LECTURE COMPLEX

Discipline:	Project activities and biostatistics
Code discipline:	PAB 2303
Name and code of the EP:	6B10115 "Medicine"
Volume of study hours/credits:	150/5
Study course and semester:	2/4
Length of lecture:	8

Shymkent, 2025 y.

The lecture complex was developed in accordance with the working curriculum of the discipline (syllabus) "Introduction to scientific research" and discussed at the department meeting:

Head of department, ass. prof, _______________ M.B. Ivanova

Protocol no. 12a from "28" 05 2025 y.

LECTURE №1

1. Theme: Introduction to biostatistics. Stage of statistical research.

2. Aim of the lecture: formation of students' ideas about the discipline "Biostatistics": its subject, tasks, methods and history of formation, as well as familiarization of students with the sequence of statistical research.

3. Lecture thesis:

Introduction to biostatistics.

Biostatistics is a science that deals with the collection, processing and analysis of data in medicine and biology.

Biostatistics plays an important role in medicine, as it allows conducting objective studies of the effectiveness of treatment, analyzing the influence of various factors on health, predicting the probability of the occurrence of diseases, etc. It helps doctors and researchers make informed decisions based on data.

Task biostatistics:

- quantitative presentation of facts (measurement) is an expression of the properties of a separate biological object in the form of a number, variant or value of a variable;
- a generalized description of a set of facts (statistical estimation) is a calculation of indicators and parameters that fully characterize the properties of a set of objects of the same type or a sample;
- the search for regularities (testing of statistical hypotheses) is proof of the non-randomness of differences between comparable groups, objects, the dependence of their characteristics on external or internal causes.

Biostatistical method:

- collection of data, which can be passive (observation) and active (experiment);
- descriptive statistics, which deals with the description and presentation of data;
- comparative statistics, which allows you to analyze data in the studied groups and compare the groups with each other in order to obtain certain conclusions. These conclusions can be formulated in the form of hypotheses or forecasts;
- methods that allow to estimate the relationship between symptoms;
- methods focused on forecasting data, based on the study of relationships between phenomena or evaluation of their dynamics.

The history of biostatistics began at the end of the 19th century, when statistical methods began to be used in medical and biological research.

The founder of biostatistics is Francis Galton, who for the first time introduced the term "biometrics" into the scientific circulation, developed the basis of correlation analysis.

Followed by F. Galton became the English statistician Carl Pearson, who developed numerous statistical tests and methods.

In the 20th century, Ronald Fischer's ideas and methods of statistics, which were based on biostatistics, were widely used. Currently, biostatistics is an integral part of medical and biological research and continues to develop with the development of new methods of data collection and analysis.

Conducting a statistical study begins with defining the problem, in accordance with which the goal and objectives of the study are set, the literature on this problem is studied, and a working hypothesis is developed. This stage of research is called preparatory.

A problem in public health and health care may be, for example, a low level of health of the population or its group, assumptions about the cause and factors affecting the health of the population or its group, detection of deficiencies in the organization of work of medical workers, etc.

The working hypothesis is the main idea about how to solve the problem facing the researcher. The researcher proposes to test the hypothesis based on the empirical data obtained during the research.

The purpose of the study is the final result to which the study is directed.

Tasks of research - step-by-step achievement of the goal of research. Research tasks reflect specific questions that need to be solved consistently in order to reach the final goal of the research.

The stage of statistical studies is presented in Fig. 1.1.

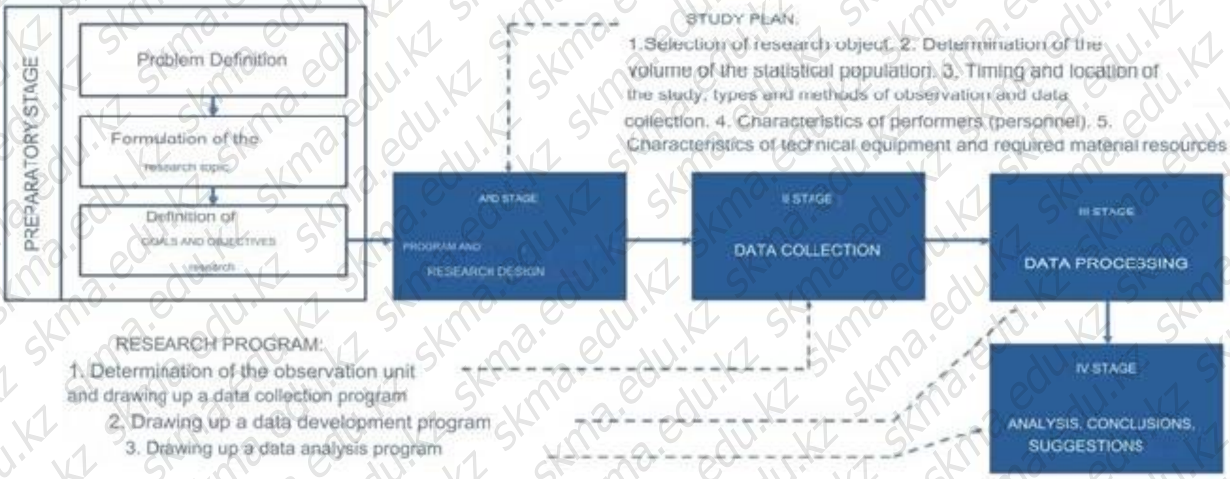

PREPARATORY STAGE
Problem Definition
Formulation of the research topic
Definition of GOALS AND OBJECTIVES research
STUDY PLAN:
1.Selection of research object. 2. Determination of the volume of the statistical population. 3. Timing and location of the study, types and methods of observation and data collection. 4. Characteristics of performers (personnel). 5. Characteristics of technical equipment and required material resources
AND STAGE
PROGRAM AND RESEARCH DESIGN
II STAGE
DATA COLLECTION
III STAGE
DATA PROCESSING
IV STAGE
ANALYSIS, CONCLUSIONS, SUGGESTIONS
RESEARCH PROGRAM:
1. Determination of the observation unit and drawing up a data collection program
2. Drawing up a data development program
3. Drawing up a data analysis program

Figure 1.1. Stages of statistical research

Stage I - compilation of the program and plan of statistical research.

Stage II - organization and collection of necessary data provided by the research program. Data may be collected by conducting surveys, examining medical records, or conducting observations. It is important to guarantee the sufficiency and quality of data for analysis.

Stage III - processing of collected data (control, grouping, encryption, calculation of statistical indicators, summary in statistical tables). Statistical programs can be used for data processing.

Stage IV – analysis and interpretation of research results. Hypothesis testing, correlation analysis, regression analysis, etc. can be used for data analysis. Based on the analysis of the results of the research, conclusions and proposals are formulated.

Stage I of statistical research - compilation of the program and plan of statistical research.

The program of statistical research provides for the solution of the following questions:
1) definition of a monitoring unit and compilation of a data collection program;
2) compilation of data processing program;
3) compilation of the program for the analysis of collected data.

A unit of observation is each primary element of a statistical population. For example, every student, every born, every patient.

The observation unit is endowed with similarities and differences.

Similarities are general accounting signs that indicate the belonging of a specific unit of observation to this population.

Differences in signs are individual features (characteristics) of each unit of observation, which are the final object of statistical research. Differences in signs are subject to study and registration, so they are called accounting signs.

Accounting features are features by which the elements of the observation unit in the statistical population differ.

Accounting features are classified by character and role in the aggregate.

According to the characteristics, it can be divided into:

• qualitative (attributive, descriptive) features - described by words. Nominal and ordinal are distinguished among qualitative signs.

Nominal - signs that can be directly measured. It consists of mutually exclusive categories. For example, diagnosis, gender, profession, marital status. Nominal data that can be attributed only to two opposite categories "yes" - "no", which take one of two meanings (lived - died, smoked - not smoked), are called dichotomous (binary).

Ordinal - signs that can be arranged in a natural order (ranked). For example, assessment of patient's severity, stage of illness, self-assessment of health status.

• quantitative signs - signs that are expressed by numbers. Quantitative signs include:

Continuous - taking any value on a continuous scale. For example, body mass, temperature, biochemical indicators of blood.

Discrete - taking values only from a certain list of certain numbers, usually integers. For example, the number of relapses, the number of children in the family, the number of diseases in one patient, the number of cigarettes smoked, the number of emergency calls, the number of patients admitted to the hospital.

Different types of scales are used for accounting signs.

The scale is a necessary, mandatory element of the measuring procedure. The following types of scales are used in medical research:

• the nominal or scale of names is used to classify the properties of the object, assigning them numerical, letter and other symbolic characteristics (gender, nationality, eye color, hair color, diagnosis, etc.);

• ordinal or rank - ranks the sign values (scale of stages of hypertensive disease according to Myasnikov, scale of degrees of heart failure according to Strajesko-Vasilenko-Langu, scale of severity of coronary insufficiency according to Vogelson, etc.);

• interval - shows the "range" of individual measurements of a symptom (time, temperature scale, test scores);

• scale of relations - shows the ratio of the measured values of the characteristic (height, weight, reaction time, number of performed task tests).

According to the role, the following are distinguished:

• factorial (independent) signs influencing the change of dependent signs;

• resultative (dependent) signs that change their value under the influence of factor signs.

For example, the number of smoked cigarettes is a factor sign, the probability of lung and heart disease is a result sign.

The data collection program is a sequential presentation of the considered signs - questions that need to be answered when conducting this study. The data collection program is made in the form of a registration document (questionnaire, form, card, etc.), which includes the characteristics that the researcher wants to study during the experiment, and which is filled in for each unit of observation.

The program for the development of the received data provides for the compilation of mock-ups of statistical tables.

The program provides a list of analysis

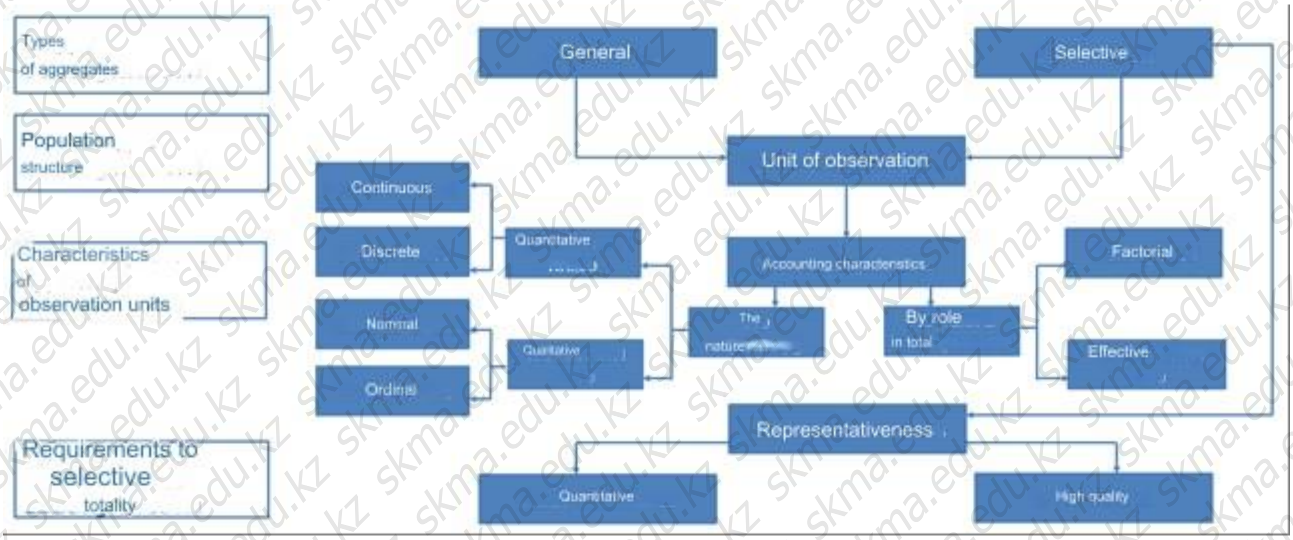


Figure 1.2. Types and structure of the statistical population

There are two types of aggregates - general and selective.

The general population is a group that consists of an infinitely large number of objects (all patients with this pathology; all residents of this territory, etc.).

Sample population (sample) is a part of the general population selected for research and intended to characterize the entire population. The sample should be representative (representative) in quantity and quality in relation to the general population.

Quantitative representativeness is based on the law of large numbers and means a sufficient number of elements of the sample population, calculated according to special formulas.

Qualitative representativeness means the correspondence of the characteristics characterizing the elements of the sample population in relation to the general one. In other words, the internal structure of the sample according to the main characteristics (gender, age, etc.) should correspond to the general population.

The volume of the statistical population (N, n) is the number of elements of the population taken for the study.

The time and place of research is the compilation of a calendar plan for the implementation of this research in a specific territory.

According to the time of registration, two types of observation are distinguished - current (or constant) and one-time (or one-moment).

Current monitoring is a type of monitoring in which registration is carried out continuously as units of monitoring occur. For example, every case of birth, death, referral to a medical institution.

One-time observation - the phenomena under study are recorded at a specific moment (hour, day of the week, date). For example, the population census, the composition of beds in the hospital.

For the researcher, it is important to determine the method of conducting the study: continuous observation or non-continuous (selective).

Continuous observation is a registration of all units of observation that make up the general population.

Non-continuous (selective) observation is the study of only a part of the population to characterize the whole.

Various methods of selection of units are used for conducting random statistical observation: random, mechanical, nested, directed, typological.

- random selection - selection carried out by drawing lots (by the first letter of the family name, by the date of birth, etc.);

• mechanical selection - selection, when every fifth (20%) or tenth (10%) unit of observation is mechanically selected for study in the entire population;

• nested (serial) selection - selection in which not individual units are selected from the general population, but nests (series) that are selected by random or mechanical selection. For example, 10 departments were randomly selected from all departments of the medical academy to participate in the survey, and then all teachers of these departments were interviewed. The department in this case is a "nest".

• directed selection is a selection in which only those observations are selected from the general population, which allow to establish the influence of unknown factors while establishing the influence of known ones. For example, when studying the influence of work experience on injuries, workers of one profession, one age, one workshop, one educational level are selected.

• typological selection is the selection of units from pre-grouped qualitative groups of the same type. For example, when studying the patterns of morbidity among the urban population, it is necessary to first divide the studied population by age structure. After that, a random selection is made in each age group.

Characteristics of performers (cadre) - how many people and what qualifications conduct the study.

Characteristics of technical equipment and required materials - laboratory equipment and instruments, corresponding research purposes, stationery (paper, forms), etc.

Stage II of statistical research - organization and collection of necessary data.

Collection of data is a process of registration, filling of officially existing or specially developed accounting documents (voucher, card, etc.).

The method of collecting statistical data:

• immediate collection of data is carried out by researchers who themselves register signs and facts by counting, measuring, weighing, etc., and then enter the data into statistical observation forms;

• the documentary method of data collection involves obtaining statistical information from documents representing the studied objects, from accounting documentation (history of illness, history of child development, sick list, etc.);

• a survey is a method of data collection, in which the researcher receives the necessary information about each unit of observation with the words asked (oral survey, questionnaire).

Stage III of statistical research - processing of collected data.

This stage of statistical research includes the following actions:

1) control of collected data;
2) encryption;
3) grouping;
4) summary of data;
5) calculation of statistical indicators and statistics

1. Control is an inspection of the collected material with the aim of selecting accounting documents with defects for their subsequent correction, addition or exclusion from the study. For example, gender, age, or no answers to other questions are not indicated in the questionnaire.

2. Encryption is the application of conditional designations of the highlighted features. For example, gender: male. - M, wife. - J; curit - 1, or curit - 0.

3. Data grouping is the distribution of the collected material by qualitative and quantitative characteristics. For example, the grouping of students by courses of study, by floor, by faculty.

4. Summary of data - entry of digital data received after calculation into tables. The table distinguishes between the subject and the predicate.

The subject is what is said in the table (signs that are the subject of research), usually placed vertically in the left part of the table.

Predicate is what characterizes the subject, placed horizontally.

Statistical tables are divided into simple (tab. 1.1), group (tab. 1.2), combination (tab. 1.3).

Table 1.1.

Distribution of smoking students by faculty

Name of the faculty	All students	
	Absolute number of students	%
Medicine		
Pharmacy		
Total		100

Table 1.2.

Distribution of smoking students of various faculties by gender and height, in which they smoked the first cigarette

Name of the faculty	Gender		Age at which you smoked your first cigarette			Total
	M	F	up to 15 years	15-18	over 18 years old	
Medicine						
Pharmacy						
Total						

Table 1.3

Distribution of smoking students of different faculties by gender and average number of cigarettes smoked per day

Name of the faculty	Average number of cigarettes smoked by students per day									Total		
	10 or less up to 15 years			11-20			more 20					
	M	F	both sexes	M	F	both sexes	M	F	both sexes	M	F	both sexes
Medicine												
Pharmacy												
Total												

4. Illustrative material: presentation, slides.

5. Literature:

- Basic:

1. Koychubekov B.K. Biostatistics. uch. allowance/ B.K. Koychubekov.- Almaty: Evero, 2016.
2. Boleshov M.Ə. Medical statistics: okulyk.– Almaty: Evero, 2015.

- Additional:

1. Fundamentals of statistical analysis in medicine: textbook. manual / ed. V.A. Reshetnikova.– M.: Medical Information Agency, 2020. – 176 p.

6. Security questions:

1. List the tasks of biostatistics and its methods.
2. Conduct a brief historical overview of the development of biostatistics as a science.
3. Indicate the sequence (stages) of conducting a statistical study.
4. List the components of a statistical research program.
5. State what a statistical research plan includes.
6. Define a statistical population.

7. Define the unit of observation and provide a classification of its accounting characteristics.
8. List the types of scales used in medical research. Give examples.
9. What are the requirements for the sample population?
10. What is the data collection process?
11. What actions does the stage of processing the received data include?
12. What is data grouping?

LECTURE №2

1. Theme: Descriptive statistics.
2. Aim of the lecture: to develop in students an understanding of the methods of descriptive statistics for assessing and analyzing a statistical population in the study of public health and the activities of medical organizations.
3. Lecture thesis:

1. Definition of ICDescriptive statistics or descriptive statistics (from the English descriptive statistics) is a section of statistics that deals with the processing of empirical data, their systematization, visual presentation in the form of graphs and tables, as well as their quantitative description through basic statistical indicators.

The first step in systematizing statistical observation materials is determining the statistical distribution of the sample.

Statistical distribution of a sample (or variation series, or frequency distribution) is a series of numerical measurements of a characteristic, differing from each other in magnitude and arranged in a certain order (increasing or decreasing).

Variant (x) is called each numerical value in the variation series.

The frequency of variants (ν) is the number of elements of a population that have the same numerical value. The total number of options in a variation series is denoted by n.

Types of variation series (Fig. 2.1):

1. Depending on the type of quantity:

• discrete – contains options represented only by integer numbers (for example, the number of relapses, the number of children in the family, the number of ambulance calls);

• continuous – can contain any value on a continuous scale for measuring a trait (for example, body weight, height, temperature, blood biochemical parameters).

2. Depending on the frequency with which each option occurs in the variation series:

• simple is a series in which each option occurs once (all numbers are different);

• weighted – this is a series in which each option occurs more than once (with different frequencies).

3. Depending on the grouping option:

• ungrouped – contains all the values of individual options;

• grouped (interval) - represented by intervals of variant values and the frequency of variants included in each of them. As a rule, an interval series is used when there are a large number of observations.

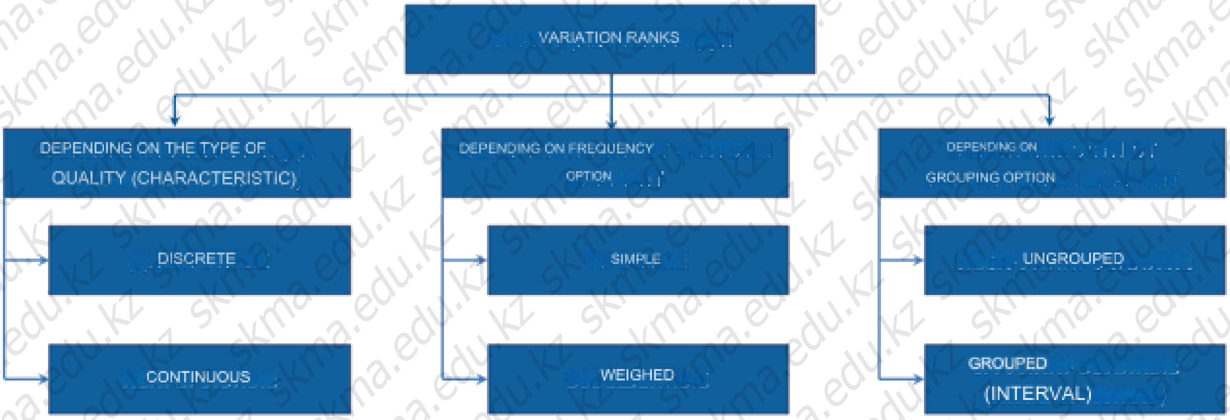

VARIATION RANKS
DEPENDING ON THE TYPE OF QUALITY (CHARACTERISTIC)
DEPENDING ON FREQUENCY OPTION
DEPENDING ON GROUPING OPTION
DISCRETE
CONTINUOUS
SIMPLE
WEIGHED
UNGROUPED
GROUPED (INTERVAL)

Figure. 2.1. Types of variation series

Example 2.1. Heart rate (HR) values were recorded in 20 patients with tachycardia: 100, 100, 100, 112, 112, 112, 112, 112, 120, 120, 120, 120, 120, 120, 124, 124, 124, 124 , 128, 128.

This is a discrete, weighted, ungrouped variation series.

This series can be presented in the form of a table (Table 2.1) and depicted graphically using a distribution polygon or frequency polygon (Fig. 2.2).

Table 2.1.

Options (xi)	Frequencies (vi)
100	3
112	5
120	6
124	4
128	2
Total	*n*=20

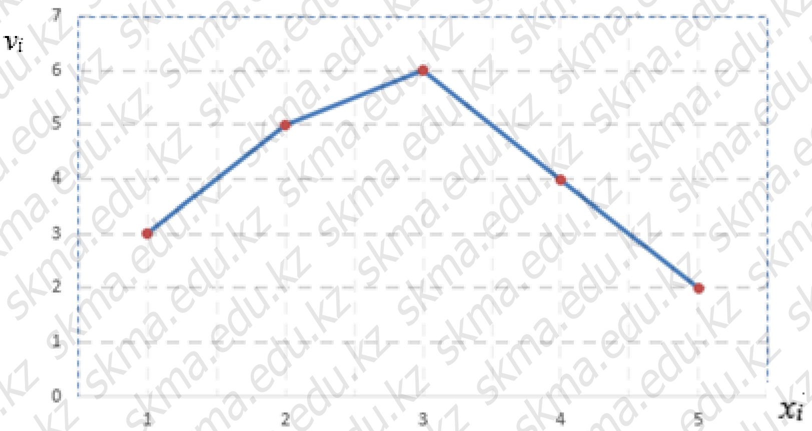


Figure. 2.2. Polygon

Example 2.2. For a sample population consisting of 100 men aged 20-25 years, the height of each observation unit was determined.

According to the measurement results, a continuous, weighted, grouped variation series was constructed (Table 2.2):

Table 2.2.

Variants (xi), cm	Frequencies (νi)
150-155	1
155-160	11
160-165	14
165-170	26
170-175	26
175-180	13
180-185	8
185-190	1
Total:	100

The number of intervals for grouped variation series is determined by the Sturges formula: k=1+3.322•lg n, where n is the sample size. (2.1)

To calculate the width of the interval, use the formula: $h = \dfrac{x_{\max} - x_{\min}}{1 + 3{,}322\lg n}$ (2.2)

where $x_{\max}$, $x_{\min}$ are the largest and smallest values of the option, respectively.
The beginning of the first interval is taken as follows: $x_{\max}$=$x_{\min}$–0.5•h. (2.3)
The grouped variation series is presented graphically in the form of a stepped figure called a histogram (Fig. 2.3).

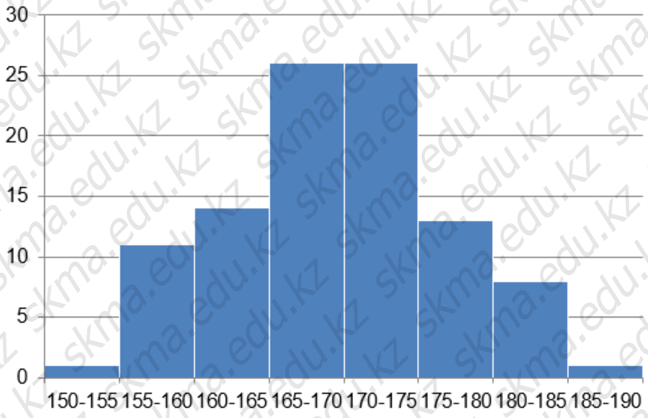
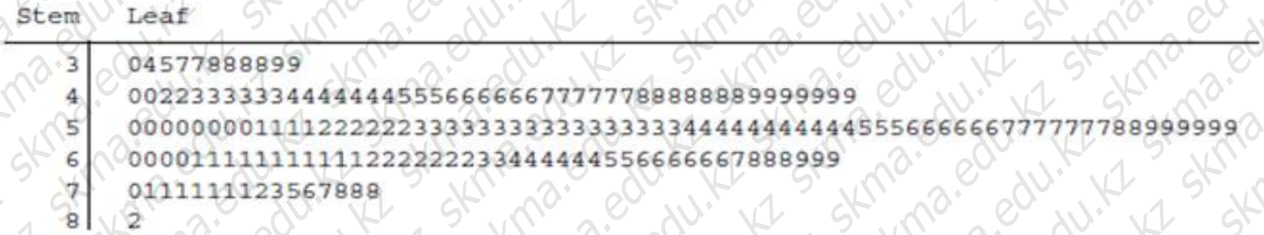


Figure. 2.3. Histogram

You can also visually assess the shape and scope of the data distribution using a *Stem and Leaf Plot.* Example 2.3. There is data on the age of patients of a cardiologist attending a hospital in Shymkent for 2023. Using the “stem with leaves” graph, determine which age patients most often and most rarely access it.

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52

53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						



```
Stem  Leaf
   3 | 04577888899
   4 | 0022333333444444455566666677777778888888999999
   5 | 00000000111122222223333333333333333344444444444555666666777777788999999
   6 | 000011111111111222222233444444556666667888999
   7 | 0111111123567888
   8 | 2
```

Figure; 2.4. Stem and Leaf Plot

The graph shows that patients aged 50 to 59 years most often apply. Patients over 80 years of age are most rarely referred. The number of visits from patients aged 40 to 49 and from 60 to 69 is almost the same.

After the variation series has been constructed, they begin to process it. It consists in finding indicators of central tendency and indicators of variability (diversity).

Indicators of central tendency include average and structural values.

Average values in medicine and healthcare can be used:

1) to assess health status - for example, parameters of physical development (average height, average weight, average vital capacity of the lungs, etc.), somatic indicators (average level of blood sugar coagulation, average pulse, average erythrocyte sedimentation rate (ESR) and etc.);

2) to assess the organization of work of medical organizations, as well as the activities of individual doctors and paramedical workers (average length of stay of a patient in a bed, average number of visits per 1 hour of appointment at a clinic, etc.).

Depending on the nature of the problem, one or another type of average is used. In this biostatistics course, only the arithmetic mean (Average, Mean) will be considered, because Power means (harmonic mean, square mean, geometric mean) are rarely used in medical research.

- Simple arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad (2.4)$$

where n is the total number of members of the series (sample size);

- Weighted arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i v_i}{\sum_{i=1}^{n} v_i}, \qquad (2.5)$$

where vi – frequencies;

The weighted arithmetic mean is used in calculations in grouped variation series, when the series is divided into separate intervals and there is data on the frequency of each of them, but the values of individual variants are not presented. In this case, the middle of each interval is taken as a variant.

Example 2.4. To determine the average body weight of men taking part in a medical study. The data is presented in the form of a grouped variation series (Table 2.3).

Table 2.3

Body weight of examined men (kg)	Number of people surveyed (vi)	Middle of the interval (xi)	x_i*v_i
66-70,9	11	68,5	753,5
71-75,9	18	73,5	1323
76-80,9	24	78,5	1884
81-85,9	14	83,5	1169
Total	67		5129,5

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i * v_i}{\sum_{i=1}^{n} v_i} = \frac{5129{,}5}{67} = 76{,}6 \text{ kg}$$

Structural quantities.

• Mode (Mode) is the value of the attribute that occurs most frequently. If all values in a variation series occur the same number of times, then such a series has no mode. If two values of a variation series have the same frequency and it is greater than the frequency of any other value, then such a variation series has two modes (bimodal).

• Median (Median, Me) – a variant located in the middle of an ordered variation series. When finding the median, two cases should be distinguished:

1. if the volume of the population n is an odd number and the options are ordered (written from smallest to largest), then the median will be the option that occupies the central position in the series. Its serial number can be found using the formula (n+1)/2, where n is the sample size;

2. if the volume of the population n is an even number, then the median is equal to half the sum of the options located in the middle of the ordered variation series:

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \qquad (2.6)$$

In medical research, the most commonly used average value is the arithmetic mean. However, if the values of the characteristics have a distribution other than normal, then to characterize the central tendency it is more reasonable to use the median rather than the arithmetic mean!

• Quantiles are separate equal parts into which the variation series is divided (Fig. 2.5):

- quartiles (Quartile) – values that divide the variation series into four equal parts;
- quintiles - values that divide the variation series into five equal parts;
- deciles (Decile) - values dividing the variation series into ten equal parts;
- percentiles (Percentile) - values that divide the variation series into one hundred equal parts

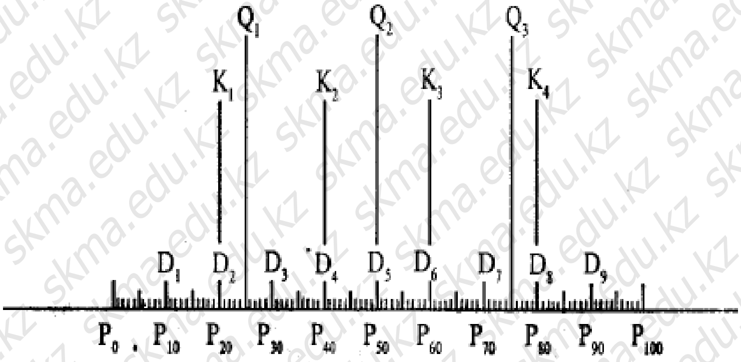


Figure2.5. Structural characteristics of the variation series

Quartiles are most often used in statistics.

The first or lower quartile (Q1) or 25th percentile (P25) is the value of a random variable below which 25% of the sample falls.

The second quartile (Q2) or 50th percentile (P50) is always equal to the median.

The third or upper quartile (Q3) or 75th percentile (P75) is the value of a random variable above which 25% of the sample falls (Figure 2.6).

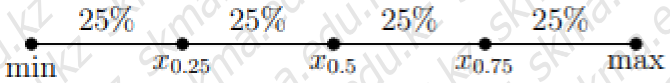


Figure. 2.6. Quartiles

To calculate quartiles, you need to divide the variation series by the median into two equal parts. If the number of options is even, then divide the row in half. If it is odd, then we divide the series into two parts and the median is included in each part. Then you need to find the middle of the row for each half. The resulting numbers will be the upper and lower quartiles, respectively.

Example 2.5. There are data on the duration of the disease in years for individual patients: 1, 7, 5, 19, 6, 12, 10, 20, 15. Determine quartiles.

Let's arrange the variation series: 1, 5, 6, 7, 10, 12, 15, 19, 20.

The number of observations is 9 – an odd number. The median of a series is a number with a serial number (n+1)/2=10/2=5. Me=10.

To determine quartiles, we divide the series into two halves: (1, 5, 6, 7, 10) and (10, 12, 15, 19, 20).

A median of 10 was included in each part. Next, we find the medians for each half:

(1, 5, 6, 7, 10) – the number is odd, the median is a number with a serial number (n+1)/2=6/2=3. Me=6. Thus, the upper quartile is Q1=6.

(10, 12, 15, 19, 20) - the number is odd, the median is a number with a serial number (n+1)/2=6/2=3. Me=15. Thus, the lower quartile of Q3=15.

Indicators of variability (diversity)

Variability indicators include: range, interquartile range, dispersion, standard deviation, coefficient of variation.

• Range of variation series (Range, R) – the difference between the largest and smallest value

$$R=x_{max}-x_{min}, \tag{2.7}$$

of the variant in the sample:

• Interquartile range (IQR) – the difference between the third and first quartiles:

$$IQR=Q3-Q1, \qquad (2.8)$$

• Dispersion (Variance, S2) is the average square of deviations of individual values of a characteristic from its average value (dimensionless value).

- simple variance: $$S^2=\frac{\sum_{n=1}^{n}(x_i-\bar{x})^2}{n-1} \qquad (2.9)$$

- weighted variance: $$S^2=\frac{\sum_{n=1}^{n}(x_i-\bar{x})^2\cdot v_i}{n-1} \qquad (2.10)$$

In the event that the dispersion is calculated for the general population in the denominator of the fraction instead of (n-1), you need to put n!

• Standard deviation (Standard Deviation, S or σ) - a measure of the spread of a random variable from its mean value, expressed in the same units as the options, defined as the square root of the

variance: $$V=\frac{S}{\bar{x}}\cdot 100\% \qquad (2.11)$$

To characterize any population that has a normal distribution, it is enough to know two parameters: the arithmetic mean and the standard deviation. This characteristic is written as follows (or).

• Coefficient of Variation (V) - a measure of the spread of a random variable, expressed as a percentage, defined as the ratio of the standard deviation to the average value of the characteristic:

$$V=\frac{S}{\bar{x}}\cdot 100\% \qquad (2.12)$$

The closer the coefficient of variation is to zero, the less variation in the values of the characteristic. The greater the coefficient of variation, the more variable the trait.

The population is considered homogeneous if the coefficient of variation does not exceed 33%.

When V<10%, the diversity of the series is considered weak, when 10%≤V≤20% - average, when V>20% - strong.

Example 2.6. In the city of Shymkent in 2023, the body weight of 7-year-old boys was measured (data are presented in Table 2.4). According to a similar study carried out in the same city, but in 2013, the average body weight of 7-year-old boys was 23.8 kg, s ±3.6 kg.

1) Calculate the arithmetic mean and indicators of diversity of the variation series (dispersion, standard deviation, coefficient of variation).

2) Evaluate the results obtained, compare their variability with the data of the previous study, and draw appropriate conclusions.

Table 2.4.

Results of measuring the body weight of 7-year-old boys in Shymkent in 2023.

Body mass	Middle of the interval (xi)	Number of boys (vi)	$x_i\cdot v_i$	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2\cdot v_i$
15-18,9	17	16	272	-7	19	784
19-22,9	21	27	567	-3	9	243
23-26,9	25	32	800	1	1	32
27-30,9	29	16	464	5	25	400
31-34,9	33	9	297	9	81	729
Total		100	2400			2188

Solution.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i * v_i}{\sum_{i=1}^{n} v_i} = \frac{2400}{100} = 24\text{kg} \qquad S^2 = \frac{\sum_{n=1}^{n}(x_i - \bar{x})^2 \cdot v_i}{n-1} = \frac{2188}{100-1} = 22{,}1$$

$$s = \sqrt{S^2} = \sqrt{22{,}1} = \pm 4{,}7 \qquad V = \frac{s}{\bar{x}} \cdot 100\% = \frac{4{,}7}{24} = 19{,}6\%$$

Conclusions:

1. The average body weight of 7-year-old boys in the city of Shymkent is 24 kg.
2. Dispersion is 22.1, standard deviation is ±4.7 kg, coefficient of variation is 19.6%.
3. The value of the coefficient of variation equal to 19.6% indicates the average diversity of the trait.

Thus, we can assume that the resulting average body weight is quite typical. Compared to 2013, in 2023 there is greater variability in body weight among 7-year-old boys (±4.7 kg versus ±3.6 kg). A similar conclusion follows from their comparison of the coefficients of variation (V in 2013 (3.6/23.8•100%) = 15.1%).

A Box and Whisker plot is often used to illustrate basic descriptive statistics.

For example, box-and-whisker plots were constructed for sample data on the heights of 30-year-old women (Figure 2.7 a, b).

When analyzing such graphs, you must definitely pay attention to the “legend”, i.e. symbols that are given at the bottom of the graph.

The first graph (Fig. 2.7, a) shows the average, minimum and maximum values, as well as the standard deviation. The second graph (Fig. 2.7, b) shows the values of the median, 25th and 75th

percentiles.

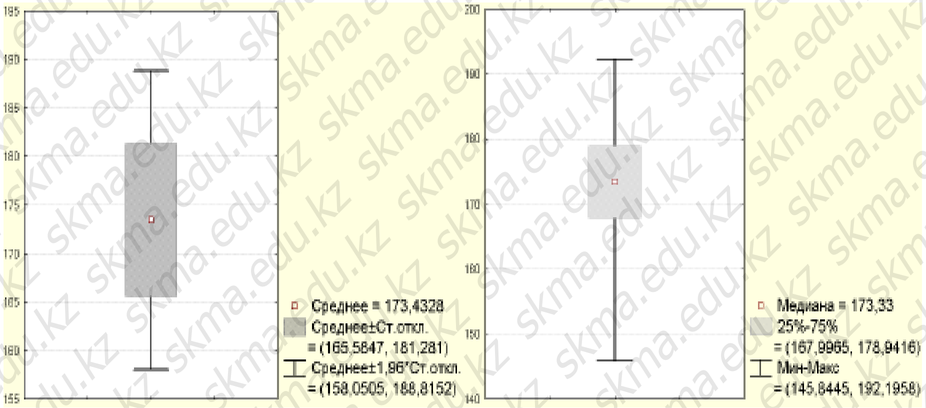


Figure 2.7. Displaying statistics on a box-and-whisker plot

4. Illustrative material: presentation, slides.

5. Literature:

• Basic:
1. Koychubekov B.K. Biostatistics. uch. allowance/ B.K. Koychubekov. - Almaty: Evero, 2016.
2. Boleshov M.Ә. Medical statistics: okulyk. – Almaty: Evero, 2015.
• Additional:
1. Fundamentals of statistical analysis in medicine: textbook. manual / ed. V.A. Reshetnikova. – M.: Medical Information Agency, 2020. – 176 p.
2. Mamaev A.N. Statistical methods in medicine. / A.N. Mamaev, D.A. Kudlay. – M.J practical medicine, 2021. – 136 p.
3. Gerasimov A.N. Medical statistics: Proc. allowance. – M.: Medical Information Agency, 2007. – 480 p.

6. **Control questions:**

1. What are descriptive statistics used for?
2. What is a variation series?
3. What types of variation series do you know?
4. What graphic tools can be used to represent a variation series?
5. What are averages used for?
6. What types of averages do you know?
7. What is the difference between simple and weighted quantities?
8. How to determine the mode, median and quartiles of a variation series?
9. By what criteria can the diversity of a trait be assessed?
10. What is the purpose of variance?
11. What is the purpose of standard deviation?
12. How is the value of the coefficient of variation interpreted?
13. What statistics can be used to construct a box-and-whisker plot?

LECTURE № 3

1. Theme: Normal distribution. Fundamentals of the theory of testing statistical hypotheses. Consent criteria.

2. Aim of the lecture: to develop students' understanding of the normal distribution and familiarize them with the basic concepts of the theory of testing statistical hypotheses.

3. Lecture thesis

Normal distribution.

In order to correctly choose a statistical method for analyzing the characteristic (variable) being studied, it is necessary to know its distribution law.

The law of distribution of a random variable is a correspondence established between all possible numerical values of a random variable and the probabilities (frequencies) of their occurrence in the aggregate.

The most frequently used in practice are the following types of distribution laws: binomial, Poisson (for discrete random variables); uniform, exponential, normal (for continuous random variables).

In the statistical analysis of medical data, the normal distribution is of greatest interest, since many biological and medical indicators (height, weight, cholesterol levels, blood pressure, temperature, blood counts, etc.) have distribution laws close to normal.

A normal (or Gaussian, or bell-shaped) distribution (Figure 3.1) is characterized by the fact that the largest number of observations have a value close to the mean, and the more the values differ from the mean, the fewer such observations.

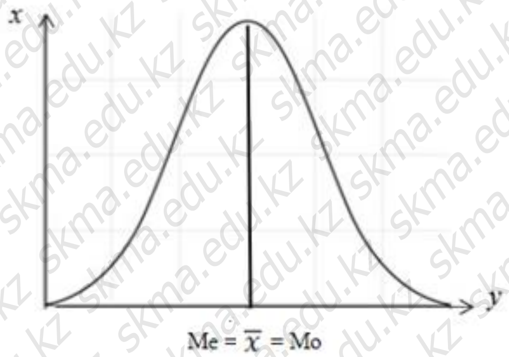


Figure. 3.1. Normal distribution of a variable

In Figure 3.1, the y-axis indicates the values that the attribute takes, and the x-axis indicates the frequency of occurrence of attribute values. The more often these values occur, the higher the curve. With a normal distribution, the highest frequency of occurrence occurs in the area of average values of the trait.

A random variable (feature, variable) x, subject to normal distribution, has a probability density function of the form:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad (3.1)$$

where μ is the mathematical expectation (- the sample mean is an estimate of the mathematical expectation);

σ – standard deviation (s – sample standard deviation is an estimate of σ).

Those. the bell-shaped curve is described by function (3.1).

The normal distribution curve has the following properties:

- has a bell shape, which means it is symmetrical relative to the average value;
- sample mean, mode and median are equal and correspond to the top of the distribution;
- asymptotically (infinitely close) approaches the x-axis;
- the area under the normal distribution curve is assumed to be 1 (or 100%);
- the shape of the curve depends on two parameters and ;
- “three sigma” rule (Fig. 3.2):

68.2% of all values of a normally distributed random variable lie in the interval ;
95.4% of all values of a normally distributed random variable lie in the interval ;
99.6%% of all values of a normally distributed random variable lie in the interval.

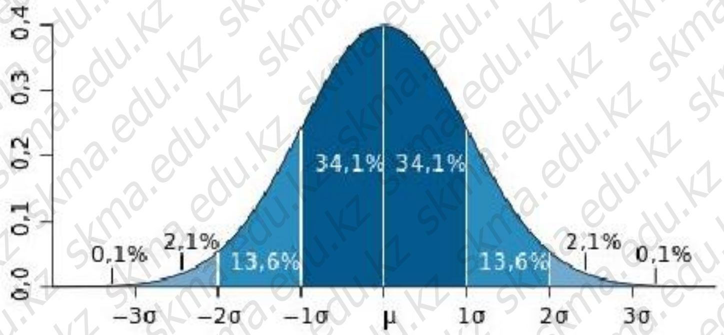


Figure. 3.2. Three Sigma Rule

In biology and medicine, asymmetric distribution is also common.

Unlike normal, this type of distribution is not symmetrical. The distribution can be “stretched” both to the left and to the right.

Asymmetry coefficient (Skewness, As) is an indicator that characterizes the skewness of the distribution:

$$As = \frac{\mu_3}{s^3}, \qquad (3.2)$$

where

$$\mu_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 \cdot \nu_i}{n}$$

If As>0, then the distribution is “stretched” to the left, and if As<0, then to the right. For a sample that has a normal distribution, the skewness coefficient is equal to or very close to zero (Fig. 3.3.a).

Kurtosis coefficient (Kurtosis, Ex) is an indicator that characterizes the severity of the distribution

peak:
$$Ex = \frac{\mu_4}{\sigma^4} - 3, \qquad (3.3)$$

where
$$\mu_4 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4 \cdot \nu_i}{n}$$

If Ex>0, then the distribution has a sharper peak than the normal distribution, if Ex<0, then the distribution has a more “flat” peak than the normal distribution. For a sample with a normal distribution, the kurtosis coefficient is close to zero (Fig. 3.3. b).

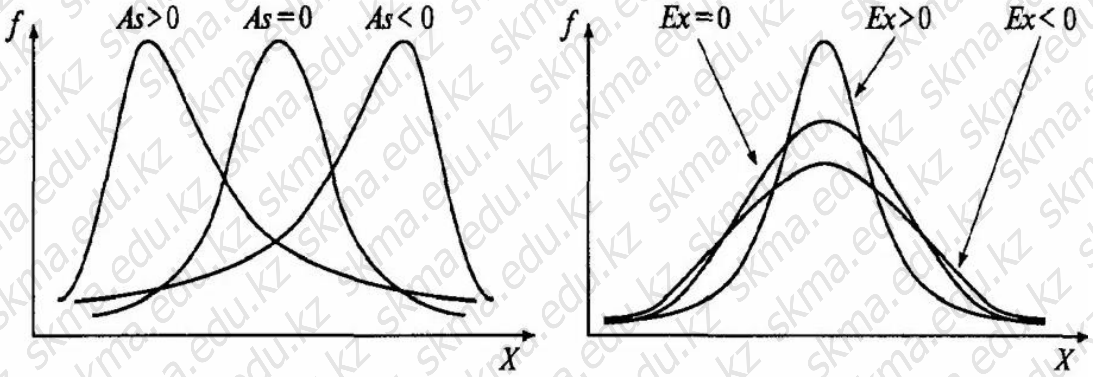


Figure 3.3. Graphic interpretation of skewness and kurtosis coefficient values

Basic concepts and definitions of the theory of statistical hypothesis testing.

When conducting scientific research in the field of medicine, health care and pharmacy, it is often necessary to make some judgment (hypothesis) based on the observations of a sample regarding the characteristics of the general population from which this sample is drawn that are of interest to the experimenter. That is, we are talking about testing statistical hypotheses.

The theory of statistical hypothesis testing is the main tool of evidence-based medicine.

A statistical hypothesis is some assumption about the numerical parameters of a known distribution (mean, variance, standard deviation) or about the form of the distribution law of a random variable.

Two hypotheses are always put forward: null (H0) and alternative (H1).

Null hypothesis H0 (main) - a hypothesis about the absence of differences between groups, or about certain parameter values, or about the correspondence of the distribution of a random variable to a certain law (for example, normal).

Alternative hypothesis H1 (competing) - a hypothesis about the existence of differences between groups, either about parameter values that differ from the given ones, or about the discrepancy of the distribution to a certain law.

The null hypothesis is formulated in such a way that it is the opposite of the research hypothesis that prompted the study. For example, the null hypothesis when comparing characteristics in two groups will always state that there are no differences between them, and the alternative that there are differences.

As a result of testing, the null hypothesis is either accepted or rejected in favor of the alternative. In this case, there is a risk of making two types of errors (Fig. 3.4).

	H_0 accepted	H_0 is rejected
H_0 is correct	The right decision	Type I error (α)
H_0 is false	Type II error (β)	The right decision

Figure 3.4. Errors of the first and second kind

An error of the first type is that the hypothesis H0 will be rejected, although in fact it is correct. The probability of making such an error is called the significance level (α).

For medical and pharmaceutical research, the significance level is α=0.05.

An error of the second type is that the hypothesis H0 will be accepted, but in fact it is incorrect (false). The probability of making such an error is called the confidence level (β).

The value 1-β is called the power of the criterion - this is the probability of rejecting an incorrect hypothesis.

For example, a patient is sick, but a blood test did not show this (false negative result). The doctor did not prescribe treatment for the patient and made a type I error.

For example, a patient is healthy, but a blood test shows the presence of a disease (false positive result). The doctor prescribed treatment for the patient and made a type II error.

To test the null hypothesis, statistical methods (tests or criteria) are used.

A statistical criterion (test) is a mathematical rule according to which it is determined whether or not the hypothesis of interest to the researcher corresponds to experimental (empirical) data.

A statistic is a function of sample observations on which the null hypothesis is accepted or rejected.

The observed (empirical, calculated) value of a criterion is a value that is calculated from a sample population that obeys a certain distribution law.

The acceptance area is the set of possible values of the statistical criterion at which the null hypothesis is accepted.

The critical region is the set of possible values of the statistical criterion at which the null hypothesis is rejected.

Critical points are points that delimit the critical area and the area where the hypothesis is accepted (Fig. 3.5).

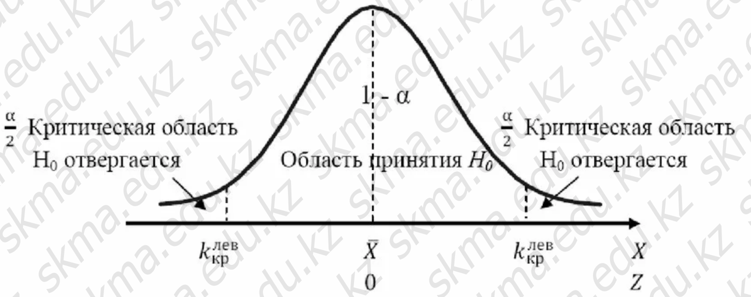


Figure 3.5. Critical region and hypothesis acceptance region

The type of area where the null hypothesis is accepted depends on the type of alternative hypothesis.

If the alternative hypothesis contains a greater than sign (for example), then the acceptance region will be right-sided.

If the alternative hypothesis contains a less than sign (for example), then the acceptance region will be left-sided.

If the alternative hypothesis contains an not equal sign (for example), then the scope of acceptance will be two-sided.

Scheme for testing statistical hypotheses:

1. Two hypotheses are put forward: the main (null) “H0” and the alternative “H1”.
2. Set the significance level α.
3. According to the initial data, i.e. based on the sample, the observed value of the criterion is calculated.
4. Using special statistical tables, the tabular value of the criterion is determined.
5. By comparing the observed and tabulated values, a conclusion is drawn about the correctness of a

particular hypothesis.

Consent criteria

Checking the sample population for compliance with normal distribution is the most important stage of scientific research, because The choice of statistical data analysis method depends on the results of testing the hypothesis about the normal distribution of the empirical population. Those. without answering the question “Is the sample to be studied normally distributed?” it is impossible to apply, much less correctly interpret, the results of statistical analysis.

To test the hypothesis about the normal distribution of the sample, goodness-of-fit tests are used.

Agreement criteria make it possible to determine when discrepancies between theoretical and empirical frequencies should be considered insignificant, i.e. random, and when - significant, i.e. non-random (Fig. 3.6).

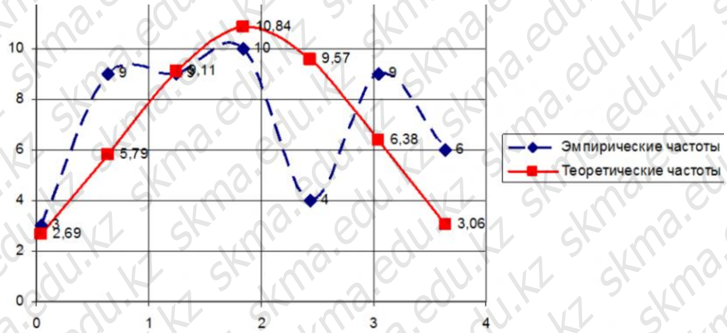


Figure 3.6. Comparison of empirical and theoretical frequencies

The most common goodness-of-fit tests are the χ2-Pearson and Kolmogorov-Smirnov tests.

Scheme for applying the χ2-Pearson goodness-of-fit test:

1) H0: random variable “X” is normally distributed.

H1: The random variable "X" is not normally distributed.

2) α=0.05 - significance level.

3) $$\chi^2_{calc} = \sum_{i=1}^{k} \frac{(\nu_i - \nu_i^*)^2}{\nu_i^*} \qquad (3.4)$$

where k is the number of intervals into which the empirical distribution is divided, is the observed frequency of the trait in the i-th group, is the theoretical frequency.

The theoretical frequency is calculated using the formula:

$$\nu_i^* = n \cdot p_i, \qquad (3.5)$$

where *pi* are the probabilities of a random variable falling into the interval [xi, xi+1], which are found by the formula:

$$p_i(x_i \le X \le x_{i+1}) = \Phi\left(\frac{x_{i+1} - \bar{x}}{s}\right) - \Phi\left(\frac{x_i - \bar{x}}{s}\right),$$

where is the sample mean, s is the standard deviation, Φ(x) is the distribution function of the normalized normal distribution.

For ease of calculation, fill out a table like 3.1. The amount calculated in the last column will be the calculated value.

Table 3.1.

Interval $[x_i, x_{i+1}]$	Empirical frequencies ν_i	Probabilities p_i	Theoretical frequencies ν_i^*	$(\nu_i - \nu_i^*)^2$	$\frac{(\nu_i - \nu_i^*)^2}{\nu_i^*}$

4) $\chi^2_{table}(\alpha; f)$,
where f=k-3 is the number of degrees of freedom for a normal distribution (tabular value), k is the number of intervals into which the sample is divided.

5) If $\chi^2_{calc} > \chi^2_{табл}$, then «$H_0$» accepted.

If $\chi^2_{calc} > \chi^2_{табл}$, then «$H_0$» rejected.

Pearson's goodness-of-fit test is used for a large number of observations (n>30), and the frequency of each group must be at least five.

Scheme for applying the Kolmogorov – Smirnov goodness of fit test:

1) H0: random variable “X” is normally distributed.

H1: The random variable "X" is not normally distributed.

2) α=0.05 - significance level.

3) $\lambda_{calc} = d_{\max}\sqrt{n}$, (3.7)

Where $d_{\max} = \max\left|F_n(x) - F(x)\right|$ - the maximum value of the absolute value of the difference between the observed distribution function Fn(x) and the corresponding theoretical distribution function F(x), n is the number of observations in the statistical series.

The values of the theoretical distribution function F(x) for the normal distribution are calculated using the formula: $F(x) = \frac{1}{2} + \Phi\left(\frac{x_{i+1} - \bar{x}}{s}\right)$, (3.8)

Where $\bar{x}$ is the sample mean, s is the standard deviation, Φ(x) is the Laplace function.

For ease of calculation, fill out a table like 3.2. The largest value in the last column will be the calculated value λ_{calc}

Interval $[x_i, x_{i+1}]$	Empirical frequencies ν_i	Accumulated frequencies $\nu^*_{i,\,накопл}$	Observed distribution function $F_n(x) = \frac{\nu_{i,накопл}}{n}$	Theoretical distribution function $F(x)$	$\left|F_n(x) - F(x)\right|$

1) $\Lambda_{table}=1, 36$ *(tabular value at α=0.05*

2) If $\lambda_{calc} \leq \lambda_{табл}$, then «$H_0$» accepted.

3) If $\lambda_{calc} > \lambda_{табл}$, then «$H_0$» rejected.

The Kolmogorov-Smirnov criterion is used for a large number of observations (n>30).

4. Illustrative material: presentation, slides.

5. Literature:

• Basic:
1. Koychubekov B.K. Biostatistics. uch. allowance/ B.K. Koychubekov. - Almaty: Evero, 2016.
2. Boleshov M.Ə. Medical statistics: okulyk. – Almaty: Evero, 2015.
• Additional:
1. Fundamentals of statistical analysis in medicine: textbook. manual / ed. V.A. Reshetnikova. – M.: Medical Information Agency, 2020. – 176 p.
2. Mamaev A.N. Statistical methods in medicine. / A.N. Mamaev, D.A. Kudlay. – M.J practical medicine, 2021. – 136 p.

3. Gerasimov A.N. Medical statistics: Proc. allowance. – M.: Medical Information Agency, 2007. – 480 p.

6. Questions:

1. What is the law of distribution of a random variable?
2. What types of distributions are most often used in practice?
3. Why is the normal distribution of greatest interest in statistical analysis of medical data?
4. What properties does a normal distribution curve have?
5. What parameters determine the shape and location of the normal distribution curve?
6. What indicators characterize asymmetric distribution?
7. What is called a statistical hypothesis?
8. What is the difference between the null and alternative hypotheses?
9. What is called an error of the first and second types?
10. What is the confidence level and significance level?
11. What is a statistical test?
12. What is the general scheme for testing statistical hypotheses?
13. What are consent criteria used for?

LECTURE № 4-5

1. Theme: Parametric methods of comparative statistics

2. Aim of the lecture: to develop students' understanding of parametric methods of comparative statistics, their practical application and interpretation of results in the context of medical research.

3. lecture thesis: Comparative statistics means conducting a comparative analysis of data in two or more groups. This is one of the main methods used in medicine and science in general to evaluate the effectiveness of various approaches, strategies, and technologies.

The most common task in conducting scientific research in the field of medicine is the comparison of data obtained through observations or experiments in different sample populations. For example, one sample is an experimental one (the researchers had an influence on the object or phenomenon being studied), and the second sample is a control sample (there was no influence on the object of observation).

If the researcher manages to notice any numerical differences in the characteristics of the compared samples, then the question arises: "What is the probability that these differences are non-random and will be systematically repeated in the future when reproducing the experimental conditions?", or in other words, "Are the identified differences statistically significant?"

The choice of an appropriate sample comparison method is determined by several factors:

- type of indicators being compared (quantitative or qualitative);
- type of distribution;
- the number of groups being compared;
- dependence or independence of samples.

Depending on the type of distribution of the samples under consideration, parametric and nonparametric methods (or statistical tests) can be applied to them.

Parametric criteria assume the presence of a normal distribution in the compared samples and use distribution parameters (means, variances, standard deviation) in the calculation process. For example, Student's t-test, Fisher's F-test, etc.

Nonparametric tests do not assume normal distribution in the samples being compared and use ranks (ordinal numbers) of attribute values in the calculation process. For example, Mann-Whitney test, Wilcoxon test, sign test, etc.

Nonparametric tests give slightly rougher estimates than parametric ones, but are more universal. Parametric methods are more accurate, but can only be used for normally distributed samples.

There are independent and dependent sample populations.

Independent (unrelated) samples are different groups of objects, characterized by the fact that the probability of selecting any object from one sample does not depend on the selection of any object from another sample.

Dependent (related) samples are the same group of objects, but studied at different points in time.

For example, a pharmaceutical company wants to test the effectiveness of a new drug to lower blood pressure. Data can be collected in two ways:

1 - One group of people is given a new drug and another a placebo, and then the blood pressure is compared between the groups. The samples are independent;

2 - blood pressure is measured in the same people before and after taking the drug. Samples are dependent.

Let's consider some parametric methods of comparative statistics.

1. Fisher's F-test

To test the hypothesis about the equality of variances of two samples, Fisher's F test is used. It is able to correctly estimate variances only if both samples are independent and have a normal distribution.

Scheme for applying the Fisher F test:

1) H_0: $s_1^2 = s_2^2$

H_1: $s_1^2 \neq s_2^2$

2) α=0.05 - significance level.

3) $$F_{calc} = \frac{s_1^2}{s_2^2} \qquad (4.1)$$

4) F_{table} (α;f1;f2) where f1=n1-1, f2=n2-1 – number of degrees of freedom.

5) If $F_{calc} \leq F_{table}$, then “Ho” is accepted.

If $F_{calc} > F_{table}$, then “Ho” is rejected.

2. Two sample t-test

To test the hypothesis about the equality of two sample means for independent samples, a two-sample Student t-test is used.

Rules for using the Student t-test:

1) both samples being compared must have a normal distribution;

2) it is possible to compare only two groups;

3) it is advisable to use this criterion for small samples (n<30), because increasing the sample size increases the sensitivity of the criterion, but with a significant increase in the number of observations it is possible to identify changes that are not significant;

4) it is necessary to take into account the presence/absence of homogeneity of variances (equality/inequality of variances) in the samples. To determine the homogeneity of variances, it is necessary to apply the Fisher F test.

Scheme for applying the two-sample Student t-test if the variances are equal:

1) H_0: $\bar{x}_1 = \bar{x}_2$

H_1: $\bar{x}_1 \neq \bar{x}_2$.

2) α=0.05 - significance level.

3) ,
$$t_{расч} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2} \cdot (n_1 + n_2 - 2)} \qquad (4.2)$$

where n1, n2 are the volumes of the samples under consideration s_1^2, s_2^2, are the variances of the samples under consideration $\bar{X}_1$, $\bar{X}_2$, and are the average values of the samples being compared.

4) $t_{tabl}(\alpha;f)$, where f= n1+n2-2 is the number of degrees of freedom.

5) If $t_{calc} > t_{табл}$, then “Ho” is accepted.

If $t_{calc} > t_{табл}$ then “Ho” is rejected.

Scheme for applying the two-sample Student t-test if the variances are unequal:

1) *H₀*: $\bar{x}_1 = \bar{x}_2$

H₁: $\bar{x}_1 \neq \bar{x}_2$.

2) *α*=0,05 – significance level

3) $$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \quad (4.3)$$

where n1, n2 are the volumes of the samples under consideration S_1^2 , S_2^2 , are the variances of the samples under consideration $\bar{X}_1$, $\bar{X}_2$, and are the average values of the samples being compared.

4) $t_{tabl}(\alpha;f)$, where $f = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}}$ - number of degrees of freedom.

5) if $t_{calc} \leq t_{табл}$, then «*H₀*» accepted.

if $t_{calc} > t_{табл}$ then «*H₀*» rejected

3. Paired t-test

To test the hypothesis about the equality of two sample means for dependent samples, the paired Student's t-test is used.

Scheme for applying the paired Student t-test:

1) H₀: $\overline{x_1} = \overline{x_2}$,

H₁: $\overline{x_1} \neq \overline{x_2}$

2) α=0,05- significance level

3) $$t_{расч} = \frac{\bar{d}}{S_d / \sqrt{n}}, \quad (4.4)$$

where d=xi1-xi2 - differences between the corresponding values of pairs of variables, n - sample size,

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}, \quad S_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$$

4) $t_{табл}(\alpha;f)$, where *f*= *n* – 1 - number of degrees of freedom.

5) if $t_{calc} \leq t_{table}$, then «*H₀ accepted*

if $t_{calc} > t_{table}$, then «*H₀*» rejected

4. One sample t-test

This criterion is intended to test the hypothesis that the sample mean is equal to any value.

Scheme for applying the one-sample t-test:

1) H₀: $\bar{x} = a$,

H₁: $\bar{x} \neq a$

2) α=0,05- significance level

3 $t_{calc} = \frac{x-a}{s/\sqrt{n}}$ (4.5)

4) $t_{tabl}(\alpha;f)$ where $f= n – 1$ - - number of degrees of freedom

5) if $t_{calc} \leq t_{table}$, then «H_0» accepted
if $t_{calc} > t_{table}$, then «H_0» rejected

5. One-way Analysis of Variance (ANOVA)

Analysis of variance is used to test the hypothesis about the equality of sample means in the case when more than two independent samples (k>2) having a normal distribution are considered.

Scheme for using one-way analysis of variance, in the case when the variances are equal:

1) H_0: $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_k$.

H_1: $\bar{x}_1 \neq \bar{x}_2 \neq \ldots \neq \bar{x}_k$.

2) α=0,05- significance level

3)

3.1) Overall average, $\bar{x} = \frac{\bar{x}_1+\bar{x}_2+\cdots+\bar{x}_k}{k}$.

3.2) factor sum of squared deviations,$SS_{\text{факт}} = r\sum\left(\bar{x}_{\text{гр }j} - \bar{x}\right)^2$, (4.6)

where r is the number of values in each sample.

3.3) residual sum of squared deviations

$$SS_{\text{ост}} = \sum_{i=1}^{r}\left(x_{i1} - \bar{x}_{\text{гр1}}\right)^2 + \sum_{i=1}^{r}\left(x_{i2} - \bar{x}_{\text{гр2}}\right)^2 + \cdots + \sum_{i=1}^{r}\left(x_{ik} - \bar{x}_{\text{гр}k}\right)^2 \quad (4.7)$$

where k is the number of samples

3.4) factor variance , $S^2{}_{\text{факт}} = \frac{SS_{\text{факт}}}{k-1}$. (4.8)

3.5) residual variance $S^2{}_{\text{ост}} = \frac{SS_{\text{ост}}}{k(r-1)}$. (4.9)

3.6) $F_{\text{расч}} = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}$. (4.10)

4) F_{table} $(\alpha;f_1;f_2)$, where $f_1=k-1$, $f_2=k(r-1)$ – number of degrees of freedom (k - number of samples, r - number of values in each sample).

5) if $F_{calc} \leq F_{table}$, then «H_0» accepted
if $F_{calc} > F_{table}$, then «H_0» rejected

4. Illustrative material: presentation, slides.

5. Literature:

- Basic:

1. Koychubekov B.K. Biostatistics. uch. allowance/ B.K. Koychubekov. - Almaty: Evero, 2016.

- Additional:

1. Mamaev A.N. Statistical methods in medicine. / A.N. Mamaev, D.A. Kudlay. – M.J practical medicine, 2021. – 136 p.

6. Control questions:

1. What factors determine the choice of an appropriate method for comparing samples?
2. What is the difference between parametric and non-parametric statistical methods?
3. What is the difference between dependent and independent samples?
4. To test which hypothesis is Fisher's F test used? What is his scheme?
5. To test which hypothesis is the Student's t-test used? What are the conditions for its use?
6. What is the difference between two-sample and paired Student's t tests?
7. What is the design of the two-sample Student's t test?

8. What is the design of the paired Student's t test?
9. What hypothesis is the one-sample t-test used to test? What is his scheme?
10. To test which hypothesis is one-way analysis of variance used? What is his scheme?

LECTURE № 6

1. Theme: Nonparametric methods of comparative statistics
2. Aim of the lecture: to develop students' understanding of nonparametric methods of comparative statistics, their practical application and interpretation of results in the context of medical research.
3. Lecture thesis: Depending on the type of distribution of the samples under consideration, parametric and nonparametric methods (or statistical tests) can be applied to them.

Nonparametric methods are used in two cases: when sample data does not have a normal distribution or when there is a small number of observations.

To calculate nonparametric criteria, a procedure is used to rank the values of the characteristic being compared, i.e. arranging values in ascending order.

Rank is the ordinal number of the attribute value.

If the numbers are not repeated, then their ranks correspond to their serial numbers. If a certain number is repeated several times, then all of them are assigned an average rank.

For each parametric criterion there is at least one non-parametric analogue.

Let's look at some nonparametric methods of comparative statistics.

1. Mann-Whitney U-test

This rank test is used to test the hypothesis of equality of sample means in the case of two independent samples that do not have a normal distribution.

This test is a nonparametric analogue of the two-sample Student t-test.

The U test is suitable for comparing small samples. Each sample must have at least 3 characteristic values. It is allowed that there are 2 values in one sample, but then the second must have at least five (n1, n2≥3 or n1=2, n2≥5).

The condition for applying the Mann-Whitney U test is the absence of matching attribute values in the compared groups (all numbers are different) or a very small number of such matches.

Scheme for applying the Mann-Whitney U test:

1) H_0: $\bar{x}_1 = \bar{x}_2$
H_1: $\bar{x}_1 \neq \bar{x}_2$.
2) α=0.05 - significance level.
3) Calculate the differences for each case (BEFORE-AFTER). The absolute values of the differences are ranked (that is, they are ranked modulo without taking into account the sign), without taking into account the zeros. Assign signs ("+" or "–") of differences to each rank. Receive iconic ranks. Trasch is defined as the smallest of the values of T+ and T-, which are the sums of the positive and negative ranks, respectively.
4) Ttable (α; n).
5) If $T_{crit} > T_{table}$, then "H_0" is accepted.
If $T_{crit} \leq T_{table}$, then "H_0" is rejected

3. Kruskal-Wallis H-test

This criterion is a nonparametric analogue of one-way analysis of variance and is used to compare three or more independent groups that do not have a normal distribution.

When comparing three samples, it is allowed that each of them has at least 3 observations, or one of them has 4 observations, and the other two have 2 each; in this case, it does not matter which sample contains the number of subjects, but the ratio of 4:2:2 is important.

The table of critical values of the H-criterion is provided only for the case when the number of samples is k≤5, and the number of subjects in each group is ni≤8. With a large number of samples and subjects in each sample, it is necessary to use the table of critical values of the χ2 test, because the Kruskal-Wallis test asymptotically approaches the "χ2" distribution.

Scheme for applying the Kruskal–Wallis H test:

1. *H₀:* $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_k$.

   H₁: $\bar{x}_1 \neq \bar{x}_2 \neq \ldots \neq \bar{x}_k$.

2. α=0.05 - significance level.
3. Calculate the differences for each case (BEFORE-AFTER). The absolute values of the differences are ranked (that is, they are ranked modulo without taking into account the sign), without taking into account the zeros. Assign signs ("+" or "–") of differences to each rank. Receive iconic ranks. Trasch is defined as the smallest of the values of T+ and T-, which are the sums of the positive and negative ranks, respectively.

4) Ttable (α; n).

5) If $T_{crit} > T_{table}$, then "H_0" is accepted.

   If $T_{crit} \leq T_{table}$, then "H_0" is rejected

Kruskal-Wallis H-test

This criterion is a nonparametric analogue of one-way analysis of variance and is used to compare three or more independent groups that do not have a normal distribution.

When comparing three samples, it is allowed that each of them has at least 3 observations, or one of them has 4 observations, and the other two have 2 each; in this case, it does not matter which sample contains the number of subjects, but the ratio of 4:2:2 is important**..**

The table of critical values of the H-criterion is provided only for the case when the number of samples is k≤5, and the number of subjects in each group is ni≤8. With a large number of samples and subjects in each sample, it is necessary to use the table of critical values of the χ2 test, because the Kruskal-Wallis test asymptotically approaches the "χ2" distribution

Scheme for applying the Kruskal–Wallis H test:

1) *H₀:* $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_k$.

   H₁: $\bar{x}_1 \neq \bar{x}_2 \neq \ldots \neq \bar{x}_k$.

2) α=0.05 - significance level.

3) $$H_{calc} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

where $n = \sum_{i=1}^{k} n_i$ is the total number of observations for all groups, R_i is the sum of the ranks of the i-th sample.

In the case when the number of sample выборок $k \leq 5$ H_{table} (α;n_1;n_2;…;n_5), where n_1, n_2, …, n_5 – are the volumes of the samples under consideration.

In the case when the number of samples $k>5$ $H_{table}=\chi^2_{tablr}(\alpha\ ;\ f)$, где $f=k$-1 is the number of degrees of freedom (tabular value).

5) if $H_{calc} < H_{table}$, then «H_0» accepted

   if $H_{calc} \geq H_{table}$, then «H_0» rejected

4. Illustrative material: presentation, slides

5. Literature:.

- Basic:

1. Koychubekov B.K. Biostatistics. uch. allowance/ B.K. Koychubekov. - Almaty: Evero, 2016.

2. Boleshov M.Ə. Medical statistics: okulyk. – Almaty: Evero, 2015.

• Additional:

1. Mamaev A.N. Statistical methods in medicine. / A.N. Mamaev, D.A. Kudlay. – M.J practical medicine, 2021. – 136 p.

2. Fundamentals of statistical analysis in medicine: textbook. manual / ed. V.A. Reshetnikova. – M.: Medical Information Agency, 2020. – 176 p.

6. Control questions:

1. In what cases are nonparametric comparison methods used?
2. What do the concepts “rank” and “ranking” mean?
3. To test which hypothesis is the Mann-Whitney U test used? What are the conditions for its use?
4. What is the design of the Mann-Whitney U test?
5. To test which hypothesis is the Wilcoxon T-test used? What are the conditions for its use?
6. What is the design of the Wilcoxon T-test?
7. To test which hypothesis is the Kruskal-Wallis H test used? What are the conditions for its use?
8. What is the design of the Kruskal-Wallis H test?

LECTURE № 7

1. Theme: Analysis of qualitative characteristics

2. Aim of the lecture: to familiarize students with some methods of analyzing qualitative features in the context of medical research.

3. Lecture thesis

Qualitative features are characteristics that are not measured by numerical values, but describe the quality of an object (for example, gender, blood type, eye color, hair color, marital status, presence/absence of disease, place of residence (urban/rural), etc.).

Analysis of qualitative characteristics allows you to systematize and classify information without reducing it to numerical values.

Qualitative data analysis is important for identifying relationships between various factors when conducting medical research. Allows you to identify risk groups, clarify diagnoses and make informed decisions in terms of treatment and prevention.

There are several types of qualitative features: binary, nominal and ordinal.

Binary features- are features that can take only two possible values. For example, gender (male/female), presence/absence of disease, blood pressure normal/high, postoperative complications present/no.

Nominal features- are features that divide objects into categories, but do not have an orderly relationship between these categories. For example, eye color, blood type, country of residence. There is no obvious order between categories.

Ordinal features- are features that divide objects into categories that have a certain order, but the differences between the values are not equal. For example, education (no, secondary, higher), severity of the disease (mild, moderate, severe). There is an order, but the spacing between values may not be equal.

Constructing contingency tables is an important method for analyzing qualitative data, especially in medical research. Such tables allow you to explore connections between two or more qualitative variables and determine the degree of their relationship.

Consider a contingency table of size 2x2 (Table 6.1). There are two binary signs: A - with outcomes A1, A2; B - with outcomes B1, B2. The central part of the table contains the frequencies of occurrence of combinations of these characteristics - a, b, c, d.

Table 6.1.

Conjugation table size 2x2

B A	B_1	B_2	Sum
A_1	*a*	*b*	*a+b*
A_2	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*n=a+b+c+d*

Example 6.1. Table 6.2 presents data on the use of two types of drugs (A and B) by patients with the same diagnosis and their condition (sick/healthy) after 5 days of use. A total of 39 patients were examined, of which 25 were healthy, 14 were sick. Drug A was taken by 18 people, drug B by 21.

Let's consider a contingency table of size rxs (Table 6.3). There are two signs: A - with outcomes A1, A2, ..., Ar and B - with outcomes B1, B2, ..., Bs. The central part of the table contains the frequencies of occurrence of various combinations of characteristics A and B - νij.

Table 6.2.

Conjugation table size 2x2

State Type of medicine	Healthy	Sick	Sum
A	10	8	18
B	15	6	21
Sum	25	14	39

Table 6.3.

Conjugation table for size 5 x 5

A B	B_1	B_2	…	B_s	Sum
A_1	ν_{11}	ν_{12}	…	ν_{1s}	$\nu_{1.}$
A_2	ν_{21}	ν_{22}	…	ν_{2s}	$\nu_{2.}$
…	…	…	…	…	…
A_r	ν_{r1}	ν_{r2}	…	ν_{rs}	$\nu_{r.}$
Sum	$\nu_{.1}$	$\nu_{.2}$	…	$\nu_{.s}$	$\nu_{..}$

Example 6.2. In table 6.4. data on the number of observations and cases of mortality for four forms of acute purulent destruction of the lungs are presented.

Table 6.4.

Conjugation table size 4x2

Exodus Form diseases	Deaths	Recovery	Sum
Purulent abscess	5	136	141
Gangrenous abscess	11	37	48
Gangrene of the lobes	7	8	15
Total gangrene	6	5	11
sum	29	186	215

Contingency tables can be used to test the hypothesis about the statistical significance of the relationship between qualitative variables.

1. Pearson's chi-squared test

This criterion is used to analyze qualitative characteristics in independent samples if the frequencies in the contingency table cells are greater than or equal to 5.

Scheme for applying the Pearson χ2 test (rxs size table):

1) H_0: there is no connection between qualitative characteristics.

H_1: there is a connection between qualitative characteristics.

2) α=0.05 - significance level.

3) $$\chi^2_{расч} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(\nu_{ij}-\nu^*_{ij}\right)^2}{\nu^*_{ij}} \qquad (6.1)$$

where ν_{ij} are observed frequencies ν^*_{ij} are theoretical (expected) frequencies.

Theoretical (expected) frequencies are calculated using the formula:

$$\nu^*_{ij} = \nu_{.i}\cdot\frac{\nu_{j.}}{\nu_{..}} \qquad (6.2)$$

where ν.i is the sum for column i, νj. – sum for line j, ν.. – total number of observations**.**

4) χ^2_{table} (α ; f), where $f=(r-1)(s-1)$ - number of degrees of freedom

5) If □^{2} $_{calc}$□ □$_{table}$, then “H0” is accepted

6) If □^{2} $_{calc}$ □ □$_{table}$ then “H0” is rejected.,

Yates's correction.

Formula (6.3) gives overestimated values. In practice, this results in the null hypothesis being rejected too often. To compensate for this effect, the Yates correction is introduced into the formula:

$$\chi^2_{calc} = \frac{\left(ad-bc-\frac{n}{2}\right)^2 n}{(a+b)(c+d)(a+c)(b+d)} \qquad (6.4)$$

When solving problems, you need to perform calculations using formulas (6.3) and (6.4)

2. Fisher’s exact test

This criterion is used to analyze qualitative characteristics in independent samples. It is suitable for comparing very small samples (if the observed frequencies are less than 5). Used only for contingency tables of size 2x2.

Scheme for applying Fisher's exact test:

1) H_0: there is no connection between qualitative characteristics.

H_1: there is a connection between qualitative characteristics

2) α=0.05 - significance level.

3) $$P_{calc} = \frac{(a+b)!\cdot(c+d)!\cdot(a+c)!\cdot(b+d)!}{a\,!\cdot b\,!\cdot c!\cdot d\,!\cdot n!} \qquad (6.5)$$

where a, b, c, d – observed frequencies, n – total number of observations, ! – factorial is the product of a number and a sequence of numbers, each of which is less than the previous one by 1 (for example, 4!=4•3•2•1).

5) The calculated value of the criterion is compared with the significance level of 0.05.

If $P_{calc} \geq 0{,}05$, then “H0” is accepted.

If $P_{calc} < 0{,}05$, then “H0” is rejected.

4. Illustrative material: presentation, slides.

5. Literature:

- Basic:

1. Koychubekov B.K. Biostatistics. uch. allowance/ B.K. Koychubekov. - Almaty: Evero, 2016.

- Additional:

1. Mamaev A.N. Statistical methods in medicine. / A.N. Mamaev, D.A. Kudlay. – M.J practical medicine, 2021. – 136 p.

2. Boslaf S. Statistics for everyone. / Per. from English P.A. Volkova, I.M. Flamer, M.V. Liberman, A.A. Galitsyn. – M.: DMK Press, 2017. – 586 p.: ill.

6. Control questions:

1. What is the peculiarity of the analysis of qualitative characteristics?
2. Why is qualitative feature analysis important in medical research?
3. What is a contingency table of 2x2 and size rxs?
4. What conditions must be met when applying the Pearson $\chi2$ test?
5. What is the scheme of the Pearson $\chi2$ test?
6. What is the Yates correction used for?
7. In what cases is Fisher's exact test used?
8. What is the design of Fisher's exact test?
9. In what cases is McNemar's $\chi2$ test used?
10. What is the design of McNemar's $\chi2$ test?

LECTURE № 8

1.Theme: Correlation Analysis

2. Aim of the lecture: To familiarize students with the fundamentals of correlation analysis.

3. Lecture thesis:

One of the important tasks of epidemiology is the analysis of morbidity in relation to risk factors. A **risk factor** in medicine is a factor that contributes to the occurrence of a disease (for example, smoking is a risk factor for myocardial infarction or cancer; the number of accidents in the water supply system is a risk factor for dysentery).

Correlation analysis is used for the quantitative assessment of risk factors in disease development.

Correlation analysis is a quantitative method for determining the strength and direction of the relationship between two or more random variables.

The term *"correlation"* was first introduced into scientific use by the French paleontologist G. Cuvier (18th century), and in statistics it was first applied by F. Galton (19th century).

To numerically characterize the relationship between variables, the concept of the **correlation coefficient** is introduced.

The **correlation coefficient** is an indicator that characterizes the strength and direction of a relationship and takes values in the interval **[-1, 1]**.

To assess the strength of the relationship, **Chaddock's scale** is used in correlation theory (Table 7.1).

Table 7.1. Quantitative Measure of the Strength of Association

Correlation coefficient	Qualitative characteristic
0.1 – 0.3	Weak
0.3 – 0.5	Moderate
0.5 – 0.7	Noticeable
0.7 – 0.9	High
0.9 – 1.0	Very strong

According to direction, **direct** and **inverse** correlation relationships are distinguished.

A **direct correlation** is a relationship in which an increase in one variable is associated with an increase in another variable (for example, an increase in dysentery incidence with an increase in the proportion of substandard water samples in the water supply).

An **inverse correlation** is a relationship in which an increase in one variable is associated with a decrease in another variable (for example, a decrease in hepatitis B incidence as vaccination coverage increases).

For a direct relationship, the correlation coefficient ranges from **0 to +1**.
For an inverse relationship, it ranges from **–1 to 0**.

If the correlation coefficient equals **0**, there is no relationship between the phenomena.
If it equals **+1 or –1**, the relationship is **functional**.

To analyze the dependence between two variables, **scatter plots** are used.

A **scatter plot** is a visual way of presenting a correlation relationship between two variables (Figure 7.1).

A scatter plot is a point diagram in the form of a graph obtained by plotting experimental observation points on a specific scale. The coordinates of the points correspond to the values of the studied variable and the influencing factor. The distribution of points shows the presence and nature of the relationship between the variables.

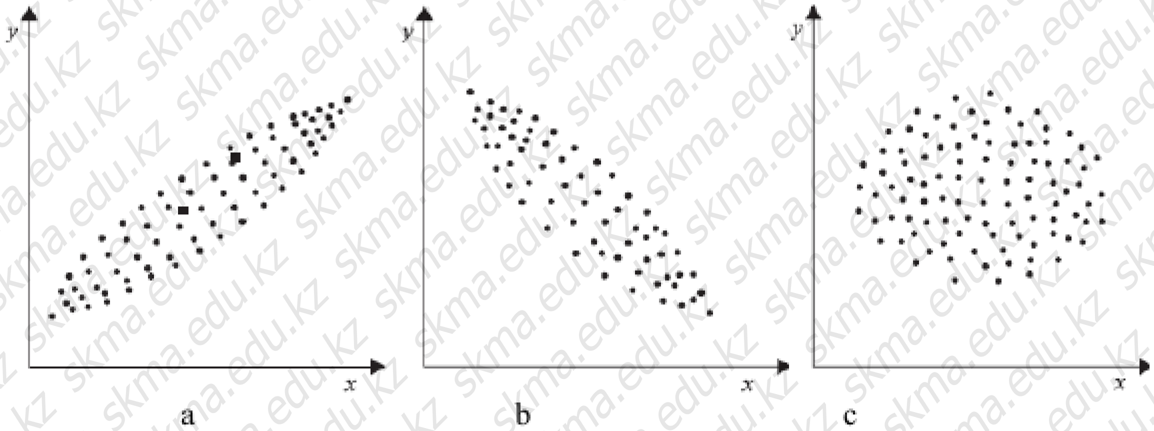


Figure 7.1. Scatter plots:
a – direct relationship; b – inverse relationship; c – no relationship.

The **linear (pairwise) Pearson correlation coefficient** characterizes the strength and direction of the relationship, $r_{xy} = \dfrac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2 \sum (y-\overline{y})^2}}$, where r_{xy} is the correlation coefficient; x and y are correlated series; $\bar{x}$ and $\bar{y}$ are mean values.

The **pairwise correlation coefficient** is a **parametric coefficient**.

The Pearson correlation coefficient can be applied if the following conditions are met:

- the compared variables are measured on an interval or ratio scale;
- the distributions of the variables are close to normal;
- the number of values for both variables is the same.

The **statistical significance** of the correlation coefficient is determined by comparing it with its **standard error**.

The **standard error of the correlation coefficient**, $m_r = \pm \dfrac{1-r_{xy}^2}{\sqrt{n}}$, where r_{xy} is the correlation coefficient and n is the number of observations.

The correlation coefficient is considered statistically significant if it exceeds its standard error by **three times**. Otherwise, the number of observations must be increased.

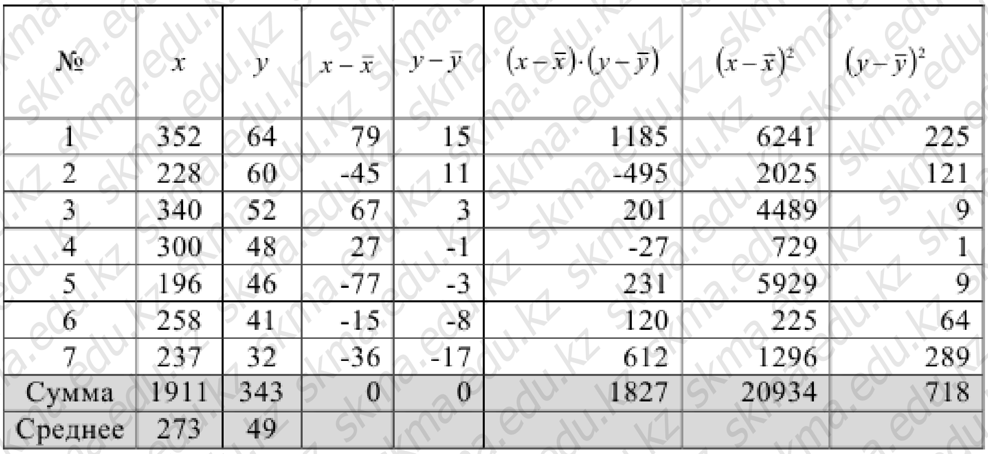
The significance of the correlation coefficient can also be determined using special tables.

Example 7.1 Calculate the Pearson linear correlation coefficient for the following data:

Incidence of ARI per 1,000 population (x):	352	228	340	300	196	258	237
Incidence of pneumonia per 1,000 population (y):	64	60	52	48	46	41	32

Solution:

1. Construct a calculation table.



№	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\cdot(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	352	64	79	15	1185	6241	225
2	228	60	-45	11	-495	2025	121
3	340	52	67	3	201	4489	9
4	300	48	27	-1	-27	729	1
5	196	46	-77	-3	231	5929	9
6	258	41	-15	-8	120	225	64
7	237	32	-36	-17	612	1296	289
Сумма	1911	343	0	0	1827	20934	718
Среднее	273	49					

2. Calculate the correlation coefficient.

$$r_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{1827}{\sqrt{20934\cdot 718}} = 0{,}47.$$

3. Analyze the result: the relationship is **direct and moderate**.

4. Calculate the standard error of the correlation coefficient: the coefficient is **not statistically significant**, as it does not exceed the standard error threefold.

$$m_r = \pm\frac{1-r_{xy}^2}{\sqrt{n}} = \pm\frac{1-0{,}47^2}{\sqrt{7}} = 0{,}3,$$

In the analysis of clinical and pharmaceutical phenomena, the following **nonparametric correlation coefficients** are often used:

- Spearman’s rank correlation;
- Kendall’s τ (tau);
- Yule’s association coefficient;
- Pearson’s contingency coefficient;
- Chuprov’s coefficient of association;
- Gamma (γ), etc.

Let us consider the **Spearman rank correlation coefficient**, developed and proposed in 1904 by **C. E. Spearman**, an English psychologist and professor at the University of London and Chesterfield University.

The **rank correlation coefficient** measures the relationship between the ranks of a given variable according to different characteristics.

Spearman’s rank correlation coefficient is used to determine the strength of relationships between both quantitative and qualitative variables, provided that their values can be ranked in ascending or descending order.

Spearman’s rank correlation coefficient, where *n* is the sample size, and $(r_{xi} – r_{yi})$ is the difference between the ranks of the *i*-th object.

The qualitative interpretation of the strength of the rank correlation coefficient, as with other correlation coefficients, can be assessed using **Chaddock’s scale**.

Spearman’s rank correlation coefficient is applied when the sample size satisfies the condition **$5 \leq n \leq 40$**.

Example 7.2 In one settlement, a chronic epidemic of Flexner dysentery was recorded. Preliminary analysis and laboratory studies showed frequent occurrences of substandard bacteriological water samples in the water supply network (risk factor). It is necessary to test the hypothesis of a relationship between these two indicators.

Month	**Number of dysentery cases (x)**	**Proportion of non-standard water samples (y)**
January	10	0
February	9	0.5
March	2	1.1
April	7	2.0
May	6	1.8
June	11	2.9
July	26	6.7
August	32	4.5
September	46	8.7
October	38	7.1
November	8	3.2
December	5	0

Solution:

№	x	y	r_x	r_y	$r_x - r_y$	$r_x - r_y$□^{2}
1	2	0	7	1,5	5,5	30,25
2	9	0,5	6	3	3	9
3	2	1,1	1	4	-3	9
4	7	2,0	4	6	-2	4
5	6	1,8	3	5	-2	4
6	11	2,9	8	7	1	1
7	26	6,7	9	10	-1	1
8	32	4,5	10	9	1	1
9	46	8,7	11	12	-1	1
10	38	7,1	12	11	1	1
11	8	3,2	5	8	-3	9

12	5	0	2	1,5	0,5	0,25
SUMM						70,5

Construct a calculation table.

$$\rho = 1-\frac{6}{n^3-n}\sum_{i=1}^{n}\left(r_{x_i}-r_{y_i}\right)^2 = 1-\frac{6}{12^3-12}\cdot 70.5 \approx 0,75.$$

$$m_r = \pm\frac{1-r_{xy}^2}{\sqrt{n}} = \pm\frac{1-0,75^2}{\sqrt{12}} \approx 0,12,$$

- Calculate the correlation coefficient.
- Analyze the result: the relationship is **direct and strong**.
- Calculate the standard error of the correlation coefficient: the coefficient is **statistically significant**, as it exceeds the standard error more than threefold.

4. Illustrative material:Presentations, slides.

5 Literature:

• Basic:

1. Koychubekov B.K. Biostatistics. uch. allowance/ B.K. Koychubekov. - Almaty: Evero, 2016.

• Additional:

1. Mamaev A.N. Statistical methods in medicine. / A.N. Mamaev, D.A. Kudlay. – M.J practical medicine, 2021. – 136 p.

2. Boslaf S. Statistics for everyone. / Per. from English P.A. Volkova, I.M. Flamer, M.V. Liberman, A.A. Galitsyn. – M.: DMK Press, 2017. – 586 p.: ill.

6. Control Questions:

1. Why is correlation analysis used in epidemiological studies?
2. Within what limits does the correlation coefficient vary?
3. Why are scatter plots needed?
4. What formula is used to calculate the Pearson correlation coefficient?
5. How is the statistical significance of the correlation coefficient determined?
6. In which cases is Spearman’s rank correlation coefficient used?
7. What formula is used to calculate Spearman’s rank correlation coefficient?