Ministry of Internal Affairs of Ukraine
DNIPROPETROVSK STATE UNIVERSITY
OF INTERNAL AFFAIRS
ECONOMIC AND INFORMATION SECURITY
DEPARTMENT

Svitlana NASONOVA

THEORY OF PROBABILITY AND MATHEMATICAL STATISTICS

Textbook for students of economic specialties

Language of presentation: **English**

Dnipro
2022

UDC 519.2
N 31

Recommended by the scientific and methodical council of Dnipropetrovsk State University of Internal Affairs (protocol No. 10 dated 15.06.2022)

AUTHOR: **Svitlana Nasonova** – Associate Professor of the Department of Economic and Information Security, Candidate of Technical Sciences, Associate professor

REVIEWERS:

Kuznetsov V. M. – Head of the Department of Higher Mathematics of the Ukrainian State University of Science and Technology, Doctor of Physical and Mathematical Sciences, Professor;

Olevskyi V. I. – Head of the Department of Higher Mathematics of the Ukrainian State University of Chemistry and Technology, Doctor of Technical Sciences, Professor.

Nasonova S. S.

N 31 **Theory of probability and mathematical statistics** : textbook for students of economic specialties / Svitlana Nasonova. Dnipro : Dnipropetrovsk State University of Internal Affairs, 2022. 152 p.

ISBN 978-617-8032-71-5

The textbook contains the necessary theoretical information and formulas on the theory of probability and mathematical statistics. Much attention is paid to statistical methods for processing experimental data. Examples of solving practical problems are considered. There are tasks for independent solution and questions for self-study.

It is intended for English-speaking students of economic specialties.

Насонова С. С.

Н 31 **Теорія ймовірностей та математична статистика** : навч. посібник для студентів економічних спеціальностей / Світлана Насонова. Дніпро : Дніпроп. держ. ун-т внутр. справ, 2022. 152 с.

У навчальному посібнику наведені необхідні теоретичні відомості та формули з теорії ймовірностей та математичної статистики. Велику увагу приділено статистичним методам обробки експериментальних даних. Розглянуті приклади розв'язання практичних завдань. Наведені завдання для самостійного виконання та питання для самоперевірки.

Призначається для англомовних студентів економічних спеціальностей.

ISBN 978-617-8032-71-5

CONTENTS

INTRODUCTION

We often have a need to solve problems in which the result of the action is not unambiguously determined. For example, the owner of the store does not know how many customers there will be during a certain period, the businessman does not know – what will be the euro exchange rate tomorrow, the banker does not know – whether the loan will be repaid.

People in practical activities encounter uncertainty. Uncertainty is a phenomenon, the observation of which gives different results.

Thus, the *subject of probability theory* is the study of the laws of mass random events. Knowledge of these regularity allows us to foresee how these events will occur.

The purpose of the studying the discipline «Theory of probability and mathematical statistics» is: 1) to provide students with theoretical knowledge and practical skills in probability theory and mathematical statistics, which will be necessary for the understanding of economic disciplines; 2) to develop of such a level of knowledge and mathematical culture, which will allow to understand and analyze processes and laws in economics and the world around.

The basis for the study of the discipline «Theory of probability and mathematical statistics» is the discipline «Further mathematics».

By studying the discipline, students will learn: 1) *at the conceptual level*: basic concepts and methods of the theory of probability and mathematical statistics, the scope of statistical methods in ensuring financial and economic security; 2) *at practical and creative levels*: features of working with specific statistical problems of professional activity. They will be able to: 1) *at the algorithmic level:* to use the acquired knowledge and practical skills to solve typical problems of professional activity; 2) *at the heuristic level:* to use the acquired knowledge and practical skills to solve atypical complex problems of professional activity using models of mathematical statistics; 3) *at the creative level*: to find new areas of application of methods of probability theory and mathematical statistics and processes to increase the efficiency of professional activities.

The textbook contains theoretical material, examples of solving practical problems, tasks for independent solution. At the end of each chapter there are conclusions, as well as questions for self-test (in the form of the test), practical tasks. In this textbook you can also find literature recommended for self-study, as well as a list of questions for the final control.

The material presented on the pages of this textbook can be considered as a general educational minimum, necessary for everyone who for the first time begins to study the basics of probability theory and mathematical statistics in order to apply the knowledge gained in practical activities.

Chapter 1
BASIC CONCEPTS OF THE THEORY OF PROBABILITY

1.1. Formation of the theory of probability as a science

«The Theory of Probability is common sense, supported by calculations» – said Marquis de Laplace.

The random nature of events and processes was noted in ancient times. The ancient Greek philosopher Epicurus believed that chance is inherent in the very nature of phenomena, and, therefore, chance is objective. There were attempts to create a mathematical approach to the study of random events, but the first mathematical calculations of probabilities appeared in written documents only in the middle of the 17th century.

The initial impetus for the development of the theory of probability was the problems related to gambling, such as craps, cards, roulette, when they began to use quantitative calculations and predicting the chances of success. The need for the development of a new science was dictated by the tasks that were set in the insurance business, as well as in demography, or, as they said then, political arithmetic. From the end of the 17th century, insurance against accidents and natural disasters began to be produced on a scientific basis. In the 16th-17th centuries, ship insurance and fire insurance became widespread in all countries of Western Europe. In the 18th century, numerous insurance companies and lotteries were created in Italy, Flanders, and the Netherlands.

In 1654, the entire scientific (and not only) community of Paris started talking about the emergence of a new science – the theory of probability. The foundations of this theory were laid not in scientific work, but in correspondence between two famous French mathematicians Blaise Pascal and Pierre de Fermat about a problem concerning the game of dice. It all started with the fact that the courtier of the French king, Chevalier de Mere, himself a gambler, turned to Blaise Pascal with questions about the problem of points. Two famous questions of de Mere to Pascal have come down to us: 1) how many times do you need to throw two dice so that there are more than half of the total number of throws of two sixes at once; 2) how to fairly divide the money wagered if the players stopped the game prematurely? In 1654, Pascal approached the mathematician Pierre de Fermat and corresponded with him about these problems. Together they established some

of the initial provisions of the theory of probability, in particular, they came to the concept of mathematical expectation and the theorems of addition and multiplication of probabilities. The results of Pascal and Fermat got acquainted with the Dutch physicist and mathematician Christian Huygens, who wrote the work “On Calculations in Gambling”. This work is considered the first book on probability theory.

The formation of the theory of probability is associated with the name of the famous Swiss mathematician Jacob Bernoulli. In his treatise "The Art of Assumptions", on which he worked for 20 years, the classical definition of probability was first introduced, as well as the statistical concept of probability. The next important stage in the development of the theory of probability is associated with the names of Moivre, Laplace, Gauss, Poisson. Further, in the 19th century, representatives of the Petersburg mathematical school, such as Bunyakovsky, Chebyshev, Markov, Lyapunov, were engaged in the development of the theory of probability.

Today, the theory of probability has developed into the universal theory that finds application in many areas of human activity. It is widely used in economics, transport, manufacturing, statistics, military affairs. Modern natural science widely uses the theory of probability as a theoretical basis in processing the results of observations.

1.2. Types of random events

In the theory of probability, as in every mathematical discipline, there are some concepts that underlie it. The main concepts in probability theory are the concepts of stochastic experiment, random event and probability of a random event.

A stochastic (random) experiment is the process of implementing a certain set of conditions an unlimited number of times, the results of which cannot be predicted in advance. Here is an example.

Let’s look at an example of a stochastic experiment:

1. Consider the possibility of winning the lottery. Before the draw, it is not possible to predict whether there will be a win on this ticket, and if there will be a win, which one. Thus, the lottery can also be considered as a stochastic experiment.

The concept of a *random event* is associated with the implementation of a certain set of conditions.

A random event is an event that may or may not occur as a result of multiple experiments. Events are indicated by capital letters of the Latin

alphabet: A, B, C.

A reliable event is an event that must occur as a result of an experiment. An event that cannot occur as a result of an experiment is *impossible.*

Let's look at an example.

If there are only white balls in the box, then getting a white ball from the box is a reliable event, and getting a ball of another color from this box is an impossible event.

Two or more random events are called *equally possible* if the conditions under which they occur are the same and they have the same chance of occurring.

Examples of *equally possible events*:

1. The appearance of the coat of arms or number in one toss of the coin;

2. Falling out of an even or odd number of points in one toss of the dice.

An example of *unequal events* is the falling out of two points and an odd number of points in one dice toss.

Two or more events are *compatible* if the occurrence of one of them in the trial does not exclude the occurrence of the other events.

Two or more events are called *incompatible* if, in the same trial, the occurrence of one of them exclude the occurrence of the remaining events.

Examples of incompatible events:

1. The appearance of the coat of arms and numbers in one toss of the coin;

2. The appearance of an even and odd number of points in one toss of the dice.

An example of compatible events – falling out of one point and an odd number of points in one dice toss.

A group of incompatible events A1, A2… An is called a *complete group of events* if one and only one event of this group is bound to occur as a result of the trial.

Examples of a complete group of incompatible events:

1. The appearance of the coat of arms and a number in one toss of a coin;

2. The appearance of black and red suits when removing the card from the deck;

3. Getting an even and odd number of points in one toss of a dice.

Two incompatible events that form a complete group are called *opposite.* An event opposite to event A is denoted by $\bar{A}$.

Consider some operations we can perform on events.

The sum of two events A and B is called an event C that means that

either event A or event B will appear, or A and B will appear together. Denote the sum of events as follows: C = A + B (or C = A ∪ B).

The sum of several events is an event that consists in the appearance of at least one of these events.

Examples.

1. If two shots are fired from a cannon and event A is a hit from the first shot, event B is a hit from the second shot then A + B is a hit from the first or second shot, or a hit from both shots.

2. A company has made a large profit through the following activities. Let event A – increase in production, event B – decrease in production costs. Then A + B is a random event, the meaning of which is that the company either increased production or reduced its cost, or at the same time increased the volume of production and reduced production costs.

The product of two events A and B is called an event C, which consists in the compatible occurrence of event A and event B. The product of events is denoted as follows: C = AB (or C = A ∩ B).

Examples.

1. Let event A means that a detail is standard, event B means that the detail is painted, then event AB means that the detail is standard and painted.

2. Let event A means that the first shooter hit the target, and event B means that the second shooter hit the target. Then event AB means that both the first and second shooters hit the target.

The difference of two events A and B is called an event C that occurs only when event A occurs, but event B does not occur. The difference of the two events is denoted as follows: C = A-B (or C = A \ B).

Let's look at an example.

If event A is a hit in domain A, and event B is a hit in domain B then A-B is a hit in the difference of these domains. It is a set of points that belong to domain A and do not belong to domain B.

An event $\overline{A}$ is called *the opposite of event A* if event $\overline{A}$ occurs if and only if event A does not occur.

It should be noted that events A and $\overline{A}$ form a complete group.

Let's look at an example:

A dice is tossed once. Event A is the appearance of the coat of arms. Then the opposite event $\overline{A}$ is the appearance of the number.

1.3. Elements of combinatorics

Let's consider two rules that will be useful to us when we solve tasks.

Sum rule.

If object A can be selected from the set of objects in m ways, and object B can be selected from the set of objects in n ways, then one of the objects A or B can be selected in m + n ways.

Product rule.

If object A can be selected from the set of objects in m ways and each time object A is selected, object B can be selected in n ways, then pair (A, B) can be selected in m·n ways.

Let's look at the next definition. *The factorial* of a nonnegative integer n (denote n!) is the product, calculated by the formula:

$$n! = 1 \cdot 2 \cdot 3 \ldots \cdot n, \text{ where } 0! = 1, 1! = 1.$$

Permutations are groups consisting of n elements, that differ only in the order of the elements. The number of all permutations of n elements is equal to

$$P_n = n!$$

Permutations with repetitions. If among n elements are the same, then the formula of permutations with repetitions has the form

$$P_n(n_1, n_2, \ldots n_k) = \frac{n!}{n_1! n_2! \ldots n_k!}$$

where n_1 is the number of repetitions of the first element; n_2 is the number of repetitions of the second element; n_3 is the number of repetitions of the third element, and so on, n_k is the number of repetitions of the k-th element.

It should be understood that

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

Placements of n elements by m elements are combinations that contain m elements taken from these n elements that differ from each other either by elements or by the order of the elements. The number of all possible placements of n elements by m elements is determined by the formula

$$A_n^m = \frac{n!}{(n-m)!}$$

Combinations are groups composed of n different elements by m elements that differ from each other by at least one element. The number of combinations on a set of n elements by m elements is calculated by the formula

$$C_n^m = \frac{n!}{m!\,(n-m)!}$$

Examples.

1. How many ways can you place 12 people at a table with 12 chairs?

Solution. The number of ways is equal to P_{12} = 12! = 479 001 600.

2. How many permutations can be obtained from the letters that make up the word «mathematics»?

Solution. The required number is equal to the number of permutations with repetitions, namely the number of permutations of 11 elements, among which the letter «m» is repeated 2 times, the letter «a» is repeated 2 times, the letter «t» is repeated 2 times. Therefore, the total number of permutations is equal to:

$$P_{11}(2,2,2) = \frac{11!}{2! \cdot 2! \cdot 2!} = 1\,663\,200$$

3. Second-year students study 8 disciplines. How many ways can you schedule classes for Friday if you need to schedule three lectures on that day?

Solutions. The number of ways is equal to the number of placements of 8 elements by 3 elements

$$A_8^3 = \frac{8!}{(8-3)!} = \frac{8 \cdot 7 \cdot 6 \cdot 5!}{5!} = 336$$

1.4. Classical and statistical definition of probability

The main concept of probability theory is probability. The word «probability» is often used in everyday life. Everyone is familiar with the phrase: «It will probably snow tomorrow», or «Most likely I will go to nature this weekend», or «It's just incredible», or «There is a chance to get a credit automatically». Such phrases intuitively estimate the probability that some random event will occur. In turn, mathematical probability gives some numerical estimate of the probability that some random event will occur.

The classical probability of event A is equal to the ratio of the number of elementary events favorable to event A to all equally possible and pairwise incompatible elementary events in this experiment:

$$P(A) = \frac{m}{n},$$

where m is the number of elementary events favorable to event A; n is the number of all equally possible and pairwise incompatible elementary events in this experiment.

The relative frequency of the occurrence of event A in a series of n trials will be called the ratio

$$W(A) = \frac{m}{n},$$

where m is the number of cases in which event A occurred.

Examples.

1. The letters E, X, E, R, C, I, S, E, S are written on the cards. Find the probability that the word "EXERCISES" will appear in a row when the cards are randomly laid out.

Solution. In this case, the number of all elementary events is equal to the number of permutations with repetitions on the set of 9 elements. Let's write down the formula

$$n = P_9(3,2) = \frac{9!}{3! \cdot 2!} = \frac{3! \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9}{3! \cdot 2} = 30240$$

The number of elementary events favorable to event A (event A is that the word "EXERCISES" will appear in a row), is equal to 1. Thus, the required probability is found by the formula:

$$P(A) = \frac{1}{30240} \approx 0{,}00003$$

2.Among 100 products, the technical control department found 8 non-standard ones. What is the relative frequency of non-standard products?

Solution. Let A is a random event, which consists in the appearance of a non-standard product, then by the definition of the relative frequency of event A we obtain

$$W(A) = \frac{m}{n} = \frac{8}{100} = 0{,}08$$

1.5. Theorems of addition and multiplication of probabilities

Let's consider theorems about the probability of the sum and product of events.

The theorem of addition of probabilities. The probability of the sum of two compatible events is equal to the sum of the probabilities of these events minus the probability of the compatible occurrence of the same events.

$$P(A+B) = P(A) + P(B) - P(AB)$$

Corollary 1. The probability of the sum of two incompatible events A and B is equal to the sum of their probabilities

$$P(A+B) = P(A) + P(B)$$

Corollary 2. For a set of pairwise incompatible events $\{A_1, A_2, \dots A_n\}$, which form a complete group, this formula has the form:

$$P(A_1 + A_2 + \cdots + A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) = 1$$

Corollary 3. If A and $\overline{A}$ are opposite events (that is A and $\overline{A}$ form a complete group of two incompatible events), then there is an equality

$$P(A) + P(\overline{A}) = 1$$

Remark. This property of opposite events is very useful in solving many tasks, when the probability of the event A is better determined by the probability of the opposite event. That is, we have a formula

$$P(A) = 1 - P(\overline{A})$$

Examples.

1. The probability that the day will be cloudy P (A) = 0,7. Find the probability that the day will be clear.

Solution. Events "the day will be clear" and "the day will be cloudy" are opposite, so the probability that the day will be clear is equal to P (A) = =1− 0,7 = 0,3.

2. The group consists of five students who learn only German and nine students who learn only English. Find the probability that two randomly selected students are learning the same language.

Solution. Event A – two students learn German. Event B – two students learn English. These events are incompatible. We apply the theorem about the probability of the sum of two incompatible events. Let's find probabilities P(A) and P(B). These probabilities are calculated using the formula of classical probability:

$$P(A) = \frac{C_5^2}{C_{14}^2} \qquad P(B) = \frac{C_9^2}{C_{14}^2}$$

$$C_{14}^2 = \frac{14!}{2! \cdot 12!} = \frac{13 \cdot 14}{1 \cdot 2} = 91 \qquad C_5^2 = \frac{5!}{2! \cdot 3!} = \frac{4 \cdot 5}{1 \cdot 2} = 10$$

$$C_9^2 = \frac{9!}{2! \cdot 7!} = \frac{8 \cdot 9}{1 \cdot 2} = 36 \qquad P(A) + P(B) = \frac{C_5^2 + C_9^2}{C_{14}^2} = \frac{10 + 36}{91} = \frac{46}{91}$$

2. The probability that a worker will make 10 details in one shift (event A1) is equal to 0,15; the probability of making 9 details in one shift (event A2) is equal to 0,2; the probability of making 8 details or less (event A3) is equal to 0,65. Find the probability that the worker will make at least 9 details in one shift.

Solution. Let event A means that the worker will make at least 9 details, that is either 9 or 10 details will be made.

Method 1: Since events A1 and A2 are incompatible, then the probability of event A is calculated by the formula

$$P(A) = P(9) + P(10) = 0{,}15 + 0{,}2 = 0{,}35$$

Method II: Let event A means that the worker will make at least 9 details in one shift. This means that either 9 or 10 details will be made. Then the opposite event $\overline{A}$ means that less than 9 parts will be made in one shift. This means that the worker will make 8 parts. So we can write: $P(\overline{A})=0{,}65$. Then

$$P(A) = 1 - P(\overline{A}) = 1 - 0{,}65 = 0{,}35$$

The theorem of multiplication of probabilities. The probability of compatible occurrence of two events is equal to the product of the probability of one of them on the conditional probability of the other, calculated under the assumption that the first event has already occurred

$$P(A \cdot B) = P(A) \cdot P_A(B) = P(B) \cdot P_B(A)$$

Corollary 1. The probability of the compatible appearance of several events is equal to the product of the probability of one of them on the conditional probabilities of all others, and the probability of each next event is calculated under the assumption that all previous events have already occurred

$$P(A_1 \cdot A_1 \cdot A_3 \cdot \ldots \cdot A_n) = P(A_1) \cdot P_{A_1}(A_2) \cdot P_{A_1 \cdot A_2}(A_3) \cdot \ldots \cdot P_{A_1 \cdot A_2 \cdot \ldots \cdot A_{n-1}}(A_n)$$

Corollary 2. If two events A and B are independent, then $P_A(B) = P(B)$. That is the probability of the compatible appearance of events A and B is equal to

$$P(A \cdot B) = P(A) \cdot P(B)$$

Let's look at an example.

1. There are 3 white and 3 black balls in the box. One ball is taken out of it twice at random, without turning them back. Find the probability of removing first the white and then the black ball.

Solution. Let's give the definition of the random events A and B. Let random event A means that the white ball was removed first, random event B – the black ball was removed by the second. The sought probability is equal to

$$P(A \cdot B) = P(A) \cdot P_A(B)$$

According to the classical definition of probability we can write

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

After the first trial, there were 5 balls left in the box, including 2 white and 3 black (we have already removed one white ball). Thus

$$P_A(B) = \frac{3}{5}$$

From here

$$P(A \cdot \mathrm{B}) = \frac{1}{2} \cdot \frac{3}{5} = \frac{3}{10} = 0{,}3$$

CONCLUSIONS ON THE TOPIC

1. *A stochastic (random) experiment* is the process of implementing a certain set of conditions an unlimited number of times, the results of which cannot be predicted in advance.

2. *A random event* is an event that may or may not occur as a result of multiple experiments. Events can be *reliable* and *impossible, equally possible*, *compatible* and *incompatible, opposite*. If one and only one event of the group of incompatible events is bound to occur as a result of the trial then this group of incompatible events is called a *complete group of events*.

3. *Combinatorics* is a branch of mathematics devoted to solving problems related to the selection and arrangement of elements of a certain

finite set in accordance with given rules. When solving combinatorial problems, the following formulas are used: permutations, placements and combinations.

4. If object A can be selected from the set of objects in m ways, and object B can be selected from the set of objects in n ways, then one of the objects A or B can be selected in m + n ways (*Sum rule*). If object A can be selected from the set of objects in m ways and each time object A is selected, object B can be selected in n ways, then pair (A, B) can be selected in m · n ways (*Product rule*).

5. *The probability* of event A is equal to the ratio of the number of elementary events favorable to event A to all equally possible and pairwise incompatible elementary events in the experiment (classical definition of probability).

6. *The probability of the sum of two events* is equal to the sum of the probabilities of these events minus the probability of the compatible occurrence of the same events.

$$P(A + B) = P(A) + P(B) - P(AB)$$

If A and B are *incompatible events* then

$$P(A + B) = P(A) + P(B)$$

If A and $\overline{A}$ are *opposite events*, then

$$P(A) + P(\overline{A}) = 1$$

7. *The probability of compatible occurrence of two events* is equal to the product of the probability of one of them on the conditional probability of the other, calculated under the assumption that the first event has already occurred

$$P(A \cdot B) = P(A) \cdot P_A(B) = P(B) \cdot P_B(A)$$

If two events A and B are *independent*, then

$$P(A \cdot B) = P(A) \cdot P(B)$$

SELF-TEST QUESTIONS

Please choose the correct answer from the options below.

1. A *reliable event* is an event that:

a. may or may not occur as a result an experiment;
b. must occur as a result of an experiment;
c. cannot occur as a result of an experiment.

2. An *impossible event* is an event that:

a. may or may not occur as a result an experiment;
b. cannot occur as a result of an experiment;
c. must occur as a result of an experiment.

3. Which of the following random events are *equally possible events*?

a. event A1: the appearance of the coat of arms in one toss of the coin; event A2: the appearance of the number in one toss of the coin.

b. event A1: falling out of two points in one dice toss; event A2: falling out of seven points in one dice toss.

c. event A1: choosing a white ball from a box containing only white balls; event A2: choosing a black ball from the box containing only white balls.

4. Which of the following random events are *compatible events*?

a. the appearance of the coat of arms (event A1) and number (event A2) in one toss of the coin;

b. the appearance of an even (event A1) and odd number of points (event A2) in one toss of the dice;

c. falling out of one point (event A1) and an odd number of points (event A2) in one dice toss.

5. If the probability P (A) of event A is known, then the probability of the *opposite event* is equal to

a. 1-P(A)
b. P(A)-1
c. P(A)

6. The probability of a random event A can be found by the formula:

a. m/n
b. n/m
c. n/(m-1)

7. The probability of a random event A is a number that belongs to the interval (segment)

a. [0,1]
b. (0,1)
c. [0,1)

8. If A and B are incompatible events, then the probability of the sum of these events P (A + B) is equal to

a. P(A)+P(B)+P(AB)
b. P(B)-P(AB)
c. P(A)+P(B)

9. If A and B are independent events, then the probability of their compatible occurrence P (AB) is equal to:

a. P(A)*P(B)
b. P(A)*P(A/B)
c. P(B)*P(B/A)

10. Which of the events is random?

a. the number pi is an irrational number
b. two plus two is four
c. departure of the airplane in bad weather

PRACTICAL TASKS

1. How many different permutations can be formed from all the letters of the word *fortune*?

Answer: $P_7 = 5040$

2. How many three-digit numbers can you make up of the digits 1, 2, 3 if each digit appears in the number record only once?

Answer: $P_3 = 6$

3. How many ways can you take 2 details from a box containing 10 details?

Answer: $C_{10}^2 = 45$

4. The coin is tossed once. Find the probability that the coat of arms will fall out.

Answer: $P = 1/2$

5. The dice are tossed once. Find the probability of getting: 1) even

number of points; 2) odd number of points.

Answer: 1) $P = 1/2$; *2)* $P = 1/2$

6. 10 tickets are drawn in the lottery. The prize falls on 1 ticket. Bought three tickets. What is the probability that one of them will win?

Answer: $P = 0{,}3$

7. A four-digit number is chosen at random. Find the probability that the number is a multiple of five?

Answer: $P = 0{,}2$

8. Two dice are tossed. What is the probability of getting two sixes?

Answer: $P = 1/36$

9. One box contains 4 white and 8 black balls, and the other one contains 3 white and 9 black balls. One ball was taken out of each box. Calculate the probability that both balls will be white.

Answer: $P = 1/12$

10. Two shooters fired one shot each. The probability of hitting the target by the first shooter is 0,6; for the second shooter this probability is equal to 0,8. Find the probability that at least one shooter will hit the target.

Answer: $P = 0{,}92$

LITERATURE FOR SELF-STUDY

1. A.V. Tyurinand, A.Yu. Akhmerov Theory of probability and mathematical statistics: Textbook. − Dusseldorf: LAP LAMBERT Academic Publishing GmbH & Co.KG., 2020. – 148 p.

2. Prasanna Sahoo PROBABILITY AND MATHEMATICAL STATISTICS: Textbook. – USA: Department of Mathematics of the University of Louisville, 2013. – 712 p.

3. J. K. Blitzstein, J. Hwang, Introduction to Probability Second Edition. – Taylor & Francis Group, LLC, 2019. – 636 p.

Chapter 2
FORMULA OF COMPLETE PROBABILITY. BAYES' FORMULA. BERNOULLI'S FORMULA. LOCAL AND INTEGRAL THEOREMS OF LAPLACE

2.1. Formula of complete probability

Let event A occur only if one of the incompatible events B_1, B_2, B_n, which form a complete group, occurs. Then the probability of event A is equal to the sum of products of the probabilities of events B_1, B_2,…,B_n by the corresponding conditional probability of event A:

$$P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) + \cdots + P(B_n) \cdot P_{B_n}(A),$$

where $P(B_1) + P(B_2) + \cdots + P(B_n) = 1.$

This equality is called the *formula of complete probability*. Events B1, B2,…, Bn are called *hypotheses.*

Let's look at an example.

A group of 30 students take a probability theory exam. Ten of them have studied all the material and therefore the probability of passing the exam is 90%. Twelve students studied 75% of the questions. For them the probability of passing the exam is 60%. The rest of the students go to the exam without studying anything. For them the probability of passing the exam is 0,001. What is the probability that a randomly selected student will pass the exam?

Solution. Event A – a randomly selected student will take an exam. Let's put forward hypotheses:

B1 – randomly selected student is a student who has studied the material by 90%;

B2 – randomly selected student is a student who has studied the material by 75%;

B3 – randomly selected student is a student who has studied nothing.

Find the probabilities of hypotheses B_1, B_2, …, B_n:

$$P(B_1) = \frac{10}{30} = \frac{1}{3} \qquad P(B_2) = \frac{12}{30} = \frac{2}{5} \qquad P(B_3) = \frac{8}{30} = \frac{4}{15}$$

Events B_1, B_2, B_3 form a *complete group*. Really,

$$P(B_1) + P(B_2) + P(B_3) = 1.$$

Determine the *conditional probabilities* of event A:

$$P_{B_1}(A) = 0{,}9; \quad P_{B_2}(A) = 0{,}6; \quad P_{B_3}(A) = 0{,}001.$$

Then, using the *formula of complete probability*, we can find the probability of event A:

$$P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) + P(B_3) \cdot P_{B_3}(A) \approx 0{,}54$$

2.2. Probability of hypotheses. Bayes' formula

Let event A occur only if one of the incompatible events B1, B2,…,Bn which form a complete group appears. If event A has already occurred then probabilities of hypotheses can be overestimated by the Bayes' formula:

$$P_A(B_i) = \frac{P(B_i) \cdot P_{B_i}(A)}{P(A)}, (i = 1, 2, \dots, n)$$

where $P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) + \cdots + P(B_n) \cdot P_{B_n}(A)$.

Let's look at an example.

There are 10 rifles in the pyramid, 4 of them are equipped with an optical sight. The probability that the shooter will hit the target when firing from a rifle with an optical sight is equal to 0,95. For a rifle without an optical sight, this probability is equal to 0,8. The shooter hit the target with a random rifle. What is more likely: the shooter fired from a rifle with or without an optical sight?

Solution. Let event A – hitting a target. Let's put forward some hypotheses: B_1- shot from a rifle equipped with an optical sight; B_2 – shot from a rifle without an optical sight. Find probabilities of the hypotheses:

$$P(B_1) = \frac{4}{10} = 0{,}4 \qquad P(B_2) = \frac{6}{10} = 0{,}6$$

Find conditional probabilities of event A:

$$P_{B_1}(A) = 0{,}95 \qquad P_{B_2}(A) = 0{,}8$$

According to the formula of complete probability, we calculate the probability of event A:

$$P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) = 0{,}4 \cdot 0{,}95 + 0{,}6 \cdot 0{,}8 = 0{,}86$$

Let's overestimate probabilities of hypotheses:

$$P_A(B_1) = \frac{P(B_1) \cdot P_{B_1}(A)}{P(A)} = \frac{0{,}4 \cdot 0{,}95}{0{,}86} \approx 0{,}44$$

$$P_A(B_2) = \frac{P(B_2) \cdot P_{B_2}(A)}{P(A)} = \frac{0{,}6 \cdot 0{,}8}{0{,}86} \approx 0{,}56$$

So, most likely, the shooter fired a rifle without an optical sight.

2.3. Bernoulli's formula. Local and integral theorems of Laplace

Until this time, we considered probabilities of events associated with *single trials.* Now let's look at such schemes in which trials are repeated.

Repeated trials are the gradual execution of the same trial (n times) or *n* identical trials at the same time. For example, you can toss a coin 10 times in a row or toss 10 identical coins at the same time.

If repeated trials are performed and the probability of event A in each trial does not depend on the results of other trials, then such trials are called *independent of event A.*

Examples of independent trials: gradual tossing (n times) of a symmetrical coin; gradual throwing (n times) of a dice; gradual throwing of a ball (n times) into the basket.

Bernoulli's formula. Suppose that n independent trials are performed, in each of which the probability of occurrence of event A is equal to p (0<p<1). The probability that event A will occur exactly m times is calculated by the formula

$$P_n(m) = C_n^m \cdot p^m \cdot q^{n-m}, \qquad q = 1 - p$$

Remark. It should be noted that the Bernoulli's formula is used for a small number of independent trials (n is approximately equal to 20 or n≈20).

If the number of trials is large enough, and the probability p of event A in each trial is small enough, you should use the *Poisson's formula*

$$P_n(m) \approx \frac{\lambda^m}{m!} e^{-\lambda}, \qquad \lambda = np.$$

It should be noted that m is the number of occurrences of event A in n independent trials, λ is the average number of occurrences of event A in n trials. When both parameters p and q are large enough (but not more than 1), the local and integral Laplace theorems are used.

Local theorem of Laplace. The probability that in n independent trials, in each of which the probability of occurrence of event A is equal to p, event A will occur exactly m times, is calculated by the formula

$$P_n(m) \approx \frac{1}{\sqrt{npq}}\varphi(x_m), \qquad x_m = \frac{m-np}{\sqrt{npq}}$$

Remark. Values of the function φ (x) for non-negative x is given in the table (see appendix); for negative values of x use the same table, taking into account, that φ (–x) = φ (x).

Integral Laplace theorem. The probability that in n independent trials, in each of which the probability of occurrence of event A is equal to p event A will occur not less than m_1 times and not more than m_2 times is calculated by the formula

$$P(m_1, m_2) = \Phi(\ddot{x}) - \Phi(\dot{x}),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}dz - \textit{Laplace function},$$

$$\dot{x} = \frac{m_1 - np}{\sqrt{npq}}$$

$$\ddot{x} = \frac{m_2 - np}{\sqrt{npq}}$$

Remark. Values of the Laplace function Φ (x) for non-negative argument x (0≤x≤5) is given in the table (see appendix); for values of x > 5 Laplace function Φ(x) = 0,5; for negative values of x we use the same table, considering that Φ (–x) = – Φ (x).

Examples.

1. The coin is tossed five times. Find the probability that: a) three times will fall coat of arms; b) coat of arms will fall at least three times.

Solution. According to the Bernoulli's formula we have:

$$a)\ P_5(3) = C_5^3 \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^2 = \frac{C_5^3}{2^5} = \frac{10}{32}$$

$$b)\ P_5(3 \le m \le 5) = P_5(3) + P_5(4) + P_5(5) = \frac{1}{2}$$

Because

$$P_5(4) = C_5^4 \cdot \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right) = \frac{C_5^4}{2^5} = \frac{5}{32}; \quad P_5(5) = C_5^5 \cdot \left(\frac{1}{2}\right)^5 \cdot \left(\frac{1}{2}\right)^0 = \frac{C_5^5}{2^5} = \frac{1}{32}$$

2. We know that 1% of the total number of products produced by the enterprise are defective. What is the probability that out of 500 randomly taken products: a) there is not a single defective; b) exactly three defective products?

Solution. By the condition of the task p = 0,01, n = 500, λ = np = 5. We need to use Poisson's formula

a) $P_{500}(0) \approx e^{-5} \approx 0{,}0067;$

b) $P_{500}(3) \approx \frac{5^3}{3!} e^{-5} \approx 0{,}1404$

3. Find the probability that event A will occur exactly 80 times in 400 trials if the probability of its occurrence in each trial is equal to 0,2.

Solution. By the condition of the task n = 400; m = 80; p = 0.2; q = 0.8. We use Laplace's local theorem

$$P_{400}(80) \approx \frac{1}{\sqrt{400 \cdot 0{,}2 \cdot 0{,}8}} \varphi(x) = \frac{1}{8} \varphi(x),$$

where

$$x = \frac{m - np}{\sqrt{npq}} = \frac{80 - 400 \cdot 0{,}2}{8} = 0$$

According to the table of values of the function φ (x) (see appendix) we can find that φ (0) = 0,3989. Therefore, the required probability is equal to

$$P_{400}(80) \approx \frac{1}{8} \cdot 0{,}3989 = 0{,}04986.$$

4. The probability of occurrence of an event in each of the 100 independent trials is constant and equal to 0,8. Find the probability that the event will occur: a) not less than 75 times and not more than 90 times; b) not less than 75 times; c) not more than 74 times.

Solution. We need to use Laplace's integral theorem:

$$P(m_1, m_2) = \Phi(\ddot{x}) - \Phi(\dot{x}),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz \text{ – Laplace function},$$

$$\dot{x} = \frac{m_1 - np}{\sqrt{npq}}$$

$$\ddot{x} = \frac{m_2 - np}{\sqrt{npq}}$$

a) n=100; p=0,8; q=0,2; m_1 =75; m_2 =90. Let's calculate ẋ, ẍ :

$$\dot{x} = \frac{m_1 - np}{\sqrt{npq}} = \frac{75 - 100 \cdot 0{,}8}{\sqrt{100 \cdot 0{,}8 \cdot 0{,}2}} = -1{,}25$$

$$ẍ = \frac{m_2 - np}{\sqrt{npq}} = \frac{90 - 100 \cdot 0{,}8}{\sqrt{100 \cdot 0{,}8 \cdot 0{,}2}} = 2{,}5$$

Laplace function is an odd function, that is $\Phi(-x) = -\Phi(x)$, so we obtain the following result:

$$P_{100}(75, 90) = \Phi(2{,}5) - \Phi(-1{,}25) = \Phi(2{,}5) + \Phi(1{,}25) =$$
$$= 0{,}4938 + 0{,}394 = 0{,}8882$$

b) the requirement that the event occur at least 75 times means that the number of occurrences of the event can be equal to 75, 76,…, 100.

Thus, to solve the problem we will use the Laplace integral theorem, considering that $m_1 = 75$; $m_2 = 100$:

$$\dot{x} = \frac{m_1 - np}{\sqrt{npq}} = \frac{75 - 100 \cdot 0{,}8}{\sqrt{100 \cdot 0{,}8 \cdot 0{,}2}} = -1{,}25$$

$$ẍ = \frac{m_2 - np}{\sqrt{npq}} = \frac{100 - 100 \cdot 0{,}8}{\sqrt{100 \cdot 0{,}8 \cdot 0{,}2}} = 5$$

$$P_{100}(75, 100) = \Phi(5) - \Phi(-1{,}25) = \Phi(5) + \Phi(1{,}25) = 0{,}5 + 0{,}3944 =$$
$$= \ 0{,}8944$$

c) Events "A appeared not less than 75 times" and "A appeared not more than 74 times" are opposite, so the sum of the probabilities of these events is equal to 1.

Therefore, the required probability is

$$P_{100}(0, 74) = 1 - P_{100}(75, 100) = 1 - 0{,}8944 = 0{,}1056$$

CONCLUSIONS ON THE TOPIC

1. If event A occur only if one of the incompatible events B_1, B_2,…,B_n, which form a complete group occurs then the probability of event A is equal to the sum of products of the probabilities of events B_1, B_2,…,B_n by the corresponding conditional probability of event A:

$$P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) + \cdots + P(B_n) \cdot P_{B_n}(A).$$

This equality is called the *formula of complete probability*. Events B1, B2,…, Bn are called *hypotheses.*

2. Bayes' formula has the form

$$P_A(B_i) = \frac{P(B_i) \cdot P_{B_i}(A)}{P(A)}, (i = 1, 2, \dots, n)$$

This formula is applied when event A has occurred and it is necessary to recalculate the probability of hypotheses.

3. Bernoulli's formula has the form

$$P_n(m) = C_n^m \cdot p^m \cdot q^{n-m}, \qquad q = 1 - p$$

This formula is used when we perform n independent trials and we need to find the probability that event A will occur exactly m times. It should be noted that the Bernoulli's formula is used for a small number of independent trials (n is approximately equal to 20 or n≈20). If the number of trials is large enough, then the probability that in n independent trials event A will occur exactly m times can be determined by the formula (Local theorem of Laplace)

$$P_n(m) \approx \frac{1}{\sqrt{npq}}\varphi(x_m), \qquad x_m = \frac{m - np}{\sqrt{npq}}$$

4. If the number of trials is large enough, and the probability p of event A in each trial is small enough, you should use the *Poisson's formula*

$$P_n(m) \approx \frac{\lambda^m}{m!}e^{-\lambda}, \qquad \lambda = np.$$

5. The probability that in n independent trials, in each of which the probability of occurrence of event A is equal to p event A will occur not less than m_1 times and not more than m_2 times is calculated by the formula (Integral theorem of Laplace)

$$P(m_1, m_2) = \Phi(\ddot{x}) - \Phi(\dot{x}),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{z^2}{2}}dz \text{ – } \textit{Laplace function.}$$

SELF-TEST QUESTIONS

1. The formula of *complete probability* has the form

$a.\ P(A) = P(A) \cdot P_A(B_1) + P(A) \cdot P_A(B_2) + \cdots + P(A) \cdot P_A(B_n)$

$b.\ P(A) = P(B_1) \cdot P_{B_1}(A) + P(B_2) \cdot P_{B_2}(A) + \cdots + P(B_n) \cdot P_{B_n}(A)$

$c.\ P_A(B_i) = \frac{P(B_i) \cdot P_{B_i}(A)}{P(A)}, (i = 1, 2, \dots, n)$

2. The *Bayesian formula* has the form

$a.\ P(A) = P(A) \cdot P_A(B_1) + P(A) \cdot P_A(B_2) + \cdots + P(A) \cdot P_A(B_n)$

$b.\ P_A(B_i) = \frac{P(B_i) \cdot P_{B_i}(A)}{P(A)}, (i = 1, 2, \dots, n)$

$c.\ P_{B_i}(A) = \frac{P(B_i) \cdot P_{B_i}(A)}{P(A)}, (i = 1, 2, \dots, n)$

3. The *Bernoulli's formula* has the form

a. $P_n(m) = C_m^n \cdot p^n \cdot q^{m-n}$

b. $P_n(m) = C_n^m \cdot p^m \cdot q^{n-m}$

c. $P_n(m) = C_m^n \cdot p^n \cdot q^{n-m}$

4. The *Bernoulli's formula* is used for

a. the small number of independent trials ($n \leq 20$);

b. the large number of independent trials;

c. any number of independent trials

5. The *parameter* λ in Poisson's formula is

a. the total number of occurrences of event A in a series of independent trials;

b. the average number of occurrences of event A in a series of independent trials;

c. another answer.

6. Find *the incorrect statement*:

a. if we toss a coin 10 times, we will get a series of independent trials;

b. if we take one ball at a time from a basket containing 10 balls (without returning the ball to the basket), then we get a series of independent trials;

c. if we toss 10 identical coins at the same time, we will get a series of independent trials;

7. A coin is tossed 100 times. To find the probability that tails (number) will fall out exactly 55 times, *you need to use the formula*:

a. Bernoulli's formula

b. Poisson's formula

c. Local Laplace theorem

8. A dice is tossed 15 times. To find the probability that six points will fall out exactly 7 times, *you need to use the formula*:

a. Bernoulli's formula

b. Poisson's formula

c. Local Laplace theorem

9. A dice is tossed 125 times. To find the probability that six points will fall out not less than 45 times and not more than 55 times, *you need to use the formula*:

a. Bernoulli's formula

b. Local Laplace theorem

c. Integral Laplace theorem

10. *The Laplace function* is

a. an even function

b. an odd function

c. a function that is neither even nor odd

PRACTICAL TASKS

1. Among the patients admitted to the hospital in the city of Dnipro with a confirmed diagnosis of Covid-19, 30% are people aged 30 to 50 years, 50% of those admitted are people over 50 years old and 20% are children. The probability that a patient between the ages of 30 and 50 will not get a complication is 0,6; the probability that a patient over the age of 50 will not get a complication is 0,4; the probability that the child will not get a complication is 0,9. The patient admitted to the hospital had no complications. What is the probability that it was a person over 50 years old?

Answer: $P = 1/28$

2. According to statistics, 50% of borrowers in the bank are government agencies, 30% – other banks, 20% – individuals. Also, statistics show that the probability of non-repayment of the loan is 0,01; 0,1; 0,3 respectively. The head of the credit department received a message about the non-repayment of the loan. What is the probability that the individual did not pay?

Answer: $P = 12/19$

3. The coin is tossed 200 times. Find the probability that the coat of arms appears exactly 100 times.

Answer: 1) $P \approx 0{,}06$

4. The insurance company has concluded 40,000 contracts. The probability of an insured event for each of them during the year is 2%. Find the probability that there will be no more than 870 such cases.

Answer: $P = 0{,}9938$

5. The probability of buying a winning ticket is 0.25. Find the probability that among 15 purchased tickets there will be exactly 5 winning ones.

Answer: $P \approx 0{,}165$

6. The probability that a test for Covid-19 will give a reliable result is equal to 0.9. How many tests need to be done so that with a probability of 0.98 it can be expected that at least 150 tests will give a reliable result?

Answer: $n = 177$

LITERATURE FOR SELF-STUDY

1. A.V. Tyurinand, A.Yu. Akhmerov Theory of probability and mathematical statistics: Textbook. − Dusseldorf: LAP LAMBERT Academic Publishing GmbH & Co.KG., 2020. – 148 p.

2. Prasanna Sahoo Probability and mathematical statistics: Textbook. – USA: Department of Mathematics of the University of Louisville, 2013. – 712 p.

3. J. K. Blitzstein, J. Hwang, Introduction to Probability Second Edition. – Taylor & Francis Group, LLC, 2019. – 636 p.

Chapter 3
DISCRETE AND CONTINUOUS RANDOM VARIABLES. INTEGRAL DISTRIBUTION FUNCTION. DIFFERENTIAL DISTRIBUTION FUNCTION

3.1. Discrete and continuous random variables

A random variable is a variable that, according to the results of an experiment, can take one or another value (which is unknown in advance).

Random variables are denoted by *capital letters* of the Latin alphabet X, Y, Z and their possible values are denoted by *lowercase letters*, respectively.

Each random variable is associated with a numerical set – a *set of values of a random variable.*

Random variables by the structure of the set of values can be divided into two categories: *discrete and continuous.*

A discrete random variable is a variable which possible values are separate numbers (that is there are no possible values between two neighboring values).

Examples.

1. Number of standard details among 100 manufactured details. This number is a discrete random variable that can take values from 0 to 100.

2. Mark obtained by students in the exam. This is a discrete random variable. It can take values: 2, 3, 4, 5.

A continuous random variable is a variable which possible values continuously fill an interval (finite or infinite).

For example, a continuous random variable is:

1. Time of trouble-free operation of the device.

2. Water level in the Dnieper river.

3. The diameter of the machined detail.

In order to completely determine a random variable, it is necessary to set the law of its distribution.

The law of distribution of a discrete random variable is a list of its possible values and the corresponding (appropriate) probabilities.

The law of distribution of a discrete random variable can be given in the form of *a table*:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

Numbers $x_1,x_2,\ldots,x_n$ are possible values of the random variable X, and numbers $p_1,p_2,\ldots,p_n$ are the probabilities of these values.

Keep in mind that

$$p_1 + p_2 + \cdots + p_n = 1$$

Let's look at an example.

There are 1000 tickets in the lottery. One win of $ 100, 10 winnings of $ 20, 20 winnings of $ 10, 100 winnings of $ 1 are drawn. Random variable X is a value of possible winnings of a holder of one lottery ticket. Draw up the law of distribution of the random variable X.

Solution. Random variable X can take the values: {0, 1, 10, 20, 100}. The corresponding probabilities in this case can be found by the classical definition of the probability

$$P(X = x_i) = \frac{m}{n}$$

So, we can calculate probabilities of events: X=0, X=1, X=10, X=20, X=100. We can write

$$x = 0, \quad P(A) = \begin{bmatrix} n = 1000 \\ m = 869 \end{bmatrix} = 0{,}869;$$

$$x = 1, \quad P(A) = \begin{bmatrix} n = 1000 \\ m = 100 \end{bmatrix} = 0{,}100;$$

$$x = 10, \quad P(A) = \begin{bmatrix} n = 1000 \\ m = 20 \end{bmatrix} = 0{,}020;$$

$$x = 20, \quad P(A) = \begin{bmatrix} n = 1000 \\ m = 10 \end{bmatrix} = 0{,}010;$$

$$x = 100, \quad P(A) = \begin{bmatrix} n = 1000 \\ m = 1 \end{bmatrix} = 0{,}001;$$

$$\sum p_i = 0{,}869 + 0{,}100 + 0{,}020 + 0{,}010 + 0{,}001 = 1.$$

The law of distribution of a random variable has the form:

X	0	1	10	20	100
p	0,869	0,100	0,020	0,010	0,001

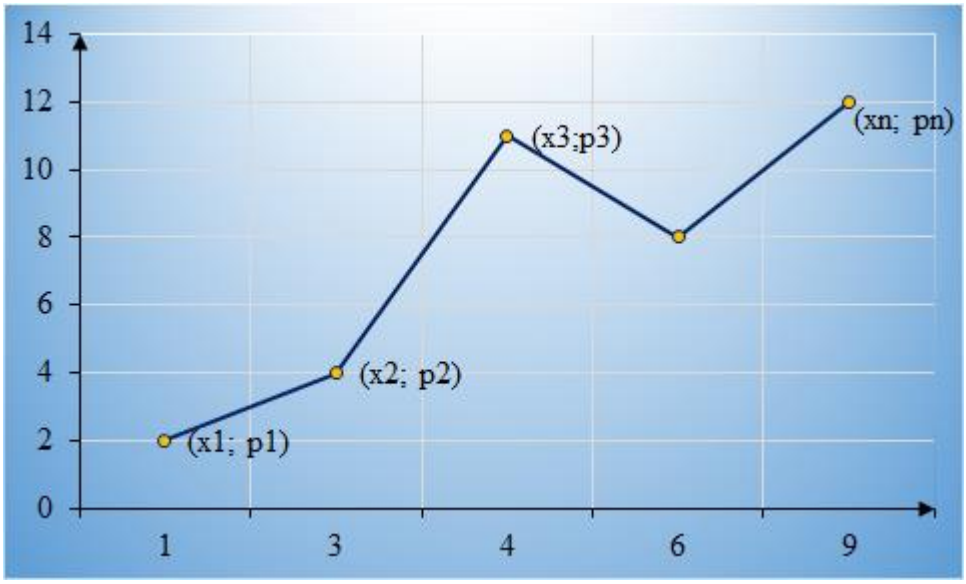
The distribution law of a discrete random variable can also be given analytically (in the form of a formula)

$$P(X = x_i) = \varphi(x_i)$$

or using the distribution function.

The law of distribution of a discrete random variable can be given graphically. If in the rectangular coordinate system on the axis OX we note the values of the discrete random variable X: $x_1, x_2,\ldots,x_n$, and on the axis OY we note the corresponding values of probabilities $p_1, p_2,\ldots,p_n$, then we get n points with coordinates: (x_1, p_1) , $(x_2, p_2),\ldots, (x_n, p_n)$. We connect these points with segments. We will get a *distribution polygon*.

Thus, the distribution polygon is a *graphical representation* of the distribution law of a discrete random variable.



Remark. These characteristics cannot be found for a continuous random variable, at least because it takes an innumerable number of values. Therefore, we consider the fundamental characteristic of any random variable – the integral function of probability distribution.

3.2. Integral distribution function

The integral distribution function of a random variable X is a function that for each value of *x* determines the probability that random variable X will take a value less than x:

$$F(x)=P(X < x)$$

The term "*integral distribution function*" is often used instead of the term "*distribution function*." Let's consider some properties of the distribution function:

Property 1. The values of the distribution function belong to the segment [0,1]:

$$0 \le F(x) \le 1$$

Property 2. The probability that a random variable will take a value from the segment [a, b] is calculated by the formula

$$P(a < X < b) = F(b) - F(a)$$

Property 3. The probability that a continuous random variable will take one value, such as x_1, is equal to zero:

$$P(X = x_1) = 0$$

Property 4. If all possible values of a random variable belong to the interval (a, b), then

$$F(x) = 0 \ \ when \ \ x \le a; \quad F(x) = 1 \ \ when \ \ x \ge b$$

Property 5. For the function F (x) the following boundary relations hold:

$$\lim_{x\to-\infty} F(x) = 0, \quad \lim_{x\to\infty} F(x) = 1$$

Examples.

1. Random variable X is given by the distribution function

$$F(x) = \begin{cases} 0, & x \le -1 \\ \dfrac{3x}{4} + \dfrac{3}{4}, & -1 < x \le \dfrac{1}{3} \\ 1, & x > \dfrac{1}{3} \end{cases}$$

Find the probability that as a result of the trial, variable X will take a value from the interval (0, 1/3).

Solution. Let's use the formula:

$$P(a < X < b) = F(b) - F(a)$$

When a = 0, b = 1/3 we get the following result:

$$P\left(0 < X < \frac{1}{3}\right) = F\left(\frac{1}{3}\right) - F(0) = \left[\frac{3x}{4} + \frac{3}{4}\right]_{x=\frac{1}{3}} - \left[\frac{3x}{4} + \frac{3}{4}\right]_{x=0} = \frac{1}{4}$$

2. The random variable is given by the distribution function

$$F(x) = \begin{cases} 0, & x \le 2 \\ 0{,}5x - 1, & 2 < x \le 4 \\ 1, & x > 4 \end{cases}$$

Find the probability that as a result of the trial it will take a value: 1) less than 0,2; 2) less than 3; 3) not less than 3; 4) not less than 5.

Solution.

1) Since for $x \le 2$ F (x) = 0, then F (0,2) = 0, that is P (x <0,2) = 0;
2) P(X<3)=F(3)=$[0{,}5x-1]_{x=3}$=1,5-1=0,5;
3) Events $X \ge 3$ and $X < 3$ are opposite, so $P(X \ge 3) + P(X < 3) = 1$. From here we get that $P(X \ge 3) = 1 - P(X < 3) = 1 - 0{,}5 = 0{,}5$
4) Events X≥5 and X <5 are opposite, so P (X≥5) + P (X <5) = 1. From here we get that $P(X \ge 5) = 1 - P(X < 5) = 1 - F(5) = 1 - 1 = 0$

3.3. Differential distribution function

The differential distribution function of a continuous random variable is called the first derivative of the distribution function

$$f(x) = F'(x)$$

The terms *"probability density*" and "*differential function*" are often used instead of the term "*distribution density*".

The *probability* that a continuous random variable will take a value belonging to the interval (a, b) can be calculated by the formula

$$P(a < x < b) = \int_a^b f(x)dx$$

Knowing the distribution density, you can find the distribution function

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

The distribution density has some properties:

Property 1. The distribution density is non-negative, that is

$$f(x) \geq 0$$

Property 2. An improper integral of the distribution density from -∞ to ∞ is equal to 1

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

In particular, if all possible values of a random variable belong to the interval (a, b), then

$$\int_{a}^{b} f(x)dx = 1$$

Examples.

1. The distribution function of a continuous random variable is given

$$F(x) = \begin{cases} 0 & x \leq 0 \\ sinx & 0 < x \leq \pi/2 \\ 1 & x > \pi/2 \end{cases}$$

Find the distribution density *f(x)* (differential function).

Solution. The distribution density is equal to the first derivative of the distribution function, that is

$$f(x) = F'(x) = \begin{cases} 0 & x \leq 0 \\ cosx & 0 < x \leq \pi/2 \\ 0 & x > \pi/2 \end{cases}$$

2. The distribution density of a continuous random variable is given

$$f(x) = \begin{cases} 0 & x \leq 0 \\ cosx & 0 < x \leq \pi/2 \\ 0 & x > \pi/2 \end{cases}$$

Find the distribution function F (x).

Solution. To find the distribution function, we need to use the formula

$$F(x) = \int_{-\infty}^{x} f(x)dx = 0$$

1. If $x \leq 0$, then $f(x) = 0$, that is

$$F(x) = \int_{-\infty}^{x} 0dx = 0$$

2. If $0 < x \leq \frac{\pi}{2}$, then

$$F(x) = \int_{-\infty}^{0} 0dx + \int_{0}^{x} cosx\, dx = sinx$$

3. If $x > \frac{\pi}{2}$, then

$$F(x) = \int_{-\infty}^{0} 0dx + \int_{0}^{\frac{\pi}{2}} cosx\, dx + \int_{\frac{\pi}{2}}^{x} 0dx + = sinx \Big|_{0}^{\pi/2} = 1$$

Thus, the sought integral function has the form

$$F(x) = \begin{cases} 0 & x \leq 0 \\ sinx & 0 < x \leq \pi/2 \\ 1 & x > \pi/2 \end{cases}$$

CONCLUSIONS ON THE TOPIC

1. *A random variable* is a variable that, according to the results of an experiment, can take one or another value (which is unknown in advance).

2. Random variables by the structure of the set of values can be divided into two categories: *discrete and continuous. A discrete random variable* is a variable which possible values are separate numbers (that is there are no possible values between two neighboring values). *A continuous random variable* is a variable which possible values continuously fill an interval (finite or infinite).

3. *The law of distribution of a discrete random variable* can be given in the form of *a table*, *analytically* (in the form of a formula). The law of distribution of a discrete random variable can also be given graphically or using the distribution function. *The law of distribution of a continuous random variable* can be given using the distribution function (integral or differential).

4. *The integral distribution function of a random variable X* is a function that for each value of *x* determines the probability that random variable X will take a value less than *x*: $F(x)=P(X < x)$.

5. *The differential distribution function* of a continuous random variable is called the first derivative of the distribution function $f(x) = F'(x)$

SELF-TEST QUESTIONS

1. Which of the following random variables is *a discrete random variable*?

a. the grade obtained by the student on the exam
b. time of trouble-free operation of the device
c. water level in the Dnieper river

2. Which of the following random variables is *a continuous random variable*?

a. the grade obtained by the student on the exam
b. the number of standard details among 500 manufactured details
c. time of trouble-free operation of the device

2. *The distribution function* takes values from the interval

a. (0, 1)
b. [0, 1]
c. [-1, 1]

3. *The distribution function* F(x) is equal to

a. P(X<x)
b. 1-P(X<x)
c. P(X>x)

4. *The distribution density function* f(x) is equal to

a. F''(x)
b. 1-F'(x)
c. F'(x)

5. *If all possible values* of a random variable X belong to the interval (a, b), then

a. $F(x) = 0 \ when \ \ x \geq a; \quad F(x) = 1 \ when \ x < b$
b. $F(x) = 0 \ when \ \ x \leq a; \quad F(x) = 1 \ when \ x \geq b$
c. $F(x) = 0 \ when \ \ x < a; \quad F(x) = 1 \ when \ x > b$

6. *The probability* that a random variable will take a value from the segment [a, b] is calculated by the formula

$a. \ P(a < X < b) = F(a) - F(b)$
$b. \ P(a < X < b) = 1 - F(a)$
$c. \ P(a < X < b) = F(b) - F(a)$

7. *All possible values* of the random variable X belong to the interval $(-1, 1)$. What is the value of the integral distribution function $F(x)$ when $x \geq 1$?

a. 0
b. 1
c. -1

8. *The integral distribution function* has the form

$$F(x)=\begin{cases}0 & x\le 0\\ kx+b & 0<x\le a\\ 1 & x>a\end{cases}$$

What is the value of the differential function on the interval

$$0<x\le a\ ?$$

a. 0
b. 1
c. k

9. A *distribution polygon* is
a. a stepped figure
b. a polyline
c. a straight line

PRACTICAL TASKS

1. A discrete random variable is given by the distribution law

X	2	5	8
p	0,4	0,25	0,35

Find the distribution function and plot it.

Answer: $F(x)=\begin{cases}0 & x\le 2\\ 0{,}4 & 2<x\le 5\\ 0{,}65 & 5<x\le 8\\ 1, & x>8\end{cases}$

2. A random variable X is given by a function

$$F(x)=\begin{cases}0 & x\le 1\\ 5x-1 & 1<x\le 3\\ 1 & x>3\end{cases}$$

Find the value of the distribution density at x=2. Plot its graph.
Answer: $f(x)=5$

3. A random variable is given by the distribution density function

$$f(x)=\begin{cases}0 & x\le 0\\ x/8 & 0<x\le 4\\ 0 & x>4\end{cases}$$

Find the integral distribution function F(x).

Answer: $F(x) = \begin{cases} 0 & x \le 0 \\ x^2/16 & 0 < x \le 4 \\ 1 & x > 4 \end{cases}$

4. A random variable X is given by the distribution function

$F(x) = \begin{cases} 0 & x \le 0 \\ 2x^2/5 & 0 < x \le 7 \\ 1 & x > 7 \end{cases}$

Find the probability that as a result of the trial, variable X will take a value from the interval (1; 5).

Answer: $P(1 < X < 5) = 9{,}6$

5. The law of distribution of a discrete random variable has the form:

X	1	2	3	4
p	4	7	2	10

Plot the distribution polygon.

LITERATURE FOR SELF-STUDY

1. James Nicholson. Complete Probability & Statistics 1 for Cambridge International AS & A Level. – Oxford University Press – Children, 2019. – 226 p.

2. James Nicholson. Complete Probability & Statistics 2 for Cambridge International AS & A Level. – Oxford University Press – Children, 2019. – 210 p.

3. A.V. Tyurinand, A.Yu. Akhmerov Theory of probability and mathematical statistics: Textbook. − Dusseldorf: LAP LAMBERT Academic Publishing GmbH & Co.KG., 2020. – 148 p.

Chapter 4
NUMERICAL CHARACTERISTICS OF RANDOM VARIABLES. DISTRIBUTION LAWS OF RANDOM VARIABLES

4.1. Numerical characteristics of discrete random variables

The law of distribution for both discrete and continuous random variables gives complete information about them. However, certain numerical parameters of random variables are very important in Probability Theory: *mathematical expectation*, which is a mean value around which possible values of a random variable are based, *dispersion and standard deviation*, which characterize the degree of scattering of random variables around mathematical expectation.

The mathematical expectation of a discrete random variable is the sum of the products of the values of a random variable by the probabilities corresponding to these values.

$$M(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n$$

The mathematical expectation of a discrete random variable has the following properties:

Property 1. The mathematical expectation of a constant value is equal to the constant value

$$\mathrm{M(C)} = \mathrm{C}, \qquad \mathrm{C} - const$$

Property 2. A constant multiplier can be taken out of a sign of the mathematical expectation:

$$\mathrm{M(CX)} = \mathrm{CM(X)}, \qquad \mathrm{C} - const$$

Property 3. The mathematical expectation of the product of pairwise independent random variables is equal to the product of the mathematical expectations of these random variables:

$$\mathrm{M}(\mathrm{X}_1 \cdot \mathrm{X}_2 \cdot \ldots \cdot \mathrm{X}_n) = \mathrm{M}(\mathrm{X}_1) \cdot \mathrm{M}(\mathrm{X}_2) \cdot \ldots \cdot \mathrm{M}(\mathrm{X}_n)$$

Property 4. The mathematical expectation of the sum of random variables is equal to the sum of the mathematical expectations of the terms:

$$\mathrm{M}(\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_n) = \mathrm{M}(\mathrm{X}_1) + \mathrm{M}(\mathrm{X}_2) + \cdots + \mathrm{M}(\mathrm{X}_n)$$

Examples.

1. Find the mathematical expectation of a discrete random variable, which is given by the distribution law:

X	-4	6	10
p	0,2	0,3	0,5

Solution. Let's use the formula:

$$M(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n$$

We obtain the following value of the mathematical expectation for a given distribution law:

$$M(X) = -4 \cdot 0{,}2 + 6 \cdot 0{,}3 + 10 \cdot 0{,}5 = 6$$

2. Find the mathematical expectation of a discrete random variable Z = X + 2Y, if the mathematical expectations of X and Y are known: M (X) = 5, M (Y) = 3.

Solution. Let's use the formula:

$$\mathrm{M}(\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_n) = \mathrm{M}(\mathrm{X}_1) + \mathrm{M}(\mathrm{X}_2) + \cdots + \mathrm{M}(\mathrm{X}_n)$$

and the formula:

$$\mathrm{M}(\mathrm{CX}) = \mathrm{CM}(\mathrm{X}), \qquad \mathrm{C} - const$$

We get the mathematical expectation of the random variable Z:

$$M(Z) = M(X + 2Y) = M(X) + M(2Y) = M(X) + 2M(Y) = 5 + 2 \cdot 3 = 11$$

3. Find the mathematical expectation of the number of points that fall out when throwing a dice.

Solution. List all possible values of the discrete random variable X − the number of points that fall out when throwing a dice.

$$\mathrm{X}: \{1, 2, 3, 4, 5, 6\}$$

We compose the distribution law of probabilities of a discrete random variable X. The probabilities of falling out of one, two, three, four, five, six points are the same and are equal to:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

Thus, the law of distribution of a random variable has the form

X	1	2	3	4	5	6
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Find the mathematical expectation of a discrete random variable X:

$$M(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} =$$
$$= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3{,}5$$

The dispersion of a discrete random variable X is the mathematical expectation of the square of the deviation of a random variable from its mathematical expectation:

$$D(X) = M[X - M(X)]^2$$

The *dispersion of a discrete random variable X* can be calculated by the formula:

$$D(X) = M(X^2) - [M(X)]^2$$

The *dispersion* has the following properties:

Property 1. The dispersion of the constant value is equal to zero:

$$D(C) = 0, \qquad \mathrm{C} - const$$

Property 2. A constant multiplier can be taken out of the sign of dispersion:

$$D(\mathrm{CX}) = C^2 D(\mathrm{X}), \qquad \mathrm{C} - const$$

Property 3. The dispersion of the sum of independent random variables is equal to the sum of the dispersions of the terms:

$$D(\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_n) = D(\mathrm{X}_1) + D(\mathrm{X}_2) + \cdots + D(\mathrm{X}_n)$$

The standard deviation of a discrete random variable X is equal to the square root of the dispersion:

$$\sigma(X) = \sqrt{D(X)}$$

Examples.

1. Find the standard deviation of a discrete random variable X is given by the distribution law:

X	-5	2	3	4
p	0,4	0,3	0,1	0,2

Solution. To find the standard deviation we need to use the formula:

$$\sigma(X) = \sqrt{D(X)}$$

Let's find the dispersion of a random variable X. In this case we need to use the formula:

$$D(X) = M(X^2) - [M(X)]^2$$

Find the mathematical expectation of a random variable X:

$$M(X) = -5 \cdot 0{,}4 + 2 \cdot 0{,}3 + 3 \cdot 0{,}1 + 4 \cdot 0{,}2 = -0{,}3$$

Write the distribution law of a random variable X^2:

X^2	25	4	9	16
p	0,4	0,3	0,1	0,2

Find the mathematical expectation of the random variable X^2:

$$M(X^2) = 25 \cdot 0{,}4 + 4 \cdot 0{,}3 + 9 \cdot 0{,}1 + 16 \cdot 0{,}2 = 15{,}3$$

Find the dispersion:

$$D(X) = M(X^2) - [M(X)]^2 = 15{,}3 - (-0{,}3)^2 = 15{,}21$$

Find the sought standard deviation:

$$\sigma(X) = \sqrt{D(X)} = \sqrt{15{,}21} = 3{,}9$$

4.2. Numerical characteristics of continuous random variables

The mathematical expectation of a continuous random variable X, the possible values of which belong to the entire numerical axis OX, is determined by the equality

$$M(X) = \int_{-\infty}^{\infty} xf(x)\,dx,$$

where f (x) −is the distribution density of a random variable X.

In particular, if all possible values of the random variable X belong to the interval (a, b), then

$$M(X) = \int_{a}^{b} xf(x)\,dx$$

The dispersion of a continuous random variable X, the possible values of which belong to the entire numerical axis OX, is determined by the equality

$$D(X) = \int_{-\infty}^{\infty} [x - M(X)]^2 f(x)\,dx,$$

or equivalent equality

$$D(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - [M(X)]^2$$

In particular, if all possible values of the random variable X belong to the interval (a, b), then

$$D(X) = \int_{a}^{b} [x - M(X)]^2 f(x)\,dx,$$

or

$$D(X) = \int_{a}^{b} x^2 f(x)\,dx - [M(X)]^2$$

The standard deviation of a continuous random variable X is equal to the square root of the dispersion:

$$\sigma(X) = \sqrt{D(X)}$$

Let's look at examples.

1. A random variable X is given by the distribution density: f (x) = 2x in the interval (0, 1); outside this interval f (x) = 0. Find the mathematical expectation of the variable X.

Solution. Let's use the formula

$$M(X) = \int_{a}^{b} xf(x)\,dx$$

We substitute in this formula a = 0, b = 1, f (x) = 2x. We will obtain

$$M(X) = 2\int_{0}^{1} x \cdot x\,dx = 2\int_{0}^{1} x^2\,dx = \frac{2x^3}{3}\Big|_0^1 = \frac{2}{3}$$

2. Find the mathematical expectation of a random variable X given by the distribution function

$$F(x) = \begin{cases} 0, & x \le 0 \\ \dfrac{x}{4}, & 0 < x \le 4 \\ 1, & x > 4 \end{cases}$$

Solution. Find the distribution density of a random variable X:

$$f(x) = F'(x) = \begin{cases} 0, & x \le 0 \\ \frac{1}{4}, & 0 < x \le 4 \\ 0, & x > 4 \end{cases}$$

Find the sought mathematical expectation:

$$M(X) = \int_0^4 xf(x)\,dx = \int_0^4 x \cdot \frac{1}{4}dx = \frac{1}{4}\int_0^4 x\,dx = \frac{x^2}{8}\Big|_0^4 = 2$$

4.3. Distribution laws of discrete random variables

Consider a discrete random variable X − the number of the occurrence of event A in *n* independent trials. The probability of the occurrence of event A is constant and equal to *p*. The random variable X can take m values, where *m* = 0, 1, 2, ..., *n*. The probability that the random variable X will take the value of *m* is calculated by the Bernoulli's formula

$$P_n(m) = P(X = m) = C_n^m \cdot p^m \cdot q^{n-m}, \qquad q = 1 - p$$

The law of distribution of a random variable X is called *binomial.*

If the number of trials is large, and the probability of the occurrence of an event in each trial is very small, then to calculate the probability $P(X = m)$ we use the approximate *Poisson's formula*

$$P_n(m) \approx \frac{\lambda^m}{m!}e^{-\lambda}, \qquad \lambda = np.$$

In this case, the law of distribution of a random variable X is called *Poisson's law.*

Examples.

1. The coin was tossed twice. Draw up the law of distribution of a discrete random variable X − the number of the appearance of the "coat of arms".

Solution. The probability of the appearance of the "coat of arms" with each toss of the coin is the same and is equal to p = 1/2, respectively, the probability of dropping the "number" is equal to q = 1− p = 1/2.

Let's write down all possible values of a discrete random variable $X = \{0, 1, 2\}$. The corresponding probabilities are found by Bernoulli's formula:

$$P_2(0) = C_2^0 \cdot \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^2 = \frac{1}{4} = 0{,}25$$

$$P_2(1) = C_2^1 \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{1}{2}\right) = 2 \cdot \frac{1}{4} = 0{,}5$$

$$P_2(2) = C_2^2 \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^0 = \frac{1}{4} = 0{,}25$$

The distribution law of a discrete random variable X has the form

X	0	1	2
P	0,25	0,5	0,25

2. The device consists of three independently operating elements. The probability of failure of each element in one trial is equal to 0,1. Draw up the distribution law of the number of elements that failed in one trial.

Solution.

A discrete random variable X (the number of elements that failed in one trial) can take the following values: x1 = 0 (no element failed); x2 = 1 (one element failed); x3 = 2 (two elements failed); x4 = 3 (three elements failed);

Element failures are independent of each other, the probabilities of failure of each element are equal, so we can use the Bernoulli's formula. Considering that

$$n = 3;\ p = 0{,}1;\ \ q = 1 - p = 0{,}9$$

we get that

$P_3(0) = q^3 = (0{,}9)^3 = 0{,}729;\ P_3(1) = C_3^1 pq^2 = 3 \cdot 0{,}1 \cdot (0{,}9)^2 = 0{,}243;$

$P_3(2) = C_3^2 p^2 q = 3 \cdot (0{,}1)^2 \cdot (0{,}9) = 0{,}027;\ P_3(3) = p^3 = (0{,}1)^3 = 0{,}001.$

So, the binomial distribution law of a random variable X has the form:

X	0	1	2	3
p	0,729	0,243	0,027	0,001

The mathematical expectation of a random variable X distributed by the binomial law is equal to

$$M(X) = np$$

The dispersion of a random variable X distributed by the binomial law is equal to

$$D(X) = npq$$

The standard deviation of a random variable X distributed by the binomial law is equal to

$$\sigma(X) = \sqrt{D(X)}$$

Numerical characteristics of the Poisson's distribution:

$$M(X) = np, \quad D(X) = np, \ \ \sigma(X) = \sqrt{D(X)}\,.$$

In view of that the Poisson parameter is equal to λ = np, we can write

$$M(X) = \ \lambda, \ D(X) = \lambda, \ \ \sigma(X) = \sqrt{D(X)} = \sqrt{\lambda}$$

Examples.

1. The probability of hitting the target when firing a gun is equal to 0.8. Find the mathematical expectation of the total number of hits if 5 shots are fired.

Solution. Let a random event (event A) is a hit in the target when firing a gun. There are 5 independent trials (shots). The probability of occurrence of event A in each trial is the same and equal to 0.8. Therefore, the distribution of a discrete random variable X – the number of hits from a gun is binomial.

Mathematical expectation of the total number of hits is equal to

$$M(X) = n \cdot p = 5 \cdot 0{,}8 = 4$$

2. Made 10 independent trials. The probability of the occurrence of the random event A in each trial is the same and equal to 0.8. Find the standard deviation of the random variable X – the number of the occurrence of the event A in these trials.

Solution. The random variable X has a binomial distribution law. To answer the question of the task, you need to use the following formulas:

$$D(X) = npq = np(1-p) = 10 \cdot 0{,}8 \cdot 0{,}2 = 1{,}6;$$

$$\sigma(X) = \sqrt{D(X)} = \sqrt{1{,}6} \approx 1{,}26$$

4.4. Laws of distribution of continuous random variables

A continuous random variable X has a *uniform law of the distribution on the interval* (a, b) if for all possible values of X belonging to this interval the density of the distribution remains constant, that is

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & if \ \ x \epsilon (a,b) \\ 0, & if \ \ x \notin (a,b) \end{cases}$$

Numerical characteristics of the uniform distribution:

$$M(X) = \frac{a+b}{2}, \qquad D(X) = \frac{(b-a)^2}{12}, \qquad \sigma(X) = \sqrt{D(X)} = \frac{b-a}{2\sqrt{3}}.$$

Examples.

1. The law of uniform distribution is given by the probability density

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & if\ \ x\epsilon(a,b) \\ 0, & if\ \ x \notin (a,b) \end{cases}$$

Find the distribution function F(x).

Solution. To find the distribution function, we use the connection between the differential and integral functions in the form

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

If $x \leq a$, then $f(x) = 0$

$$F(x) = \int_{-\infty}^{x} 0dx = 0$$

If $a < x \leq b$, then

$$F(x) = \int_{-\infty}^{a} 0dx + \int_{a}^{x} \frac{1}{b-a}dx = \frac{1}{b-a}\int_{a}^{x} dx = \frac{x-a}{b-a}$$

If $x > b$, then

$$F(x) = \int_{-\infty}^{a} 0dx + \int_{a}^{b} \frac{1}{b-a}\ dx + \int_{b}^{x} 0dx\ = \frac{1}{b-a}\ x\Big|_{a}^{b} = 1$$

Thus, the sought integral function has the form

$$F(x) = \begin{cases} 0, & x \leq a \\ \dfrac{x-a}{b-a}, & a < x \leq b \\ 1, & x > b \end{cases}$$

2. Find the numerical characteristics of a random variable X, which has a uniform distribution law on the interval (2, 8).

Solution. Let's use the formulas:

$M(X) = \frac{a+b}{2}$, $D(X) = \frac{(b-a)^2}{12}$, $\sigma(X) = \sqrt{D(X)} = \frac{b-a}{2\sqrt{3}}$. So, we get

$M(X) = \frac{2+8}{2} = 5$, $D(X) = \frac{(8-2)^2}{12} = 3$, $\sigma(X) = \sqrt{D(X)} = \frac{8-2}{2\sqrt{3}} = \sqrt{3} \approx 1{,}7$.

3. A continuous random variable X is distributed according to a uniform law. Find the probability density of a random variable X and plot it, if $M(X) = 4, D(X) = 3$.

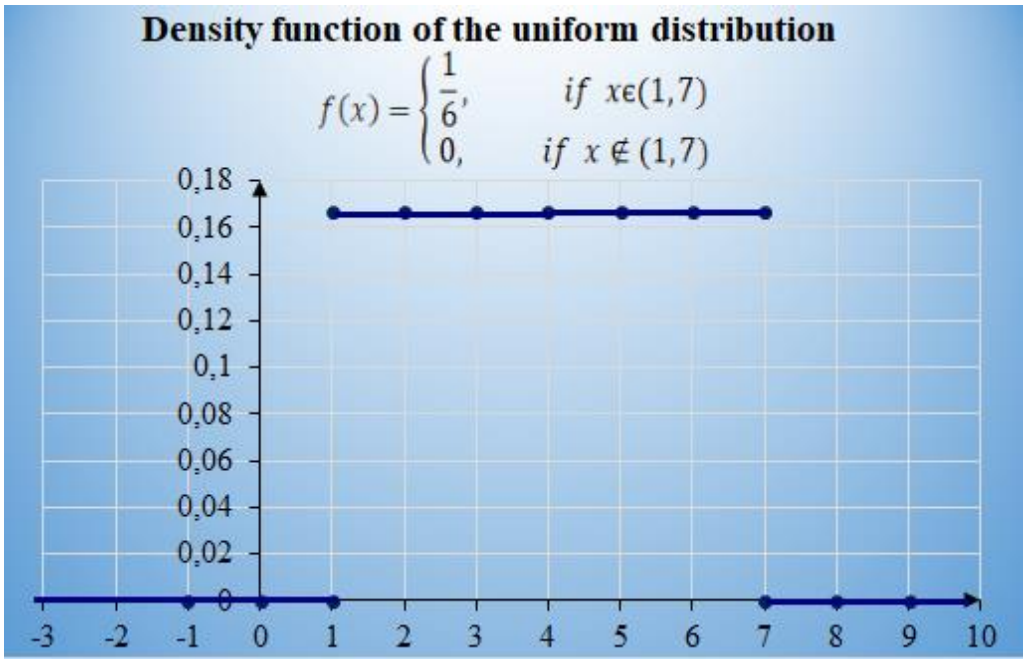
Solution. Considering that the mathematical expectation of a uniformly

distributed random variable X is equal to $M(X) = \frac{a+b}{2}$, and its dispersion is calculated by the formula $D(X) = \frac{(b-a)^2}{12}$, we have a system of equations:

$$\begin{cases} \frac{a+b}{2} = 4 \\ \frac{(b-a)^2}{12} = 3 \end{cases} \Rightarrow \begin{cases} a+b = 8 \\ (b-a)^2 = 36 \end{cases} \Rightarrow \begin{cases} a+b=8 \\ b-a=6 \end{cases} \Rightarrow \begin{cases} a=1 \\ b=7 \end{cases}$$

The probability density of the random variable X has the form

$$f(x) = \begin{cases} \frac{1}{6}, & if \ \ x \epsilon (1,7) \\ 0, & if \ \ x \notin (1,7) \end{cases}$$



The distribution of probabilities of a continuous random variable X, the density of which has the form

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

is called the *normal distribution of probabilities.* In this formula a is the mathematical expectation, and σ is the standard deviation of X.

Thus, *the numerical characteristics* of the normal distribution are completely determined by the distribution density.

The *probability* that X will take the value belonging to the interval (α, β) is determined by the formula

$$P(\alpha < x < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right),$$

where F (x) is the Laplace function.

It is often required to calculate the probability that the absolute value of the deviation of a normally distributed random variable X is less than a given positive number δ, that is it is required to find the probability that the inequality $|X - a| < \delta$ is true.

Let us replace this inequality with the equivalent double inequality

$$-\delta < X - a < \delta, \quad \text{or} \quad a - \delta < X < a + \delta.$$

Then we can write the following equality

$$P(|X - a| < \delta) = P(a - \delta < X < a + \delta)$$
$$= \Phi\left[\frac{(a + \delta) - a}{\sigma}\right] - \Phi\left[\frac{(a - \delta) - a}{\sigma}\right] = \Phi\left(\frac{\delta}{\sigma}\right) - \Phi\left(-\frac{\delta}{\sigma}\right).$$

Taking into account the equality

$$\Phi\left(-\frac{\delta}{\sigma}\right) = -\Phi\left(\frac{\delta}{\sigma}\right),$$

(the Laplace function is odd), we get

$$P(|X - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right).$$

In particular, if $a = 0$

$$P(|X - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right).$$

Thus, we can conclude that if two random variables are normally distributed and $a = 0$, then the probability of taking a value belonging to the interval $(-\delta, \delta)$ is greater for the value that has a smaller value of σ. This fact fully corresponds to the probabilistic meaning of the parameter σ (σ is the standard deviation; it characterizes the dispersion of a random variable around its mathematical expectation).

Let's transform the formula

$$P(|X - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right).$$

We set $\delta = \sigma t$. As a result, we get

$$P(|X - a| < \sigma t) = 2\Phi(t).$$

If $t = 3,$ therefore, $\sigma t = 3\sigma$ and we can write the following equality

$$P(|X - a| < 3\sigma) = 2\Phi(3) = 2 \cdot 0{,}49865 = 0{,}9973.$$

Thus, the probability that the deviation in absolute value will be less than three times the standard deviation is 0.9973.

In other words, the probability that the absolute value of the deviation will exceed three times the standard deviation is very small (it is equal to 0.0027). This means that in 0.27% of cases this can happen.

Thus, we can formulate the *three-sigma rule*: if a random variable is normally distributed, then the absolute value of its deviation from the mathematical expectation does not exceed three times the standard deviation.

In practice, the three-sigma rule is applied as follows: if the distribution of the random variable under study is unknown, but the above condition is met, then there is reason to assume that the studied variable is distributed normally; otherwise, it has a non-normal distribution.

Examples

1. Normally distributed random variable X is given by density

$$f(x) = \frac{1}{5\sqrt{2\pi}} e^{-\frac{(x-1)^2}{50}}$$

Find the mathematical expectation and dispersion of X.

Solution. By the form of the function f (x) we determine that

$$a = M(X) = 1, \qquad \sigma = 5.$$

Dispersion of X is found by the formula:

$$D(X) = \sigma^2 = 25.$$

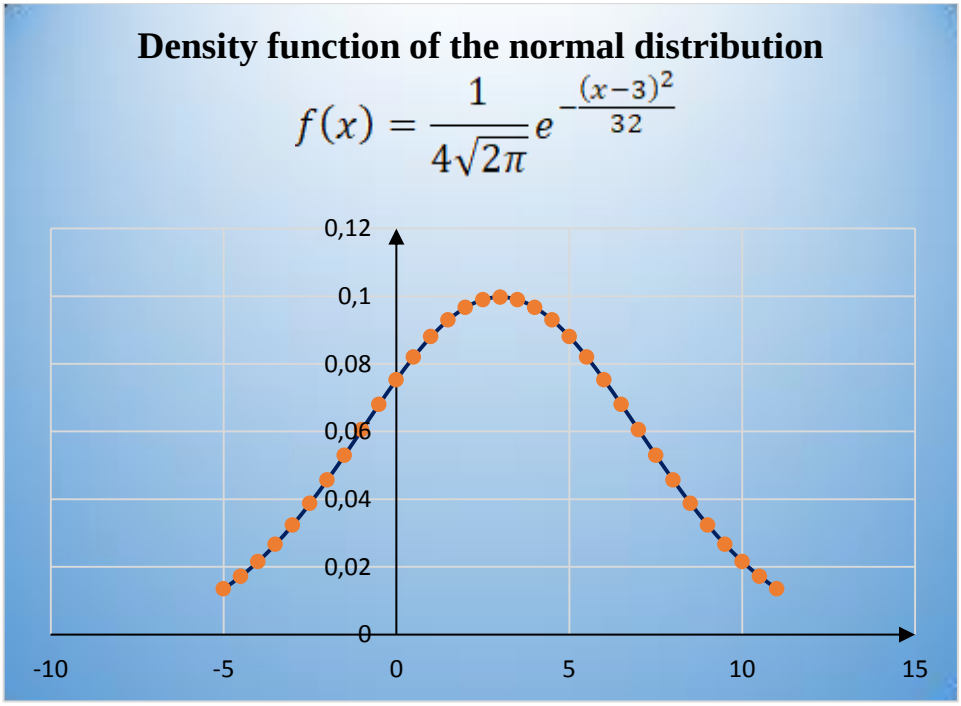
2. The mathematical expectation of a normally distributed random variable X is equal to 3, and the dispersion of it is equal to 16. Write down the distribution density of X and plot it.

Solution.

Normal distribution parameters: $a = M(X) = 3,\ \sigma = \sqrt{D(X)} = 4$. The density of the distribution has the form

$$f(x) = \frac{1}{4\sqrt{2\pi}} e^{-\frac{(x-3)^2}{32}}$$

Let's plot this function. The function graph is as follows



3. The mathematical expectation of the normally distributed random variable X is equal to 10 and the standard deviation is equal to 2. Find the probability that X will take a value from the interval (12, 14).

Solution. Let's use the formula

$$P(\alpha < x < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right)$$

Let's substitute $\alpha = 12,\ \beta = 14, a = 10, \sigma = 2$. Then

$$P(12 < x < 14) = \Phi(2) - \Phi(1)$$

According to the table of values of the Laplace function we find

$$\Phi(2) = 0{,}4772,\ \Phi(1) = 0{,}3413.$$

Thus, the sought probability is equal to

$$P(12 < x < 14) = 0{,}1359$$

A continuous random variable X has an *exponential probability distribution* if the distribution density is determined by the equality

$$f(x) = \begin{cases} 0, & if\ x < 0 \\ \lambda e^{-\lambda x}, & if\ x \geq 0 \end{cases}$$

where λ – is a constant value (λ> 0).

The distribution function of the exponential law has the following form

$$F(x) = \begin{cases} 0, & if \quad x < 0 \\ 1 - e^{-\lambda x}, & if \quad x \geq 0 \end{cases}$$

Let a continuous random variable X be distributed exponentially. *The probability* that X belongs to the interval (a, b) distributed by the exponential law is equal to

$$P(a < x < b) = e^{-\lambda a} - e^{-\lambda b}$$

Numerical characteristics of the exponential distribution have the following form:

$$M(X) = \frac{1}{\lambda}, \qquad D(X) = \frac{1}{\lambda^2}, \qquad \sigma(X) = \sqrt{D(X)} = \frac{1}{\lambda}.$$

Let's look at examples:

1. A continuous random variable X is distributed according to the exponential law given by the distribution function

$$F(x) = \begin{cases} 0, & if \; x < 0 \\ 1 - e^{-0,6x}, & if \quad x \geq 0 \end{cases}$$

Find the probability that X belongs to the interval (2, 5).

Solution. Let's use the formula

$$P(a < x < b) = e^{-\lambda a} - e^{-\lambda b}$$

Let's substitute $\lambda = 0,6; \quad a = 2; \; b = 5$. We get

$$P(2 < x < 5) = e^{-1,2} - e^{-3} = 0,251$$

2. Find the dispersion and standard deviation of the exponential distribution given by the probability density

$$f(x) = \begin{cases} 0, & if \quad x < 0 \\ 10e^{-10x}, & if \quad x \geq 0 \end{cases}$$

Solution. The parameter of the exponential distribution λ = 10. The dispersion is determined by the formula

$$D(X) = \frac{1}{\lambda^2} = \frac{1}{10^2} = 0,01$$

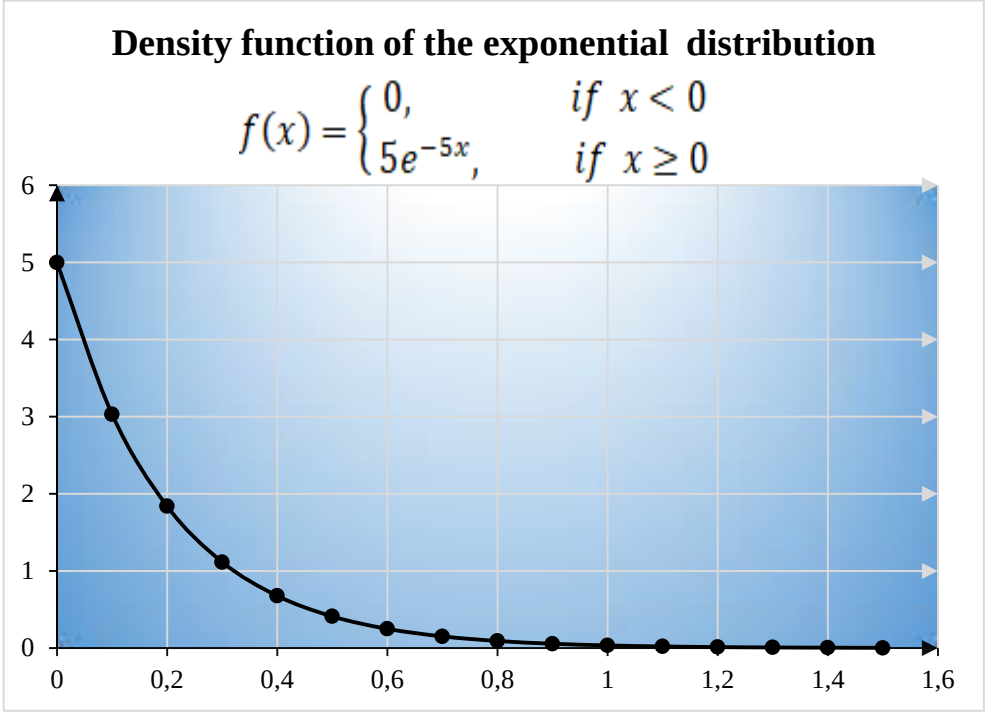
$$\sigma(X) = \sqrt{D(X)} = \frac{1}{\lambda} = 0,1.$$

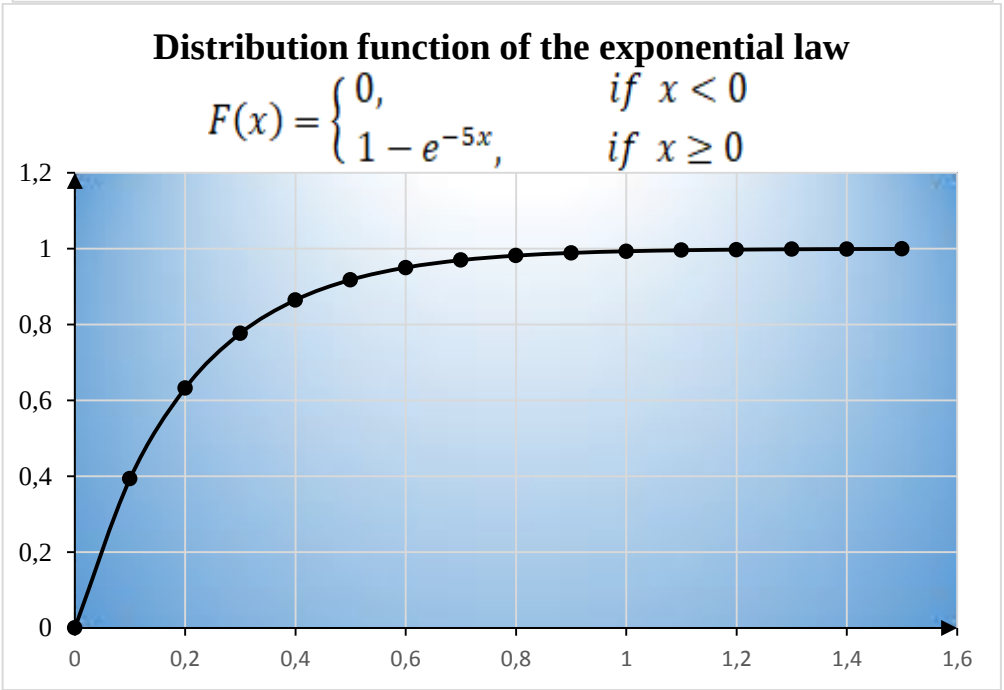
3. Write down the density and distribution function of the exponential law and plot them, if the parameter λ = 5.

Solution. We use the known formulas for density and distribution function of the exponential law. We get:

$$f(x) = \begin{cases} 0, & if \ \ x < 0 \\ 5e^{-5x}, & if \ \ x \geq 0 \end{cases}$$

$$F(x) = \begin{cases} 0, & if \ \ x < 0 \\ 1 - e^{-5x}, & if \ \ x \geq 0 \end{cases}$$





Let X_i $(i = 1, 2, \dots, n)$ are normal independent random variables; the mathematical expectation of each of them is equal to zero, and the standard deviation is equal to one. Then the sum of the squares of these variables

$$\chi^2 = \sum_{i=1}^{n} X_i^2$$

distributed according to the χ^2 law with $k = n$ degrees of freedom; if these quantities are connected by a linear relation, for example $\sum_{i=1}^{n} \mathrm{X}_i = n\overline{X}$, then the number of degrees of freedom is equal to $k = n - 1$.

The density of the χ^2 distribution has the form

$$f(x) = \begin{cases} 0, & x \le 0 \\ \dfrac{1}{2^{\frac{k}{2}} G\left(\frac{k}{2}\right)} e^{-\frac{x}{2}} x^{\frac{k}{2}-1}, & x > 0, \end{cases}$$

Where G(x) is the gamma function. There is an equality

$$G(n+1) = n!$$

It can be seen from this equality that the χ^2 distribution is determined by one parameter – the number of degrees of freedom k. It should be noted that as the number of degrees of freedom increases, the χ^2 distribution slowly approaches normal.

4.5. Reliability function

Let's introduce the concept of *reliability function*. We will call some device an element, regardless of whether it is “simple” or “complex”.

Let the element start working at time t_0 and after this time with duration t a failure occurs. Denote by T a continuous random variable − the duration of the element's uptime. If the element worked flawlessly (before the failure) for a time less than t, then, consequently, a failure will occur in a time of duration t.

Thus, the distribution function $F(t) = P(T < t)$ determines the probability of failure over time t. Therefore, the probability of failure-free operation for the same time of duration t, that is, the probability of the opposite event $T > t$ is equal to

$$R(t) = P(T > t) = 1 - F(t).$$

The reliability function $R(t)$ is a function that determines the probability of failure-free operation of an element for a time of duration t.

Often, the duration of the uptime of an element has an exponential distribution, the distribution function of which has the form

$$F(t) = 1 - e^{-\lambda t}.$$

Therefore, in the case of an exponential distribution of the uptime of the element, the reliability function has the form

$$R(t) = 1 - F(t) = 1 - \left(1 - e^{-\lambda t}\right) = e^{-\lambda t}.$$

The exponential law of reliability is the function of reliability, which is defined by the equality

$$R(t) = e^{-\lambda t},$$

where λ is the failure rate.

This formula allows you to find the probability of failure-free operation of an element on a time interval of duration t, if the uptime has an exponential distribution.

CONCLUSIONS ON THE TOPIC

1. *The mathematical expectation, dispersion and standard deviation are* numerical characteristics of both discrete and continuous random variables.

2. *The mathematical expectation of a discrete random variable* is the sum of the products of the values of a random variable by the probabilities corresponding to these values: $M(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n$. The *dispersion of a discrete random variable X* can be calculated by the formula: $D(X) = M(X^2) - [M(X)]^2$. *The standard deviation of a discrete random variable X* is equal to the square root of the dispersion: $\sigma(X) = \sqrt{D(X)}$.

3. *The mathematical expectation of a continuous random variable X* is determined by the equality: $M(X) = \int_a^b xf(x)\,dx$. *The dispersion of a continuous random variable X* is determined by the formula: $D(X) = \int_a^b x^2 f(x)\,dx - [M(X)]^2$. *The standard deviation of a continuous random variable X* is equal to the square root of the dispersion: $\sigma(X) = \sqrt{D(X)}$.

4. From previous lectures, we know that *the law of distribution of a discrete random variable* is a list of its possible values and the corresponding probabilities. If each value of a random variable X is assigned a probability value, which we can calculate using the formula: $P_n(m) = P(X = m) = C_n^m \cdot p^m \cdot q^{n-m}$, $q = 1 - p$ then X has *a binomial distribution*. If each value of a random variable X is assigned a probability value, which we can calculate

using the formula: $P_n(m) \approx \frac{\lambda^m}{m!} e^{-\lambda}$, $\lambda = np$ then X has *a Poisson's distribution law*.

5. *The law of distribution of a continuous random variable* can be given using the distribution function (integral or differential). A continuous random variable X has a *uniform law of the distribution on the interval* (a, b) if for all possible values of X belonging to this interval the density of the distribution remains constant

$$f(x) = \begin{cases} \frac{1}{b-a}, & if \ \ x \epsilon (a,b) \\ 0, & if \ \ x \notin (a,b) \end{cases}$$

6. A continuous random variable X has a *normal law of the distribution* if its density has the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

7. A continuous random variable X has an *exponential probability distribution* if the distribution density is determined by the equality

$$f(x) = \begin{cases} 0, & if \ \ x < 0 \\ \lambda e^{-\lambda x}, & if \ \ x \geq 0 \end{cases}$$

8. *The reliability function* $R(t)$ is a function that determines the probability of failure-free operation of an element for a time of duration t. In the general case, this function is determined by the formula

$$R(t) = P(T > t) = 1 - F(t).$$

SELF-TEST QUESTIONS

1. *The mathematical expectation* of a discrete random variable is calculated by the formula:

a. $M(X) = {x_1}^2 \cdot p_1 + {x_2}^2 \cdot p_2 + \cdots + {x_n}^2 \cdot p_n$
b. $M(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_n \cdot p_n$
c. $M(X) = x_1 \cdot {p_1}^2 + x_2 \cdot {p_2}^2 + \cdots + x_n \cdot {p_n}^2$

2. *The dispersion* of a discrete random variable is calculated by the formula:

a. $D(X) = M(X^2) - [M(X)]^2$
b. $D(X) = M(X) - [M(X)]^2$
c. $D(X) = [M(X)]^2 - M(X^2)$

3. *The mathematical expectation* of a constant value M(C) is equal to

a. C
b. 0
c. C^2

4. If X is a random variable, C is a constant, then *the mathematical expectation* M(CX) is equal to

a. CM(CX)
b. CM(X)
c. 1

5. *The dispersion* of a constant value D(C) is equal to

a. 1
b. C
c. 0

6. Among the distribution laws below, choose the *distribution law for a discrete random variable*

a. uniform
b. normal
c. binomial

7. The mathematical expectation of a random variable X distributed by the *binomial law* is equal to

a. $M(X) = npq$
b. $M(X) = np$
c. $M(X) = \sqrt{np}$

8. *The dispersion* of a random variable X distributed by the binomial law is

a. $D(X) = npq$
b. $D(X) = np$
c. $D(X) = \sqrt{npq}$

9. A continuous random variable X has a *uniform law of the distribution* on the interval (a, b). Then the mathematical expectation of X is equal to

a. $M(X) = \frac{a+b}{2}$
b. $M(X) = \frac{a-b}{2}$
c. $M(X) = \frac{(a+b)^2}{2}$

10. *Normally distributed random variable* X is given by the density $f(x) = \frac{1}{5\sqrt{2\pi}} e^{-\frac{(x-1)^2}{50}}$. Find the mathematical expectation of X.

a. 5
b. 1
c. 50

PRACTICAL TASKS

1. A random variable X is given by the distribution law

X	1	3	4	7	8
p	0,2	0,2	0,3	0,2	0,1

Find $M(x),\ D(x),\ \sigma(x)$.
Answer: $M(x) = 4{,}2;\ D(X) = 5{,}36;\ \sigma(x) \approx 2{,}32$

2. Find $M(2x + y)$, if $M(x) = 3,\ M(y) = 2$.
Answer: $M(2x + y) = 8$

3. Find $D(3x - 2y)$, if $D(x) = 2,\ D(y) = 3$.
Answer: $D(3x - 2y) = 6$

4. A random variable X is given by the distribution density function

$$f(x) = \begin{cases} 0 & x \le 0 \\ cosx & 0 < x \le \pi/2 \\ 0 & x > \pi/2 \end{cases}$$

Find mathematical expectation $M(x)$.
Answer: $M(X) \approx 2{,}14$

5. A basketball player throws the ball into the basket with a probability of 0.8. He throws the ball 10 times. Find: 1) the average number (mathematical expectation) of hits; 2) the variance of the number of hits; 3) the standard deviation of the number of hits.
Answer: $M(X) = 8;\ D(X) = 1{,}6;\ \ \sigma(X) \approx 1{,}26$

6. A continuous random variable X is given by the distribution density function

$$f(x) = \begin{cases} 0 & x \le 5 \\ A & 5 < x \le 11 \\ 0 & x > 11 \end{cases}$$

Find: 1) constant value A; 2) mathematical expectation M(x); 3) dispersion D(x); 4) standard deviation $\sigma(x)$.
Answer: $A = 1/6;\ \ M(X) = 8;\ D(X) = 3;\ \ \sigma(X) = \sqrt{3}$

7. A continuous random variable X is given by a distribution density function

$$f(x) = \begin{cases} 0, & if \quad x < 0 \\ \frac{1}{5} e^{-\frac{1}{5}x}, & if \quad x \geq 0 \end{cases}$$

Find: 1) the distribution function; 2) mathematical expectation M(x); 3) dispersion D(x); 4) standard deviation $\sigma(x)$.

Answer:

$$F(x) = \begin{cases} 0, & if \quad x < 0 \\ 1 - e^{-\frac{1}{5}x}, & if \quad x \geq 0 \end{cases}; \quad M(X) = 5; \; D(X) = 25; \quad \sigma(X) = 5$$

LITERATURE FOR SELF-STUDY

1. James Nicholson. Complete Probability & Statistics 1 for Cambridge International AS & A Level. – Oxford University Press – Children, 2019. – 226 p.

2. James Nicholson. Complete Probability & Statistics 2 for Cambridge International AS & A Level. – Oxford University Press – Children, 2019. – 210 p.

3. A.V. Tyurinand, A.Yu. Akhmerov Theory of probability and mathematical statistics: Textbook. − Dusseldorf: LAP LAMBERT Academic Publishing GmbH & Co.KG., 2020. – 148 p.

Chapter 5
ELEMENTS OF MATHEMATICAL STATISTICS

5.1. The importance of statistics in everyday life

«Statistics knows everything» – with these words begins the second part of the novel by Ilf and Petrov «The Twelve Chairs». To emphasize the importance of statistics in everyday life, they give an example of predicting the results of the US presidential election in 1936. Roosevelt and Landon were then candidates for the election. The editors of a venerable magazine decided to conduct a survey of voters in the telephone directory. 10 million postcards were sent across the country asking for the name of the future president. Soon, the magazine informed that Landon would be elected president of the United States by a large margin in future elections.

A parallel survey was conducted by sociologists Gallup and Roper, based on a sample of only 4,000 respondents. Despite the fact that the editors of the magazine polled 10 million voters, spending heavily on postcard distribution, data collection and processing, their forecast turned out to be wrong. After all, he relied on the point of view of only those voters who had a telephone. The forecast of sociologists almost coincided with the results of the elections.

Let us analyze the situation that has developed during the collection and processing of statistical data. As it was said, the publishing house of the magazine interviewed only those voters who had a telephone. That is, they polled only "wealthy voters." At the same time, the opinion of the poor ones was not taken into account. Sociologists interviewed voters from different social groups, so their forecast turned out to be accurate. Speaking in the language of statistics, it turns out that sociologists have collected a representative sample, that is, a sample that can be used to judge the general population (the population of USA in this example). And the publisher of the journal for research selected a sample that was not representative. Therefore, their prediction about the outcome of the elections turned out to be erroneous.

The first statistical studies were carried out in England and Germany. In the middle of the 17th century, a scientific direction arose in England, called "political arithmetic". It was founded by Petty and Graunt, who, on the

basis of information about mass social processes, tried to discover the patterns of social life. Along with the school of “political arithmetic”, the school of descriptive statistics developed in England, and the school of “state studies” developed in Germany. The development of "political arithmetic" and "state science" contributed to the emergence of the science of statistics.

5.2. The main tasks of mathematical statistics

Mathematical statistics is a branch of mathematics that studies the methods of collecting, systematizing and analyzing the results of observations of mass random phenomena in order to identify existing regularities. Methods of probability theory are used to learn these regularities.

The main tasks of mathematical statistics are as follows:

− To indicate methods of collecting and grouping statistical data obtained from observations to get scientific and practical conclusions;

− To develop methods for analyzing statistical data depending on the purpose of the study.

5.3. General and sample populations

Let it be required to study a set of homogeneous objects with respect to some qualitative or quantitative feature that characterizes these objects. For example, if there is a batch of parts, then the standard part can serve as a qualitative sign, and the controlled size of the part can serve as a quantitative sign.

Sometimes a complete survey is carried out, that is, each of the objects in the population is examined with respect to the feature that is of interest. In practice, however, a continuous survey is used relatively rarely. For example, if the set of objects contains a very large number of objects, then it is physically impossible to conduct a continuous survey. If the survey of the object requires large material costs, then it makes no sense to conduct a continuous survey. In such cases, a limited number of objects are randomly selected from the entire population and subjected to study.

A general population is a set of objects of the same type that are studied on some sign. Let’s look at examples.

Examples of the general population.

1.A set of private banks in Ukraine by profit;

2. A set of products of a particular type of goods by quality;

3. A set of people by age.

Suppose that it is not possible to investigate some sign in all elements of the general population (or there are many of them, or for other reasons). In this case, we use the *selection method*, according to which from the general population on some sign randomly select k elements: $x_1, x_2, \ldots, x_k$. The set of this elements is called the *sample.* Further study of the general population is related to the analysis of the sample. The results of the sample analysis are extended to the studied general population.

5.4. Representative sample. Selection methods

When we compiling a sample, we can proceed in two ways: after an object is selected and observed over it, it can be returned or not returned to the general population. In accordance with the foregoing, the samples are classified into repeated and non-repeated.

A sample is called *repeated*, which is obtained under the following condition: the selected object (before selecting the next object) is returned to the general population.

A *non-repetitive* sample is one that is obtained under the following condition: the selected object (before the selection of the next object) is not returned to the general population. In practice, non-repetitive random selection is usually used.

In order for the data of the sample to be sufficiently confident in judging the feature of interest in the general population, it is necessary that the objects of the sample represent it correctly. In other words, the sample must correctly represent the proportions of the population. This requirement is briefly formulated as follows: the sample must be *representative*.

According to the law of large numbers, it can be argued that the sample will be representative if it is selected randomly: each object of the sample is selected randomly from the general population if all objects have the same probability of being included in the sample.

If the size of the general population is large enough, and the sample is only a small part of this population, then the difference between repeated and non-repeated samples becomes insignificant.

As for the methods of selecting elements in the sample, there are nuances that you need to know. In practice, various selection methods are used. Basically, these methods can be divided into two types:

- Selection that does not require the division of the general population into parts. These include: simple random nonrepetitive selection and simple

random reselection.

- Selection in which the population is divided into parts. These include: typical, mechanical and serial selection.

A simple random selection is a selection, in which objects are extracted one by one from the entire general population. Simple selection can be done in a variety of ways. For example, to extract n objects from the general population of volume N, they do this: they write out numbers from 1 to N on cards that are thoroughly mixed, and one card is randomly taken out. An object that has the same number as the extracted card is subjected to examination, then the card is returned to the pack and the process is repeated, that is, the cards are mixed, one is taken out at random, and so on. This is done n times, resulting in a simple random resampling of size n.

If the extracted cards are not returned to the pack, then the sample will be a simple random non-repetitive one.

With a large volume of the general population, the described process turns out to be very laborious. In this case, ready-made tables of "random numbers" are used, in which the numbers are arranged in random order. In order to select, for example, 50 objects from a numbered general population, open any page of the table of random numbers and write out 50 numbers in a row. The sample will include those objects whose numbers match the written random numbers. If the random number of the table is greater than the number N, then the random number is skipped. When performing non-repetitive sampling, the random numbers of the table that have already been encountered before should also be skipped.

A typical selection is a selection in which objects are selected not from the entire general population, but from each of its “typical” parts. For example, if parts are made on several machines, then the selection is made not from the entire set of parts produced by all machines, but from the products of each machine separately. Typical selection is used when the trait being examined fluctuates markedly in different typical parts of the general population. For example, if products are made on several machines, among which there are more and less worn out ones, then a typical selection will be appropriate here.

A mechanical selection is a selection, in which the general population is “mechanically” divided into as many groups as there are objects to be included in the sample, and one object is selected from each group. For example, if you need to select 20% of the parts made by the machine, then every fifth part is selected. If it is required to select 5% of the parts, then every twentieth part is selected, and so on. It should be borne in mind that in some cases mechanical selection may not ensure the representativeness of the

sample.

A serial selection is a selection, in which objects are selected from the general population not one at a time, but in "series", which are subjected to a continuous examination. Serial selection is used when the examined trait fluctuates slightly in different series. In practice, combined selection is often used, in which the above methods are combined.

5.5. Statistical distribution of the sample

Suppose that a sample $x_1, x_2, \ldots, x_k$ of volume n is selected from the general population to study the quantitative (discrete or continuous) sign X. Moreover, the value of x_1 was observed n_1 times, the value of x_2 was observed n_2 times,…,the value of x_k was observed n_k times.

The observed values of x_i of the sign X are called *variants*, and the list of variants written in increasing order is called the *variation series*.

Numbers $n_1, n_2, \ldots n_k$ are called *frequencies* and the numbers $w_1 = {}^{n_1}/_{n}$, $w_2 = {}^{n_2}/_{n}$, …, $w_k = {}^{n_k}/_{n}$ are called *relative frequencies.*

The statistical distribution of the sample is the set of variants x_i of the variation series and the corresponding frequencies n_i or relative frequencies w_i.

It should be understood that the *sum of all frequencies* is equal to the sample size n, that is

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

and the sum of the relative frequencies is equal to one

$$w_1 + w_2 + \cdots + w_k = 1$$

The statistical distribution of the sample can also be set as a *list of intervals and corresponding frequencies.* The frequency of the interval is equal to the sum of frequencies of the variants that fell into this interval.

Examples.

1. The sample is given as a frequency distribution

x_i	2	5	7
n_i	1	3	6

Find the relative frequency distribution.

Solution.

Let's calculate the sample size: n = 1 + 3 + 6 = 10. Find the relative frequencies:

$$w_1 = \frac{1}{10} = 0{,}1; \qquad w_2 = \frac{3}{10} = 0{,}3; \qquad w_3 = \frac{6}{10} = 0{,}6.$$

The distribution of relative frequencies has the form of the table

x_i	2	5	7
w_i	0,1	0,3	0,6

2. In the study of the random variable X got the following values: -2, -4, -5, -4, -2, -4, 1, 0, 0, 1, 1, 2, 4, -4, -5, 4, 0, -2, -5, 0, 2, -5, 0, 1, 2. Find: 1) frequency distribution; 2) distribution of relative frequencies.

Solution.

Find the frequency distribution:

x_i	-5	-4	-2	0	1	2	4
n_i	4	4	3	5	4	3	2

Let's calculate the sample size: $n = 4 + 4 + 3 + 5 + 4 + 3 + 2 = 25$.

Find the relative frequencies:

$$w_1 = \frac{4}{25};\ w_2 = \frac{4}{25}; w_3 = \frac{3}{25}; w_4 = \frac{5}{25}; w_5 = \frac{4}{25}; w_5 = \frac{3}{25}; w_6 = \frac{2}{25}.$$

The distribution of relative frequencies has the form of the table

x_i	-5	-4	-2	0	1	2	4
w_i	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$

5.6. Empirical distribution function

Let the statistical distribution of the frequencies of the quantitative sign X be known. Let's introduce the notation: n_x is the number of observations in which the value of the sign X less than x was observed; n is the total number of observations (sample size). It is clear that the relative frequency of the event $X < x$ is equal to n_x/n. If x changes, then the relative frequency also changes, that is, the relative frequency is a function of x. Since this function is found empirically, it is called *empirical*.

So, the empirical distribution function (sample distribution function) is the function $F^*(x)$, which for each value of x determines the relative frequency of the event $X < x$. We can write

$$F^*(x) = \frac{n_x}{n},$$

where n_x −is the number of variants less than x; n −is the sample size.

In contrast to the empirical distribution function of the sample, the distribution function $F(x)$ of the population is called the theoretical

distribution function. The difference between the empirical and theoretical functions is that the theoretical function $F(x)$ determines the probability of an event $X < x$, while the empirical function $F^*(x)$ determines the relative frequency of the same event.

It is proved that for large n the values of the functions $F^*(x)$ and $F(x)$ differ little. This implies the expediency of using the empirical distribution function of the sample for an approximate representation of the theoretical (integral) distribution function of the general population. This conclusion is also confirmed by the fact that the function $F^*(x)$ has the same properties as the function $F(x)$. Consider some properties of the empirical distribution function.

The empirical function has the following *properties*:

Property 1. The values of the empirical function belong to the segment [0; 1].

Property 2. $F^*(x) -$ is non-decreasing function.

Property 3. If $x_1 -$ is the smallest variant, and $x_k -$ is the largest variant, then $F^*(x) = 0$, if $x \le x_1$ and $F^*(x) = 1$, if $x > x_k$.

Examples.

1. Find the empirical function for this statistical distribution of the sample

x_i	1	4	6
n_i	10	15	25

Solution. Let's find the sample size: n = 10 + 15 + 25 = 50. The smallest variant $x_1 = 1$, so

$$F^*(x) = 0, \quad if \quad x \le 1$$

The value of $X < 4$, namely $x_1 = 1$, was observed 10 times, then

$$F^*(x) = \frac{10}{50} = 0{,}2 \quad if \quad 1 < x \le 4$$

The values of $X < 6$, namely $x_1 = 1$, $x_2 = 4$ were observed 10 + 15 = 25 times, so

$$F^*(x) = \frac{25}{50} = 0{,}5 \quad if \quad 4 < x \le 6$$

The largest variant $x_3 = 6$, therefore

$$F^*(x) = 1, \ if \quad x > 6$$

The empirical distribution function has the form:

$$F^*(x) = \begin{cases} 0 & x \le 1 \\ 0{,}2 & 1 < x \le 4 \\ 0{,}5 & 4 < x \le 6 \\ 1 & x > 6 \end{cases}$$

5.7. Graphic image of the sample. Polygon and histogram

Let the quantitative sign X has a discrete distribution. *A frequency polygon* is a polyline which segments connect points (x_1, n_1), (x_2, n_2),…, (x_k, n_k), where x_i – variants of the sample, n_i –corresponding frequencies.

A relative frequency polygon is a polyline which segments connect points (x_1, w_1), (x_2, w_2),…, (x_k, w_k), where x_i – variants of the sample, w_i –corresponding relative frequencies.

Let the quantitative sign X has a continuous distribution. Then the interval, which includes the observed values of the sign, is divided into several intervals of length h and find n_i – the sum of frequencies of variants that fall into the i-th interval.

The histogram of frequencies is called a stepped figure, which is composed of rectangles, the bases of which are partial intervals of length h, and the heights are equal to the ratio n_i/h (frequency density).

The histogram of relative frequencies is called a stepped figure, which is composed of rectangles, the bases of which are partial intervals of length h, and the heights are equal to the ratio w_i/h (relative frequency density).

Examples.

1. Build a frequency polygon and an empirical distribution function for a given discrete statistical distribution

x_i	2,5	4,5	6,5	8,5	10,5	12,5	14,5
n_i	5	10	15	20	25	15	10

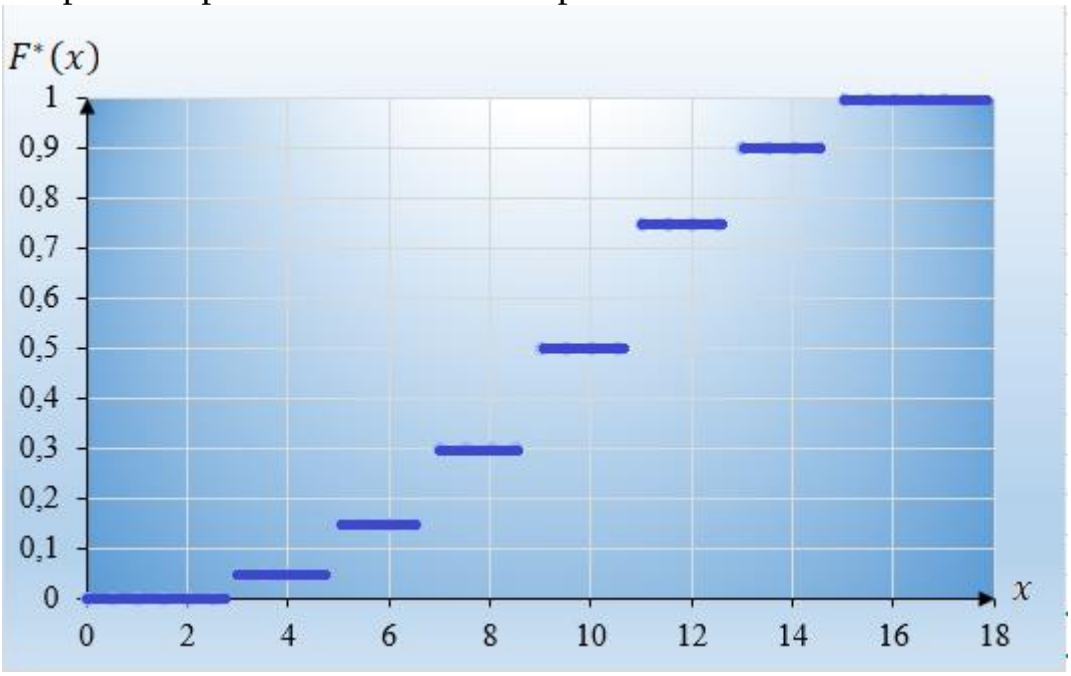
Solution. Let's calculate the frequencies, relative frequencies and values of the empirical distribution function. The results of the calculations are recorded into the table:

x_i	n_i	w_i	$F^*(x_i)$
2,5	5	0,05	0,05
4,5	10	0,1	0,15
6,5	15	0,15	0,30
8,5	20	0,2	0,50
10,5	25	0,25	0,75
12,5	15	0,15	0,90
14,5	10	0,1	1
$\sum$	100	1	-

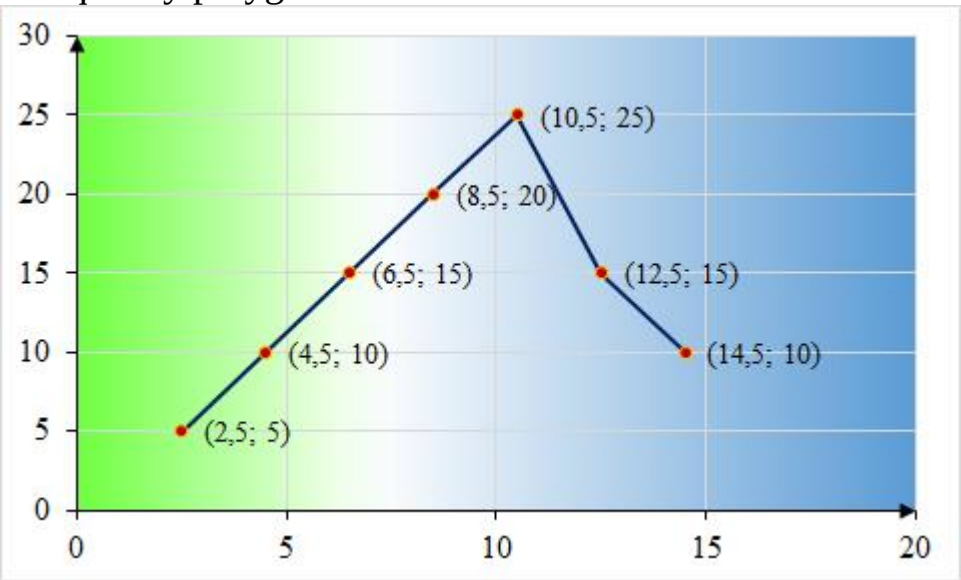
Analyzing the table data, we write down the empirical distribution function. The empirical distribution function has the form

$$F^*(x) = P(X < x) = \frac{n_x}{n} = \begin{cases} 0 & x \le 2{,}5 \\ 0{,}05 & 2{,}5 < x \le 4{,}5 \\ 0{,}15 & 4{,}5 < x \le 6{,}5 \\ 0{,}3 & 6{,}5 < x \le 8{,}5 \\ 0{,}5 & 8{,}5 < x \le 10{,}5 \\ 0{,}75 & 10{,}5 < x \le 12{,}5 \\ 0{,}90 & 12{,}5 < x \le 14{,}5 \\ 1 & x > 14{,}5 \end{cases}$$

Graphical representation of the empirical distribution function



The frequency polygon has the form



2. In the study of the random variable X got the following values: 0,6; 0,8; 0,61; 0,72; 0,69; 0,68; 0,68; 0,72; 0,74; 0,68. Compose an interval series with a step h = 0.05.

Solution.

Find the frequency distribution:

x_i	0,6	0,61	0,68	0,69	0,72	0,74	0,8
n_i	1	1	3	1	2	1	1

The smallest variant $x_{min} = 0{,}6$ and the largest variant $x_{max} = 0{,}8$. Count the number of partial intervals:

$$k = \frac{x_{max} - x_{min}}{h} = \frac{0{,}8 - 0{,}6}{0{,}05} = 4$$

For each partial interval we calculate the frequencies n_i. The results are recorded into the table:

$x_i - x_{i+1}$	0,6-0,65	0,65-0,7	0,7-0,75	0,75-0,8
n_i	2	4	3	1

Based on the results of 50 measurements of the values of some continuous random variable (see table below). 1) group the results of observations (construct an interval statistical series of frequencies and interval statistical series of relative frequencies with a step h = 12); 2) construct a histogram of frequencies.

-24	-23	-20	-28	-32	28	12	19	18	40
-17	-39	-38	-22	-24	23	31	22	30	23
-16	-14	-24	-30	-23	12	34	27	28	22
-13	-11	-24	-8	-24	26	19	7	24	27
-22	-29	-24	-18	-26	25	24	15	20	25

Solution. Find the number of partial intervals. Let's define

$$x_{min} = -39, \quad x_{max} = 40.$$

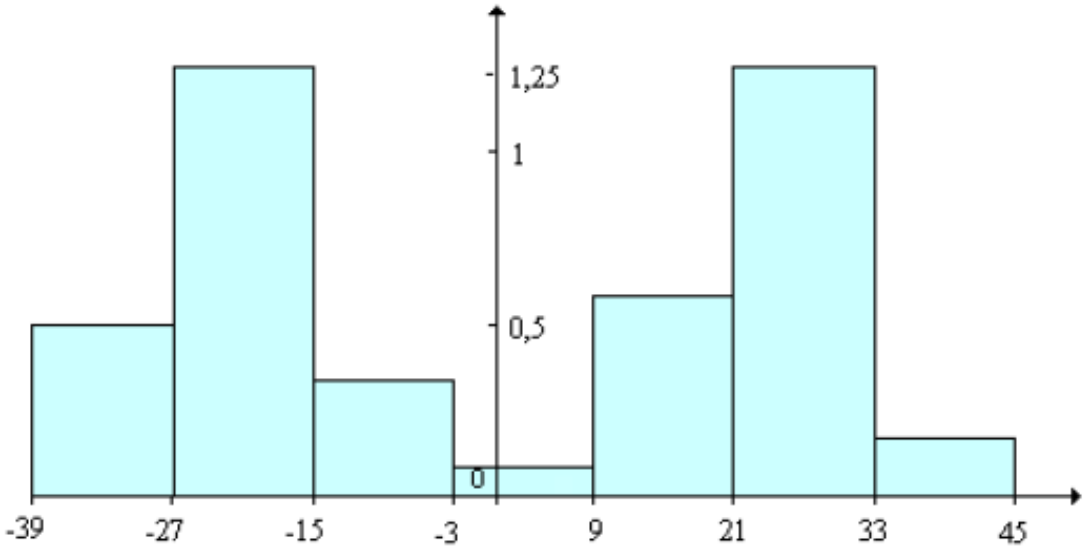
Then the number of partial intervals will be equal to

$$k = \frac{x_{max} - x_{min}}{h} = \frac{79}{12} \approx 7.$$

For each interval, calculate the frequencies n_i, the relative frequencies w_i and the density n_i/h. The data are recorded into the table:

	1	2	3	4	5	6	7	$\sum$
I_i	-39;-	-27;-	-15; -3	-3; 9	9; 21	21; 33	33; 45	
n_i	6	15	4	1	7	15	2	50
w_i	0,12	0,3	0,08	0,02	0,14	0,3	0,04	1
n_i/h	0,5	1,25	0,33	0,08	0,58	1,25	0,17	

The histogram has the form



CONCLUSIONS ON THE TOPIC

1. *The main tasks* of mathematical statistics are as follows: 1) to indicate methods of collecting and grouping statistical data obtained from observations to get scientific and practical conclusions; 2) to develop methods for analyzing statistical data depending on the purpose of the study.

2. If it is not possible to investigate some sign in all elements of the general population then we use the *selective method*, according to which from the general population on some sign randomly select k elements: $x_1, x_2, \ldots, x_k$. Elements $x_1, x_2, \ldots, x_k$ are called *variants*, and the set of this elements is called the *sample.*

3. The sample must correctly represent the proportions of the general population. This requirement is briefly formulated as follows: the sample must be *representative*.

4. *The empirical distribution function* is the function $F^*(x)$, which values can be calculated by the formula: $F^*(x) = \frac{n_x}{n}$, where n_x −is the number of variants less than x; n −is the sample size (number of variants of the sample).

5. *The difference* between the empirical and theoretical functions is that the theoretical function $F(x)$ determines the probability of an event $X < x$, while the empirical function $F^*(x)$ determines the relative frequency of the same event.

6. *For large n*, the empirical distribution function of the sample can be used to approximate the theoretical (integral) distribution function of the population.

7. *Polygon* and *histogram* of frequencies are a graphic representation of the sample. Moreover, if X is a discrete sign, then a polygon is built. If X is a continuous feature, then a histogram is constructed.

SELF-TEST QUESTIONS

1. A histogram is
a. a stepped figure
b. a polyline
c. a straight line
2. The sum of the relative frequencies is equal to
a. the sample size n
b. 1
c. 0

3. The sum of all frequencies is equal to

a. the sample size n

b. 1

c. 0

4. If $x_1 -$ is the smallest variant, and $x_k -$ is the largest variant, then

a. $F^*(x) = 0$, if $x \le x_1$ and $F^*(x) = 1$, if $x > x_k$

b. $F^*(x) = 0$, if $x > x_k$ and $F^*(x) = 1$, if $x \le x_1$

c. $F^*(x) = 0$, if $x > x_k$ and $F^*(x) = -1$, if $x \le x_1$

5. The empirical distribution function is determined by the formula

a. $F^*(x) = \frac{n_x}{n}$

b. $F^*(x) = \frac{n}{n_x}$

c. $F^*(x) = \frac{1}{n_x}$

6. A frequency polygon is a polyline which segments connect points

a. (x_1, w_1), (x_2, w_2),…, (x_k, w_k), where $x_i -$ variants of the sample, $w_i -$corresponding relative frequencies

b. (x_1, n_1), (x_2, n_2),…, (x_k, n_k), where $x_i -$ variants of the sample, $n_i -$corresponding frequencies

c. $(x_1, n_1/n)$, $(x_2, n_2/n)$,…, $(x_k, n_k/n)$, where $x_i -$ variants of the sample, $n_i -$corresponding frequencies

7. A relative frequency polygon is a polyline which segments connect points

a. $(x_1, n_1/n)$, $(x_2, n_2/n)$,…, $(x_k, n_k/n)$, where $x_i -$ variants of the sample, $n_i -$corresponding frequencies

b. (x_1, n_1), (x_2, n_2),…, (x_k, n_k), where $x_i -$ variants of the sample, $n_i -$corresponding frequencies

c. (x_1, w_1), (x_2, w_2),…, (x_k, w_k), where $x_i -$ variants of the sample, $w_i -$corresponding relative frequencies

8. The heights of rectangles of the histogram of frequencies are equal to the ratio

a. w_i/h b. h/w_i

c. n_i/h d. h/n_i

9. The heights of rectangles of the histogram of relative frequencies are equal to the ratio

a. w_i/h b. h/w_i

c. n_i/h d. h/n_i

10. Before the elections of the mayor of the city, it is planned to conduct a sociological survey. Sociologists were interested in the question: "Which candidate will win the election?". 500 people engaged in business (with an income level above the average for the country) were selected as interviewees. Will this sample be representative?

a. yes
b. no
c. hard to say

PRACTICAL TASKS

1. In the study of the random variable X got the following values: -5, -3, 0, -4, 1, 7, -5, -3, 1, -4, 5, 0, -5, 1, -3. Find the sample size. Make: a) the frequency distribution; b) the distribution of relative frequencies.

Answer:

1) the sample size $n = 15$;

2) the distribution of frequencies has the form

x_i	-5	-4	-3	0	1	5	7
n_i	3	2	3	2	3	1	1

3) the distribution of relative frequencies has the form

x_i	-5	-4	-3	0	1	5	7
w_i	3/15	2/15	3/15	2/15	3/15	1/15	1/15

2. For a given statistical distribution

x_i	3	5	6	9	10	11	19	21	25	26
n_i	4	6	7	10	8	6	4	3	1	1

Find: 1) the relative frequency of the variant x = 21; 2) the relative frequency of the interval [9, 19]; 3) the relative frequency of the segment [21, 26]; 4) the ordinate of the frequency polygon for variant x = 5; 5) the ordinate of the polygon of relative frequencies for variant x = 6; 6) the value of the empirical function on the interval (6, 9]; 7) the value of the empirical function on the interval (11, 19];

Answer: 1) 3/50*; 2)* 28/50*; 3)* 5/50*; 4)* 6*; 5)* 7/50; *6)* 17/50*; 7)* 41/50.

3. In the study of the random variable X got the following values: -0,01; -0,15; -0,15; -0,15; 0,32; 0,32; 0,45; -0,06; -0,06. Find 1) the ordinate of the frequency polygon for the variant x = -0.06; 2) the ordinate of the polygon of relative frequencies for the variable x = 0.45; 3) the height of the k-th rectangle of the frequency histogram, if k = 3, h = 0,2; 4) the height of the k-th rectangle of the histogram of relative frequencies, if k = 1, h = 0,2; Build the histogram of frequencies and the histogram of relative frequencies.

Answer: 1) 2; *2)* 0,1; *3)* 20; *4)* 3

LITERATURE FOR SELF-STUDY

1. R. J. Larsen, M. L. Marx. An introduction to mathematical statistics and its applications. – Pearson Education, Inc. 2018. – 753 p.

2. D. Rasch, D. Schott. Mathematical Statistics. – John Wiley & Sons Ltd, 2018. – 676 p.

3. Thomas A. Garrity All the Math you missed Second Edition. – Cambridge University Press, 2021. – 416 p.

Chapter 6
STATISTICAL ESTIMATES OF DISTRIBUTION PARAMETERS

6.1. Point estimates

The function $f(X_1, X_2, \dots, X_n)$ rom the observed quantities $X_1, X_2, \dots, X_n$ is called a *statistical estimate* of the unknown parameter Θ of the theoretical distribution.

A point estimate is a statistical estimate determined by a single number $\Theta^* = f(x_1, x_2, \dots, x_n)$, where $x_1, x_2, \dots, x_n$ −are the results of observations on the quantitative sign X (that is the set of $x_1, x_2, \dots, x_n$ is a sample).

An unbiased estimate is an estimate which mathematical expectation is equal to a parameter that is estimated for any sample size.

A biased estimate is an estimate which mathematical expectation is not equal to the parameter being estimated.

The unbiased estimate of the general mean (mathematical expectation) is the *sample mean*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n}$$

In this formula: $x_i -$ is the sample variant, $n_i -$ is the frequency of variant x_i, $n = \sum_{i=1}^{k} n_i -$ is the sample size.

Remark. If the variants x_i −are large numbers, then to simplify the calculations it makes sense to subtract from each variant the same number C, that is go to the *conditional variants* by the formula

$$u_i = x_i - C$$

Then *the sample mean* will be equal to

$$\bar{x}_u = C + \frac{1}{n}\sum_{i=1}^{k} n_i u_i$$

The biased estimate of the general dispersion is the sample dispersion

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i$$

The sample dispersion can be calculated by the formula

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} n_i {x_i}^2 - (\bar{x})^2$$

Remark. If the variants $x_i -$ are large numbers, then to simplify the calculations it makes sense to subtract from each variant the same number C (we choose the C the way we want), that is go to the *conditional variants* by the formula

$$u_i = x_i - C$$

Then the sample dispersion will be equal to

$$D_{\text{в}}(X) = D_{\text{в}}(u) = \frac{1}{n}\sum_{i=1}^{k} n_i {u_i}^2 - \left[\frac{1}{n}\sum_{i=1}^{k} n_i u_i\right]^2$$

Remark. If the variants $x_i -$ are decimal fractions, then to simplify the calculations it makes sense to go to the *conditional variants* by the formula

$$u_i = Cx_i$$

The constant C is chosen so that u_i will be integers. Then the sample dispersion will be equal to

$$D_{\text{в}}(X) = D_{\text{в}}(u)/C^2$$

The unbiased estimate of the general dispersion is the corrected sample dispersion, which is calculated by the formula

$$S^2 = \frac{n}{n-1} D_{\text{в}} = \frac{1}{n-1}\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i$$

or by the formula

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} n_i\, x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)^2\right)$$

Let us determine another numerical characteristic of the variation of a random variable – *the sample standard deviation*, which for *the biased dispersion estimate* is determined by the formula

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}}$$

and for *the corrected dispersion* it is given by the formula

$$S_{\text{в}} = \sqrt{S_{\text{в}}^2}$$

Examples.

1. A sample of volume $n = 50$ was obtained from the general population

x_i	2	5	7	10
n_i	16	12	8	14

Find the unbiased estimate of the general mean.

Solution. The unbiased estimate of the general mean is the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{1}{50}(16 \cdot 2 + 12 \cdot 5 + 8 \cdot 7 + 14 \cdot 10) = \frac{288}{50} = 5{,}76$$

2. A sample of volume $n = 10$ was obtained from the general population

x_i	1250	1270	1280
n_i	2	5	3

Find the sample mean for this statistical distribution of the sample.

Solution. Variants x_i are large numbers, so to simplify the calculations, we go to the conditional variants by the formula

$$u_i = x_i - 1270$$

The distribution of conditional variants has the form

u_i	-20	0	10
n_i	2	5	3

Find the sample mean

$$\bar{x} = \text{C} + \frac{1}{n}\sum_{i=1}^{k} n_i u_i = 1270 + \frac{1}{10}(2 \cdot (-20) + 5 \cdot 0 + 3 \cdot 10) =$$

$$= 1270 - 1 = 1269$$

3. A sample of volume $n = 10$ was obtained from the general population

x_i	2560	2600	2620	2650	2700
n_i	2	3	10	4	1

Find the sample mean for this statistical distribution of the sample.

Solution. Variants x_i are large numbers, so to simplify the calculations, we go to the conditional variants by the formula

$$u_i = x_i - 2620$$

The distribution of conditional variants has the form

u_i	-60	-20	0	30	80
n_i	2	3	10	4	1

Find the sample mean

$$\bar{x} = \mathrm{C} + \frac{1}{n}\sum_{i=1}^{k} n_i u_i$$

$$\bar{x} = 2620 + \frac{1}{20}(2 \cdot (-60) + 3 \cdot (-20) + 10 \cdot 0 + 4 \cdot 30 + 1 \cdot 80) =$$
$$= 2620 + 1 = 2621$$

4. Sample size $n = 41$. The biased estimate of the general dispersion $D_{\text{в}} = 3$. Find the unbiased estimate of the dispersion of the general population.

Solution. The unbiased estimate of the general dispersion is calculated by the formula

$$S^2 = \frac{n}{n-1} D_{\text{в}}$$

Taking into account that $D_{\text{в}} = 3$, $n = 41$ we obtain:

$$S^2 = \frac{41}{40} \cdot 3 = 3{,}075$$

5. A sample of volume $n = 100$ was obtained from the general population

x_i	340	360	375	380
n_i	20	50	18	12

Find the sample dispersion for this statistical distribution of the sample.

Solution. Let's go to the conditional variants by the formula

$$u_i = x_i - 360$$

The distribution of conditional variants has the form

u_i	-20	0	15	20
n_i	20	50	18	12

Let's fill in the table

n_i	u_i	${u_i}^2$	$n_i u_i$	$n_i {u_i}^2$
20	-20	400	-400	8000
50	0	0	0	0
18	15	225	270	4050
12	20	400	240	4800
$\sum$			110	16850

Find the sample dispersion. To do this, use the formula

$$D_{\text{в}}(X) = D_{\text{в}}(u) = \frac{1}{n}\sum_{i=1}^{k} n_i {u_i}^2 - \left[\frac{1}{n}\sum_{i=1}^{k} n_i u_i\right]^2$$

We get the following result

$$D_{\text{в}} = \frac{16850}{100} - \left(\frac{110}{100}\right)^2 = 168{,}5 - 1{,}21 = 167{,}29$$

6. A sample of volume $n = 50$ was obtained from the general population

x_i	0,1	0,5	0,6	0,8
n_i	5	15	20	10

Find the sample dispersion for this statistical distribution of the sample.
Solution. Let's go to the conditional variants by the formula

$$u_i = 10x_i$$

The distribution of conditional variants has the form

u_i	1	5	6	8
n_i	5	15	20	10

Let's fill in the table

n_i	u_i	${u_i}^2$	$n_i u_i$	$n_i {u_i}^2$
5	1	1	5	5
15	5	25	75	375
20	6	36	120	720
10	8	64	80	640
$\sum$			280	1740

Find the sample dispersion. To do this, use the formula

$$D_{\text{в}}(X) = \frac{D_{\text{в}}(u)}{C^2},$$

where $D_{\text{в}}(u)$ can be calculated by the formula

$$D_{\text{в}}(u) = \frac{1}{n}\sum_{i=1}^{k} n_i {u_i}^2 - \left[\frac{1}{n}\sum_{i=1}^{k} n_i u_i\right]^2$$

We get the following result

$$D_{\text{в}} = \frac{1}{100}\left(\frac{1740}{50} - \left(\frac{280}{50}\right)^2\right) = \frac{1}{100}(34{,}8 - 31{,}36) = \frac{3{,}44}{100} = 0{,}0344$$

7. A sample of volume $n = 10$ was obtained from the general population

x_i	102	104	108
n_i	2	3	5

Find the corrected sample dispersion for this statistical distribution of the sample.

Solution. Let's go to the conditional variants by the formula

$$u_i = x_i - 104$$

The distribution of conditional variants has the form

u_i	-2	0	4
n_i	2	3	5

Let's fill in the table

n_i	u_i	${u_i}^2$	$n_i u_i$	$n_i {u_i}^2$
2	-2	4	-4	8
3	0	0	0	0
5	4	16	20	80
$\sum$			16	88

Find the sample dispersion. To do this, use the formula

$$D_{\text{в}}(X) = D_{\text{в}}(u) = \frac{1}{n}\sum_{i=1}^{k} n_i {u_i}^2 - \left[\frac{1}{n}\sum_{i=1}^{k} n_i u_i\right]^2$$

We get the following result

$$D_{\text{в}} = \frac{1}{10}\cdot 88 - \left(\frac{16}{10}\right)^2 = 8{,}8 - 2{,}56 = 6{,}24$$

The corrected sample dispersion is calculated by the formula

$$S^2 = \frac{n}{n-1} D_{\text{в}}$$

Taking into account that $D_{\text{в}} = 6{,}24; n = 10$ we get

$$S^2 = \frac{10}{9}\cdot 6{,}24 \approx 6{,}93$$

8. A sample of volume $n = 10$ was obtained from the general population

x_i	0,01	0,05	0,09
n_i	2	3	5

Find the corrected sample dispersion for this statistical distribution of the sample.

Solution. Let's go to the conditional variants by the formula

$$u_i = 100x_i$$

The distribution of conditional variants has the form

u_i	1	5	9
n_i	2	3	5

Let's fill in the table

n_i	u_i	${u_i}^2$	$n_i u_i$	$n_i {u_i}^2$
2	1	1	2	2
3	5	25	15	75
5	9	81	45	405
$\sum$			62	482

Find the sample dispersion. To do this, use the formula

$$D_{\text{в}}(X) = \frac{D_{\text{в}}(u)}{C^2},$$

where $D_{\text{в}}(u)$ can be calculated by the formula

$$D_{\text{в}}(u) = \frac{1}{n}\sum_{i=1}^{k} n_i {u_i}^2 - \left[\frac{1}{n}\sum_{i=1}^{k} n_i u_i\right]^2$$

We get

$$D_{\text{в}} = \frac{1}{10000}\left(\frac{482}{10} - \left(\frac{62}{10}\right)^2\right) = \frac{1}{10000}(48{,}2 - 38{,}44) = \frac{9{,}76}{10000} \approx 0{,}001$$

The corrected sample dispersion is calculated by the formula

$$S^2 = \frac{n}{n-1} D_{\text{в}}$$

Taking into account that $D_{\text{в}} = 0{,}001; n = 10$ we get the following result

$$S^2 = \frac{10}{9} \cdot 0{,}001 \approx 0{,}001$$

6.2. Interval estimates of distribution parameters

Point estimates have the disadvantage that they can't be used to judge the accuracy of the estimation. But the accuracy of the estimation in statistical calculations is of great importance. Therefore, it is necessary to determine such a random interval (θ_1, θ_2), which would cover the unknown value of the parameter θ with a given probability γ. So, unlike point estimate, interval estimate is determined by *two numbers* θ_1 *and* θ_2.

The probability γ is called the *confidence level or reliability*. In practice, the reliability of the estimate is set in advance, it is chosen close to one: $\gamma = 0{,}9;\ \gamma = 0{,}95;\ \gamma = 0{,}99$.

The interval (θ_1, θ_2) that covers the unknown value of the parameter θ with a given probability γ is called the *confidence interval.*

6.3. Interval estimation of parameters of normally distributed general population

The interval estimate (with reliability γ) of the mathematical expectation a of a normally distributed quantitative sign X *for a known value of the standard deviation* σ of the general population is called the confidence interval, that we can write in the form

$$\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}},$$

where $\frac{t\sigma}{\sqrt{n}}$ *– estimation accuracy*, n – sample size, t –value of the argument of the Laplace function $\Phi(t)$, which can be found by the formula

$$\Phi(t) = \frac{\gamma}{2}$$

If the value of the standard deviation σ of the general population is not known, and the sample size $n < 30$, then *the interval estimate of the mathematical expectation a* of a normally distributed quantitative sign X is called the confidence interval like this

$$\bar{x} - \frac{t_\gamma S}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma S}{\sqrt{n}},$$

where *S* is the *corrected sample standard deviation*; the values of *t* are found in the table for the given values of *n* and γ.

The interval estimate (with reliability γ) of the standard deviation σ of the normally distributed quantitative sign X is called the confidence interval like this

$$S(1-q) < \sigma < S(1+q) \quad if \quad q < 1$$

or like this

$$0 < \sigma < S(1+q) \quad if \quad q > 1,$$

where *q* is found in the table for given values of *n* and γ.

If it is necessary to estimate the mathematical expectation with a given accuracy and reliability, then the minimum sample size that will ensure this accuracy is found by the formula

$$n = \frac{t^2\sigma^2}{\delta^2}$$

Examples.

1. Find the confidence interval for estimating with reliability 0.95 the unknown mathematical expectation of a normally distributed sign X, if the general standard deviation $\sigma = 5$, the sample mean $\bar{x} = 14$, the sample size

$n = 25$.

Solution. We need to find a confidence interval

$$\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}}$$

All components of this formula are known to us, except for the value of *t*. Find *t* from the relation:

$$\Phi(t) = \frac{\gamma}{2}$$

According to the table of values of the Laplace function, we find the value of *t* at which the equality holds

$$\Phi(t) = 0{,}475$$

We get the value *t*=1,96.

Calculate the left boundary of the confidence interval:

$$\bar{x} - \frac{t\sigma}{\sqrt{n}} = 14 - \frac{1{,}96 \cdot 5}{\sqrt{25}} = 14 - 1{,}96 = 12{,}04.$$

Calculate the right boundary of the confidence interval:

$$\bar{x} + \frac{t\sigma}{\sqrt{n}} = 14 + \frac{1{,}96 \cdot 5}{\sqrt{25}} = 14 + 1{,}96 = 15{,}96.$$

We obtain the following confidence interval

$$12{,}04 < a < 15{,}96$$

2. Find the minimum sample size, if with a reliability of 0,975; the accuracy of estimating the mathematical expectation a of the general population is equal to δ = 0,3. We know that the standard deviation of the normally distributed general population is equal to σ = 1,2.

Solution. The minimum sample size can be found by the formula

$$n = \frac{t^2\sigma^2}{\delta^2}$$

We know all the components of this formula, except for the value of *t*. Find *t* from the relation:

$$\Phi(t) = \frac{\gamma}{2}$$

According to the table of values of the Laplace function, we find the value of t at which the equality holds

$$\Phi(t) = 0{,}4875$$

We get the value *t*=2,24.

Therefore, the minimum sample size

$$n = \frac{t^2\sigma^2}{\delta^2} \approx \frac{7{,}23}{0{,}09} \approx 81$$

3. Find the minimum sample size at which, with a reliability of 0,925; the accuracy of estimating the mathematical expectation a of a normally distributed population is equal to δ = 0,2. We know the standard deviation σ = 1,5 of the population.

Solution. The minimum sample size can be found by the formula

$$n = \frac{t^2\sigma^2}{\delta^2}$$

We know all the components of this formula, except for the value of *t*. Find *t* from the relation:

$$\Phi(t) = \frac{\gamma}{2}$$

According to the table of values of the Laplace function, we find the value of *t* at which the equality holds

$$\Phi(t) = 0{,}4625$$

We get the value *t*=1,78.

Therefore, the minimum sample size

$$n = \frac{t^2\sigma^2}{\delta^2} \approx \frac{7{,}13}{0{,}04} \approx 179$$

4. A sample of volume n = 10 was obtained from the population

x_i	-2	1	2	3	4	5
n_i	2	1	2	2	2	1

Estimate with reliability γ = 0.95 the mathematical expectation *a* of a normally distributed population using a confidence interval.

Solution. According to the given statistical distribution we will find the sample mean, the sample dispersion, the corrected sample dispersion and the corrected standard deviation:

$$\bar{x} = \frac{1}{10}(2 \cdot (-2) + 1 \cdot 1 + 2 \cdot 2 + 2 \cdot 3 + 2 \cdot 4 + 1 \cdot 5) = \frac{20}{10} = 2;$$

$$D_{\text{в}} = \frac{1}{10}(2 \cdot 4 + 1 \cdot 1 + 2 \cdot 4 + 2 \cdot 9 + 2 \cdot 16 + 1 \cdot 25) - (2)^2 = 9{,}2 - 4 = 5{,}2$$

$$S^2 = \frac{10}{9} \cdot 5{,}2 \approx 5{,}78$$

$$S = \sqrt{5{,}78} \approx 2{,}4$$

The confidence interval for the mathematical expectation of the general population has the form

$$\bar{x} - \frac{t_\gamma S}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma S}{\sqrt{n}}$$

All components of this formula are known to us, except for the value of t_γ. Using the table of values of t_γ (you can find these values in the appendix), for $\gamma = 0{,}95,\ n = 10$ we find t_γ.

We obtain the value of $t_\gamma = 2{,}26$.

Calculate the left boundary of the confidence interval:

$$\bar{x} - \frac{t_\gamma S}{\sqrt{n}} = 2 - \frac{2{,}26 \cdot 2{,}4}{\sqrt{10}} \approx 2 - 1{,}7 = 0{,}3.$$

Calculate the right boundary of the confidence interval:

$$\bar{x} + \frac{t_\gamma S}{\sqrt{n}} = 2 + \frac{2{,}26 \cdot 2{,}4}{\sqrt{10}} \approx 2 + 1{,}7 = 3{,}7.$$

We obtain the following confidence interval

$$0{,}3 < a < 3{,}7$$

5. A sample of volume n = 12 was obtained from the population

x_i	-0,5	-0,4	-0,2	0	0,2	0,6	0,8	1	1,2	1,5
n_i	1	2	1	1	1	1	1	1	2	1

Estimate with reliability γ = 0.95 the mathematical expectation *a* of a normally distributed population using a confidence interval.

Solution. According to the given statistical distribution we will find the sample mean, the sample dispersion, the corrected sample dispersion and the corrected standard deviation:

$$\bar{x} = \frac{1}{12}(1 \cdot (-0{,}5) + 2 \cdot (-0{,}4) + 1 \cdot (-0{,}2) + 1 \cdot 0 + 1 \cdot 0{,}2 + 1 \cdot 0{,}6 + 1 \cdot 0{,}8 + 1 \cdot 1 + 2 \cdot 1{,}2 + 1 \cdot 1{,}5) = \frac{5}{12} \approx 0{,}42;$$

$$D_{\text{в}} = \frac{1}{12}(1 \cdot 0{,}25 + 2 \cdot 0{,}16 + 1 \cdot 0{,}04 + 1 \cdot 0 + 1 \cdot 0{,}04 + 1 \cdot 0{,}36 + 1 \cdot 0{,}64 + 1 \cdot 1 + 2 \cdot 1{,}44 + 1 \cdot 2{,}25) - (0{,}42)^2 \approx 0{,}65 - 0{,}18 = 0{,}47$$

$$S^2 = \frac{10}{9} \cdot 0{,}47 \approx 0{,}52$$

$$S = \sqrt{5{,}78} \approx 0{,}72$$

The confidence interval for the mathematical expectation of the general population has the form

$$\bar{x} - \frac{t_\gamma S}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma S}{\sqrt{n}}$$

All components of this formula are known to us, except for the value of t_γ. Using the table of values of t_γ (you can find these values in the appendix), for $\gamma = 0{,}95$, $n = 12$ we find t_γ. We obtain the value of $t_\gamma = 2{,}2$.

Calculate the left boundary of the confidence interval:

$$\bar{x} - \frac{t_\gamma S}{\sqrt{n}} = 0{,}42 - \frac{2{,}2 \cdot 0{,}72}{\sqrt{12}} \approx 0{,}42 - 0{,}46 = -0{,}04.$$

Calculate the right boundary of the confidence interval:

$$\bar{x} + \frac{t_\gamma S}{\sqrt{n}} = 0{,}42 + \frac{2{,}2 \cdot 0{,}72}{\sqrt{12}} \approx 0{,}42 + 0{,}46 = 0{,}88.$$

We obtain the following confidence interval

$$-0{,}04 < a < 0{,}88$$

6. According to the data of *nine* independent measurements of some physical quantity, the arithmetic mean of the measurement results is equal to $\bar{x} = 30{,}1$ and the corrected standard deviation S = 6 were found. Estimate the true value of the measured value using a confidence interval. The reliability of the estimate should be equal to $\gamma = 0{,}99$. It is considered that the measurement results are normally distributed.

Solution. The true value of the measured quantity is equal to its mathematical expectation a. Therefore, the problem is to find the confidence interval for the mathematical expectation a (for an unknown value of the standard deviation σ)

$$\bar{x} - \frac{t_\gamma S}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma S}{\sqrt{n}}$$

All components of this formula are known to us, except for the value of t_γ. Using the table of values of t_γ (you can find these values in the appendix), for $\gamma = 0{,}99$, $n = 9$ we find t_γ. We obtain the value of $t_\gamma = 3{,}36$.

Calculate the left boundary of the confidence interval:

$$\bar{x} - \frac{t_\gamma S}{\sqrt{n}} = 30{,}1 - \frac{3{,}36 \cdot 6}{\sqrt{9}} = 30{,}1 - 6{,}72 = 23{,}38.$$

Calculate the right boundary of the confidence interval:

$$\bar{x} + \frac{t_\gamma S}{\sqrt{n}} = 30{,}1 + \frac{3{,}36 \cdot 6}{\sqrt{9}} = 30{,}1 + 6{,}72 = 36{,}82.$$

We obtain the following confidence interval

$$25{,}38 < a < 36{,}82$$

7. According to the sample, the volume of which is equal to n = 16, the corrected value of the standard deviation S = 1 of the normally distributed quantitative sign X was found. Find the confidence interval for the general standard deviation if the reliability of the estimate is equal to γ = 0.95.

Solution. The task is to find the confidence interval for the general standard deviation

$$S(1-q) < \sigma < S(1+q) \quad if \quad q < 1 \quad \text{or}$$

$$0 < \sigma < S(1+q) \quad if \quad q > 1$$

Using the table of values of q (you can find these values in the appendix), for $\gamma = 0{,}95, \quad n = 16$ we can find q. We obtain the value q=0,44: q <1, so the confidence interval for the general standard deviation calculated by the formula

$$S(1-q) < \sigma < S(1+q)$$

Calculate the left boundary of the confidence interval

$$S(1-q) = 1 \cdot (1 - 0{,}44) = 0{,}56$$

Calculate the right boundary of the confidence interval:

$$S(1+q) = 1 \cdot (1 + 0{,}44) = 1{,}44$$

We get the following result

$$0{,}56 < \sigma < 1{,}44$$

CONCLUSIONS ON THE TOPIC

1. The function $f(X_1, X_2, \dots, X_n)$ from the observed quantities $X_1, X_2, \dots, X_n$ is called a *statistical estimate* of the unknown parameter Θ of the theoretical distribution.

2. *A point estimate* is a statistical estimate determined by a single number $\Theta^* = f(x_1, x_2, \dots, x_n)$, where $x_1, x_2, \dots, x_n$ −are the results of observations on the quantitative sign X (that is the set of $x_1, x_2, \dots, x_n$ is a sample).

3. *An unbiased estimate* is an estimate which mathematical expectation is equal to a parameter that is estimated for any sample size. *A biased estimate* is an estimate which mathematical expectation is not equal to the parameter being estimated.

4. *The unbiased estimate of the general mean* (mathematical expectation) is the *sample mean*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{n}$$

The unbiased estimate of the general dispersion is the corrected sample dispersion, which is calculated by the formula

$$S^2 = \frac{n}{n-1} D_{\text{в}} = \frac{1}{n-1}\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i$$

The biased estimate of the general dispersion is the sample dispersion

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i$$

The sample standard deviation is determined by the formula

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}}$$

or by the formula

$$S_{\text{в}} = \sqrt{S_{\text{в}}^2}$$

5. Point estimates have the disadvantage that they can't be used to judge the accuracy of the estimation. Therefore, it is necessary to determine such a random interval (θ_1, θ_2), which would cover the unknown value of the parameter θ with a given probability γ. The probability γ is called the *confidence level or reliability*. The interval (θ_1, θ_2) that covers the unknown value of the parameter θ with a given probability γ is called the *confidence interval.*

6. *The interval estimate (with reliability γ) of the mathematical expectation a* of a normally distributed quantitative sign X *for a known value of the standard deviation* σ of the general population is called the confidence interval, that we can write in the form

$$\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}}$$

7. *If the value of the standard deviation σ of the general population is not known*, and the sample size $n < 30$, then *the interval estimate of the mathematical expectation a* of a normally distributed quantitative sign X is called the confidence interval like this

$$\bar{x} - \frac{t_\gamma S}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma S}{\sqrt{n}}$$

8. *The interval estimate (with reliability γ) of the standard deviation* σ of the normally distributed quantitative sign X is called the confidence

interval like this

$$S(1-q) < \sigma < S(1+q) \quad if \quad q < 1$$

or like this

$$0 < \sigma < S(1+q) \quad if \quad q > 1$$

9. If it is necessary to estimate the mathematical expectation with a given accuracy and reliability, then the minimum sample size that will ensure this accuracy is found by the formula

$$n = \frac{t^2\sigma^2}{\delta^2}$$

SELF-TEST QUESTIONS

1. *A point estimate* is a statistical estimate determined by
a. a single number
b. a random interval
c. another answer
2. *The interval* (θ_1, θ_2) that covers the unknown value of the parameter θ with a given probability γ is
a. a point estimate of the unknown parameter Θ
b. an interval estimate of the unknown parameter Θ
c. another answer
3. *An unbiased estimate* is an estimate which mathematical expectation is equal to
a. a parameter that is estimated for any sample size
b. a parameter that is estimated for the sample size more than 100
c. another answer
4. *The sample mean is*
a. an unbiased estimate of the mathematical expectation
b. a biased estimate of the mathematical expectation
c. another answer
5. *The sample dispersion is*
a. an unbiased estimate of the general dispersion
b. a biased estimate of the general dispersion
c. another answer
6. *The corrected sample dispersion is*
a. an unbiased estimate of the general dispersion
b. a biased estimate of the general dispersion
c. another answer

7. We know an unbiased estimate of the general dispersion. To find the *sample standard deviation* we should use the formula

a. $\sigma_{\text{в}} = \sqrt{D_{\text{в}}}$

b. $S_{\text{в}} = \sqrt{S_{\text{в}}^2}$

c. another formula

8. We know a biased estimate of the general dispersion. To find the *sample standard deviation* we should use the formula

a. $\sigma_{\text{в}} = \sqrt{D_{\text{в}}}$

b. $S_{\text{в}} = \sqrt{S_{\text{в}}^2}$

c. another formula

9. *The interval estimate* (with reliability γ) of the mathematical expectation *a* of a normally distributed quantitative sign X for a known value of the standard deviation σ of the general population is found by the formula:

a. $\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}}$

b. $\bar{x} - \frac{t_\gamma S}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma S}{\sqrt{n}}$

c. another formula

10. *The interval estimate (with reliability γ) of the standard deviation* σ of the normally distributed quantitative sign X is found by the formula:

a. $S(1-q) < \sigma < S(1+q) \;\; if \;\; q < 1$

b. $0 < \sigma < S(1-q) \;\; if \;\; q > 1$

c. another formula

PRACTICAL TASKS

1. A sample of volume $n = 10$ was obtained from the general population

x_i	12	14	18
n_i	2	3	5

Find the corrected sample dispersion for this statistical distribution of the sample.

Answer: $S^2 \approx 6{,}93$

2. The biased estimate of the general dispersion $D_{в}$=3. Find the unbiased estimate of the dispersion of the general population. The sample size is equal to 41.

Answer: $S^2 = 3{,}075$

3. A sample of volume $n = 20$ was obtained from the general population

x_i	-4	-1	2	3	5	6
n_i	1	3	4	7	3	2

Find: 1) the sample mean; 2) the standard deviation for the sample dispersion; 3) the standard deviation for the corrected sample dispersion.

Answer: $\bar{x} = 2{,}45$; $\sigma_{в} \approx 2{,}08$; $S^2 \approx 4{,}56$

4. A sample is selected from the general population. Its volume is 10 elements. The sample mean is equal to 0,7. Find the unbiased estimate of the general mean.

Answer: $\bar{x} = 0{,}7$

5. There are 30 students in the eighth grade of the gymnasium. 15 students are 13 years old, 12 students are 14 years old, the remaining 3 students are 12 years old. Find the average age of an eighth grader.

Answer: $\bar{x} = 13{,}3$

6. The analysis of the exam results is as follows: the grade "5" was given to 4 students; no one got the grade "4", 6 students got the grade "3"; The grade "2" was given to 2 students. Find the sample dispersion of the grade that students received on the exam.

Answer: $D_{в} = 1{,}25$

LITERATURE FOR SELF-STUDY

1. R. J. Larsen, M. L. Marx. An introduction to mathematical statistics and its applications. – Pearson Education, Inc. 2018. – 753 p.

2. D. Rasch, D. Schott. Mathematical Statistics. – John Wiley & Sons Ltd, 2018. – 676 p.

3. Thomas A. Garrity All the Math you missed Second Edition. – Cambridge University Press, 2021. – 416 p.

Chapter 7
ELEMENTS OF CORRELATION THEORY

7.1. Statistical and correlation dependences. General concepts

Functional dependence is a dependence between quantities (both deterministic and random), when each value of one variable corresponds to a certain value of another.

Statistical (or probabilistic) dependence is a dependence between random variables, when each value of one variable corresponds to a certain (conditional) distribution of another variable. That is, each value of one variable corresponds not to one, but to many possible values of another variable.

The statistical dependence between Y and X is ambiguous, so the researcher is interested in the regularity of change of the *conditional mathematical expectation* $M_x(Y)$ (mathematical expectation of the random variable Y, calculated under the assumption that the variable X took the value of x) depending on x.

The statistical dependence between two variables, in which each value of one variable corresponds to a certain value of the conditional mathematical expectation of another variable, is called *correlation*.

It should be borne in mind that every *correlation dependence is statistical*, but *not every statistical dependence is correlation*.

The correlation dependence can be represented in the form of the following equations:

$$M_x(Y) = \varphi(x)$$
$$M_y(X) = \psi(y)$$

These equations are called *regression equations Y on X and X on Y*, respectively.

Functions φ (x), ψ (y) are called *regression functions*, and their graphs are called *regression lines*. Consider the concept of *the correlation coefficient*.

In mathematical statistics, *the correlation coefficient* is an indicator that characterizes the strength of the statistical dependence between two or more random variables.

A sample correlation coefficient is used to estimate the degree of linear dependence between two continuous random variables. It is calculated by the formula

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – the average value of X,

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ - the average value of Y,

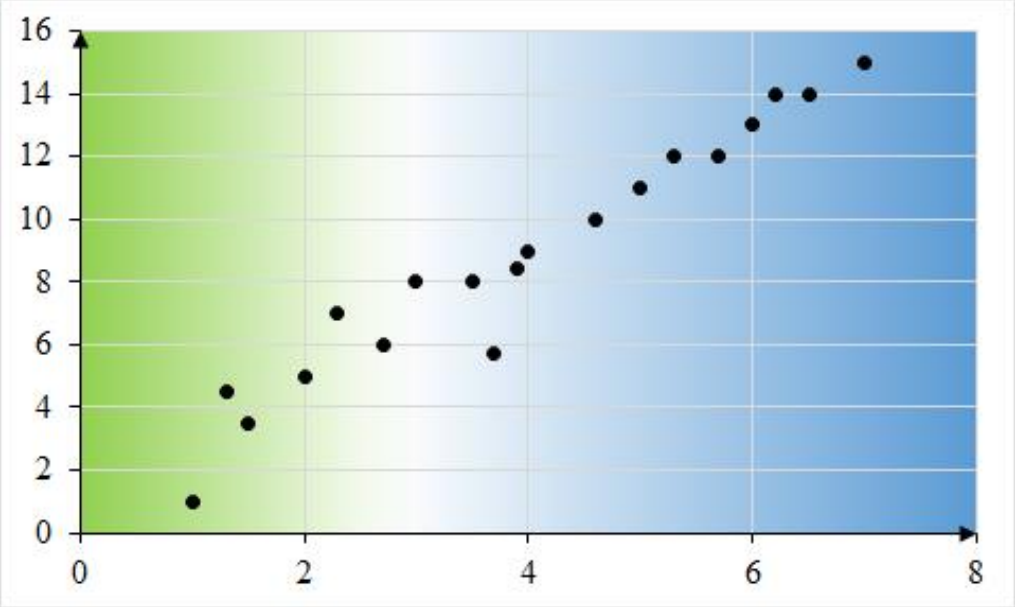
$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ – the average value of the product of X on Y,

$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}$- standard deviation of X,

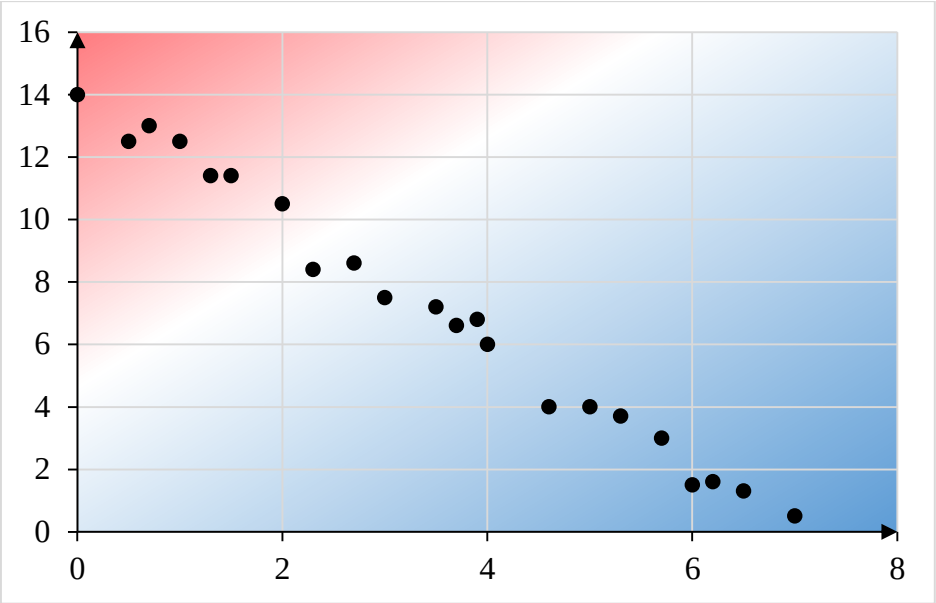
$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}$- standard deviation of Y.

The values of the correlation coefficient are always in the range from -1 to 1 and are interpreted as follows:

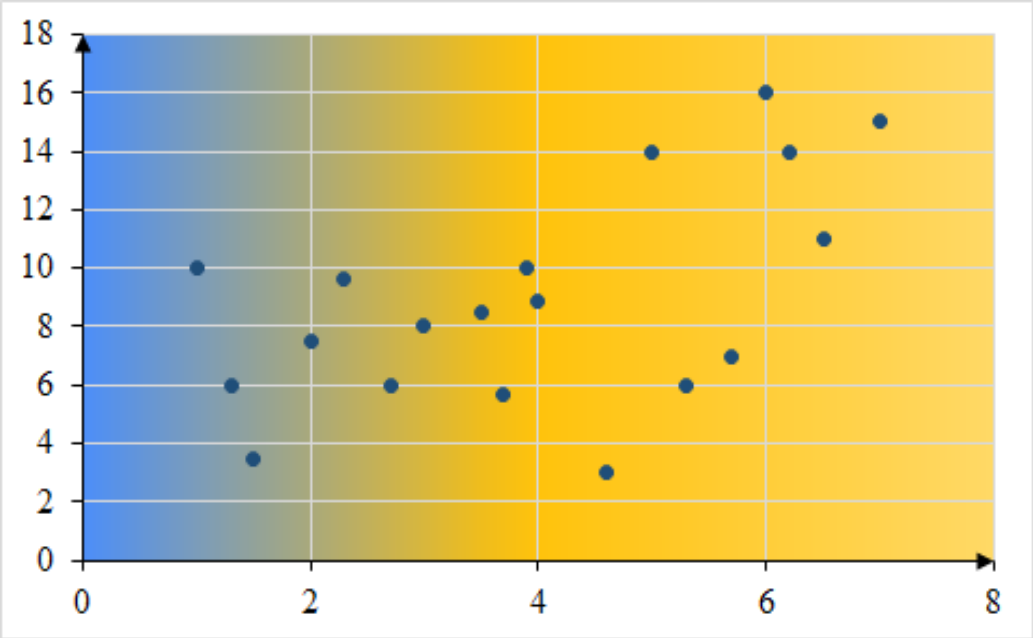
- *if the correlation coefficient is close to 1*, then there is a positive correlation between the variables. That is, if the values of the variable *x* increase, then the variable *y* will also increase. The correlation field in this case has the form



- *if the correlation coefficient is close to -1*, it means that there is a strong negative correlation between the variables. That is, if the value of x will increase, then y will decrease. The correlation field in this case has the form



- intermediate values close to 0 will indicate a weak correlation between the variables and, accordingly, a low dependence between them. That is, the behavior of the variable x completely (or almost completely) will not affect the behavior of y. The correlation field in this case has the form



Example

According to the statistical distribution of values of random variables X and Y:

X	0	10	20	30	40	50	60	70	80
Y	33,5	37,0	41,2	46,1	50,0	52,9	56,8	64,3	69,9

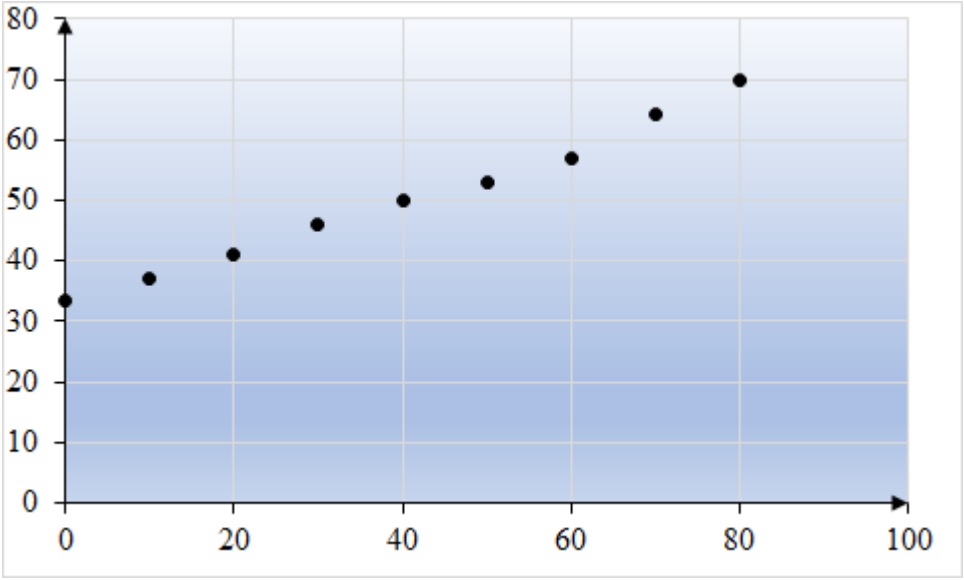
1) build a correlation field;

2) find the sample correlation coefficient;

3) draw a conclusion about the degree of correlation between X and Y.

Solution.

1) The correlation field of the dependence of the sign Y on X has the form



2) For convenience, we will write some intermediate results of calculations in the following table

№	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	0	33,5	0	0	1122,25
2	10	37,0	100	307	1369,00
3	20	41,2	400	824	1697,44
4	30	46,1	900	1383	2125,21
5	40	50,0	1600	2000	2500,00
6	50	52,9	2500	2645	2798,41
7	60	56,8	3600	3408	3226,24
8	70	64,3	4900	4501	4134,49
9	80	69,9	6400	5592	4886,01
$\sum$	360	451,7	20400	20723	23859,05

$$\bar{x} = \frac{1}{9} \cdot 360 = 40$$

$$\bar{y} = \frac{1}{9} \cdot 451{,}7 \approx 50{,}19$$

$$\overline{xy} = \frac{1}{9} \cdot 20723 \approx 2302{,}56$$

$$\overline{x^2} = \frac{1}{9} \cdot 20400 \approx 2266{,}67$$

$$\overline{y^2} = \frac{1}{9} \cdot 23859{,}05 \approx 2651{,}01$$

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{2266{,}67 - 40^2} = \sqrt{666.67} \approx 25.82$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{2651{,}01 - (50{,}19)^2} \approx 11{,}49$$

The sample correlation coefficient is determined by the formula

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y},$$

We obtain the following result:

$$r = \frac{2302{,}56 - 40 \cdot 50{,}19}{25.82 \cdot 11{,}49} = \frac{294{,}96}{296{,}67} \approx 0{,}99$$

3) r = 0.99, so dependence between X and Y is almost linear.

7.2. Linear pair regression. Formulas for calculating the coefficients of the sample linear regression equation

We will look for the sample linear regression equation in the form

$$y = \rho x + b$$

where $\rho, b -$ unknown coefficients to be determined.

The least squares method can be used to determine the unknown coefficients of the regression equation (3). *According to this method*, the unknown coefficients are selected so that the sum of the squares of the deviations of the empirical group averages from the values found by the regression equation was minimal.

The system of normal equations for determining the unknown coefficients of the linear regression equation has the form

$$\begin{cases} \sum_{i=1}^{n} {x_i}^2 \cdot \rho + \sum_{i=1}^{n} x_i \cdot b = \sum_{i=1}^{n} x_i \cdot y_i \\ \sum_{i=1}^{n} x_i \cdot \rho + nb = \sum_{i=1}^{n} y_i \end{cases}$$

Solving system (4), we obtain the following formulas for determining the unknown coefficients of the sample linear regression equation

$$\rho = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = \bar{y} - \rho \cdot \bar{x}$$

Examples

1. Given a statistical distribution of values of random variables X and Y:

X	1	2	3	4	5
Y	3,4	4,4	2,9	0,9	1,4

Using the method of least squares, find the sample equation of the linear regression Y on X.

Solution. There is a linear dependence between the two studied signs Y and X, so the sample regression equation is defined as a linear dependence

$$y = \rho x + b,$$

where

$$\rho = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = \bar{y} - \rho \cdot \bar{x}$$

For convenience, we will write some intermediate results of calculations in the following table

№	x_i	y_i	${x_i}^2$	$x_i y_i$
1	1	3,4	1	3,4
2	2	4,4	4	8,8
3	3	2,9	9	8,7
4	4	0,9	16	3,6
5	5	1,4	25	7
$\sum$	15	13	55	31,5

$$\bar{x} = \frac{1}{5} \cdot 15 = 3$$

$$\bar{y} = \frac{1}{5} \cdot 13 = 2{,}6$$

$$\overline{xy} = \frac{1}{5} \cdot 31{,}5 = 6{,}3$$

$$\overline{x^2} = \frac{1}{5} \cdot 55 = 11$$

We find the unknown coefficients of the linear regression equation:

$$\rho = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{6{,}3 - 3 \cdot 2{,}6}{11 - 3^2} = -\frac{1{,}5}{2} = -0{,}75$$

$$b = \bar{y} - \rho \cdot \bar{x} = 2{,}6 + 0{,}75 \cdot 3 = 4{,}85$$

Therefore, the sample regression equation has the form

$$y = -0{,}75x + 4{,}85$$

2. Given a statistical distribution of values of random variables X and Y:

X	1,5	2	2,5	3	3,5	4
Y	1,3	1,9	2,6	4	5	6,5

Using the method of least squares, find the sample equation of the linear regression Y on X.

Solution. There is a linear relationship between the two studied signs Y and X, so the sample regression equation is defined as a linear dependence

$$y = \rho x + b,$$

where

$$\rho = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = \bar{y} - \rho \cdot \bar{x}$$

For convenience, we will write some intermediate results of calculations in the following table

№	x_i	y_i	${x_i}^2$	$x_i y_i$
1	1,5	1,3	2,25	1,95
2	2	1,9	4	3,8
3	2,5	2,6	6,25	6,5
4	3	4	9	12
5	3,5	5	12,25	17,5
6	4	6,5	16	26
$\sum$	16,5	21,3	49,75	67,75

$$\bar{x} = \frac{1}{6} \cdot 16{,}5 = 2{,}75$$

$$\bar{y} = \frac{1}{6} \cdot 21{,}3 = 3{,}55$$

$$\overline{xy} = \frac{1}{6} \cdot 67{,}75 \approx 11{,}29$$

$$\overline{x^2} = \frac{1}{6} \cdot 49{,}75 \approx 8{,}29$$

We find the unknown coefficients of the linear regression equation:

$$\rho = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{11{,}29 - 2{,}75 \cdot 3{,}55}{8{,}29 - (2{,}75)^2} \approx \frac{1{,}53}{0{,}73} \approx 2{,}10$$

$$b = \bar{y} - \rho \cdot \bar{x} = 3{,}55 — 2{,}10 \cdot 2{,}75 \approx -2{,}23$$

Therefore, the sample regression equation has the form

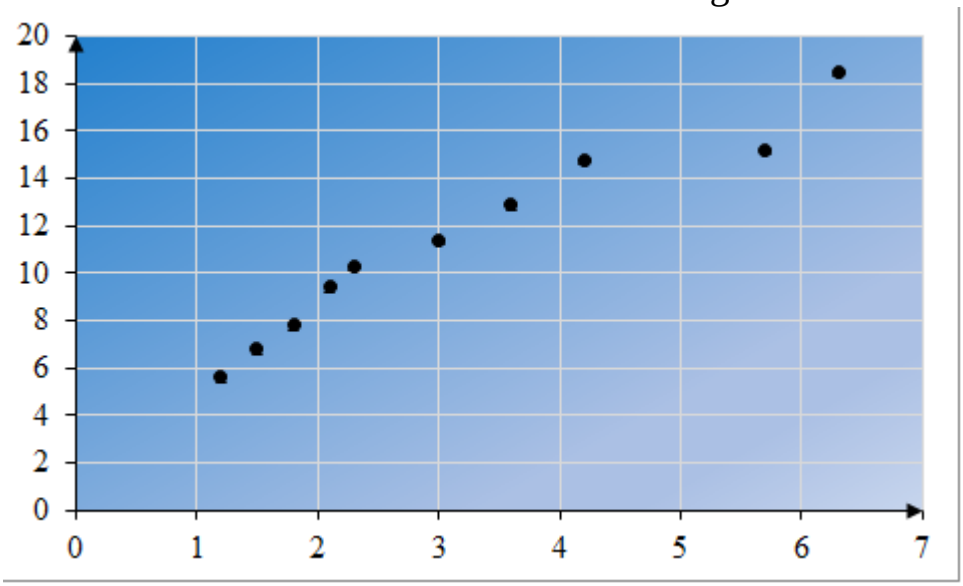
$$y = 2{,}1x - 2{,}23$$

3. Given a statistical distribution of values of random variables X and Y:

x_i	1,2	1,5	1,8	2,1	2,3	3,0	3,6	4,2	5,7	6,3
y_i	5,6	6,8	7,8	9,4	10,3	11,4	12,9	14,8	15,2	18,5

Using the method of least squares, find the sample equation of the linear regression Y on X.

Solution. Here is the correlation field of the original data.



There is a linear dependence between the two studied signs Y and X, so the sample regression equation is defined as a linear dependence

$$y = \rho x + b,$$

where

$$\rho = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = \bar{y} - \rho \cdot \bar{x}$$

For convenience, we will write some intermediate results of calculations in the following table

№	x_i	y_i	${x_i}^2$	$x_i y_i$
1	1,2	5,6	1,44	6,72
2	1,5	6,8	2,25	10,2
3	1,8	7,8	3,24	14,04
4	2,1	9,4	4,41	19,74
5	2,3	10,3	5,29	23,69
6	3	11,4	9	34,2
7	3,6	12,9	12,96	46,44
8	4,2	14,8	17,64	62,16
9	5,7	15,2	32,49	86,64
10	6,3	18,5	39,69	116,55
$\sum$	31,7	112,7	128,41	420.38

$$\bar{x} = \frac{1}{10} \cdot 31{,}7 = 3{,}17$$

$$\bar{y} = \frac{1}{10} \cdot 112{,}7 = 11{,}27$$

$$\overline{xy} = \frac{1}{10} \cdot 420{,}38 = 42{,}038$$

$$\overline{x^2} = \frac{1}{10} \cdot 128{,}41 = 12{,}848$$

We find the unknown coefficients of the linear regression equation:

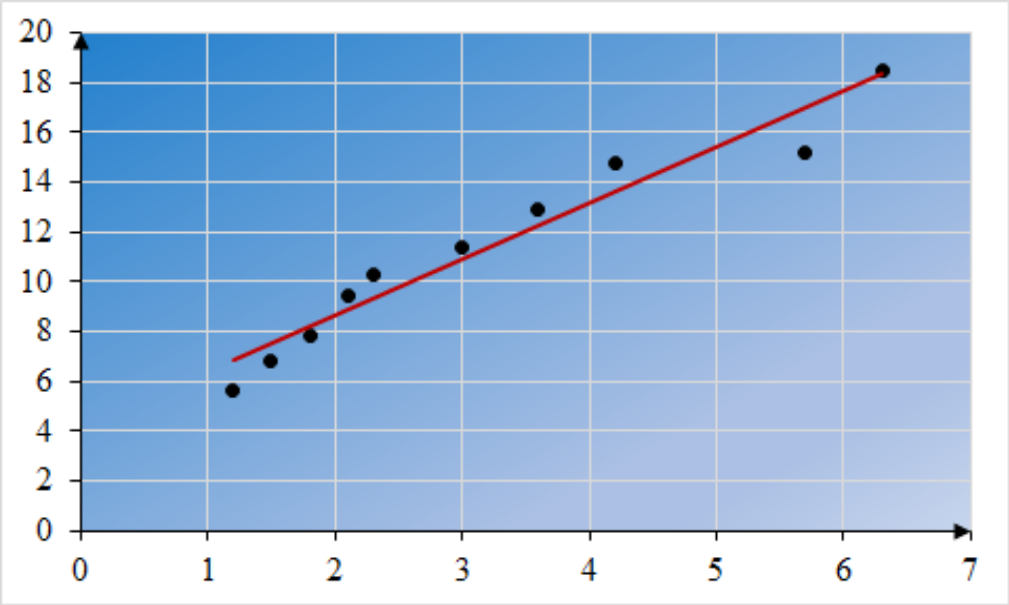
$$\rho = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{42{,}038 - 3{,}17 \cdot 11{,}27}{12{,}848 - 3{,}17^2} \approx \frac{6{,}312}{2{,}799} \approx 2{,}26$$

$$b = \bar{y} - \rho \cdot \bar{x} = 11{,}27 — 2{,}26 \cdot 3{,}17 \approx 4{,}11$$

Therefore, the sample regression equation has the form

$$y = 2{,}26x + 4{,}11$$

Let’s draw a regression line on the correlation field



7.3. Equation of a linear regression on the grouped data

The primary task of the statistical processing of experimental material is *systematization of the received data*. The data obtained from the results of the experiment are disordered, or *ungrouped*. To organize data, they should be divided into groups or classes according to some criteria. Ordered data are called *grouped*.

If the result of the experiment is a set of ordered pairs (x_j, y_i) of two signs X and Y, between which there is a correlation, the experimental data are recorded in a table, which is called *correlation table*.

Example

According to observations of two signs X and Y was obtained a set of pairs (x_j, y_i): (10; 15) – 5 times, (20; 15) – 7 times, (20; 25) – 20 times, (30; 25) – 23 times, (30; 35) – 30 times, (30; 45) – 10 times, (40; 35) – 47 times, (40; 45) – 11 times, (40; 55) – 9 times, (50; 35) – 2 times, (50; 45) – 20 times, (50; 55) – 7 times, (60; 45) – 6 times, (60 ; 55) – 3 times. Make a correlation table.

Solution. Let's form a correlation table (or a table of grouped data)

Y\X	10	20	30	40	50	60	n_y
15	5	7	-	-	-	-	12
25	-	20	23	-	-	-	43
35	-	-	30	47	2	-	79
45	-	-	10	11	20	6	47
55	-	-	-	9	7	3	19
n_x	5	27	63	67	29	9	*n=200*

Let's understand what information the correlation table carries. The top line shows the possible values of X: 10, 20, 30, 40, 50, 60. The bottom line shows the frequencies with which the corresponding variants appeared. For example, variant X = 30 occurred 63 times, X = 50 occurred 29 times, and so on. Thus, the correlation table contains the law of *distribution of the sign X*.

Similarly, the first column contains all the values of the sign Y: 15, 25, 35, 45, 55, and the last column contains the corresponding frequencies. For example, variant Y = 35 occurred in the distribution 79 times and so on. Therefore, the correlation table contains the *distribution law of the sign Y*.

Non-empty cells of the correlation table determine how many times the corresponding pair (x_j, y_i) met: for example, the pair (30,25) met in the distribution 23 times, and the pair (50, 35) met in the distribution 2 times. The sample size is recorded in the lower right corner of the table. In our example, the sample size is n = 200.

The sample equation of the linear regression Y on X for the grouped data has the form

$$\overline{y_x} - \bar{y} = r_{\text{в}} \frac{\sigma_y}{\sigma_x}(x - \bar{x}),$$

where $\overline{y_x}$ – is the conditional mean; $\bar{x}$ and $\bar{y}$ – sample means of X and Y; σ_x, σ_y – sample standard deviations; $r_{\text{в}}$ –sample correlation coefficient, which is calculated by the formula

$$r_{\text{в}} = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n\sigma_x\sigma_y}$$

Similarly, you can write a *sample equation of the linear regression X on Y*. For grouped data, it looks like this

$$\overline{x_y} - \bar{x} = r_{\text{в}} \frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

If in equation of the linear regression Y on X we put $r_{\text{в}}\frac{\sigma_y}{\sigma_x}=\rho$, we get an equation

$$\overline{y_x}-\bar{y}=\rho(x-\bar{x}),$$

where

$$\rho=\frac{\sum n_{xy}xy-n\bar{x}\bar{y}}{n{\sigma_x}^2}=\frac{\sum n_{xy}xy-n\bar{x}\bar{y}}{n(\overline{x^2}-\bar{x}^2)}$$

The sample equation of the linear regression X on Y can also be written in the form

$$\overline{x_y}-\bar{x}=\rho(y-\bar{y})$$

The coefficient ρ will be equal to

$$\rho=\frac{\sum n_{xy}xy-n\bar{x}\bar{y}}{n{\sigma_y}^2}=\frac{\sum n_{xy}xy-n\bar{x}\bar{y}}{n(\overline{y^2}-\bar{y}^2)}.$$

Examples

1. Find the sample equation of a linear regression Y on X according to the correlation table

Y\X	15	20	25	30	35	40	45	n_y
10	-	-	-	-	-	6	1	7
30	-	-	-	-	-	4	2	6
50	-	-	8	10	5	-	-	23
70	3	5	2	-	-	-	-	10
90	2	1	-	1	-	-	-	4
n_x	5	6	10	11	5	10	3	*n=50*

Solution. The sample equation of the linear regression Y on X for the grouped data has the form

$$\overline{y_x}-\bar{y}=\rho(x-\bar{x}),$$

where

$$\rho=\frac{\sum n_{xy}xy-n\bar{x}\bar{y}}{n(\overline{x^2}-\bar{x}^2)}.$$

Find all components of the formula for calculating ρ:

$$\bar{x}=\frac{1}{50}(15\cdot 5+20\cdot 6+25\cdot 10+30\cdot 11+35\cdot 5+40\cdot 10+45\cdot 3)=\frac{1485}{50}=29{,}7$$

$$\bar{y}=\frac{1}{50}(10\cdot 7+30\cdot 6+50\cdot 23+70\cdot 10+90\cdot 4)=\frac{2460}{50}=49{,}2$$

$$\overline{x^2} = \frac{1}{50}(225 \cdot 5 + 400 \cdot 6 + 625 \cdot 10 + 900 \cdot 11 + 1225 \cdot 5 + 1600 \cdot 10 + 2025 \cdot 3)$$
$$= \frac{47875}{50} \approx 957{,}5$$
$$\sum n_{xy}xy = 6 \cdot 10 \cdot 40 + 1 \cdot 10 \cdot 45 + 4 \cdot 30 \cdot 40 + 2 \cdot 30 \cdot 45 + 8 \cdot 50 \cdot 25 +$$
$$+10 \cdot 50 \cdot 30 + 5 \cdot 50 \cdot 35 + 3 \cdot 70 \cdot 15 + 5 \cdot 70 \cdot 20 + 2 \cdot 70 \cdot 25 + 2 \cdot 90 \cdot 15 +$$
$$+1 \cdot 90 \cdot 20 + 1 \cdot 90 \cdot 30 = 64950$$
$$n\bar{x}\bar{y} = 50 \cdot 29{,}7 \cdot 49{,}2 = 73062$$
$$n(\overline{x^2} - \bar{x}^2) = 50(957{,}5 - 29{,}7^2) = 50 \cdot 75{,}41 = 3770{,}5$$
$$\rho = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n(\bar{x}^2 - \bar{x}^2)} = \frac{64950 - 73062}{3770{,}5} = -\frac{8112}{3770{,}5} \approx -2{,}15$$

The sample equation of the linear regression Y on X looks like this

$$\overline{y_x} - 49{,}2 = -2{,}15 \cdot (x - 29{,}7)$$
$$\overline{y_x} - 49{,}2 = -2{,}15x + 63{,}86$$
$$\boldsymbol{\overline{y_x} = -2{,}15x + 113{,}06}$$

2. Find the sample equation of a linear regression Y on X according to the correlation table

Y/X	12	17	22	27	32	37	n_y
25	2	4	-	-	-	-	6
35	-	6	3	-	-	-	9
45	-	-	6	35	4	-	45
55	-	-	2	8	6	-	16
65	-	-	-	14	7	3	24
n_x	2	10	11	57	17	3	*n=100*

Solution. The sample equation of the linear regression Y on X for the grouped data has the form

$$\overline{y_x} - \bar{y} = \rho(x - \bar{x}),$$

where

$$\rho = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n(\overline{x^2} - \bar{x}^2)}.$$

Find all components of the formula for calculating ρ:

$$\bar{x} = \frac{1}{100}(12 \cdot 2 + 17 \cdot 10 + 22 \cdot 11 + 27 \cdot 57 + 32 \cdot 17 + 37 \cdot 3) = \frac{2630}{100} = 26{,}3$$
$$\bar{y} = \frac{1}{100}(25 \cdot 6 + 35 \cdot 9 + 45 \cdot 45 + 55 \cdot 16 + 65 \cdot 24) = \frac{4930}{100} = 49{,}3$$

$$\overline{x^2} = \frac{1}{100}(144 \cdot 2 + 289 \cdot 10 + 484 \cdot 11 + 729 \cdot 57 + 1024 \cdot 17 + 1369 \cdot 3) \approx$$
$$\approx 715{,}7$$

$$\sum n_{xy}xy = 2 \cdot 25 \cdot 12 + 4 \cdot 25 \cdot 17 + 35 \cdot 17 \cdot 6 + 3 \cdot 35 \cdot 22 + 6 \cdot 45 \cdot 22 +$$
$$+35 \cdot 45 \cdot 27 + 4 \cdot 45 \cdot 32 + 2 \cdot 55 \cdot 22 + 8 \cdot 55 \cdot 27 + 6 \cdot 55 \cdot 32 + 14 \cdot 65 \cdot 27 +$$
$$+7 \cdot 65 \cdot 32 + 3 \cdot 65 \cdot 37 = 133610$$

$$n\bar{x}\bar{y} = 100 \cdot 26{,}3 \cdot 49{,}3 \approx 129659$$
$$n(\overline{x^2} - \bar{x}^2) = 100(715{,}7 - 26{,}3^2) = 2401$$

$$\rho = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n(\bar{x}^2 - \bar{x}^2)} = \frac{133610 - 129659}{2401} = \frac{3951}{2401} \approx \mathbf{1{,}65}$$

The sample equation of the linear regression Y on X look like this

$$\overline{y_x} - 49{,}3 = 1{,}65 \cdot (x - 26{,}3)$$
$$\overline{y_x} - 49{,}3 = 1{,}65x - 43{,}39$$
$$\boldsymbol{\overline{y_x} = 1{,}65x + 5{,}91}$$

CONCLUSIONS ON THE TOPIC

1. *Functional dependence* is a dependence between quantities (both deterministic and random), when each value of one variable corresponds to a certain value of another. *Statistical (or probabilistic) dependence* is a dependence between random variables, when each value of one variable corresponds to a certain (conditional) distribution of another variable. The statistical dependence between two variables, in which each value of one variable corresponds to a certain value of the conditional mathematical expectation of another variable, is called *correlation*.

2. *The correlation dependence* can be represented in the form of the following equations:

$$M_x(Y) = \varphi(x)$$
$$M_y(X) = \psi(y)$$

3. These equations are called *regression equations Y on X and X on Y*, respectively. Functions φ (x), ψ (y) are called *regression functions*, and their graphs are called *regression lines*.

4. *A sample correlation coefficient* is used to estimate the degree of linear dependence between two continuous random variables. It is calculated

by the formula

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}$$

5. *The values of the correlation coefficient* are always in the range from -1 to 1 and are interpreted as follows: if the correlation coefficient is close to 1, then there is a positive correlation between the variables. That is, if the values of the variable x increase, then the variable y will also increase; if the correlation coefficient is close to -1, it means that there is a strong negative correlation between the variables. That is, if the value of x will increase, then y will decrease; intermediate values close to 0 will indicate a weak correlation between the variables. That is, the behavior of the variable x completely (or almost completely) will not affect the behavior of y.

6. *The least squares method* can be used to determine the unknown coefficients of the regression equation. According to this method, the unknown coefficients are selected so that the sum of the squares of the deviations of the empirical group averages from the values found by the regression equation will be minimal.

7. *To organize data*, they should be divided into groups or classes according to some criteria. Ordered data are called *grouped*. If the result of the experiment is a set of ordered pairs (x_j, y_i) of two signs X and Y, between which there is a correlation, the experimental data are recorded in a table, which is called *correlation table*.

SELF-TEST QUESTIONS

1. *Statistical (or probabilistic) dependence* is a dependence between random variables

a. which each value of one variable corresponds to a certain value of the conditional mathematical expectation of another variable

b. when each value of one variable corresponds to a certain (conditional) distribution of another variable

c. when each value of one variable corresponds to a certain value of another one

2. *Choose* the correct statement:

a. every *correlation dependence is statistical*

b. every *statistical dependence is correlation*

c. every *functional dependence is correlation*

3. *Regression equation Y on X* has the form

a. $M_x(Y) = \varphi(x)$

b. $M_y(X) = \psi(y)$

c. another formula

4. *A sample correlation coefficient* is used to estimate the degree of dependence between two continuous random variables and this dependence is

a. linear
b. non-linear
c. another answer

5. *A sample correlation coefficient* is calculated by the formula:

a. $r = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}$

b. $r = \dfrac{\overline{xy} - \bar{x}}{\sigma_x}$

c. $r = \dfrac{\overline{xy} - \bar{y}}{\sigma_y}$

6. *If the correlation coefficient* is close to 1, then it means that

a. there is a positive correlation between variables that is, if values of variable *x* increase, then variable *y* will also increase

b. there is a strong negative correlation between variables that is, if the value of x will increase, then y will decrease

c. there is a weak correlation between variables that is, behavior of variable x completely (or almost completely) will not affect behavior of y.

7. *The least squares method* can be used to determine

a. the unknown coefficients of the regression equation
b. the unknown distribution parameters
c. the degree of correlation between two random variables

8. *A correlation table* is a table

a. for grouped data
b. for ungrouped data
c. another answer

9. *The conditional mean* $\overline{y_x}$ is

a. the arithmetic mean of the values of X corresponding to Y=y
b. the arithmetic mean of the values of Y corresponding to X=x
c. another answer

10. *The sample equation of the linear regression Y on X* for the grouped data has the form

a. $\overline{x_y} - \bar{x} = r_{\text{в}} \dfrac{\sigma_x}{\sigma_y}(y - \bar{y})$

b. $\overline{y_x} - \bar{y} = r_{\text{в}} \dfrac{\sigma_x}{\sigma_y}(x - \bar{x})$

c. $\overline{y_x} - \bar{y} = r_{\text{в}} \dfrac{\sigma_y}{\sigma_x}(x - \bar{x})$

PRACTICAL TASKS

1. According to the statistical distribution of values of random variables X and Y we can write:

X	0	1	2	3
Y	5	4	3	7

Find the sample correlation coefficient.

Answer: $r \approx 0{,}38$

2. We know the statistical distribution of values of random variables X and Y:

X	5	8	10	12
Y	11	20	26	32

Build a correlation field. Make assumption about the nature of the dependence between quantitative signs X and Y.

Answer: there is a strict linear dependence between X and Y

3. Given a statistical distribution of values of random variables X and Y:

X	-1	0	2	4
Y	3	4	6	8

Using the method of least squares, find the sample equation of the linear regression Y on X.

Answer: $y = x + 4$

4. Given a statistical distribution of values of random variables X and Y:

X	2	4	7	8
Y	2,5	7,5	12	14,5

Using the least squares method, find the unknown coefficients of the equation of the linear regression Y on X and write this equation.

Answer: $\rho \approx 1{,}9;\ b = -0{,}85;\ y = 1{,}9x - 0{,}85$

LITERATURE FOR SELF-STUDY

1. D. Selvamuthu, D. Das. Introduction to Statistical Methods, Design of Experiments and Statistical Quality Control. – Springer Nature Singapore Pte Ltd., 2018. – 445 p.

2. Gerald Keller. Statistics for Management and Economics, Eleventh Edition. Cengage Learning. 2018. 998 p.

3. S.C. Gupta, I. Gupta. Business statistics. – Himalaya Publishing House Pvt. Ltd, 2019. – 788 p.

Chapter 8
STATISTICAL VERIFICATION OF HYPOTHESES

8.1. Basic concepts

It is often necessary to know the law of distribution of the general population. If the distribution law is unknown, but there is reason to assume that it has a certain form (let's call it A), then a hypothesis is put forward: the general population is distributed according to the law A. Thus, in this hypothesis we are talking about the type of expected distribution.

A case is possible when the distribution law is known, but its parameters are unknown. If there is reason to assume that the unknown parameter θ is equal to a certain value θ_0, put forward a hypothesis: $\theta = \theta_0$. Thus, in this hypothesis we are talking about the estimated value of the parameter of one known distribution.

Other hypotheses are also possible: about the equality of the parameters of two or more distributions, about the independence of samples, and many others.

A statistical hypothesis is a hypothesis about the form of an unknown distribution, or about the parameters of known distributions.

Consider an example of statistical hypotheses:

- the population is distributed according to Poisson's law;
- the variances of two normal populations are equal.

In the first hypothesis, an assumption is made about the form of the unknown distribution, in the second hypothesis, an assumption is made about the parameters of two known distributions.

Along with the hypothesis put forward, a contradictory hypothesis is also considered. If the proposed hypothesis is rejected, then there is a contradictory hypothesis.

The null (main) hypothesis is the hypothesis put forward. It is usually denoted H_0.

A competing (alternative) hypothesis is a hypothesis that contradicts the null hypothesis. The competing hypothesis is designated H_1.

For example, if the null hypothesis is to assume that the expectation value a of the normal distribution is 10, then the competing hypothesis might be to

assume that $a \neq 10$. Briefly it is written like this: $H_0: a = 10;\ H_1: a \neq 10$.

There are hypotheses containing both one and several assumptions.

A simple hypothesis is one that contains only one assumption. For example, if λ is the parameter of the exponential distribution, then the hypothesis $H_0: \lambda = 3$ is simple. Let's give another example of a simple hypothesis: the mathematical expectation of a normal distribution is equal to 2 (the standard deviation is known). The hypothesis in this case will be written as follows: $H_0: a = 2$.

A complex hypothesis is a hypothesis that consists of a finite or infinite number of simple hypotheses. For example, a complex hypothesis: $H_0: \lambda > 5$ consists of an infinite number of simple hypotheses of the form: $H_i: \lambda = b_i$, where b_i is any number more than 5.

The hypothesis put forward may be *correct or incorrect*, so there is a need to verify it. Since the verification is carried out by statistical methods, it is called statistical.

During the statistical verification of the hypothesis, *errors of the first and second types* can be made.

An error of the first kind is that the correct hypothesis will be rejected.

An error of the second kind is that the wrong hypothesis will be accepted.

It should be borne in mind that the consequences of these errors can be different. For example, if the correct decision to "continue building a residential building" is rejected, then this mistake of the first kind will entail material damage; if the wrong decision is made to "continue construction", despite the danger of a collapse of a residential building, then this can lead to casualties.

To verify the null hypothesis, a specially selected random variable is used, the exact or approximate value of which is known. This value is denoted by U or Z if it is normally distributed, χ^2 – if the random variable is distributed according to the law χ^2 and so on.

A statistical criterion (or a criterion) is a random variable (U, Z, χ^2, and so on) that serves to verify the null hypothesis.

To verify hypotheses, the partial values of the quantities included in the criterion are calculated from the data of the samples, and thus the private (observed) value of the criterion is obtained.

The observed value of the criterion is the value calculated from the samples.

After choosing a certain criterion, the set of all its possible values is divided into two non-overlapping subsets: one of them contains the criterion values under which the null hypothesis is rejected, and the other, under which

it is accepted.

The critical area is the set of test values for which the null hypothesis is rejected.

The acceptance area of a hypothesis is the set of criterion values under which the hypothesis is accepted.

The basic principle of testing statistical hypotheses can be formulated as follows: if the observed value of the criterion belongs to the critical area, then the hypothesis is rejected. If the observed value of the criterion belongs to the acceptance area of the hypothesis, then the hypothesis is accepted.

Since the criterion is a one-dimensional random variable, all its possible values belong to a certain interval. Therefore, the critical area and the hypothesis acceptance area are also intervals and, therefore, there are points that separate them.

Critical points (boundaries) are the points separating the critical area from the area of acceptance of the hypothesis.

There are one-sided (right-sided and left-sided) and two-sided critical areas.

The right-sided area is the critical area defined by the inequality $K > k_{cr}$, where K is the criterion, k_{cr} is a positive number.

The left-sided area is the critical area defined by the inequality $K < k_{cr}$, where k_{cr} is a negative number.

One-sided areas are called right-sided and left-sided critical areas.

A two-sided critical area defined by the inequalities $K < k_1$, $K > k_2$, where $k_2 > k_1$. In particular, if the critical points are symmetric with respect to zero, then the two-sided critical area is determined by the inequality $|K| > k_{cr}$.

Consider the question of finding the right-sided critical area. In order to find it, it is enough to find the critical point. To find the critical point, a sufficiently small probability is chosen. This probability is called the significance level α. The critical point is chosen from the condition

$$P(K > k_{cr}) = \alpha, \quad k_{cr} > 0.$$

That is, from the condition that the criterion K will take a value more than k_{cr} with a probability equal to the accepted level of significance.

For each criterion, there are corresponding tables, according to which the critical point is found.

Finding the left-sided and two-sided critical areas is reduced (as well as for the right-handed one) to finding the corresponding critical points.

If we are considering the left- sided area, then the critical point is found from the condition

$$P(K < k_{cr}) = \alpha, \qquad k_{cr} < 0.$$

That is, from the condition that the criterion K will take a value less than k_{cr} with a probability equal to the accepted level of significance.

If the area is two-sided, then the critical points are found from the condition

$$P(K < k_1) + P(K > k_2) = \alpha.$$

If the distribution of the criterion is symmetrical with respect to zero, then the critical points can be found from the condition

$$P(K > k_{cr}) = \alpha/2.$$

If the distribution law is unknown, but there is reason to assume that it has a certain form (for example A), then the null hypothesis is checked: the population is distributed according to the law A.

The verification of the hypothesis about the alleged law of the unknown distribution is carried out using a specially selected value – the goodness of fit criterion.

The goodness of fit criterion is the criterion for testing the hypothesis about the supposed law of the unknown distribution.

There are several goodness of fit criteria: χ^2 criterion Pearson's criterion, Kolmogorov's criterion and other criteria.

In the framework of this lecture, we will consider the Pearson's criterion for testing the hypothesis of an unknown distribution law. We will compare empirical (observed) and theoretical (calculated under the assumption that the distribution law has a certain form) frequencies. Usually empirical and theoretical frequencies differ. The question arises: is this discrepancy coincidental? It is possible that the discrepancy is random (insignificant) and is explained either by a small number of observations, or by the way they are grouped, or by other reasons. It is possible that the frequency discrepancy is not accidental (significant) and is explained by the fact that the theoretical frequencies were calculated based on an incorrect hypothesis about the type of distribution of the general population.

Pearson's criterion answers the above question. However, like any criterion, it does not prove the validity of the hypothesis, but only establishes its agreement (or disagreement) with the observational data at the accepted level of significance.

8.2 Verification of the hypothesis about the normal law of distribution of the general population by Pearson's criterion

Let the empirical distribution be given in the form of a sequence of variants and their corresponding frequencies

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

Verify the hypothesis about a normal distribution law of the general population of X using Pearson's criterion.

Rule.

In order to verify the hypothesis of a normal law of distribution of the general population at a given level of significance α you should:

1. Calculate the sample mean $\bar{x}$ and the sample standard deviation $\sigma_в$.
2. Calculate the theoretical frequencies

$$n_i' = \frac{nh}{\sigma_в} \cdot \varphi(u_i),$$

where n −is the sample size (sum of all frequencies); $h -$ is the step (the difference between two adjacent variants); u_i −are arguments of the function

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2},$$

and the values of u_i are calculated by the formula

$$u_i = (x_i - \bar{x})/\sigma_в.$$

3. Compare empirical and theoretical frequencies using Pearson's criterion. To do this:

a) make a calculation table and find the observed value of the criterion

$$\chi^2_{observed} = \sum \frac{(ni\ - n_i')^2}{n_i'}$$

b) according to the table of critical points of the χ^2 distribution find the number of degrees of freedom

$$k = s - 3,$$

where $s -$ is number of groups of the sample.

c) according to the values of α and k find the critical point

$$\chi^2_{cr}(\alpha; k)$$

If

$$\chi^2_{observed} < \chi^2_{cr}$$

then there is no reason to reject the hypothesis of a normal distribution law of the general population, that is the empirical and theoretical frequencies differ not significantly (accidentally).

If

$$\chi^2_{observed} > \chi^2_{cr,}$$

then the hypothesis about the normal law of the general population is rejected, that is, the empirical and theoretical frequencies differ significantly.

Examples.

1. Using Pearson's criterion, verify whether the hypothesis of a normal distribution of the general population of X agrees with the empirical distribution of the sample, the volume of which is equal to *n*=200. The level of significance should be equal to $\alpha = 0{,}05$.

x_i	5	7	9	11	13	15	17	19	21
n_i	15	26	25	30	26	21	24	20	13

Solution. Let's make a calculation table

i	n_i	x_i	$x_i\, n_i$	x_i^2	$x_i^2 \cdot n_i$	$u_i = {(x_i - \bar{x})}/{\sigma_в}$	$\varphi(u_i)$	n_i'
1	15	5	75	25	375	-1,62	0,1074	9,1
2	26	7	182	49	1274	-1,20	0,1942	16,5
3	25	9	225	81	2025	-0,77	0,2966	25,3
4	30	11	330	121	3630	-0,35	0,3752	32,0
5	26	13	338	169	4394	0,08	0,3977	33,9
6	21	15	315	225	4725	0,51	0,3503	29,8
7	24	17	408	289	6936	0,93	0,2589	22,0
8	20	19	380	361	7220	1,36	0,1582	13,5
9	13	21	273	441	5733	1,78	0,0818	7,0
$\sum$	200		2526		36312			

Find the sample mean and the sample standard deviation:

$$\bar{x} = \frac{1}{200} \cdot 2526 = 12{,}63$$

$$\overline{x^2} = \frac{1}{200} \cdot 36312 = 181{,}56$$

$$\sigma_в = \sqrt{\overline{x^2} - \bar{x}^2} \approx \sqrt{22{,}04} \approx 4{,}695$$

Define step h:

$$h = x_{i+1} - x_i = 2$$

Theoretical frequencies are calculated by the formula

$$n_i' = \frac{nh}{\sigma_в} \cdot \varphi(u_i) = \frac{200 \cdot 2}{4{,}695} \cdot \varphi(u_i) \approx 85{,}2 \cdot \varphi(u_i)$$

Let's compare empirical and theoretical frequencies. To do this, make a calculation table

i	n_i	n_i'	n_i-n_i'	$(ni - n_i')^2$	$(ni - n_i')^2 / n_i'$
1	15	9,1	5,9	34,81	3,8
2	26	16,5	9,5	90,25	5,5
3	25	25,3	-0,3	0,09	0,0
4	30	32,0	-2,0	4,00	0,1
5	26	33,9	-7,9	62,41	1,8
6	21	29,8	-8,8	77,44	2,6
7	24	22,0	2,0	4,00	0,2
8	20	13,5	6,5	42,25	3,1
9	13	7,0	6,0	36,00	5,1
$\sum$	200				$\chi^2_{observed} = 22{,}2$

According to the table of critical distribution points χ^2 at the level of significance α = 0.05 and the number of degrees of freedom k = s-3 = 9-3 = 6 we find

$$\chi^2_{cr}(0{,}05; 6) = 12{,}6$$

Because the

$$\chi^2_{observed} > \chi^2_{cr}$$

then we reject the hypothesis of a normal distribution of the general population.

2. Using Pearson's criterion, verify whether the hypothesis of a normal distribution of the general population X agrees with the empirical distribution of the sample, the volume of which is equal to *n*=200. The level of significance should be equal to $\alpha = 0{,}05$.

x_i	0,3	0,5	0,7	0,9	1,1	1,3	1,5	1,7	1,9	2,1	2,3
n_i	6	9	26	25	30	26	21	24	20	8	5

Solution. Let's make a calculation table

i	n_i	x_i	$x_i\, n_i$	x_i^2	$x_i^2 \cdot n_i$	u_i	$\varphi(u_i)$	n_i'
1	6	0,3	1,8	0,09	0,54	-1,96	0,0584	4,76
2	9	0,5	4,5	0,25	2,25	-1,55	0,1200	9,80
3	26	0,7	18,2	0,49	12,74	-1,15	0,2059	16,81
4	25	0,9	22,5	0,81	20,25	-0,74	0,3034	24,77
5	30	1,1	33	1,21	36,3	-0,33	0,3778	30,84
6	26	1,3	33,8	1,69	43,94	0,08	0,3980	32,49
7	21	1,5	31,5	2,25	47,25	0,49	0,3538	28,88
8	24	1,7	40,8	2,89	69,36	0,89	0,2685	21,91
9	20	1,9	38	3,61	72,2	1,30	0,1714	13,99
10	8	2,1	16,8	4,41	35,28	1,71	0,0925	7,55
11	5	2,3	11,5	5,29	26,45	2,12	0,0422	3,44
$\sum$	200		252,4		366,6			

Find the sample mean and the sample standard deviation:

$$\bar{x} = \frac{1}{200} \cdot 252{,}4 \approx 1{,}26$$

$$\overline{x^2} = \frac{1}{200} \cdot 366{,}6 \approx 1{,}83$$

$$\sigma_{\text{в}} = \sqrt{\overline{x^2} - \bar{x}^2} \approx \sqrt{0{,}24} \approx 0{,}49$$

Define step h:

$$h = x_{i+1} - x_i = 0{,}2$$

Theoretical frequencies are calculated by the formula

$$n_i' = \frac{nh}{\sigma_{\text{в}}} \varphi(u_i) = \frac{200 \cdot 0{,}2}{0{,}49} \cdot \varphi(u_i) \approx 81{,}63 \cdot \varphi(u_i)$$

Let's compare empirical and theoretical frequencies. To do this, make a calculation table

i	n_i	n_i'	$n_i - n_i'$	$(n_i - n_i')^2$	$(ni - n_i')^2 / n_i'$
1	6	4,76	1,24	1,54	0,32
2	9	9,80	-0,8	0,64	0,07
3	26	16,81	9,19	84,46	5,02
4	25	24,77	0,23	0,05	0,002
5	30	30,84	-0,84	0,71	0,02
6	26	32,49	-6,49	42,12	1,30
7	21	28,88	-7,88	62,09	2,15
8	24	21,91	2,09	4,37	0,20
9	20	13,99	6,01	36,12	2,58
10	8	7,55	0,45	0,20	0,03
11	5	3,44	1,56	2,43	0,71
$\sum$	200				$\chi^2_{observed}$ =12,40

According to the table of critical distribution points χ^2 at the level of significance α=0.05 and the number of degrees of freedom $k = s - 3 = 11 - 3 = 8$, we find

$$\chi^2_{cr}(0{,}05; 8) = 15{,}5$$

Because the

$$\chi^2_{observed} < \chi^2_{cr}$$

then we accept the hypothesis of a normal distribution of the general population.

3. Using Pearson's criterion, check the random or significant difference between the empirical frequencies n_i and the theoretical frequencies n_i', which are calculated based on the hypothesis of the normal distribution of the general population X. The level of significance should be equal to $\alpha = 0{,}01$.

n_i	8	16	40	72	36	18	10
n_i'	6	18	36	76	39	18	7

Solution. Let's make a calculation table

i	n_i	n_i'	$n_i - n_i'$	$(ni - n_i')^2$	$(ni - n_i')^2 / n_i'$
1	8	6	2	4	0,667
2	16	18	-2	4	0,222
3	40	36	4	16	0,444
4	72	76	-4	16	0,211
5	36	39	-3	9	0,231
6	18	18	0	0	0
7	10	7	3	9	1,286
$\sum$	200				$\chi^2_{observed} = 3{,}061$

According to the table we find $\chi^2_{observed} = 3{,}061$.

According to the table of critical distribution points χ^2 for a given level of significance α = 0.01 and the number of degrees of freedom

$$k = s - 3 = 7 - 3 = 4,$$

find the critical point

$$\chi^2_{cr}(0{,}01; 4) = 13{,}3$$

Because the

$$\chi^2_{observed} < \chi^2_{cr},$$

then we accept the hypothesis of the normal distribution of the general population, that is the difference between the empirical and theoretical frequencies is not significant (accidentally).

4. Using Pearson's criterion, check the random or significant difference between the empirical frequencies n_i and the theoretical frequencies n_i', which are calculated based on the hypothesis of the normal distribution of the general population X. The level of significance should be equal to $\alpha = 0{,}05$.

n_i	5	10	20	8	7
n_i'	6	14	18	7	5

Solution. Let's make a calculation table

i	n_i	n_i'	$n_i - n_i'$	$(ni - n_i')^2$	$(ni - n_i')^2 / n_i'$
1	5	6	-1	1	0,167
2	10	14	-4	16	1,143
3	20	18	2	4	0,222
4	8	7	1	1	0,143
5	7	5	2	4	0,8
$\sum$	50				$\chi^2_{observed} = 2{,}475$

According to the table we find $\chi^2_{observed} = 2{,}475$.

According to the table of critical distribution points χ^2 for a given level of significance $\alpha = 0.05$ and the number of degrees of freedom

$$k = s - 3 = 5 - 3 = 2,$$

find the critical point

$$\chi^2_{cr}(0{,}05; 2) = 6$$

Because the

$$\chi^2_{observed} < \chi^2_{cr},$$

then we accept the hypothesis of the normal distribution of the general population, that is the difference between the empirical and theoretical frequencies is not significant (accidentally).

5. Using Pearson's criterion, check the random or significant difference between the empirical frequencies n_i and the theoretical frequencies n_i', which are calculated based on the hypothesis of the normal distribution of the general population X. The level of significance should be equal to $\alpha = 0{,}05$.

n_i	14	18	32	70	20	36	10
n_i'	10	24	34	80	18	22	12

Solution. Let's make a calculation table

i	n_i	n_i'	n_i -n_i'	$(ni - n_i')^2$	$(ni - n_i')^2 / n_i'$
1	14	10	4	16	1,6
2	18	24	-6	36	1,5
3	32	34	-2	4	0,118
4	70	80	-10	100	1,25
5	20	18	2	4	0,222
6	36	22	14	196	8,909
7	10	12	-2	4	0,333
$\sum$	200				$\chi^2_{observed} = 13{,}932$

According to the table we find $\chi^2_{observed} = 13,932.$

According to the table of critical distribution points χ^2 for a given level of significance α = 0.05 and the number of degrees of freedom

$$k = s - 3 = 7 - 3 = 4,$$

find the critical point

$$\chi^2_{cr}(0{,}05; 4) = 9{,}5$$

Because the

$$\chi^2_{observed} > \chi^2_{cr},$$

then we reject the hypothesis of a normal distribution of the general population, that is the difference between the empirical and theoretical frequencies is significant.

8.3. Verification the hypothesis of the exponential law of distribution of the general population

An empirical distribution of a continuous random variable X in the form of a sequence of intervals $x_i - x_{i+1}$ and the corresponding frequencies n_i are given, and $\sum n_i = n$ (sample size). It is necessary, using Pearson's criterion, to verify the hypothesis that the random variable X has an exponential distribution.

Rule.

In order to verify the hypothesis that the random variable X has an exponential distribution at the significance level α, it is necessary:

1. Find the sample mean $\bar{x}$ by a given empirical distribution. To do this, use the formula

$$\bar{x} = \frac{1}{n}\sum x_i^* n_i,$$

where

$$x_i^* = \frac{x_i + x_{i+1}}{2}.$$

2. Take as an estimate of the parameter λ of the exponential distribution the value of λ^* which is calculated by the formula

$$\lambda^* = \frac{1}{\bar{x}}$$

3. Find the probabilities of X in the intervals $(x_i,\ x_{i+1})$ by the formula

$$P_i = P(x_i < X < x_{i+1}) = e^{-\lambda^* x_i} - e^{-\lambda^* x_{i+1}}$$

4. Calculate the theoretical frequencies

$$n_i' = n \cdot P_i,$$

where $n = \sum n_i -$ is the sample size.

5. Using Pearson's criterion, compare empirical and theoretical frequencies. The number of degrees of freedom is equal

$$k = s - 2,$$

where *s* – is the number of sampling intervals.

Remark.

Small frequencies (n <5) and their corresponding intervals should be combined. In this case, the number of degrees of freedom k = s-2, where s is the number of sampling intervals remaining after merging.

Examples

1. As a result of testing 200 elements for the duration of work, an empirical distribution was obtained (see the table). The first column shows the time intervals in hours, and the second column shows frequencies, that is the number of elements which operating time belongs to the corresponding time interval. It is necessary at the level of significance α = 0.05 to verify the hypothesis that the operating time of the elements is distributed according to the exponential law.

$x_i - x_{i+1}$	n_i
0-5	133
5-10	45
10-15	15
15-20	4
20-25	2
25-30	1

Solution.

1. Find the average operating time of all elements (as the average operating time of one element we will consider the middle of the interval to which the element belongs):

$$\bar{x} = \frac{1}{200} \cdot (133 \cdot 2{,}5 + 45 \cdot 7{,}5 + 15 \cdot 12{,}5 + 4 \cdot 17{,}5 + 2 \cdot 22{,}5 + 1 \cdot 27{,}5) = \\ = \frac{1000}{200} = 5$$

2. Find the estimate of the parameter λ of the exponential distribution

$$\lambda^* = \frac{1}{\bar{x}} = \frac{1}{5} = 0{,}2$$

3. Find the probabilities of X in each of the intervals $(x_i,\ x_{i+1})$ by the formula

$$P_i = P(x_i < X < x_{i+1}) = e^{-\lambda^* x_i} - e^{-\lambda^* x_{i+1}}$$

$$P_1 = P(0 < X < 5) = e^{-0{,}2 \cdot 0} - e^{-0{,}2 \cdot 5} = 1 - e^{-1} = 0{,}6321$$
$$P_2 = P(5 < X < 10) = e^{-0{,}2 \cdot 5} - e^{-0{,}2 \cdot 10} = e^{-1} - e^{-2} = 0{,}2326$$
$$P_3 = P(10 < X < 15) = e^{-0{,}2 \cdot 10} - e^{-0{,}2 \cdot 15} = e^{-2} - e^{-3} = 0{,}0855$$
$$P_4 = P(15 < X < 20) = e^{-0{,}2 \cdot 15} - e^{-0{,}2 \cdot 20} = e^{-3} - e^{-4} = 0{,}0315$$
$$P_5 = P(20 < X < 25) = e^{-0{,}2 \cdot 20} - e^{-0{,}2 \cdot 25} = e^{-4} - e^{-5} = 0{,}0116$$
$$P_6 = P(25 < X < 30) = e^{-0{,}2 \cdot 25} - e^{-0{,}2 \cdot 30} = e^{-5} - e^{-6} = 0{,}0042$$

4. Find the theoretical frequencies by the formula

$$n_i' = n \cdot P_i,$$

$$n_1' = n \cdot P_1 = 200 \cdot 0{,}6321 = 126{,}42$$
$$n_2' = n \cdot P_2 = 200 \cdot 0{,}2326 = 46{,}52$$
$$n_3' = n \cdot P_3 = 200 \cdot 0{,}0855 = 17{,}10$$
$$n_4' = n \cdot P_4 = 200 \cdot 0{,}0315 = 6{,}3$$
$$n_5' = n \cdot P_5 = 200 \cdot 0{,}0116 = 2{,}32$$
$$n_6' = n \cdot P_6 = 200 \cdot 0{,}0042 = 0{,}84$$

Small frequencies (n <5) and their corresponding intervals should be combined. That is, $n_4 = 4 + 2 + 1 = 7$. Let's combine the corresponding theoretical frequencies. We obtain: $n_4' = 6{,}3 + 2{,}32 + 0{,}84 = 9{,}46$.

5. Compare empirical and theoretical frequencies. To do this, make a table

i	n_i	n_i'	n_i -n_i'	$(ni - n_i')^2$	$(ni - n_i')^2 / n_i'$
1	133	126,42	6,58	43,2964	0,3425
2	45	46,52	-1,52	2,3104	0,0497
3	15	17,10	-2,10	4,4100	0,2579
4	7	9,46	-2,46	6,0516	0,6397
$\sum$	200				$\chi^2_{observed} = 1{,}29$

According to the table of critical distribution points at the level of significance $\alpha = 0{,}05$ and the number of degrees of freedom

$$k = s - 2 = 4 - 2 = 2$$

we find

$$\chi^2_{cr}(0{,}05; 2) = 6.$$

Because the

$$\chi^2_{observed} < \chi^2_{cr}$$

then we accept the hypothesis of the exponential distribution of the general population.

8.4. Verification the hypothesis about the distribution of the general population according to the binomial law

Conducted *n* experiments. Each experiment consists of *k* independent trials, in each of which the probability of event A is the same. The number of occurrences of event A in each trial is recorded. The result is the following distribution of the random variable X – the number of occurrences of event A in each trial.

x_i	0	1	2	…	k
n_i	0	1	2	…	n_k

Using Pearson's criterion, verify the hypothesis of a binomial distribution law of a discrete random variable X.

Rule.

In order to verify the hypothesis at the level of significance α that the discrete random variable X (the number of occurrences of event A in each trial) is distributed according to the binomial law, it is necessary:

1. According to Bernoulli's formula, find the probabilities P_i that event A will appear exactly *i* times in *k* trials.
2. Find the theoretical frequencies

$$n_i' = n \cdot P_i,$$

where *n*- is the number of experiments.

3. Using Pearson's criterion, compare empirical and theoretical frequencies. The number of degrees of freedom is equal

$$k = s - 1,$$

where *s* – is the maximum number of occurrences of the event A in one experiment.

Remark.

The number of degrees of freedom will be equal *k = s-1*, if the probability of event A is known. If the probability of occurrence of event A was estimated by the sample, then the number of degrees of freedom will be equal to *k = s-2*.

Examples.

1. Conducted n = 100 experiments. Each experiment consisted of k = 10 trials, in each of which the probability of occurrence of event A is equal to p = 0.3. The result of the experiment is the following empirical distribution

x_i	0	1	2	3	4	5
n_i	2	10	27	32	23	6

The first line indicates the number of occurrences of event A in one experiment (that is x_i); *the second line* indicates the frequency n_i, that is the number of experiments in which the event A was occurred. It is necessary at the level of significance α to verify the hypothesis that the discrete random variable X (the number of occurrences of event A in each trial) is distributed according to the binomial law.

Solution.

1. According to Bernoulli's formula

$$P_i = P_k\ (i) = C_k^i p^i q^{k-i},\ \ i = 0,1,2,3,4,5$$

We find the probability P_i that event A will appear in $k = 10$ trials exactly i times.

Since $p = 0{,}3;\ \ q = 1 - 0{,}3 = 0{,}7$ we obtain

$P_0 = P_{10}(0) = C_{10}^0 p^0 q^{10} = (0{,}7)^{10} = 0{,}0282$

$P_1 = P_{10}(1) = C_{10}^1 pq^9 = 10 \cdot 0{,}3 \cdot (0{,}7)^9 = 0{,}1211$

$P_2 = P_{10}(2) = C_{10}^2 p^2 q^8 = 45 \cdot (0{,}3)^2 \cdot (0{,}7)^8 = 0{,}2335$

$P_3 = P_{10}(3) = C_{10}^3 p^3 q^7 = 120 \cdot (0{,}3)^3 \cdot (0{,}7)^7 = 0{,}2668$

$P_4 = P_{10}(4) = C_{10}^4 p^4 q^6 = 210 \cdot (0{,}3)^4 \cdot (0{,}7)^6 = 0{,}2001$

$P_5 = P_{10}(5) = C_{10}^5 p^5 q^5 = 252 \cdot (0{,}3)^5 \cdot (0{,}7)^5 = 0{,}1029$

2. Find the theoretical frequencies

$$n_i' = n \cdot P_i,$$

$n_0' = 100 \cdot 0{,}0282 = 2{,}82$

$n_1' = 100 \cdot 0{,}1211 = 12{,}11$

$n_2' = 100 \cdot 0{,}2335 = 23{,}35$

$n_3' = 100 \cdot 0{,}2668 = 26{,}68$

$n_4' = 100 \cdot 0{,}2001 = 20{,}01$

$n_5' = 100 \cdot 0{,}1029 = 10{,}29$

We compare empirical and theoretical frequencies using Pearson's criterion. Let's make a calculation table. Since the frequency $n_0 = 2$ is small, we combine it with the frequency $n_1 = 10$. As a result

$$n_1 = 2 + 10 = 12,$$

and the corresponding theoretical frequency

$$n_1' = 2{,}82 + 12{,}11 = 14{,}93.$$

The calculation table has the form

i	n_i	n_i'	n_i-n_i'	$(ni - n_i')^2$	$(ni - n_i')^2 / n_i'$
1	12	14,93	-2,93	8,5849	0,5750
2	27	23,35	3,65	13,3225	0,5706
3	32	26,68	5,32	28,3024	1,0608
4	23	20,01	2,99	8,9401	0,4468
5	6	10,29	-4,29	18,4041	1,7886
$\sum$	100				$\chi^2_{observed} = 4{,}44$

So,

$$\chi^2_{observed} = 4{,}44$$

According to the table of critical distribution points χ^2 at the level of significance $\alpha = 0{,}05$ and the number of degrees of freedom $k = s - 1 = 5 - 1 = 4$ we find

$$\chi^2_{cr}(0{,}05; 4) = 9{,}5$$

Because

$$\chi^2_{observed} < \chi^2_{cr}$$

then we accept the hypothesis of a binomial distribution law of the random variable X.

2. The experiment, consisting of the simultaneous tossing of four coins, was repeated 100 times. The empirical distribution of the discrete random variable X (the number of coats of arms that appeared) is given in the form of a table

x_i	0	1	2	3	4
n_i	8	20	42	22	8

The first row of the table shows the number of coats of arms, which fell with one tossing coins. The second row of the table indicates the frequency, that is the number of throws at which x_i coats of arms fell. At the significance level α, verify the hypothesis that the random variable X has a binomial distribution law.

Solution.

1. According to Bernoulli's formula

$$P_i = P_k\ (i) = C_k^i p^i q^{k-i},\ \ i = 0{,}1{,}2{,}3{,}4$$

we find the probability P_i that event A (falling out of the coat of arms) will appear in $k=4$ trials exactly i times.

The probability that the coat of arms fall out with one toss of a coin is equal to

$$p = \frac{1}{2} = 0{,}5;$$

The probability of the opposite event (falling out of the number) is equal to

$$q = 1 - 0{,}5 = 0{,}5$$

We obtain the following values of probabilities

$P_0 = P_4\ (0) = q^4 = (0{,}5)^4 = 0{,}0625$

$P_1 = P_4\ (1) = C_4^1 pq^3 = 4 \cdot 0{,}5 \cdot (0{,}5)^3 = 0{,}25$

$P_2 = P_4\ (2) = C_4^2 p^2 q^2 = 6 \cdot (0{,}5)^2 \cdot (0{,}5)^2 = 0{,}375$

$P_3 = P_4\ (3) = C_4^3 p^3 q\ \ = 4 \cdot (0{,}5)^3 \cdot 0{,}5 = 0{,}25$

$P_4 = P_4\ (4) = p^4 = 0{,}0625$

2. Find the theoretical frequencies by the formula: $n_i' = n \cdot P_i$. We can write

$n_0' = 100 \cdot 0{,}0625 = 6{,}25$

$n_1' = 100 \cdot 0{,}25 = 25$

$n_2' = 100 \cdot 0{,}375 = 37{,}5$

$n_3' = 100 \cdot 0{,}25 = 25$

$n_4' = 100 \cdot 0{,}0625 = 6{,}25$

3. We compare empirical and theoretical frequencies using Pearson's criterion. Let's make a calculation table.

i	n_i	n_i'	n_i -n_i'	$(ni\ - n_i')^2$	$(ni\ - n_i')^2 \Big/ n_i'$
1	8	6,25	1,75	3,0625	0,49
2	20	25	-5	25	1
3	42	37,5	4,5	20.25	0,54
4	22	25	-3	9	0,36
5	8	6,25	1,75	3,0625	0,49
$\sum$	100				$\chi^2_{observed} = 2{,}88$

According to the table of critical distribution points χ^2 at the level of significance $\alpha = 0{,}05$ and the number of degrees of freedom $k = s - 1 = 5 - 1 = 4$ we find

$$\chi^2_{cr}(0{,}05; 4) = 9{,}5.$$

Because

$$\chi^2_{observed} < \chi^2_{cr}$$

then we accept the hypothesis of a binomial distribution law of the random variable X.

8.5. Verification the hypothesis about the uniform distribution of the general population

Let's formulate the task. An empirical distribution of a continuous random variable X is given as a sequence of intervals $x_{i-1} - x_i$ and their corresponding frequencies n_i. It is required, using the Pearson's criterion, to test the hypothesis that the random variable X is uniformly distributed.

Rule

In order to verify the hypothesis of a uniform distribution of X, that is, according to the law

$$f(x) = \begin{cases} \dfrac{1}{b-a} & if \;\; x\epsilon(a,b) \\ 0 & in\ other\ cases, \end{cases}$$

it is necessary:

1. Estimate parameters a and b, that is, the ends of the interval in which possible values of X were observed

$$a^* = \bar{x} - \sqrt{3} \cdot \sigma_b; \;\; b^* = \bar{x} + \sqrt{3} \cdot \sigma_b.$$

2. Find the probability density of the estimated distribution

$$f(x) = \frac{1}{b^* - a^*}.$$

3. Find theoretical frequencies:

$$n_1' = nP_1 = nf(x) \cdot (x_1 - a^*) = n \cdot \frac{1}{b^* - a^*}(x_1 - a^*);$$

$$n_2' = n_3' = \cdots n_{s-1}' = n \cdot \frac{1}{b^* - a^*}(x_i - x_{i-1});$$

$$n_s' = n \cdot \frac{1}{b^* - a^*}(b^* - x_{s-1}).$$

4. Compare empirical and theoretical frequencies using Pearson's criterion, considering the number of degrees of freedom equal to $k = s - 3$, where $s -$ the number of intervals into which the sample is divided.

Example.

200 trials were made, as a result of each of which the event A appeared at different points in time. As a result, an empirical distribution was obtained (see table below). It is required at a significance level of 0.05 to verify the hypothesis that the time of occurrence of events is uniformly distributed.

$x_{i-1} - x_i$	n_i	$x_{i-1} - x_i$	n_i
8-9	12	13-14	6
9-10	40	14-15	11
10-11	22	15-16	33
11-12	16	16-17	18
12-13	28	17-18	14

Solution.

Find estimates for the parameters distributed by the uniform law:

$$a^* = \bar{x} - \sqrt{3} \cdot \sigma_b;\ \ b^* = \bar{x} + \sqrt{3} \cdot \sigma_b.$$

In order to find the sample mean and standard deviation, we take the midpoints of the intervals as the observed values of X. We obtain an empirical distribution of equally spaced variants:

x_i^*	8,5	9,5	10,5	11,5	12,5	13,5	14,5	15,5	16,5	17,5
n_i	12	40	22	16	28	6	11	33	18	14

Let's calculate $\bar{x}$ and σ_b:

$$\bar{x} = \frac{1}{200}(102 + 380 + 231 + 184 + 350 + 81 + 159{,}5 + 511{,}5 + 297 + 245) \approx$$
$$\approx 12{,}71$$

$$\overline{x^2} = \frac{1}{200}(867 + 3610 + 2425{,}5 + 2116 + 4375 + 1093{,}5 + 2312{,}75 + 7928{,}25$$
$$+ 4900{,}5 + 4287{,}5) = 169{,}58$$

$$\bar{x}^2 = (12{,}71)^2 \approx 161{,}54$$

$$D_b = \overline{x^2} - \bar{x}^2 = 169{,}58 - 161{,}54 = 8{,}04$$

$$\sigma_b = \sqrt{D_b} \approx 2{,}84$$

So, we can write $a^* = 12{,}71 - \sqrt{3} \cdot 2{,}84 \approx 7{,}79$;

$$b^* = 12{,}71 + \sqrt{3} \cdot 2{,}84 \approx 17{,}63.$$

Find the density of the assumed uniform distribution

$$f(x) = \frac{1}{b^* - a^*} = \frac{1}{17{,}63 - 7{,}79} \approx 0{,}10$$

Let's find theoretical frequencies:

$n_1' = nf(x) \cdot (x_1 - a^*) = 200 \cdot 0{,}10 \cdot (9 - 7{,}79) = 24{,}2;$

$n_2' = 200 \cdot 0{,}10 \cdot (10 - 9) = 20;$

$n_3' = n_4' = \cdots n_9' = 20;$

$n_{10}' = 200 \cdot 0{,}10 \cdot (17{,}63 - 17) = 12{,}6$

Let's compare empirical and theoretical frequencies using Pearson's criterion. The number of degrees of freedom is equal to $k = s - 3 = 10 - 3 = 7$. Let's make a calculation table.

i	n_i	n_i'	n_i-n_i'	$(ni - n_i')^2$	$(ni - n_i')^2 / n_i'$
1	12	24,2	-12,2	148,84	6,15
2	40	20	20	400	20
3	22	20	2	4	0,2
4	16	20	-4	16	0,8
5	28	20	8	64	3,2
6	6	20	-14	196	9,8
7	11	20	-9	81	4,05
8	33	20	13	169	8,45
9	18	20	-2	4	0,2
10	14	12,6	1,4	1,96	0,16
$\sum$	200				$\chi^2_{observed}$ =53,01

Thus, from the calculation table, we learn that $\chi^2_{observed}$ =53,01. According to the table of critical χ^2distribution points, at the significance level of 0,05 and the number of degrees of freedom $k = 7$, we find the critical point of the critical area:

$$\chi^2_{cr}(0{,}05; 7) = 14{,}1$$

decause

$$\chi^2_{observed} > \chi^2_{cr}$$

we reject the hypothesis of a uniform distribution of the general population.

8.6. Verification the hypothesis about the distribution of the general population according to the Poisson's law

Let the empirical distribution of a discrete random variable X be given. It is required, using the Pearson's criterion, to verify the hypothesis about the distribution of the general population according to the Poisson's law.

Rule

In order to verify the hypothesis that the random variable X is distributed according to the Poisson's law at the significance level α, we need:

1. Find the sample mean $\bar{x}$ from a given empirical distribution.
2. Take the Poisson's distribution parameter λ equal to $\bar{x}$, that is we can write

$$\lambda = \bar{x}.$$

3. Use the Poisson's formula to find the probabilities of exactly i events occurring in n trials ($i = 0, 1, 2, \dots, r$), where r is the largest number of observed events, n is the sample size.
4. Find theoretical frequencies using the formula: $n_i' = n \cdot P_i$
5. Compare empirical and theoretical frequencies using the Pearson's criterion, if the number of degrees of freedom is equal to $k = s - 2$, where s is the number of different sample groups.

Let's look at example. As a result of an experiment consisting of 200 trials, in each of which the number of occurrences of some event was recorded, the following empirical distribution was obtained

x_i	0	1	2	3	4
n_i	132	43	20	3	2

It is required at the significance level $\alpha = 0{,}05$ to verify the hypothesis about the distribution of the general population according to the Poisson's law.

Solution.

Let's calculate the sample mean. So,

$$\bar{x} = \frac{1}{200}(43 + 40 + 9 + 8) = 0{,}5.$$

Let's find the Poisson's distribution parameter

$$\lambda = \bar{x} = 0{,}5.$$

Use the Poisson's formula to ind the probabilities of exactly i events $(i = 0, 1, 2, 3, 4)$ occurring in 200 trials:

$$P_n(i) = \frac{(0{,}5)^i \cdot e^{-0{,}5}}{i!} \quad (i = 0, 1, 2, 3, 4)$$

So,

$$P_0 = P_{200}(0) = \frac{(0{,}5)^0 \cdot e^{-0{,}5}}{0!} \approx 0{,}607$$

$$P_1 = P_{200}(1) = \frac{0{,}5 \cdot e^{-0{,}5}}{1!} = 0{,}305$$

$$P_2 = P_{200}(2) = \frac{(0{,}5)^2 \cdot e^{-0{,}5}}{2!} \approx 0{,}076$$

$$P_3 = P_{200}(3) = \frac{(0{,}5)^3 \cdot e^{-0{,}5}}{3!} \approx 0{,}013$$

$$P_4 = P_{200}(4) = \frac{(0{,}5)^4 \cdot e^{-0{,}5}}{4!} \approx 0{,}002$$

Find theoretical frequencies:

$n_0' = n \cdot P_0 = 200 \cdot 0{,}607 = 121{,}4$
$n_1' = n \cdot P_1 = 200 \cdot 0{,}305 = 61$
$n_2' = n \cdot P_2 = 200 \cdot 0{,}076 = 15{,}2$
$n_3' = n \cdot P_3 = 200 \cdot 0{,}013 = 2{,}6$
$n_4' = n \cdot P_4 = 200 \cdot 0{,}002 = 0{,}4$

Let's combine small theoretical frequencies: $n_3' + n_4' = 2{,}6 + 0{,}4 = 3$.

Let's combine small empirical frequencies: $n_3 + n_4 = 3 + 2 = 5$.

Compare empirical and theoretical frequencies using the Pearson's criterion. To do this, we will make a table

i	n_i	n_i'	n_i -n_i'	$(ni - n_i')^2$	$(ni - n_i')^2 / n_i'$
0	132	121,4	10,6	112,36	0,93
1	43	61	-18	324	5,31
2	20	15,2	4,8	23,04	1,52
3	5	3	2	4	1,33
$\sum$	200				$\chi^2_{observed}$ =9,09

Thus, from the calculation table, we learn that $\chi^2_{observed}$ =9,09. According to the table of critical χ^2distribution points, at the significance level of 0,05 and the number of degrees of freedom $k = 4 - 2 = 2$, we find the critical point of the critical area:

$$\chi^2_{cr}(0{,}05; 7) = 14{,}1.$$

Because

$$\chi^2_{observed} > \chi^2_{cr}$$

we reject the hypothesis of a uniform distribution of the general population.

CONCLUSIONS ON THE TOPIC

1. *A statistical hypothesis* is a hypothesis about the form of an unknown distribution, or about the parameters of known distributions.
2. *The null (main) hypothesis* is the hypothesis put forward. *A competing (alternative) hypothesis* is a hypothesis that contradicts the null hypothesis.
3. The hypothesis put forward may be *correct or incorrect*, so there is a need to verify it. Since the verification is carried out by statistical methods, it is called statistical.
4. *The significance level* is the probability that a correct null hypothesis will be rejected
5. *A statistical criterion* (or a criterion) is a random variable that serves to verify the null hypothesis.
6. *To verify hypotheses*, the partial values of the quantities included in the criterion are calculated from the data of the samples, and thus the private (observed) value of the criterion is obtained.
7. *The critical area* is the set of test values for which the null hypothesis is rejected.
8. *The acceptance area of a hypothesis* is the set of criterion values under which the hypothesis is accepted.
9. *The basic principle of testing statistical hypotheses* can be formulated as follows: if the observed value of the criterion belongs to the critical area, then the hypothesis is rejected. If the observed value of the criterion belongs to the acceptance area of the hypothesis, then the hypothesis is accepted.
10. *To verify the statistical hypothesis* about the form of the unknown distribution of the general population, *Pearson's criterion* is used. If the observed value of the criterion is less than the critical one, that is $\chi^2_{observed} < \chi^2_{cr}$, then we accept the proposed hypothesis about the form of the unknown distribution. Otherwise, we reject the proposed hypothesis.

SELF-TEST QUESTIONS

1. A statistical hypothesis is a hypothesis

a. about the form of an unknown distribution or parameters of a known distribution

b. about the parameters of an unknown distribution

c. another answer

2. A simple hypothesis is a hypothesis consisting of

a. one assumption

b. two or more assumptions

c. another answer

3. Significance level is the probability that

a. a correct null hypothesis will be rejected

b. an incorrect null hypothesis will be accepted

c. another answer

4. The critical area is the set of all criterion values under which

a. the null hypothesis is rejected

b. the null hypothesis is accepted

c. another answer

5. The acceptance area of a hypothesis is the set of all values of the criterion under which

a. the null hypothesis is rejected

b. the null hypothesis is accepted

c. another answer

6. The hypothesis about the exponential distribution of the general population according to the Pearson's criterion is tested. What formula can be used to find theoretical frequencies?

a. $n_i' = n_i \cdot P_{i,} \quad where\ P_i = P_N(i) = C_N^i \cdot p^i \cdot q^{N-i}$

b. $n_i' = n_i \cdot P_{i,} \quad where\ P_i = P_n(i) = \frac{\lambda^i \cdot e^{-\lambda}}{i!}$

c. $n_i' = n_i \cdot P_{i,} \quad where\ \ P_i = e^{-\lambda x_i} - e^{-\lambda x_{i+1}}$

7. The hypothesis about the form of distribution of the general population according to the Pearson's criterion is tested. What formula should be used to calculate the observed value of the criterion?

a. $\chi^2_{observed} = \sum_i \frac{(n_i - n_i')^2}{n_i}$

b. $\chi^2_{observed} = \sum_i \frac{(n_i - n_i')^2}{n_i'}$

c. $\chi^2_{observed} = \sum_i \frac{n_i - n_i'}{n_i^2}$

8. The hypothesis about the form of distribution of the general population according to the Pearson's criterion is tested. What formula should be used in this case to calculate the number of degrees of freedom? The number of sample groups S and parameters r estimated from the sample are known.

a. $k = S - 1 - r$
b. $k = S - 2 - r$
c. $k = S - 3 - r$

9. When testing the hypothesis of the uniform distribution of the general population, it was found that with the significance level α and the number of degrees of freedom k, the inequality $\chi^2_{observed} > \chi^2_{cr}$ is satisfied. What conclusion can be drawn?

a. should accept the hypothesis of a uniform distribution
b. reject the hypothesis of a uniform distribution
c. nothing definite can be said about the distribution of the general population

10. If the observed value of the criterion is less than the critical one, that is $\chi^2_{observed} < \chi^2_{cr}$, then this means that

a. empirical and theoretical frequencies differ randomly
b. empirical and theoretical frequencies differ significantly
c. nothing definite can be said about the difference between empirical and theoretical frequencies

PRACTICAL TASKS

1. Given a statistical distribution of the random variable X

x_i	1	4	7	10	13
n_i	7	11	9	8	15

Using Pearson's criterion, verify whether the hypothesis of a normal distribution of the general population of X agrees with the empirical distribution of the sample. The level of significance should be equal to $\alpha = 0{,}01$.

Answer: $\chi^2_{observed} = 10{,}995$; $\chi^2_{cr}(0{,}01; 2) = 9{,}2$; $\chi^2_{observed} > \chi^2_{cr}$, *that is we reject the hypothesis of a normal distribution of the general population of X.*

2. As a result of an experiment consisting of 70 trials the following empirical distribution was obtained

$x_{i-1} - x_i$	n_i
1-3	10
3-5	15
5-7	25
7-9	20

It is required at a significance level of 0.05 to verify the hypothesis that the quantitative sign X is uniformly distributed.

Answer: $\chi^2_{observed} = 2{,}438; \quad \chi^2_{cr}(0{,}05; 1) = 3{,}8; \quad \chi^2_{observed} < \chi^2_{cr,}$ *that is we accept the hypothesis of the uniform distribution of the general population of X.*

3. As a result of an experiment consisting of 50 trials the following empirical distribution was obtained

$x_{i-1} - x_i$	n_i
0-1	8
1-2	7
2-3	12
3-4	10
4-5	13

Using Pearson's criterion, verify whether the hypothesis of the exponential distribution of the general population of X agrees with the empirical distribution of the sample. The level of significance should be equal to $\alpha = 0{,}05$.

Answer: : $\chi^2_{observed} = 37{,}91; \quad \chi^2_{cr}(0{,}05; 3) = 7{,}8; \quad \chi^2_{observed} > \chi^2_{cr},$ *that is we reject the hypothesis of the exponential distribution of the general population of X.*

LITERATURE FOR SELF-STUDY

1. D. Selvamuthu, D. Das. Introduction to Statistical Methods, Design of Experiments and Statistical Quality Control. – Springer Nature Singapore Pte Ltd., 2018. – 445 p.

2. Gerald Keller. Statistics for Management and Economics, Eleventh Edition. Cengage Learning. 2018. 998 p.

3. S.C. Gupta, I. Gupta. Business statistics. – Himalaya Publishing House Pvt. Ltd, 2019. – 788 p.

ANSWERS TO SELF-TEST QUESTIONS

	CHAPTER							
	1	2	3	4	5	6	7	8
1	b	b	a	b	a	a	b	a
2	b	b	c	a	b	b	a	b
3	a	b	b	a	a	a	a	a
4	c	a	a	b	a	a	a	a
5	a	b	c	c	a	b	a	b
6	a	b	b	c	b	a	a	c
7	a	c	c	b	c	b	a	b
8	c	a	b	a	c	a	a	a
9	a	c	c	a	a	a	b	b
10	c	b	b	b	b	a	c	a

QUESTIONS FOR FINAL CONTROL

1. What is the *subject of probability theory?*
2. What is a *random event*?
3. What is a *reliable event*? Give examples of the *reliable event*.
4. What event is called *impossible*? Give examples of the *impossible event*.
5. Give examples of *equally possible events*.
6. What are *permutations*? Give a definition and write down the calculation formula.
7. What are *combinations*? Give a definition and write down the calculation formula.
8. What are *placements*? Give a definition and write down the calculation formula.
9. Formulate the *classical* and *statistical* definitions of probability.
10. Formulate the *addition theorem* of the probabilities of random events.
11. Formulate a *multiplication theorem* of the probabilities of random events.
12. Define an *opposite event*. Write down the formula.
13. What *events* are called *independent*?

14. What *trials* are called *independent*? Write down the Bernoulli's formula.
15. Formulate the *local and integral theorems of Laplace*.
16. What is a *random variable*? What is the difference between continuous and discrete random variables?
17. How can the *law of distribution* of a discrete random variable be given?
18. What is the *integral distribution function*? Formulate the definition and write down the formula.
19. Formulate the *properties* of the integral distribution function.
20. What is a *differential distribution function*? Formulate the definition and write down the formula.
21. What *numerical characteristics* does a random variable have?
22. Write down the *formulas for calculating* the numerical characteristics of a *discrete random variable.*
23. Write down the *formulas for calculating* the numerical characteristics of a *continuous random variable*.
24. What *laws of distribution* of a discrete random variable do you know?
25. Write down the *formulas for calculating* the numerical characteristics of a random variable that has a *binomial distribution*.
26. Write down the *formulas for calculating* the numerical characteristics of a random variable distributed according to the *Poisson's law*.
27. What *laws of distribution* of a continuous random variable do you know?
28. Write down the *formulas for calculating* the density and distribution function of a continuous random variable *distributed uniformly* over the interval (a, b).
29. The random variable X is *uniformly* distributed over the interval (a, b). Write down the *formulas for calculating* the mathematical expectation, variance and standard deviation X.
30. Write down the *formulas for calculating* the density and distribution function of a continuous random variable distributed according to the *normal law.*
31. The random variable X has a *normal distribution law* with parameters a, σ. What is the *mathematical expectation* and *standard deviation* of X?
32. Write down *formulas for calculating* the density and distribution function of a continuous random variable distributed according to an

exponential law.

33. Write down the *formulas for calculating* the numerical characteristics of a continuous random variable distributed according to an *exponential law*.

34. Define the *reliability function*. Write down the formula to calculate it.

35. Formulate the *main tasks* of mathematical statistics.

36. What is a *general population*? Give examples of a *general population*.

37. What is a *sample*? What is a *representative sample*?

38. Give the definition of the *empirical distribution function* and write down the formula for its calculation.

39. Formulate the *main properties* of the empirical distribution function.

40. How do you understand what a *frequency polygon* is? What is a *histogram*?

41. What is the *difference* between *point and interval estimates* of unknown distribution parameters?

42. What is an *unbiased estimate* of the mathematical expectation, an *unbiased estimate* of the general dispersion, a *biased estimate* of the general dispersion? Write down the formulas.

43. How to find the *confidence interval for the mathematical expectation* of a normally distributed quantitative sign X *with a known standard deviation*?

44. How to find the *confidence interval for the mathematical expectation* of a normally distributed quantitative sign X with *an unknown standard deviation*?

45. How to find the *confidence interval for the standard deviation* of a normally distributed feature X?

46. What is a *statistical dependence*?

47. What kind of dependence is called *correlation*?

48. There is a *correlation* between X and Y. Can this dependence be considered *statistica*l? Can any *statistical dependence* be considered as a *correlation*?

49. What is a *sample correlation coefficient* and what formula is used to find it?

50. What conclusion about the nature of the *dependence* between X and Y can be made if the value of the *sample correlation coefficient is close to one*?

51. What conclusion about the nature of the *dependence* between two

random variables can be made if the value of the *sample correlation coefficient is close to zero*?

52. How to use the *least squares method* to find unknown coefficients in the equation of the linear regression?

53. What is a *correlation table* and what is its assignment?

54. What is a *statistical hypothesis*? Give examples of a statistical hypothesis.

55. What is a *simple hypothesis*, a *complex hypothesis*? Give examples of the simple and complex hypotheses.

56. What is a *statistical criterion*? What is the *basic principle* for testing statistical hypotheses.

57. Formulate the concepts: a *right-sided critical area*, *left-sided critical area*, *two-sided critical area.*

58. How to apply the *Pearson's criterion* to verify the hypothesis of a *normal distribution* of the general population? Formulate a rule.

59. How to apply the *Pearson's criterion* to verify the hypothesis of an *exponential distribution* of the general population? Formulate a rule.

60. How to apply the Pearson's *criterion* to verify the hypothesis of a *binomial distribution* of the general population? Formulate a rule.

61. How, using *Pearson's criterion*, to verify the hypothesis of a *uniform distribution* of the general population? Formulate a rule.

INDEX

LITERATURE

BASIC

1. James Nicholson. Complete Probability & Statistics 1 for Cambridge International AS & A Level. – Oxford University Press – Children, 2019. – 226 p.

2. James Nicholson. Complete Probability & Statistics 2 for Cambridge International AS & A Level. – Oxford University Press – Children, 2019. – 210 p.

3. J. K. Blitzstein, J. Hwang, Introduction to Probability Second Edition. – Taylor & Francis Group, LLC, 2019. – 636 p.

4. D. Rasch, D. Schott. Mathematical Statistics. – John Wiley & Sons Ltd, 2018. – 676 p.

5. R. J. Larsen, M. L. Marx. An introduction to mathematical statistics and its applications. – Pearson Education, Inc. 2018. – 753 p.

6. A.V. Tyurinand, A.Yu. Akhmerov Theory of probability and mathematical statistics: Textbook. − Dusseldorf: LAP LAMBERT Academic Publishing GmbH & Co.KG., 2020. – 148 p.

7. Prasanna Sahoo probability and mathematical statistics: Textbook. – USA: Department of Mathematics of the University of Louisville, 2013. – 712 p.

8. Mary C. Meyer Probability and Mathematical Statistics: Theory, Applications, and Practice in R. – USA: Society for Industrial and Applied Mathematics, 2019. – 707 p.

9. Thomas A. Garrity All the Math you missed Second Edition. – Cambridge University Press, 2021. – 416 p.

FOR IN-DEPTH STUDY OF THE COURSE

1. D. Selvamuthu, D. Das. Introduction to Statistical Methods, Design of Experiments and Statistical Quality Control. – Springer Nature Singapore Pte Ltd., 2018. – 445 p.

2. Gerald Keller. Statistics for Management and Economics, Eleventh Edition. Cengage Learning. 2018. 998 p.

3. S.C. Gupta, I. Gupta. Business statistics. – Himalaya Publishing House Pvt. Ltd, 2019. – 788 p.

ANNEX 1

TABLE OF VALUES OF THE LOCAL LAPLACE FUNCTION

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	39986	3984	3982	3980	3977	3973
0,1	0,3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	0,3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	0,3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	0,3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	0,3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	0,3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	0,3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	0,2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	0,2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	0,2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	0,1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	0,1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	0,1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	0,1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	0,1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0,0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0,0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0,0656	0644	0632	0620	0608	0596	0584	0573	0562	0551

x	0	1	2	3	4	5	6	7	8	9
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0,0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0,0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0,0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0,0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0,0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0,0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0,0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0,0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0,0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0,0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0,0024	0023	0022	0022	0021	0020	0020	0019	0018	0018

x	0	1	2	3	4	5	6	7	8	9
3,3	0,0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0,0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0,0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0,0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0,0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0,0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0,0002	0002	0002	0002	0002	0002	0002	0002	0001	0001
4,0	0,0001338		0000589		0000249		0000101		0000040	
5,0	0,0000015									

ANNEX 2

TABLE OF VALUES OF THE INTEGRAL LAPLACE FUNCTION

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	**0,50**	0,1915	**1,00**	0,3413	**1,50**	0,4332	**2,00**	0,4772	**3,00**	0,49865
0,01	0,0040	**0,51**	0,1950	**1,01**	0,3438	**1,51**	0,4345	**2,02**	0,4783	**3,20**	0,49931
0,02	0,0080	**0,52**	0,1985	**1,02**	0,3461	**1,52**	0,4357	**2,04**	0,4793	**3,40**	0,49966
0,03	0,0120	**0,53**	0,2019	**1,03**	0,3485	**1,53**	0,4370	**2,06**	0,4803	**3,60**	0,499841
0,04	0,0160	**0,54**	0,2054	**1,04**	0,3508	**1,54**	0,4382	**2,08**	0,4812	**3,80**	0,499928
0,05	0,0199	**0,55**	0,2088	**1,05**	0,3531	**1,55**	0,4394	**2,10**	0,4821	**4,00**	0,499968
0,06	0,0239	**0,56**	0,2123	**1,06**	0,3554	**1,56**	0,4406	**2,12**	0,4830	**4,50**	0,499997
0,07	0,0279	**0,57**	0,2157	**1,07**	0,3577	**1,57**	0,4418	**2,14**	0,4838	**5,00**	0,499997
0,08	0,0319	**0,58**	0,2190	**1,08**	0,3599	**1,58**	0,4429	**2,16**	0,4846		
0,09	0,0359	**0,59**	0,2224	**1,09**	0,3621	**1,59**	0,4441	**2,18**	0,4854		
0,10	0,0398	**0,60**	0,2257	**1,10**	0,3643	**1,60**	0,4452	**2,20**	0,4861		
0,11	0,0438	**0,61**	0,2291	**1,11**	0,3665	**1,61**	0,4463	**2,22**	0,4868		
0,12	0,0478	**0,62**	0,2324	**1,12**	0,3686	**1,62**	0,4474	**2,24**	0,4875		
0,13	0,0517	**0,63**	0,2357	**1,13**	0,3708	**1,63**	0,4484	**2,26**	0,4881		
0,14	0,0557	**0,64**	0,2389	**1,14**	0,3729	**1,64**	0,4495	**2,28**	0,4887		
0,15	0,0596	**0,65**	0,2422	**1,15**	0,3749	**1,65**	0,4505	**2,30**	0,4893		
0,16	0,0636	**0,66**	0,2454	**1,16**	0,3770	**1,66**	0,4515	**2,32**	0,4898		
0,17	0,0675	**0,67**	0,2486	**1,17**	0,3790	**1,67**	0,4525	**2,34**	0,4904		
0,18	0,0714	**0,68**	0,2517	**1,18**	0,3810	**1,68**	0,4535	**2,36**	0,4909		
0,19	0,0753	**0,69**	0,2549	**1,19**	0,3830	**1,69**	0,4545	**2,38**	0,4913		
0,20	0,0793	**0,70**	0,2580	**1,20**	0,3849	**1,70**	0,4554	**2,40**	0,4918		
0,21	0,0832	**0,71**	0,2611	**1,21**	0,3869	**1,71**	0,4564	**2,42**	0,4922		

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,22	0.0871	**0,72**	0.2642	**1,22**	0.3883	**1,72**	0.4573	**2,44**	0.4927		
0,23	0.0910	**0,73**	0.2673	**1,23**	0.3907	**1,73**	0.4582	**2,46**	0.4931		
0,24	0.0948	**0,74**	0.2703	**1,24**	0.3925	**1,74**	0.4591	**2,48**	0.4934		
0,25	0.0987	**0,75**	0.2734	**1,25**	0.3944	**1,75**	0.4599	**2,50**	0.4938		
0,26	0.1026	**0,76**	0.2764	**1,26**	0.3962	**1,76**	0.4608	**2,52**	0.4941		
0,27	0.1064	**0,77**	0.2794	**1,27**	0.3980	**1,77**	0.4616	**2,54**	0.4945		
0,28	0.1103	**0,78**	0.2823	**1,28**	0.3997	**1,78**	0.4625	**2,56**	0.4948		
0,29	0.1141	**0,79**	0.2852	**1,29**	0.4015	**1,79**	0.4633	**2,58**	0.4951		
0,30	0.1179	**0,80**	0.2881	**1,30**	0.4032	**1,80**	0.4641	**2,60**	0.4953		
0,31	0.1217	**0,81**	0.2910	**1,31**	0.4049	**1,81**	0.4649	**2,62**	0.4956		
0,32	0.1255	**0,82**	0.2939	**1,32**	0.4066	**1,82**	0.4656	**2,64**	0.4959		
0,33	0.1293	**0,83**	0.2967	**1,33**	0.4082	**1,83**	0.4664	**2,66**	0.4961		
0,34	0.1331	**0,84**	0.2995	**1,34**	0.4099	**1,84**	0.4671	**2,68**	0.4963		
0,35	0.1368	**0,85**	0.3023	**1,35**	0.4115	**1,85**	0.4678	**2,70**	0.4965		
0,36	0.1406	**0,86**	0.3051	**1,36**	0.4131	**1,86**	0.4686	**2,72**	0.4967		
0,37	0.1443	**0,87**	0.3078	**1,37**	0.4147	**1,87**	0.4693	**2,74**	0.4969		
0,38	0.1480	**0,88**	0.3106	**1,38**	0.4162	**1,88**	0.4699	**2,76**	0.4971		
0,39	0.1517	**0,89**	0.3133	**1,39**	0.4177	**1,89**	0.4706	**2,78**	0.4973		
0,40	0.1554	**0,90**	0.3159	**1,40**	0.4192	**1,90**	0.4713	**2,80**	0.4974		
0,41	0.1591	**0,91**	0.3186	**1,41**	0.4207	**1,91**	0.4719	**2,82**	0.4976		
0,42	0.1628	**0,92**	0.3212	**1,42**	0.4222	**1,92**	0.4726	**2,84**	0.4977		
0,43	0.1664	**0,93**	0.3238	**1,43**	0.4236	**1,93**	0.4732	**2,86**	0.4979		
0,44	0.1700	**0,94**	0.3264	**1,44**	0.4251	**1,94**	0.4738	**2,88**	0.4980		
0,45	0.1736	**0,95**	0.3289	**1,45**	0.4265	**1,95**	0.4744	**2,90**	0.4981		
0,46	0.1772	**0,96**	0.3315	**1,46**	0.4279	**1,96**	0.4750	**2,92**	0.4982		
0,47	0.1808	**0,97**	0.3340	**1,47**	0.4292	**1,97**	0.4756	**2,94**	0.4984		
0,48	0.1844	**0,98**	0.3365	**1,48**	0.4306	**1,98**	0.4761	**2,96**	0.4985		
0,49	0.1879	**0,99**	0.3389	**1,49**	0.4319	**1,99**	0.4767	**2,98**	0.4986		

ANNEX 3

CRITICAL POINTS OF THE χ^2 DISTRIBUTION

K	The level of significance					
	0,01	0,025	0,05	0,95	0,975	0,89
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

ANNEX 4

TABLE OF VALUES

$$t_\gamma = t(\gamma, n)$$

n				n			
	0,95	0,99	0,999		0,95	0,99	0,999
5	2,78	4,60	8,61	20	2,093	2,861	2,861
6	2,57	4.03	6,86	25	2,064	2.797	2.797
7	2,45	3,71	5,96	30	2,045	2.756	2.756
8	2,37	3,50	5,41	35	2,032	2,720	2,720
9	2,31	3,36	5,04	40	2,023	2.708	2.708
10	2.26	3,25	4,78	45	2,016	2,692	2,692
11	2.23	3,17	4,59	50	2,009	2,679	2,679
12	2,20	3,11	4,44	60	2,001	2,662	2,662
13	2,18	3,06	4,32	70	1,996	2,649	2,649
14	2,16	3,01	4,22	80	1,001	2,640	2,640
15	2,15	2,98	4,14	90	1,987	2.633	2.633
16	2,13	2.95	4,07	100	1,984	2,627	2,627
17	2,12	2,92	4,02	120	1,980	2,617	2,617
18	2,11	2,90	3,97	∞	1,960	2.576	2.576
19	2,10	2,88	3,92				

TABLE OF VALUES

$$q = q(\gamma, n)$$

n \ γ	0,95	0,99	0,999	n \ γ	0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Навчальне видання

Насонова Світлана Сергіївна

ТЕОРІЯ ЙМОВІРНОСТЕЙ ТА МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник для студентів економічних спеціальностей

Англійською мовою

Редактори, оригінал-макет –
Коваленко-Марченкова Є. В., Самотуга А. В.

Підп. до друку 14.11.2022. Формат 60х84/16. Друк – цифровий. Гарнітура – Times.
Ум.-друк. арк. 8,86. Обл.-вид. арк. 9,50. Наклад – 35 прим. Зам. № 11/22-нп

Надруковано у Дніпропетровському державному університеті внутрішніх справ
49005, м. Дніпро, просп. Гагаріна, 26, rvv_vonr@dduvs.in.ua
Свідоцтво про внесення до Державного реєстру ДК № 6054 від 28.02.2018