL ERRORS.

Lecturer: Maxim V. Khorosh, PhD



ANALYSIS OF RESEARCH RESULTS IS CARRIED OUT BY USING OF MATHEMATICAL AND STATISTICAL METHODS UNDER THE CONDITIONS OF THEIR RESPONSIBILITY WITH THE NATURE OF RESEARCHED PHENOMENON

THE NEXT GROUPS OF METHODS ARE DISTINGUISHED:

1. Methods of calculating generalizing coefficients that characterize the different aspects of each of the features of the program:

- methods of calculating of the **relative values**
- methods of calculating of the **average values**
- methods for **assessing the reliability** of relative and average values

2. Methods of comparing of the different statistical aggregates:

- methods for assessing the reliability of the differences **of generalizing coefficients**;
- methods for assessing the reliability of differences **in the distribution of characteristics**;
- methods of **standardization** of generalizing coefficients.

3. Methods of differentiation, assessment of interaction and integration of factors.

analysis of variance

correlation analysis

regression analysis

factor analysis

principal components method

discriminant analysis

sequential analysis

allow to solve the following tasks:
a) to decompose a multifactorial complex into constituent factors, highlighting important and insignificant;
b) to study the interaction of factors
c) obtain an integrated assessment based on a set of factors.

4. Methods of analysis of the dynamics of phenomena (analysis of dynamic or time series).

STAGES OF DESCRIPTION OF QUANTITATIVE SIGN

1. Determining the type of distribution of the sign.

2. Estimation of the central tendency of the studied population.

3. Assessment of its diversity (spread).

TO CORRECTLY CHOOSE THE PATH OF STATISTICAL ANALYSIS, IT IS NECESSARY TO KNOW THE TYPE OF DISTRIBUTION OF THE RESEARCH SIGN.

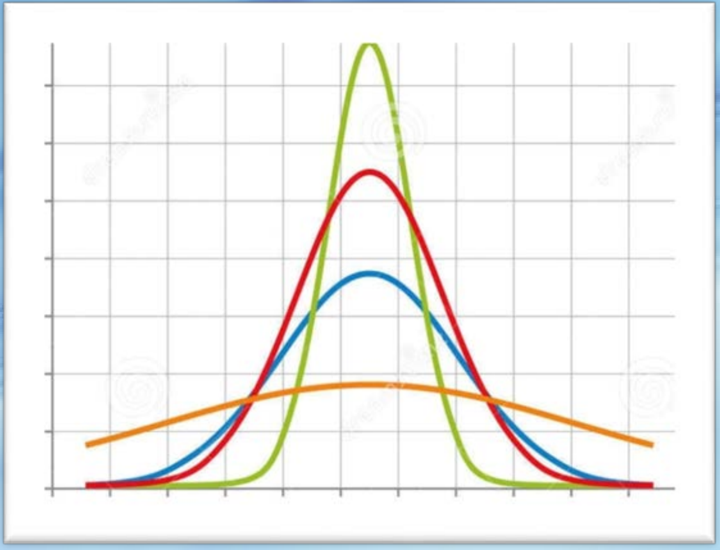
Under the **type of distribution** of a random variable means the correspondence that is established between all possible numerical values of a random variable and the possibility of their occurrence in the aggregate.

The analysis of the statistical aggregate begins with the establishment of the type of distribution of the studied trait.

⬇

The obtained data are presented in the form of a variation series, depicted graphically and appropriate calculations are made.

- In the case of a **distribution close to normal**, **parametric statistics** are used for further statistical analysis,
- If the **distribution is different from normal or with an unknown distribution**, it is recommended to use **non-parametric statistics**.

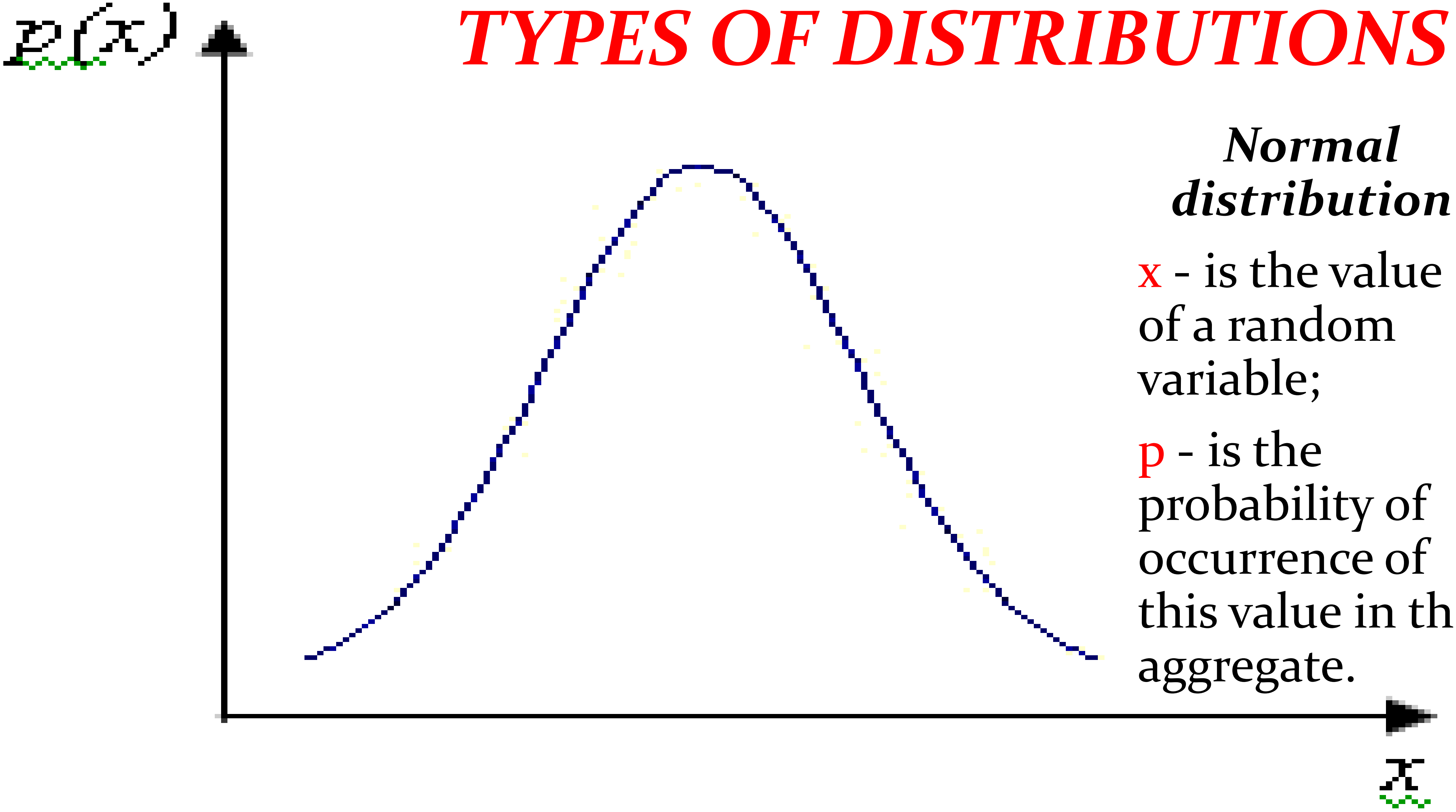


TYPES OF DISTRIBUTIONS

NORMAL (Gaussian distribution) - describes the combined effect on the studied phenomenon of a small number of randomly combined factors (compared to the total sum of factors), the number of which is infinitely large.

- Occurs most often in nature, for which it is called "normal".
- Characterizes the distribution of continuous random variables.

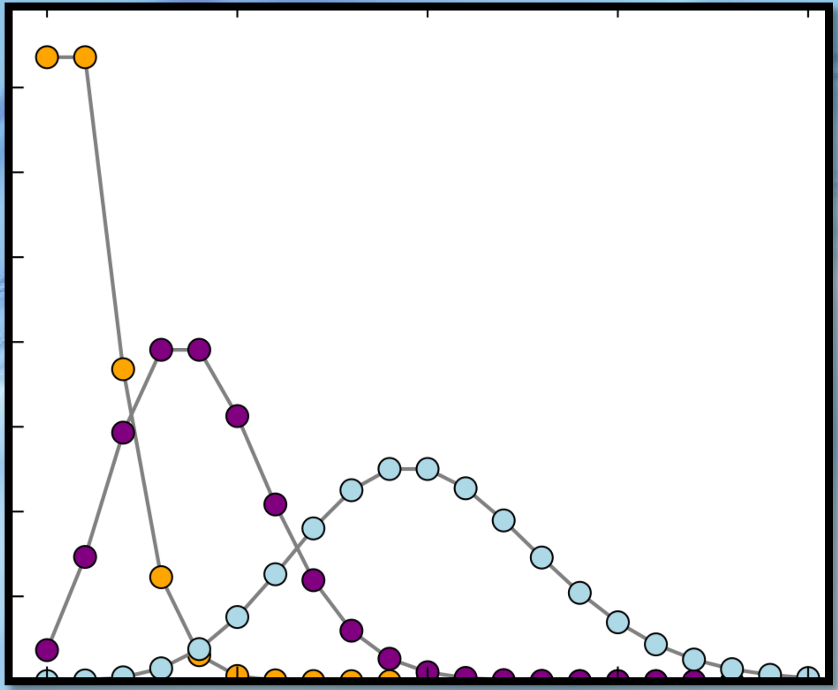
TYPES OF DISTRIBUTIONS



Normal distribution

x - is the value of a random variable;

p - is the probability of occurrence of this value in the aggregate.

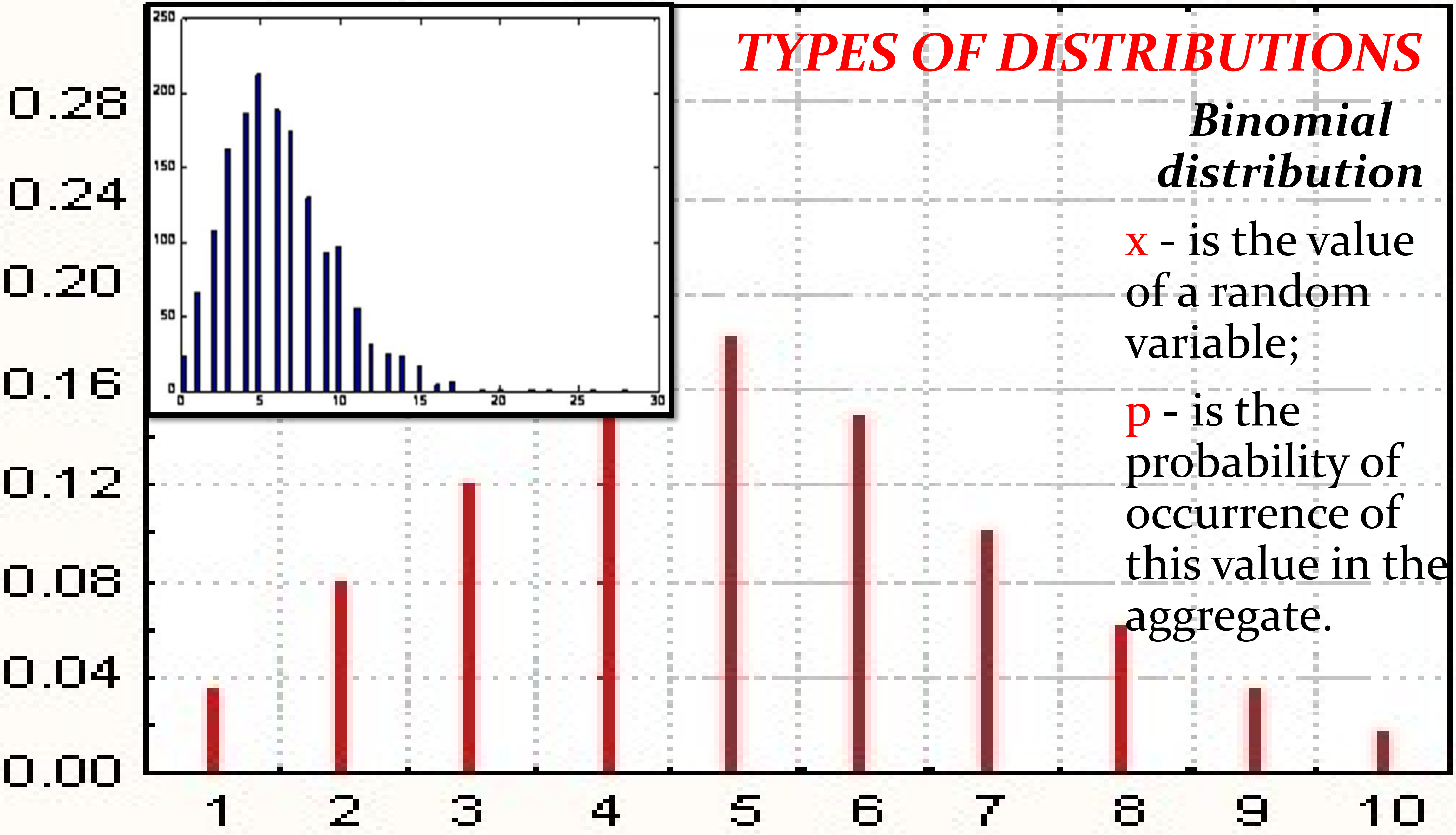


TYPES OF DISTRIBUTIONS

BINOMIAL (Bernoulli distribution)

- Intermediate type, with a large number of tests tends to normal.
- The binomial distribution characterizes the distribution of discrete random variables.

TYPES OF DISTRIBUTIONS

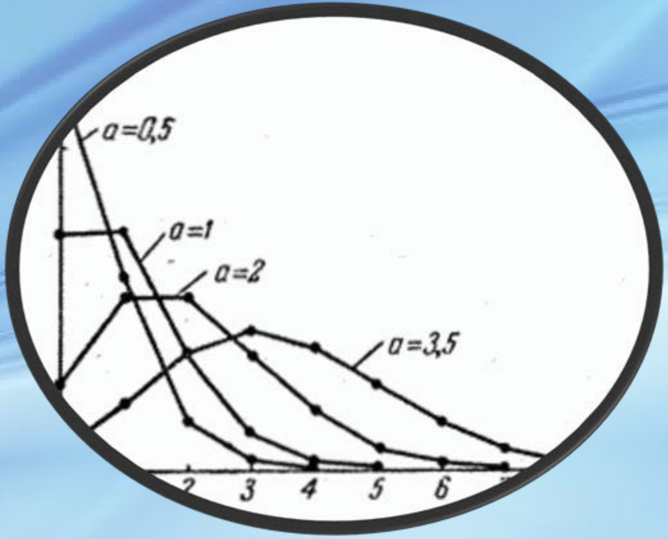


Binomial distribution

x - is the value of a random variable;

p - is the probability of occurrence of this value in the aggregate.

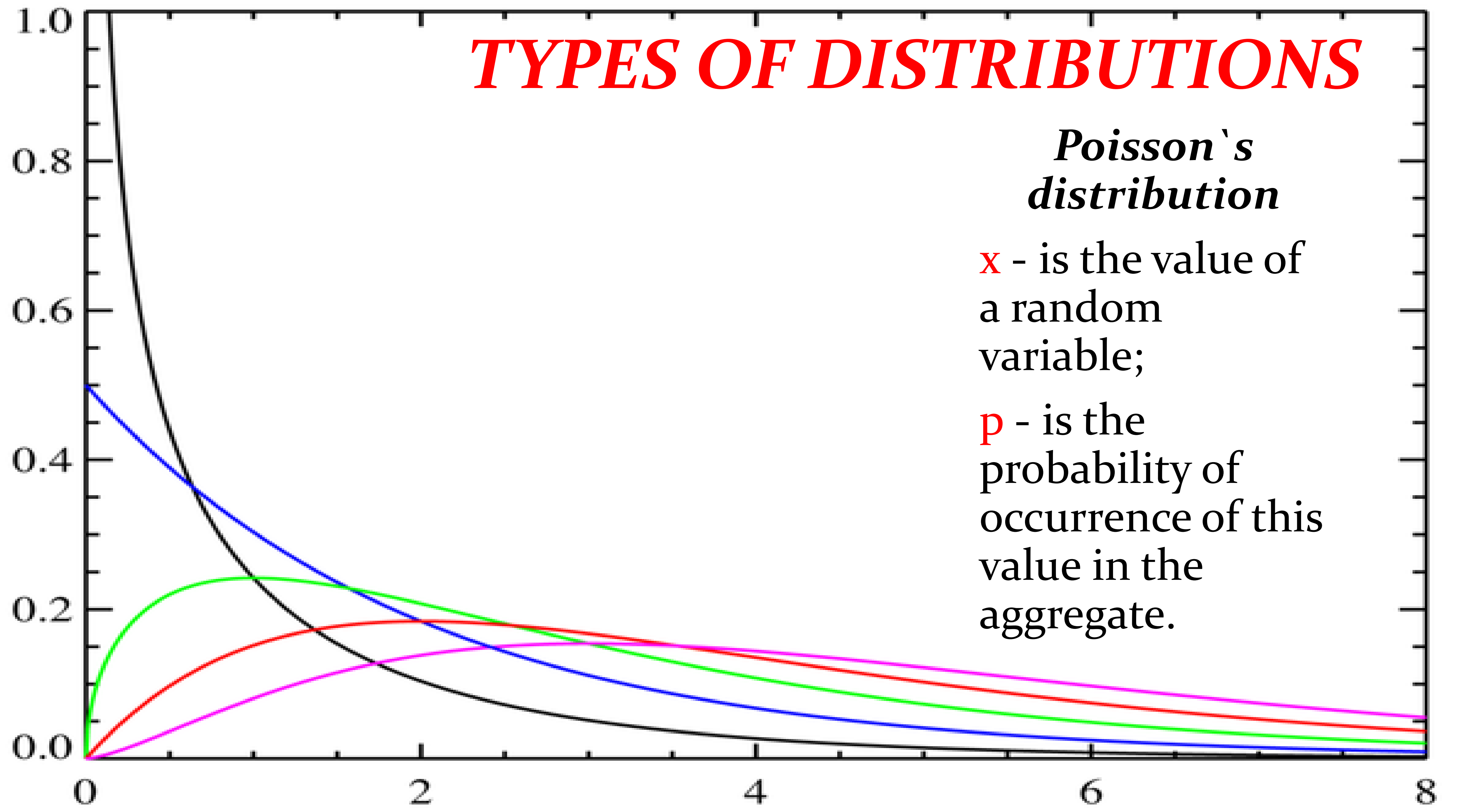
TYPES OF DISTRIBUTIONS



POISSON`s DISTRIBUTION - describes events in which with increasing value of a random variable, the probability of its occurrence in the aggregate decreases sharply.

- The Poisson`s distribution is characteristic of rare events and can also be considered as an extreme variant of binomial.
- Characterizes the distribution of discrete and specific random variables.

TYPES OF DISTRIBUTIONS



Poisson`s distribution

x - is the value of a random variable;

p - is the probability of occurrence of this value in the aggregate.

Variatian row (frequency table) - it is a disordered or ranked (ordered) in ascending or descending order of a number of statistical frequencies.

The variation series by its design has 2 characteristics:

- the value of the sign - **variance** x_i, $i = 1,2,\ldots,m$;
- number of cases of variance: absolute - frequencies p_i (f_i), relative - frequencies w_i (relative shares of frequencies in the total sum of frequencies).

$$\frac{x}{f} : \frac{x1}{f1} - \frac{\ldots}{\ldots} - \frac{x\infty}{f\infty}$$

TYPES OF VARIATION ROWS :

1. Depending on the type of random variable:

- **discrete;**
- **continuous.**

2. Depending on the group option:

- **not grouped;**
- **grouped (interval).**

3. Depending on the frequency with which each variant occurs in the variation series:

- **simple (p = 1);**
- **suspended (p> 1).**

Example: in a variation row of systolic blood pressure measured in ten patients:

- **110, 140, 120, 130, 120, 170, 130, 140, 130, 160;**
  - **110, 120, 120, 130, 130, 130, 140, 140, 160, 170;**

variances are only 6 values :

- **110, 120, 130, 140, 160, 170.**

The frequencies will take the following values:

- for variance 110 frequency P = 1 (value 110 occurs in one patient)
- for variance 120 frequency P = 2 (value 120 occurs in two patients)
- for variance 130 frequency P = 3 (value 130 occurs in three patients)
- for variance 140 frequency P = 2 (value 140 occurs in two patients)
- for variance 160 frequency P = 1 (value 160 occurs in one patient)
- for variance 170 frequency P = 1 (value 170 occurs in one patient)

110, 140, 120, 130, 120, 170, 130, 140, 130, 160

110, 120, 120, 130, 130, 130, 140, 140, 160, 170

110	120	130	140	160	170
1	2	3	2	1	1

...−120	121–140	141−160	161−...
3	5	2	1

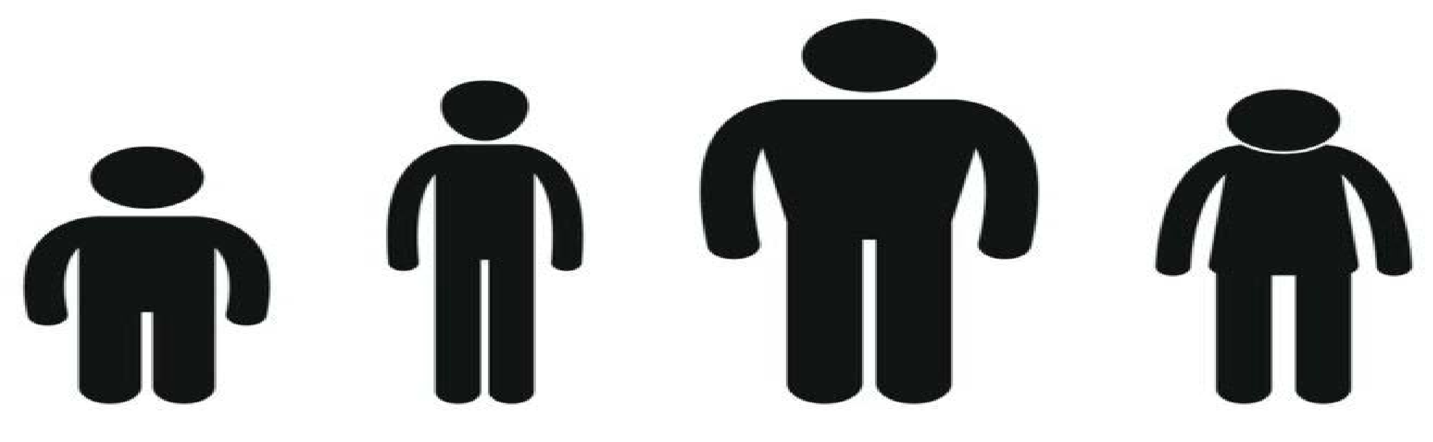
The variation series can be divided into separate (possibly equal) parts, which are called **quantiles**.

The following quantiles are most often used :

- Median - 2 parts
- Tertsil - 3 parts
- Quarter - 4 parts
- Decile - 10 parts
- Percentile - 100 parts

AVERAGE VALUES

GENERAL CONCEPT OF AVERAGE VALUES



The average value is a generalized quantitative characteristic of the population on the basis of the studied feature in specific conditions of place and time.

The average value - generalizing characteristics of the set of individual values of some quantitative feature.

The average value reflects the common and typical that is inherent in the units of this population.

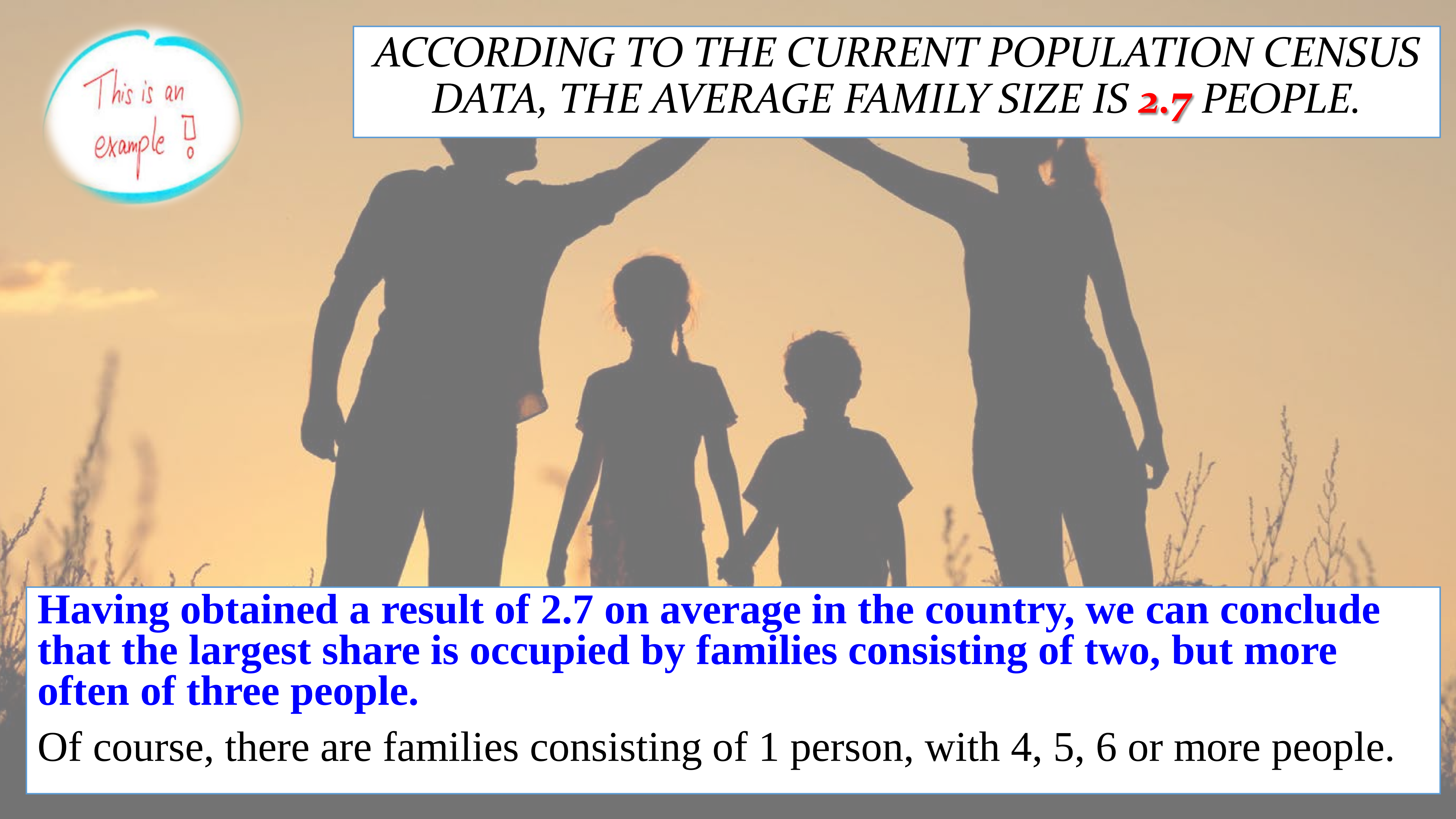
In the average values, individual deviations corresponding to individual units of the population are repaid.

For the mean to make sense, it must be calculated for a homogeneous population.

Using the average, we can single-handedly describe the phenomenon under study.

This is an example !

ACCORDING TO THE CURRENT POPULATION CENSUS DATA, THE AVERAGE FAMILY SIZE IS **2.7** *PEOPLE.*



Having obtained a result of 2.7 on average in the country, we can conclude that the largest share is occupied by families consisting of two, but more often of three people.

Of course, there are families consisting of 1 person, with 4, 5, 6 or more people.

Necessary conditions for calculating the average value - qualitative homogeneity of the population: all units of the population must have the studied feature.

It is impossible to calculate the average scholarship in Poltava, because not all residents of Poltava, and not even all students living in the city, receive the same scholarship. The same can be said about pensions in Kyiv or salaries in Kharkiv.

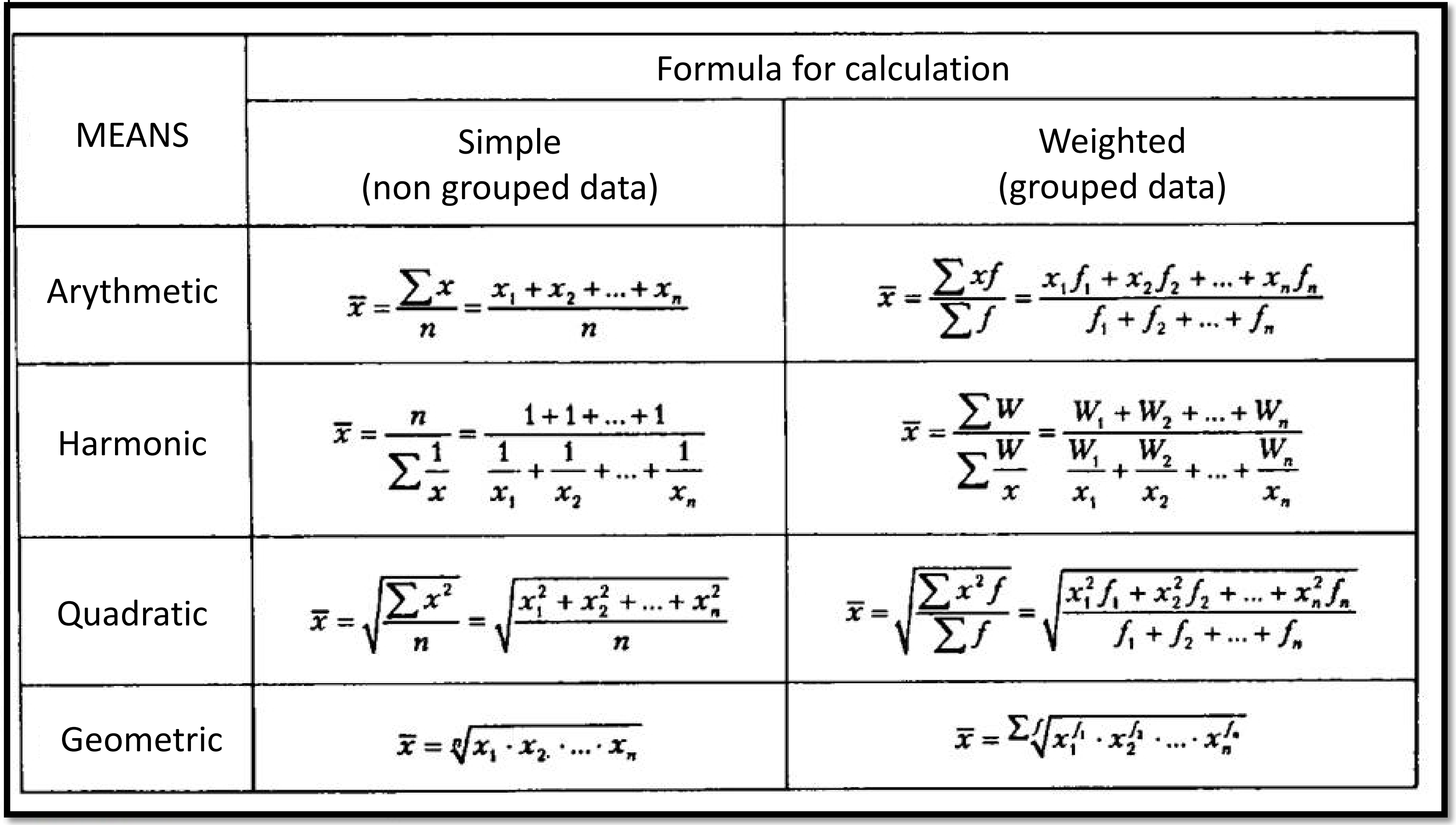
Regarding such a statistical population as the population of a locality, it is more correct to talk about the average income per capita.

The average scholarship (salary, pension) can be calculated only among those who receive a scholarship (salary, pension).

AVERAGES

MATHEMATICAL

STRUCTURAL



MEANS	Formula for calculation	
	Simple (non grouped data)	Weighted (grouped data)
Arythmetic	$\bar{x}=\frac{\sum x}{n}=\frac{x_1+x_2+\ldots+x_n}{n}$	$\bar{x}=\frac{\sum xf}{\sum f}=\frac{x_1f_1+x_2f_2+\ldots+x_nf_n}{f_1+f_2+\ldots+f_n}$
Harmonic	$\bar{x}=\frac{n}{\sum\frac{1}{x}}=\frac{1+1+\ldots+1}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}}$	$\bar{x}=\frac{\sum W}{\sum\frac{W}{x}}=\frac{W_1+W_2+\ldots+W_n}{\frac{W_1}{x_1}+\frac{W_2}{x_2}+\ldots+\frac{W_n}{x_n}}$
Quadratic	$\bar{x}=\sqrt{\frac{\sum x^2}{n}}=\sqrt{\frac{x_1^2+x_2^2+\ldots+x_n^2}{n}}$	$\bar{x}=\sqrt{\frac{\sum x^2f}{\sum f}}=\sqrt{\frac{x_1^2f_1+x_2^2f_2+\ldots+x_n^2f_n}{f_1+f_2+\ldots+f_n}}$
Geometric	$\bar{x}=\sqrt[n]{x_1\cdot x_2\cdot\ldots\cdot x_n}$	$\bar{x}=\sqrt[\sum f]{x_1^{f_1}\cdot x_2^{f_2}\cdot\ldots\cdot x_n^{f_n}}$

MODA

MEDIAN

TO CALCULATE THE DEGREE (MATHEMATICAL) AVERAGE IT IS NECESSARY TO USE ALL THE AVAILABLE VALUES OF THE SIGN

- **Simple means** - are used provided that the frequency of the option is equal to 1.
- **Weighted means** are called values that take into account that some variants of the values of the feature may have a different number, and therefore each option has to be multiplied by this number.

ARITHMETIC MEAN

$$\bar{x} = \frac{\sum x_i}{n}$$

simple arithmetic mean

Used when the calculation is performed on ungrouped data

$$\bar{x} = \frac{\sum x_i \cdot f_i}{\sum f_i}$$

weighted arithmetic mean

Used when data is presented in the form of distribution series or groupings

EXAMPLE OF SIMPLE ARITHMETIC MEAN

The student passed 4 exams and received the following grades: 3, 4, 4 and 5.

The average score according to the formula of the simple arithmetic mean of idle time:

$$(3+4+4+5)/4 = 16/4 = 4.$$

EXAMPLE OF WEIGHTED ARITHMETIC MEAN

The student passed 4 exams and received the following grades: 3, 4, 4 and 5.

The average score according to the formula of the weighted arithmetic mean of idle time:

$$(3*1 + 4*2 + 5*1)/4 = 16/4 = 4.$$

EXAMPLE OF WEIGHTED ARITHMETIC MEAN

If the values of X are given in the form of intervals, then the calculations use the midpoints of the intervals X, which are defined as the half-sum of the upper and lower limits of the interval.

If there is no lower or upper limit (open interval) in the interval X, then the scope (difference between the upper and lower limit) of the adjacent interval X is used to find it..

EXAMPLE OF WEIGHTED ARITHMETIC MEAN

The company has 10 employees with experience up to 3 years, 20 - with experience from 3 to 5 years, 5 employees - with experience more than 5 years.

Average length of service of employees according to the formula of weighted arithmetic mean, taking as X the middle of the intervals of experience (2, 4 and 6 years):

(2*10+4*20+6*5)/(10+20+5) = 3,71 years.

STRUCTURAL AVERAGE

Usually the mathematical mean is not enough to analyze the distribution.

Structural means **(mode, median, quartile, decile and percentile)** are used for the primary analysis of the distributio.n of features in the aggregate.

MODA

Moda - is the most common variant of the variation series.

For a discrete series, this is the variant to which the highest frequency corresponds.

Number of retakin of exem, x	Number of students, f	X*f
1	12	12
2	34	68
3	8	24

Mo = 2

For an interval series with equal intervals, the mode is determined using the following formula :

$$\boxed{M_o = x_{M_o} + h_{M_o} \cdot \frac{f_2 - f_1}{2f_2 - f_1 - f_3}}$$

where x_{M_o} - the beginning of the modal interval;

h_{Mo} - the magnitude of the modal interval;

f_2 - modal interval frequency;

f_1 - premodal interval frequency;

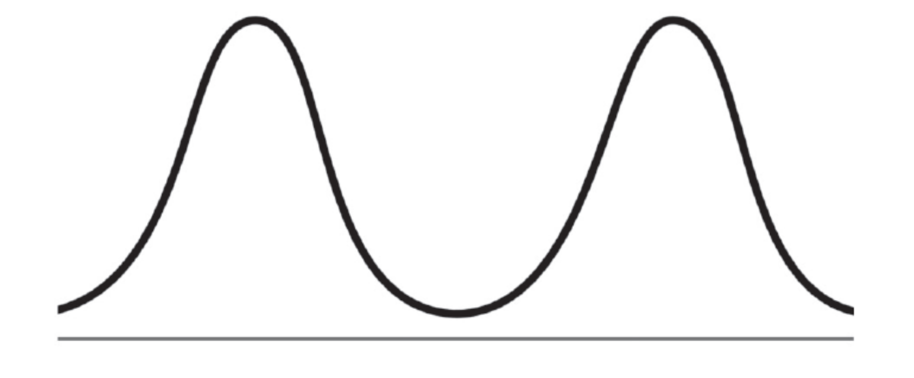
f_3 - postmodal interval frequency

Выработка, м.	Число рабочих, *f*	*x*	*x·f*	*x`*	*x`·f*	*S*	$x-\bar{x}$	$(x-\bar{x})^2\cdot f$	*x`* $^2\cdot f$
до 200	3	190	570	-3	-9	3	-63,9	12249,63	27
200-220	12	210	2520	-2	-24	15	-43,9	23126,52	48
220-240	50	230	11500	-1	-50	65	-23,9	28560,50	50
240-260	56	250	14000	0	0	121	-3,9	851,76	0
260-280	47	270	12690	1	47	168	16,1	12182,87	47
280-300	23	290	6670	2	46	191	36,1	29973,83	92
300-320	7	310	2170	3	21	198	56,1	22030,47	63
320 и более	2	330	660	4	8	200	76,1	11582,42	32
Итого:	200		50780		39			140558	359

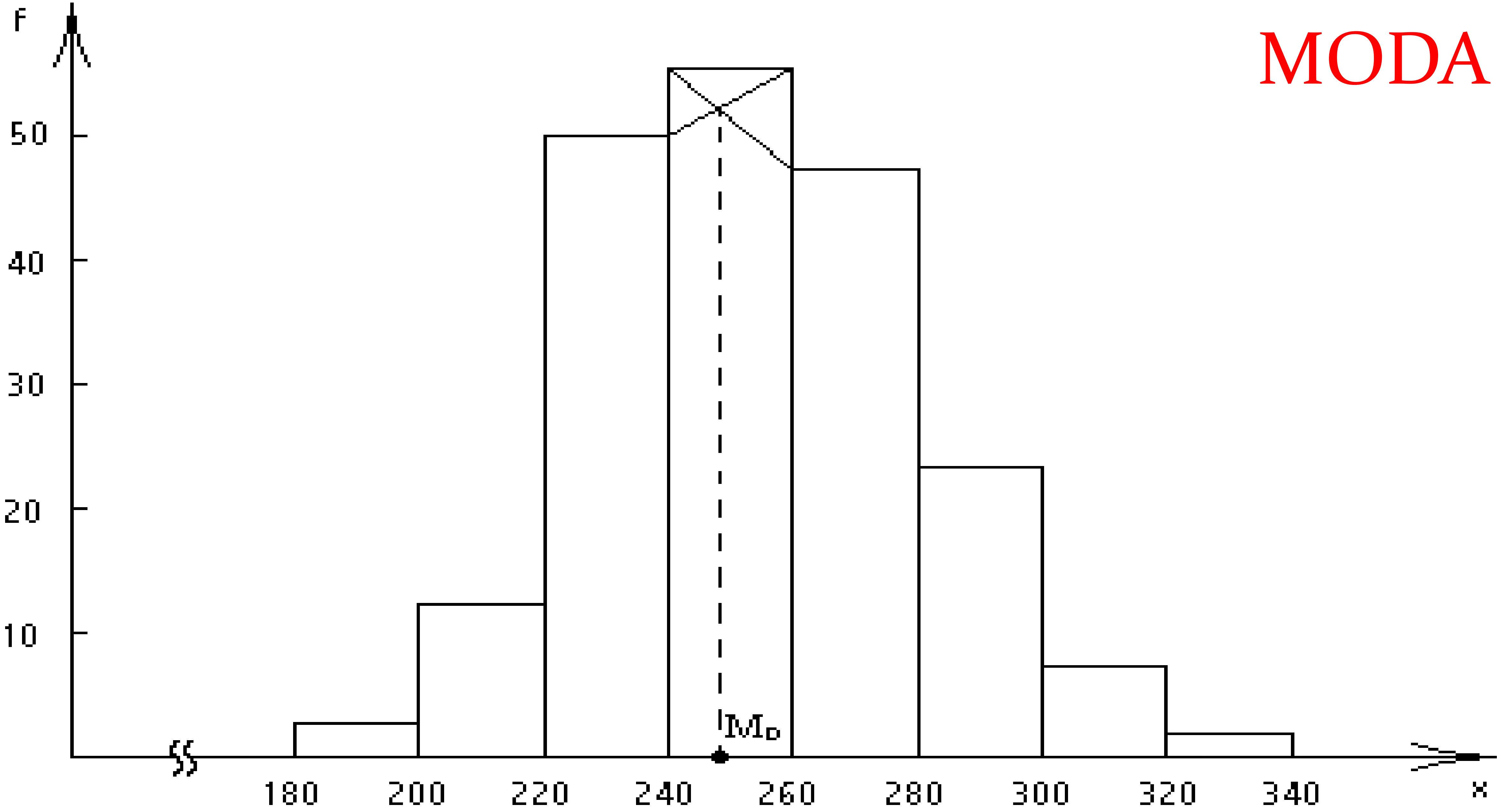
$$Mo = 240 + 20\cdot\frac{56-50}{2\cdot 56-50-47} = 248\textit{м}.$$

MODA

- If the modal interval is the first or last, then the missing frequency (premodal or postmodal) is taken equal to zero.
- If in a discrete series several variants have the highest frequency (which is quite rare), then we are talking about a bimodal or multimodal distribution.



MODA



MEDIAN

- This is the central, middle value of the series.
- *Me* - is the value of the attribute in the unit located in the middle of the ranked (ordered) population.

This is a variant lying in the middle of the variation series and dividing it into two equal parts.

- In the discrete series *Me* is by definition, and in the interval series by formula.

MEDIAN

If a discrete variation series contains an odd number of variants, then there is a single variant, to the right and left of which is the same number of variants :

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

MEDIAN

- If the discrete variation series contains an even number of variants, then there are two variants, to the right and left of which the same number of variants is located.
- ***Me*** is equal to the arithmetic mean of the two values:

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

For a discrete series, the median is the variant for which the accumulated frequency for the first time exceeds half of the sum of the frequencies

For the interval series, the median is determined by the following formula:

$$Me = x_{Me} + h_{Me} \cdot \frac{\frac{\sum f}{2} - S_{Me-1}}{f_{Me}},$$

where x_{Me} - the beginning of the median interval;

h_{Me} - the value of the median interval;

f_{Me} - the frequency of the median interval;

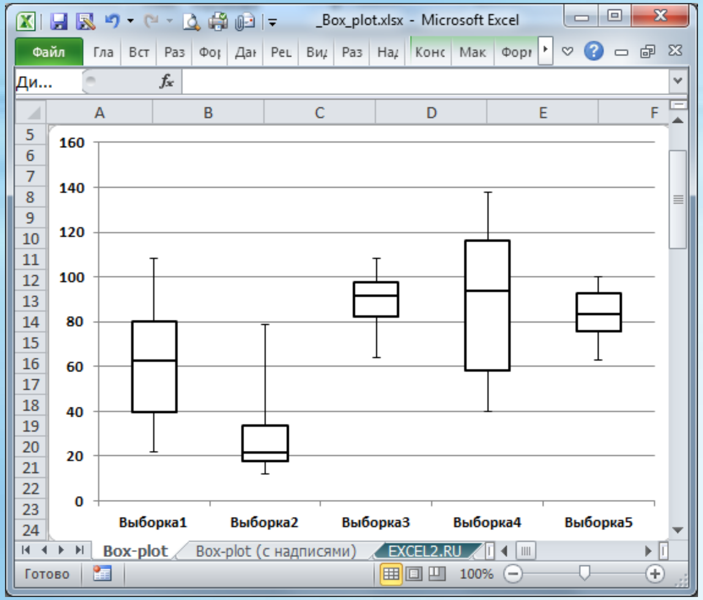
S_{Me-1} - the accumulated frequency of the premedian interval

Выработка, м.	Число рабочих, f	x	$x\cdot f$	$x`$	$x`\cdot f$	S	$x-\bar{x}$	$(x-\bar{x})^2\cdot f$	$x`^2\cdot f$
до 200	3	190	570	-3	-9	3	-63,9	12249,63	27
200-220	12	210	2520	-2	-24	15	-43,9	23126,52	48
220-240	50	230	11500	-1	-50	65	-23,9	28560,50	50
240-260	56	250	14000	0	0	121	-3,9	851,76	0
260-280	47	270	12690	1	47	168	16,1	12182,87	47
280-300	23	290	6670	2	46	191	36,1	29973,83	92
300-320	7	310	2170	3	21	198	56,1	22030,47	63
320 и более	2	330	660	4	8	200	76,1	11582,42	32
Итого:	200		50780		39			140558	359

$$Me = 240 + 20\cdot\frac{\frac{200}{2}-65}{56} = 252{,}5$$

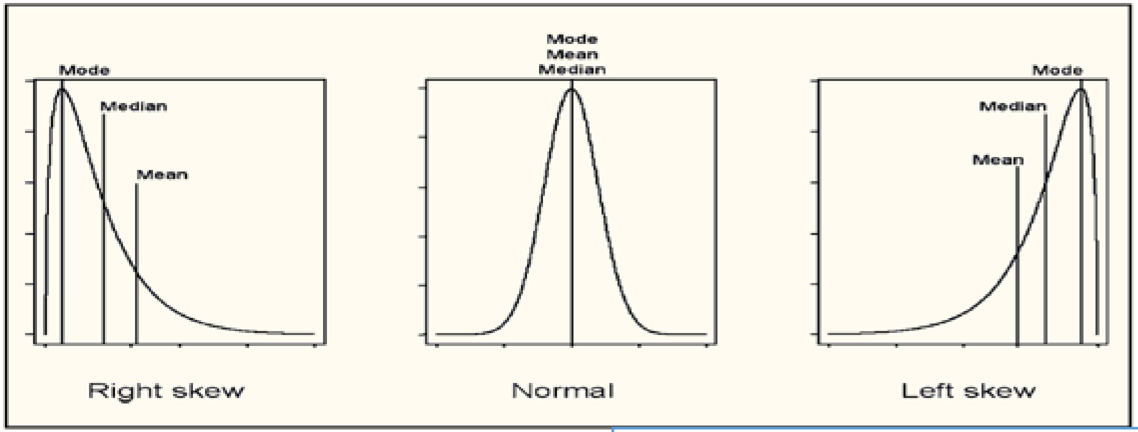
This means that half of the workers have a productivity of less than 252.5 m, and the other half more

Mo & Me



In practical calculations, Mo and Me can be values that are far apart from each other.

For a clearer fixation of the nature of the distribution use other structural averages - quartiles, deciles, percentiles.



In practical calculations, Mo and Me can be values that are far apart from each other.

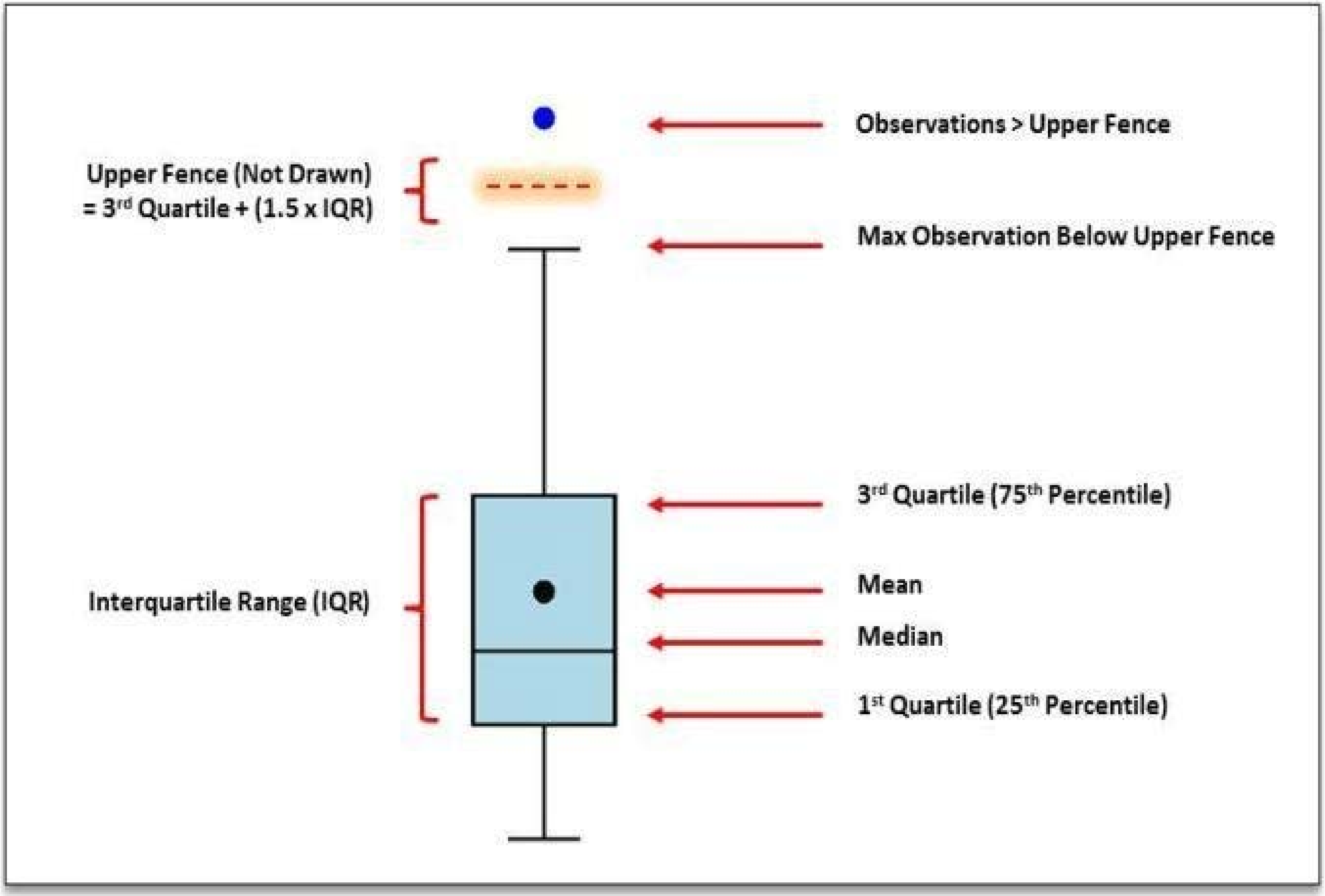
The following ratio of mode, median, and arithmetic mean takes place in the variation series

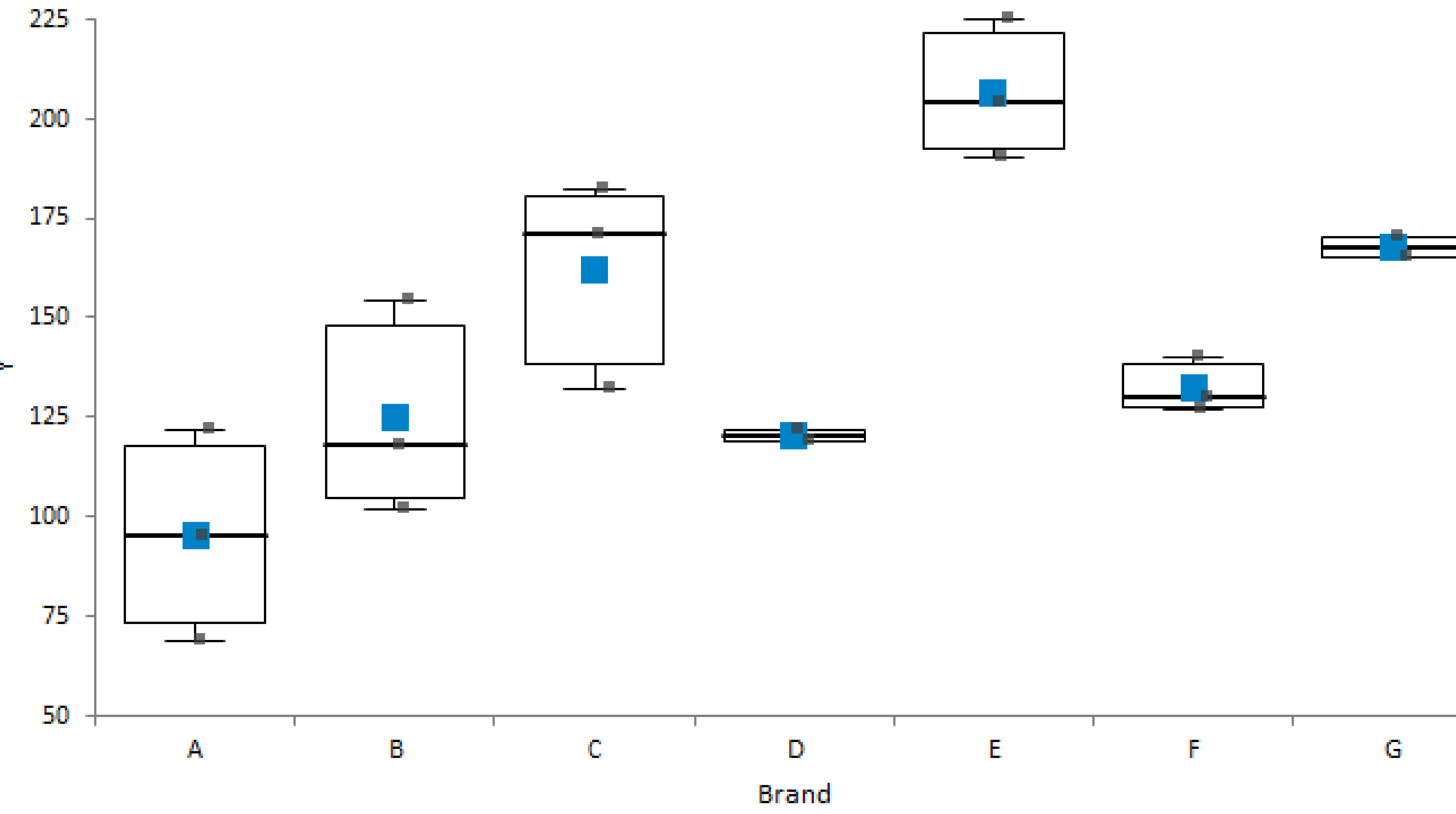
Left-hand asymmetry of the series $\bar{x} < Me < Mo$

Right-hand asymmetry of the series $Mo < Me < \bar{x}$

Symmetry of the series $\bar{x} = Me = Mo$

Mo и Me


Observations > Upper Fence
Upper Fence (Not Drawn)
= 3rd Quartile + (1.5 x IQR)
Max Observation Below Upper Fence
3rd Quartile (75th Percentile)
Mean
Interquartile Range (IQR)
Median
1st Quartile (25th Percentile)


225
200
175
150
125
100
75
50
Y
A
B
C
D
E
F
G
Brand

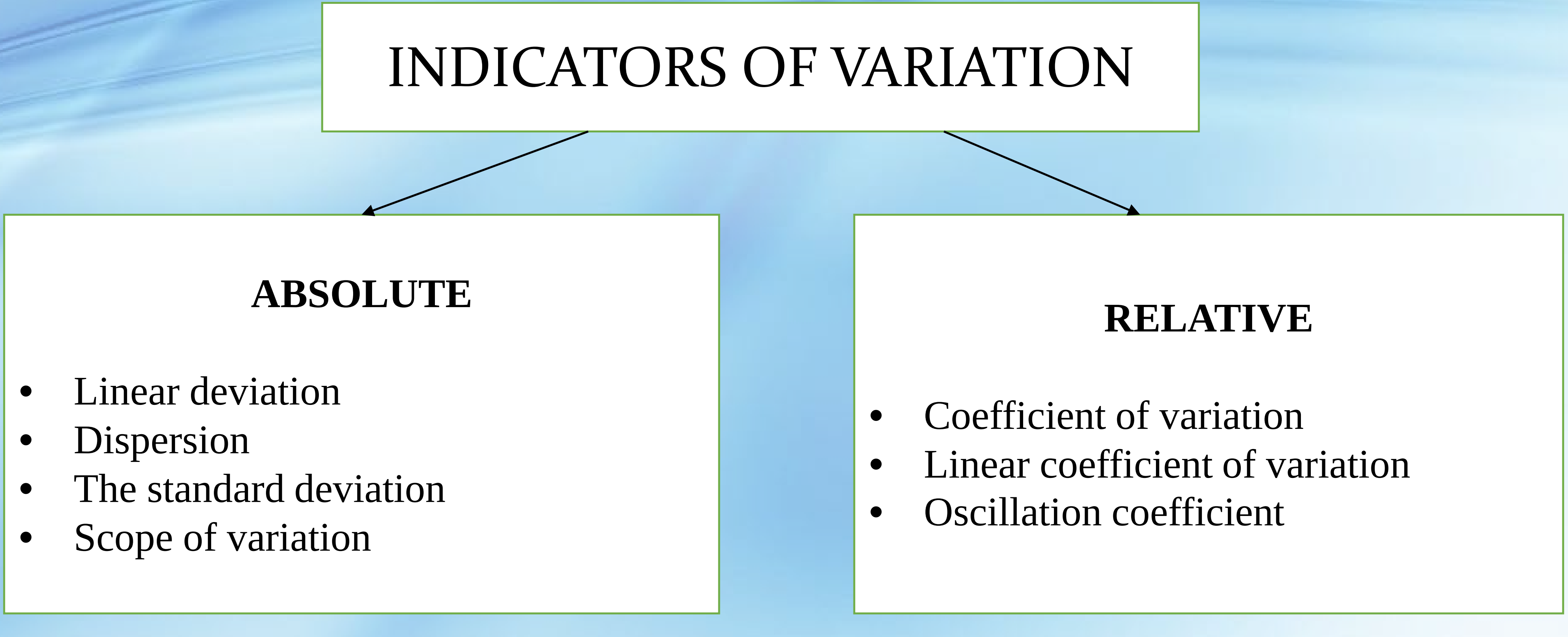
CHARACTERISTICS OF DIVERSITY OF VARIATION ROW

Variation is called variability, the variability of the magnitude of the trait.

Variation is manifested in deviations from the mean and depends on many factors influencing the socio-economic phenomenon.

Variation is random and systematic, exists in space and time.

Indicators of variation are divided into absolute and relative.



INDICATORS OF VARIATION

ABSOLUTE

- Linear deviation
- Dispersion
- The standard deviation
- Scope of variation

RELATIVE

- Coefficient of variation
- Linear coefficient of variation
- Oscillation coefficient

SCOPE OF VARIATION (AMPLITUDE)

$$R = x_{\max} - x_{\min}$$

- **shows how large the difference is between the population units having the smallest and largest value of the attribute.**
- **does not take into account the frequency option.**

An example is the difference between the maximum and minimum pensions of different groups of the population, the level of income of different categories of workers or the production rates of workers of a certain specialty or qualification.

SCOPE OF VARIATION (AMPLITUDE)

- depends only on the two extreme values of the attribute.

For this reason, it is advisable to use it in cases where either the minimum or maximum option is of particular importance, ie. when the scope of variation is of great semantic importance.

For example, they determine the limits within which the size of certain indicators can vary; it is used to assess various types of risks.

- the magnitude of the range of variation is greatly influenced by chance.

Since only two values of the attribute are taken from the statistical series, and the extremes in the series, the magnitude of these values may be influenced by causes of random nature, the magnitude of variation may depend on the causes of random nature.

LINEAR DEVIATION

- The linear deviation is calculated in order to give a generalized characteristic to the distribution of deviations, which takes into account the differences of all units of the studied statistical population.
- The linear deviation is defined as the arithmetic mean of the deviations of individual values from the average without taking into account the sign of these deviations.

LINEAR DEVIATION

- Simple

$$\overline{d} = \frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right|}{n}$$

- Weighted

$$\overline{d} = \frac{\sum_{i=1}^{m} \left| x_i - \overline{x} \right| \cdot f_i}{\sum_{i=1}^{m} f_i}$$

LINEAR DEVIATION

- ***the student passed 4 exams and received the following grades: 3, 4, 4 and 5.***
- ***arithmetic mean = 4.***
- ***Calculate the simple linear deviation:***

d = (|3-4|+|4-4|+|4-4|+|5-4|)/4 = 0,5.

- ***the student passed 4 exams and received the following grades: 3, 4, 4 and 5.***
- ***arithmetic mean = 4.***
- ***Calculate the weighted average linear deviation:***
- ***d = (|3-4|*1+|4-4|*2+|5-4|*1)/4 = 0,5.***

DISPERSION

- **The dispersion -** is the mean square of the deviations of the X values from the arithmetic mean.

- The total dispersion reflects the variation of the trait due to all the factors acting in this set.

DISPERSION

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

$$\sigma^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 f_i}{\sum_{i=1}^{m} f_i}$$

- ***the student passed 4 exams and received the following grades: 3, 4, 4 and 5.***
- ***arithmetic mean = 4.***
- ***the simple dispersion =***

$((3-4)^2+(4-4)^2+(4-4)^2+(5-4)^2)/4 = 0,5$.

- ***the student passed 4 exams and received the following grades: 3, 4, 4 and 5.***
- ***arithmetic mean = 4.***
- ***the weighted dispersion =***

$((3-4)^2*1+(4-4)^2*2+(5-4)^2*1)/4 = 0,5$.

THE STANDARD DEVIATION

The standard deviation shows how much the value of the characteristic varies on the units of the studied population, and is expressed in the same units as the variants.

THE STANDARD DEVIATION

- Simple

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

- Weighted

$$\sigma = \sqrt{\frac{\sum_{i=1}^{m}(x_i - \bar{x})^2 f_i}{\sum_{i=1}^{m} f_i}}$$

IF THE DISPERSION IS PREVIOUSLY CALCULATED, THE SQUARE DEVIATION CAN BE CALCULATED BY EXTRACTING THE SQUARE ROOT FROM THE INDICATOR TO

THE STANDARD DEVIATION

- In the example about the student in which the variance was calculated above, we find the standard deviation as the square root of it:

$$\sigma = \sqrt{0{,}5} = 0{,}707$$

THE STANDARD DEVIATION

The standard deviation is most often used in determining the norm and pathology, which is based on the "rule of three sigmas", valid only for the normal distribution.

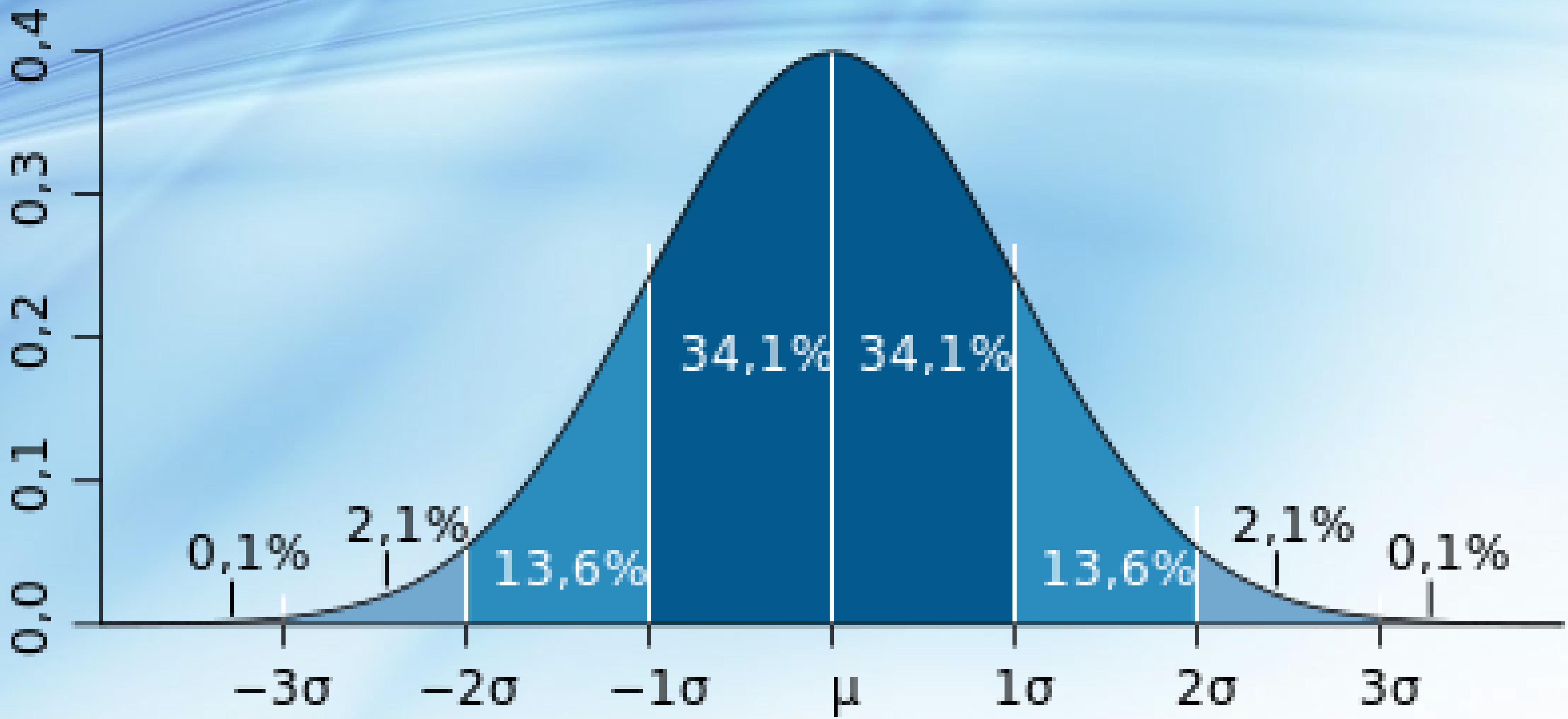
“Rule of three sigmas"

- 68.3% all variants deviate from its average by no more than on X ± σ
- 95.4% all variants are within X ± 2σ
- 99.7% all variants are within X ± 3σ

THE STANDARD DEVIATION

Deviation of a parameter from its arithmetic mean

- within **1σ** is regarded as the norm,
- a deviation within is considered subnormal **± 2σ**
- pathological - beyond this limit, ie **> ± 2σ**


0,4
0,3
0,2
0,1
0,0
34,1%
34,1%
13,6%
13,6%
2,1%
2,1%
0,1%
0,1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ

PROBABILITY OF RESEARCH RESULTS

PROBABILITY

In medicine, the mathematical apparatus of probability theory and mathematical statistics is used for:

- *predicting the occurrence and determining the consequences of the disease;*
- *forecasting the development of epidemics;*
- *for processing the accumulated long-term experience, etc.*

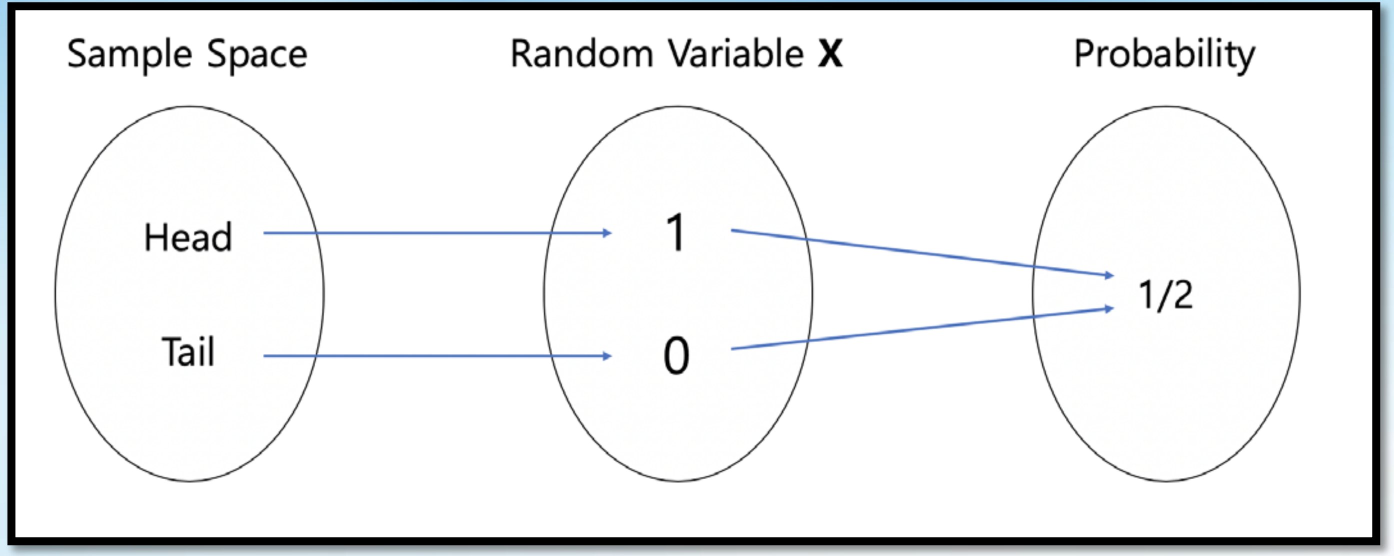
Events: probable, impossible, accidental

Probability - allows you to judge how often you can expect the manifestation of an event with a given number of experiments.

A random phenomenon is a phenomenon that, with the repeated reproduction of the same study, can proceed somewhat differently each time.

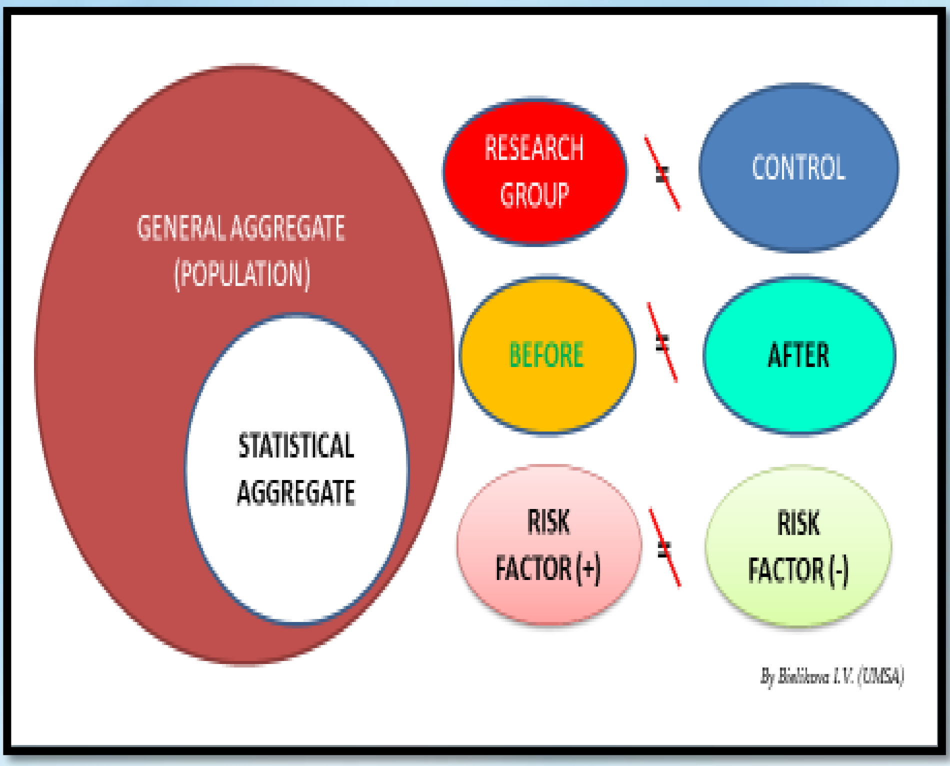
There are 3 types of random events:

- incompatible – which can not occur simultaneously.
- independent – If the fact of occurrence of one of them does not cancel the possibility of occurrence of another.
- dependent – when the result A affects the initial events B.



To assess the probability of the results of a sample survey means to determine the extent to which the conclusions (results) made for it can be transferred to the general population.

That is, for part of the phenomenon to think about the phenomenon as a whole and the main inherent patterns.

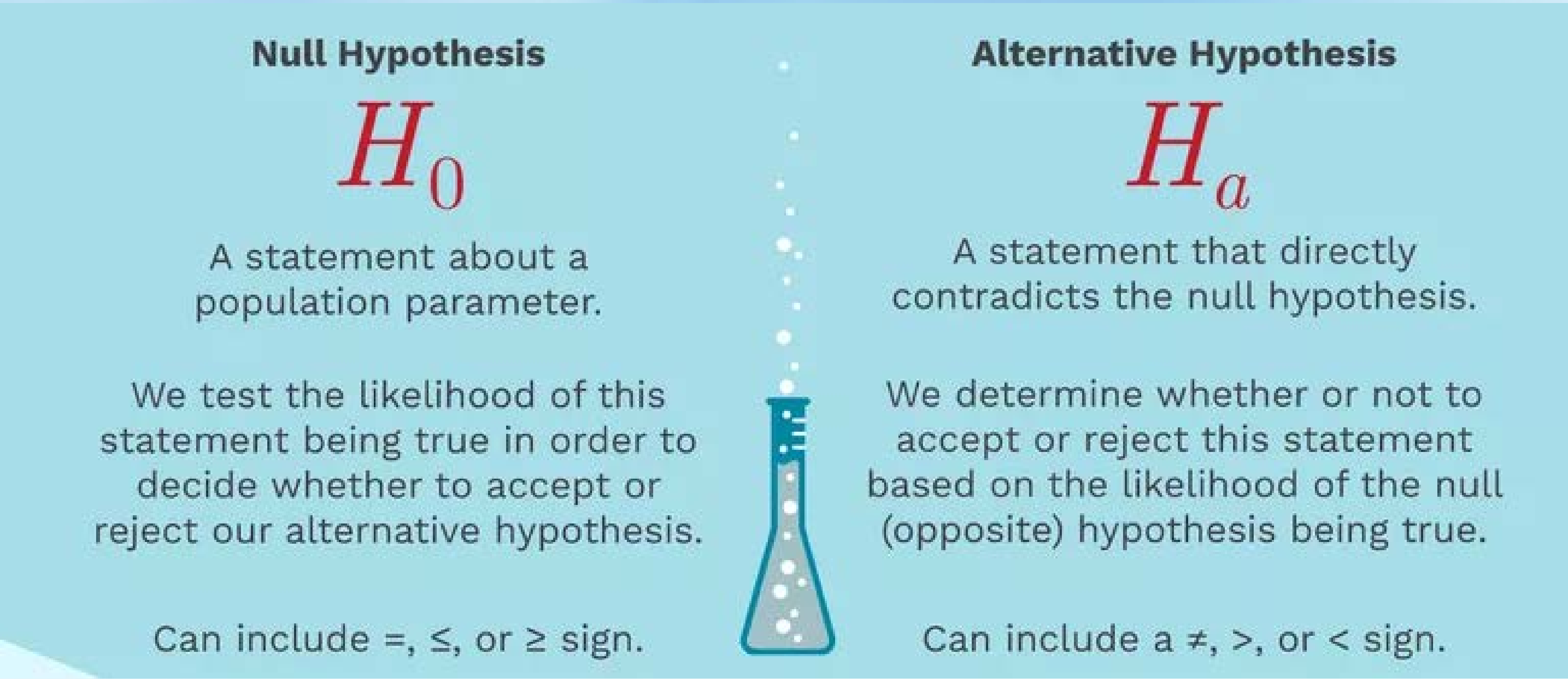


By Bielikova I.V. (UMSA)

Example: Researchers are interested in blood pressure in medical students aged 20 years. To solve this and similar problems, they receive information on the parts of the general population (UMSA students) and judge the general population on the characteristics of the part.

HYPOTHESIS

Hypothesis - an assumption at a certain level of statistical significance about the properties of the general population according to estimates of the sample.



Null Hypothesis

H_0

A statement about a population parameter.

We test the likelihood of this statement being true in order to decide whether to accept or reject our alternative hypothesis.

Can include =, ≤, or ≥ sign.

Alternative Hypothesis

H_a

A statement that directly contradicts the null hypothesis.

We determine whether or not to accept or reject this statement based on the likelihood of the null (opposite) hypothesis being true.

Can include a ≠, >, or < sign.

THE NULL HYPOTHESIS

The null hypothesis reflects that there will be no observed effect in our experiment. In a mathematical formulation of the null hypothesis, there will typically be an equal sign. This hypothesis is denoted by H_0.

The null hypothesis is what we attempt to find evidence against in our hypothesis test. We hope to obtain a small enough p-value that it is lower than our level of significance alpha and we are justified in rejecting the null hypothesis. If our p-value is greater than alpha, then we fail to reject the null hypothesis.

If the null hypothesis is not rejected, then we must be careful to say what this means. The thinking on this is similar to a legal verdict. Just because a person has been declared "not guilty", it does not mean that he is innocent. In the same way, just because we failed to reject a null hypothesis it does not mean that the statement is true.

For example, we may want to investigate the claim that despite what convention has told us, the mean adult body temperature is not the accepted value of 98.6 degrees Fahrenheit. The null hypothesis for an experiment to investigate this is “The mean adult body temperature for healthy individuals is 98.6 degrees Fahrenheit.” If we fail to reject the null hypothesis, then our working hypothesis remains that the average adult who is healthy has a temperature of 98.6 degrees. We do not prove that this is true.

If we are studying a new treatment, the null hypothesis is that our treatment will not change our subjects in any meaningful way. In other words, the treatment will not produce any effect in our subjects.

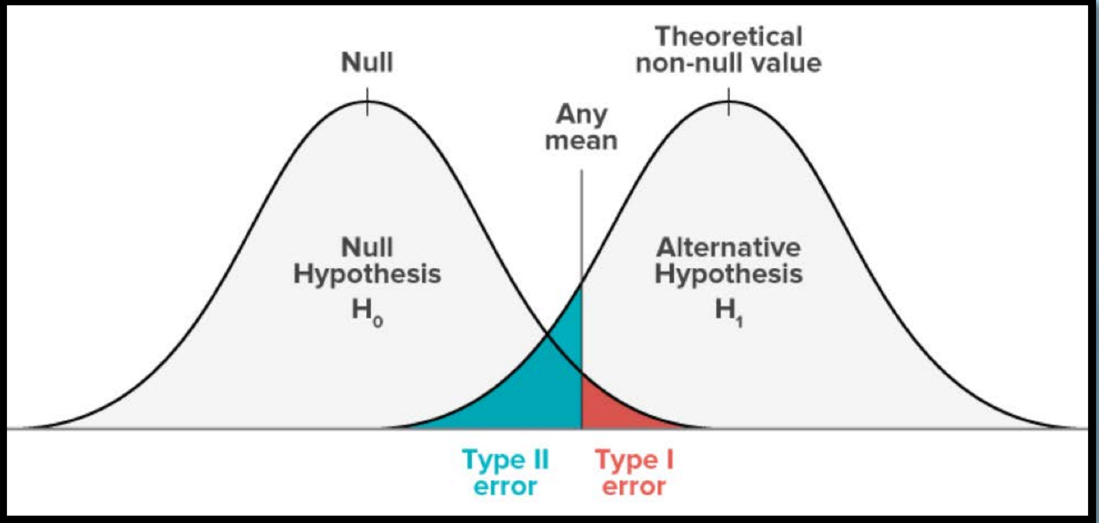
THE ALTERNATIVE HYPOTHESIS

The alternative or experimental hypothesis reflects that there will be an observed effect for our experiment. In a mathematical formulation of the alternative hypothesis, there will typically be an inequality, or not equal to symbol. This hypothesis is denoted by either H_a or by H_1.

The alternative hypothesis is what we are attempting to demonstrate in an indirect way by the use of our hypothesis test. If the null hypothesis is rejected, then we accept the alternative hypothesis. If the null hypothesis is not rejected, then we do not accept the alternative hypothesis. Going back to the above example of mean human body temperature, the alternative hypothesis is “The average adult human body temperature is not 98.6 degrees Fahrenheit.”

If we are studying a new treatment, then the alternative hypothesis is that our treatment does, in fact, change our subjects in a meaningful and measurable way.

Statistical Error (m) shows how the result obtained in the sample study differs from the result that could be obtained in a continuous study of the whole population.



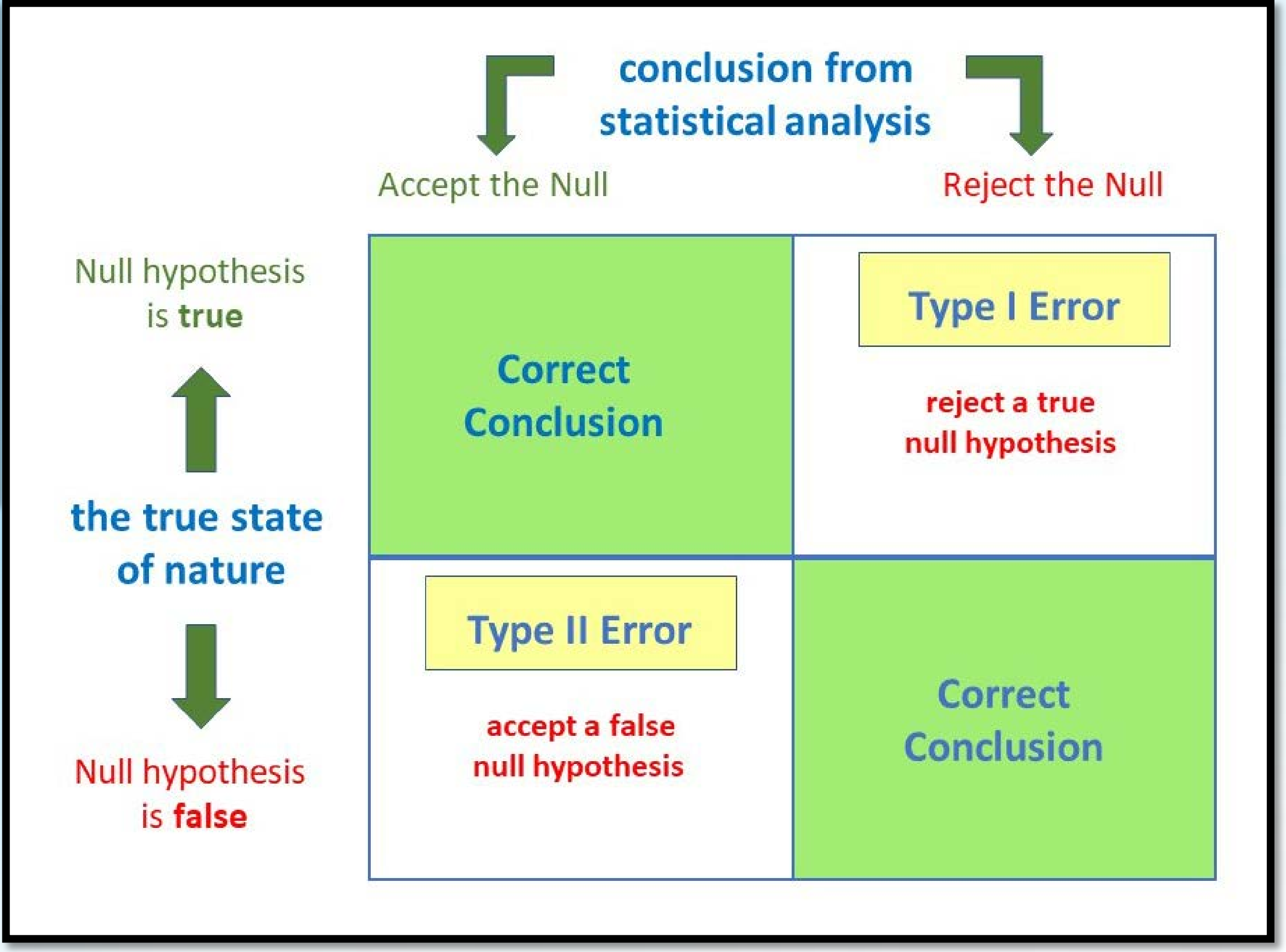
$$m = \frac{\sigma}{\sqrt{n}}$$

For average values

$$m = \sqrt{\frac{p \bullet q}{n}}$$

For relative values

n – number of observations
σ – standard deviation
P – relative value
q - (100-P) – the probability that the phenomenon will not be registered



conclusion from statistical analysis

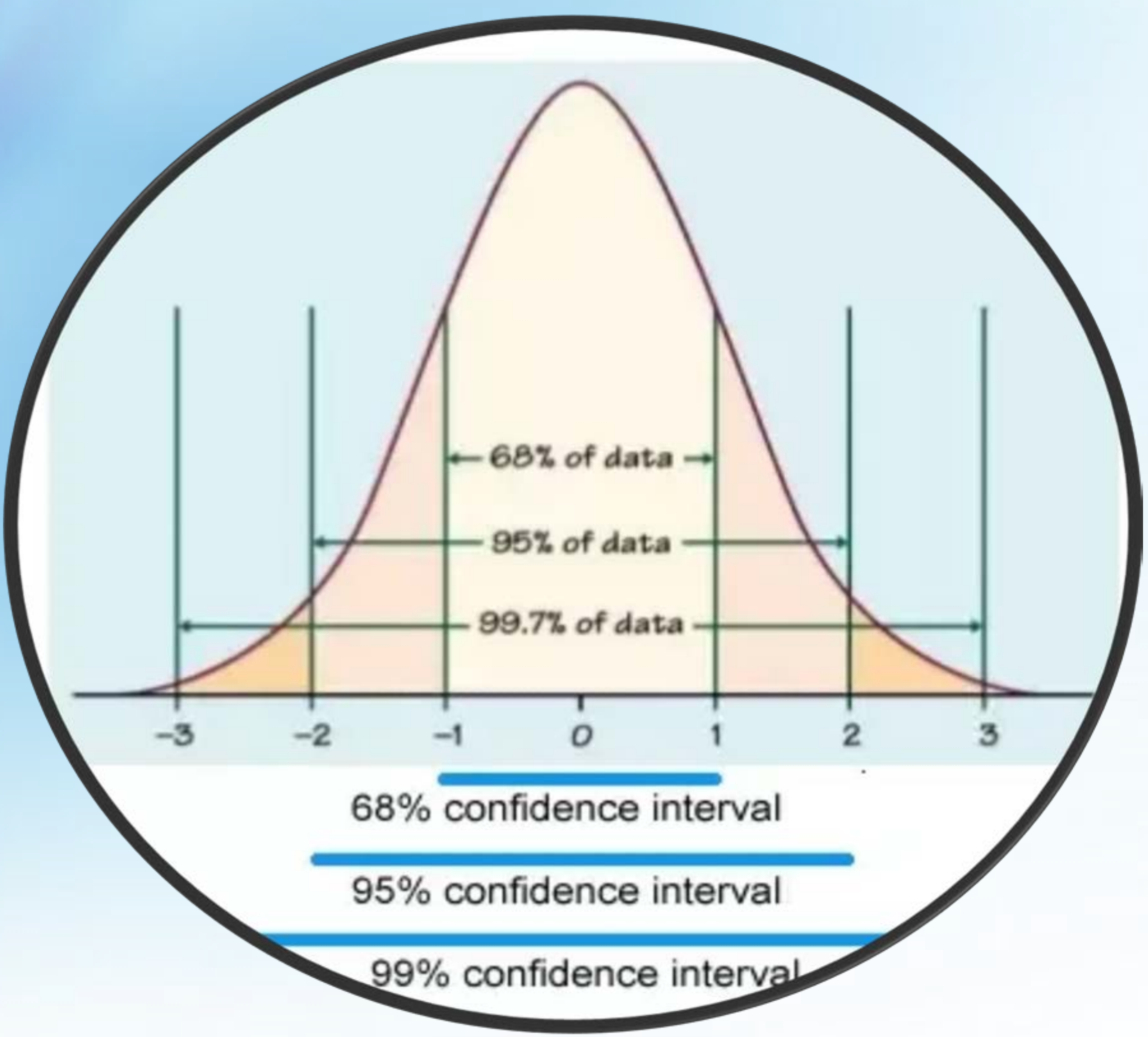
the true state of nature

	Accept the Null	Reject the Null
Null hypothesis is true	Correct Conclusion	Type I Error reject a true null hypothesis
Null hypothesis is false	Type II Error accept a false null hypothesis	Correct Conclusion

If the study aims to assess the overall (population) average value of the quantitative trait, it is advisable to present the result in the form of the arithmetic mean (M) and its 95% confidence interval (CI).

Confidence limits *are the limits of average (or relative) values, beyond which due to random fluctuations there is a small probability***.**

Confidence interval (CI) *is a term used in mathematical statistics for interval estimation of statistical parameters, which is more preferable with a small sample size than a point. An interval that covers an unknown parameter with a given reliability.*



IN MEDICAL PRACTICE, THE ACCEPTED PROBABILITY BECOMES (p<0,05)

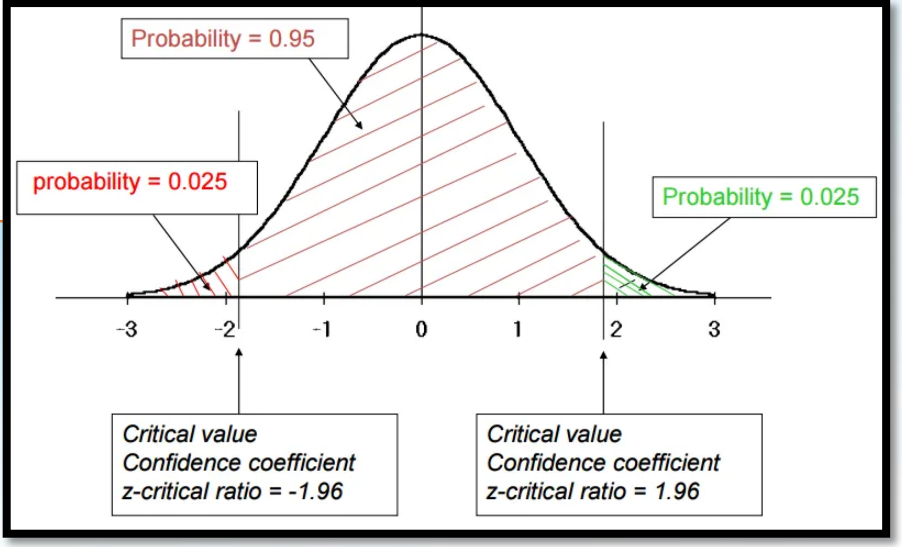
$$P_{general} = P_{stat.agr} \pm tm$$

(for relative values)

$$M_{general} = M_{stat.agr} \pm tm$$

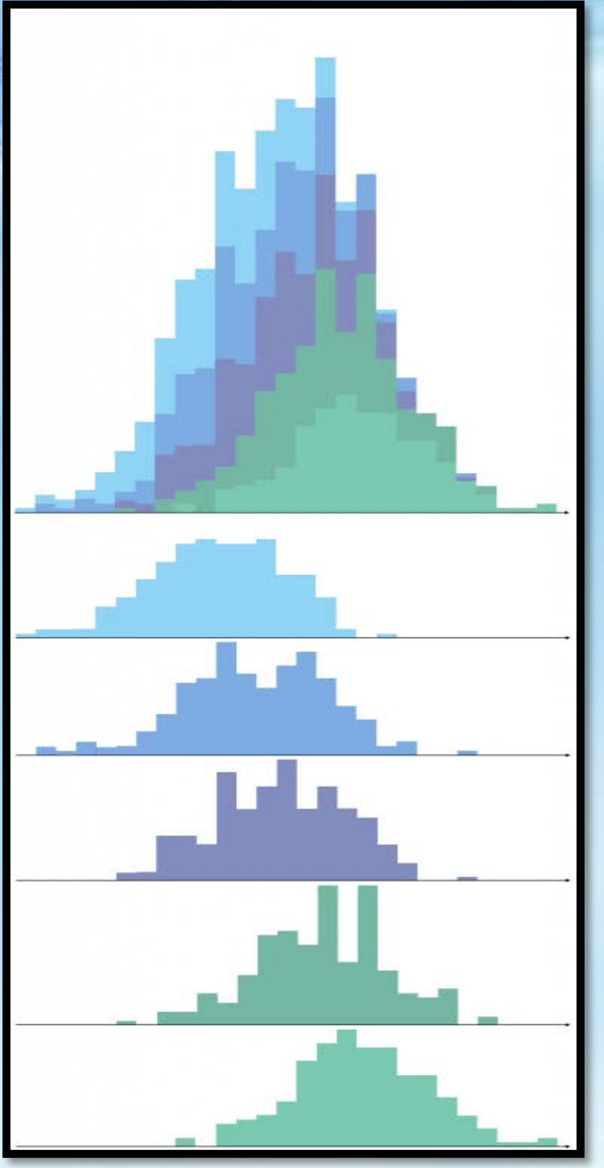
(for average values)

- $P_{general}$ and $M_{general}$ - required general parameters of frequency and average level,
- $P_{stat.agr}$ and $M_{stat.agr}$ - sample values were found,
- m - display error,
- t - confidence criterion.



The accepted probability is 95.5% (t = 2) or 99.7 (t = 3)

SELECTION OF METHODS TO BE USED FOR FURTHER STATISTICAL EXAMINATION



DETERMINING THE TYPE OF DISTRIBUTION

⇩ ⇩

NORMAL DISTRIBUTION (GAUSSIAN)

ABNORMAL DISTRIBUTION (POISSON`S)

⇩ ⇩

PARAMETRIC STATISTICS

NONPARAMETRIC STATISTICS

Student's t-test is used to determine the statistical significance of differences in averages in the normal distribution

Paired t-test

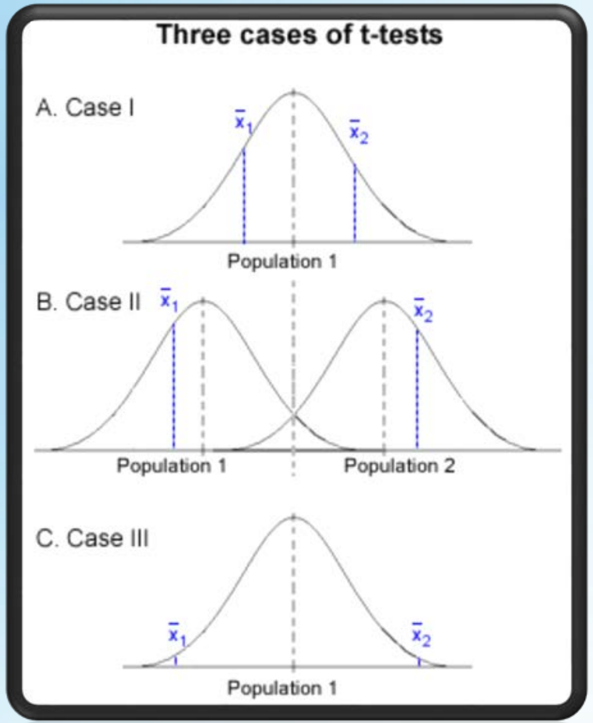
(for dependent samples)

The **paired** sample **t-test**, sometimes called the dependent sample **t-test**, is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. In a **paired** sample **t-test**, each subject or entity is measured twice, resulting in pairs of observations.

Student's t-test

(for independent samples)

The Student’s t Test is used to compare the mean of two normally distributed samples, preferably of equal size and variance. More specifically, the Student’s t Test gives you a probabilistic estimate of the likelihood that your samples were randomly selected from the same population, i.e. that the Null Hypothesis is true. If the Student’s t Test suggests that your samples were probably not taken from the same population, you can conclude, with some certainty (how certain depends on your p value), that your samples were taken from different populations.



$$\boldsymbol{t = \frac{X}{\sigma/\sqrt{n}}}$$

$$\boldsymbol{t = \frac{X_1 - X_2}{\sqrt{{m_1}^2 + {m_2}^2}}}$$

t ≥ 2 probability of error prediction 95% or more (p <0,05)
t <2 probability of error prediction less than 95% (p> 0,05)

Comparison of different sets using nonparametric methods

The difference between related groups

1. Criterion of signs for comparison of groups with pairwise connected variants

2. Wilcoxon's criterion

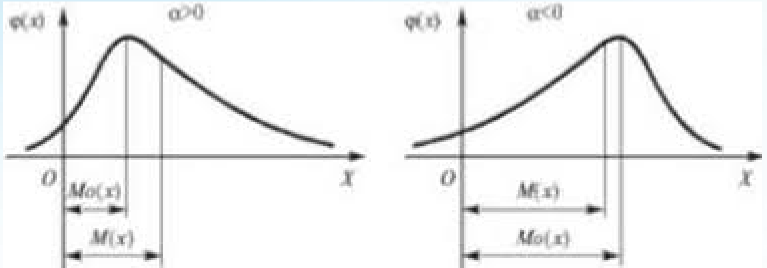
The difference for independent populations

1. Mann-Whitney test

2. Kolmogorov-Smirnov criterion

For comparison of more than two indicators, for an estimation of qualitative signs:

Use Pearson's matching criterion (x2) - xi square (for relative values)



Criterion of signs

The sign test is a statistical method to test for consistent differences between pairs of observations, such as the weight of subjects before and after treatment. Given pairs of observations (such as weight pre- and post-treatment) for each subject, the sign test determines if one member of the pair (such as pre-treatment) tends to be greater than (or less than) the other member of the pair (such as post-treatment).

The paired observations may be designated x and y. For comparisons of paired observations (x,y), the sign test is most useful if comparisons can only be expressed as $x > y$, $x = y$, or $x < y$. If, instead, the observations can be expressed as numeric quantities ($x = 7$, $y = 18$), or as ranks (rank of x = 1st, rank of y = 8th), then the paired t-test[1] or the Wilcoxon signed-rank test[2] will usually have greater power than the sign test to detect consistent differences.

If X and Y are quantitative variables, the sign test can be used to test the hypothesis that the difference between the X and Y has zero median, assuming continuous distributions of the two random variables X and Y, in the situation when we can draw paired samples from X and Y.[3]

The sign test can also test if the median of a collection of numbers is significantly greater than or less than a specified value. For example, given a list of student grades in a class, the sign test can determine if the median grade is significantly different from, say, 75 out of 100.

The sign test is a non-parametric test which makes very few assumptions about the nature of the distributions under test – this means that it has very general applicability but may lack the statistical power of the alternative tests.

The two conditions for the paired-sample sign test are that a sample must be randomly selected from each population, and the samples must be dependent, or paired. Independent samples cannot be meaningfully paired. Since the test is nonparametric, the samples need not come from normally distributed populations. Also, the test works for left-tailed, right-tailed, and two-tailed tests.

Wilcoxon's criterion

The **Wilcoxon signed-rank test** is a non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's *t*-test (also known as "*t*-test for matched pairs" or "*t*-test for dependent samples") when the distribution of the difference between two samples' means cannot be assumed to be normally distributed. A Wilcoxon signed-rank test is a nonparametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution.

Mann-Whitney test

Further Information

The Mann-Whitney *U* test is a nonparametric test that allows two groups or conditions or treatments to be compared without making the assumption that values are normally distributed. So, for example, one might compare the speed at which two different groups of people can run 100 metres, where one group has trained for six weeks and the other has not.

Requirements

Two random, independent samples

The data is continuous - in other words, it must, in principle, be possible to distinguish between values at the nth decimal place

Scale of measurement should be ordinal, interval or ratio

For maximum accuracy, there should be no ties, though this test - like others - has a way to handle ties

Null Hypothesis

The null hypothesis asserts that the *medians* of the two samples are identical.

Kolmogorov-Smirnov criterion

In statistics, the **Kolmogorov–Smirnov test** (**K–S test** or **KS test**) is a nonparametric test of the equality of continuous (or discontinuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). It is named after Andrey Kolmogorov and Nikolai Smirnov.

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case). In the one-sample case, the distribution considered under the null hypothesis may be continuous, purely discrete or mixed. In the two-sample case, the distribution considered under the null hypothesis is a continuous distribution but is otherwise unrestricted.

The two-sample K–S test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

The Kolmogorov–Smirnov test can be modified to serve as a goodness of fit test. In the special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution. This is equivalent to setting the mean and variance of the reference distribution equal to the sample estimates, and it is known that using these to define the specific reference distribution changes the null distribution of the test statistic. Various studies have found that, even in this corrected form, the test is less powerful for testing normality than the Shapiro–Wilk test or Anderson–Darling test. However, these other tests have their own disadvantages. For instance the Shapiro–Wilk test is known not to work well in samples with many identical values.

Pearson's matching criterion

Pearson's chi-squared test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. It is the most widely used of many chi-squared tests (e.g., Yates, likelihood ratio, portmanteau test in time series, etc.) – statistical procedures whose results are evaluated by reference to the chi-squared distribution. Its properties were first investigated by Karl Pearson in 1900. In contexts where it is important to improve a distinction between the test statistic and its distribution, names similar to Pearson χ-squared test or statistic are used.

It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have total probability 1. A common case for this is where the events each cover an outcome of a categorical variable. A simple example is the hypothesis that an ordinary six-sided die is "fair" (i. e., all six outcomes are equally likely to occur.)