MINISTRY OF SCIENCE AND HIGHER EDUCATION
OF THE RUSSIAN FEDERATION
Federal State Autonomous Educational Institution
of Higher Education
«SOUTHERN FEDERAL UNIVERSITY»

Institute of Mathematics, Mechanics and Computer Sciences

named after I. I. Vorovich

Probability Theory and Mathematical Statistics

Study guide

ISBN

Rostov-on-Don – 2026

UDC 519.2(075.8)

LBC 22.17 я73

Published by decision of the Department of Optimization Methods and Machine Learning of Southern Federal University (Record No. 6 of July 4, 2026).

Reviewers:

Natalia V. Danilova, Doctor of Physical and Mathematical Sciences, Associate Professor, Associate Professor at the Department of Optimization Methods and Machine Learning, Southern Federal University

Oleg E. Kudryavtsev, Doctor of Physical and Mathematical Sciences, Associate Professor, Head of the Department of Informatics and Information Customs Technologies, Rostov Branch of the Russian Customs Academy

Probability Theory and Mathematical Statistics: study guide / I. A. Zemliakova; Southern Federal University. – Rostov-on-Don, 2026. - 193 p.

This study guide is intended for undergraduate students enrolled in the degree program 02.03.02 Theoretical Computer Science and Information Technologies, as well as for students of other specializations taking the course “Probability Theory and Mathematical Statistics”. The guide provides a systematic exposition of the section "Probability Theory and Mathematical Statistics," including theoretical material, examples of problems with detailed solutions, including problems both for in-class collaborative work and for self-study. The aim of this study guide is to present some of the theoretical foundations of probability theory and mathematical statistics necessary for problem-solving, as well as to provide a large number of problems for both learning and independent solution.

UDC 519.2
LBC 22.17

TABLE OF CONTENTS

1. Set theory

Probability theory makes extensive use of set theory. Let us begin with the basic definitions and operations from set theory.

Definition 1 [9]. A **set** is a collection of objects, which we will call the elements of the set. We write $x \in A$, if an element x belongs to the set A. Let us denote $x \notin A$ if the element x does not belong to the set A.

Example 1. The number $x = 0.5$ belongs to the set of rational numbers and does not belong to the set of integers. So, $0.5 \in \mathrm{Q}$, $0.5 \notin \mathrm{Z}$.

If a set contains a finite number of elements, then this set can be specified by listing all of its elements: $A = \{a_1, a_2, \ldots, a_n\}$.

Example 2. Consider an experiment of rolling a six-sided die once. Then the set of possible outcomes is $\{1, 2, 3, 4, 5, 6\}$. In the experiment of tossing a coin once the set of possible outcomes is $\{H, T\}$, where H denotes heads and T denotes tails.

If the set A is **countably infinite,** that is, if it contains an infinite number of elements, then we will write it as follows: $A = \{a_1, a_2, \ldots\}$.

Example 3. The set of natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$ is countably infinite because its elements can be listed in a sequence, or equivalently, placed in one-to-one correspondence with the positive integers.

A set can also be specified by indicating a certain property that is satisfied by all elements of the given set.

Example 4. The set $A = \{x \mid x \in Z, x > 0, x < 10\}$. In this example the set A consists of elements x such that x belongs to the integers, x is greater than 0, and x is less than 10.

Definition 2 [9]. The **empty set** is a set that contains no elements. Let us denote this set by $\emptyset$. For example, the set of outcomes with a score greater than 6 when rolling a die once is an example of an empty set.

Definition 3 [9]. A set B is called a **subset** or a part of a set A if every element of set B is an element of set A. Let us write this as $B \subseteq A$ or $A \supseteq B$.

Two sets A and B are **equal** if and only if $A \subseteq B$ and $B \subseteq A$. In this case, we write $A = B$.

Definition 4 [16]. A universal set is a set that contains all elements under consideration in a particular problem. We denote it by Ω.

Next, let us recall the basic operations on sets.

Definition 5 [16]. The **complement** of a set A, with respect to the universal set Ω, is the set $A^c = \{x \in \Omega | x \notin A\}$.

That is, the complement of set A consists of those elements that are contained in the set Ω but are not contained in set A.

Example 5. For example, $A = \{x \mid x \in \mathrm{R}, x > 0\}$ is a set of all positive real numbers, the universal set R is the set of all real numbers, then the complement of set A is $A^c = \{x \mid x \in \mathrm{R}, x \leq 0\}$, so these are all non-positive numbers.

The graphical representation of the universal set as a rectangle allows us to visually represent operations on sets using Venn diagrams, where subsets of the universal set are depicted as circles located inside the rectangle. Thus, let us show the complement using the Venn diagram:

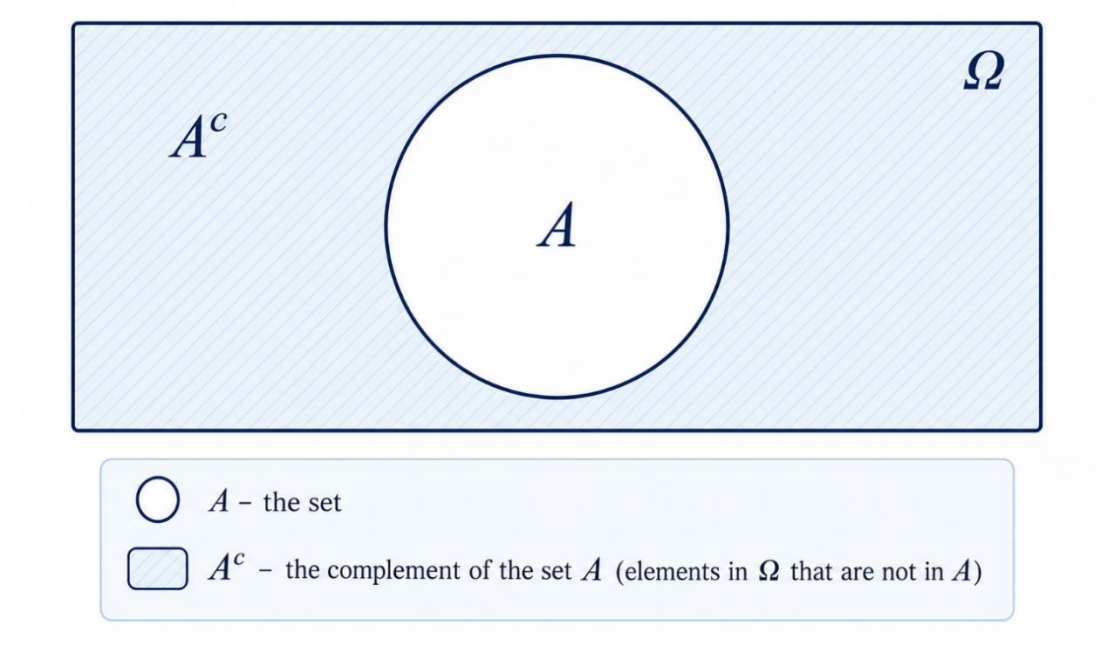


Figure1. The complement of a set A

Definition 6 [16]. The **union** of two sets A and B is the set of all elements that belong to A or B (or both), and is denoted by $A \cup B = \{x | x \in A \text{ or } x \in B\}$.

Example 6. If $A = \{1, 2, 3, 4\}$, $B = \{3, 4, 5, 6\}$, the union of these sets $A \cup B = \{1, 2, 3, 4, 5, 6\}$. The union is illustrated by the following Venn diagram:

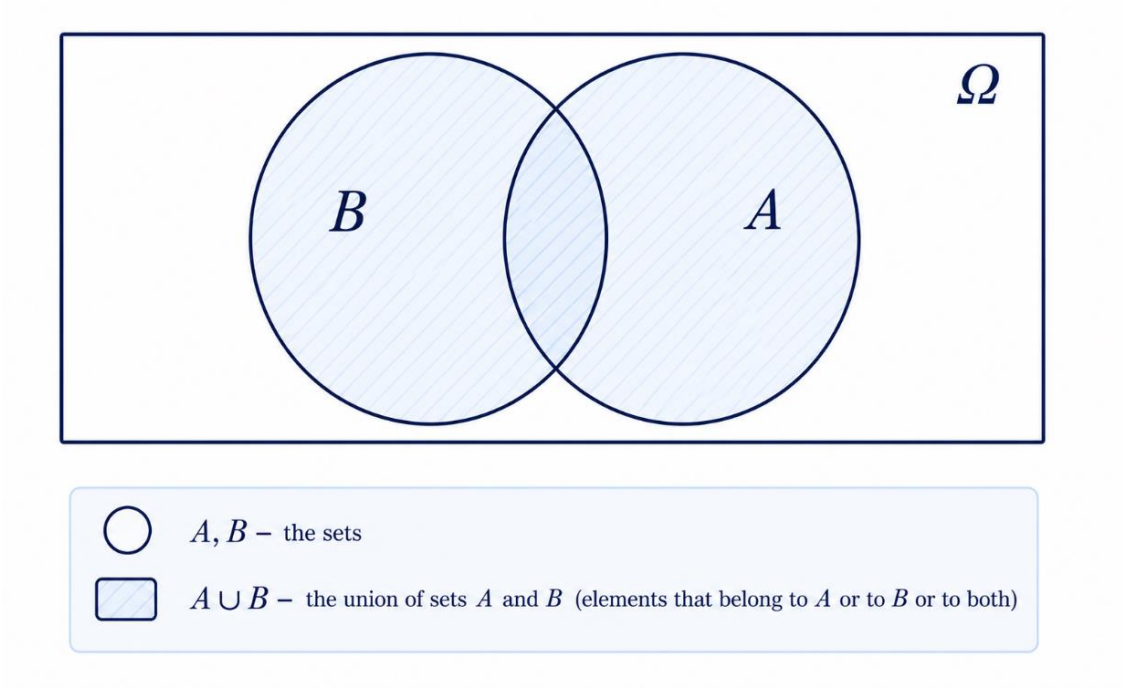


Figure 2. The union of sets A and B

Definition 7 [16]. The **intersection** of two sets A and B is the set of all elements that belong to both A and B, and is denoted by $A \cap B = \{x | x \in A \; and \; x \in B\}$.

Example 7. If $A = \{1, 2, 3, 4\}$, $B = \{3, 4, 5, 6\}$, the intersection of these sets $A \cap B = \{3, 4\}$. Let us show the intersection using the Venn diagram:

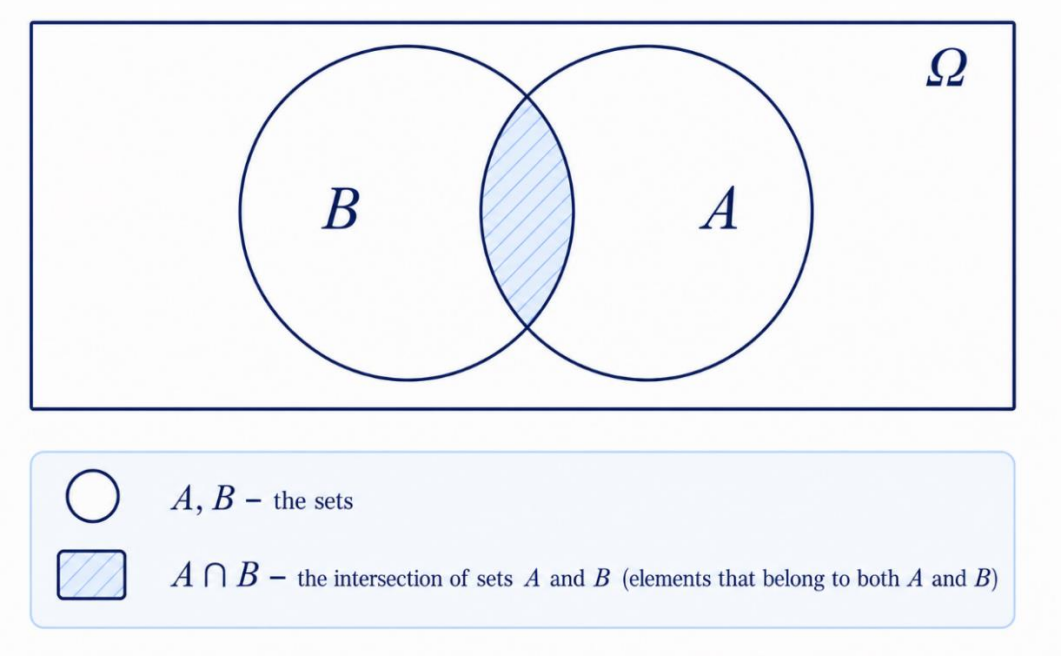


Figure 3. The intersection of sets A and B

Definition 8 [9]. The **difference** of two sets A and B is a set consisting of elements that belong to the set A and do not belong to the set B: $A \backslash B = \{x | x \in A \; and \; x \notin B\}$

Example 8. If $A = \{1, 2, 3, 4\}$, $B = \{3, 4, 5, 6\}$, the difference of these sets $A \setminus B = \{1, 2\}$. Let us show the difference using the Venn diagram:

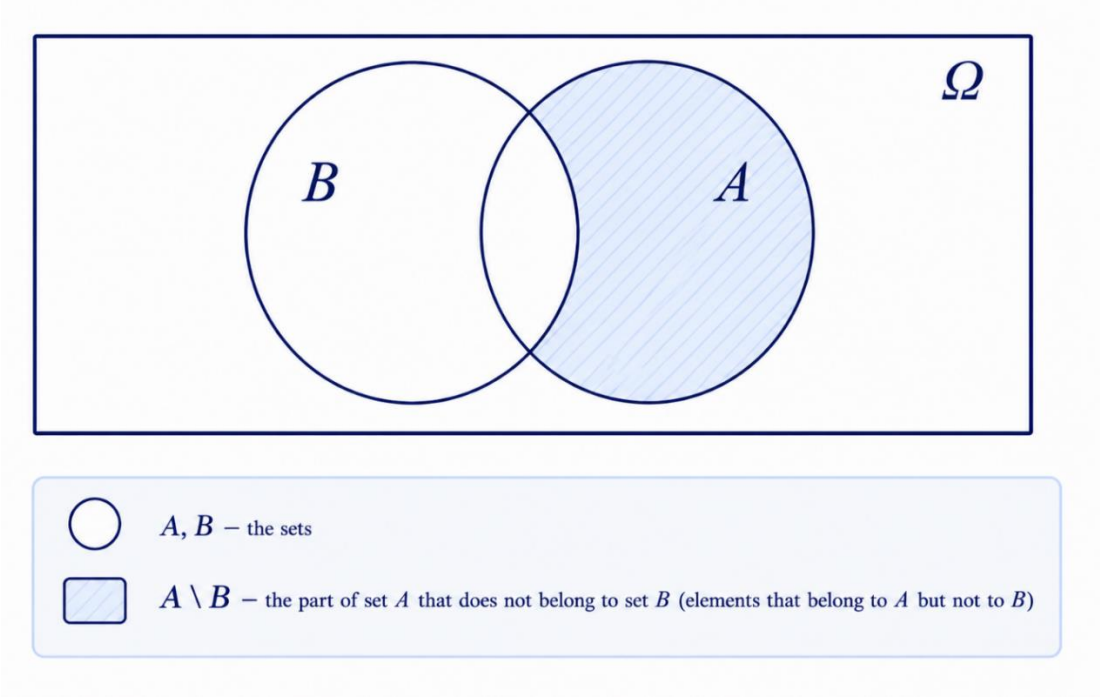


Figure 4. The difference of sets A and B

Finally, let us recall the basic laws of set algebra [9,16]:

1. Commutative laws:

$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A.$$

2. Associative laws:

$$(A \cup B) \cup C = A \cup (B \cup C),$$
$$(A \cap B) \cap C = A \cap (B \cap C).$$

3. Distributive laws**:**

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

4. Identity laws:

$$A \cup \emptyset = A,$$
$$A \cap \Omega = A,$$

5. Domination laws:

$$A \cup \Omega = \Omega,$$
$$A \cap \emptyset = \emptyset.$$

6. Idempotent laws:

$$A \cup A = A,$$
$$A \cap A = A.$$

7. Complement laws:

$$A \cup A^c = \Omega,$$
$$A \cap A^c = \emptyset.$$

8. Double complement law:

$$(A^c)^c = A.$$

9. Absorption laws:

$$A \cup (A \cap B) = A,$$
$$A \cap (A \cup B) = A.$$

10. De Morgan's laws:

$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c.$$

11. Difference as intersection with the complement:

$$A \setminus B = A \cap B^c.$$

Set theory provides the mathematical foundation for constructing a probabilistic model.

Let us now move on to solving problems related to set theory.

Problem 1. Prove the De Morgan's law: $(A \cup B)^c = A^c \cap B^c$ (formally by parsing boolean expressions and with Venn diagram).

Proof.

1) Let us start with Boolean algebra.

Let *x* be an arbitrary element of the universal set. Then $x \in (A \cup B)^c \Leftrightarrow \neg(x \in A \cup B) \Leftrightarrow \neg((x \in A) \vee (x \in B)) \Leftrightarrow (\neg(x \in A)) \wedge \big(\neg(x \in B)\big)$ ($here$ De Morgan's law for Boolean logic) $\Leftrightarrow (x \in A^c) \wedge (x \in B^c) \Leftrightarrow x \in A^c \cap B^c$. So, for every element *x*, $(A \cup B)^c = A^c \cap B^c$. Hence, the first De Morgan's law is proved.

2) Proof using a Venn diagram.

The proof using a Venn diagram involves sequentially depicting all the operations contained in the expression. Let us show the left-hand side $(A \cup B)^c$:

1. $A \cup B$:

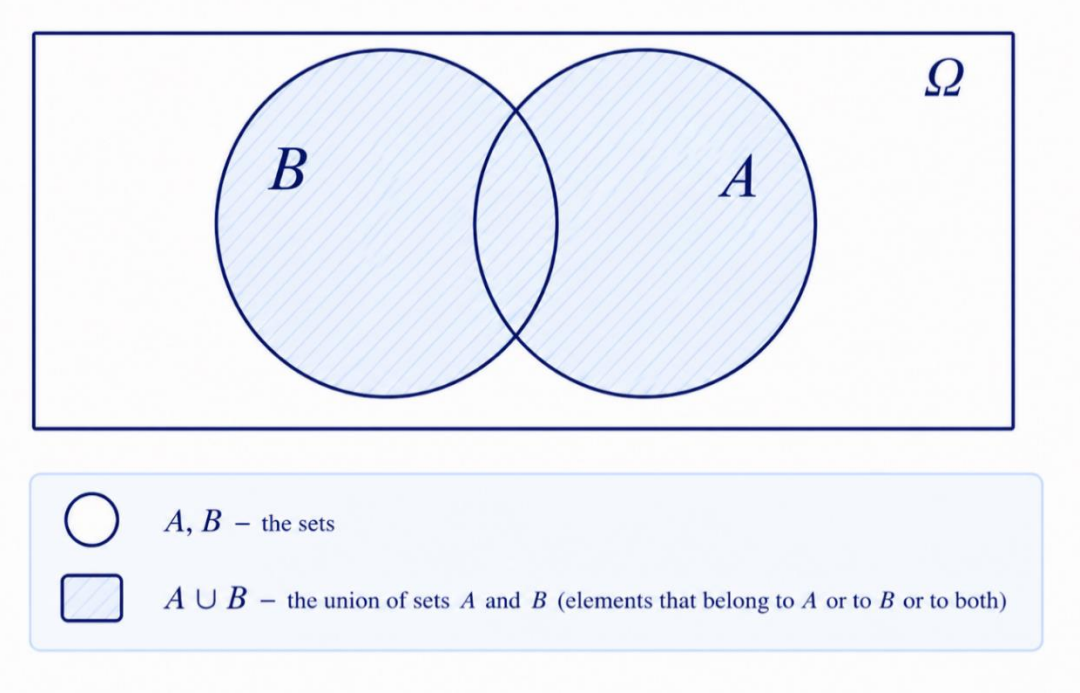


Figure 5. $A \cup B$

2. $(A \cup B)^c$:

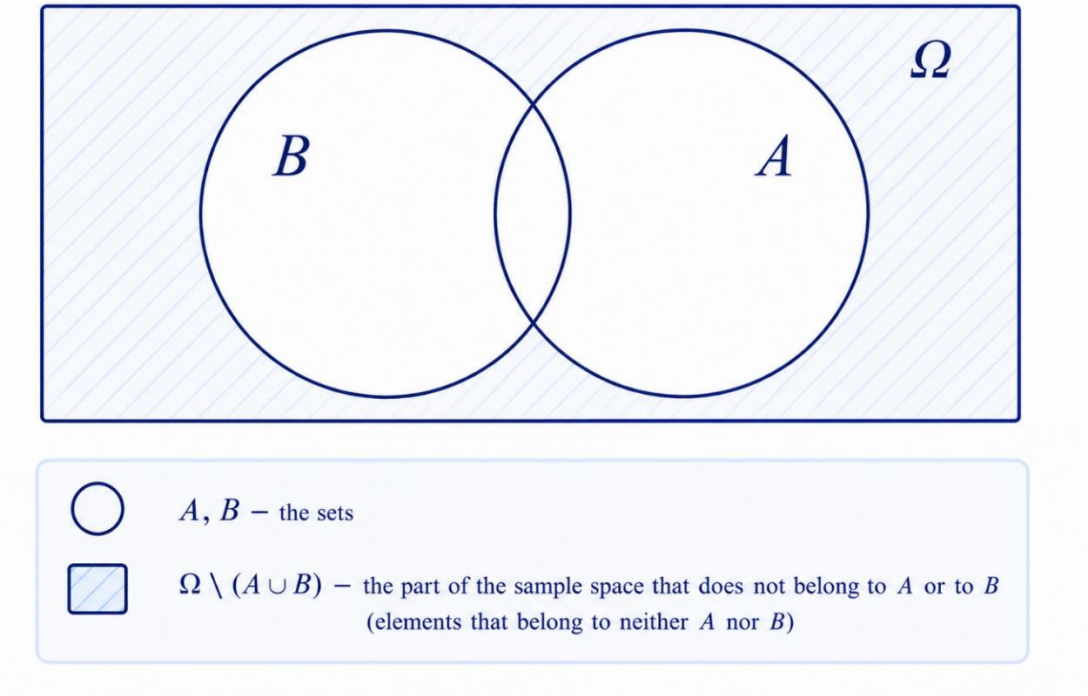


Figure 6. $(A \cup B)^c$

This is the result for the left-hand side**.** Let us show the right-hand side $A^c \cap B^c$:

1. A^C:

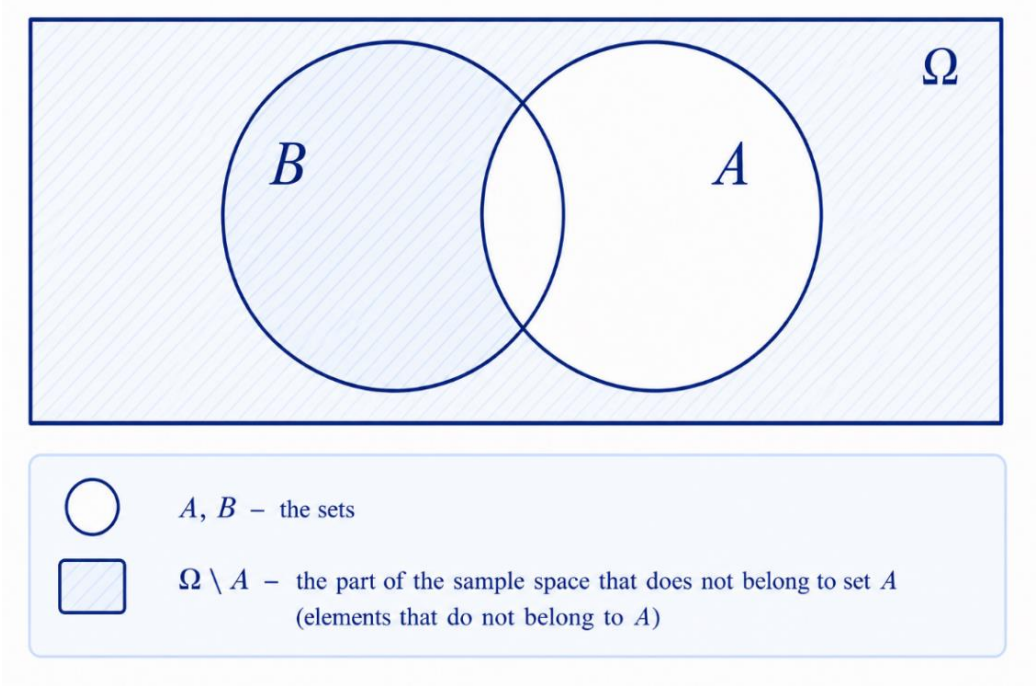


Figure 7. A^C

2. B^C:

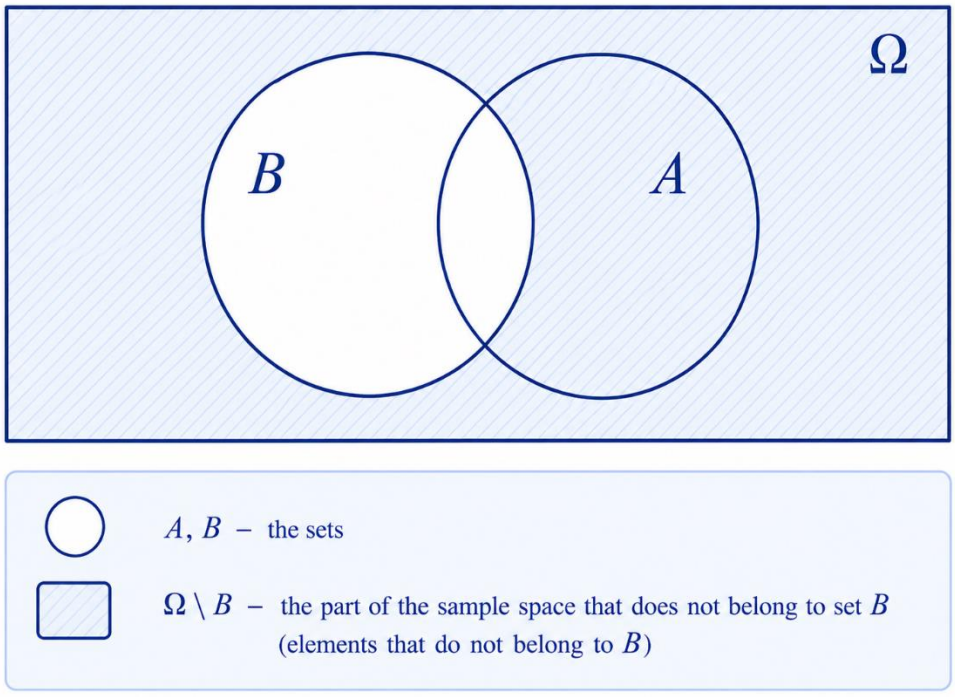


Figure 8. B^C

3. Let us combine the two previous diagrams to see the desired intersection $A^c \cap B^c$:

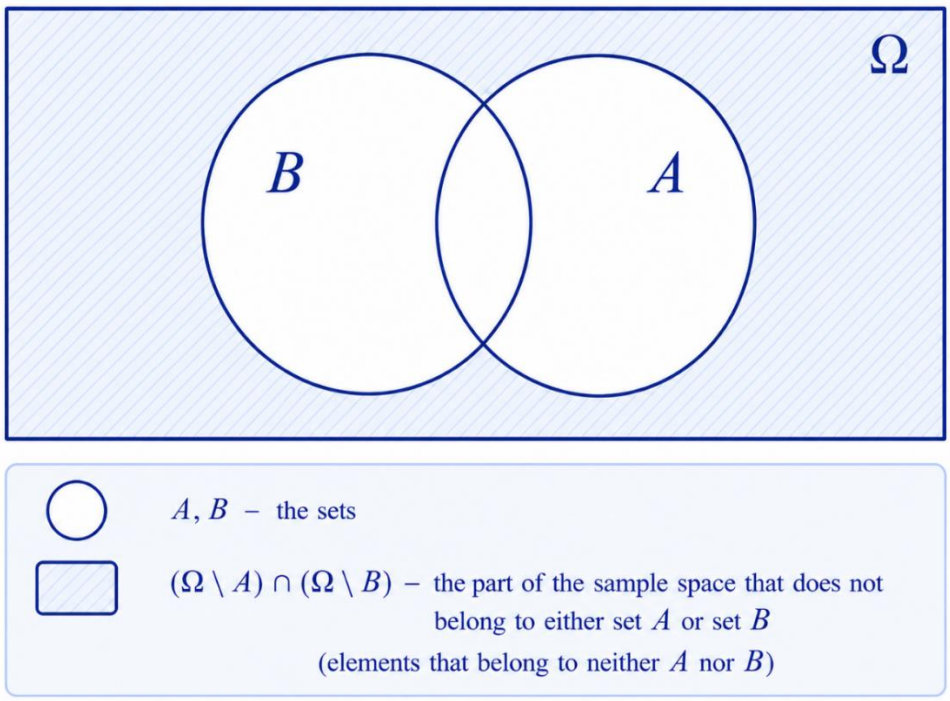


Figure 9. $A^c \cap B^c$

The intersection of the shadings is the desired set.

Note that the regions corresponding to the left-hand side and the right-hand side coincide. Hence, the first De Morgan's law is proved.

Problem 2. For a six-sided die, consider the events

A = {the number of points is even},

B = {the number of points is greater than 3}.

Find the sets represented by the left-hand side and the right-hand side of the equality $(A \cup B)^c = A^c \cap B^c$, and verify that they are equal.

Solution.

Let's define the set corresponding to the left-hand side, the right-hand side, and compare them. Let us start with the left-hand side. Let's start with sets A and B.

A = {the number of points is even}, so A = {2, 4, 6}; B = {the number of points is greater than 3}, so B = {4, 5, 6}. Then $A \cup B$ = {2, 4, 5, 6} and $(A \cup B)^C$ = {1, 3}. This is the left-hand side. Let's move on to the right-hand side: A^C = {1, 3, 5}, B^C = {1, 2, 3}, $A^C \cap B^C$ = {1, 3}. Thus, the result obtained for the right-hand side coincides with the result obtained for the left-hand side. Thus, the set on the left-hand side is equal to the set on the right-hand side.

Problem 3. Three coins are tossed and a four-sided die is rolled. Describe a suitable sample space and determine its cardinality. Describe the set of outcomes of the following events:

1. A = {exactly two Heads are obtained and the die shows an even number};
2. B = {the die shows a number greater than 2};

3. C = {exactly one Head is obtained and the die shows an odd number};

Describe the sets of outcomes for the following events:

1. A and B;

2. only in B;

3. B and C;

4. A but not B.

Solution:

Let H denote *Heads* and T denote *Tails*.

The suitable sample space Ω is

Ω = {(HHH,1), (HHH,2), (HHH,3), (HHH,4), (HHT,1), (HHT,2), (HHT,3), (HHT,4), (HTH,1), (HTH,2), (HTH,3), (HTH,4), (HTT,1), (HTT,2), (HTT,3), (HTT,4), (THH,1), (THH,2), (THH,3), (THH,4), (THT,1), (THT,2), (THT,3), (THT,4), (TTH,1), (TTH,2), (TTH,3), (TTH,4), (TTT,1), (TTT,2), (TTT,3), (TTT,4)}.

Number of outcomes**:** 32.

The elements of Ω do not need to be listed individually in order to determine the cardinality of the sample space. Their number can be calculated using the *Multiplication Principle* from combinatorics (it has certainly been studied before; for more details, see *Section 3*): if action 1 can be performed in n_1 ways, action 2 in n_2 ways, ..., action k in n_k ways, then the sequence of all actions can be performed in $n_1 \cdot n_2 \cdot \ldots \cdot n_k$ ways.

So, by the Multiplication Principle, each of the three coins has 2 possible outcomes and the four-sided die has 4 possible outcomes. Therefore, the sample space contains

$2^3 \cdot 4 = 32$ elements.

Let us move on to describing sets:

1. A = {exactly two Heads are obtained and the die shows an even number}

A = {(HHT,2), (HHT,4), (HTH,2), (HTH,4), (THH,2), (THH,4)}

2. B = {the die shows a number greater than 2};

B = {(HHH,3), (HHH,4), (HHT,3), (HHT,4), (HTH,3), (HTH,4), (HTT,3), (HTT,4), (THH,3), (THH,4), (THT,3), (THT,4), (TTH,3), (TTH,4), (TTT,3), (TTT,4)}

3. C = {exactly one Head is obtained and the die shows an odd number};

C = {(HTT,1), (HTT,3), (THT,1), (THT,3), (TTH,1), (TTH,3)}

Let us move on to describing the sets of outcomes for the following events:

1. A and B: A ∩ B = {(HHT,4), (HTH,4), (THH,4)}
2. only in B: $B \setminus (A \cup C)$ = {(HHH,3), (HHH,4), (HHT,3), (HTH,3), (HTT,4), (THH,3), (THT,4), (TTH,4), (TTT,3), (TTT,4)}
3. B and C: $B \cap C$ = {(HTT,3), (THT,3), (TTH,3)}

4. A but not B: $A \setminus B = \{(HHT,2), (HTH,2), (THH,2)\}$

Problem 4. Give an example of sets A, B, and C such that $A \cap B = A \cap C$ but $B \neq C$.

Solution.

For example, $A = \{1,2\}$, $B = \{1,3\}$, $C = \{1,4\}$.

$A \cap B = \{1\}$, $A \cap C = \{1\}$, $B \neq C$.

Problem 5. There are 50 programmers in a software company who specialize in Python or Java or do not specialize in either language. Some programmers specialize in both languages. It is known that 30 programmers specialize in Python, 25 specialize in Java, and 10 programmers specialize in both languages. Determine the number of programmers who:

- specialize only in Python;
- do not specialize in Java;
- specialize in Python or Java;
- do not specialize in any of these programming languages.

Solution:

Let:

P = {programmers who specialize in Python}

J = {programmers who specialize in Java}

The total number of programmers is: $|\Omega| = 50$, $|P| = 30$, $|J| = 25$, $|P \cap J| = 10$ (10 programmers specialize in both languages).

- specialize only in Python. Let's show them using the Venn diagram:

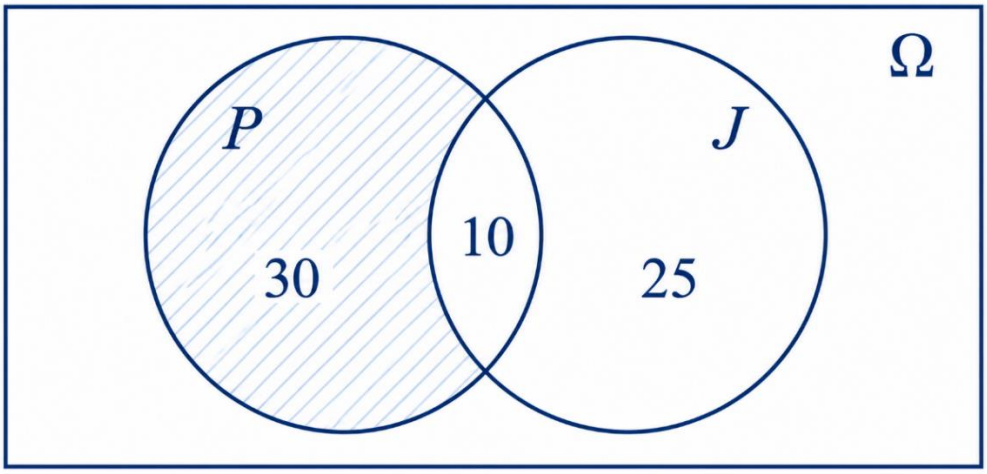


Figure 10. Programmers who specialize only in Python

As we can see, this is a difference $P \setminus J$. The Python circle contains 30 programmers, including those who also know Java. Therefore, the Python-only part is: $|P \setminus J| = 30 - 10 = 20$.

- do not specialize in Java. Let's show them using the Venn diagram:

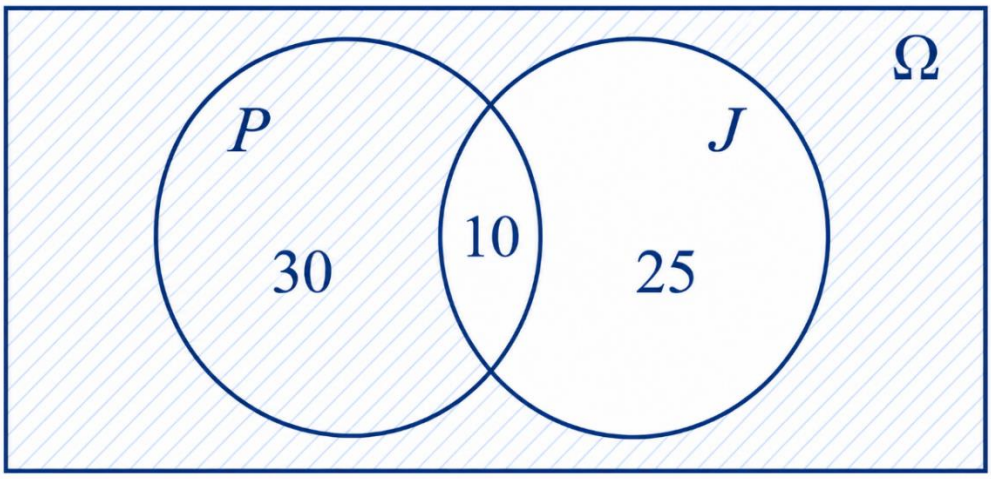


Figure 11. Programmers who do not specialize in Java

As we can see, this is a complement of a set *J*. First, let us count how many programmers specialize in neither language: $|(P \cup J)^C| = 50 - (30 + 25 - 10) = 5$. The Python-only part is: $|P \setminus J| = 30 - 10 = 20$ (from the previous section). Thus, based on these considerations, we obtain: $|J^C| = 50 - 20 - 5 = 25$.

- specialize in Python or Java. Let's show them using the Venn diagram:

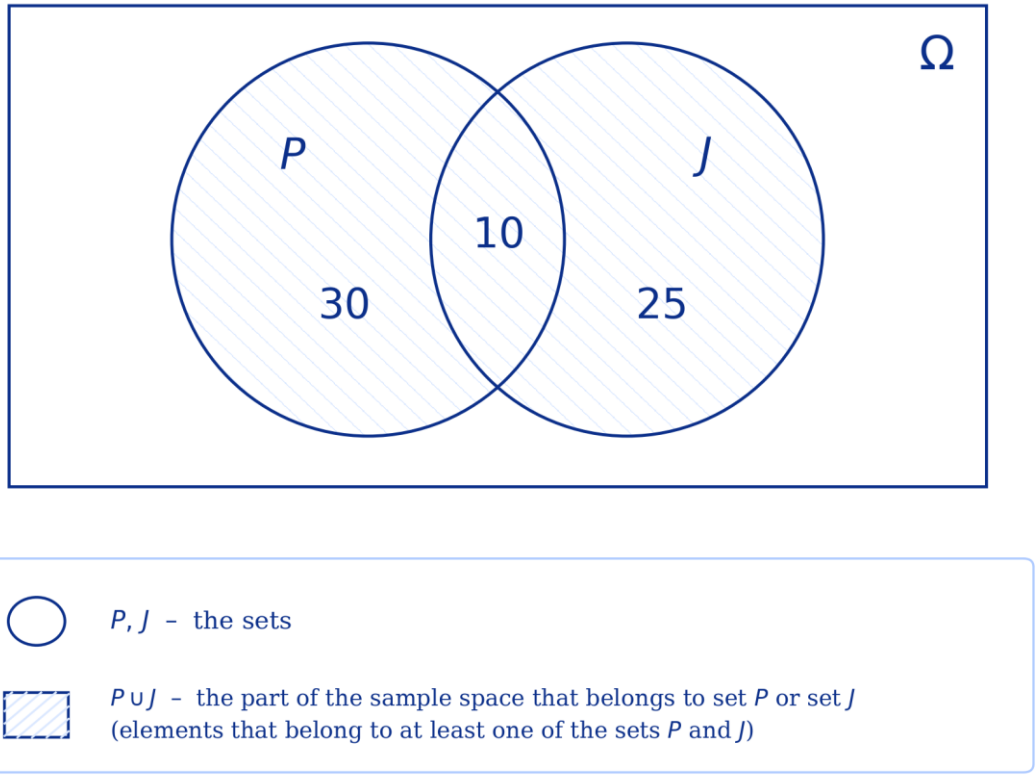


Figure 12. Programmers who specialize in Python or Java

As we can see, this is a union $P \cup J$. The number of programmers who specialize in Python or Java is $|P \cup J| = 30 + 25 - 10 = 45$.

- do not specialize in any of these programming languages. The necessary reasoning was presented in the second section, from which $|(P \cup J)^C| = 50 - (30 + 25 - 10) = 5$.

Answer:

- 20 programmers specialize only in Python;
- 25 do not specialize in Java;
- 45 specialize in Python or Java;
- 5 specialize in neither language.

Tasks for independent work (homework).

1. Prove De Morgan's law: $(A \cap B)^c = A^c \cup B^c$ (formally by parsing boolean expressions and with Venn diagram).
2. For a six-sided die, consider the events A = {the number of points is even}, B = {the number of points is greater than 3}. Find and compare sets on both sides of the equality: $(A \cap B)^c = A^c \cup B^c$.
3. A fair coin is tossed twice. Consider the events
   A = {at least one Head is obtained};
   B = {the second toss is Heads}.

Find and compare the sets on both sides of the equality

$$(A \cup B)^c = A^c \cap B^c.$$

4. Two coins and a six-sided die are tossed. Describe a suitable sample space of all possible outcomes and determine the number of elements in it.

Describe the set of outcomes of the following events:

- A = {the first coin shows Heads, and the die shows an odd number};
- B = {at least one Head is obtained, and the die shows a number greater than 4};
- C = {both coins show the same result, and the die shows an even number}.

Describe the sets of outcomes for the following events:

- A and B;
- only in B;
- B and C;
- A but not B.

5. Give examples of sets for which $A \cup B = A \cup C$, but $B \neq C$.

6. There are 70 students in a university group who study Mathematics or Statistics or do not study either subject. Some students study both subjects at the same time. It is known that 40 students study Mathematics; 30 students study Statistics; 15 students study both Mathematics and Statistics.

Determine the number of students who:

- study only Mathematics;

- do not study Statistics;
- study Mathematics or Statistics;
- do not study any of these subjects.

7. Let $\Omega = \{1, 2, 3, \ldots, 30\}$. Consider the following subsets of Ω:
   - $A = \{x \in \Omega \mid x \text{ is divisible by } 2\}$,
   - $B = \{x \in \Omega \mid x \text{ is divisible by } 3\}$,
   - $C = \{x \in \Omega \mid x \text{ is a prime number}\}$.

   Find the following sets:

   a) $A \cap B$;

   b) $(A \cup B)^C$;

   c) $A \cap C$;

   d) $(A \setminus B) \cup (B \setminus A)$;

   e) $(A \cap B^C) \cup (A^C \cap B)$;

   f) $(A \cup C)^C$ and $A^C \cap C^C$.

Compare the sets obtained in parts d) and e), and verify De Morgan's law in part f).

8. A technology company employs 120 artificial intelligence specialists. Among them:
   - 78 specialists use Python;
   - 54 specialists use R;
   - 36 specialists use Julia;
   - 32 specialists use both Python and R;
   - 20 specialists use both Python and Julia;
   - 18 specialists use both R and Julia;
   - 12 specialists use all three programming languages.

   Determine the number of specialists who:

   a) use only Python;

   b) use only R;

   c) use only Julia;

   d) use exactly two of the three programming languages;

   e) use at least one of the three programming languages;

   f) use none of these programming languages;

   g) use Python but use neither R nor Julia;

   h) do not use Python.

   Represent the information using a Venn diagram.

9. Let Ω denote the set of messages analyzed by an AI-based security system. Define the events:
   - A = {messages flagged by the anomaly detector};

- B = {messages flagged by the spam detector};
- C = {messages sent for manual review}.

Express each of the following events using set notation:

a) messages flagged by both detectors;

b) messages flagged by exactly one of the two detectors;

c) messages flagged by neither detector;

d) messages flagged by at least one detector and sent for manual review;

e) messages flagged by the anomaly detector but neither flagged by the spam detector nor sent for manual review;

f) messages sent for manual review but not flagged by either detector;

g) messages belonging to exactly one of the three sets A, B, and C;

h) messages belonging to exactly two of the three sets.

2. Probabilistic models

Definition 9 [1]. A **probabilistic model** is a mathematical description of a random experiment that includes all possible outcomes, the rules by which we identify events, and a method for calculating their probabilities.

The possible results of a random experiment are called **outcomes**. The set of all possible outcomes of an experiment is denoted by Ω. A **probability law** assigns a nonnegative number $P(A)$ to each event A, where A is a set of possible outcomes. The value $P(A)$ describes the likelihood of the event A based on available information or assumptions. A valid probability law must satisfy several basic properties, which will be discussed further.

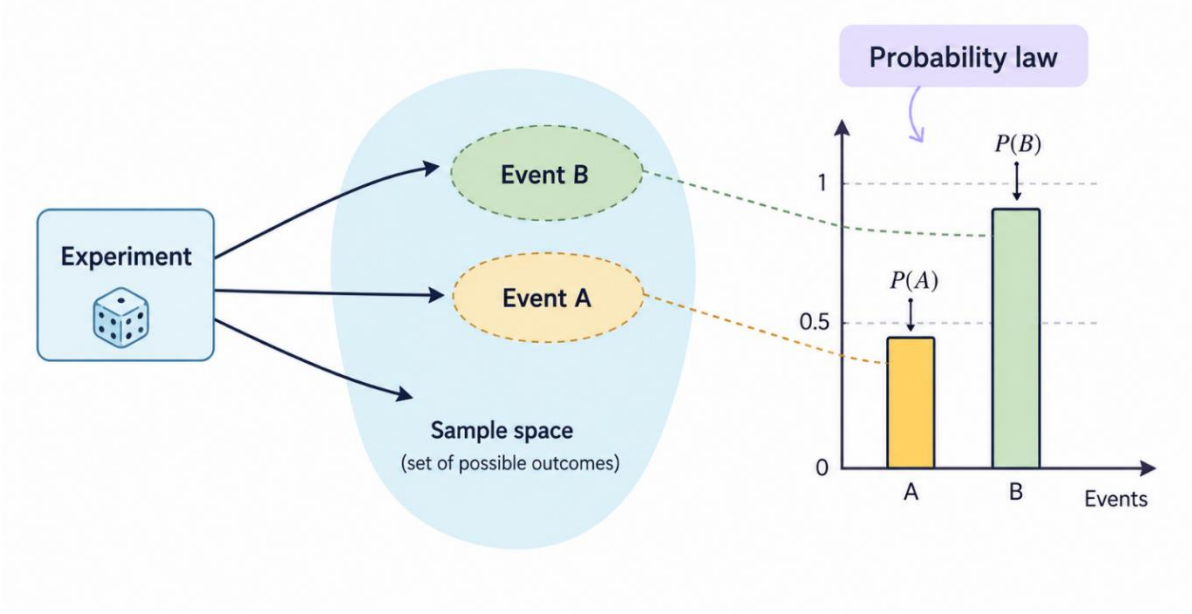


Figure 13. A probabilistic model

Definition 10 [1]. An **event** is any subset of the sample space. The event occurs if the outcome of the experiment belongs to this subset.

Example 9. Consider a random experiment in which a six-sided die is rolled once. The result of the experiment is the number that appears on the upper face of the die. The possible results are: 1, 2, 3, 4, 5, 6. These results are called **outcomes**. The set of all possible outcomes forms the **sample space** $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Thus, Ω contains all possible results of the random experiment. A **probabilistic model** for this experiment consists of the sample space Ω, the collection of possible events and the probability law that assigns probabilities to these events. Let's consider some events.

Example 10. Consider the following events:

A = {an even number is obtained} = {2, 4, 6},

B = {a number greater than 4 is obtained} = {5, 6}.

The outcomes in the sample space must be distinct and mutually exclusive, and exactly one outcome must occur whenever the experiment is performed. The sample space must include every possible outcome while containing only the information relevant to the probabilistic model.

Definition 11 [1]. The **probability law** assigns to every event A a number $P(A)$, called the probability of A, satisfying the following axioms:

Probability Axioms [1, 17].

i. Nonnegativity

The probability of any event A cannot be negative. $P(A) \geq 0$ for every event A.

ii. Additivity

If A and B are two disjoint events (events that cannot occur simultaneously), then the probability of their union is equal to the sum of their individual probabilities: $P(A \cup B) = P(A) + P(B)$. More generally, if the sample space contains an infinite sequence of disjoint events: $A_1, A_2, \ldots$ then the probability of their union is equal to the sum of their probabilities: $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$.

iii. Normalization

The probability of the entire sample space Ω is equal to one: $P(\Omega) = 1$**.** This means that the outcome of a random experiment always belongs to the sample space.

Basic Properties of Probability [1,17].

1. The impossible event, or equivalently the empty set, has probability zero: $P(\emptyset) = 0$.
2. **(Complement Rule)** For any event A, the probability of its complement is: $P(A^C) = 1 - P(A)$.

Example 11. A software company has 50 programmers. Among them, 8 programmers are currently working on a critical project. One programmer is selected at random. Find the probability that the selected programmer is not working on the critical project.

Solution.

Let A be the event that the selected programmer is working on the critical project. Then A^C is the event that the selected programmer is not working on the critical project. The probability of event A is $P(A) = 8/50$. According to the Complement Rule: $P(A^c) = 1 - P(A) = 1 - 8/50 = 42/50 = 21/25$.

Answer: 21/25.

3. For any event A, the probability of this event lies between zero and one $0 \leq P(A) \leq 1$.
4. If $A \subseteq B$, then $P(A) \leq P(B)$.
5. For any two events A and B $P(A \setminus B) = P(A) - P(A \cap B)$.
6. (Addition Rule) For any two events A and B $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

We now illustrate how a probability law can be constructed for a simple finite experiment.

Example 12. Consider an experiment involving a single roll of a four-sided die. There are four possible outcomes: 1, 2, 3, and 4. The sample space is $\Omega = \{1, 2, 3, 4\}$. The possible events include $\{1, 2, 3, 4\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\emptyset$. If the die is fair, all possible outcomes are considered equally likely. Therefore, each elementary outcome is assigned the same probability: $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = 0.25$.

According to the additivity axiom, the probability of the entire sample space is the sum of the probabilities of all mutually exclusive outcomes: $P(\{1,2,3,4\}) = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) = 1$.

This result agrees with the normalization axiom, which states that the probability of the sample space must be equal to one. Therefore, the probability law for this experiment is: $P(\{1,2,3,4\}) = 1$, $P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = 0.25$, $P(\emptyset) = 0$.

This probability law satisfies all three probability axioms: nonnegativity, additivity, and normalization.

Example 13. Consider an experiment in which two fair six-sided dice are rolled once. The outcome is an ordered pair (i, j), where i is the number shown by the first die and j is the number shown by the second die. The sample space is: $\Omega = \{(1,1), (1,2), ..., (6,5), (6,6)\}$. The sample space contains 36 possible outcomes. We assume that each possible outcome has the same probability $P(\{(i,j)\}) = 1/36$ for every possible outcome in Ω.

Let us construct a probability law that satisfies the three probability axioms. Consider, as an example, the event A = {the sum of the two dice is equal to 7}.

The event A consists of the following outcomes: $A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$.

Using the additivity axiom, the probability of event A is the sum of the probabilities of its individual outcomes:

$P(A) = P(\{(1,6)\}) + P(\{(2,5)\}) + P(\{(3,4)\}) + P(\{(4,3)\}) + P(\{(5,2)\}) + P(\{(6,1)\}) = 1/36 + 1/36 + 1/36 + 1/36 + 1/36 + 1/36 = 6/36 = 1/6$

Similarly, the probability of any event is equal to the number of possible outcomes contained in the event divided by the total number of outcomes in the sample space.

Thus, let's formulate the **Discrete Uniform Probability Law.**

Definition 12 [1]. A **discrete uniform probability law** is a probability model in which all possible outcomes of a finite sample space are equally likely.

The probability of any event A is determined by the number of outcomes contained in this event:

$$\boldsymbol{P(A) = \frac{\text{number of outcomes in A}}{\text{number of outcomes in } \Omega}}$$

or equivalently:

$$\boldsymbol{P(A) = \frac{|A|}{|\Omega|}}$$

If the sample space $\boldsymbol{\Omega}$ contains $\boldsymbol{n}$ possible outcomes, event A contains $\boldsymbol{m}$ possible outcomes, thus:

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

Let us consider an example.

Example 14. Two cards are drawn with replacement from a set of four cards numbered 1, 2, 3, and 4. The outcome is an ordered pair (i, j), where i is the number on the first card drawn and j is the number on the second card drawn. Since each draw has four possible results, the sample space contains $4 \cdot 4 = 16$ equally likely outcomes.

We assume that each card is equally likely to be drawn. Therefore, there are sixteen possible outcomes: $\Omega = \{(1,1), (1,2), (1,3), (1,4), (2,1), ..., (4,4)\}$

Each possible outcome has the same probability: $P(\{(i,j)\}) = 1/16$, $i = 1,..,4$, $j = 1,..,4$.

To calculate the probability of an event, we count the number of outcomes belonging to this event and divide by the total number of possible outcomes.

For example:

$P(\{\text{the sum of the two numbers is even}\}) = 8/16 = ½$;

$P(\{\text{the sum of the two numbers is greater than 5}\}) = 6/16 = 3/8$;

$P(\{\text{the first number is equal to the second}\}) = 4/16 = ¼$;

$P(\{\text{the first number is smaller than the second}\}) = 6/16 = 3/8$;

$P(\{\text{at least one drawn number is equal to 1}\}) = 7/16$.

Let us now proceed to solving problems.

Problem 6.

A three-digit number is randomly chosen from all possible three-digit numbers. Find the probability that the chosen number will be:

a) a number whose last digit is even;

b) a three-digit number with all digits different.

Solution.

a) The sample space is $\Omega = \{100, 101, 102, ..., 999\}$, $n = |\Omega| = 900$. Let A be the event that the selected number has an even last digit. The last digit can be: 0, 2, 4, 6, 8. For each possible last digit, there are 90 possible choices of the first two digits (in the first position, there can be any digit from 1 to 9 (9 options), and in the second position, any digit from 0 to 9 (10 options)).

Therefore, the number of favorable outcomes is:

$m = |A| = 5 \cdot 90 = 450$

Hence:

$P(A) = 450/900 = 1/2$

Answer: ½

b) The sample space is $\Omega = \{100, 101, 102, ..., 999\}$, $n = |\Omega| = 900$. Let's consider an event *B:* three-digit numbers with all digits different. The first digit cannot be zero, so it can be chosen in 9 ways (1, 2, 3, ..., 9). The second digit can be any digit except the first one, so we have 9 ways. The third digit can be any digit except the first two digits, so 8 ways.

Therefore, the number of three-digit numbers with different digits is:

$m = |B| = 9 \cdot 9 \cdot 8 = 648$

Hence:

$P(B) = 648/900 = 18/25$

Answer: 18/25

Problem 7. A cube, all sides of which are painted, is sawn into a thousand cubes of the same size, which are then thoroughly mixed. Find the probability that a randomly drawn cube has painted faces:

a) one;

b) two;

c) three.

Solution.

A cube is painted on all six faces and then divided into 1000 smaller cubes of the same size, so $n = |\Omega| = 1000$.

We assume that each small cube is equally likely to be selected.

To find the required probabilities, we count the number of small cubes with one, two, or three painted faces.

a) cubes with one colored face

Cubes with exactly one painted face are located in the middle of each large face, excluding edges and corners.

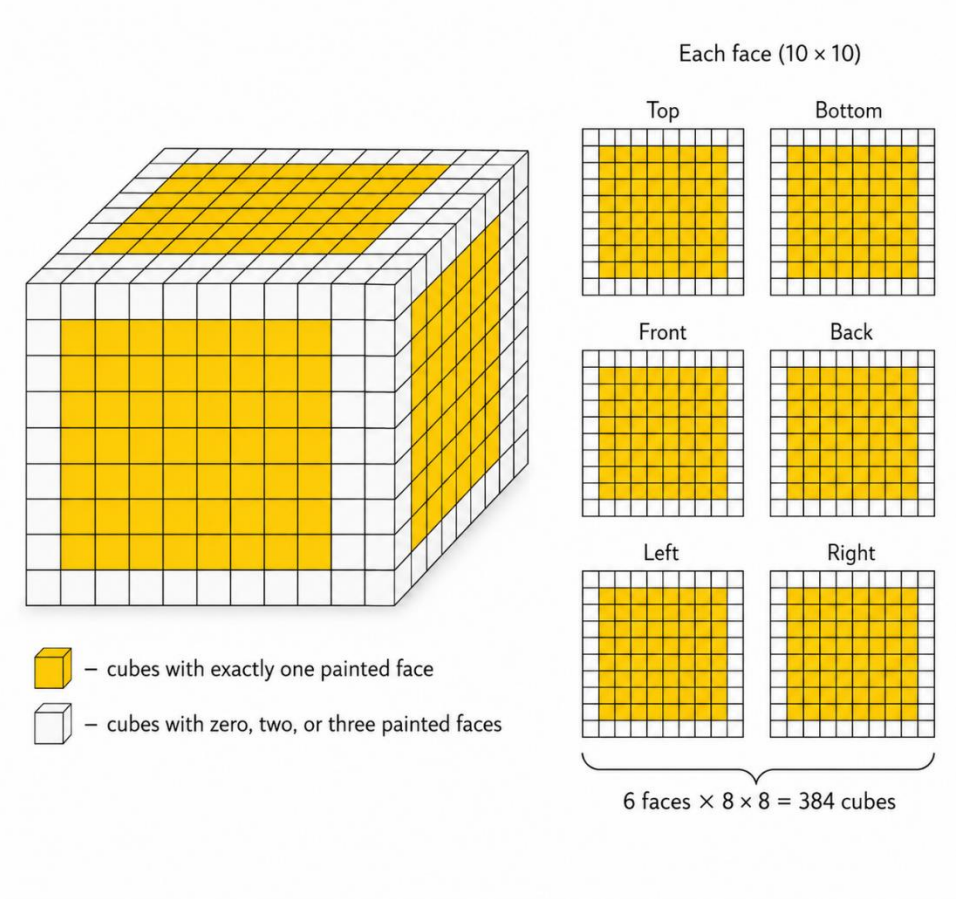


Figure 14. Cubes with exactly one painted face

Each face contains 64 such cubes. Since the original cube has 6 faces, $6 \cdot 64 = 384$
small cubes have exactly one colored face. Therefore:

$$P(\{\text{one colored face}\}) = 384/1000 = 0.384$$

b) cubes with two painted faces

Cubes with exactly two painted faces are located along the edges, excluding the corner cubes.

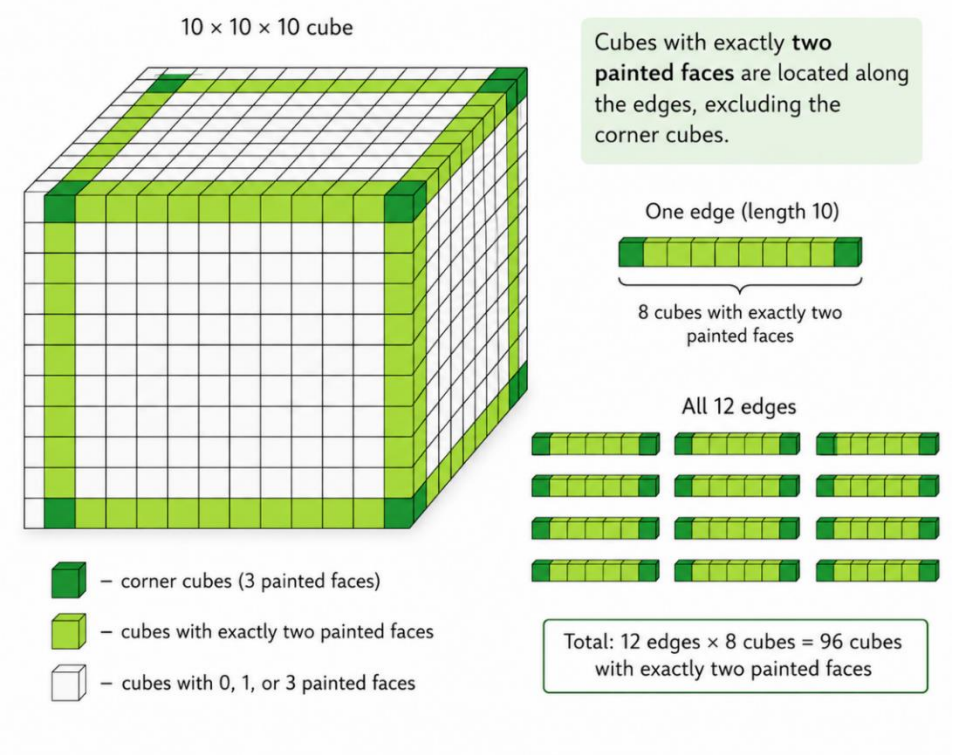


Figure 15. Cubes with exactly two painted faces

Each edge contains 8 such cubes, because the two corner cubes are excluded. The cube has 12 edges. Therefore $12 \cdot 8 = 96$ small cubes have exactly two painted faces.

Hence:

$$P(\{\text{two painted faces}\}) = 96/1000 = 0.096$$

c) cubes with three painted faces

Cubes with three painted faces are located at the corners of the original cube.

A cube has 8 corners. Therefore 8 small cubes have three painted faces.

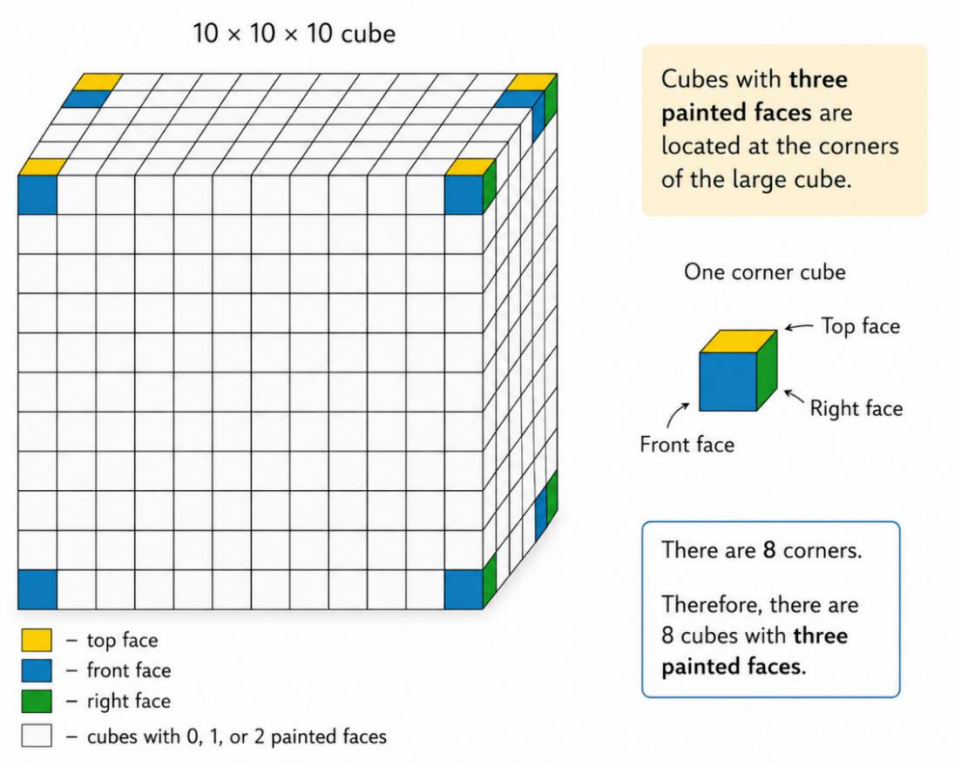


Figure 16. Cubes with three painted faces

Hence:

$$P(\{\text{three painted faces}\}) = 8/1000 = 0.008$$

Answer:

a) 0.384

b) 0.096

c) 0.008

Problem 8. Two fair dice are rolled. Find the probabilities of the following events:

- the sum of the numbers rolled is equal to 9;
- the sum of the numbers rolled is equal to 10, and the difference between the numbers on the dice is 2;
- the sum of the numbers rolled is equal to 10, if it is known that the absolute difference between the numbers on the dice is 2;
- the sum of the numbers rolled is equal to 6, and the product of the numbers on the dice is 8.

Solution.

Consider an experiment in which two fair dice are rolled. The sample space consists of all ordered pairs $\Omega = \{(1,1), (1,2), ..., (6,6)\}$, $n = |\Omega| = 6 \cdot 6 = 36$ possible outcomes.

a) The sum of the rolled points is equal to 9. Let A = {the sum of the numbers on the dice is 9}. The possible outcomes are A = {(3,6), (4,5), (5,4), (6,3)}. The number of favorable outcomes is $m = |A| = 4$.

$P(A) = 4/36 = 1/9$

b) The sum of the rolled points is equal to 10, and the difference between the numbers on the dice is 2. Let B = {the sum is 10 and the difference is 2}; B = {(4,6), (6,4)}. The number of favorable outcomes is $m = |B| = 2$.

$P(B) = 2/36 = 1/18$

c) The sum is equal to 10, if it is known that the difference between the numbers on the dice is 2. In this problem, we must select favorable outcomes from the set of all outcomes that satisfy the condition. Therefore, let us start with the outcomes that satisfy the condition.

C = {the difference between the numbers on the dice is 2}. The outcomes satisfying this condition are C = {(1,3), (2,4), (3,1), (3,5), (4,2), (4,6), (5,3), (6,4)}, $n = |C| = 8$.

Among these outcomes, the ones where the sum is 10 are {(4,6), (6,4)}. The number of favorable outcomes is equal to 2 ($m = 2$).

Therefore, the required probability is 2/8 = 1/4.

d) The sum of the rolled points is equal to 6, and the product of the numbers on the dice is 8. Let D = {the sum is 6 and the product is 8}; D = {(2,4), (4,2)}; $|D| = 2$.

$P(D) = 2/36 = 1/18$.

Answer:

a) 1/9

b) 1/18

c) 1/4

d) 1/18

Problem 9. A software company has 20 programming tasks in its database, 12 of which are marked as completed. A developer randomly selects 4 tasks for review. Find the probability that all selected tasks will be completed.

Solution.

The total number of possible selections is the number of combinations of 20 tasks taken 4 at a time. Let us recall (and we will elaborate further in Section 3) that **combinations** are responsible for the number of ways to select a certain set of objects from some other, larger set. A combination is a way of selecting k objects from n objects without considering the order of selection. The number of combinations is denoted by $C(n, k)$ and represents the number of possible groups of k elements chosen from a set of n elements. Let's recall the formula for calculating combinations:

$$C(n,k) = \frac{n!}{k!\,(n-k)!}$$

where:

- ***n*** is the total number of objects;
- ***k*** is the number of selected objects;
- **!** denotes the factorial operation.

A developer selects four distinct tasks uniformly at random.

So, the total number of possible selections is *C(20,4)*. The number of favorable selections (all selected tasks are completed) is *C(12,4)*.

Therefore, the required probability is $P(A) = \frac{C(12,4)}{C(20,4)} = \frac{12!}{4!(12-4)!} \div \frac{20!}{4!(20-4)!} = \frac{12!}{4!8!} \cdot \frac{4!16!}{20!} = \frac{8!\cdot 9\cdot 10\cdot 11\cdot 12}{4!8!} \cdot \frac{4!16!}{16!\cdot 17\cdot 18\cdot 19\cdot 20} = \frac{9\cdot 10\cdot 11\cdot 12}{17\cdot 18\cdot 19\cdot 20} = \frac{33}{323} \approx 0.102.$

The following representation was used in the calculations:

$12! = 1\cdot 2\cdot 3\cdot \ldots \cdot 12 = 1\cdot 2\cdot \ldots \cdot 8\cdot \ldots \cdot 12 = 8!\cdot 9\cdot 10\cdot 11\cdot 12$

$20! = 1\cdot 2\cdot 3\cdot \ldots \cdot 20 = 1\cdot 2\cdot \ldots \cdot 16\cdot \ldots \cdot 20 = 16!\cdot 17\cdot 18\cdot 19\cdot 20$

Answer: 0.102

Tasks for independent work (homework).

1. Two fair dice are rolled. Find the probabilities of the following events:
   a) the sum of the numbers rolled is equal to 8;
   b) the sum of the numbers rolled is equal to 11, and the difference between the numbers on the dice is 1;
   c) the sum of the numbers rolled is equal to 11, if it is known that the difference between the numbers on the dice is 1;
   d) the sum of the numbers rolled is equal to 7, and the product of the numbers on the dice is 10.
2. A library has 25 books on its shelf, 15 of which are works of fiction. A reader randomly chooses 5 books to borrow. Find the probability that all selected books will be fiction.
3. A factory produces 30 electronic components, 18 of which are of high quality. A quality inspector randomly selects 6 components for testing. Find the probability that all selected components will be of high quality.
4. Three fair coins are tossed simultaneously. Find the probabilities of the following events:
   a) exactly two heads are obtained;
   b) the number of heads is equal to 3, and the number of tails is equal to 0;
   c) the number of heads is equal to 2, if it is known that at least one coin shows tails;
   d) the number of heads is equal to 1, and the number of tails is equal to 2.

5. A three-digit number is randomly chosen from all possible three-digit numbers. Find the probability that the chosen number is:

a) a number whose first digit is odd;

b) a number whose sum of digits is less than 5.

6. A three-digit number is randomly chosen from all possible three-digit numbers. Find the probability that the chosen number will be:

a) a number divisible by 5;

b) a number in which the tens digit is greater than both the hundreds and units digits.

7. A fair six-sided die is tossed twice. Find the probability of the following events:

a) the sum of the two numbers is 7;

b) the product of the two numbers is 12;

c) both numbers are even;

d) the first number is greater than the second.

8. A four-digit access code is randomly selected from all codes from 0000 to 9999. Leading zeros are allowed, and all codes are equally likely.

Find the probability that the selected code:

a) contains four different digits;

b) contains exactly one pair of equal digits, while the other two digits are different from each other and from the repeated digit;

c) contains at least two equal digits;

d) contains at least one digit 0 and at least one digit 7;

e) contains neither 0 nor 7.

9. A cube is painted on all six faces and then divided into $8 \cdot 8 \cdot 8 = 512$ smaller cubes of equal size. The smaller cubes are thoroughly mixed, and two of them are selected simultaneously.

Find the probability that:

a) both selected cubes have no painted faces;

b) both selected cubes have exactly one painted face;

c) one selected cube has exactly one painted face and the other has exactly two painted faces;

d) at least one selected cube has three painted faces;

e) the two selected cubes together have exactly three painted faces.

The order in which the two cubes are selected is not taken into account.

3. Enumerative combinatorics

Enumerative combinatorics studies methods for counting the number of possible outcomes in different types of arrangements and selections. It provides the

basic tools for solving problems involving combinations and permutations, which are widely used in probability theory.

We begin with the fundamental counting rules.

Sum rule [16]. If two actions A and B are mutually exclusive, and action A can be performed in m ways, and B in n ways, then any one of these actions (either A or B) can be performed in $\boldsymbol{m + n}$ ways.

Problem 10. A software development team consists of 15 backend developers and 12 frontend developers. In how many ways can one developer be selected as a team leader?

Solution.

There are 15 possible choices for a backend developer and 12 possible choices for a frontend developer. By the sum rule, the number of possible choices is $15 + 12 = 27$.

Answer: 27.

The Fundamental Counting Principle, also known as the **Multiplication Principle**, states that if a process consists of k successive steps, and the i-th step can be performed in n_i ways for every possible result of the preceding steps, then the entire process can be completed in $\boldsymbol{N = n_1 \cdot n_2 \cdot \ldots \cdot n_k}$ different ways [16].

Problem 11. A software company has 18 backend developers and 12 frontend developers. In how many ways can one developer be appointed as the lead project coordinator and another as the deputy project coordinator?

Solution.

The total number of developers is equal to 30. The number of ways to choose a first developer $n_1 = 30$, the number of ways to choose a second developer $n_2 = 30 - 1 = 29$. Then, according to the Fundamental Counting Principle, the total number of ways will be equal to $30 \cdot 29 = 870$.

Answer: 870.

A permutation of *n* distinct elements is an arrangement of all n elements in a specific order. Two permutations are different if the elements occur in different orders. The number of permutations of n distinct elements is $\boldsymbol{P(n) = n!}$ [16].

Problem 12. Find the number of distinct arrangements of the letters in the word HORSE.

Solution.

Since all five letters are distinct, the number of possible arrangements is $P(5) = 5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$.

Answer: 120.

A permutation of a multiset is an arrangement of elements in which some elements are identical and appear multiple times. If an element a_1 occurs n_1 times,

an element a_2 occurs n_2 times, and so on, with the element a_s occurring n_s times, then the total number of distinct permutations is calculated as:

$$\boldsymbol{P(n;\ n_1, n_2, \ldots, n_s) = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_s!}}$$

where n is the total number of elements; $n_1, n_2, \ldots, n_s$ are the numbers of repetitions of each distinct element [16].

Problem 13. Find the number of distinct arrangements of the letters in the word FINANCE.

Solution.

The word FINANCE contains seven letters, and the letter N occurs twice. All other letters occur once. Therefore, the number of distinct arrangements is:

$$P(7;\ 2) = \frac{7!}{2!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7}{1 \cdot 2} = 2520.$$

Answer: 2520.

An ordered selection with replacement is a method of choosing k elements from a set of n elements, where the same element can be selected more than once. The number of possible ordered selections is $\boldsymbol{N = n^k}$ where n is the number of available elements and k is the number of selections [16].

Problem 14. A box contains 8 different programming books. Three books are selected successively with replacement, and the order of selection is recorded. Find the number of possible selections if the books are chosen **with replacement** (after each selection, the book is returned to the box).

Solution.

Since each selected book is returned to the box, the same book can be chosen more than once. So, this is a selection with repetition. The number of possible ordered selections is $N = 8^3 = 512$.

Let's break down in more detail how this result was obtained. For the first selection there are 8 choices; for the second selection there are 8 choices (because the first book was returned back); for the third selection there are 8 choices. Using the Fundamental Counting Principle $N = 8 \cdot 8 \cdot 8 = 512$.

Answer: 512.

A *k*-permutation of *n* distinct elements is an ordered selection of k elements from a set of n elements, where each element can be chosen at most once. Since the order of selected elements is important, different arrangements of the same elements are considered different outcomes.

The number of possible k-permutations of n elements is:

$$\boldsymbol{P(n, k) = \frac{n!}{(n-k)!}}$$

where n is the total number of available elements; k is the number of selected elements; the order of elements is taken into account [16].

Problem 15. A box contains 8 different programming books. Three books are selected successively without replacement, and the order of selection is recorded. Find the number of possible selections if the books are chosen **without replacement** (after each selection, the book is not returned to the box).

Solution.

A box contains 8 different programming books. Three books are selected without replacement, which means that after a book is selected, it is not returned to the box. Since the books are different and the order of selection is important, we use the formula for k-permutations without repetition:

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

Let's break down in more detail how this result was obtained. Since the selected book is not returned, the number of available books decreases after each selection. For the first selection there are 8 choices; for the second selection there are 7 choices; for the third selection there are 6 choices. Using the Fundamental Counting Principle $N = 8 \cdot 7 \cdot 6 = 336$.

Answer: 336.

A combination of *n* elements taken *k* at a time is a selection of k elements from a set of n elements, where the order of selected elements does not matter and each element can be chosen at most once.

The number of possible combinations of n elements taken k at a time is given by:

$$\boldsymbol{C(n,k) = \frac{n!}{k!\,(n-k)!}}$$

where n is the total number of elements in the set; k is the number of selected elements; the order of elements is not considered important [16].

Problem 16. A software company has 25 programmers: 15 backend developers and 10 frontend developers. A team of 5 programmers must be selected to work on a new project. Find the number of possible teams with no restrictions on the composition of the team.

Solution.

Since the order of selection does not matter:

$$C(25,5) = \frac{25!}{5!\,(25-5)!} = \frac{25!}{5!\,20!} = \frac{20! \cdot 21 \cdot 22 \cdot 23 \cdot 24 \cdot 25}{5!\,20!}$$
$$= \frac{21 \cdot 22 \cdot 23 \cdot 24 \cdot 25}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 53130.$$

Answer: 53130.

Let us consider a few more problems with detailed solutions.

Problem 17. Find the number of distinct arrangements of the letters in each of the following words:

-FATHER

-MATHEMATICS

Solution.

1) Word: FATHER

The word FATHER contains 6 different letters. Since all letters are different, this is a **a permutation of *6* distinct elements** and the number of possible arrangements is $P(6) = 6! = 720$.

2) Word: MATHEMATICS

The word MATHEMATICS contains 11 letters. Some letters are repeated:

- M appears 2 times;
- A appears 2 times;
- T appears 2 times.

Since there are repeated letters, we use the formula for **a permutation of a multiset**: $P(11; 2, 2, 2) = \frac{11!}{2!2!2!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11}{2 \cdot 2 \cdot 2} = 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 9 \cdot 10 \cdot 11 = 4989600$.

Answer: 720; 4989600.

Problem 18. Two programmers, Alex and Ben, solve coding tasks one at a time. Each task is solved by exactly one of them. The competition ends as soon as either programmer:

- solves two consecutive tasks;
- solves three tasks in total.

Using a tree diagram, determine the number of possible sequences in which the competition can end.

Solution.

The corresponding tree diagram is shown in Figure 17:

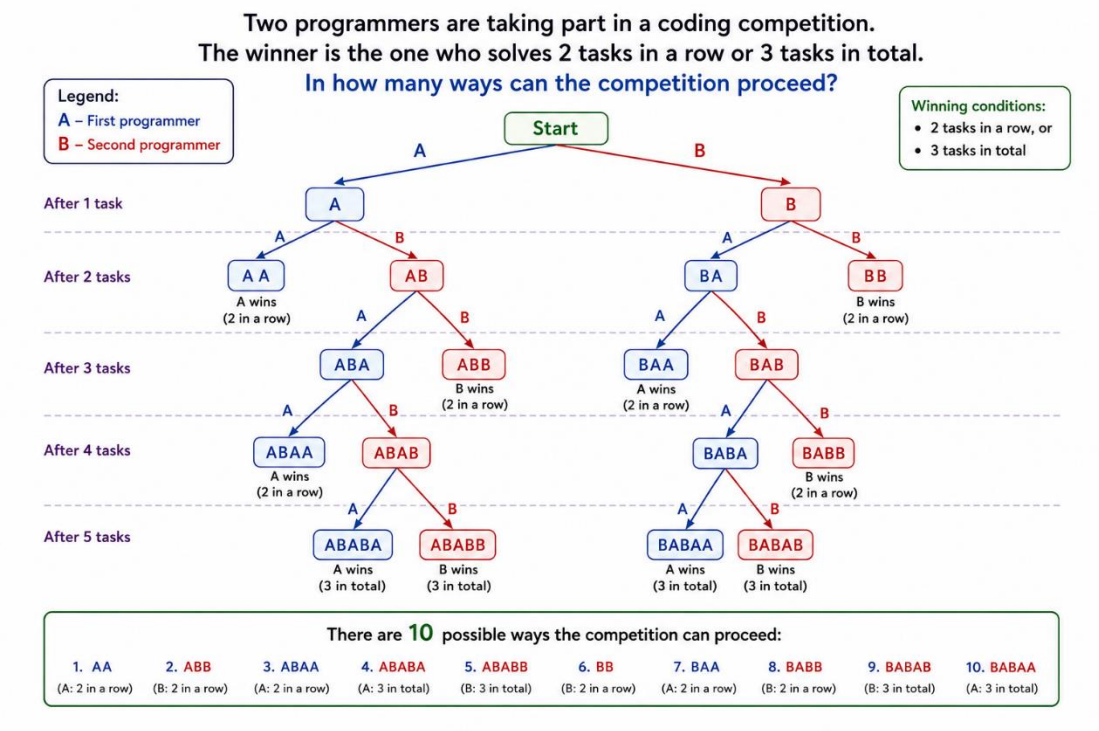


Figure 17. Tree diagram for Problem 18

Answer: 10.

Problem 19. A class consists of 10 boys and 12 girls. Determine the number of ways to choose:

a) 1 person from the class;

b) 1 boy and one girl;

c) 1 president and 1 vice president

Solution.

a) By the sum rule: $N = 10 + 12 = 22$.

b) By the Fundamental Counting Principle: $N = 10 \cdot 12 = 120$.

c) Since the two offices are distinct, order matters. We need to select a president from 22 students and a vice president from the remaining 21 students.

Since the roles are different, we use a k-permutation:

$$P(22,2) = \frac{22!}{(22-2)!} = \frac{22!}{20!} = \frac{20! \cdot 21 \cdot 22}{20!} = 21 \cdot 22 = 462.$$

This solution can also be described as follows: there are 22 ways to choose the president, and since the president has already been chosen (that is, the position is occupied by one person), there are 21 people left for the position of vice president. Then, according to the Fundamental Counting Principle (Rule of Product), the number of ways to choose 1 president and 1 vice president is: $N = 22 \cdot 21 = 462$.

Answer: a) 22; b) 120; c) 462.

Problem 20. A software company wants to create a new development team. The company has 6 backend developers, 5 frontend developers, and 8 QA engineers. The

manager needs to select 3 backend developers, 2 frontend developers, and 4 QA engineers for the project. In how many different ways can the team be formed?

Solution.

Since the order of selected employees does not matter, we use combinations without repetition. We need to select 3 developers from 6:

$$C(6,3) = \frac{6!}{3!\,(6-3)!} = \frac{6!}{3!\,3!} = \frac{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6}{1\cdot 2\cdot 3\cdot 1\cdot 2\cdot 3} = 20$$

So, there are 20 ways to choose backend developers.

We need to select 2 developers from 5:

$$C(5,2) = \frac{5!}{2!\,(5-2)!} = \frac{5!}{2!\,3!} = \frac{1\cdot 2\cdot 3\cdot 4\cdot 5}{1\cdot 2\cdot 1\cdot 2\cdot 3} = 10$$

So, there are 10 ways to choose frontend developers.

We need to select 4 QA engineers from 8:

$$C(8,4) = \frac{8!}{4!\,(8-4)!} = \frac{8!}{4!\,4!} = \frac{4!\cdot 5\cdot 6\cdot 7\cdot 8}{4!\cdot 1\cdot 2\cdot 3\cdot 4} = 70$$

The three selections can be made separately. Therefore, by the Fundamental Counting Principle: $N = C(6, 3) \cdot C(5, 2) \cdot C(8,4) = 20 \cdot 10 \cdot 70 = 14000$.

Answer: 14000.

Problem 21. A teacher has four identical red pens and three identical blue pens. On each of seven consecutive days, she gives one pen to a student. How many different sequences of pen colors are possible?

Solution.

We have a multiset of items: 4 identical red pens and 3 identical blue pens. We need to arrange them in a sequence of 7 days. Since there are identical pens, this is a permutation with repetition: $P(7;4,3) = \frac{7!}{4!\cdot 3!} = \frac{4!\cdot 5\cdot 6\cdot 7}{4!\cdot 1\cdot 2\cdot 3} = 5\cdot 7 = 35$

So, there are 35 ways to distribute pens over the 7 days.

Answer: 35.

Tasks for independent work (homework).

1. A box contains seven distinct blue balls and five distinct red balls. Determine the number of ways:

   a) choose 2 balls;

   b) choose 2 balls of the same color.

2. There are 10 married couples at the party. Find the number of ways to choose 1 man and 1 woman so that they are not married.

3. Twelve distinct candies are distributed among five children. Each of the first three children must receive two candies, and each of the remaining two children must receive three candies. In how many ways can the candies be distributed?

4. The company has 15 employees — 9 programmers and 6 designers. Determine the number of ways you can select:

a) 6 employees for a project;

b) 4 programmers and 2 designers;

c) a team leader, a deputy, and a secretary.

5. How many different arrangements of 5 flags are possible if there are 2 identical blue, 2 identical green, and 1 yellow flag?

6. There are 8 red marbles and 6 blue marbles. In how many ways can 5 marbles be chosen so that 2 of them are blue?

7. Eight spectators are standing in a line at a theater box office. How many different queues of these 8 people are possible? How many queues are possible in which:

a) four specific spectators A, B, C, D stand together in the order ABCD;

b) four specific spectators A, B, C, D stand together (in any order).

8. Suppose the code consists of 3 letters followed by 2 digits. Find the number of:

a) codes

b) codes with different letters

c) codes in which all three letters are the same.

9. A software company has 8 machine-learning engineers, 6 data engineers, and 5 QA engineers. A project team must consist of exactly:

a) 3 machine-learning engineers;

b) 2 data engineers;

c) 2 QA engineers.

After the team has been formed, one team member is appointed as the project manager, and one of the selected machine-learning engineers is appointed as the technical lead. The project manager and the technical lead must be different people. In how many different ways can the team be formed and the two positions assigned?

10. An access code consists of six symbols. Each symbol is chosen from the letters A, B, C, D, and E or from the digits 1, 2, 3, 4, 5, and 6. Symbols may be repeated. Find the number of access codes that satisfy all of the following conditions:

a) the first symbol is a letter;

b) the last symbol is a digit;

c) the code contains exactly two letters.

11. Find the number of distinct arrangements of the letters of the word PROBABILITY in which all the vowels appear consecutively. The two occurrences of the letter B and the two occurrences of the letter I are identical.

12. A shelf contains:

- 5 different books on artificial intelligence;

- 4 different books on statistics;
- 3 different books on programming.

Find the number of ways to arrange all 12 books if:

a) all books on the same subject must be placed together;

b) all books on the same subject must be placed together, and the artificial intelligence group must not be adjacent to the programming group.

13. A research center has 10 different datasets:

- 3 medical datasets;
- 7 non-medical datasets.

Four datasets are selected one after another for testing, without replacement. The order in which the datasets are selected is important. In how many ways can the four datasets be selected if exactly two of them must be medical datasets?

14. Two programmers, Alex and Ben, solve coding tasks one at a time. Each task is solved by exactly one of them. The competition ends as soon as one programmer:

- solves three tasks in total; or
- solves two consecutive tasks.

Using a tree diagram, determine the number of different possible sequences in which the competition can proceed. The sequence must stop immediately when one of the winning conditions is satisfied.

15. An artificial intelligence system analyzes eight requests. For each request, the result is recorded as:

- C — the response is correct;
- I — the response is incorrect.

How many different sequences of eight results contain exactly five correct responses and three incorrect responses, provided that no two incorrect responses occur consecutively?

16. Six different developers and four different data analysts must stand in a row for a company photograph. In how many ways can they be arranged if no two data analysts are allowed to stand next to each other? All ten employees are distinct.

4. Conditional probability

Definition 13 [2]. The conditional probability of an event A given an event B is the probability that event A occurs given that event B has occurred. It allows us to update the probability of an event when new information about the experiment is available. The event B changes the original sample space by restricting our attention only to outcomes that belong to B.

It is known that event B has occurred. Knowing this, calculate the probability of event A. The conditional probability of event A relative to event B is calculated by the formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},\ P(B) > 0$$

The notation P(A | B) is read as: "the probability of A given B" or "the probability of event A, given that event B has occurred."

The vertical line "|" is called the conditional bar and means "given that" or "under the condition that." So, P(A | B) means 'the probability that event A occurs, assuming that event B has already occurred.'

In this notation *A* is the event of interest (the event whose probability we want to find); *B* is called the conditioning event (the event that is already known to have occurred).

For example, let *R* be the event that it rains and let *C* be the event that the weather is cloudy. Then *P(R | C)* is read as "the probability of rain given cloudy weather." It describes the likelihood of rain under the condition that cloudy weather is observed.

The following examples illustrate situations in which conditional probability may be used.

Example 15. Two numbers are selected successively with replacement from the set {1, 2, 3, 4, 5, 6}. You are told that their sum is 8. What is the probability that the first number is 3?

Example 16. Given that the first submitted solution contains a syntax error, what is the probability that the second submitted solution also contains a syntax error?

Example 17. Given that a randomly selected computer program has failed a test, what is the probability that it contains a critical error?

Example 18. Given that a particular stock price decreases during a trading day, what is the probability that the market index also decreases on the same day?

Example 19. Two fair six-sided dice are rolled. What is the probability that the sum of the points is 10 if it is known that at least one die shows 6.

Let A be the event that the sum of the numbers on the dice is 10; *B* – the event that one of the dice shows 6.

We need to find the conditional probability *P(A | B)* which is read as: "the probability of A given B." We know that one die already shows 6, and we need to find the probability that the total sum is 10.

Let's find the possible outcomes under condition *B.* When two dice are rolled, there are $6 \cdot 6 = 36$ possible outcomes. But we know that one die shows 6. The possible outcomes are: (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), (1,6), (2,6), (3,6), (4,6), (5,6).

Therefore, the number of possible outcomes under condition B is 11. So, *P(B) = 11/36.*

We need the sum of the dice to be equal to 10. Among all possible outcomes we have (6,4), (4,6). So, there are 2 favorable outcomes from 36. *P(A ∩ B) = 2/36.*

The conditional probability:

$$P(A \mid B) = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11} \approx 0.18.$$

For this example, one can solve it more quickly without writing out the conditional probability formula, by using the idea of changing the sample space.

Example 20 (alternative solution to Example 19). We know that one of the dice shows 6. Therefore, we do not consider all possible outcomes of two dice. We only consider outcomes where at least one die is equal to 6.

Possible outcomes are (6,1), (6,2), (6,3), (6,4), (6,5), (6,6), (1,6), (2,6), (3,6), (4,6), (5,6). There are 11 possible outcomes.

The sum of the dice is 10 in two cases: (6,4), (4,6).

Therefore, among the 11 possible outcomes, only 2 are favorable, so *P(the sum of the points is 10 if it is known that the number 6 fell on one die)* = 2/11.

In finite models with equally likely outcomes, the reduced-sample-space approach is often simpler.

Problem 22. A fair six-sided die is rolled twice. Let *X* and *Y* be the results of the first and second throws, respectively.

Determine *P(A | B),* where $A = \{max(X, Y) = 5\}$, $B = \{min(X, Y) = 3\}$.

Solution.

We need to find the probability that the maximum value is 5, given that the minimum value is 3.

Condition *B* gives us additional information: the smaller of the two numbers rolled is equal to 3. The possible outcomes satisfying *B* are (3,3), (3,4), (3,5), (3,6), (4,3), (5,3), (6,3). So, the reduced sample space contains 7 possible outcomes.

Now we consider event $A = \{max(X, Y) = 5\}$. The maximum value is equal to 5 for exactly two outcomes: (3,5) and (5,3).

The required probability is the ratio of favorable outcomes to all outcomes satisfying the condition: *P(A | B)* = 2 / 7.

Answer: 2 / 7.

Tasks for independent work (homework).

1. A fair six-sided die is rolled twice. Let *X,Y* be the results of the first and second throws, respectively. Determine *P(A | B)* and *P(B | A),* where $A=\{max(X,Y)=6\}$, $B=\{min(X,Y)=3\}$.
2. Two cards are drawn successively without replacement from a 36-card deck containing nine cards of each suit. Find the probability that the second card is a diamond, given that the first card:

   a) is a diamond;

   b) is not a diamond.

3. Two fair six-sided dice are rolled. Find the probability that both dice show even numbers, given that the sum of the numbers shown is even.

4. A three-digit number is selected uniformly at random from all three-digit numbers with distinct digits. Find the probability that it is divisible by 5.

5. A database contains 12 programming tasks, 8 of which have been completed and 4 of which have not been completed. Four tasks are selected at random without replacement. Find the probability that all four selected tasks are completed, given that at least three of them are completed.

6. One card is randomly selected from a standard deck of 52 playing cards. It is known that the selected card is a face card, that is, a jack, queen, or king. Find the probability that the selected card is a heart or a queen.

7. Two distinct cards are selected successively without replacement from a set of eight cards numbered 1 through 8. Find the probability that at least one selected card is numbered 1, given that the sum of the two numbers is odd.

5. Basic Rules of Probability

Theorem 1 (Multiplication Rule of Probability) [1,2]. The probability that two events A and B occur simultaneously is equal to the probability of one event multiplied by the conditional probability of the other event, calculated under the assumption that the first event has occurred. Thus,

$$\boldsymbol{P(A \cap B) = P(A) \cdot P(B|A)}$$

where $P(A \cap B)$ is the probability that events A and B occur simultaneously; $P(A)$ is the probability of event A; $P(B \mid A)$ is the conditional probability of event B given that event A has already occurred.

Corollary 1 (Chain Rule of Probability) [1,2]. Consider an event A that occurs if and only if all of the events $A_1, A_2, \ldots, A_n$ occur, that is $A = A_1 \cap A_2 \cap \ldots \cap A_n$, so

$$\boldsymbol{P(A_1 \cap A_2 \cap \ldots \cap A_n)} = \boldsymbol{P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \ldots P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1})}$$

Example 21. A software company employs 12 programmers, including seven backend developers and five frontend developers. Three programmers are selected successively without replacement. Find the probability that all three selected programmers are backend developers.

Solution.

Let A_1, A_2, and A_3 denote the events that the first, second, and third selected programmers, respectively, are backend developers.

The event that all three selected programmers are backend developers can be written as:

$$A = A_1 \cap A_2 \cap A_3$$

Using the General Multiplication Rule:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2)$$

Now calculate each probability.

The probability that the first selected programmer is a backend developer is $P(A_1) = 7/12$. If the first selected programmer is a backend developer, then there are 6 backend developers left among 11 programmers: $P(A_2 \mid A_1) = 6/11$. If the first two selected programmers are backend developers, then there are 5 backend developers left among 10 programmers: $P(A_3 \mid A_1 \cap A_2) = 5/10$.

So, $P(A_1 \cap A_2 \cap A_3) = \frac{7}{12} \cdot \frac{6}{11} \cdot \frac{5}{10} = \frac{7}{44}$.

Answer: 7/44.

We now turn to the concept of independent events.

Corollary 2 (Multiplication Rule for Independent Events) [1,2]. If events $A_1, A_2, \ldots, A_n$ are mutually independent, then:

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)$$

Definition 14 [1,2]. Two events A and B are called **independent** if the occurrence of one event does not affect the probability of the other event. Mathematically, events A and B are independent if:

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

If this equality does not hold, the events are called dependent.

Example 22. A computer system contains three servers. Each server can fail during a working day.

The probability that:

- the first server fails is 0.1;
- the second server fails is 0.2;
- the third server fails is 0.15.

The probability that all three servers fail simultaneously is 0.003.

Determine whether the events A_1, A_2, and A_3 (failure of the first, second, and third servers) are mutually independent.

Solution.

Let us introduce the following events into consideration: A_1 — the first server fails; A_2 — the second server fails; A_3 — the third server fails.

For mutually independent events, the following condition must hold:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3)$$

Calculate the product of individual probabilities:

$$P(A_1) \cdot P(A_2) \cdot P(A_3) = 0.1 \cdot 0.2 \cdot 0.15 = 0.003$$

The given probability of simultaneous failure is $P(A_1 \cap A_2 \cap A_3) = 0.003$.

Since $P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3)$, the events A_1, A_2, and A_3 are mutually independent.

Answer: The events are mutually independent.

Example 23. A company has three warehouses containing computer components. Each warehouse contains 20 components. The first warehouse contains 16 working components; the second warehouse contains 18 working components; the third warehouse contains 15 working components. One component is selected independently and uniformly at random from each warehouse. Find the probability that all selected components are working.

Solution.

Let A_1, A_2, and A_3 denote the events that the components selected from the first, second, and third warehouses, respectively, are working. Since the components are selected from different warehouses, these events are independent. The probabilities of the events are: $P(A_1) = 16/20$, $P(A_2) = 18/20$, $P(A_3) = 15/20$.

Using the Multiplication Rule for Independent Events:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3) = \frac{16}{20} \cdot \frac{18}{20} \cdot \frac{15}{20} = 0.54$$

Answer: 0.54.

Definition 15 [17]. Two events A and B are called **compatible, or non-mutually exclusive**, if they can occur together. Equivalently, $A \cap B \neq \emptyset$. This means that there is at least one outcome belonging to both A and B.

If $A \cap B = \emptyset$, the events are called **mutually exclusive, or disjoint**.

Example 24. A fair six-sided die is rolled once. Let A be the event that the die shows 4, and let B be the event that the die shows an even number. Then $A = \{4\}$, $B = \{2, 4, 6\}$. Since $A \cap B = \{4\} \neq \emptyset$, the events A and B are compatible.

Theorem 2 (Addition Rule of Probability) [1]. For any two events A and B, the probability of the occurrence of at least one of these events is given by:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where $P(A \cap B)$ is the probability that both events occur simultaneously.

Example 25. A class consists of 30 students. Of these, 18 study Python, 12 study Java, and five study both languages. One student is selected at random. Find the probability that the selected student studies Python or Java.

Solution.

Let A be the event that the selected student studies Python, and let B be the event that the selected student studies Java. Since five students study both languages, the events are not mutually exclusive. Therefore,

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 18/30 + 12/30 - 5/30 = 25/30 = 5/6$.

Answer: 5/6.

Theorem 3 (Addition Rule for Mutually Exclusive Events) [1]. If A and B are mutually exclusive events, then $A \cap B = \emptyset$ and $P(A \cup B) = P(A) + P(B)$.

Example 26. A software company has 40 programmers. Among them 18 programmers are backend developers; 12 programmers are frontend developers.

Each programmer belongs to exactly one of the following categories: backend development, frontend development, or another area. One programmer is selected at random. Find the probability that the selected programmer is either a backend developer or a frontend developer.

Solution.

Since no programmer belongs to both the backend and frontend categories, $A \cap B = \emptyset$. Therefore, $P(A \cup B) = 18/40 + 12/40 = 3/4 = 0.75$.

Answer: 0.75.

Let us consider some examples of problems.

Problem 23. A software company employs 14 programmers, including 8 backend developers and 6 frontend developers. Five programmers are randomly selected for a new project. Find the probability that exactly three of the selected programmers are frontend developers.

Solution

Let A denotes the event that exactly three selected programmers are frontend developers. The total number of ways to select 5 programmers from 14 is $C(14,5)$; to choose 3 frontend developers from 6 frontend developers is $C(6,3)$; to choose the remaining 2 backend developers from 8 backend developers is $C(8,2)$. Therefore, the number of favorable selections is $C(6,3) \cdot C(8,2)$. Using the classical probability formula:

$$P(A) = \frac{\text{number of favorable outcomes}}{\text{number of all outcomes}}$$

we obtain:

$$P(A) = \frac{C(6,3) \cdot C(8,2)}{C(14,5)}$$

Now calculate:

$$C(6,3) = \frac{6!}{3!\,(6-3)!} = 20$$

$$C(8,2) = \frac{8!}{2!\,(8-2)!} = 28$$

$$C(14,5) = \frac{14!}{5!\,(14-5)!} = 2002$$

Therefore:

$$P(A) = \frac{20 \cdot 28}{2002} = \frac{280}{1001} \approx 0.28$$

Answer: 0.28.

Problem 24. The probabilities of three independent events A_1, A_2, and A_3 are p_1, p_2, and p_3, respectively. Determine the probability that only one of these events will occur.

Solution

Let A_1, A_2, and A_3 denote the events that the first, second, and third outcomes of interest occur, respectively.

The probabilities of these events are: $P(A_1) = p_1$, $P(A_2) = p_2$, $P(A_3) = p_3$.

Since the events are independent, the occurrence or non-occurrence of one event does not affect the probabilities of the others. We need to find the probability that only one of the three events occurs.

This can happen in three mutually exclusive ways:

1. Only A_1 occurs, while A_2 and A_3 do not occur:
$$P(A_1 \cap A_2^C \cap A_3^C) = p_1(1 - p_2)(1 - p_3)$$
2. Only A_2 occurs, while A_1 and A_3 do not occur:
$$P(A_1^C \cap A_2 \cap A_3^C) = (1 - p_1)p_2(1 - p_3)$$
3. Only A_3 occurs, while A_1 and A_2 do not occur:
$$P(A_1^C \cap A_2^C \cap A_3) = (1 - p_1)(1 - p_2)p_3$$

Because these cases are mutually exclusive, we use the Addition Rule of Probability:

$$P = p_1(1 - p_2)(1 - p_3) + (1 - p_1)p_2(1 - p_3) + (1 - p_1)(1 - p_2)p_3$$

Answer: $p_1(1 - p_2)(1 - p_3) + (1 - p_1)p_2(1 - p_3) + (1 - p_1)(1 - p_2)p_3$.

Problem 25. Two independently operating alarm devices are installed to signal an accident. If an accident occurs, the first device operates with probability 0.8 and the second with probability 0.7. Find the probability that exactly one device operates.

Solution

Let A denote the event that the first device operates, $P(A) = p_1 = 0.8$; B - event that the second device works $P(B) = p_2 = 0.7$. Since they operate independently, $P(A^C) = 1 - p_1 = 1 - 0.8 = 0.2$; $P(B^C) = 1 - p_2 = 1 - 0.7 = 0.3$.

We need the probability that exactly one device works. This can happen if first works and second does not work or first does not work and second works. Since these two events are mutually exclusive, we add their probabilities.

Total probability (exactly one works): $P(\textit{exactly one}) = p_1(1 - p_2) + (1 - p_1)p_2 = 0.8 \cdot 0.3 + 0.2 \cdot 0.7 = 0.38$.

Answer: 0.38.

Problem 26. A library shelf contains 15 textbooks, five of which are hardcover editions. The librarian takes 3 textbooks at random. Find the probability that at least one of the selected textbooks will be bound.

Solution

Total number of ways to choose 3 textbooks from 15 is *C(15, 3).*

"At least one bound" means: 1 bound or 2 bound or 3 bound. It is easier to calculate the complement event: "No bound textbooks" — i.e., all 3 selected are from the 10 unbound ones.

Number of ways to choose 3 unbound textbooks is *C(10, 3).* So, the probability of the complement (no bound):

$$P(\text{no bound}) = \frac{C(10,3)}{C(15,3)} = \frac{120}{455} = \frac{24}{91}.$$

Therefore, $P(A) = 1 - P(A^C) = 1 - 24/91 = 67/91 \approx 0.736$.

Answer: 67/91.

Tasks for independent work (homework).

1. A software company has 12 programmers. Among them, 7 programmers are experienced developers and 5 programmers are interns. Two programmers are selected uniformly at random without replacement. Find the probability that among the two selected programmers there will be:
   a) one experienced developer and one intern;
   b) two experienced developers;
   c) at least one experienced developer.
2. A software company has a database containing 100 test cases for a new application. Among them, 10 test cases contain errors and 90 test cases are correct**.** Four test cases are selected uniformly at random without replacement for detailed analysis. Find the probability that among the selected test cases:
   a) there are no test cases with errors**;**
   b) all selected test cases contain errors**.**
3. A software company has 20 programmers, including 12 experienced developers. Five programmers are selected uniformly at random without replacement from the list to participate in a new project. Find the probability that among the selected programmers there will be exactly three experienced developers.
4. A software company has 10 programming tasks in a queue. Among them, 4 tasks are related to backend development, while the remaining 6 tasks are related to frontend development. Three tasks are selected uniformly at random without replacement for a development sprint. Find the probability that among the selected tasks there will be at least one backend task**.**
5. A printing facility has four flatbed printing presses. Each press is operating at a given time with probability 0.8, independently of the others. Find the probability that at least one press is operating.
6. Two students take an exam. The probability that the first student passes the exam is 0.8, and for the second student it is 0.7. Assume that the students' exam

results are independent. Find the probability that only one of the students passes the exam.

7. Three students independently solve a programming problem. The probabilities that the first, second, and third students will solve it correctly are 0.6**,** 0.7, and 0.8, respectively. Find the probability that during a given test:

a) exactly one student solves it correctly;

b) exactly two students solve it correctly;

c) all three students solve it correctly.

8. There are 10 balls in a box, 4 of which are colored. Three balls are selected uniformly at random without replacement. Find the probability that at least one of the balls taken is colored.

9. Two guns are fired independently at the same target. The probability that at least one shot hits the target is 0.6. The probability that the second gun hits the target is 0.4. Find the probability that the first gun hits the target.

10. Five kittens were born in a shelter. The probability that a kitten has a genetic disorder is 0.1 (independently for each kitten). Find the probability that it will be necessary to provide special medical care for:

a) none of the kittens;

b) at least one kitten;

c) exactly one kitten;

d) more than two kittens.

11. There are 20 candies in the package, 15 of which are chocolate, and 5 are lollipops. Seven candies are selected uniformly at random without replacement. Find the probability that all the extracted candies are chocolate.

12. Three shooters shoot at the same target simultaneously. The probabilities of hitting with one shot are 0.7, 0.8, and 0.9, respectively. Assume that the shooters hit the target independently. Find the probability that with a simultaneous volley of these shooters the targets will be:

a) exactly one shooter hits the target;

b) at least one shooter hits the target.

13. A box contains 18 computer components, 14 of which are working and 4 of which are defective. Three components are selected one after another without replacement. Find the probability that all three selected components are working.

14. Three independent artificial intelligence modules process the same request. The probabilities that the modules operate correctly are:

a) 0.92 for the first module;

b) 0.85 for the second module;

c) 0.90 for the third module.

Find the probability that all three modules operate correctly.

15. Four independent servers provide the same online service. The probabilities that the servers operate without failure during a working day are: 0.95, 0.90, 0.92, and 0.88.

Find the probability that:

a) all four servers operate without failure;

b) exactly three servers operate without failure;

c) at least one server fails;

d) exactly one server fails.

16. A software company employs 20 developers:

a) 9 backend developers;

b) 7 frontend developers;

c) 4 mobile developers.

17. Four developers are selected at random to form a project team. Find the probability that the selected team contains:

a) exactly two backend developers and two frontend developers;

b) exactly two backend developers and one developer from each of the other two groups;

c) at least one developer from each group;

d) no mobile developers.

18. Five components are selected one after another without replacement from a box containing 12 working components and 5 defective components. Find the probability that:

a) the first component is defective and the remaining four are working;

b) exactly one of the five selected components is defective;

c) the fifth selected component is the only defective component;

d) at least one selected component is defective.

For part b), consider all possible positions of the defective component.

6. Law of Total Probability and Bayes' Theorem.

Theorem 4 (Law of Total Probability) [1,17]. Suppose that the sample space Ω is divided into mutually exclusive and exhaustive events $B_1, B_2, \ldots, B_n$. Then the probability of any event A is equal to the sum of the conditional probabilities of A in each case multiplied by the probability of the corresponding case:

$$P(A) = P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + \ldots + P(B_n)P(A \mid B_n).$$

It can be written as:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$$

Example 27. A software company has three development teams:

- Team A develops 40% of all software modules;
- Team B develops 35% of all software modules;
- Team C develops 25% of all software modules.

The probability that a module contains a programming error is:

- 2% for modules developed by Team A;
- 4% for modules developed by Team B;
- 5% for modules developed by Team C.

A module is selected at random.

Find the probability that the selected module contains a programming error.

Solution.

Let E denote the event that the selected module contains a programming error. Let A_1, A_2, and A_3 denote the events that the module was developed by Teams A, B, and C, respectively. The events A_1, A_2, and A_3 form a partition of the sample space because every module belongs to exactly one development team. $P(A_1) = 0.4$, $P(A_2) = 0.35$, $P(A_3) = 0.25$.

Conditional probabilities: $P(E \mid A_1) = 0.02$, $P(E \mid A_2) = 0.04$, $P(E \mid A_3) = 0.05$.

Using the Total Probability Theorem:

$$P(E) = P(E \mid A_1)P(A_1) + P(E \mid A_2)P(A_2) + P(E \mid A_3)P(A_3)$$

Substitute the values:

$$P(E) = 0.4 \cdot 0.02 + 0.35 \cdot 0.04 + 0.25 \cdot 0.05 = 0.0345.$$

Answer: 0.0345.

Problems involving the Law of Total Probability are often effectively solved by constructing a probability tree diagram, which clearly represents all possible outcomes and their corresponding probabilities.

Example 27 (alternative solution to Example 27 using a probability tree diagram).

The same probability can be obtained from the probability tree shown in Figure 18:

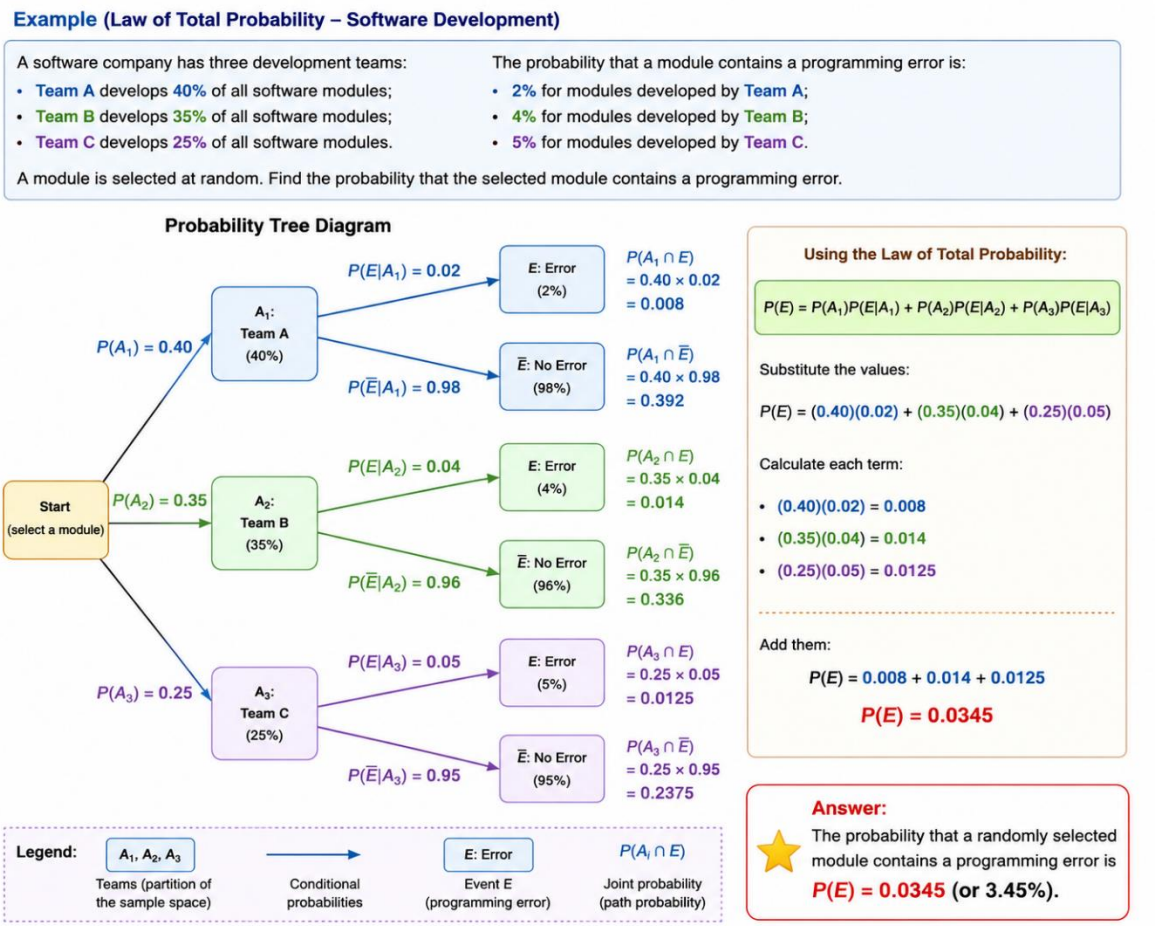


Figure18. Probability tree diagram for Example 27

Theorem 5 (Bayes' Theorem) [1,17]. Let $A_1, A_2, \ldots, A_n$ form a partition of the sample space, with $P(A_i) > 0$ for every i. For any event B with $P(B) > 0$,

$$\boldsymbol{P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}}$$

Here, $P(A_i)$ is the prior probability of hypothesis A_i, $P(B \mid A_i)$ is the probability of the observed event B under this hypothesis, and $P(A_i \mid B)$ is the posterior probability of A_i after B has been observed.

Example 28. A software company has three development teams:

- Team A develops 40% of all software modules;
- Team B develops 35% of all software modules;
- Team C develops 25% of all software modules.

The probability that a module contains a programming error is:

- 2% for modules developed by Team A;
- 4% for modules developed by Team B;
- 5% for modules developed by Team C.

It is known that the selected module contains a programming error. Find the probability that this module was developed by team A.

Solution.

Let E denote the event that the selected module contains a programming error, and let A_1 denote the event that it was developed by Team A; A_2 denote the event that it was developed by Team B; A_3 denote the event that it was developed by Team C.

The events A_1, A_2, and A_3 form a partition of the sample space because every module belongs to exactly one development team. $P(A_1) = 0.4$, $P(A_2) = 0.35$, $P(A_3) = 0.25$.

Conditional probabilities: $P(E \mid A_1) = 0.02$, $P(E \mid A_2) = 0.04$, $P(E \mid A_3) = 0.05$.

Using the Law of Total Probability:

$$P(E) = P(A_1)P(E \mid A_1) + P(A_2)P(E \mid A_2) + P(A_3)P(E \mid A_3)$$

Substitute the values:

$$P(E) = 0.4 \cdot 0.02 + 0.35 \cdot 0.04 + 0.25 \cdot 0.05 = 0.0345.$$

We need to find the probability that the erroneous module was developed by Team A. Let's use Bayes' theorem:

$$P(A_1 \mid E) = \frac{P(A_1)\,P(E \mid A_1)}{P(E)}$$

Substitute the values:

$$P(A_1 \mid E) = \frac{0.4 \cdot 0.02}{0.0345} = \frac{0.008}{0.0345} \approx 0.232$$

Answer: 0.232.

Problems involving the Bayes' Theorem are often effectively solved by constructing a probability tree diagram.

Example 28 (alternative solution to Example 28 using a probability tree diagram).

The posterior probability can also be obtained from the probability tree shown in Figure 19:

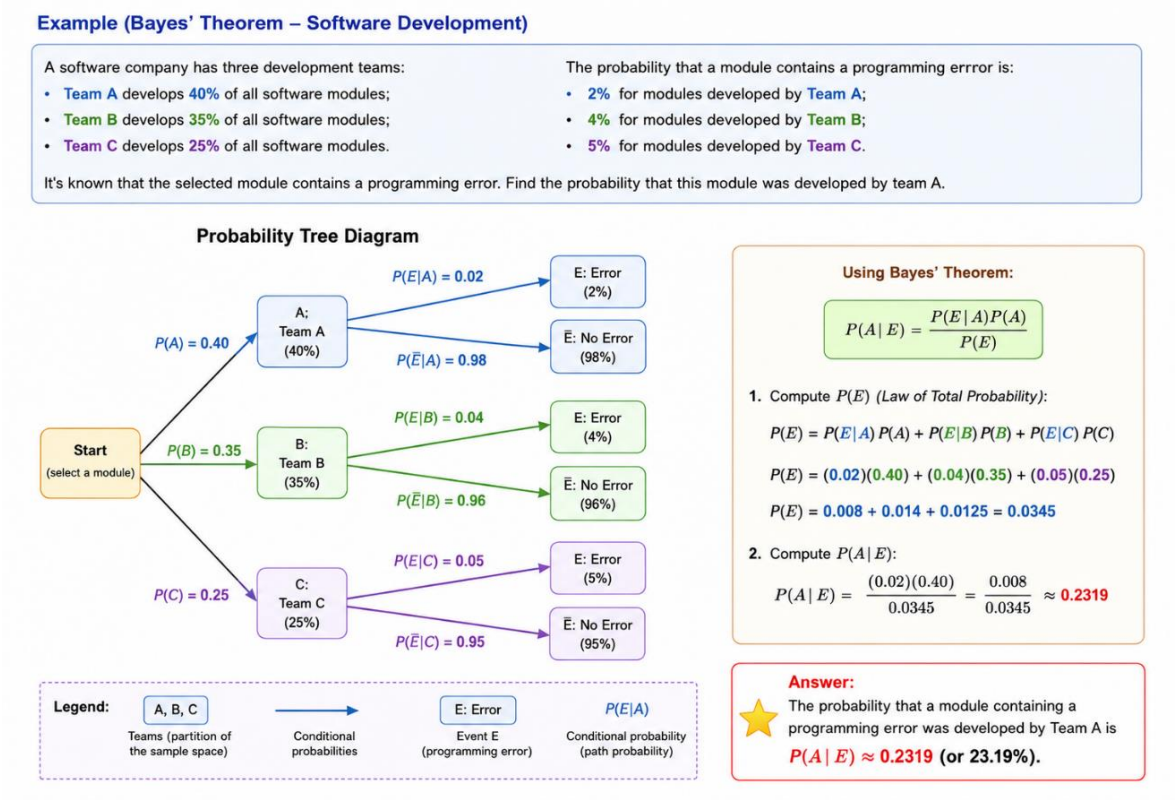


Figure19. Probability tree diagram for Example 28

Let us consider a few more examples.

Problem 27. The tournament participants are divided into three types:

- 40% are Type I players, and your probability of defeating a Type I player is 0.25;
- 35% are Type II players, and your probability of defeating a Type II player is 0.45;
- the remaining 25% are Type III players, and your probability of defeating a Type III player is 0.70.

You are paired uniformly at random with one of the participants. Find the probability that you win the game.

Solution.

Let W denote the event that you win, and let A_1, A_2, and A_3 denote the events that your opponent is a Type I, Type II, or Type III player, respectively.

The probabilities of selecting each type of opponent are: $P(A_1) = 0.4$; $P(A_2) = 0.35$; $P(A_3) = 0.25$.

The conditional probabilities of winning are: $P(W \mid A_1) = 0.25$; $P(W \mid A_2) = 0.45$; $P(W \mid A_3) = 0.7$.

By the law of total probability:

$$P(W) = P(A_1)P(W \mid A_1) + P(A_2)P(W \mid A_2) + P(A_3)P(W \mid A_3)$$

Substitute the given values:

$$P(W) = 0.4 \cdot 0.25 + 0.35 \cdot 0.45 + 0.25 \cdot 0.7 = 0.4325.$$

Answer: 0.4325.

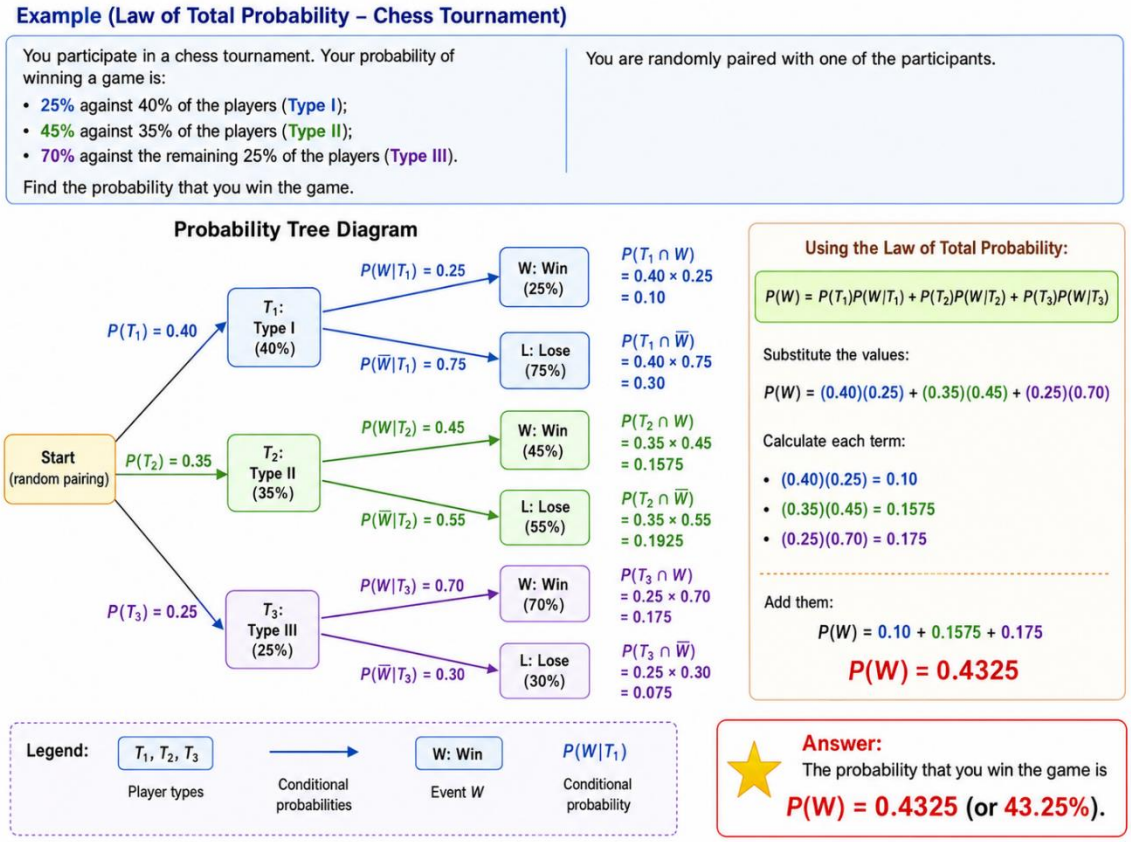


Figure 20. Probability tree diagram for Problem 27

Problem 28. The tournament participants are divided into three types:

- 40% are Type I players, and your probability of defeating a Type I player is 0.25;
- 35% are Type II players, and your probability of defeating a Type II player is 0.45;
- the remaining 25% are Type III players, and your probability of defeating a Type III player is 0.70.

You are paired uniformly at random with one of the participants. Given that you won the game, find the probability that your opponent was a Type I player.

Solution.

Let W denote the event that you win, and let A_1, A_2, and A_3 denote the events that your opponent is a Type I, Type II, or Type III player, respectively.

The probabilities of selecting each type of opponent are: $P(A_1) = 0.4$; $P(A_2) = 0.35$; $P(A_3) = 0.25$.

The conditional probabilities of winning are: $P(W \mid A_1) = 0.25$; $P(W \mid A_2) = 0.45$; $P(W \mid A_3) = 0.7$.

By the law of total probability:

$$P(W) = P(A_1)P(W \mid A_1) + P(A_2)P(W \mid A_2) + P(A_3)P(W \mid A_3)$$

Substitute the given values:

$$P(W) = 0.4 \cdot 0.25 + 0.35 \cdot 0.45 + 0.25 \cdot 0.7 = 0.4325.$$

We need to find the probability that your opponent was a Type I player given that you won. Let's use Bayes' Theorem:

$$P(A_1 \mid W) = \frac{P(A_1)\,P(W \mid A_1)}{P(W)}$$

Substitute the values:

$$P(A_1 \mid W) = \frac{0.4 \cdot 0.25}{0.4325} \approx 0.23$$

Answer: 0.23.

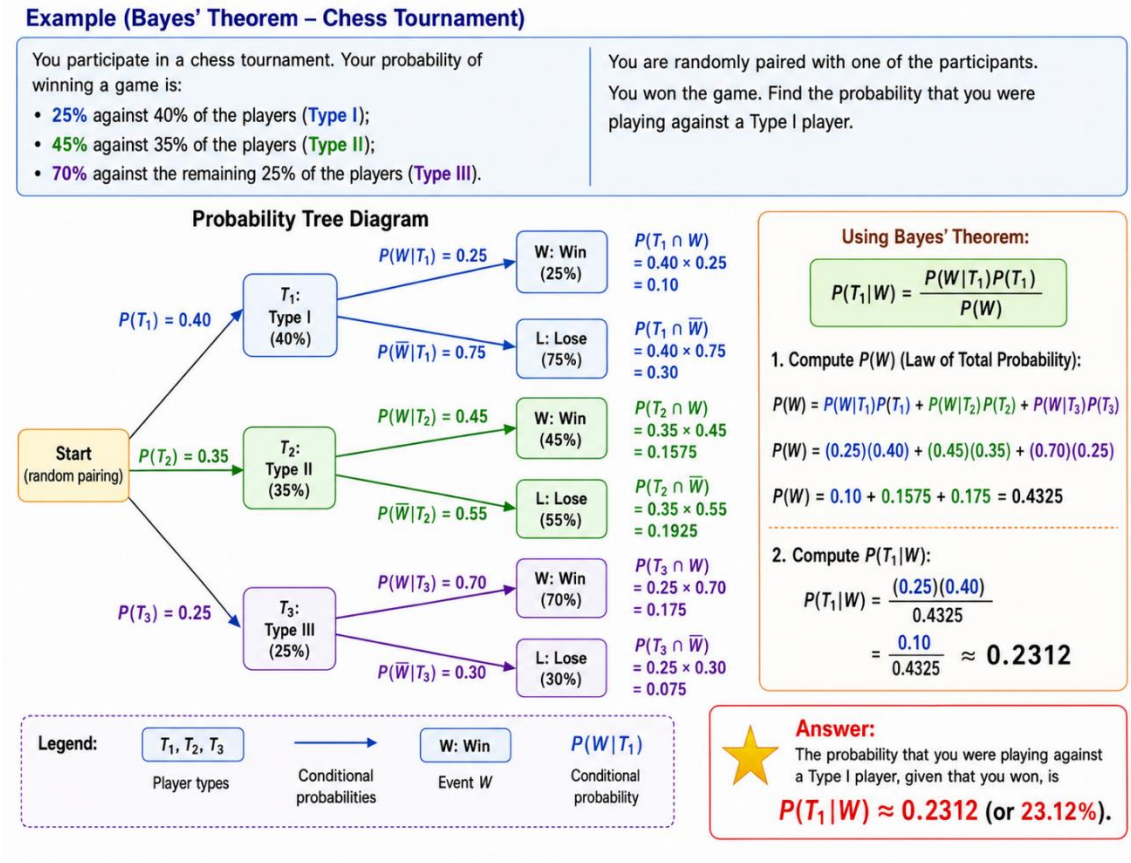


Figure 21. Probability tree diagram for Problem 28

Tasks for independent work (homework).

1. A software company has three development teams, X, Y, and Z. Team X has developed 10 software modules, 4 of which contain programming errors. Team Y has developed 6 software modules, 1 of which contains a programming error. Team Z has developed 8 software modules, 3 of which contain programming errors. One module is selected uniformly at random from the 24 modules developed by the three teams.

a) Find the probability that the selected module is error-free.

b) Given that the selected module is error-free, find the probability that it was developed by Team Z.

2. A veterinary clinic treats animals from three shelters, X, Y, and Z. Shelter X provides 50% of all animals admitted to the clinic, and 3% of these animals are diagnosed with an infectious disease. Shelter Y provides 30% of all animals admitted to the clinic, and 4% of these animals are diagnosed with an infectious disease. Shelter Z provides the remaining 20% of the animals, and 5% of these animals are diagnosed with an infectious disease. An animal is selected at random from all the animals admitted to the clinic.

a) Find the probability that the selected animal is diagnosed with an infectious disease.

b) Given that the selected animal is diagnosed with an infectious disease, find the probability that it came from Shelter Z.

3. A factory produces two types of machine parts: Type A and Type B**.** Type A and Type B parts are produced in the ratio 4 : 3**.** The probability that a Type A part fails the quality inspection is 0.08, while for a Type B part this probability is 0.15. A randomly selected part has failed the quality inspection**.** Find the probability that it is a Type B part.

4. Research conducted by psychologists has shown that men and women may respond differently to certain life situations. The study found that: 75% of women respond positively to these situations. 35% of men respond negatively to these situations. A survey was completed by 18 women and 12 men.

a) Determine the probability that a randomly selected questionnaire contains a positive response.

b) A randomly selected questionnaire contains a negative response. Find the probability that it was completed by a man.

5. Five students from Group 1, seven from Group 2, and four from Group 3 participate in a programming contest. The probabilities that a student from the respective groups solves a particular difficult problem are 0.6, 0.5, and 0.4.

a) What is the probability that a randomly selected contestant will solve the difficult problem?

b) A randomly selected contestant solved the difficult problem. What is the probability that he is from the second group?

6. A factory receives electronic components from three suppliers:

- Supplier A provides 25% of all components;
- Supplier B provides 45% of all components;
- Supplier C provides 30% of all components.

The probabilities that a component is defective are:

- 0.02 for Supplier A;

- 0.03 for Supplier B;
- 0.04 for Supplier C.

Find:

a) The probability that a randomly selected component is defective;

b) The probability that a defective component was supplied by Supplier C.

Construct a probability tree diagram.

7. An artificial intelligence platform receives two types of requests:
   - 60% are simple;
   - 40% are complex.

For a simple request:

- Model A is used with probability 0.70;
- Model B is used with probability 0.30.

For a complex request:

- Model A is used with probability 0.25;
- Model B is used with probability 0.75.

The probability of a correct response is:

- 0.96 when Model A processes a simple request;
- 0.91 when Model B processes a simple request;
- 0.78 when Model A processes a complex request;
- 0.88 when Model B processes a complex request.

Find:

a) The probability that the platform gives a correct response;

b) Given that the response is incorrect, find the probability that the request was complex;

c) Given that the response is correct, find the probability that Model B was used.

Construct a probability tree diagram containing all possible branches.

8. An artificial intelligence system operates in one of three versions:
   - Version I is used in 50% of cases;
   - Version II is used in 30% of cases;
   - Version III is used in 20% of cases.

The probability of correctly classifying any given image is:

- 0.9 for Version I;
- 0.8 for Version II;
- 0.7 for Version III.

The same version of the system classifies two images. Conditional on the version being used, the two classification results are independent.

Find:

a) the probability that both images are classified correctly;

b) the probability that exactly one image is classified correctly;

c) the probability that neither image is classified correctly;

d) the probability that Version III was used, given that exactly one image was classified correctly.

7. The binomial probability formula. The de Moivre–Laplace Theorem: Local and Integral Forms.

A **Bernoulli trial** is a random experiment with exactly two mutually exclusive outcomes, conventionally called success and failure. The probability of success is denoted by p, and the probability of failure is denoted by q, where $q = 1 - p$.

A **binomial experiment** consists of a fixed number n of Bernoulli trials and satisfies the following conditions:

1. The number of trials, n, is fixed in advance.
2. Each trial has exactly two possible outcomes: success and failure.
3. The probability of success remains constant from trial to trial and is equal to p. Consequently, the probability of failure is $q = 1 - p$.
4. The trials are independent. This means that the outcome of one trial does not affect the outcome of any other trials.

The probability of obtaining exactly k successes in n trials is given by **the binomial probability formula**:

$$\boldsymbol{P_n(k) = C(n,k)p^k q^{n-k}}$$

where $C(n, k)$ is the binomial coefficient, representing the number of ways to choose the k trials in which success occurs. The factor p^k is the probability of success in those k trials, and q^{n-k} is the probability of failure in the remaining $n - k$ trials [1, 2, 17].

Example 29. A fair coin is tossed ten times. Find the probability of obtaining exactly six heads.

Why the Binomial Formula Can Be Applied

The Binomial Formula can be used because all the necessary conditions are satisfied:

1. The number of trials is fixed: the coin is tossed exactly 10 times.
2. Each trial has exactly two possible outcomes: heads or tails.
3. The probability of success remains constant in every trial: $p = 0.5$.
4. The trials are independent: the result of one toss does not affect the results of the other tosses.
5. We are interested in the probability of obtaining exactly 6 successes.

Solution.

Each coin toss is a Bernoulli trial because it has exactly two possible outcomes: heads and tails. Let heads represent success and tails represent failure. Since the coin is fair, $p = 0.5$, $q = 1 - p = 0.5$.

The coin is tossed 10 times, so $n = 10$. We need to find the probability of obtaining exactly 6 heads, so $k = 6$.

The probability of obtaining exactly k successes in n independent Bernoulli trials is calculated using the binomial probability formula:

$$P_{10}(6) = C(10,6) \cdot 0.5^6 \cdot 0.5^{10-6} = \frac{10!}{6!\,4!} \cdot \left(\frac{1}{2}\right)^6 \cdot \left(\frac{1}{2}\right)^4 = \frac{6! \cdot 7 \cdot 8 \cdot 9 \cdot 10}{6! \cdot 1 \cdot 2 \cdot 3 \cdot 4} \cdot \left(\frac{1}{2}\right)^{10}$$

$$= \frac{7 \cdot 3 \cdot 10}{1024} = \frac{210}{1024} \approx 0.21.$$

Answer: 0.21.

Local de Moivre–Laplace Theorem.

The local de Moivre–Laplace theorem is used to approximate the probability of obtaining exactly k successes in a **large number** of independent Bernoulli trials.

The theorem can be applied under the following conditions:

1. The experiment consists of n independent Bernoulli trials.
2. Each trial has exactly two possible outcomes, conventionally called success and failure.
3. The probability of success is the same in every trial and is denoted by p.
4. The probability of failure is $q = 1 - p$.
5. The number of trials n is sufficiently large.
6. Both np and nq are sufficiently large. A commonly used practical guideline is: $np \geq 5$ and $nq \geq 5$. For a more accurate approximation, it is preferable that $np \geq 10$ and $nq \geq 10$.
7. The number k of successes should not be too far from the expected number np. The approximation is most accurate when k is close to np and may become less accurate in the tails of the distribution.

Theorem 6 (The Local Form of the de Moivre–Laplace Theorem) [6, 14]. Suppose that n independent Bernoulli trials are performed and that the probability p of the occurrence of an event A is the same in each trial, where $0 < p < 1$. Let $q = 1 - p$. If n is sufficiently large, then the probability $P_n(k)$ that the event A occurs exactly k times in n trials can be approximated by the formula:

$$\boldsymbol{P_n(k) \approx \frac{1}{\sqrt{npq}} \varphi(x)}$$

where $\boldsymbol{x = \frac{k-np}{\sqrt{npq}}}$ and $\boldsymbol{\varphi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}}$ is called the **standard normal probability density function**. It is also referred to as the Gaussian density function.

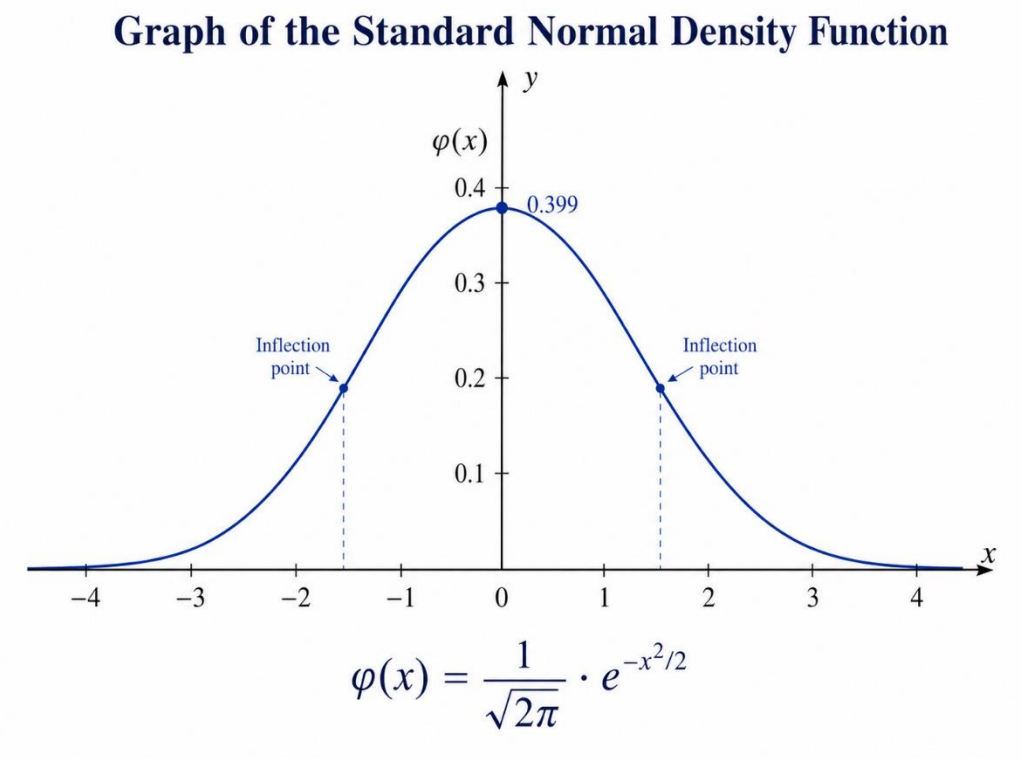


Figure 22. The graph of standard normal probability density function

The standard normal density function has the following properties [6, 14]:

1. *φ(x)* is defined for all real *x*.
2. *φ(x) > 0* for all *x*.
3. ***φ(−x) = φ(x)***, so the function is even.
4. The graph is symmetric about the *y*-axis.
5. The maximum value is attained at *x = 0*.
6. The function increases on (−∞, 0) and decreases on (0, +∞).
7. *φ(x)* approaches zero as *x* approaches ±∞.
8. Although *φ(x)* is positive for every finite value of *x*, its values become negligibly small for large |*x*|. In most practical calculations, *φ(x)* may be taken **to be approximately zero when $|x| \geq 4$**.

In traditional hand calculations, values of φ(x) are obtained from a table. In such table, for each value of *x*, the corresponding value *φ(x)* is provided. Since the function *φ(x)* is even, *φ(−x) = φ(x),* for a negative argument, its absolute value is used. For example: *φ(*−1.27) = φ(1.27) ≈ 0.1781.

The table of values of the function is given in Appendix I.

Example 30. A cloud service processes 500 independent user requests. The probability that a single request results in a server error is 0.12. Find the probability that exactly 50 of the 500 requests result in server errors.

Solution.

Each user request can be regarded as a Bernoulli trial with two possible outcomes: the request results in a server error or the request is processed without a server error. Let the occurrence of a server error be considered a success.

The requests are assumed to be independent, and the probability of a server error is the same for each request. Therefore, the local de Moivre–Laplace theorem can be used to approximate the required probability.

The total number of requests is $n = 500$. The probability that a single request results in a server error is $p = 0.12$. The probability that a request is processed without a server error is: $q = 1 - p = 1 - 0.12 = 0.88$.

We need to find the probability that exactly 50 requests result in server errors, so $k = 50$. The local de Moivre–Laplace theorem is appropriate when the number of trials is large and both np and nq are sufficiently large.

Calculate:

$np = 500 \cdot 0.12 = 60$

$nq = 500 \cdot 0.88 = 440$

Both values are much greater than 10. Therefore, the binomial distribution can be approximated by the normal distribution, and the local de Moivre–Laplace theorem can be applied.

The probability of exactly k successes in n independent Bernoulli trials is approximately:

$$P_n(k) \approx \frac{1}{\sqrt{npq}}\varphi(x)$$

where $x = \frac{k-np}{\sqrt{npq}}$.

The standardized value x is:

$$x = \frac{k - np}{\sqrt{npq}} = \frac{50 - 500 \cdot 0.12}{\sqrt{500 \cdot 0.12 \cdot 0.88}} \approx -1.3762$$

Rounding to two decimal places $x \approx -1.38$

The standard normal density function is even, so, $\varphi(-1.38) = \varphi(1.38)$ and using the table of values of the standard normal density function (Appendix I): $\varphi(1.38) \approx 0.1539$.

Substitute the obtained values into the local de Moivre–Laplace formula:

$$P_{500}(50) \approx \frac{1}{\sqrt{500 \cdot 0.12 \cdot 0.88}} \cdot 0.1539 \approx 0.02.$$

Answer: 0.02.

The Integral Form of the de Moivre–Laplace Theorem.

The integral form of the de Moivre–Laplace theorem is used to approximate the probability that the number of successes in a **large number** of independent Bernoulli trials lies within a specified interval.

Unlike the local form, which is used to find the probability of obtaining exactly *k* successes, the integral form is used when the number of successes is required to be between two given values.

The theorem can be applied under the following conditions:

- The experiment consists of *n* independent Bernoulli trials.
- Each trial has exactly two possible outcomes, conventionally called success and failure.
- The probability of success is the same in every trial and is denoted by *p*.
- The probability of failure is $q = 1 - p$.
- The number of trials n is sufficiently large.
- Both *np* and nq are sufficiently large. A commonly used practical guideline is: $np \geq 5$ and $nq \geq 5$.

Theorem 7 (The Integral Form of the de Moivre–Laplace Theorem) [6, 17]. Suppose that *n* independent Bernoulli trials are performed and that the probability *p* of the occurrence of an event A is the same in each trial, where $0 < p < 1$ and $q = 1 - p$. If *n* is sufficiently large, then the probability that the event A occurs at least k_1 times and at most k_2 times can be approximated by the formula:

$$\boldsymbol{P(k_1, k_2) \approx \Phi(x_2) - \Phi(x_1)},$$

where

$$\boldsymbol{x_1 = \frac{k_1 - np}{\sqrt{npq}}},$$

$$\boldsymbol{x_2 = \frac{k_2 - np}{\sqrt{npq}}},$$

and

$$\boldsymbol{\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\, dt}$$

is the Laplace function.

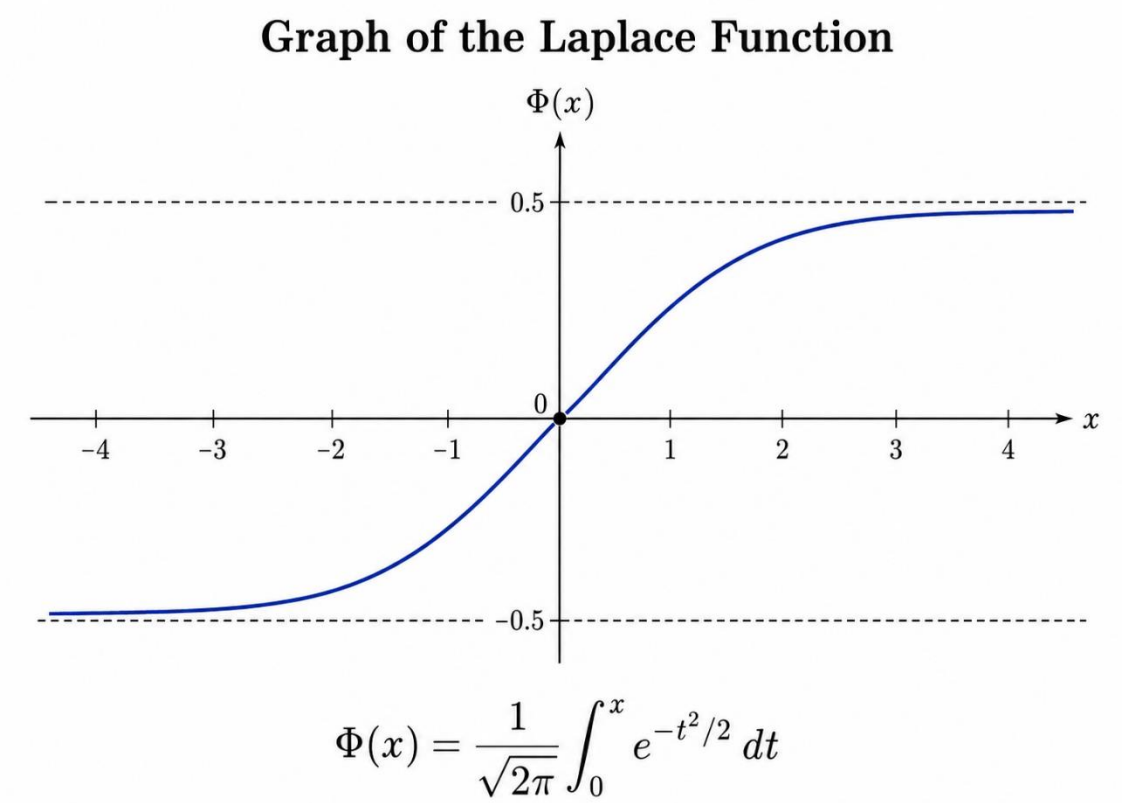


Figure 23 .The graph of Laplace function

The Laplace function has the following properties [6]:

- *Φ(x)* is defined for all real values of *x*.
- *Φ(0)* = 0.
- The function *Φ(x)* is odd: ***Φ(-x) = -Φ(x).***
- The function is strictly increasing on the entire real line.
- The values of the Laplace range from -0.5 to 0.5.
- As *x* approaches positive infinity:

$$\lim_{x\to+\infty} \Phi(x) = \frac{1}{2}.$$

- As *x* approaches negative infinity:

$$\lim_{x\to-\infty} \Phi(x) = -\frac{1}{2}.$$

Although *Φ(x)* does not reach ±0.5 for any finite value of *x*, its values become very close to these limits when $|x|$ is large. In most practical calculations, the following approximations may be used:

$$\Phi(x) \approx 0.5 \text{ when } x \geq 4,$$
$$\Phi(x) \approx -0.5 \text{ when } x \leq -4.$$

In problems involving the integral form of the de Moivre–Laplace theorem, the values of the Laplace function are usually found using a table **(Appendix II)**. The table generally contains the values of *Φ(x)* for non-negative arguments only. Since

the Laplace function is odd, $\Phi(-x) = -\Phi(x)$, the value for a negative argument can be obtained by changing the sign of the corresponding value for the positive argument.

For example, suppose that $x = -1.27$, $\Phi(-1.27) = -\Phi(1.27) = -0.398$.

The table of values of the Laplace function is given in Appendix II.

Difference Between the Local and Integral Forms

The local form of the de Moivre–Laplace theorem is used when the problem asks for the probability of obtaining exactly k successes. The integral form is used when the problem asks for the probability that the number of successes belongs to an interval.

Thus:

- exactly k successes - the local form of the de Moivre-Laplace theorem;
- from k_1 to k_2 successes - the integral form of the de Moivre-Laplace theorem;
- at least or at most a given number of successes - the integral form of the de Moivre–Laplace theorem.

Example 31. An artificial intelligence system classifies 300 images independently. The probability that the system classifies any given image correctly is 0.8. Find the probability that between 230 and 250 images, inclusive, are classified correctly.

Solution.

Each image classification can be regarded as an independent Bernoulli trial with two possible outcomes: the image is classified correctly or the image is classified incorrectly.

Let a correct classification be considered a success.

The number of trials is $n = 300$. The probability of success is $p = 0.8$, and the probability of failure is $q = 1 - p = 1 - 0.8 = 0.2$.

We need to find the probability that the number of correctly classified images is between 230 and 250, inclusive: $P(230, 250)$.

Since the problem asks for the probability that the number of successes lies within an interval, we use the integral form of the de Moivre–Laplace theorem.

Check the applicability of the theorem: $np = 300 \cdot 0.8 = 240$, $nq = 300 \cdot 0.2 = 60$.

Both values are greater than 10. Therefore, the normal approximation can be applied.

The required probability is approximated by

$$P(k_1, k_2) \approx \Phi(x_2) - \Phi(x_1),$$

where

$$x_1 = \frac{k_1 - np}{\sqrt{npq}}, x_2 = \frac{k_2 - np}{\sqrt{npq}},$$

and $\Phi(x)$ is the Laplace function.

In this problem $k_1 = 230$, $k_2 = 250$.

Let's calculate x_1 and x_2:

$$x_1 = \frac{k_1 - np}{\sqrt{npq}} = \frac{230 - 300 \cdot 0.8}{\sqrt{300 \cdot 0.8 \cdot 0.2}} \approx -1.44$$

$$x_2 = \frac{k_2 - np}{\sqrt{npq}} = \frac{250 - 300 \cdot 0.8}{\sqrt{300 \cdot 0.8 \cdot 0.2}} \approx 1.44$$

Thus,

$$P(230, 250) \approx \Phi(1.44) - \Phi(-1.44).$$

Let's use the table of values of the Laplace function (Appendix II):

$$P(230, 250) \approx \Phi(1.44) - \Phi(-1.44) \approx 0.4251 - (-0.4251) = 0.8502.$$

Answer: 0.85.

Let us consider a few more problems from this section.

Problem 29. A software tester checks ten program modules. Each module independently contains an error with probability 0.15. Find the probability that exactly two of the modules contain errors.

Solution.

Checking each software module can be regarded as a Bernoulli trial because there are exactly two possible outcomes: the module contains an error or the module does not contain an error.

Let the event that a module contains an error be considered a success.

The probability of success in each trial is $p = 0.15$. The probability that a module does not contain an error is $q = 1 - p = 1 - 0.15 = 0.85$. The tester checks 10 independently selected modules, so $n = 10$. We need to find the probability that errors are found in exactly 2 modules. Therefore $k = 2$.

The probability of exactly k successes in n independent Bernoulli trials is calculated using the binomial probability formula:

$$P_{10}(2) = C(10,2) \cdot 0.15^2 \cdot 0.85^{10-2} = \frac{10!}{2!\,8!} \cdot 0.15^2 \cdot 0.85^8$$

$$= \frac{8! \cdot 9 \cdot 10}{8! \cdot 1 \cdot 2} \cdot 0.15^2 \cdot 0.85^8 = 45 \cdot 0.15^2 \cdot 0.85^8 \approx 0.28.$$

Answer: 0.28.

Problem 30. A university administers an online test to 2500 students. Each student passes the test independently with probability 0.64. Find the probability that exactly 1550 students pass the test.

Solution.

Each student's test result can be regarded as a Bernoulli trial with two possible outcomes: the student passes the test or the student does not pass the test.

The students' test results are assumed to be independent. Since the number of students is large and the problem asks for the probability that exactly 1550 students pass, we use the local de Moivre–Laplace theorem.

The number of students is $n = 2500$. The probability that a student passes the test is $p = 0.64$. The probability that a student does not pass the test is $q = 1 - p = 1 - 0.64 = 0.36$. The required number of students who pass is $k = 1550$.

First of all, let's calculate x:

$$x = \frac{k - np}{\sqrt{npq}} = \frac{1550 - 2500 \cdot 0.64}{\sqrt{2500 \cdot 0.64 \cdot 0.36}} \approx -2.0833$$

Rounding to two decimal places $x \approx -2.08$.

The standard normal density function is even, so, $\varphi(-2.08) = \varphi(2.08)$ and using the table of values of the standard normal density function (Appendix I): $\varphi(2.08) \approx 0.0459$.

Substitute the obtained values into the local de Moivre–Laplace formula:

$$P_{2500}(1550) \approx \frac{1}{\sqrt{2500 \cdot 0.64 \cdot 0.36}} \cdot 0.0459 \approx 0.0019.$$

Answer: 0.0019.

Problem 31. An artificial intelligence system analyzes 200 images independently. The probability that the system classifies any given image correctly is 0.70. Find the probability that the number of correctly classified images is:

a) at least 130 and no more than 150;

b) at least 130;

c) no more than 129.

Solution.

Each image classification can be regarded as a Bernoulli trial with two possible outcomes: the image is classified correctly or the image is classified incorrectly.

Let a correct classification be considered a success. The classifications are assumed to be independent, and the probability of success is the same for every image.

The given values are $n = 200$, $p = 0.7$, $q = 1 - p = 1 - 0.7 = 0.3$.

Since the number of trials is large and the problem asks for probabilities involving intervals and inequalities, we use the integral form of the de Moivre–Laplace theorem.

Checking the applicability of the theorem

Calculate $np = 200 \cdot 0.7 = 140$, $nq = 200 \cdot 0.3 = 60$. Both values are greater than 10. Therefore, the normal approximation can be applied.

a) The probability of at least 130 and no more than 150 correct classifications

We need to find $P(130, 150) = \Phi(x_2) - \Phi(x_1)$.

Let's calculate x_1 and x_2:

$$x_1 = \frac{k_1 - np}{\sqrt{npq}} = \frac{130 - 200 \cdot 0.7}{\sqrt{200 \cdot 0.7 \cdot 0.3}} \approx -1.54$$

$$x_2 = \frac{k_2 - np}{\sqrt{npq}} = \frac{150 - 200 \cdot 0.7}{\sqrt{200 \cdot 0.7 \cdot 0.3}} \approx 1.54$$

Thus,

$$P(130, 150) \approx \Phi(1.54) - \Phi(-1.54).$$

Let's use the table of values of the Laplace function (Appendix II):

$$(130, 150) \approx \Phi(1.54) - \Phi(-1.54) \approx 0.4382 - (-0.4382) = 0.8764.$$

b) The probability of at least 130 correct classifications

We need to find $P(130, 200) = \Phi(x_2) - \Phi(x_1)$.

Let's calculate x_1 and x_2:

$$x_1 = \frac{k_1 - np}{\sqrt{npq}} = \frac{130 - 200 \cdot 0.7}{\sqrt{200 \cdot 0.7 \cdot 0.3}} \approx -1.54$$

$$x_2 = \frac{k_2 - np}{\sqrt{npq}} = \frac{200 - 200 \cdot 0.7}{\sqrt{200 \cdot 0.7 \cdot 0.3}} \approx 9.26$$

Thus,

$$P(130, 200) \approx \Phi(9.26) - \Phi(-1.54).$$

Let's use the table of values of the Laplace function (Appendix II):

$$P(130, 200) \approx \Phi(9.26) - \Phi(-1.54) \approx 0.5 - (-0.4382) = 0.9382.$$

c) The probability of no more than 129 correct classifications

We need to find $P(0, 129)$. The events from b) and c) are complementary. Therefore $P(0, 129) = 1 - P(130, 200) = 1 - 0.9382 = 0.0618$.

Answer: a) 0.8764; b) 0.9382; c) 0.0618

Tasks for independent work (homework).

1. Two equally skilled chess players play a series of independent games. Assuming that draws are impossible, which event is more likely for a given player: winning exactly 3 out of 6 games or winning exactly 4 out of 8 games?

2. A cloud service processes five independent user requests. The probability that a single request is completed successfully is 0.8. Find the probability that:

   a. fewer than two requests are completed successfully;
   b. at least two requests are completed successfully.

3. A litter consists of eight puppies. Assume that each puppy is female independently of the others with probability 0.47. Find the probability that the litter contains:

   a) exactly three female puppies;

b) no more than four female puppies;

c) more than five female puppies;

d) at least two but no more than six female puppies.

4. The test consists of eight problems and is considered passed if a student solves at least five of them. The student solves each problem independently with probability 0.7. Find the probability that the student passes the test.

5. A wildlife rehabilitation center treats 150 injured birds. Each bird recovers independently with probability 0.7. Find the probability that exactly 100 of the birds recover.

6. An artificial intelligence system classifies 200 images independently. The probability of correctly classifying any given image is 0.85. Find the probability that exactly 165 images are classified correctly.

7. A test consists of 70 independent true-or-false questions. To pass the test, a student must answer at least 35 questions correctly. Assuming that the student selects each answer at random, find the probability that they pass the test.

8. A fair coin is tossed 300 times. Find the probability that the number of heads is between 100 and 250, inclusive.

9. An event occurs independently in each trial with probability 0.7. Find the smallest number of trials n such that the probability of at least 70 occurrences is at least 0.8?

10. An artificial intelligence system analyzes 600 medical images independently. The probability that the system classifies any given image correctly is 0.82. Using the integral form of the de Moivre–Laplace theorem and the continuity correction, find:

a) the probability that between 475 and 510 images, inclusive, are classified correctly;

b) the probability that the number of correctly classified images differs from 492 by no more than 15;

c) the smallest positive integer d such that the probability that the number of correctly classified images differs from 492 by no more than d is at least 0.95.

11. A cybersecurity system independently checks 12 incoming files. The probability that any given file contains malicious code is 0.10. Find the probability that exactly two of the checked files contain malicious code.

12. A production line manufactures electronic sensors independently. The probability that a sensor is defective is 0.04. Ten sensors are selected for inspection. Find the probability that:

a) none of the selected sensors is defective;

b) exactly one selected sensor is defective;

c) at least one selected sensor is defective.

13. A cloud platform processes 800 independent requests. The probability that a request cannot be processed successfully is 0.05. Find the probability that exactly 35 requests cannot be processed successfully.

14. An automated translation system independently translates 500 sentences. The probability that any given sentence is translated correctly is 0.86. Using the integral form of the de Moivre–Laplace theorem, find the probability that between 420 and 445 sentences, inclusive, are translated correctly.

15. A medical artificial intelligence system independently examines 400 images. The probability that it correctly identifies the condition shown in any given image is 0.75. Find the probability that the system correctly identifies the condition:

a) in at least 285 and no more than 315 images;

b) in at least 285 images;

c) in no more than 284 images.

Use the integral form of the de Moivre–Laplace theorem.

8. Discrete Random Variables.

Definition 16 [1, 17]. A **random variable** is a real-valued function defined on the sample space of a random experiment that assigns exactly one real number to each possible outcome.

If Ω is the sample space of the experiment, a random variable X is a function $X: \Omega \to \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers. For every outcome $\omega \in \Omega$, the value $X(\omega)$ is the numerical value assigned to that outcome.

The value of a random variable does not represent an additional outcome of the experiment. Rather, it is a numerical characteristic of the outcome that has occurred. Different outcomes may be assigned the same numerical value if they share the characteristic being measured.

For example, suppose that three coins are tossed and X denotes the number of heads obtained. Then X can take the values 0, 1, 2, and 3:

- $X = 0$ for the outcome TTT;
- $X = 1$ for the outcomes HTT, THT, and TTH;
- $X = 2$ for the outcomes HHT, HTH, and THH;
- $X = 3$ for the outcome HHH.

Thus, X does not record the complete outcome of the experiment. It records only the numerical characteristic of interest.

Thus, the random variable does not record the complete result of the experiment. It records only the numerical characteristic of interest – in this case, the number of heads.

Definition 17 [2]. The set of all values that a random variable X can take is called **the range of X**, or its **set of possible values**. For the example above, the possible values of X are $\{0, 1, 2, 3\}$.

An event involving a random variable is determined by the outcomes for which the value of the random variable satisfies a specified condition. For example, the event $\{X = 2\}$ consists of all outcomes in which exactly two heads are obtained: {HHT, HTH, THH}.

Similarly, the event $\{X \geq 2\}$ consists of all outcomes in which at least two heads are obtained: {HHT, HTH, THH, HHH}.

Definition 18 [1]. A random variable is called **discrete** if it takes a finite or countably infinite set of possible values. In the example above, X is a discrete random variable because it can take only the four values 0, 1, 2, and 3.

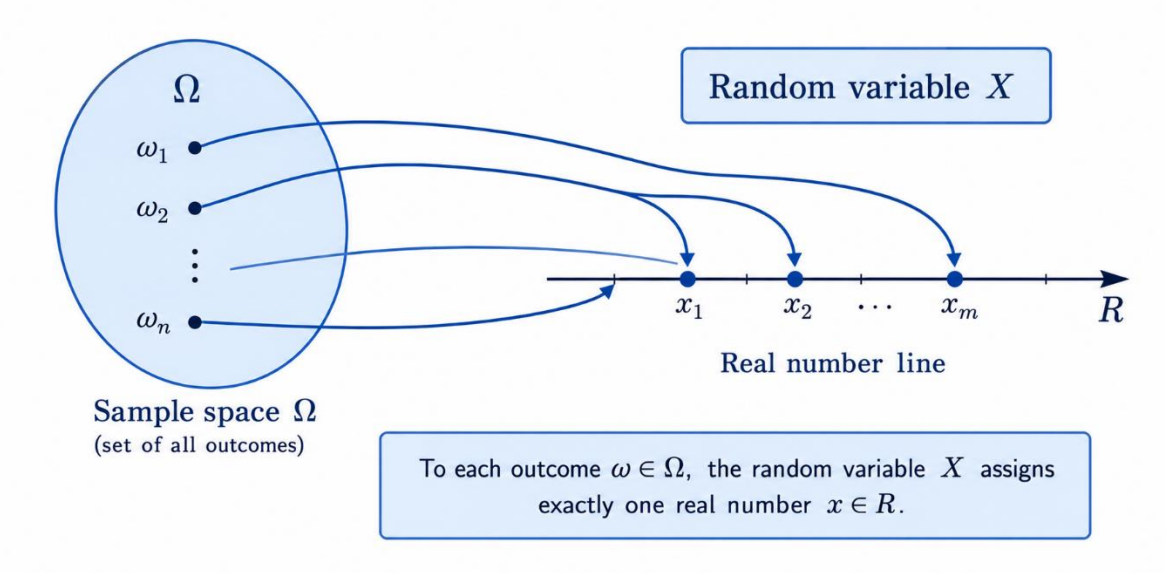


Figure 24. Mapping from the sample space to the values of a discrete random variable

The following are examples of discrete random variables

Example 32. Let X be the number of heads obtained when three fair coins are tossed. The random variable X can take one of the following values: $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$.

Example 33. In an experiment involving two rolls of a six-sided die, the following quantities are examples of random variables:

- the sum of the numbers obtained;
- the product of the numbers obtained;
- the larger of the two numbers;
- the number of times a six is obtained.

Example 34. Let X be the sum of the numbers obtained when two fair dice are rolled. The smallest possible value of X is 2, and the largest possible value is 12. Therefore, X can take the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12.

Random variables are denoted by uppercase letters, such as X, Y, and Z, while their possible values are denoted by the corresponding lowercase letters, such as x, y, and z.

Definition 19 [6]. The probability mass function (PMF) of a discrete random variable X is a function that assigns to each real number x the probability that X takes the value x.

It is defined by:

$$\boldsymbol{p_X(x) = P(X = x)}$$

The probability mass function satisfies the following properties:

- $p_X(x) \geq 0$ for every x
- the sum of the probabilities over all possible values of X is equal to 1: $\sum_x p_X(x) = 1$

Thus, if a discrete random variable X can take the values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$, then

$$\boldsymbol{p_X(x_i) = P(X = x_i) = p_i}$$

where $p_i \geq 0$ and $p_1 + p_2 + \ldots = 1$.

Example 35. Three fair coins are tossed simultaneously. Let X be the number of heads obtained. Construct the probability mass function of X and verify that it satisfies the properties of a probability mass function.

Solution.

The sample space of the experiment is Ω = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

The random variable X represents the number of heads obtained. Its possible values are 0, 1, 2, and 3.

Let us determine which outcomes correspond to each possible value of X:

- $X = 0$ for the outcome TTT.
- $X = 1$ for the outcomes HTT, THT, and TTH.
- $X = 2$ for the outcomes HHT, HTH, and THH.
- $X = 3$ for the outcome HHH.

Therefore, $P(X = 0) = 1/8$, $P(X = 1) = 3/8$, $P(X = 2) = 3/8$, $P(X = 3) = 1/8$.

Thus, the probability mass function of X is given by the following table:

Table 1

Probability mass function of X in Example 35

X	0	1	2	3
p	1/8	3/8	3/8	1/8

Equivalently, the probability mass function can be written as

$$p_X(x) = \begin{cases} \frac{1}{8}, \ if \ x = 0 \\ \frac{3}{8}, \ if \ x = 1 \\ \frac{3}{8}, \ if \ x = 2 \\ \frac{1}{8}, \ if \ x = 3 \\ 0, \ otherwise \end{cases}$$

Let us verify the properties of a probability mass function. First, all probabilities are nonnegative. Second, the sum of all probabilities is equal to 1: 1/8 + 3/8 + 3/8 + 1/8 = 8/8 = 1.

Therefore, the obtained function is a valid probability mass function.

The probability mass function completely determines the probability distribution of a discrete random variable.

The instruction "Find the probability distribution of *X*" means that all possible values of the discrete random variable *X* and the probabilities associated with these values must be determined.

To find the probability distribution of *X*, one should:

1. identify all possible values that *X* can take;
2. calculate the probability of each possible value;
3. present the results in the form of a table or a formula.

Thus, finding the probability distribution of X is equivalent to determining the probability mass function of X at all its possible values. In practical problems, one is usually asked to find the probability distribution of a discrete random variable.

Example 36. An examination consists of three problems. A student solves the first, second, and third problems correctly with probabilities 0.8, 0.5, and 0.2, respectively. Assume that the results for the three problems are independent. Find the probability distribution of the number of problems solved correctly.

Solution.

Let random variable *X* denote the number of problems solved correctly by the student. Assume that the outcomes of solving the three problems are independent.

The random variable *X* can take the values 0, 1, 2, and 3. Let $p_1 = 0.8$, $p_2 = 0.5$, $p_3 = 0.2$ be the probabilities of solving the first, second, and third problems correctly. Then $q_1 = 1 - p_1 = 0.2$, $q_2 = 1 - p_2 = 0.5$, $q_3 = 1 - p_3 = 0.8$ are the corresponding probabilities of an incorrect solution.

Probability that no problem is solved correctly: $P(X = 0) = q_1 \cdot q_2 \cdot q_3 = 0.2 \cdot 0.5 \cdot 0.8 = 0.08$.

Probability that exactly one problem is solved correctly: $P(X = 1) = p_1 \cdot q_2 \cdot q_3 + q_1 \cdot p_2 \cdot q_3 + q_1 \cdot q_2 \cdot p_3 = 0.8 \cdot 0.5 \cdot 0.8 + 0.2 \cdot 0.5 \cdot 0.8 + 0.2 \cdot 0.5 \cdot 0.2 = 0.32 + 0.08 + 0.02 = 0.42$.

Probability that exactly two problems are solved correctly: $P(X = 2) = p_1 \cdot p_2 \cdot q_3 + p_1 \cdot q_2 \cdot p_3 + q_1 \cdot p_2 \cdot p_3 = 0.8 \cdot 0.5 \cdot 0.8 + 0.8 \cdot 0.5 \cdot 0.2 + 0.2 \cdot 0.5 \cdot 0.2 = 0.32 + 0.08 + 0.02 = 0.42$.

Probability that all three problems are solved correctly: $P(X = 3) = p_1 \cdot p_2 \cdot p_3 = 0.8 \cdot 0.5 \cdot 0.2 = 0.08$.

Therefore, the probability distribution of X is:

Table 2

Probability distribution of X in Example 36

X	0	1	2	3
p	0.08	0.42	0.42	0.08

Verification:

$0.08 + 0.42 + 0.42 + 0.08 = 1$.

Thus, the obtained table defines a valid probability distribution.

Among discrete random variables, several special types occur particularly frequently in probability theory and practical applications. Each of these random variables is characterized by a specific set of possible values and a particular form of probability distribution. We begin with the simplest case – the Bernoulli random variable – and then consider other important discrete distributions.

Bernoulli random variable.

Definition 20 [2, 17]. A Bernoulli random variable is a discrete random variable that takes only two possible values: 0 and 1. The value 0 represents a failure, while the value 1 represents a success.

If the probability of success is p, where $0 \le p \le 1$, then the probability of failure is $q = 1 - p$.

Thus, $P(X = 1) = p$, $P(X = 0) = 1 - p = q$.

A Bernoulli random variable with parameter p is denoted by $X \sim Bernoulli(p)$.

The probability distribution is given by the following table:

Table 3

Probability distribution of a Bernoulli random variable

X	0	1
p	$1 - p$	p

Example 37. An AI system generates a response to a user's question. Let X indicate whether the response meets the required quality standard:

- $X = 1$ if the response meets the quality standard;
- $X = 0$ if the response does not meet the quality standard.

Suppose that the probability that the response meets the quality standard is $p = 0.9$. Then $P(X = 1) = 0.9$, $P(X = 0) = 0.1$. Thus, X is a Bernoulli random variable with parameter $p = 0.9$: $X \sim Bernoulli(0.9)$.

Binomial Random Variable.

Definition 21 [6, 17]. A **binomial** random variable is a discrete random variable that represents the number of successes obtained in a fixed number n of independent Bernoulli trials, where each trial has the same probability of success p.

If X denotes the number of successes, then X can take the values 0, 1, 2, …, n.

The probability that X takes the value k is given by:

$$\boldsymbol{P(X = k) = C(n, k)p^k(1 - p)^{n-k}}$$

where $k = 0, 1, 2, \ldots, n$.

Here:

- n is the number of trials;
- k is the number of successes;
- p is the probability of success in each trial;
- $1 - p$ is the probability of failure in each trial;
- $C(n, k)$ is the number of ways to choose k successes from n trials.

A binomial random variable with parameters n and p is denoted by $X \sim Binomial(n, p)$.

A Bernoulli random variable is a special case of a binomial random variable with $n = 1$.

Example 38. An AI system generates 5 responses independently. The probability that any given response meets the required quality standard is 0.8. Let X be the number of responses that meet the standard.

Each response represents a Bernoulli trial with two possible outcomes:

- success — the response meets the standard;
- failure — the response does not meet the standard.

Since the five trials are independent and the probability of success is the same for each trial, X is a binomial random variable with parameters $n = 5$ and $p = 0.8$: $X \sim Binomial(5, 0.8)$.

Geometric random variable.

Definition 22 [2, 17]. A **geometric random variable** represents the number of independent Bernoulli trials required to obtain the first success, including the trial on which the first success occurs.

If X denotes the trial on which the first success occurs, then X can take the values1, 2, 3, … For X to take the value k, the first k - *1* trials must result in failure, and the k-th trial must result in success. The probability that X takes the value k is given by:

$$\boldsymbol{P(X = k) = (1 - p)^{k-1} p}$$

where $k = 1, 2, 3, \ldots$ and $0 < p \leq 1$.

A geometric random variable with parameter p is denoted by $X \sim Geometric(p)$.

Example 39. An AI system repeatedly generates responses until the first response that meets the required quality standard is obtained. Assume that the responses are generated independently and that the probability that any response meets the standard is $p = 0.7$.

Let X be the number of responses generated up to and including the first response that meets the standard. Then X is a geometric random variable with parameter $p = 0.7$: $X \sim Geometric(0.7)$.

Poisson random variable.

Definition 23 [15, 17]. A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$ if:

$$\boldsymbol{P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!},\ \lambda = np}$$

where $k = 0, 1, 2, \ldots$

The parameter λ represents the average number of events occurring within the specified interval; p - the probability of success.

A Poisson random variable with parameter λ is denoted by $X \sim Poisson(\lambda)$.

The Poisson distribution is commonly used to model the number of events occurring within a fixed interval of time, space, area, or volume when events occur at a constant average rate and independently over disjoint intervals.

Main conditions for applying the Poisson distribution:

- events occur independently of each other;
- the average intensity of events is constant;
- the probability of two or more events occurring simultaneously in a very small interval is negligible.

When a binomial distribution with parameters n and p is approximated by a Poisson distribution, the Poisson parameter is taken to be $\lambda = np$.

Example 40. An AI service receives user requests at an average rate of 4 requests per minute. Assume that requests arrive independently and at a constant average rate. Let X be the number of requests received during a one-minute interval. Then X is a Poisson random variable with parameter $\lambda = 4$, and $X \sim Poisson(4)$.

Hypergeometric Random Variable.

Definition 24 [15, 17]. A **hypergeometric** random variable is a discrete random variable that represents the number of successes obtained when a fixed number of

objects is selected **without replacement** from a finite population containing a known number of successes and failures.

Suppose that:

- *N* is the total number of objects in the population;
- *K* is the number of objects classified as successes;
- *n* is the number of objects selected;
- *k* is the number of successes among the selected objects.

Then *X* is said to have a hypergeometric distribution with parameters *N, K*, and *n*. This is denoted by *X ~ Hypergeometric(N, K, n).*

The probability that *X* takes the value *k* is given by:

$$\boldsymbol{P(X = k) = \frac{C(K, k) \cdot C(N - K, n - k)}{C(N, n)}}$$

Here:

- *C(K, k)* is the number of ways to select *k* successes from the *K* available successes;
- *C(N − K, n − k)* is the number of ways to select *n − k* failures from the *N − K* available failures;
- *C(N, n)* is the total number of ways to select n objects from the population of *N* objects.

Conditions for Using the Hypergeometric Distribution:

The hypergeometric distribution is applicable when:

1. the population is finite;
2. each object belongs to one of two categories, usually called success and failure;
3. a fixed number of objects is selected;
4. the objects are selected without replacement;
5. every sample of size n is equally likely to be selected.

Because the objects are selected without replacement, the outcomes of successive selections are not independent. After each selection, the composition of the remaining population changes.

As an example of a random variable having a hypergeometric distribution, we can consider the random variable *X* from Problem 32.

Let us consider some examples of solving problems.

Problem 32. An AI system has generated six responses, four of which meet the required quality standard. Three responses are selected uniformly at random without replacement. Let X denote the number of selected responses that meet the standard. Find the probability distribution of X.

Solution.

Let X denote the number of selected responses that meet the required quality standard.

Since only two responses fail to meet the standard, at least one of the three selected responses must meet it. Therefore, the possible values of X are 1, 2, 3.

The total number of ways to select three responses from six is $C(6, 3) = 20$.

Probability that exactly one selected response meets the standard (that is, one of them meets the condition, while the other two do not):

$$P(X = 1) = \frac{C(4,1) \cdot C(2,2)}{C(6,3)} = \frac{4}{20} = 0.2$$

Probability that exactly two selected responses meet the standard (that is, two of the selected responses meet the required quality standard, while the remaining one does not):

$$P(X = 2) = \frac{C(4,2) \cdot C(2,1)}{C(6,3)} = \frac{6 \cdot 2}{20} = 0.6$$

Probability that all three selected responses meet the standard:

$$P(X = 3) = \frac{C(4,3) \cdot C(2,0)}{C(6,3)} = \frac{4}{20} = 0.2$$

Therefore, the probability distribution of X is:

Table 4

Probability distribution of X in Problem 32

X	1	2	3
p	0.2	0.6	0.2

Verification:

0.2 + 0.6 + 0.2 = 1.

Thus, the obtained table defines a valid probability distribution.

Problem 33. A programmer repeatedly submits versions of a software module for automated testing. Each version passes the test independently with probability 0.8. Testing continues until the first version fails. Let X denote the total number of versions tested, including the first version that fails.

a) Determine the probability distribution of X.

b) Find the most probable value of X.

Solution.

The probability that a version passes the test is $p = 0.8$, and the probability that it fails is $q = 1 - p = 0.2$. For X to take the value k, the first $k - 1$ versions must pass the test, while the k-th version must fail. Therefore,

$$P(X = k) = 0.8^{k-1} \cdot 0.2$$

where k = 1, 2, 3, …

Thus, X has a geometric distribution with parameter 0.2: $X \sim Geometric(0.2)$.

Therefore, the probability distribution of X is:

Table 5

Probability distribution of X in Problem 33

X	1	2	3	…	k	…
p	0.2	0.16	0.128	…	$0.8^k \cdot 0.2$	…

The probabilities decrease as k increases. Therefore, the greatest probability corresponds to $k = 1$. Hence, the most probable number of versions tested is 1.

This means that the most likely outcome is that the first submitted version fails the test.

Answer: 1.

Tasks for independent work (homework).

1. An examination consists of three problems. A student solves the first, second, and third problems correctly with probabilities 0.8, 0.7, and 0.2, respectively. Assume that the results for the three problems are independent. Find the probability distribution of the number of problems solved correctly.

2. Two students are taking an examination in probability theory. Their probabilities of passing are 0.8 and 0.4, respectively. Assuming that their results are independent, let X denote the number of students who pass the examination. Determine the probability distribution of X.

3. A software team consists of three programmers working independently. The probability that any one of them introduces an error into their assigned module is 0.2. Let X denote the number of programmers who introduce an error. Find the probability distribution of X.

4. A software company estimates that 5% of newly developed modules contain a critical error. Four newly developed modules are inspected. Let X denote the number of selected modules that contain a critical error. Determine the probability distribution of X.

5. Write a binomial distribution law for a discrete random variable X – the number of occurrences of the “heads” in 3 tosses of a coin.

6. A software company tests eight modules. Assume that the test results are independent and that each module passes all quality checks with probability 0.8.

7. A software project contains 12 modules, 5 of which contain critical errors. Four modules are selected at random for inspection without replacement. Let X denote the number of selected modules that contain critical errors. Construct the probability distribution of X.

9. Cumulative Distribution Function and Numerical Characteristics of a Discrete Random Variable.

Definition 25 [6, 17]. The cumulative distribution function (CDF) of a random variable X is the function $F(x)$ defined by:

$$\boldsymbol{F(x) = P(X \leq x), x \in R.}$$

Thus, the CDF gives the probability that the random variable X takes a value less than or equal to x.

For a discrete random variable X with possible values $x_1, x_2, \ldots$, the cumulative distribution function is given by:

$$F_X(x) = \Sigma P(X = x_i)$$

where the summation is taken over all values x_i satisfying $x_i \leq x$.

Equivalently, if $p_i = P(X = x_i)$, then $F_X(x) = \Sigma p_i, \quad x_i \leq x$.

The subscript X may be omitted when no ambiguity arises.

The cumulative distribution function of a discrete random variable is a step function.

Properties of the Cumulative Distribution Function [6, 17].

1. *Boundedness.* For every real number x, $0 \leq F(x) \leq 1$.
2. *Monotonicity.* The CDF is a nondecreasing function. Thus, if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$.
3. *Limits at infinity*:

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

4. *Probability of an interval.* For any real numbers a and b such that $a < b$:

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

$$P(a \leq X \leq b) = F(b) - F(a) + P(X = a)$$

6. *Probability at a single point.* The probability at a specific value is equal to the magnitude of the jump in the cumulative distribution function (CDF) at that value:

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

where $F(x-)$ denotes the left-hand limit of F at x.

7. *Bounded Support.* If all possible values of X lie in the interval (a, b), then $F(x) = 0$ for $x < a$ and $F(x) = 1$ for $x \geq b$.

Example 41. Let X denote the number of critical errors found in a randomly selected software module. Its probability distribution is given in Table 6. Construct the cumulative distribution function of X.

Table 6

The probability distribution for Example 41

X	0	1	2	3
p	0.2	0.4	0.3	0.1

Solution.

The cumulative distribution function of X is defined by:

$$F(x) = P(X \leq x).$$

We calculate its values by successively adding the probabilities from the distribution table.

For $x < 0$, which is below the smallest possible value of X, none of the possible values satisfies $X \leq x$. Therefore, $F(x) = 0$.

For $0 \leq x < 1$, only the value $X = 0$ is included: $F(x) = P(X = 0) = 0.2$.

For $1 \leq x < 2$, the values $X = 0$ and $X = 1$ are included: $F(x) = P(X = 0) + P(X = 1) = 0.2 + 0.4 = 0.6$.

For $2 \leq x < 3$, the values $X = 0$, $X = 1$, and $X = 2$ are included: $F(x) = 0.2 + 0.4 + 0.3 = 0.9$.

For $x \geq 3$, all possible values of X are included: $F(x) = 0.2 + 0.4 + 0.3 + 0.1 = 1$.

Thus, the cumulative distribution function is

$$F(x) = \begin{cases} 0, x < 0; \\ 0.2, 0 \leq x < 1; \\ 0.6, 1 \leq x < 2; \\ 0.9, 2 \leq x < 3; \\ 1, x \geq 3. \end{cases}$$

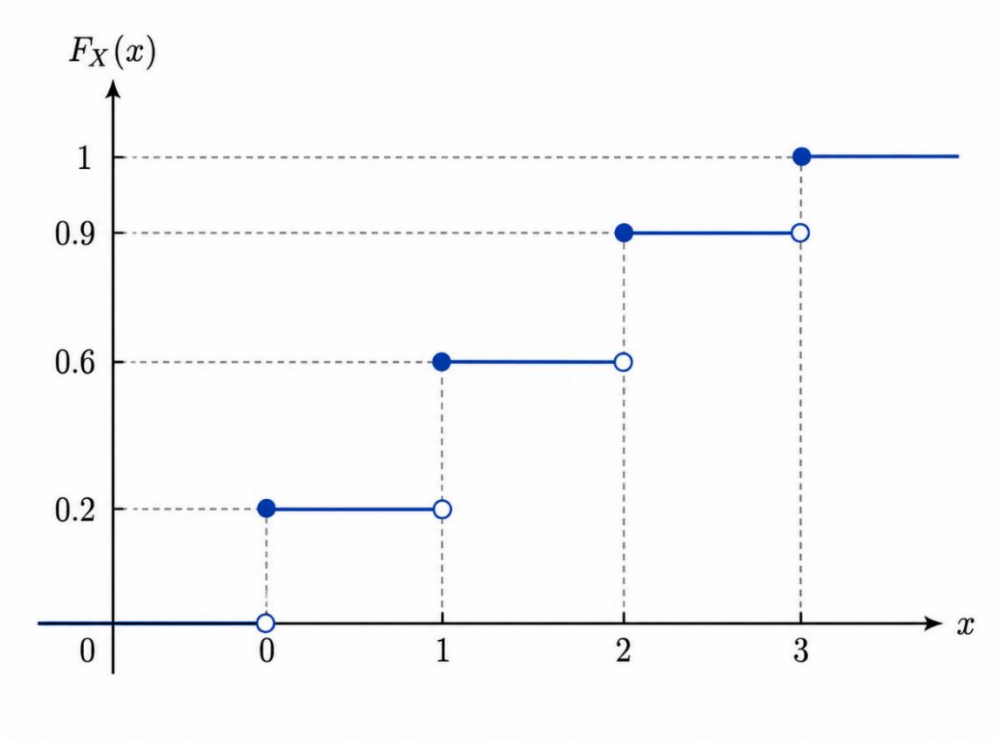


Figure 25. Cumulative distribution function of X in Example 41

Definition 26 [6, 17]. Let X be a discrete random variable with possible values $x_1, x_2, x_3, \ldots$ and corresponding probabilities $p_i = P(X = x_i)$. The **expected value**, or **expectation**, of X is defined by:

$$E(X) = x_1 p_1 + \cdots + x_n p_n + \cdots = \sum_i x_i p_i$$

where the summation is taken over all possible values of X.

The expected value is a probability-weighted average of the possible values of X. Under repeated independent realizations of the same experiment, it can be interpreted as the long-run average value of X. The expected value need not be one of the possible values of X.

Properties of Expected Value [15, 17].

1. *Expected Value of a Constant.* If C is a constant, then $E(C) = C$.

2. *Multiplication by a Constant.* For any constant C, $E(C \cdot X) = C \cdot E(X)$.

3. *Additivity.* For any random variables X and Y, $E(X + Y) = E(X) + E(Y)$. More generally, $E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$.

4. *Expected Value of a Product of Independent Random Variables*. If X and Y are independent, then $E(XY) = E(X) \cdot E(Y)$. More generally, if $X_1, X_2, \ldots, X_n$ are mutually independent, then $E(X_1 X_2 \ldots X_n) = E(X_1) \cdot E(X_2) \cdot \ldots \cdot E(X_n)$. Independence is essential for this property.

Example 42. Find the expected value of the random variable X from Example 41.

Solution.

The expected value of X is

$E(X) = \Sigma\, x \cdot P(X = x) = 0 \cdot 0.2 + 1 \cdot 0.4 + 2 \cdot 0.3 + 3 \cdot 0.1 = 0 + 0.4 + 0.6 + 0.3 = 1.3$.

Thus, $E(X) = 1.3$.

Answer: 1.3.

Expected Values of Common Discrete Distributions [15, 17].

Below are the derivations of the expected values of the discrete random variables considered earlier.

1. Bernoulli Random Variable.

Let $X \sim Bernoulli(p)$. The random variable X takes the values 0 and 1 with probabilities $P(X = 0) = 1 - p$, $P(X = 1) = p$. By the definition of expected value, $E(X) = 0 \cdot (1 - p) + 1 \cdot p = p$.

Thus, $\boldsymbol{E(X) = p.}$

This result is natural because a Bernoulli random variable is equal to 1 when success occurs and 0 otherwise. Therefore, its expected value is equal to the probability of success.

2. Binomial Random Variable.

Let $X \sim Binomial(n, p)$, where X represents the number of successes in n independent Bernoulli trials. Let $q = 1 - p$.

$$E(X) = \sum_{k=0}^{n} k \cdot C(n,k) p^k q^{n-k} = np \cdot \sum_{k=1}^{n} C(n-1, k-1) p^{k-1} q^{n-k}$$

Let $j = k - 1$. Then j takes the values $0, 1, \ldots, n - 1$. Hence,

$$E(X) = np \cdot \sum_{j=0}^{n-1} C(n-1, j) p^j q^{n-1-j}$$

By the binomial theorem, $\sum_{j=0}^{n-1} C(n-1, j) p^j q^{n-1-j} = (p+q)^{n-1}$. Since $p + q = 1$, $(p+q)^{n-1} = 1$.

Therefore, ***E(X) = np.***

3. Geometric Random Variable.

Let $X \sim Geometric(p)$, where X represents the number of trials required to obtain the first success, including the successful trial.

The possible values of X are 1, 2, 3, … Let $q = 1 - p$.

By the definition of expected value,

$$E(X) = \sum_{k=1}^{\infty} k \cdot q^{k-1} p = p \cdot \sum_{k=1}^{\infty} k \cdot q^{k-1}$$

For $|q| < 1$, the geometric series is:

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$$

Differentiating both sides with respect to q gives:

$$\sum_{k=1}^{\infty} k \cdot q^{k-1} = \frac{1}{(1-q)^2}$$

Therefore,

$$E(X) = \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}.$$

Thus,

$$\boldsymbol{E(X) = \frac{1}{p}}$$

The expected number of trials required to obtain the first success is the reciprocal of the probability of success.

4. Poisson Random Variable.

Let $X \sim Poisson(\lambda)$, where $\lambda > 0$. By the definition of expected value,

$$E(X) = \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda} \lambda^k}{k!} = \sum_{k=1}^{\infty} k \cdot \frac{e^{-\lambda} \lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^k}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

Let $j = k - 1$. Then

$$E(X) = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}$$

Using the exponential-series expansion,

$$\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{\lambda}$$

Therefore,

$$E(X) = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

Thus, ***E(X) = λ.***

The parameter λ is both the parameter of the Poisson distribution and the expected number of events occurring in the specified interval.

5. Hypergeometric Random Variable.

Let $X \sim Hypergeometric(N, K, n)$. The expected value can be derived using indicator random variables.

For each selected object, define $I_i = 1$ if the ith selected object is a success, $I_i = 0$ otherwise.

Then $X = I_1 + I_2 + \ldots + I_n$. For every selection position, the probability of selecting a success is $P(I_i = 1) = K / N$.

Therefore, $E(I_i) = 1 \cdot K/N + 0 \cdot (1 - K/N) = K/N$.

Using linearity of expectation,

$E(X) = E(I_1 + I_2 + \ldots + I_n) = E(I_1) + E(I_2) + \ldots + E(I_n) = K/N + K/N + \ldots + K/N$.

There are n terms, so

E(X) = n · K/N.

The selected objects are not independent because sampling is performed without replacement. However, independence is not required for the linearity of expectation.

The expected number of successes in the sample is therefore equal to the sample size multiplied by the proportion of successes in the population.

Example 43. Example 43. A software company tests 20 modules. Assume that the test outcomes are independent and that each module passes all quality checks with probability 0.85.

Solution.

Each module test is an independent Bernoulli trial with probability of success $p = 0.85$. Since X counts the number of successful outcomes in 20 independent trials, $X \sim Binomial(20, 0.85)$. The expected value of a binomial random variable is

$$E(X) = n \cdot p = 20 \cdot 0.85 = 17.$$

This means that, on average, 17 out of 20 modules are expected to pass all quality checks when the testing procedure is repeated many times.

Answer: 17.

Definition 27 [6, 20]. Let X be a discrete random variable with possible values $x_1, x_2, x_3, \ldots$ and corresponding probabilities $p_i = P(X = x_i)$. The **variance** of X is defined as the expected value of the squared deviation of X from its expected value:

$$\boldsymbol{Var(X) = E[(X - E(X))^2]}$$

For a discrete random variable, this definition can be written as

$$Var(X) = \sum_i (x_i - E(X))^2 \cdot p_i$$

where the summation is taken over all possible values of X.

The variance can also be calculated using the computational formula

$$\boldsymbol{Var(X) = E(X^2) - [E(X)]^2,}$$

where $\boldsymbol{E(X^2) = \Sigma_i\, x_i^2 \cdot p_i}$.

The variance measures the degree to which the possible values of a random variable are spread around its expected value. A small variance indicates that the values of X tend to be close to $E(X)$, whereas a large variance indicates greater variability.

The variance is measured in the square of the original units. For this reason, the **standard deviation** of X is defined by

$$\boldsymbol{\sigma(X) = \sqrt{Var(X)}}$$

The standard deviation is measured in the same units as the random variable.

Properties of Variance [6, 20].

1. *Nonnegativity.* For every random variable X, $Var(X) \geq 0$. This follows from the fact that $(X - E(X))^2 \geq 0$.

2. *Variance of a Constant.* If C is a constant, then $Var(C) = 0$. A constant random variable has no variability because it always takes the same value.

3. *Multiplication by a Constant.* For any constant C, $Var(C \cdot X) = C^2 \cdot Var(X)$. Thus, multiplying a random variable by a constant multiplies its variance by the square of that constant.

4. *Variance of a Sum of Independent Random Variables.* If X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$. More generally, if $X_1, X_2, \ldots, X_n$ are mutually independent, then $Var(X_1 + X_2 + \ldots + X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n)$. Independence is essential for this property.

Example 44. Find the variance and standard deviation of the random variable X from Example 41.

Solution. From Example 42, $E(X) = 1.3$. First, calculate $E(X^2)$:

$$E(X^2) = 0^2 \cdot 0.2 + 1^2 \cdot 0.4 + 2^2 \cdot 0.3 + 3^2 \cdot 0.1 = 0 + 0.4 + 1.2 + 0.9 = 2.5.$$

Using the computational formula, $Var(X) = E(X^2) - [E(X)]^2 = 2.5 - 1.3^2 = 2.5 - 1.69 = 0.81$.

The standard deviation is $\sigma(X) = \sqrt{Var(X)} = \sqrt{0.81} = 0.9$.

Answer: $Var(X) = 0.81$; $\sigma(X) = 0.9$

Variances of Common Discrete Distributions [15, 17].

Below are the derivations of the variances of the discrete random variables considered earlier. Here we will omit the derivation of the formulas and leave only the final result.

1. ***Bernoulli Random Variable.***

$$Var(X) = p \cdot q$$

2. ***Binomial Random Variable***

$$Var(X) = n \cdot p \cdot q$$

3. ***Geometric Random Variable***

$$Var(X) = \frac{q}{p^2}$$

4. ***Poisson Random Variable***

$$Var(X) = \lambda$$

5. Hypergeometric Random Variable

$$Var(X) = n \cdot \frac{K}{N} \cdot \left(1 - \frac{K}{N}\right) \cdot (N - n)/(N - 1)$$

Example 45. A software company tests 20 independently developed modules. Each module passes all quality checks with probability 0.85. Let X denote the number of modules that pass all quality checks. Find the variance and standard deviation of X.

Solution.

Each module test is an independent Bernoulli trial with probability of success $p = 0.85$. Since X counts the number of successful outcomes in 20 independent trials, $X \sim Binomial(20, 0.85)$. For a binomial random variable, $Var(X) = n \cdot p \cdot (1 - p) = 20 \cdot 0.85 \cdot (1 - 0.85) = 20 \cdot 0.85 \cdot 0.15 = 2.55$.

The standard deviation is $\sigma(X) = \sqrt{2.55} \approx 1.6$.

Answer: $Var(X) = 2.55$; $\sigma(X) = 1.6$.

Let's consider some examples of solving problems.

Problem 34. For the random variable X from Problem 32, find the expected value, variance, and standard deviation. Construct the cumulative distribution function (CDF) of X.

Solution.

Let's use the probability distribution obtained in Problem 32:

X	1	2	3

p	0.2	0.6	0.2

First, calculate the expected value: $E(X) = 1 \cdot 0.2 + 2 \cdot 0.6 + 3 \cdot 0.2 = 2$.

Next, calculate $E(X^2)$: $E(X^2) = 1^2 \cdot 0.2 + 2^2 \cdot 0.6 + 3^2 \cdot 0.2 = 4.4$.

Using the computational formula for variance, $Var(X) = E(X^2) - [E(X)]^2 = 4.4 - 2^2 = 4.4 - 4 = 0.4$.

The standard deviation is $\sigma(X) = \sqrt{Var(X)} = \sqrt{0.4} \approx 0.63$.

The cumulative distribution function is

$$F(x) = \begin{cases} 0, x < 1; \\ 0.2, 1 \le x < 2; \\ 0.8, 2 \le x < 3; \\ 1, x \ge 3. \end{cases}$$

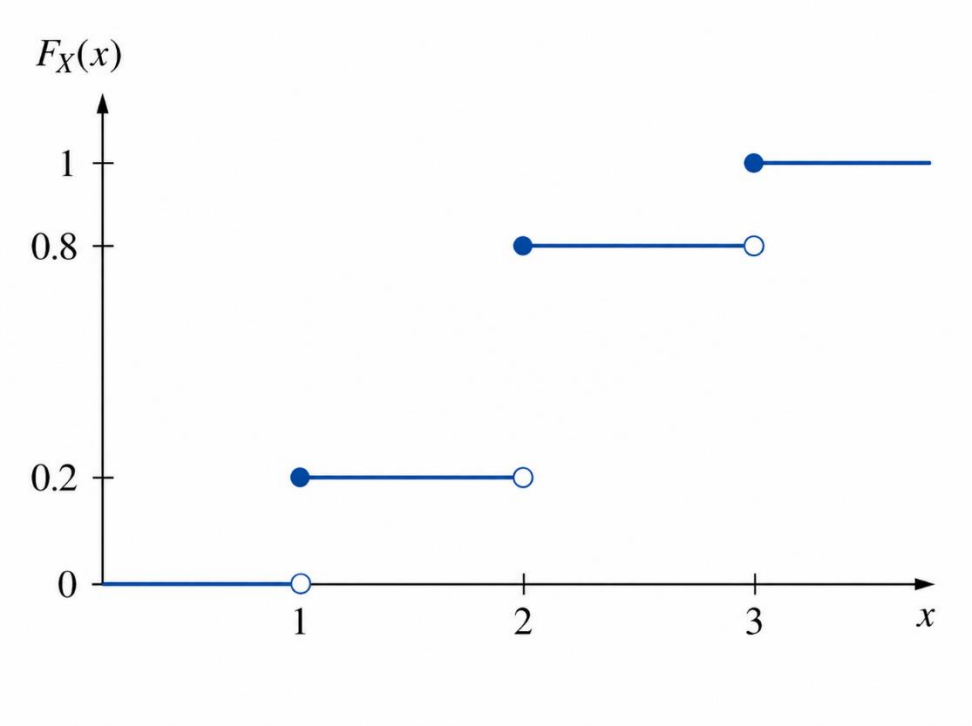


Figure 26. Cumulative distribution function of X in Problem 34

Answer: $E(X) = 2$; $Var(X) = 0.4$; $\sigma(X) = 0.63$.

Problem 35. A programmer repeatedly runs an automated testing tool until the tool detects a software defect. The probability that the tool detects a defect during any given run is 0.25. The runs are independent. Let X denote the number of runs required to detect the first defect. Find $P(X = 4)$, $P(X \le 5)$, $E(X)$, and $Var(X)$.

Solution.

The random variable X represents the number of trials required to obtain the first success. Therefore, $X \sim Geometric(0.25)$. Here, $p = 0.25$, $q = 1 - p = 0.75$.

- For the first defect to be detected on the fourth run, the tool must fail to detect a defect during the first three runs and detect one on the fourth run: $P(X = 4) = 0.75^3 \cdot 0.25 = 0.10546875 \approx 0.11$.

- The event $X \leq 5$ means that the first defect is detected during one of the first five runs. It is easier to use the complementary event: $P(X \leq 5) = 1 - P(X > 5)$. The event $X > 5$ means that the tool does not detect a defect during any of the first five runs. Therefore, $P(X \leq 5) = 1 - 0.75^5 = 1 - 0.2373046875 = 0.7626953125 \approx 0.76$.
- For a geometric random variable, $E(X) = 1/p$. Therefore, $E(X) = 1/0.25 = 4$.
- For a geometric random variable, $Var(X) = q/p^2$. Therefore, Var(X) = $0.75/0.25^2 = 0.75/0.0625 = 12$.

Answer: 0.11; 0.76; 4; 12.

Tasks for independent work (homework).

1. For each random variable X introduced in the independent-work problems in Section 8, find its expected value, variance, and standard deviation, and construct its cumulative distribution function.

2. The probability distribution of the random variable X is given in the following table:

Table 7

The probability distribution of the random variable X

X	-1	1	3	5
p	0.1	0.5	0.2	0.2

Find the expected value, variance, and standard deviation of X. Construct the cumulative distribution function (CDF) of X and plot its graph.

3. A zoology quiz consists of four questions about animals. Each question has five possible answers, only one of which is correct. A student answers each question by guessing independently and at random. Let X denote the number of correct answers. Find the probability distribution of X, its expected value, and its variance. Construct the cumulative distribution function (CDF) of X.

4. Discrete random variable X takes three possible values: $x_1 = 5$ with a probability $p_1 = 0.5$; $x_2 = 2$ with a probability $p_2 = 0.2$ and x_3 with a probability p_3. Find x_3 and p_3 knowing that $E(X) = 10$.

5. A wildlife camera is used on five independent nights. The probability that it photographs a fox on any given night is 0.2. Let X denote the number of nights on which a fox is photographed. Find the variance of X.

6. An AI system generates responses to two independent user queries. The probability that a generated response meets the required quality standard is the same for both queries. Let X denote the number of responses that meet the quality standard. Given that $E(X) = 1.2$, find the variance of X.

7. A programmer repeatedly runs a debugging tool until it detects an error. The probability that the tool detects an error during any given run is 0.4. The runs are independent. Let X denote the number of runs required to detect the first error. Find $P(X = 3)$, $P(X \leq 4)$, $E(X)$, and $Var(X)$.

8. Two independent development teams test different software components. The number X of components that pass testing in the first team has the distribution $X \sim Binomial(4, 0.75)$, and the number Y of components that pass testing in the second team has the distribution $Y \sim Binomial(6, 0.75)$. Let $Z = X + Y$ be the total number of components that pass testing. Determine the probability distribution of Z. Find $E(Z)$, $Var(Z)$, $\sigma(Z)$, and $P(Z \geq 8)$.

10. Continuous Random Variables.

Definition 28 [6, 17]. A random variable X is called **continuous** if there exists a nonnegative function $f(x)$, called the **probability density function (PDF)** of X, such that

$$P(X \in B) = \int_B f(x)dx$$

for every measurable set $B \subseteq \mathbb{R}$.

In particular, for any real numbers $a < b$, $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Since X is continuous, $P(X = a) = P(X = b) = 0$. Therefore,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

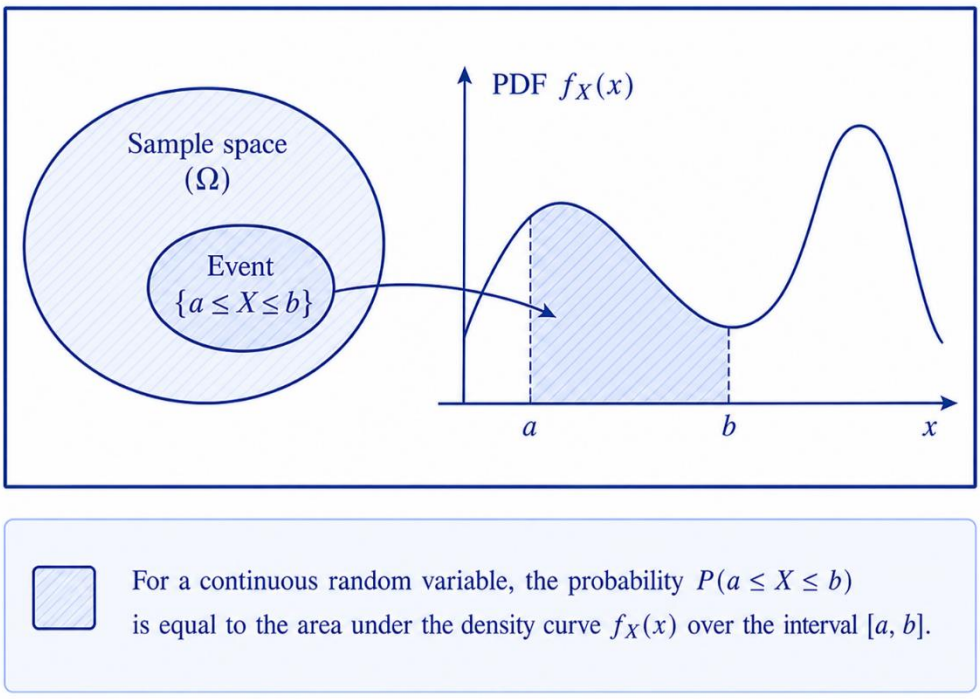


Figure 27. Probability density function (PDF)

For a continuous random variable *X*, probabilities are represented by areas under the graph of its probability density function *f(x)*.

The total area under the probability density curve is equal to 1:

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

It is important to note that the value *f(x)* itself is not a probability. Probability is obtained by calculating the area under the density curve over an interval.

Properties of the Probability Density Function [6, 15, 17].

Let *X* be a continuous random variable with probability density function *f(x)*. The probability density function has the following properties.

1. *Non-Negativity.* For every real number x, $f(x) \geq 0$.
2. *Total Area.* The total area under the graph of the probability density function is equal to 1: $\int_{-\infty}^{+\infty} f(x)dx = 1$
3. *Probability of an Interval.* For any real numbers $a < b$, the probability that *X* takes a value in the interval *[a, b]* is equal to the area under the density curve over this interval: $P(a \leq X \leq b) = \int_a^b f(x)dx$.
4. *Probability at a Single Point.* For every real number *x*, $P(X = x) = 0$. Therefore, the inclusion or exclusion of the endpoints does not affect the probability: $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$.
5. ***Relationship with the Cumulative Distribution Function.*** The cumulative distribution function of *X* is $\boldsymbol{F(x) = \int_{-\infty}^{x} f(t)dt}$**.** At every point where *F(x)* is differentiable, $\boldsymbol{f(x) = F'(x)}$**.**

Example 46. Let *X* be a continuous random variable with the probability density function

$$f(x) = \begin{cases} 2x, 0 \leq x \leq 1 \\ 0, \text{otherwise} \end{cases}$$

First, verify that *f(x)* is a valid probability density function. For all *x*, $f(x) \geq 0$.

Also, $\int_{-\infty}^{+\infty} f(x)dx = \int_0^1 2xdx = x^2|_0^1 = 1$.

Therefore, *f(x)* is a valid probability density function.

Definition 29 [15, 17]. Let *X* be a continuous random variable. The **cumulative distribution function (CDF)** of X is defined by

$$\boldsymbol{F(x) = P(X \leq x), x \in \mathbb{R}}$$

The cumulative distribution function (CDF) represents the area under the probability density curve to the left of the point *x*.

For any real numbers $a < b$,

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

Since a continuous random variable takes any particular value with probability zero, P(X = a) = P(X = b) = 0.

Therefore,

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

Properties of the Cumulative Distribution Function [6, 15, 17]

The cumulative distribution function of a continuous random variable has the following properties.

1. *Boundedness.* For every real number x, $0 \leq F(x) \leq 1$.
2. *Monotonicity***.** The function $F(x)$ is non-decreasing. Thus, if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$.
3. *Limits at Infinity.* $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$
4. *Continuity.* The cumulative distribution function of a continuous random variable is continuous.
5. *Relationship with the Probability Density Function.*

The cumulative distribution function is obtained by integrating the probability density function:

$$\boldsymbol{F(x) = \int_{-\infty}^{x} f(t)dt}$$

$$\boldsymbol{f(x) = F'(x).}$$

6. *Probability as an Area.*

For any $a < b$,

$$\boldsymbol{F(b) - F(a) = \int_{a}^{b} f(x)dx}$$

Thus, the increase in the cumulative distribution function from a to b is equal to the area under the density curve over the interval *[a, b]*.

Example 47. Let X have the probability density function

$$f(x) = \begin{cases} 2x, 0 \leq x \leq 1 \\ 0, \text{otherwise} \end{cases}$$

Find the cumulative distribution function of X.

Solution.

By definition, $F(x) = \int_{-\infty}^{x} f(t)dt$. For $x < 0$, the density is zero, so $F(x) = 0$. For $0 \leq x \leq 1$, $F(x) = \int_0^x 2tdt = t^2|_0^x = x^2$**.** For $x > 1$, the entire area under the density curve has been accumulated, so $F(x) = 1$.

Therefore,

$$F(x) = \begin{cases} 0, x < 0 \\ x^2, 0 \leq \mathrm{x} \leq 1 \\ 1, x > 1 \end{cases}$$

Expected Value, Variance, and Standard Deviation of a Continuous Random Variable [6, 15, 20].

Expected Value.

Let *X* be a continuous random variable with probability density function *f(x)*. The **expected value** of *X* is defined by:

$$\boldsymbol{E(X) = \int_{-\infty}^{+\infty} xf(x)dx}$$

The expected value represents the probability-weighted average of all possible values of *X*. It can also be interpreted as the long-run average value of *X* over a large number of repetitions of the random experiment.

The properties of variance are the same for discrete and continuous random variables.

Variance.

The **variance** of a continuous random variable *X* is defined as the expected value of the squared deviation of *X* from its expected value (the same for discrete and continuous random variables):

$$\boldsymbol{Var(X) = E[(X - E(X))^2].}$$

The variance can also be calculated using the computational formula

$$\mathbf{Var(X) = E(X^2) - (E(X))^2,}$$

where $\boldsymbol{E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx.}$

The variance measures how widely the possible values of *X* are spread around the expected value. A small variance indicates that the values tend to be close to the mean, whereas a large variance indicates greater variability.

The properties of expected value are the same for discrete and continuous random variables.

Standard Deviation

The standard deviation of a continuous random variable *X* is defined as the positive square root of its variance (the same for discrete and continuous random variables):

$$\boldsymbol{\sigma = \sqrt{Var(X)}.}$$

The variance is measured in squared units, while the standard deviation is measured in the same units as the random variable. Therefore, the standard deviation is often easier to interpret than the variance.

Example 48. Let X be a continuous random variable with the probability density function

$$f(x) = \begin{cases} 2x, 0 \le x \le 1 \\ 0, \text{otherwise} \end{cases}$$

- *Expected Value.* $E(X) = \int_0^1 x \cdot 2x dx = 2\int_0^1 x^2 dx = 2 \cdot \frac{x^3}{3}|_0^1 = \frac{2}{3}$.
- *Variance.* $Var(X) = E(X^2) - [E(X)]^2$.

$$E(X^2) = \int_0^1 x^2 \cdot 2x dx = 2\int_0^1 x^3 dx = 2 \cdot \frac{x^4}{4}|_0^1 = \frac{1}{2}$$

$Var(X) = E(X^2) - [E(X)]^2 = ½ - (2/3)^2 = 1/2 - 4/9 = 1/18.$

- *Standard Deviation.* $\sigma = \sqrt{(1/18)} \approx 0.236$.

Answer: $E(X) = 2/3$; $Var(X) = 1/18$; $\sigma \approx 0.236$.

We now consider several important continuous probability distributions.

Continuous Uniform Distribution [20].

Definition 30. A continuous random variable X is said to have a **uniform distribution** on the interval $[a, b]$, where $a < b$, if its probability density function is constant on this interval and equal to zero outside it.

This is denoted by $X \sim Uniform(a, b)$.

The **probability density function** of X is:

$$\boldsymbol{f(x) = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & a \le x \le b \\ 0, & x > b \end{cases}}$$

The constant $\frac{1}{b-a}$ is chosen so that the total area under the density curve is equal to 1: $\int_a^b \frac{1}{b-a} dx = 1$.

The graph of the probability density function is a horizontal line of height $\frac{1}{b-a}$ over the interval $[a, b]$ and is equal to zero outside this interval.

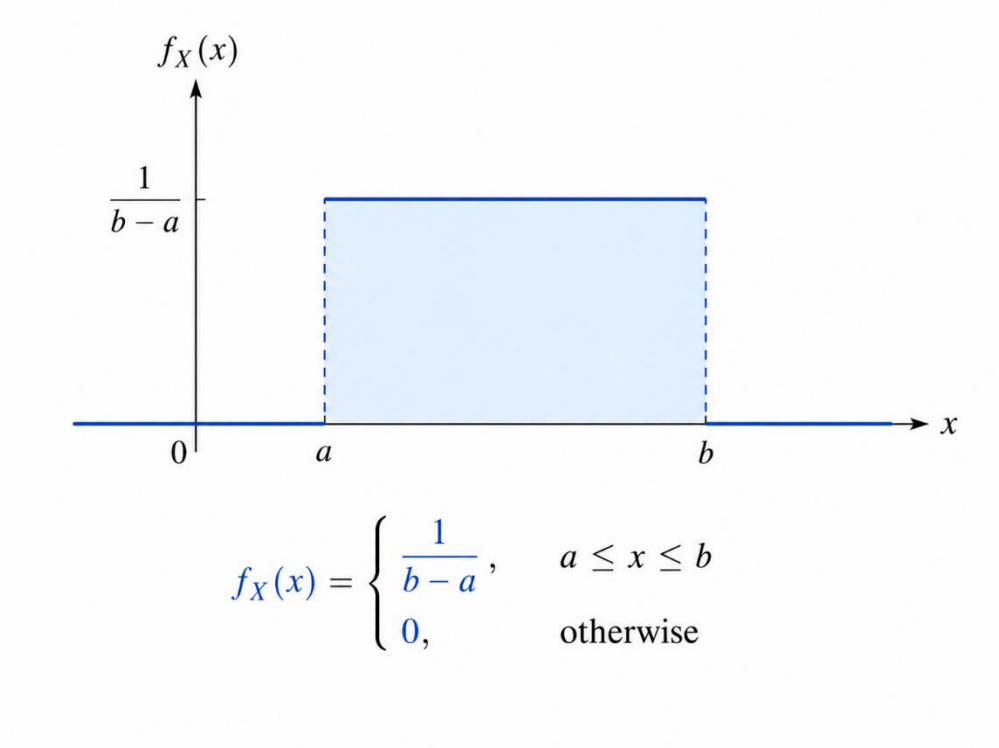


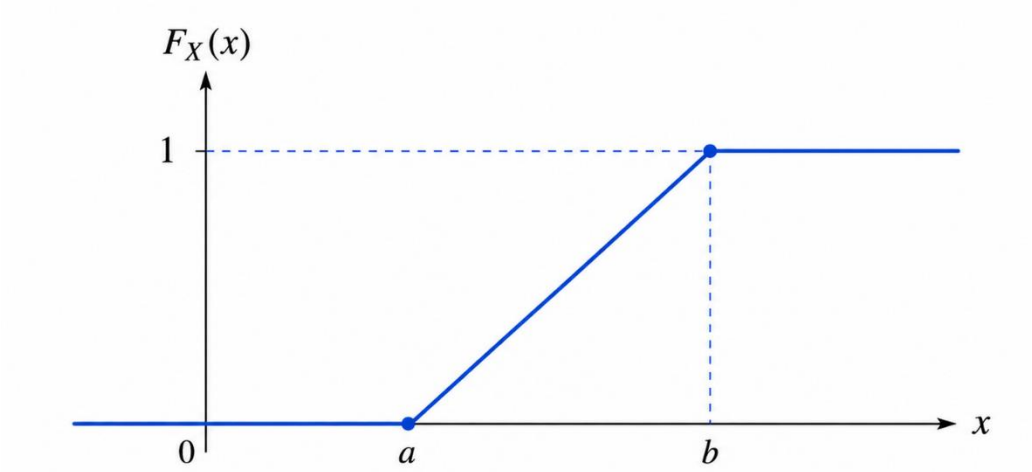
Figure 28. Probability density function of a uniform distribution

For a uniformly distributed random variable, all subintervals of equal length contained in *[a, b]* have the same probability. The probability that X belongs to an interval is equal to the length of that interval divided by the total length of *[a, b]*.

The **cumulative distribution function** of *X* is defined by:

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

The CDF is continuous and increases linearly from 0 to 1 over the interval *[a, b]*.



$$F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$$

Figure 29. Cumulative distribution function of a uniform distribution

Expected Value.

The expected value of a uniformly distributed random variable is $E(X) = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a}\int_a^b x dx = \frac{1}{b-a} \cdot \frac{x^2}{2}|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$.

Thus, $\boldsymbol{E(X) = \frac{a+b}{2}}$.

Variance.

The variance of X is $Var(X) = E(X^2) - [E(X)]^2$.

First, calculate $E(X^2)$: $E(X^2) = \frac{1}{b-a}\int_a^b x^2 dx = \frac{1}{b-a} \cdot \frac{x^3}{3}|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(a^2+ab+b^2)}{3(b-a)} = \frac{a^2+ab+b^2}{3}$.

$$Var(X) = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4}$$
$$= \frac{4a^2+4ab+4b^2-3a^2-6ab-3b^2}{12} = \frac{b^2-2ab+a^2}{12}$$
$$= \frac{(b-a)^2}{12}.$$

Thus,

$$\boldsymbol{Var(X) = \frac{(b-a)^2}{12}}.$$

Standard Deviation.

The standard deviation is $\sigma = \sqrt{Var(X)} = \sqrt{\frac{(b-a)^2}{12}} = \frac{b-a}{2\sqrt{3}} = \frac{\sqrt{3}(b-a)}{6}$.

So, $\boldsymbol{\sigma = \frac{\sqrt{3}(b-a)}{6}}$.

Example 49. The response time of an AI system is uniformly distributed between 2 and 6 seconds. Let *X* denote the response time. Thus, *X ~ Uniform(2, 6)*.

- The probability density function is

$$f(x) = \begin{cases} 0, & x < 2 \\ \frac{1}{4}, & 2 \le x \le 6 \\ 0, & x > 6 \end{cases}$$

- Find the probability that the response time is between 3 and 5 seconds. CDF:

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

Using the formula for the uniform distribution, P(3 ≤ X ≤ 5) = F(5) - F(3) = ¾ - ¼ = ½.

- The expected response time is $E(X) = \frac{2+6}{2} = 4$ seconds.
- The variance is $Var(X) = \frac{(6-2)^2}{12} = \frac{16}{12} = \frac{4}{3}$.
- The standard deviation is $\sigma = \sqrt{\frac{4}{3}} \approx 1.155$.

Normal Distribution [14, 17].

Definition 31 [17]. A continuous random variable *X* is said to have a normal distribution with parameters μ and σ^2, where $\mu \in R$ and $\sigma > 0$, if its probability density function is

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

This is denoted by $X \sim Normal(\mu, \sigma^2)$ or, more briefly, $X \sim N(\mu, \sigma^2)$.

The parameter $\boldsymbol{\mu}$ **is the expected value of** ***X*****, and the parameter** $\boldsymbol{\sigma^2}$ **is its variance.** The positive number σ is the standard deviation.

Thus, $\boldsymbol{E(X) = \mu, Var(X) = \sigma^2}$.

The graph of the probability density function is called the **normal curve** or the **bell-shaped curve**.

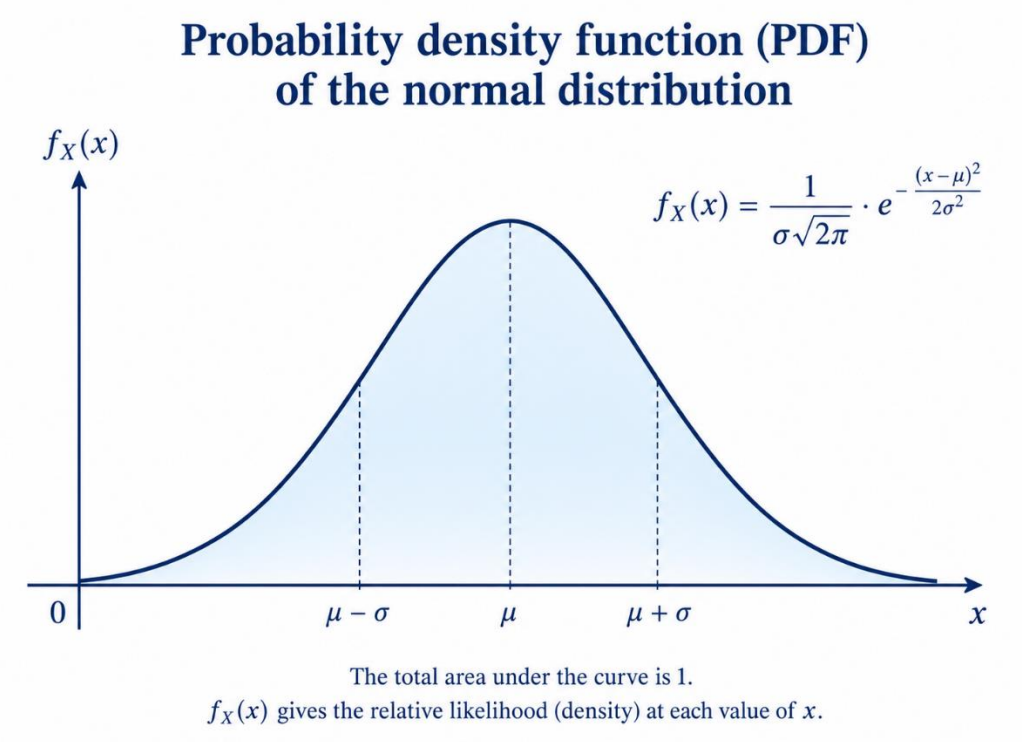


Figure 30. Probability density function (PDF) of Normal Distribution

The parameter μ determines the location of the normal curve. Increasing μ shifts the entire curve to the right, while decreasing μ shifts it to the left. The shape and spread of the curve do not change.

The standard deviation σ determines the spread of the distribution.

- A smaller value of σ produces a narrower and higher curve.
- A larger value of σ produces a wider and lower curve.

Regardless of the values of μ and σ, the total area under the curve remains equal to 1.

The cumulative distribution function of a normal random variable X is ***F(x)*** = ***P(X ≤ x)*** and is given by:

$$F(X) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

This integral cannot generally be expressed using elementary functions. Therefore, probabilities involving a normal random variable are usually calculated using standard normal distribution tables, statistical software, or a calculator.

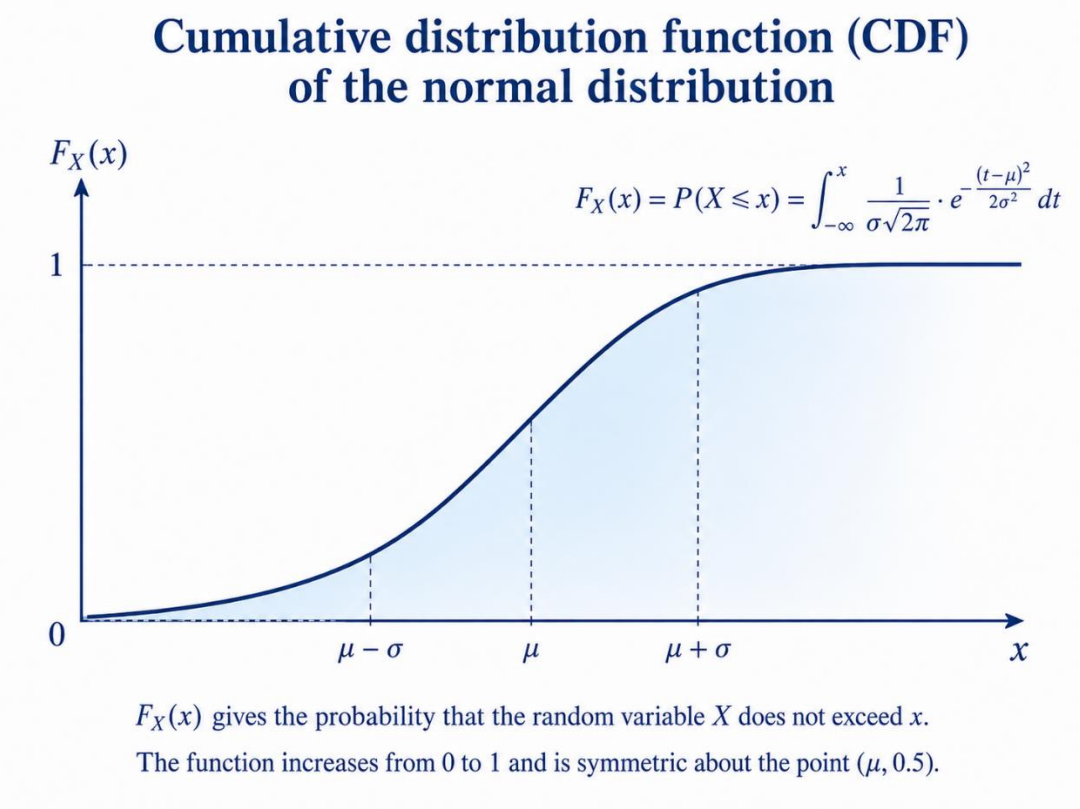


Figure 31. Cumulative distribution function (CDF) of Normal Distribution

Standard Normal Distribution [14, 17].

Definition 32 [17]. A normal random variable Z with expected value 0 and variance 1 is called a **standard normal random variable**. It is denoted by $Z \sim N(0, 1)$.

The Empirical Rule.

For a normally distributed random variable, approximately:

$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.6827;$

$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545;$

$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973.$

Thus:

- approximately 68.27% of the values lie within one standard deviation of the mean;
- approximately 95.45% lie within two standard deviations;
- approximately 99.73% lie within three standard deviations.

This is often called the **68–95–99.7 rule** or the **empirical rule**.

The Three-Sigma Rule. Let X be a normally distributed random variable with expected value μ and standard deviation σ: $X \sim N(\mu, \sigma^2)$. The **three-sigma rule** states that approximately 99.73% of all possible values of X lie within three standard deviations of the expected value:

$$P(|X - \mu| < 3\sigma) \approx 0.9973.$$

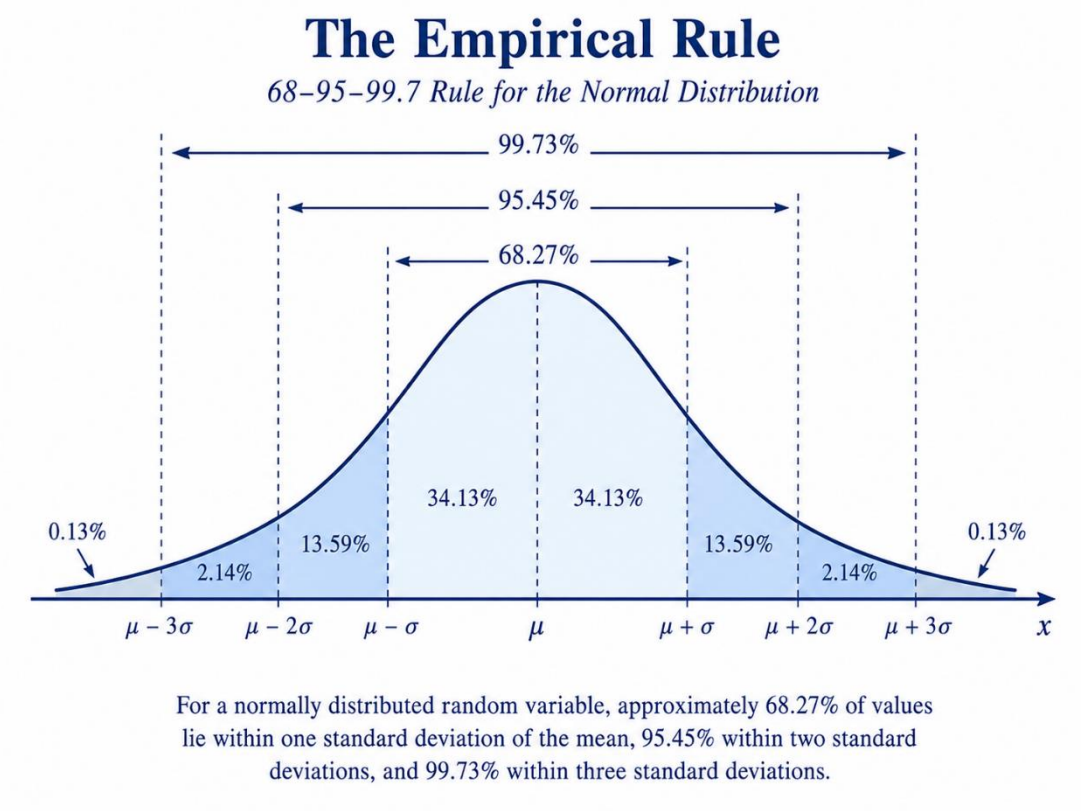


Figure 32. The Empirical Rule

Probabilities for a Normally Distributed Random Variable Using the Laplace Function.

Let X be a normally distributed random variable with expected value μ and standard deviation σ: $X \sim N(\mu, \sigma^2)$, where $E(X) = \mu$, $Var(X) = \sigma^2$, and $\sigma > 0$.

Let α and β be real numbers such that $\alpha < \beta$. The probability that X takes a value between α and β:

$$\boldsymbol{P(\alpha \le X \le \beta) = \Phi\left(\frac{\beta - \mu}{\sigma}\right) - \Phi\left(\frac{\alpha - \mu}{\sigma}\right)}$$

This formula gives the area under the normal density curve between the points α and β. In this formula, $\Phi(x)$ denotes the Laplace function. Its properties were given earlier in Section 7, and the table of its values is provided in Appendix II.

Probability of a Symmetric Deviation from the Mean.

Consider the event $|X - \mu| < \delta$, where $\delta > 0$.

This event means that the value of X differs from its expected value μ by less than δ. The inequality $|X - \mu| < \delta$ is equivalent to $-\delta < X - \mu < \delta$, or $\mu - \delta < X < \mu + \delta$. Therefore, $P(|X - \mu| < \delta) = P(\mu - \delta < X < \mu + \delta)$.

Using the interval probability formula, we have:

$$\boldsymbol{P(|X - \mu| < \delta) = 2\Phi(\delta/\sigma)}$$

This formula gives the probability that a normally distributed random variable lies within a distance δ of its mean.

Example 50. The response time X of an AI system is normally distributed with an expected value of 5 seconds and a standard deviation of 0.8 seconds: $X \sim N(5, 0.8^2)$. Find the probability that the response time differs from its expected value by less than 1 second.

Solution.

We need to calculate $P(|X - 5| < 1)$. Here, $\mu = 5$, $\sigma = 0.8$, $\delta = 1$.

Using the formula, $P(|X - \mu| < \delta) = 2\Phi(\delta/\sigma)$, we obtain $P(|X - 5| < 1) = 2\Phi(1/0.8) = 2\Phi(1.25)$.

From the Laplace function table, $\Phi(1.25) \approx 0.3944$. Therefore, $P(|X - 5| < 1) \approx 2 \cdot 0.3944 = 0.7888$.

Thus, the probability that the AI response time is between 4 and 6 seconds is approximately 0.79.

Answer: 0.79.

Let us now consider some problems.

Problem 36. The CDF of the random variable X has the form:

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{33}(2x^2 + 5x), & 0 \le x \le 3, \\ 1, & x > 3. \end{cases}$$

Find *PDF; E(X); var(X);* $\sigma\,(X)$, $P(1 < X < 2)$; plot graphs of CDF and PDF.

Solution.

1. *Probability density function.*

For a continuous random variable, $f(x) = F'(x)$.

Differentiating on the interval $0 < x < 3$, we obtain $f(x) = [(2x^2 + 5x) / 33]' = (4x + 5) / 33$.

Thus, the probability density function is

$$f(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{33}(4x + 5), & 0 \le x \le 3, \\ 0, & x > 3. \end{cases}$$

2. *Expected value.*

The expected value is $E(X) = \int_0^3 x \cdot \frac{1}{33}(4x + 5)dx = \frac{1}{33}\left(\frac{4x^3}{3} + \frac{5x^2}{2}\right)|_0^3 = \frac{1}{33}\left(36 + \frac{45}{2}\right) = \frac{1}{33} \cdot \frac{117}{2} = \frac{39}{22} \approx 1.77.$

3. *Variance.*

To calculate the variance, first find $E(X^2)$:

$$E(X^2) = \int_0^3 x^2 \cdot \frac{1}{33}(4x+5)dx = \frac{1}{33}\left(\frac{4x^4}{4} + \frac{5x^3}{3}\right)\Big|_0^3 = \frac{1}{33}(81+45) = \frac{126}{33} = \frac{42}{11}$$

Using the computational formula, $Var(X) = E(X^2) - [E(X)]^2$. Therefore, $Var(X) = 42/11 - (39/22)^2 \approx 0.68$.

4. *Standard deviation.*

The standard deviation is $\sigma(X) = \sqrt{Var(X)} = \sqrt{(327/484)} \approx 0.82$.

5. Probability $P(1 < X < 2)$

Using the cumulative distribution function,

$$P(1 < X < 2) = F(2) - F(1) = \frac{1}{33}(8+10) - \frac{1}{33}(2+5) = \frac{1}{3}.$$

6. *Graph of CDF.*

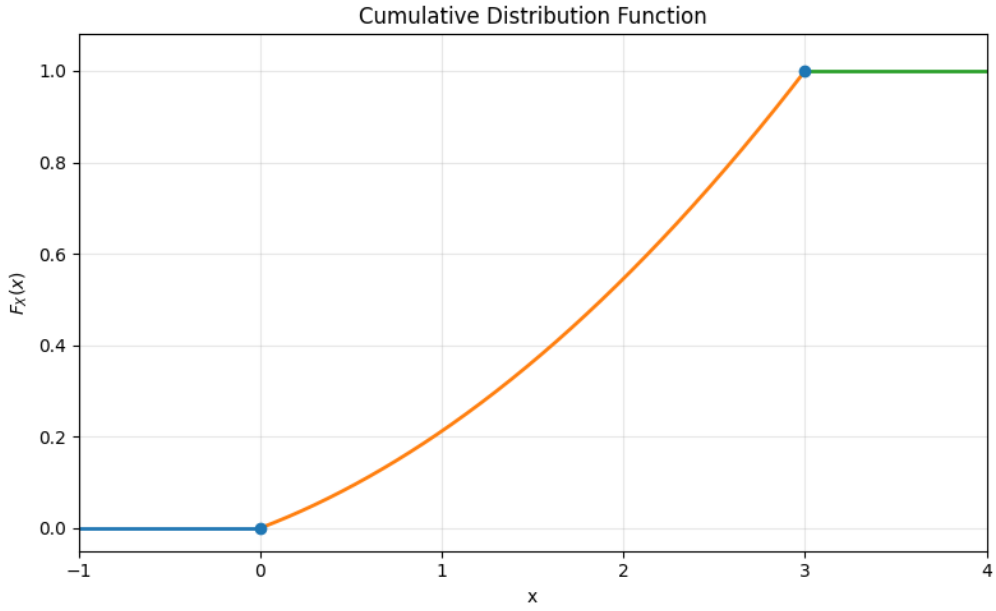


Figure 33. Cumulative distribution function in Problem 36

Graph of PDF.

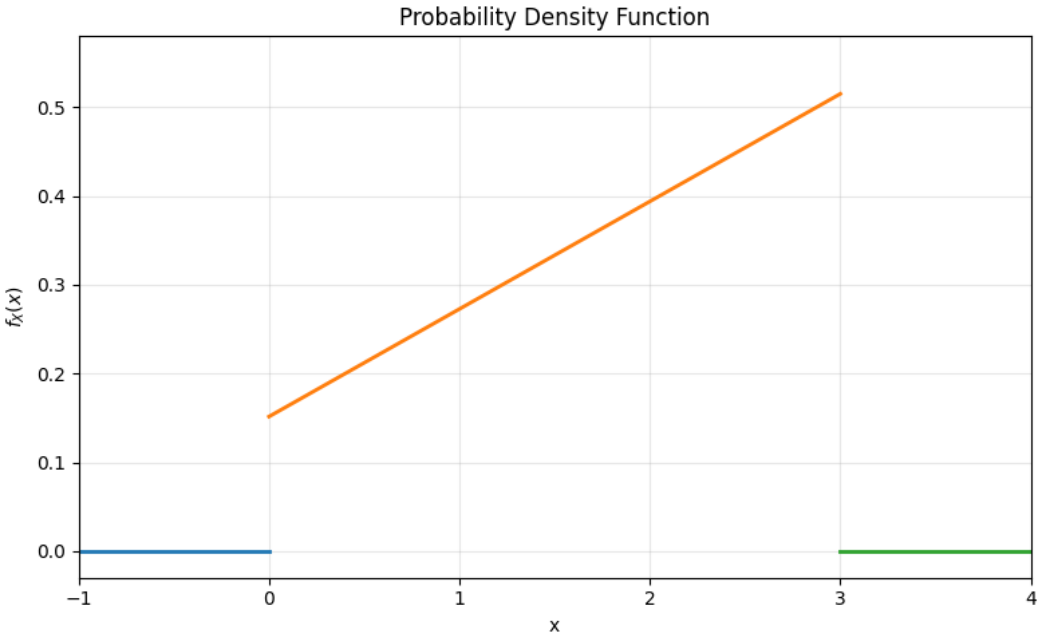


Figure 34. Probability density function in Problem 36

Problem 37. The probability density function of a continuous random variable X has the form $f(x) = \frac{3C}{4+x^2}, -\infty < x < \infty$. Find the constant parameter C.

Solution.

For $f(x)$ to be a probability density function, it must satisfy the normalization condition: $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Therefore, $\int_{-\infty}^{+\infty} \frac{3C}{4+x^2}dx = 1$. $3C\int_{-\infty}^{+\infty} \frac{1}{4+x^2}dx = 3C \cdot \frac{1}{2} arctan\frac{x}{2}|_{-\infty}^{+\infty} = 3C \cdot \frac{1}{2} \cdot \frac{\pi}{2} - 3C \cdot \frac{1}{2} \cdot \left(-\frac{\pi}{2}\right) = 3C \cdot \frac{\pi}{2} = 1 \Rightarrow C = \frac{2}{3\pi}$.

Answer: $C = \frac{2}{3\pi}$.

Tasks for independent work (homework).

1. The cumulative distribution function of a random variable X is defined for all real numbers by $F(x) = 1/2 + (1/\pi) \cdot arctan(x/2)$. Find the probability that X takes a value in the interval (0, 2).
2. The probability density function of a continuous random variable X has the form $f(x) = \frac{4C}{9+x^2}, -\infty < x < \infty$. Find the constant parameter C.
3. The PDF of the random variable X has the form:

$$f(x) = \begin{cases} 0, & x < 1, \\ A\sqrt{x}, & 1 \le x \le 4, \\ 0, & x > 4. \end{cases}$$

Find *A; P(2<X<3); CDF; E(X); Var(X); σ(X);* plot graphs of CDF and PDF.

4. A machine produces metal rods. The length *X* of a rod is normally distributed with an expected value equal to the design length of 120 mm. Practically all manufactured rods have lengths between 96 mm and 144 mm. Using the three-sigma rule, find the probability that the length of a randomly selected rod:

a) is greater than 126 mm;

b) is less than 108 mm.

5. The cumulative distribution function of a continuous random variable *X* is given by

$$F(x) = \begin{cases} 0, & x \le 0, \\ \sin\left(\frac{\pi x}{4}\right), & 0 < x \le 2, \\ 1, & x > 2. \end{cases}$$

Find:

- the probability density function *f(x)*;
- *E(X);*
- *Var(X);*
- *σ(X);*
- *P(0.5<X<1.5);*
- plot the graphs of the CDF and PDF.

6. The probability density function of a continuous random variable *X* has the form

$$f(x) = \begin{cases} A(x+1), & 0 \le x \le 2, \\ 0, & otherwise. \end{cases}$$

Find:

- the constant *A*;
- the cumulative distribution function *F(x);*
- *P(0.5<X<1.5);*
- *E(X);*
- *Var(X);*
- *σ(X);*
- plot the graphs of the CDF and PDF.

7. The cumulative distribution function of a continuous random variable *X* is given by

$$F(x) = \begin{cases} 0, & x < 0, \\ \dfrac{x^2 + 2x}{15}, & 0 \le x \le 3, \\ 1, & x > 3. \end{cases}$$

Find:

- the probability density function *f(x)*;
- *E(X);*
- *Var(X);*
- *σ(X);*
- *P(1<X<2);*
- plot the graphs of the CDF and PDF.

8. A continuous random variable *X* is uniformly distributed on the interval [5; 15]. Find the probability that *X* falls within the interval (7; 11).

9. A continuous random variable *X* is uniformly distributed on the interval [3; 11].

Find:

- the probability density function *f(x)*;
- the cumulative distribution function *F(x);*
- the expected value *E(X);*
- the variance *Var(X);*
- the standard deviation *σ*;
- the probability *P(5 < X < 9);*
- plot the graphs of the PDF and CDF.

10. The probability density function of a continuous random variable X is given by:

$$f(x) = \begin{cases} \dfrac{3x^2}{8}, & 0 \le x \le 2, \\ 0, & otherwise. \end{cases}$$

Find:

- the cumulative distribution function *F(x);*
- *P(0.5<X<1.5);*
- *E(X);*
- *Var(X);*
- *σ(X);*
- plot the graphs of the CDF and PDF.

11. The operating time X of a computer component is normally distributed with an expected value of 50 hours and a standard deviation of 4 hours. Find:

- $P(46 < X < 54)$;
- $P(X > 58)$;
- $P(X < 44)$;
- $P(|X - 50| < 6)$.

12. A continuous random variable X is uniformly distributed on an unknown interval $[a, b]$. It is known that $E(X) = 5$ and $Var(X) = 3$. Find:

- the endpoints a and b;
- the probability density function $f(x)$;
- the cumulative distribution function $F(x)$;
- the standard deviation $\sigma(X)$;
- $P(3 < X < 6)$;
- $P(X > 7)$;
- plot the graphs of the PDF and CDF.

13. The response time X of an information system is normally distributed with an expected value of 100 milliseconds. It is known that $P(92 < X < 108) \approx 0.9545$. Find:

- the standard deviation σ;
- $P(96 < X < 104)$;
- $P(X > 106)$;
- $P(X < 88)$;
- $P(|X - 100| \geq 12)$.

14. The cumulative distribution function of a continuous random variable X has the form:

$$F(x) = \begin{cases} 0, & x < 0, \\ C(3x^2 - x^3), & 0 \leq x \leq 2, \\ 1, & x > 2. \end{cases}$$

Find:

- the constant C;
- the probability density function $f(x)$;
- verify that $f(x)$ is a valid probability density function;
- $E(X)$;
- $E(X^2)$;
- $Var(X)$;
- $\sigma(X)$;
- $P(0.5 < X < 1.5)$;
- plot the graphs of the CDF and PDF.

11. Mathematical Statistics. Laboratory Work 1 “Basic Methods of Statistical Data Analysis Using Python”.

Definition 33 [15]. Mathematical statistics is a branch of mathematics concerned with the collection, organization, analysis, and interpretation of data and with drawing conclusions about populations on the basis of observations and experimental results.

Mathematical statistics uses sample data to study the properties of populations and random phenomena.

The main tasks of mathematical statistics include:

- collecting, describing, organizing, and presenting statistical data;
- selecting an appropriate probability distribution or statistical model for the observed data;
- estimating unknown parameters of the population distribution;
- testing statistical hypotheses about the population and assessing whether the observed data are consistent with a proposed theoretical model.

Definition 34 [15]. The complete set of objects, individuals, measurements, or observations under study is called the **population.**

The population may be finite or infinite. It may consist of physically existing objects or of all possible observations that could be obtained under specified conditions.

Definition 35 [15]. The number of elements in a finite population is called the **population size** and is usually denoted by N. For an infinite or conceptual population, a finite population size N is not specified.

Definition 36 [15]. A **sample** is a collection of observations obtained from a population for statistical analysis. The number of observations in the sample is called the **sample size** and is usually denoted by n. A random sample of size n from a distribution F is a collection of independent random variables $X_1, X_2, \ldots, X_n$, each having distribution F. After the observations have been obtained, their realized values are denoted by $x_1, x_2, \ldots, x_n$.

Definition 37 [15]. A sample is called **representative** if it adequately reflects the characteristics of the population that are relevant to the study. Representativeness is not guaranteed by random sampling: even a properly selected random sample may differ from the population because of sampling variability.

Definition 38 [14]. The **sample mean** is the sum of all values in the dataset divided by the total number of values. It represents the average value of the dataset.

Definition 39 [14]. A **mode** of a dataset is a value that occurs with the greatest frequency. A dataset may have one mode, several modes, or no unique mode.

Definition 40 [14]. The **sample median** is the middle value of a dataset arranged in ascending or descending order. If the number of observations is odd, the median

is the value located in the middle of the ordered dataset. If the number of observations is even, the median is calculated as the arithmetic mean of the two middle values.

Definition 41 [14]. The population **variance** is the mean squared deviation of the population values from the population mean:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

The **sample variance** is defined by

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

The divisor $n - 1$ is used when the sample variance serves as an estimator of the population variance.

Definition 42 [14]. Standard deviation is the positive square root of the variance. It measures how far the data values are typically located from their mean and is expressed in the same units as the original data.

A small standard deviation indicates that the observations are concentrated close to the mean, whereas a large standard deviation indicates that the observations are more widely dispersed. If greater consistency and homogeneity of the data are desirable, a smaller standard deviation is generally preferable. However, the interpretation of standard deviation always depends on the context of the study.

Definition 43 [6, 15]. A **quantile of order *p*** is a value that divides an ordered dataset or a probability distribution so that approximately a proportion p of the observations lies at or below that value. Quantiles describe the relative position of values within a distribution. The median is the 0.5-quantile, while the first and third quartiles are the 0.25- and 0.75-quantiles, respectively.

Definition 44 [6, 15]. The **interquartile range**, denoted by **IQR**, is a measure of dispersion equal to the difference between the third quartile and the first quartile: IQR = Q3 − Q1. It describes the spread of the middle 50% of the observations. Since it does not directly depend on the smallest and largest values, the interquartile range is less sensitive to extreme observations than the ordinary range.

Definition 45 [6, 15]. The **coefficient of variation**, denoted by **CV**, is a relative measure of dispersion calculated as the ratio of the standard deviation to the mean. It is usually expressed as a percentage: *CV = standard deviation / mean · 100%.* The coefficient of variation shows the size of the standard deviation relative to the mean and can be used to compare the variability of datasets measured in different units or having substantially different mean values. A larger coefficient of variation indicates greater relative variability.

The coefficient of variation is most meaningful for variables measured on a ratio scale with a meaningful zero and a positive mean. It may be misleading when the mean is close to zero or negative.

Definition 46 [6, 15]. Covariance is a measure of the joint variability of two numerical variables. It indicates whether the deviations of the variables from their respective means tend to have the same direction or opposite directions.

A positive covariance indicates that the two variables generally increase or decrease together. A negative covariance indicates that an increase in one variable tends to be associated with a decrease in the other. A covariance equal or close to zero may indicate little linear association, but its magnitude cannot be interpreted without considering the units and scales of the variables.

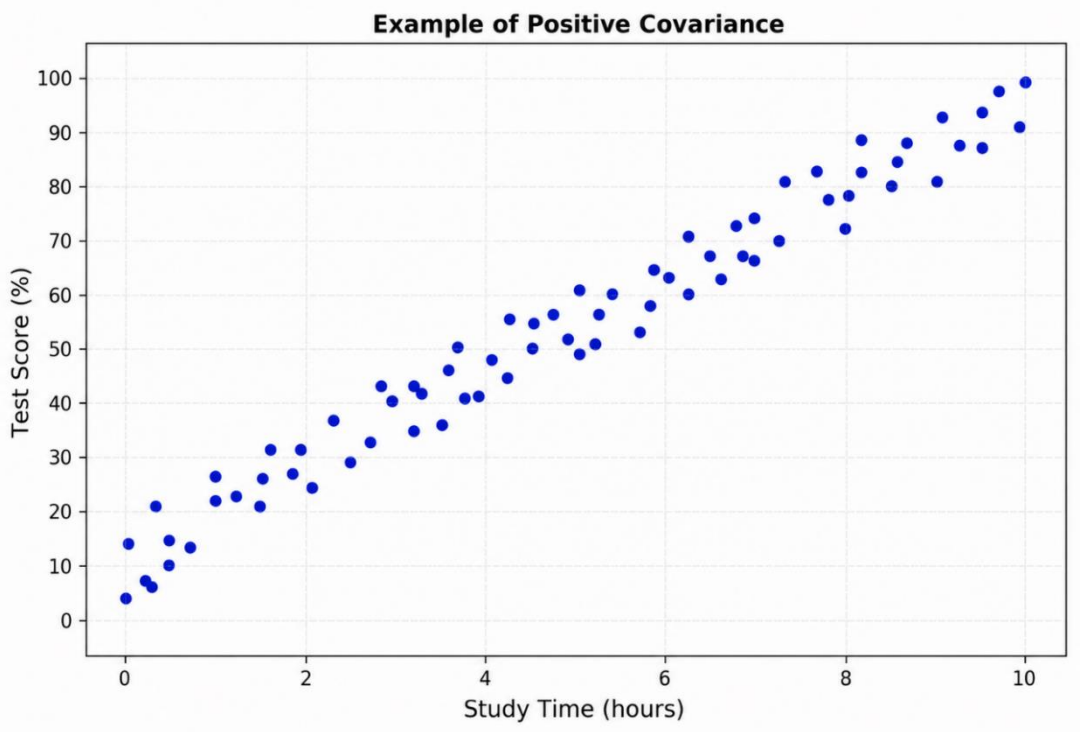


Figure 35. Example of positive covariance

Figure 35 shows a scatter plot illustrating the relationship between two numerical variables, X and Y. Each point represents a pair of observed values. The points form a clear upward pattern: as the value of X increases, the value of Y also tends to increase. This indicates a strong positive covariance between the variables.

Although the points do not lie exactly on a straight line, most of them are concentrated around an increasing linear trend. Therefore, the variables have a strong positive linear relationship.

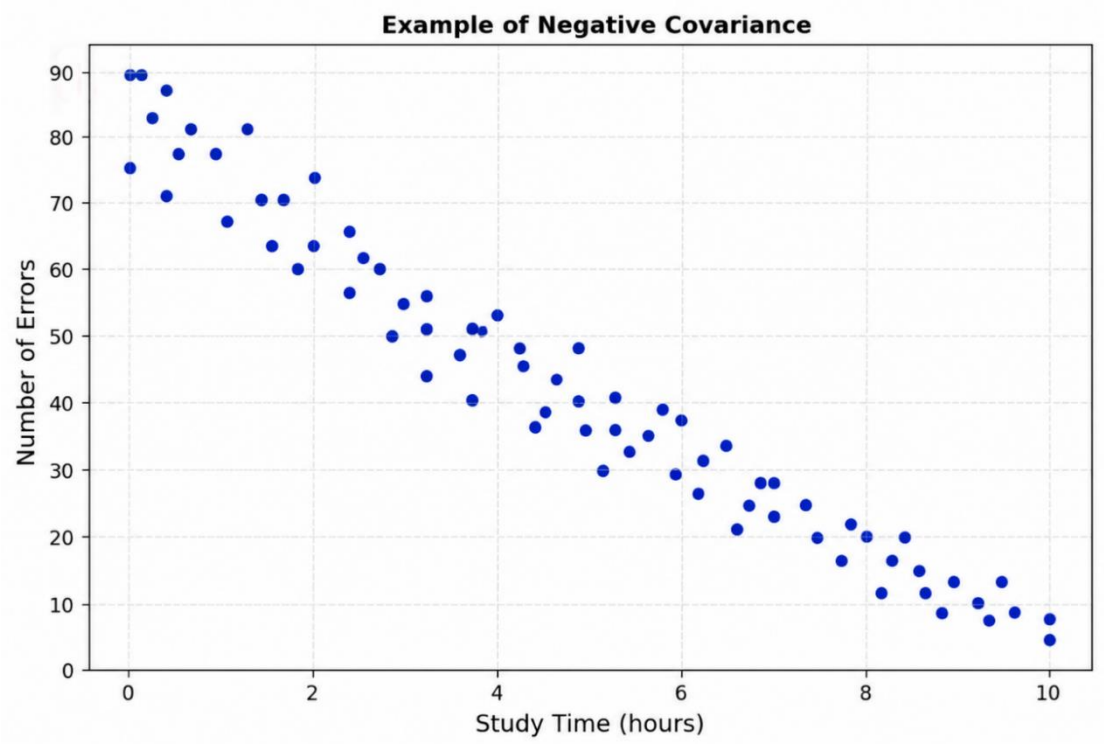


Figure 36. Example of negative covariance

The figure presents a scatter plot illustrating the relationship between study time and the number of errors. Each point represents one observation. The points form a clear downward pattern: as study time increases, the number of errors tends to decrease. This indicates a strong negative covariance between the two variables.

Although the points do not lie exactly on a straight line, most of them are concentrated around a decreasing linear trend. Therefore, the variables demonstrate a strong negative linear relationship. However, this relationship does not by itself prove that an increase in study time directly causes the reduction in the number of errors.

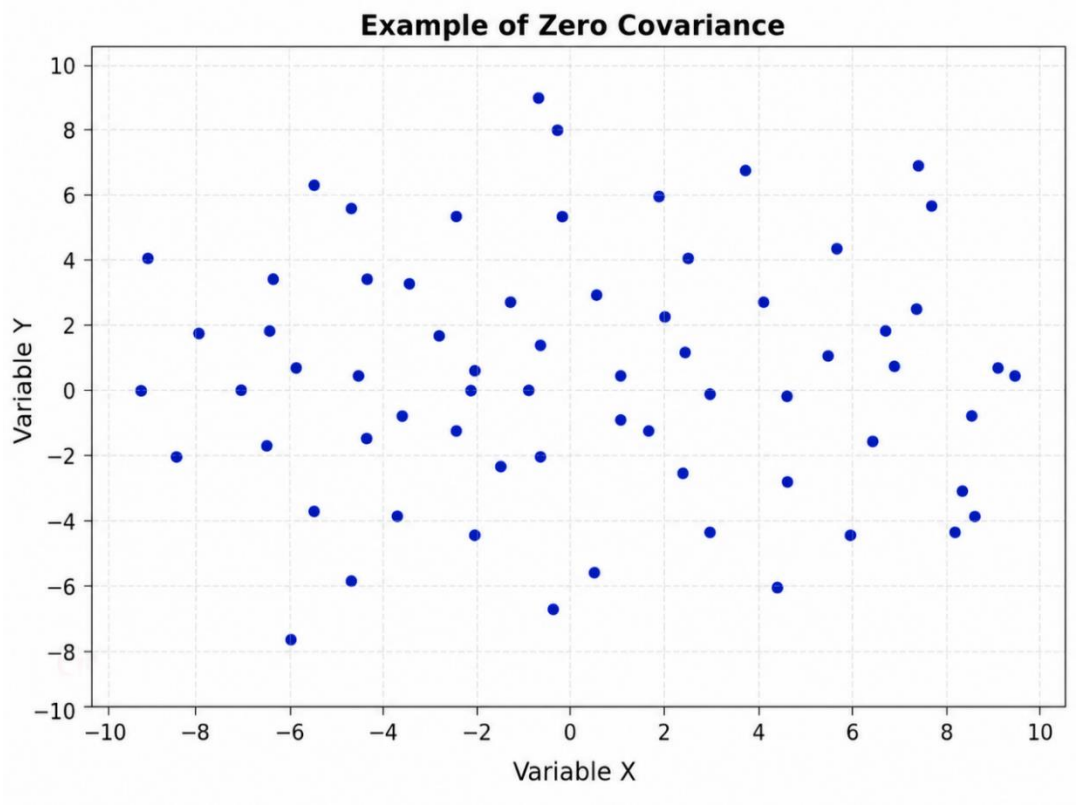


Figure 37. Example of covariance close to zero

The figure presents a scatter plot illustrating the relationship between two numerical variables, *X* and *Y*. Each point represents one pair of observed values. The points are scattered without a clear upward or downward pattern. As the value of *X* increases, the value of *Y* does not show a systematic tendency to increase or decrease.

This distribution of points represents covariance close to zero and indicates the absence of a linear relationship between the variables. However, zero covariance does not necessarily mean that the variables are completely independent, since a nonlinear relationship may still exist.

Covariance indicates the direction in which two numerical variables change together. However, the magnitude of covariance depends on the units in which the variables are measured. For this reason, covariance values obtained for different pairs of variables cannot be compared directly.

To overcome this limitation, covariance is standardized by dividing it by the product of the standard deviations of the two variables. The resulting measure is called the Pearson correlation coefficient. Unlike covariance, the correlation coefficient does not depend on the units of measurement and always takes values from −1 to 1. Therefore, it describes not only the direction but also the strength of the linear relationship between two numerical variables.

Definition 47 [15, 20]. The Pearson correlation coefficient, denoted by r, is a standardized numerical measure of the strength and direction of the linear relationship between two numerical variables. Its value ranges from −1 to 1.

A value close to 1 indicates a strong positive linear relationship, a value close to −1 indicates a strong negative linear relationship, and a value close to 0 indicates a weak or absent linear relationship.

Definition 48 [15, 20]. A **percentile** is a number that divides ordered data into hundredths. The k-th percentile is a value at or below which approximately k% of the observations lie. Percentiles describe the relative position of an observation within an ordered dataset.

The study of mathematical statistics in this course will be supported by laboratory work involving statistical data analysis and the implementation of statistical methods using the Python programming language [10, 13].

Laboratory Work 1 "Basic Methods of Statistical Data Analysis Using Python".

1. The first step of the laboratory work is to generate a random dataset using the Python programming language. The generated dataset will be used for subsequent statistical analysis. The following Python code generates a random dataset consisting of 100 integer values uniformly distributed between 1 and 500. The random.seed() function is used to ensure that the generated dataset can be reproduced.

Listing 1.1. Random Dataset Generation

```
import random
random.seed(42)
n = 100
random_dataset = [random.randint(1, 500) for i in range(n)]
print(random_dataset)
```

The first line imports the built-in `random` module, which provides functions for generating random numbers. Line `random.seed(42)` sets the initial state of the random number generator. The value 42 is called the random seed. Using the same seed makes it possible to reproduce the same sequence of random numbers each time the program is run. `n = 100` is the number of observations that will be generated. Therefore, the sample size is equal to 100. Next line generates the random dataset using a list comprehension. The function `random.randint(1, 500)` generates a random integer between 1 and 500. Both endpoints, 1 and 500, are included in the possible values. The square brackets collect all generated values into a Python list. The resulting list is assigned to the variable random_dataset. The last line outputs the generated list.

The generated dataset is displayed as a single long line, which makes it difficult to inspect visually. Therefore, we will present the data in a more convenient and readable format. The following code displays the generated dataset in rows of five values, making it easier to read and visually inspect.

Listing 1.2. Displaying the Dataset in a Readable Format

```
for index, value in enumerate(random_dataset, start=1):
    print(value, end=" ")
    if index % 5 == 0:
        print()
```

The code iterates through all elements of the generated dataset and displays them in a structured format. The `enumerate()` function is used to keep track of the position of each value. The values are printed on the same line and separated by spaces, while a line break is added after every fifth value. As a result, the dataset is displayed in rows of five values, making it easier to read and visually inspect.

Output:

```
328 58 13 380 141
126 115 72 378 53
347 380 457 280 45
303 217 17 16 48
112 120 259 309 14
288 102 367 333 360
280 215 113 230 302
143 415 446 4 389
413 82 358 217 175
143 80 111 491 391
173 53 48 195 50
184 434 177 310 136
414 23 374 236 275
64 499 473 194 41
283 151 425 322 317
454 442 186 296 99
361 36 24 339 117
396 149 41 438 120
444 52 195 143 233
326 428 187 84 190
```

2. We now proceed to calculate the main numerical characteristics of the dataset. The `statistics` module is part of the Python Standard Library. It provides functions for calculating common statistical characteristics of numerical datasets. Since the module is included with Python, it does not require any additional installation and can be imported using the command `import statistics`. In this laboratory work, the `statistics` module is used to calculate the *mean, mode,* and *median* of the generated dataset.

We will use the `mean(data)` function to calculate the *mean* of some data.

Listing 1.3. Calculation of the Sample Mean

```
import statistics
mean_value = statistics.mean(random_dataset)
print("Sample mean:", mean_value)
```

Output:

```
Sample mean: 226.67
```

The *mean* provides a useful measure of the average value of a dataset. However, it can be strongly affected by extreme observations and may not accurately represent the typical value of the data. In such cases, it may be useful to determine which value occurs most frequently. The *mode* is the value that appears most often in a dataset. It can be calculated for both numerical and categorical data. A dataset may have one mode, two modes, or several modes. Such datasets are called unimodal, bimodal, and multimodal, respectively.

In Python, the `statistics.mode()` function can be used to determine the most frequently occurring value in a dataset.

Listing 1.4. Calculation of the Sample Mode

```
import statistics
mode_value = statistics.mode(random_dataset)
print("Mode:", mode_value)
```

Output:

```
Mode: 143
```

The *mode* does not always provide a complete description of the central value of a dataset, since it depends only on the most frequently occurring observation. Another way to describe the center of numerical data is to use the *median*. In Python, the `statistics.median()` function is used to calculate the *median* of numerical data.

Listing 1.5. Calculation of the Sample Median

```
import statistics
median_value = statistics.median(random_dataset)
print("Median:", median_value)
```

Output:

```
Median: 205.0
```

3. Let’s calculate the main numerical characteristics of the variables contained in a DataFrame. The DataFrame contains information about 11 apartments. Each row represents one apartment, while the columns contain the following characteristics:

- area – the area of the apartment in square meters;
- rooms – the number of rooms;
- floor – the floor on which the apartment is located;
- price – the monthly rental price.

The pandas.DataFrame() function is used to organize the data in a tabular form. The resulting DataFrame consists of 11 rows and 4 numerical columns. Numerical characteristics such as the mean, median, mode, variance, and standard deviation can be calculated for each column.

Listing 1.6. Creation of a Dataframe

```
import pandas as pd
df = pd.DataFrame({
    'area': [42, 55, 68, 47, 73, 61, 39, 84, 58, 76, 50],
    'rooms': [1, 2, 3, 2, 3, 2, 1, 4, 2, 3, 2],
    'floor': [3, 7, 5, 10, 4, 8, 2, 6, 9, 12, 1],
    'price': [5200, 6800, 8500, 6100, 9200, 7600, 4900, 10800, 7100, 9700, 6400]
})
print(df)
```

Output:

```
    area  rooms  floor  price
0     42      1      3   5200
1     55      2      7   6800
2     68      3      5   8500
3     47      2     10   6100
4     73      3      4   9200
5     61      2      8   7600
6     39      1      2   4900
7     84      4      6  10800
8     58      2      9   7100
9     76      3     12   9700
10    50      2      1   6400
```

We can calculate the mean, median, and mode separately for each numerical column of the Pandas DataFrame, for example, for the apartment area variable:

Listing 1.7. Calculation of Descriptive Statistics for the Apartment Area Variable

```
area_mean = df['area'].mean()
area_median = df['area'].median()
area_mode = df['area'].mode()
print("Mean area:", area_mean)
print("Median area:", area_median)
print("Mode of area:", area_mode.tolist())
```

Output:

```
Mean area: 59.36363636363637
Median area: 58.0
Mode of area: [39, 42, 47, 50, 55, 58, 61, 68, 73, 76, 84]
```

In the current dataframe, all area values occur exactly once, so there is no single mode, and `mode()` will return all values.

Or we can get descriptive statistics for all columns with numeric data by applying the `mean()`, `median()`, and `mode()` methods to the entire dataframe.

Listing 1.8. Calculation of Descriptive Statistics for All Columns

```
mean_values = df.mean()
median_values = df.median()
mode_values = df.mode()
print("Mean values:")
print(mean_values)
print("\nMedian values:")
print(median_values)
print("\nMode values:")
print(mode_values)
```

Output:

```
Mean values:
area       59.363636
rooms       2.272727
floor       6.090909
price    7481.818182
dtype: float64

Median values:
area       58.0
rooms       2.0
floor       6.0
price    7100.0
dtype: float64

Mode values:
    area  rooms  floor  price
0     39    2.0      1   4900
1     42    NaN      2   5200
2     47    NaN      3   6100
3     50    NaN      4   6400
4     55    NaN      5   6800
5     58    NaN      6   7100
6     61    NaN      7   7600
7     68    NaN      8   8500
8     73    NaN      9   9200
9     76    NaN     10   9700
10    84    NaN     12  10800
```

The `.mode()` method returns all values with the highest frequency. If all values in a column occur only once, the method returns all of them. Since different columns may have different numbers of modes, Pandas fills the remaining cells with NaN.

The same statistical methods can also be applied to a separate Pandas Series. A Series is a one-dimensional data structure containing a sequence of values. In fact, each individual column of a DataFrame is represented as a Series.

For example, the following code creates a Series containing apartment areas and calculates its *mean*, *median* and *mode.*

Listing 1.9. Calculation of Descriptive Statistics for Series

```
area = pd.Series([42, 55, 68, 47, 73, 61, 39, 84, 58, 76, 50])
mean_area = area.mean()
median_area = area.median()
mode_area = area.mode()
print("Mean area:", mean_area)
print("Median area:", median_area)
print("Mode of area:", mode_area.tolist())
```

Output:

```
Mean area: 59.36363636363637
Median area: 58.0
Mode of area: [39, 42, 47, 50, 55, 58, 61, 68, 73, 76, 84]
```

In addition to the methods provided by Pandas, descriptive statistics can also be calculated using the NumPy and SciPy libraries. NumPy provides functions for numerical calculations, such as the mean and median, while the `scipy.stats` module contains functions for statistical analysis, including the calculation of the mode.

The following example calculates the mean, median, and mode for the rooms column of the apartment DataFrame.

Listing 1.10. Calculation of Descriptive Statistics for Series

```
import numpy as np
from scipy import stats
mean_rooms = np.mean(df['rooms'])
median_rooms = np.median(df['rooms'])
mode_result = stats.mode(df['rooms'], keepdims=False)
print("Mean number of rooms:", mean_rooms)
print("Median number of rooms:", median_rooms)
print("Mode:", mode_result.mode)
print("Mode frequency:", mode_result.count)
```

Output:

```
Mean number of rooms: 2.272727272727273
Median number of rooms: 2.0
Mode: 2
Mode frequency: 5
```

The `stats.mode()` function returns both the mode and its frequency. Therefore:

- `mode_result.mode` is the modal value;
- `mode_result.count` is the number of times the modal value occurs in the dataset.

Let’s calculate the median monthly rental price of the apartments in the dataset. Calculate the average area of the apartments.

Listing 1.11. Calculation of median monthly rental price of the apartments and the average area of the apartments

```
median_price = df['price'].median()
print(f"Median monthly rental price: {median_price:,.0f}")
mean_area = df['area'].mean()
print(f"Mean apartment area: {mean_area:.2f} m²")
```

Output:

```
Median monthly rental price: 7,100
Mean apartment area: 59.36 m²
```

4. In some statistical problems, numerical characteristics must be calculated not for the entire dataset, but only for observations that satisfy a specified condition. In Pandas, such observations can be selected using a Boolean mask.

A Boolean mask is a sequence of True and False values obtained by checking a condition for each row of a DataFrame. Only the rows for which the condition is True are included in further calculations.

For example, the following expression selects apartments with an area greater than 60 square meters: `df['area'] > 60.`

Let's calculate the average monthly rental price of apartments with an area greater than 60 square meters.

Listing 1.12. Calculation of the Mean Rental Price Using a Boolean Mask

```
mean_price = df.loc[df['area'] > 60, 'price'].mean()
print(f"Mean rental price: {mean_price:,.2f}")
```

Output:

```
Mean rental price: 9,160.00
```

- `df['area'] > 60` creates a Boolean mask;
- `df.loc[area_mask, 'price']` selects values from the price column only for apartments with an area greater than 60 m²;
- .mean() calculates the average rental price of the selected apartments.

In some cases, data must be selected using more than one condition. Pandas allows several Boolean conditions to be combined in a single mask. The **&** operator is used when all specified conditions must be satisfied simultaneously. In the following example, we will calculate the average monthly rental price of apartments that satisfy two conditions: their area is greater than 60 square meters, and they are located above the fifth floor.

Listing 1.13. Calculation of the Mean Rental Price Using Two Conditions

```
condition_mask = (df['area'] > 60) & (df['floor'] > 5)
selected_prices = df.loc[condition_mask, 'price']
mean_price = selected_prices.mean()
```

```
print(f"Mean rental price: {mean_price:,.2f}")
```

Output:

```
Mean rental price: 9,366.67
```

5. In Pandas, variance is calculated using the `.var()` method, while standard deviation is calculated using the `.std()` method.

For example, we can calculate the variance and standard deviation of apartment rental prices:

Listing 1.14. Calculation of Variance and Standard Deviation

```
price_variance = df['price'].var()
price_std = df['price'].std()
print(f"Variance of rental prices: {price_variance:,.2f}")
print(f"Standard deviation of rental prices: {price_std:,.2f}")
```

Output:

```
Variance of rental prices: 3,549,636.36
Standard deviation of rental prices: 1,884.05
```

Variance and standard deviation can also be calculated for all numerical columns of the DataFrame at once:

Listing 1.15. Calculation of Variance and Standard Deviation for all numerical columns of the DataFrame

```
variance_values = df.var()
std_values = df.std()
print("Sample variance:")
print(variance_values.round(2))
print("\nSample standard deviation:")
print(std_values.round(2))
```

Output:

```
Sample variance:
area         212.45
rooms          0.82
floor         12.09
price    3549636.36
dtype: float64

Sample standard deviation:
area       14.58
rooms       0.90
floor       3.48
price    1884.05
dtype: float64
```

We will also calculate the variance and standard deviation for the dataset presented in Listing 1.2. In the previous examples, the .var() and .std() methods were applied to objects created with the Pandas library, such as a Series or a DataFrame.

These methods are available because Pandas objects provide built-in tools for statistical calculations.

However, the variable random_dataset is an ordinary Python list. Python lists are designed to store collections of values, but they do not have the `.var()` and `.std()` methods. Therefore, these methods cannot be applied directly to `random_dataset`.

To calculate the variance and standard deviation of a Python list, we use the statistics module from the Python Standard Library. The `statistics.variance()` function calculates the sample variance, while the `statistics.stdev()` function calculates the sample standard deviation. The statistics module can work directly with lists, so it is not necessary to convert `random_dataset` into a Pandas Series or a NumPy array. In addition, this module is included in the Python Standard Library and does not require separate installation.

Listing 1.16. Calculation of Variance and Standard Deviation

```
import statistics
variance_value = statistics.variance(random_dataset)
std_value = statistics.stdev(random_dataset)
print(f"Sample variance: {variance_value:.2f}")
print(f"Sample standard deviation: {std_value:.2f}")
```

Output:

```
Sample variance: 20851.68
Sample standard deviation: 144.40
```

6. In addition to the mean, median, mode, variance, and standard deviation, we can calculate several other descriptive statistics. The **minimum** and **maximum** values describe the boundaries of the dataset, while the **range** shows the difference between them. **Quartiles** divide the ordered dataset into four parts, and the **interquartile range** measures the spread of the central 50% of the observations. The **coefficient of variation** expresses the standard deviation as a percentage of the mean.

Listing 1.17. Calculation of Some Another Characteristics

```
count_value = len(random_dataset)
minimum_value = min(random_dataset)
maximum_value = max(random_dataset)
range_value = maximum_value - minimum_value
q1, q2, q3 = statistics.quantiles(
    random_dataset,
    n=4,
    method='inclusive'
)
iqr_value = q3 - q1
mean_value = statistics.mean(random_dataset)
std_value = statistics.stdev(random_dataset)
coefficient_of_variation = std_value / mean_value * 100
print(f"Number of observations: {count_value}")
```

```
print(f"Minimum value: {minimum_value}")
print(f"Maximum value: {maximum_value}")
print(f"Range: {range_value}")
print(f"First quartile: {q1:.2f}")
print(f"Median: {q2:.2f}")
print(f"Third quartile: {q3:.2f}")
print(f"Interquartile range: {iqr_value:.2f}")
print(f"Coefficient of variation: {coefficient_of_variation:.2f}%")
```

Output:

```
Number of observations: 100
Minimum value: 4
Maximum value: 499
Range: 495
First quartile: 108.75
Median: 205.00
Third quartile: 358.50
Interquartile range: 249.75
Coefficient of variation: 63.71%
```

Quartiles divide an ordered dataset into four approximately equal parts:

- *Q1* is the value below which approximately 25% of the observations lie;
- *Q2* is the median;
- *Q3* is the value below which approximately 75% of the observations lie.

The interquartile range is calculated as: $IQR = Q3 - Q1$. It measures the spread of the middle 50% of the data and is less affected by extreme values than the ordinary range.

The coefficient of variation is calculated as: $CV = \text{standard deviation} / \text{mean} \cdot 100\%$. It expresses the standard deviation as a percentage of the mean. In this example, the coefficient of variation is 63.71%, indicating a high degree of relative variability in the data.

7. Let's move on to calculating the Pearson correlation coefficient. Let us estimate the Pearson correlation coefficient between apartment area and monthly rental price using the `pearsonr()` function.

Listing 1.18. Calculation of the Pearson correlation coefficient

```
from scipy import stats
area = df['area']
price = df['price']
corr = stats.pearsonr(area, price).statistic
print(f"Pearson correlation coefficient: {corr:.4f}")
```

Output:

```
Pearson correlation coefficient: 0.9977
```

The value of the Pearson correlation coefficient is close to 1. This indicates a very strong positive linear relationship between apartment area and monthly rental price in the given dataset. In general, larger apartments have higher rental prices.

In Pandas, pairwise correlations can be calculated simultaneously for several numerical variables contained in the columns of a DataFrame. The resulting values are presented as a correlation matrix. Each element of the matrix shows the Pearson correlation coefficient between a pair of variables.

We will estimate the correlations between apartment area, number of rooms, floor, and monthly rental price. Since the dataset contains four numerical variables, the result will be presented as a 4 × 4 correlation matrix.

Listing 1.19. Calculation of the Correlation Matrix for Apartment Data

```
correlation_matrix = df[['area', 'rooms', 'floor', 'price']].corr()
print("Correlation matrix:")
print(correlation_matrix.round(3))
```

Output:

```
Correlation matrix:
         area  rooms  floor  price
area   1.000  0.955  0.374  0.998
rooms  0.955  1.000  0.309  0.966
floor  0.374  0.309  1.000  0.371
price  0.998  0.966  0.371  1.000
```

Overall, apartment area and number of rooms have the strongest relationships with rental price, whereas the floor has a much weaker linear relationship with the other variables. These results describe only the given dataset and do not establish causal relationships between the variables.

8. Let's calculate the quartiles for the apartment dataset. Quartiles divide an ordered dataset into four approximately equal parts and help describe the distribution of numerical values. In particular, the first quartile *Q1* corresponds to the 25th percentile, the second quartile *Q2* is the median, and the third quartile *Q3* corresponds to the 75th percentile.

For the apartment dataset, we will calculate the quartiles of the monthly rental price variable. This will allow us to determine the price levels below which approximately 25%, 50%, and 75% of the apartments are located.

Let's calculate the first quartile, median, and third quartile of the monthly rental prices of the apartments presented in the DataFrame. Interpret the obtained values.

Listing 1.20. Calculation of Quartiles for Monthly Rental Prices

```
# list of quartiles
Q = [0.25, 0.5, 0.75]
price_quartiles = df['price'].quantile(Q)
print("Quartiles of monthly rental prices:")
print(price_quartiles)
```

Output:

```
Quartiles of monthly rental prices:
0.25    6250.0
0.50    7100.0
0.75    8850.0
Name: price, dtype: float64
```

The first quartile is 6250, which means that approximately 25% of the apartments have a monthly rental price not exceeding 6250. The second quartile is 7100. This value is also the median, which means that approximately half of the apartments have a rental price not exceeding 7100. The third quartile is 8850, which means that approximately 75% of the apartments have a monthly rental price not exceeding 8850.

The rooms column can be used to divide the apartments into groups, while the required quantiles can be calculated for the corresponding values in the price column. Let's calculate the 0.25-quantile of the monthly rental prices of apartments with two rooms and display the resulting floating-point value rounded to one decimal place.

Listing 1.21. Calculation of the 0.25-quantile

```
two_room_prices = df.loc[df['rooms'] == 2, 'price']
q25_price = two_room_prices.quantile(0.25)
print(f"0.25-quantile: {q25_price:.1f}")
```

Output:

```
0.25-quantile: 6400.0
```

The `quantile()` method can be applied not only to an entire DataFrame column but also to a subset of values selected according to a specified condition. Boolean masks can be used to select the required observations directly. Alternatively, the `groupby()` method can be used to divide the data into groups according to the values of another variable and then calculate the required quantiles separately for each group.

9. Now let's calculate 0.25-quantile and 0.5-quantile of the monthly rental prices of apartments with three rooms. Let's display the maximum of the two resulting values.

Listing 1.22. Calculation of the 0.25-quantile and 0.5-quantile

```
grouped_prices = df.groupby('rooms')['price']
quantile_values = grouped_prices.quantile([0.25, 0.5]).loc[3]
maximum_quantile = quantile_values.max()
print(f"Maximum quantile value: {maximum_quantile:.1f}")
```

Output:

```
Maximum quantile value: 9200.0
```

10. Let’s calculate the 53rd percentile of the monthly rental prices of all apartments in the dataset and display the result rounded to one decimal place. Since the 53rd percentile corresponds to the quantile of order 0.53, the calculation is performed as follows:

Listing 1.23. Calculation of the 53rd percentile

```
percentile_53 = df['price'].quantile(0.53)
print(f"53rd percentile: {percentile_53:.1f}")
```

Output:

```
53rd percentile: 7250.0
```

In this laboratory work, Python tools were used to generate, organize, and analyze numerical data. The main descriptive statistics were calculated, including the mean, median, mode, variance, standard deviation, quartiles, interquartile range, and coefficient of variation. Boolean masks and the groupby() method were also applied to analyze selected groups of observations.

The apartment dataset was used to study relationships between variables. The results showed a very strong positive correlation between apartment area and monthly rental price, while the floor had a weaker relationship with the other variables. Quantiles and percentiles were calculated to describe the position and distribution of rental prices. Overall, the work demonstrated the basic methods of statistical data analysis using Python, Pandas, NumPy, SciPy, and the statistics module.

Tasks for independent work (homework).
Theoretical knowledge test.

1. A real estate agency wants to study the rental prices of all apartments in a city. It collects information about 500 randomly selected apartments. What is the population in this study?

A. The 500 selected apartments

B. All apartments in the city

C. Only apartments with average rental prices

D. The rental prices calculated from the sample

2. A sample was selected using a simple random sampling procedure, but it accidentally contained an unusually large number of expensive apartments. Which conclusion is correct?

A. A simple random sample is always perfectly representative.

B. The sample cannot be random because it contains expensive apartments.

C. A randomly selected sample may still differ from the population because of random variation.

D. The population must contain the same proportion of expensive apartments as the sample.

3. The monthly rental prices in a dataset are relatively similar, except for one apartment with an extremely high price. Which statement is most likely to be correct?

A. The mean will be affected more strongly than the median.

B. The median will always become greater than the mean.

C. The mode will necessarily become equal to the extreme value.

D. The mean and median will remain unchanged.

4. Dataset A has a mean of 100 and a standard deviation of 10. Dataset B has a mean of 20 and a standard deviation of 10. Which dataset has greater relative variability?

A. Dataset A, because its mean is larger.

B. Dataset B, because the same standard deviation represents a larger proportion of its mean.

C. Both datasets have the same relative variability because their standard deviations are equal.

D. Relative variability cannot be compared using the coefficient of variation.

5. An extremely large observation is added to a dataset. Which measure is likely to change more significantly?

A. The interquartile range, because it uses the middle 50% of the data.

B. The ordinary range, because it depends directly on the minimum and maximum values.

C. The median, because it always depends on the largest value.

D. The first quartile, because it is always equal to the minimum value.

6. The monthly rental price is converted from basic currency units to smaller units by multiplying every price by 100. What happens to the covariance and Pearson correlation between apartment area and rental price?

A. Both covariance and correlation remain unchanged.

B. Covariance changes, but the Pearson correlation remains unchanged.

C. Correlation changes, but covariance remains unchanged.

D. Both values become equal to zero.

7. The Pearson correlation coefficient between study time and the number of errors is equal to -0.91. Which interpretation is correct?

A. There is a strong negative linear relationship between the variables.

B. Study time causes exactly 91% of the errors.

C. The variables have no relationship because the coefficient is negative.

D. The number of errors always decreases by exactly 0.91 when study time increases by one hour.

8. Two variables have a Pearson correlation coefficient close to zero, but their scatter plot has a clear U-shaped pattern. What does this result mean?

A. The variables are completely independent.

B. There is no linear relationship, but a nonlinear relationship may exist.

C. The covariance must be strongly positive.

D. The data must contain an error.

9. The 80th percentile of monthly rental prices is 9000. Which statement correctly interprets this result?

A. Exactly 80% of the apartments cost 9000.

B. Approximately 80% of the apartments have rental prices at or below 9000.

C. Approximately 20% of the apartments have rental prices at or below 9000.

D. The average rental price is equal to 9000.

10. The 0.25-quantile of rental prices for two-room apartments is 6400. Which interpretation is correct?

A. Approximately 25% of all apartments in the dataset have two rooms.

B. Approximately 25% of the two-room apartments have rental prices at or below 6400.

C. Every two-room apartment costs exactly 6400.

D. Approximately 75% of the two-room apartments have rental prices at or below 6400.

Tasks for Independent Work.

Objective

The purpose of this work is to apply the basic methods of descriptive statistical analysis using the Python programming language. Students will generate and analyze numerical data, work with a Pandas DataFrame, calculate measures of central tendency and dispersion, select observations according to specified conditions, examine relationships between variables, and calculate quantiles and percentiles.

Use the Python Standard Library and the Pandas, NumPy, and SciPy libraries.

For each task, include:

- the program code;
- the obtained output;
- a brief interpretation of the results.

Round numerical results to two decimal places unless another rounding rule is specified.

1. Random Dataset Generation

Generate a random dataset containing 100 integer values from 20 to 300, inclusive.

Use the following parameters:

- random seed: 2;
- sample size: 100;
- variable name: random_dataset.

Display the generated dataset as a Python list.

Then display the values in a more readable format so that each row contains five values.

Determine the sample size and explain why the use of a fixed random seed makes the generated dataset reproducible.

2. Measures of Central Tendency

Using the statistics module, calculate the following characteristics of random_dataset:

1. mean;
2. median;
3. mode.

Display all results on the screen.

Compare the mean and median and explain what they indicate about the central position of the data. Determine how many times the modal value occurs in the dataset.

3. Creation and Analysis of a DataFrame

The following dataset contains information about the academic performance of 11 students:

```
import pandas as pd
df = pd.DataFrame({
    'study_hours': [4, 6, 8, 5, 9, 7, 3, 10, 6, 8, 5],
    'assignments': [6, 8, 9, 7, 10, 8, 5, 10, 8, 9, 6],
    'absences': [5, 2, 3, 4, 1, 5, 6, 0, 1, 4, 3],
    'exam_score': [64, 78, 82, 71, 91, 75, 58, 95, 80, 84, 69]
})
print(df)
```

Each row represents one student. The columns contain the following information:

- study_hours — the average number of study hours per week;
- assignments — the number of completed assignments;
- absences — the number of missed classes;
- exam_score — the final examination score.

Complete the following tasks:

1. Display the DataFrame.
2. Determine the number of rows and columns.
3. Calculate the mean, median, and mode of the `study_hours` variable.
4. Calculate the mean, median, and mode for all numerical columns.
5. Explain why the `.mode()` method may return several values for one column.

6. Identify the modal number of completed assignments.

4. Application of Pandas, NumPy, and SciPy

Create a separate Pandas Series containing the values from the `exam_score` column.

Calculate its:

- mean;
- median;
- mode.

Then use NumPy and SciPy to calculate the mean, median, and mode of the assignments variable. Display both the modal value and its frequency.

Finally, calculate the median examination score and the average number of study hours per week. Interpret the obtained results.

5. Conditional Data Selection

Use Boolean masks to select observations that satisfy specified conditions.

Complete the following tasks:

1. Calculate the average examination score of students who study more than six hours per week.
2. Calculate the average examination score of students who study more than six hours per week and have fewer than four absences.
3. Display the rows of the DataFrame that satisfy both conditions.
4. Determine how many students satisfy both conditions.

Explain how the operators & and `.loc[]` are used in these calculations.

6. Variance and Standard Deviation

Calculate the sample variance and sample standard deviation of the `exam_score` variable using Pandas. Then calculate the variance and standard deviation for all numerical columns of the DataFrame. For `random_dataset`, calculate the sample variance and sample standard deviation using the statistics module.

7. Additional Descriptive Statistics

For `random_dataset`, calculate:

1. the number of observations;
2. the minimum value;
3. the maximum value;
4. the range;
5. the first quartile;
6. the median;
7. the third quartile;
8. the interquartile range;
9. the coefficient of variation.

Use the inclusive method when calculating quartiles with `statistics.quantiles()`.

Interpret the interquartile range and the coefficient of variation. Explain why the interquartile range is less sensitive to extreme values than the ordinary range.

8. Covariance and Correlation

Calculate the sample covariance between `study_hours` and `exam_score` using the Pandas `.cov()` method. Then calculate the Pearson correlation coefficient between these variables using `scipy.stats.pearsonr()`.

Answer the following questions:

1. Is the relationship positive or negative?
2. Is the linear relationship weak, moderate, or strong?
3. Is the sign of the covariance consistent with the sign of the correlation coefficient?
4. Does the obtained correlation prove that additional study time directly causes higher examination scores?

Next, construct a correlation matrix for the following variables:

- study_hours;
- assignments;
- absences;
- exam_score.

Round the matrix values to three decimal places.

Identify:

- the strongest positive correlation between two different variables;
- the strongest negative correlation;
- the variable that has the weakest linear relationship with the examination score.

Provide a brief interpretation of the matrix.

9. Quantiles and Grouped Calculations

Calculate the following quartiles of the exam_score variable:

- 0.25-quantile;
- 0.5-quantile;
- 0.75-quantile.

Interpret each obtained value in the context of student examination results.

Then complete the following tasks:

1. Calculate the 0.25-quantile of examination scores for students who completed exactly eight assignments. Display the result rounded to one decimal place.
2. Use the groupby() method to calculate the 0.25-quantile and 0.5-quantile of examination scores for students who completed exactly nine assignments.
3. Display the maximum of the two quantile values obtained in the previous task.

4. Explain the difference between selecting data with a Boolean mask and grouping data with groupby().

10. Percentile and Final Analysis

Calculate the 53rd percentile of the examination scores of all students. Display the result rounded to one decimal place.

Explain what the obtained percentile means in the context of the dataset.

Write a conclusion of approximately five to seven sentences. The conclusion should summarize:

- the central values of the generated dataset;
- the typical characteristics of the students;
- the variability of examination scores;
- the main relationships identified in the correlation matrix;
- the meaning of the calculated quartiles and percentile;
- the Python tools used in the analysis.

12. Mathematical Statistics. Laboratory Work 2 "Statistical Analysis of Normally Distributed Data".

Definition 49 [6, 14, 15]. A **histogram** is a graphical representation of the distribution of numerical data. The horizontal axis represents the values of the variable divided into intervals, while the vertical axis represents the frequency or relative frequency of observations in each interval. A histogram consists of adjacent rectangles whose heights correspond to the frequencies of the intervals.

To construct a histogram, the numerical data are divided into non-overlapping intervals, also called classes or bins. The number and width of the intervals should be selected so that the distribution can be interpreted clearly. Histograms commonly contain approximately 5 to 15 intervals.

The construction procedure includes the following steps:

1. Determine the minimum and maximum values of the dataset.
2. Select the number of intervals.
3. Determine the boundaries and width of each interval.
4. Count the number of observations belonging to each interval.
5. Place the intervals on the horizontal axis.
6. Place frequencies or relative frequencies on the vertical axis.
7. Draw adjacent rectangles whose heights represent the corresponding frequencies.

Histograms are particularly useful for displaying large datasets because they summarize many individual observations in a compact graphical form. They allow the researcher to obtain an initial understanding of the data before applying more formal statistical methods.

Definition 50 [5, 6, 15]. Skewness is a numerical measure of the asymmetry of a distribution about its center. A symmetric distribution has approximately the same shape on both sides of its central value and therefore has skewness close to zero. Positive skewness indicates that the distribution has a longer right tail, whereas negative skewness indicates a longer left tail. Skewness is dimensionless and characterizes the shape of a distribution rather than its measurement units.

The values are interpreted as follows:

- skewness ≈ 0 approximately symmetric distribution
- skewness > 0 right-skewed distribution
- skewness < 0 left-skewed distribution

For a normal distribution, theoretical skewness is equal to zero.

Definition 51 [5, 6, 15]. Kurtosis is a numerical measure of how heavy or light the tails of a distribution are in comparison with the tails of a normal distribution. A distribution with high kurtosis tends to have heavier tails and a greater tendency to produce extreme observations. A distribution with low kurtosis has lighter tails and fewer extreme observations.

Excess kurtosis is interpreted as follows:

- excess kurtosis ≈ 0 tails are similar to those of a normal distribution
- excess kurtosis > 0 heavier tails and more extreme observations
- excess kurtosis < 0 lighter tails and fewer extreme observations

Definition 52 [3,6]. A **confidence interval for the mean** is an interval estimate constructed from sample data and used to estimate the unknown mean of a population.

The sample mean provides a point estimate of the population mean. However, different random samples from the same population usually produce slightly different sample means. A confidence interval takes this sampling variability into account. When the population standard deviation is unknown, the confidence interval is constructed using the Student's t-distribution: $confidence\ interval = \bar{x} \pm t^* \cdot s / \sqrt{n}$

where:

- $\bar{x}$ is the sample mean;
- s is the sample standard deviation;
- n is the sample size;
- $s / \sqrt{n}$ is the standard error of the mean;
- t^* is the critical value of the t-distribution;
- $n - 1$ is the number of degrees of freedom.

A commonly used confidence level is 95%. The correct interpretation of a 95% confidence interval is based on repeated sampling. If the sampling procedure were repeated many times and a confidence interval were constructed from every sample,

approximately 95% of those intervals would contain the true population mean. It is common to say that we are 95% confident that the calculated interval contains the population mean. However, the confidence level describes the reliability of the interval estimation procedure rather than the probability of the population mean changing or being random [3, 6, 15, 20].

Laboratory Work 2 "Statistical Analysis of Normally Distributed Data".

The normal distribution is one of the most important probability distributions in mathematical statistics. Many continuous characteristics observed in nature and society, such as human height, measurement errors, and some biological indicators, can often be described approximately by a normal distribution [15, 17].

The normal distribution is symmetric and is completely determined by two parameters: the mean and the standard deviation. The mean determines the center of the distribution, while the standard deviation describes the spread of observations around the mean. Most observations are concentrated near the mean, whereas values far from the mean occur less frequently.

The normal distribution is important for several reasons. It is used to calculate probabilities and standard scores, construct confidence intervals, test statistical hypotheses, and develop many regression and forecasting methods. In addition, according to the central limit theorem, the distributions of many sample statistics become approximately normal when the sample size is sufficiently large, even when the original population is not normally distributed [6, 14].

1. Let's generate a synthetic dataset containing 100 human height measurements. The assumed mean height is 170 cm, and the standard deviation is 8 cm. These parameters are selected for educational purposes and do not describe every population. For convenience, we will consider only integer values. The `np.rint()` function rounds the obtained values to the nearest integers, and the `.astype(int)` method converts them to the integer data type. The following code generates 100 observations from a normal distribution with mean 170 cm and standard deviation 8 cm. The generated values are then rounded to the nearest integer for convenience. Therefore, the final dataset is discrete, although it is derived from an underlying normal model.

Listing 2.1. Generation of a Normally Distributed Human Height Dataset

```
import numpy as np
rng = np.random.default_rng(42)
n = 100
mean_height = 170
std_height = 8

height_dataset = np.rint(
    rng.normal(
```

```
        loc=mean_height,
        scale=std_height,
        size=n
    )
).astype(int)
print(height_dataset)
```

Output:

```
[172 162 176 178 154 160 171 167 170 163 177 176 171 179 174 163 173 162
 177 170 169 165 180 169 167 167 174 173 173 173 187 167 166 163 175 179
 169 163 163 175 176 174 165 172 171 172 177 172 175 171 172 175 158 167
 166 165 168 182 163 178 157 167 171 175 176 176 167 166 177 168 160 161
 163 174 171 176 167 171 175 168 174 165 167 167 160 174 166 170 174 174
 175 169 167 169 157 158 159 162 173 163]
```

The `rng.normal()` function generates values from a normal distribution. The parameter `loc=170` specifies the mean height, `scale=8` specifies the standard deviation, and `size=100` specifies the number of observations. The `np.rint()` function rounds the values to the nearest integers, after which `.astype(int)` converts them to an integer data type.

2. The `plt.hist()` function from the `Matplotlib` library is used to construct a histogram. The first argument specifies the numerical dataset to be displayed. The `bins` parameter determines the number of intervals into which the range of data values is divided. The `edgecolor='black'` parameter adds a black outline to the bars, and `linewidth=1.2` specifies the thickness of the outline. The completed histogram is displayed using the `plt.show()` function.

Listing 2.2. Construction of a Histogram of Human Height Measurements

```
import matplotlib.pyplot as plt
plt.hist(height_dataset, bins=10, edgecolor='black', linewidth=1.2)
plt.title("Histogram of Human Height Measurements")
plt.xlabel("Height (cm)")
plt.ylabel("Frequency")
plt.show()
```

Output:

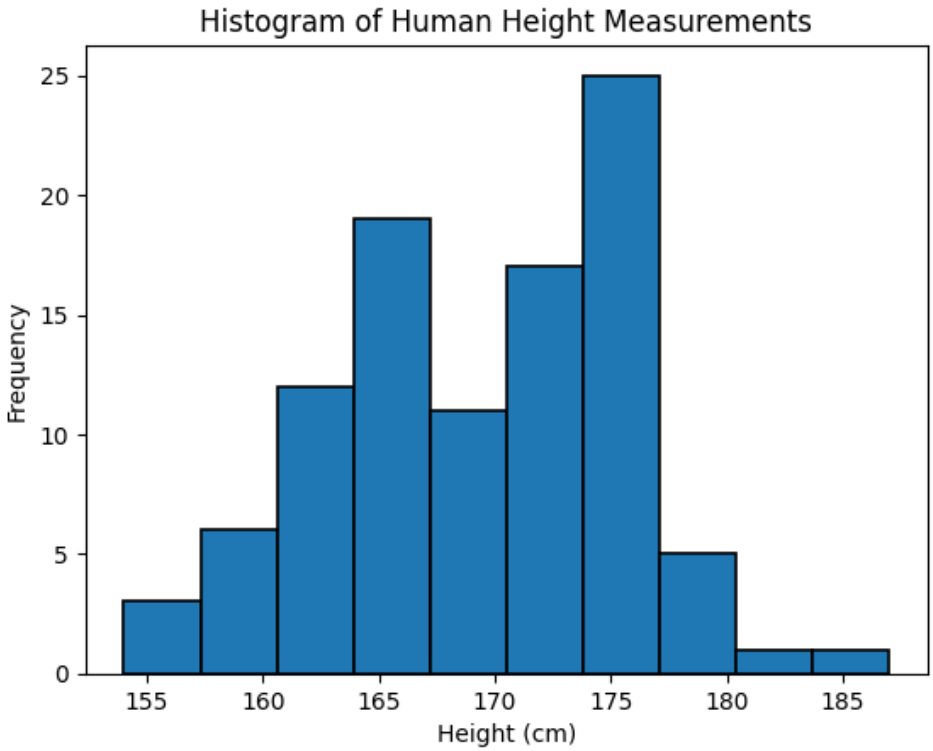


Figure 38. Histogram of Human Height Measurements

The histogram shows that most height measurements are concentrated near the center of the distribution, approximately between 165 and 177 cm. Heights that are considerably smaller or larger than the mean occur less frequently. The general shape of the histogram is approximately bell-shaped, which is characteristic of a normal distribution.

However, the bars are not perfectly symmetric because the dataset contains only 100 randomly generated observations. Differences in bar heights are caused by random sampling variability. Therefore, the histogram provides an approximate, rather than exact, representation of the theoretical normal distribution.

3. In addition to a histogram, the distribution of a numerical variable can be visualized using a **density plot**. A density plot represents the distribution as a smooth continuous curve and makes it easier to examine its general shape, center, symmetry, and spread.

For a normally distributed variable, the theoretical density function has a symmetric bell-shaped form. Its highest point is located at the mean, while the curve gradually decreases on both sides. The exact position and width of the curve depend on the mean and standard deviation. The density function of a normal distribution always has a symmetric bell-shaped form, although its position and width depend on the mean and standard deviation.

Listing 2.3. Construction of a Density Plot for Human Height Measurements

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
height_series = pd.Series(height_dataset)
height_series.plot.density(linewidth=2)
plt.title("Density Plot of Human Height Measurements")
plt.xlabel("Height (cm)")
plt.ylabel("Density")
plt.show()
```

Output:

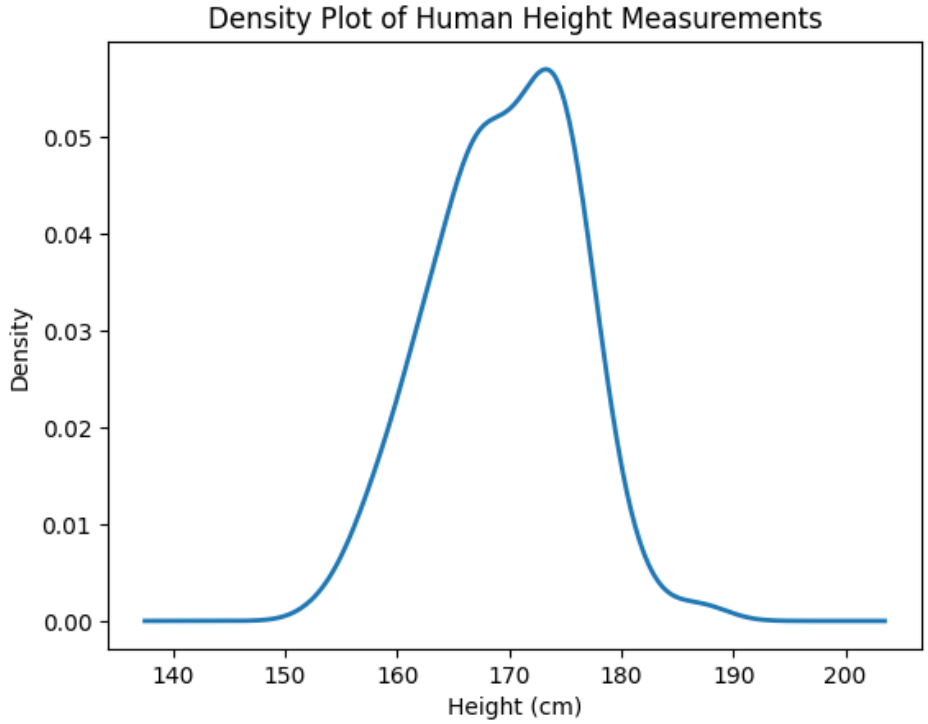


Figure 39. Density Plot for Human Height Measurements

The `pd.Series()` function converts the `height_dataset` array into a Pandas Series. A Series is a one-dimensional data structure used for storing and analyzing numerical observations.

The `height_series.plot.density()` method constructs a density plot using kernel density estimation. Instead of dividing the data into intervals, as a histogram does, this method produces a smooth curve that estimates how the observations are distributed over the range of possible values. The parameter `linewidth=2` specifies the thickness of the density curve.

The density plot has an approximately symmetric bell-shaped form. The highest part of the curve is located near 170 cm, which means that height values close to 170 cm occur most frequently in the generated dataset.

The density gradually decreases on both sides of the central value. Therefore, very short and very tall observations occur less frequently than heights near the mean. This shape is generally characteristic of a normal distribution.

The curve may not be perfectly symmetric because it is estimated from a finite sample containing only 100 observations. Random sampling variability and

rounding the height values to whole centimetres may produce small irregularities in the density curve.

4. Now let us calculate the probability that the height of a randomly selected person is less than 190 cm. Assume that human height follows a normal distribution with a mean of 170 cm and a standard deviation of 8 cm.

Listing 2.4. Calculation of the Probability That Human Height Is Less Than 190 cm Using the Probability Density Function

```
import numpy as np
from scipy.stats import norm
from scipy.integrate import quad
mean, std = 170, 8
probability = quad(
    lambda x: norm.pdf(x, mean, std),
    -np.inf,
    190
)[0]
print("Probability:", probability)
```

Output:

```
Probability: 0.9937903346742073
```

The `norm` object from the `scipy.stats` module provides the probability density function of the normal distribution. The variables *mean* and *std* store the parameters of the normal distribution. In this example, the mean height is 170 cm and the standard deviation is 8 cm. The expression `lambda x: norm.pdf(x, mean, std)` defines the normal probability density function that must be integrated. The function `norm.pdf()` calculates the density value for each possible value of *x*. The `quad()` function calculates the area under the density curve from negative infinity to 190 cm. This area is equal to the probability that the height of a randomly selected person is less than 190 cm. The `quad()` function normally returns two values: the result of integration and an estimate of the numerical error. The index [0] selects only the first value, which is the calculated probability.

Listing 2.5. Graphical Representation of the Probability That Human Height Is Less Than 190 cm

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
plt.figure(figsize=(10, 6))
mean, std = 170, 8
x = np.linspace(mean - 4 * std, mean + 4 * std, 1000)
f = norm.pdf(x, mean, std)
plt.plot(x, f)
plt.fill_between(x, f, where=(x < 190), alpha=0.3)
plt.axvline(190, linestyle="--")
plt.title("Probability That Human Height Is Less Than 190 cm")
plt.xlabel("Height (cm)")
plt.ylabel("Probability Density")
plt.show()
```

Output:

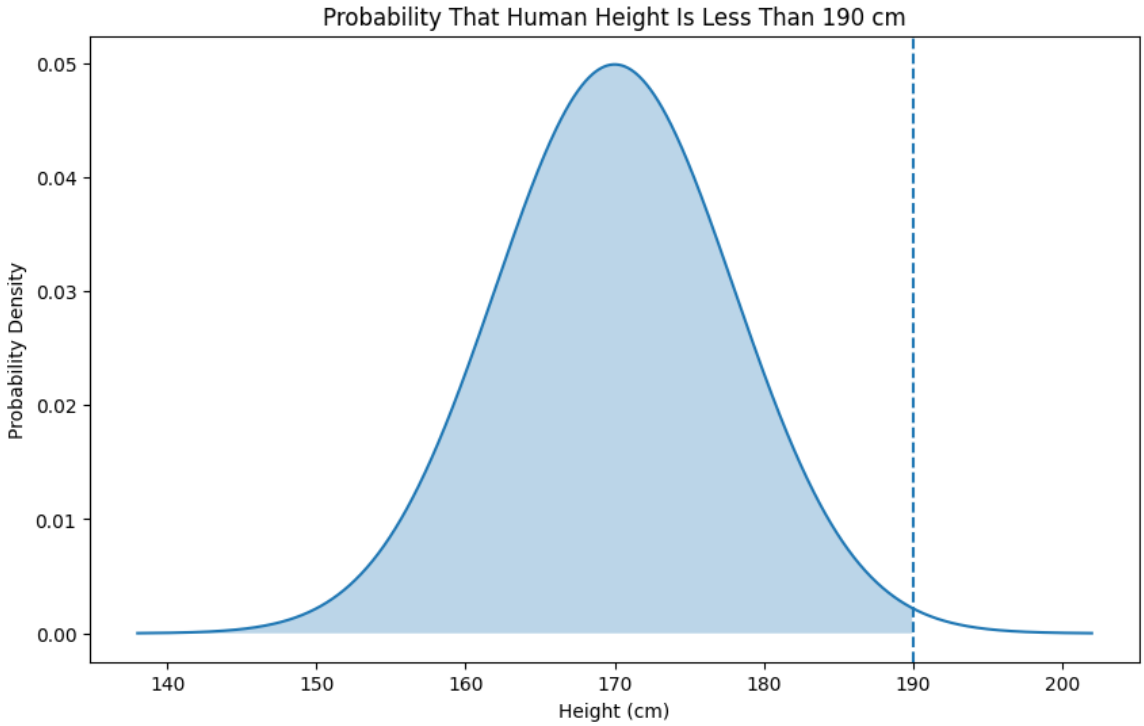


Figure 40. Graphical Representation of the Probability That Human Height Is Less Than 190 cm

The `np.linspace()` function creates 1000 evenly spaced values over the interval from four standard deviations below the mean to four standard deviations above the mean. These values are stored in the variable *x* and are used to construct a smooth normal density curve.

The `norm.pdf()` function calculates the values of the normal probability density function for each value of *x*. The parameters *mean*=170 and *std*=8 specify the mean and standard deviation of the distribution.

The `plt.plot()` function draws the normal density curve. The `plt.fill_between()` function shades the area under the curve for which $x < 190$. This shaded area represents the probability that a randomly selected person has a height of less than 190 cm.

The `plt.axvline()` function draws a vertical dashed line at 190 cm and indicates the boundary of the considered region.

The same probability can also be calculated directly using the cumulative distribution function, `norm.cdf()`.

Listing 2.6. Calculation of the Theoretical Probability That Height Is Less Than 190 cm

```
from scipy.stats import norm
mean_height = 170
std_height = 8
```

```
height_limit = 190
theoretical_probability = norm.cdf(
    height_limit,
    loc=mean_height,
    scale=std_height
)
print("Theoretical probability:", theoretical_probability)
```

Output:

```
Theoretical probability: 0.9937903346742238
```

The `norm.cdf()` function from the `scipy.stats` module calculates the cumulative probability of a normally distributed random variable. The parameter `height_limit=190` specifies the upper boundary of the required interval. The parameter `loc=170` sets the mean of the normal distribution, while `scale=8` sets its standard deviation. The function calculates the probability that the random variable takes a value less than or equal to 190 cm: $P(X \leq 190) \approx 0.9938$.

After calculating the theoretical probability, it is useful to represent the result graphically. For a continuous random variable, probability corresponds to the area under the probability density curve over a specified interval.

In this problem, the probability $P(X < 190)$ is represented by the area under the normal density curve to the left of 190 cm. Shading this area makes it possible to see which part of the distribution corresponds to people whose height is less than 190 cm. Since the shaded region covers almost the entire area under the curve, we can visually confirm that the required probability is very high.

The vertical line at 190 cm indicates the boundary of the considered interval. The area to the right of this line represents the small probability that a person's height is 190 cm or greater.

5. Now let us calculate the probability that a randomly selected person has a height greater than 190 cm. Assume that human height follows a normal distribution with a mean of 170 cm and a standard deviation of 8 cm.

Listing 2.7. Calculation of the Probability That Human Height Is Greater Than 190 cm

```
from scipy.stats import norm
mean, std = 170, 8
probability = 1 - norm.cdf(190, mean, std)
print("Probability:", probability)
```

Output:

```
Probability: 0.006209665325776159
```

Listing 2.7. Graphical Representation of the Probability That Human Height Is Greater Than 190 cm

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
```

```
plt.figure(figsize=(10, 6))
mean, std = 170, 8
x = np.linspace(mean - 4 * std, mean + 4 * std, 1000)
f = norm.pdf(x, mean, std)
plt.plot(x, f)
# Shade the area corresponding to P(X > 190)
plt.fill_between(
    x,
    f,
    where=(x > 190),
    alpha=0.3
)
plt.axvline(190, linestyle="--")
plt.title("Probability That Human Height Is Greater Than 190 cm")
plt.xlabel("Height (cm)")
plt.ylabel("Probability Density")
```

Output:

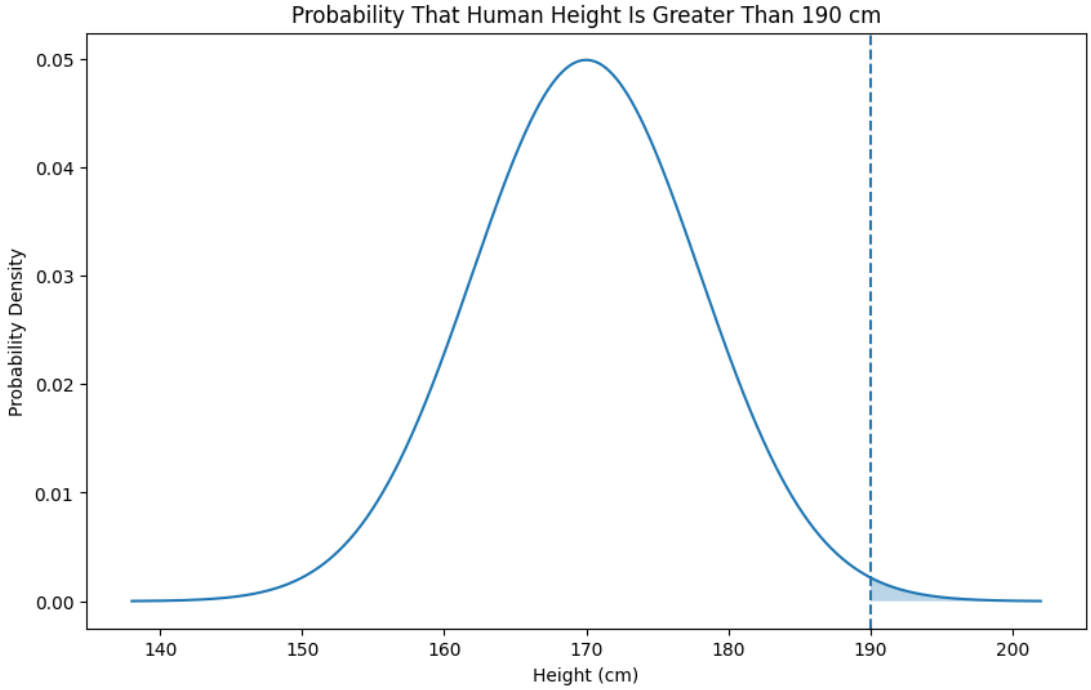


Figure 41. Graphical Representation of the Probability That Human Height Is More Than 190 cm

6. Now let us calculate the probability that a randomly selected person has a height between 150 and 170 cm. Assume that human height follows a normal distribution with a mean of 170 cm and a standard deviation of 8 cm. We need to calculate: $P(150 < X < 170)$. The required probability is found as the difference between two cumulative probabilities: $P(150 < X < 170) = P(X < 170) - P(X < 150)$.

Listing 2.8. Calculation of the Probability That Human Height Is Between 150 and 170 cm

```
from scipy.stats import norm
mean, std = 170, 8
probability = (
    norm.cdf(170, mean, std)
    - norm.cdf(150, mean, std)
```

```
)
print("Probability:", probability)
```

Output:

```
Probability: 0.49379033467422384
```

The expression `norm.cdf(170, mean, std)` calculates the probability that a person's height is less than 170 cm. The expression `norm.cdf(150, mean, std)` calculates the probability that a person's height is less than 150 cm. By subtracting the second probability from the first, we obtain the probability that height lies between 150 and 170 cm.

Listing 2.9. Graphical Representation of the Probability That Human Height Is Between 150 and 170 cm

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
plt.figure(figsize=(10, 6))
mean, std = 170, 8
x = np.linspace(mean - 4 * std, mean + 4 * std, 1000)
f = norm.pdf(x, mean, std)
plt.plot(x, f)
# Shade the area corresponding to P(150 < X < 170)
plt.fill_between(
    x,
    f,
    where=(x > 150) & (x < 170),
    alpha=0.3
)
plt.axvline(150, linestyle="--")
plt.axvline(170, linestyle="--")
plt.title("Probability That Human Height Is Between 150 and 170 cm")
plt.xlabel("Height (cm)")
plt.ylabel("Probability Density")
plt.show()
```

Output:

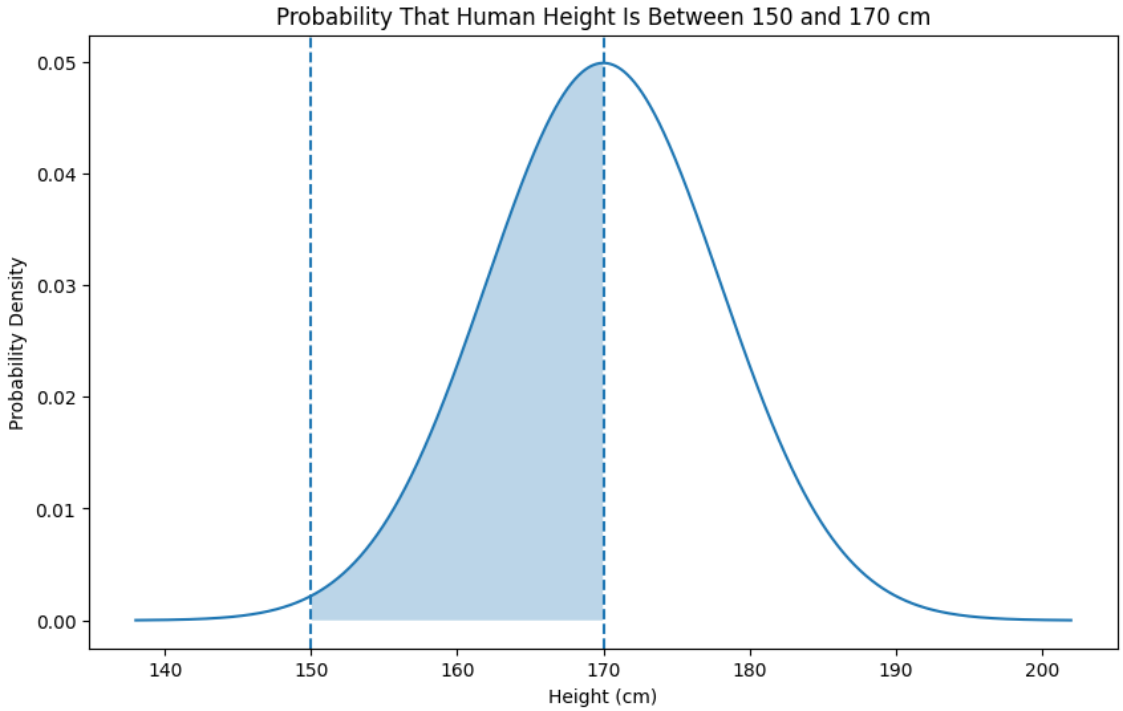


Figure 42. Graphical Representation of the Probability That Human Height Is Between 150 and 170 cm

7. Histograms and density plots provide a visual representation of the shape of a distribution. They make it possible to assess whether the data are approximately symmetric and bell-shaped. However, visual analysis alone is subjective and may not be sufficient to determine whether a dataset follows a normal distribution. Therefore, the next tasks will focus on statistical tests of normality. These tests make it possible to evaluate the assumption of normality using an objective numerical criterion.

Let us examine several methods that can be used to determine whether a dataset follows an approximately normal distribution.

Method 1. Normal Probability Plot. A normal probability plot, also known as a Q-Q plot, compares the ordered values of a dataset with the corresponding theoretical quantiles of a normal distribution. When the data are approximately normally distributed, the plotted points should lie close to a straight line. Noticeable curvature, systematic deviations, or points located far from the line may indicate that the distribution differs from a normal distribution [14].

The general procedure for constructing a normal probability plot is as follows:

1. Arrange the observed values in ascending order.
2. Assign a cumulative probability to each ordered observation.
3. Determine the theoretical quantile of the normal distribution corresponding to each cumulative probability.

4. Plot the theoretical normal quantiles along the x-axis and the ordered observed values along the y-axis.

Let us construct a normal probability plot for the previously generated human height dataset.

Listing 2.10. Construction of a Normal Probability Plot for Human Height Data

```
import matplotlib.pyplot as plt
from scipy.stats import norm

# First, sort the height data in ascending order
height_dataset_sorted = sorted(height_dataset)

# Calculate the number of observations
n = len(height_dataset_sorted)

# Calculate the cumulative probability for each observation
cumulative_probability = [
    (i - 0.375) / (n + 0.25)
    for i in range(1, n + 1)
]

# Calculate the corresponding theoretical normal quantiles
theoretical_quantiles = norm.ppf(cumulative_probability)

print(
    "First and last cumulative probabilities:",
    cumulative_probability[0],
    cumulative_probability[-1]
)

print(
    "First and last theoretical quantiles:",
    theoretical_quantiles[0],
    theoretical_quantiles[-1]
)

plt.figure(figsize=(10, 8))

plt.scatter(
    theoretical_quantiles,
    height_dataset_sorted
)

plt.xlabel("Theoretical Quantiles", fontsize=15)
plt.ylabel("Sorted Height Data", fontsize=15)
plt.title("Normal Probability Plot for Human Height", fontsize=17)

plt.show()
```

Output:

```
First and last cumulative probabilities: 0.006234413965087282 0.9937655860349127
First and last theoretical quantiles: -2.498590560962256 2.498590560962256
```

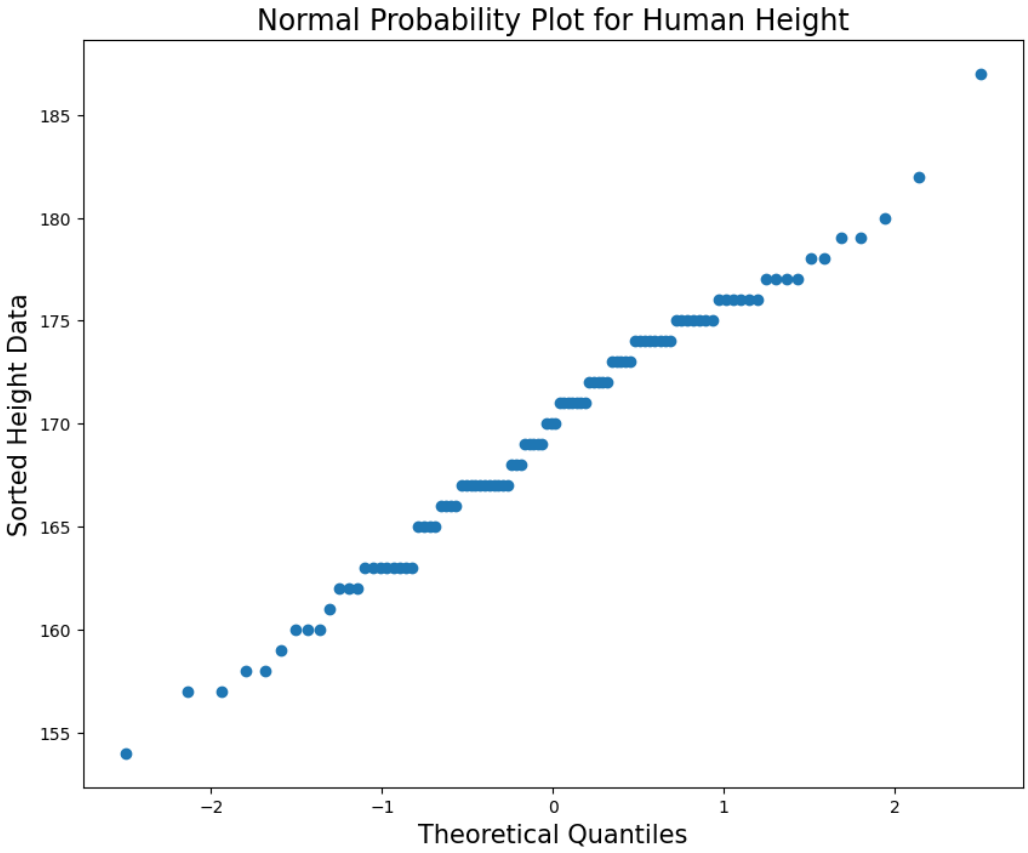


Figure 43. Normal Probability Plot for Human Height

The `sorted()` function arranges the height measurements in ascending order. The sorted values are stored in the variable `height_dataset_sorted.`

The `len()` function determines the number of observations in the dataset.

The cumulative probability assigned to each ordered observation is calculated using the expression: (i - 0.375) / (n + 0.25). Here, i represents the position of an observation in the ordered dataset, while n is the total number of observations.

The *norm.ppf()* function converts the cumulative probabilities into the corresponding theoretical quantiles of the standard normal distribution. The *plt.scatter()* function constructs the normal probability plot. The theoretical normal quantiles are placed along the x-axis, while the ordered height measurements are placed along the y-axis.

If the height data are approximately normally distributed, the points on the graph should form an approximately straight line. Systematic deviations from a straight-line pattern may indicate that the distribution is not normal.

The points on the normal probability plot form an approximately straight increasing pattern. Most observations are located close to an imaginary straight line, without strong systematic curvature. This indicates that the distribution of the generated height measurements is approximately normal.

The horizontal groups of points are caused by rounding the height measurements to integer values. Several observations therefore have the same height, producing short horizontal steps on the graph.

Slight deviations can be observed at the lower and upper ends of the plot. In particular, the smallest and largest height measurements are located somewhat farther from the general linear pattern. Such deviations are expected in a finite random sample and do not necessarily indicate a substantial departure from normality.

Overall, the normal probability plot provides visual evidence that the human height dataset is reasonably consistent with a normal distribution. However, this graphical conclusion should be supported by a formal statistical test of normality.

8. **Method 2. Shapiro-Wilk test [18].** The Shapiro-Wilk test is a formal statistical method used to determine whether a dataset is consistent with a normal distribution. Unlike a histogram or a normal probability plot, it provides a numerical test result.

The test compares the ordered observations with the values that would be expected if the data followed a normal distribution. The result is represented by the Shapiro-Wilk statistic W and the *p*-value.

The hypotheses are formulated as follows:

- H_0: the data follow a normal distribution.
- H_1: the data do not follow a normal distribution.

The *p*-value is the probability of obtaining the observed test result, or a more extreme result, assuming that the null hypothesis is true. In a normality test, the null hypothesis states that the data follow a normal distribution. A small *p*-value indicates that the observed data would be unusual under the assumption of normality and provides evidence against the null hypothesis. A large *p*-value indicates that the data are reasonably consistent with a normal distribution. The *p*-value is compared with a selected significance level, commonly $\alpha = 0.05$. If the *p*-value is less than or equal to 0.05, the null hypothesis is rejected. If the *p*-value is greater than 0.05, there is insufficient evidence to reject the null hypothesis. A *p*-value greater than 0.05 does not prove that the data are normally distributed. It only indicates that the sample does not provide sufficient evidence against the assumption of normality.

A significance level must be selected before interpreting the result. It is commonly set to $\alpha = 0.05$. The decision is made according to the following rule:

- if *p*-value > 0.05, there is insufficient evidence to reject H_0;
- if *p*-value ≤ 0.05, H_0 is rejected, and the distribution is considered significantly different from normal.

A value of W close to 1 generally indicates that the data are close to a normal distribution. However, the final conclusion should be based primarily on the *p*-value.

It is important to note that *p*-value > 0.05 does not prove that the distribution is perfectly normal. It only means that the available data do not provide sufficient evidence against the assumption of normality.

Let us apply the Shapiro–Wilk test to the previously generated human height data.

Listing 2.11. Application of the Shapiro–Wilk Normality Test to Human Height Data

```
from scipy.stats import shapiro
statistic, p_value = shapiro(height_dataset)
print("Shapiro-Wilk statistic:", statistic)
print("p-value:", p_value)
alpha = 0.05
if p_value > alpha:
    print("The data are consistent with a normal distribution.")
else:
    print("The data are not consistent with a normal distribution.")
```

Output:

```
Shapiro-Wilk statistic: 0.9855478602973372
p-value: 0.3475656062502669
The data are consistent with a normal distribution.
```

The obtained *p*-value is approximately 0.3476. Since this value is greater than the significance level of 0.05, the null hypothesis is not rejected. Therefore, the sample does not provide sufficient evidence that the distribution of the generated height measurements differs from a normal distribution. This conclusion agrees with the normal probability plot, where the points formed an approximately straight pattern.

9. **D'Agostino-Pearson Test [5].**

The D'Agostino-Pearson test is a formal statistical test used to determine whether a sample is consistent with a normal distribution. It is an omnibus normality test because it evaluates two important characteristics of the distribution simultaneously: skewness and kurtosis.

The test transforms the sample skewness and kurtosis into standardized values and combines them into a single test statistic:

$$K^2 = Z^2skewness + Z^2kurtosis$$

A large value of the test statistic indicates that the sample differs considerably from a normal distribution. A small value indicates that no strong departure from normality has been detected. In SciPy, the test statistic is compared with a chi-squared distribution with two degrees of freedom to calculate the *p*-value.

The hypotheses of the D'Agostino-Pearson test are formulated as follows:

- H_0: the data follow a normal distribution.
- H_1: the data do not follow a normal distribution.

A significance level is selected before the test is performed. In this example, we use $\alpha = 0.05$.

The decision rule is:

- if p-value > 0.05, there is insufficient evidence to reject H_0;
- if p-value ≤ 0.05, H_0 is rejected.

Let us apply the D'Agostino-Pearson test to the previously generated human height data.

Listing 2.12. Application of the D'Agostino-Pearson Test to Human Height Data

```
from scipy.stats import normaltest
statistic, p_value = normaltest(height_dataset)
print("D'Agostino-Pearson statistic:", statistic)
print("p-value:", p_value)
alpha = 0.05
if p_value > alpha:
    print("There is insufficient evidence to reject normality.")
else:
    print("The data are not normally distributed.")
```

Output:

```
D'Agostino-Pearson statistic: 0.5158354518129924
p-value: 0.7726587992942257
There is insufficient evidence to reject normality.
```

The `normaltest()` function applies the D'Agostino–Pearson test to the values stored in `height_dataset.`

The function returns two results:

- statistic contains the combined test statistic based on skewness and kurtosis;
- *p_value* contains the probability used to make a statistical decision.

The obtained p-value is compared with the significance level *alpha* = 0.05. p-value = 0.7727 > 0.05, the null hypothesis is not rejected.

The sample does not provide sufficient evidence that the distribution of the generated height measurements differs from a normal distribution. Therefore, the height data are considered consistent with the assumption of normality.

This result agrees with the histogram, density plot, normal probability plot, and Shapiro–Wilk test considered earlier.

The Shapiro–Wilk test is generally preferred as a primary normality test for small and moderate samples. It evaluates the overall agreement between the ordered observations and the values expected under a normal distribution.

The D'Agostino–Pearson test is especially useful for moderate and large samples. It combines standardized measures of skewness and kurtosis and therefore detects departures from normality associated with asymmetry, unusual tails, or both.

For the human height dataset containing 100 observations, the Shapiro–Wilk test may be used as the primary formal test, while the D'Agostino–Pearson test may serve as an additional check. For a dataset containing 1000 measurements, both tests can be applied, with particular attention to the D'Agostino–Pearson test.

The final conclusion should not be based on one test alone. Formal test results should be interpreted together with a histogram, a density plot, and a normal probability plot.

10. After checking whether the height data are approximately normally distributed, it is useful to determine how far each individual observation is located from the center of the distribution. However, the difference between an observation and the mean is expressed in the original units of measurement and does not take the variability of the data into account. To express this distance in units of standard deviation, the observations can be standardized using the ***z*-score**. A ***z*-score**, also called a **standard score**, indicates how many standard deviations an observation lies above or below the mean.

For a sample, the z-score is calculated as $z = (x - \bar{x}) / s$ where:

- x is the observed value;
- $\bar{x}$ is the sample mean;
- s is the sample standard deviation.

The z-score is interpreted as follows:

- $z = 0$: the observation is equal to the mean;
- $z > 0$: the observation is above the mean;
- $z < 0$: the observation is below the mean;
- $|z|$ describes the distance from the mean in standard deviation units.

A z-score is not a formal normality test. It is used primarily to standardize observations, compare values measured on different scales, and identify unusually distant observations. When the data are approximately normally distributed, observations with $|z| > 3$ are often regarded as potential outliers [6, 14, 15].

Let us calculate the z-score of every observation in the previously generated human height dataset.

Listing 2.13. Calculation of Z-Scores for the Human Height Dataset

```
import numpy as np
height_array = np.array(height_dataset)
sample_mean = np.mean(height_array)
sample_std = np.std(height_array, ddof=1)
z_scores = (height_array - sample_mean) / sample_std
print("Sample mean:", round(sample_mean, 2))
print("Sample standard deviation:", round(sample_std, 2))
print("First 10 z-scores:", np.round(z_scores[:10], 2))
print("Minimum z-score:", round(np.min(z_scores), 2))
print("Maximum z-score:", round(np.max(z_scores), 2))
```

Output:

```
Sample mean: 169.6
Sample standard deviation: 6.23
First 10 z-scores: [ 0.39 -1.22  1.03  1.35 -2.5  -1.54  0.22 -0.42  0.06 -1.06]
Minimum z-score: -2.5
Maximum z-score: 2.79
```

The `np.array()` function converts the height dataset into a NumPy array. The `np.mean()` function calculates the sample mean. For the generated data, the mean height is 169.6 cm. The `np.std()` function calculates the sample standard deviation. The parameter `ddof=1` indicates that the sample standard deviation is used. The expression `(height_array - sample_mean) / sample_std` subtracts the sample mean from every height measurement and divides the result by the sample standard deviation. The resulting array contains the standardized values. The `np.round()` function rounds the displayed z-scores to two decimal places.

The first height measurement is 172 cm and its *z*-score is approximately 0.39. This means that 172 cm is 0.39 standard deviations above the sample mean. The height of 162 cm has a *z*-score of approximately −1.22. Therefore, it is 1.22 standard deviations below the mean. The smallest observation, 154 cm, has a *z*-score of approximately −2.50. The largest observation, 187 cm, has a *z*-score of approximately 2.79. Since all calculated *z*-scores lie between −3 and 3, no observations are identified as potential outliers according to the commonly used three-standard-deviation rule.

11. After examining the distribution of the height measurements and confirming that the data are approximately normal, we can proceed from descriptive analysis to statistical inference. The sample mean is 169.6 cm, which is close to the assumed population mean of 170 cm. However, a small difference between these values may arise simply because of random sampling. To determine whether the observed difference is statistically significant, we can apply the **one-sample *t*-test**. This test is used to determine whether the mean of a sample differs significantly from a specified or hypothesized population mean [3, 6, 19, 20].

The test is appropriate when:

- one quantitative sample is analysed;
- the observations are independent;
- the population standard deviation is unknown;
- the data are approximately normally distributed, especially when the sample is small;
- the sample does not contain influential extreme outliers.

Suppose that the expected population mean height is 170 cm. We want to determine whether the mean of the generated sample differs significantly from this value. The hypotheses are: H_0: $\mu = 170$ cm; H_1: $\mu \neq 170$ cm. The null hypothesis

states that the population mean is equal to 170 cm. The alternative hypothesis states that the population mean differs from 170 cm. Because the alternative hypothesis includes differences in both directions, a **two-sided one-sample *t*-test** is used.

We select the significance level $\alpha = 0.05$. The decision rule is:

- if p-value > 0.05, there is insufficient evidence to reject H_0;
- if p-value ≤ 0.05, H_0 is rejected.

Let us apply the one-sample t-test to the previously generated height dataset.

Listing 2.14. Application of the One-Sample t-Test to the Human Height Dataset

```
from scipy.stats import ttest_1samp
import numpy as np
hypothesized_mean = 170
alpha = 0.05
t_statistic, p_value = ttest_1samp(
    height_dataset,
    popmean=hypothesized_mean
)
sample_mean = np.mean(height_dataset)
sample_std = np.std(height_dataset, ddof=1)
degrees_of_freedom = len(height_dataset) - 1
print("Sample mean:", round(sample_mean, 2))
print("Sample standard deviation:", round(sample_std, 2))
print("Hypothesized mean:", hypothesized_mean)
print("t-statistic:", round(t_statistic, 4))
print("Degrees of freedom:", degrees_of_freedom)
print("p-value:", round(p_value, 4))
if p_value > alpha:
    print(
        "There is insufficient evidence that "
        "the population mean differs from 170 cm."
    )
else:
    print(
        "The population mean differs significantly "
        "from 170 cm."
    )
```

Output:

```
Sample mean: 169.6
Sample standard deviation: 6.23
Hypothesized mean: 170
t-statistic: -0.6421
Degrees of freedom: 99
p-value: 0.5223
There is insufficient evidence that the population mean differs from 170 cm.
```

The function `ttest_1samp()` from the `scipy.stats` module performs the one-sample *t*-test. The first argument contains the sample observations: `height_dataset`. The parameter `popmean` specifies the hypothesized population mean: `popmean`=170. The sample standard deviation is calculated using

`np.std(height_dataset, ddof=1)`. The parameter `ddof=1` indicates that the sample standard deviation is required.

The obtained results are:

- $t = -0.6421$
- p-value = 0.5223

Since 0.5223 > 0.05 the null hypothesis is not rejected.

The sample does not provide sufficient evidence that the population mean height differs from 170 cm. The difference between the sample mean of 169.6 cm and the hypothesized mean of 170 cm is small and can reasonably be explained by random sampling variation.

12. After performing the one-sample *t*-test, it is useful to estimate the population mean rather than only test a single hypothesized value. A confidence interval provides a range of plausible values for the unknown population mean [3, 6, 15, 20].

For the generated human height dataset, we will construct a 95% confidence interval for the population mean height.

Listing 2.15. Construction of a 95% Confidence Interval for the Mean Human Height

```
import numpy as np
from scipy.stats import t
height_array = np.array(height_dataset)

# Calculate the sample characteristics
n = len(height_array)
sample_mean = np.mean(height_array)
sample_std = np.std(height_array, ddof=1)

# Set the confidence level
confidence_level = 0.95
alpha = 1 - confidence_level

# Calculate the standard error
standard_error = sample_std / np.sqrt(n)

# Determine the critical value of the t-distribution
critical_value = t.ppf(
    1 - alpha / 2,
    df=n - 1
)

# Calculate the margin of error
margin_of_error = critical_value * standard_error

# Calculate the confidence interval
lower_limit = sample_mean - margin_of_error
upper_limit = sample_mean + margin_of_error

print("Sample size:", n)
print("Sample mean:", round(sample_mean, 2))
print("Sample standard deviation:", round(sample_std, 2))
print("Standard error:", round(standard_error, 4))
```

```
print("Critical t-value:", round(critical_value, 4))
print("Margin of error:", round(margin_of_error, 4))
print(
    "95% confidence interval:",
    round(lower_limit, 2),
    "to",
    round(upper_limit, 2)
)
```

Output:

```
Sample size: 100
Sample mean: 169.6
Sample standard deviation: 6.23
Standard error: 0.623
Critical t-value: 1.9842
Margin of error: 1.2361
95% confidence interval: 168.36 to 170.84
```

Based on the generated sample, the population mean height is estimated to lie between 168.36 cm and 170.84 cm with a confidence level of 95%. The hypothesized mean of 170 cm lies inside this confidence interval. Therefore, 170 cm is a plausible value for the population mean. This conclusion agrees with the result of the one-sample *t*-test: $t(99) = -0.642$, $p = 0.522$.

The *t*-test did not detect a statistically significant difference between the sample mean and 170 cm. Similarly, the confidence interval contains 170 cm.

The two methods express the same statistical conclusion in different ways:

- the one-sample *t*-test evaluates whether a specified mean should be rejected;
- the confidence interval provides a range of plausible values for the population mean.

Tasks for independent work (homework).

Theoretical knowledge test.

1. Which statement correctly describes the roles of the mean and standard deviation in a normal distribution?

A. The mean determines the spread, while the standard deviation determines the center.

B. The mean determines the center, while the standard deviation determines the spread.

C. Both parameters determine only the symmetry of the distribution.

D. The standard deviation determines the sample size.

2. Why may the histogram of a normally generated sample fail to have a perfectly symmetric bell-shaped form?

A. A normal distribution is never symmetric in a finite sample.

B. The histogram always changes the original distribution into a uniform distribution.

C. Random sampling variability, rounding, and the selected histogram intervals may produce irregularities.

D. The mean and standard deviation cannot be used for integer data.

3. What is the main difference between norm.pdf() and norm.cdf()?

A. norm.pdf() returns a probability directly, while norm.cdf() returns a density value.

B. norm.pdf() calculates density values, while norm.cdf() calculates cumulative probabilities.

C. Both functions always return the same result.

D. norm.cdf() can only be used for sample data.

4. Why is the probability $P(150 < X < 170)$ calculated as `norm.cdf(170) - norm.cdf(150)`?

A. Because the probability between two values is the difference between their cumulative probabilities.

B. Because both values must first be converted into frequencies.

C. Because the probability density function cannot be used for normal distributions.

D. Because probabilities greater than the mean must always be subtracted.

5. Which pattern in a normal probability plot provides the strongest visual evidence that the data are approximately normally distributed?

A. The points form a horizontal line.

B. The points form a circular pattern.

C. The points follow an approximately straight increasing line without strong systematic curvature.

D. All points have identical y-coordinates.

6. The Shapiro–Wilk test gives p-value = 0.3476 at $\alpha = 0.05$. Which conclusion is correct?

A. The null hypothesis must be rejected because the p-value is positive.

B. The data have a 34.76% probability of being normally distributed.

C. There is insufficient evidence to reject the hypothesis of normality.

D. The data are proven to follow a perfect normal distribution.

7. What additional information does the D'Agostino–Pearson test use when assessing normality?

A. The sample mean and sample size only.

B. The minimum and maximum values.

C. Skewness and kurtosis.

D. The median and interquartile range only.

8. A height measurement has $z = -2.5$. What does this value mean?

A. The observation is 2.5 cm below the mean.

B. The observation is 2.5 standard deviations below the mean.

C. The observation is 2.5 standard deviations above the mean.

D. The observation must be removed from the dataset.

9. In the one-sample *t*-test, the result is $t(99) = -0.642$ and $p = 0.522$. What does the negative *t*-statistic indicate?

A. The sample mean is lower than the hypothesized population mean.

B. The sample standard deviation is negative.

C. The null hypothesis must be rejected.

D. The population mean cannot be estimated.

10. The 95% confidence interval for the population mean is [168.36, 170.84] cm. What conclusion is consistent with this interval and the one-sample *t*-test?

A. The population mean must be exactly 169.6 cm.

B. The value 170 cm is not a plausible population mean.

C. The sample proves that every person has a height between 168.36 and 170.84 cm.

D. The value 170 cm is a plausible population mean, and the sample does not show a statistically significant difference from it.

Tasks for Independent Work. Statistical Analysis of Resting Heart Rate Measurements.

Objective.

The purpose of this assignment is to develop practical skills in:

- generating normally distributed data;
- visualizing a numerical distribution;
- calculating theoretical and empirical probabilities;
- assessing the normality of a dataset;
- calculating and interpreting z-scores;
- applying the one-sample t-test;
- constructing and interpreting a confidence interval for the population mean.

Dataset Description

Assume that the resting heart rate of adults can be approximately described by a normal distribution with:

- Mean resting heart rate: 72 beats per minute
- Standard deviation: 9 beats per minute
- Sample size: 120 observations

These parameters are selected for educational purposes and should not be interpreted as universal medical standards. The observations must be rounded to the nearest integer.

Task 1. Generate the Dataset.

Generate a synthetic dataset containing 120 resting heart rate measurements.

Use the following initial code:

```
import numpy as np
rng = np.random.default_rng(24)
n = 120
mean_heart_rate = 72
std_heart_rate = 9
heart_rate_dataset = np.rint(
    rng.normal(
        loc=mean_heart_rate,
        scale=std_heart_rate,
        size=n
    )
).astype(int)
print(heart_rate_dataset)
```

Requirements

1. Run the program and display the generated dataset.
2. Explain all variables
3. Explain why the random seed is used.

Task 2. Calculate Descriptive Statistics.

Calculate the following sample characteristics:

- sample size;
- sample mean;
- sample median;
- sample standard deviation;
- minimum value;
- maximum value;
- range.

Use the sample standard deviation by specifying *ddof*=1.

Questions

1. Is the sample mean close to the assumed mean of 72 beats per minute?
2. Are the mean and median close to each other?
3. What may a noticeable difference between the mean and median indicate?
4. How does the sample standard deviation compare with the assumed standard deviation of 9?

Task 3. Construct a Histogram.

Construct a histogram of the generated resting heart rate measurements.

Describe:

1. where most observations are concentrated;
2. whether the distribution appears approximately symmetric;
3. whether it has an approximately bell-shaped form;
4. whether any unusually small or large values are visible;

5. why a finite random sample may not produce a perfectly symmetric histogram.

Task 4. Construct a Density Plot.

Convert the dataset into a Pandas Series and construct a density plot.

Explain:

- what a Pandas Series is;
- what kernel density estimation represents;
- how a density plot differs from a histogram;
- where the highest part of the density curve is located;
- whether the curve appears approximately symmetric.

Task 5. Calculate Theoretical Probabilities.

Assume that resting heart rate follows a normal distribution with $\mu = 72$, $\sigma = 9$. Calculate the following theoretical probabilities:

Part A. Calculate the probability that a randomly selected person has a resting heart rate below 60 beats per minute.

Part B. Calculate the probability that a randomly selected person has a resting heart rate above 85 beats per minute.

Part C. Calculate the probability that a randomly selected person has a resting heart rate between 65 and 80 beats per minute.

Task 6. Calculate a Probability Using Integration.

Calculate $P(X < 60)$ again by integrating the normal probability density function.

Questions

1. What does norm.pdf() calculate?
2. What does the quad() function calculate?
3. Why does the result correspond to an area under the density curve?
4. Why is [0] placed after the quad() function?
5. Does the result agree with the value obtained using norm.cdf()?

Task 7. Visualize the Theoretical Probabilities.

Construct three graphs of the normal probability density function.

Graph 1. Shade the area corresponding to $P(X < 60)$.

Graph 2. Shade the area corresponding to $P(X > 85)$.

Graph 3. Shade the area corresponding to $P(65 < X < 80)$.

Task 8. Calculate Empirical Probabilities.

Using the generated sample, calculate the empirical proportions of observations satisfying the following conditions: $X < 60$, $X > 85$, $65 < X < 80$.

Explain:

1. why the theoretical and empirical probabilities are not necessarily identical;
2. how sample size influences the difference;
3. how rounding the observations may influence the empirical result.

Suggested listing title:

Task 9. Construct a Normal Probability Plot.

Construct a normal probability plot manually.

The procedure must include:

1. sorting the observations;
2. calculating cumulative probabilities;
3. calculating theoretical normal quantiles using norm.ppf();
4. plotting theoretical quantiles against sorted observations.

Discuss:

- whether the points form an approximately straight increasing pattern;
- whether systematic curvature is present;
- whether deviations occur in the tails;
- whether horizontal groups of points appear because the data were rounded;
- whether the plot supports the assumption of normality.

Task 10. Apply the Shapiro–Wilk Test.

Apply the Shapiro–Wilk normality test to *heart_rate_dataset.*

Formulate the hypotheses:

- H_0: the data follow a normal distribution.
- H_1: the data do not follow a normal distribution.

Use the significance level $\alpha = 0.05$.

Do not write that a large p-value proves normality. State only that there is insufficient evidence to reject the assumption of normality.

Task 11. Apply the D'Agostino–Pearson Test.

Apply the D'Agostino–Pearson normality test.

Report:

- sample skewness;
- excess kurtosis;
- the D'Agostino–Pearson statistic;
- the p-value;
- the statistical decision.

Questions

1. What does positive skewness indicate?
2. What does negative skewness indicate?
3. What does excess kurtosis close to zero indicate?
4. Why does the D'Agostino–Pearson test analyse both skewness and kurtosis?
5. Does its conclusion agree with the Shapiro–Wilk test?

Task 12. Calculate *Z*-Scores.

Calculate the *z*-score for every observation. Use the sample mean and sample standard deviation.

Display:

- the first ten z-scores;
- the minimum z-score;
- the maximum z-score.

Explain the meaning of:

- a positive z-score;
- a negative z-score;
- $z = 0$;
- $|z| > 3$.

Identify all observations satisfying $|z| > 3$. State whether the sample contains potential outliers according to the three-standard-deviation rule.

Task 13. Apply the One-Sample *t*-Test.

Use the one-sample t-test to determine whether the population mean resting heart rate differs from 72 beats per minute.

Formulate the hypotheses:

- $H_0: \mu = 72$
- $H_1: \mu \neq 72$

Use $\alpha = 0.05$.

Report:

- the sample mean;
- the hypothesized mean;
- the t-statistic;
- the number of degrees of freedom;
- the p-value;
- the statistical decision;
- the practical conclusion.

Explain what the sign of the t-statistic indicates. Remember that the sign indicates whether the sample mean is above or below the hypothesized mean, but does not determine statistical significance by itself.

Task 14. Construct a 95% Confidence Interval for the Mean.

Construct a 95% confidence interval for the population mean resting heart rate using the Student's t-distribution.

Calculate:

- the sample mean;
- the sample standard deviation;
- the standard error;
- the critical t-value;
- the margin of error;
- the lower confidence limit;

- the upper confidence limit.

Questions

1. Does the interval contain 72 beats per minute?
2. Is the confidence interval conclusion consistent with the one-sample t-test?
3. What happens to the interval width when the sample size increases?
4. What happens to the interval width when the confidence level increases?
5. Why is the Student's t-distribution used instead of the standard normal distribution?

13. Mathematical Statistics. Laboratory Work 3 "A/B Testing, Big Data Sampling, and Statistical Analysis of User Behavior".

A/B testing is one of the most common applications of mathematical statistics in software development, web analytics, product management, and machine learning. It allows developers and analysts to compare two versions of a digital product and determine whether an observed difference in user behavior is caused by a real improvement or by random sampling variability.

In a typical A/B experiment, users are randomly divided into two groups. The control group uses the current version of the system, while the experimental group uses a modified version. The groups are then compared using statistical methods.

Examples of changes that can be evaluated through A/B testing include:

- a new user interface;
- a modified recommendation algorithm;
- a different search ranking model;
- a new payment page;
- a different notification strategy;
- an optimized server-side algorithm;
- a new machine learning model.

A/B testing is especially important in large-scale web systems because even a small change in conversion rate or user engagement may affect thousands or millions of users [11].

In this laboratory work, a synthetic dataset containing 20000 users will be analysed. Although this dataset is small compared with industrial Big Data systems, it is sufficiently large to demonstrate the principles of statistical analysis, sampling, hypothesis testing, effect size estimation, and engineering decision-making.

Experimental Scenario.

A large online store is testing a new recommendation algorithm.

The users are divided into two groups:

- Group A uses the current recommendation algorithm;
- Group B uses the new recommendation algorithm.

The purpose of the experiment is to determine whether the new algorithm:

- increases session duration;
- increases the number of viewed pages;
- improves conversion;
- increases the average order value;
- does not cause an unacceptable increase in server response time.

Does the new recommendation algorithm improve user engagement and conversion without causing an unacceptable decrease in system performance?

Definition 53 [11]. An **A/B test** is a controlled statistical experiment in which users or observations are randomly assigned to two or more groups in order to compare different versions of a system, product, or algorithm. The control group represents the current version, whereas the experimental group receives the new version.

The statistical analysis is used to determine whether the observed differences between the groups are sufficiently large to be considered statistically significant.

1. Let us generate a synthetic dataset containing 20000 users. Each experimental group will contain 10000 users.

The dataset will contain the following variables:

- user_id (unique user identifier)
- group (experimental group A or B)
- session_duration (session duration in seconds)
- pages_viewed (number of pages viewed during the session)
- converted (whether the user completed a purchase)
- order_value (purchase amount)
- response_time (average server response time in milliseconds)

Session duration is generated using a gamma distribution because time-based user metrics are often non-negative and right-skewed.

The conversion variable is generated using a Bernoulli model:

- 0 means that no purchase was completed;
- 1 means that the user completed a purchase.

Order values are generated only for users who completed a purchase. Server response time is generated from an approximately normal distribution. Positive order values are generated from lognormal distributions with median parameters of approximately 65 and 68 for Groups A and B, respectively.

Listing 3.1. Generation of a Synthetic A/B Testing Dataset

```
import numpy as np
import pandas as pd
rng = np.random.default_rng(42)
n_per_group = 10000
```

```
# Generate session duration in seconds
session_a = rng.gamma(
    shape=4.0,
    scale=45.0,
    size=n_per_group
)
session_b = rng.gamma(
    shape=4.2,
    scale=45.0,
    size=n_per_group
)
# Generate the number of viewed pages
pages_a = rng.poisson(
    lam=6.4,
    size=n_per_group
)
pages_b = rng.poisson(
    lam=7.0,
    size=n_per_group
)
# Generate conversion indicators
converted_a = rng.binomial(
    n=1,
    p=0.120,
    size=n_per_group
)
converted_b = rng.binomial(
    n=1,
    p=0.132,
    size=n_per_group
)
# Generate order values
order_a = np.where(
    converted_a == 1,
    rng.lognormal(
        mean=np.log(65),
        sigma=0.45,
        size=n_per_group
    ),
    0
)
order_b = np.where(
    converted_b == 1,
    rng.lognormal(
        mean=np.log(68),
        sigma=0.45,
        size=n_per_group
    ),
    0
)
# Generate server response time in milliseconds
response_a = np.clip(
    rng.normal(
        loc=220,
        scale=35,
        size=n_per_group
    ),
    80,
    None
)
response_b = np.clip(
    rng.normal(
```

```
        loc=235,
        scale=38,
        size=n_per_group
    ),
    80,
    None
)
# Create a DataFrame
df = pd.DataFrame({
    "user_id": np.arange(
        1,
        2 * n_per_group + 1
    ),
    "group": np.repeat(
        ["A", "B"],
        n_per_group
    ),
    "session_duration": np.concatenate(
        [session_a, session_b]
    ).round(1),
    "pages_viewed": np.concatenate(
        [pages_a, pages_b]
    ),
    "converted": np.concatenate(
        [converted_a, converted_b]
    ),
    "order_value": np.concatenate(
        [order_a, order_b]
    ).round(2),
    "response_time": np.concatenate(
        [response_a, response_b]
    ).round(1)
})
print(df.head())
print("Dataset shape:", df.shape)
```

Output:

```
   user_id group  session_duration  pages_viewed  converted  order_value  \
0        1     A             192.7             8          1         82.0
1        2     A             238.5             3          0          0.0
2        3     A             176.3             6          0          0.0
3        4     A             163.6            10          0          0.0
4        5     A             253.0             7          0          0.0

   response_time
0          229.9
1          221.9
2          189.2
3          282.3
4          195.3
Dataset shape: (20000, 7)
```

The function `np.random.default_rng(42)` creates a random number generator. The value 42 is used as a random seed, which makes the generated dataset reproducible.

The gamma distribution is used for session duration because it produces positive and usually right-skewed values. The Poisson distribution is used to generate the number of viewed pages because this variable represents a count. The binomial

distribution with *n=1* generates a binary conversion indicator [2, 6, 15, 17]. The `np.where()` function assigns an order value only to users who completed a purchase. The `np.clip()` function prevents server response time from becoming unrealistically small. The Pandas DataFrame combines all generated variables into one tabular dataset.

2. Before performing statistical analysis, the dataset should be checked for missing values, duplicate identifiers, and impossible observations.

In real Big Data systems, collected datasets may contain:

- missing records;
- repeated users;
- incorrect timestamps;
- negative durations;
- impossible category values;
- corrupted measurements.

Listing 3.2. Data Quality Assessment

```
print("Missing values:")
print(df.isna().sum())

duplicate_users = df.duplicated(
    subset="user_id"
).sum()

invalid_sessions = (
    df["session_duration"] <= 0
).sum()

invalid_response_times = (
    df["response_time"] <= 0
).sum()

invalid_conversion_values = (
    ~df["converted"].isin([0, 1])
).sum()

print("Duplicate user IDs:", duplicate_users)
print("Invalid session durations:", invalid_sessions)
print(
    "Invalid response times:",
    invalid_response_times
)
print(
    "Invalid conversion values:",
    invalid_conversion_values
)
```

Output:

```
Missing values:
user_id             0
group               0
session_duration    0
pages_viewed        0
converted           0
```

```
order_value         0
response_time       0
dtype: int64
Duplicate user IDs: 0
Invalid session durations: 0
Invalid response times: 0
Invalid conversion values: 0
```

The dataset does not contain missing values or duplicate user identifiers. All session durations and response times are positive. The conversion variable contains only the values 0 and 1. Therefore, the generated dataset can be used for further statistical analysis without additional cleaning.

3. Before applying statistical tests, the main sample characteristics should be calculated separately for groups A and B. Descriptive statistics provide an initial comparison between the groups but do not determine whether the differences are statistically significant [6, 14, 15].

Listing 3.3. Calculation of Descriptive Statistics by Experimental Group

```
summary = df.groupby("group").agg(
    users=("user_id", "count"),
    mean_session=(
        "session_duration",
        "mean"
    ),
    median_session=(
        "session_duration",
        "median"
    ),
    std_session=(
        "session_duration",
        "std"
    ),
    mean_pages=(
        "pages_viewed",
        "mean"
    ),
    conversion_rate=(
        "converted",
        "mean"
    ),
    mean_response_time=(
        "response_time",
        "mean"
    )
)

mean_order_value = (
    df[df["converted"] == 1]
    .groupby("group")["order_value"]
    .mean()
)

summary["mean_order_value"] = mean_order_value

print(summary.round(4))
```

Output:

```
        users  mean_session  median_session  std_session  mean_pages  \
group
A       10000      180.8152           165.7      90.6404      6.4236
B       10000      189.7992           174.9      93.4256      6.9673

        conversion_rate  mean_response_time  mean_order_value
group
A                0.1188            220.3291           71.8653
B                0.1289            235.2082           76.5196
```

The mean session duration is approximately 180.82 seconds in group A and 189.80 seconds in group B. The difference in sample means is 189.80 − 180.82 = 8.98 seconds. Users in group B also viewed more pages on average.

The conversion rates are:

- Group A: 0.1188 = 11.88%
- Group B: 0.1289 = 12.89%

The average order value among converted users is higher in group B.

However, the average server response time is also higher in group B:

- Group A: 220.33 ms
- Group B: 235.21 ms

At this stage, the new algorithm appears to improve several product metrics but may reduce system performance. Formal statistical tests are required to determine whether the observed differences are statistically significant.

4. A histogram can be used to compare the distributions of session duration in the two experimental groups [10].

Listing 3.4. Histograms of Session Duration by Experimental Group

```
import matplotlib.pyplot as plt

session_group_a = df.loc[
    df["group"] == "A",
    "session_duration"
]

session_group_b = df.loc[
    df["group"] == "B",
    "session_duration"
]

plt.figure(figsize=(10, 6))

plt.hist(
    session_group_a,
    bins=40,
    alpha=0.5,
    density=True,
    label="Group A"
)

plt.hist(
    session_group_b,
    bins=40,
    alpha=0.5,
```

```
    density=True,
    label="Group B"
)

plt.title(
    "Distribution of Session Duration by Experimental Group"
)
plt.xlabel("Session Duration (seconds)")
plt.ylabel("Density")
plt.legend()

plt.show()
```

Output:

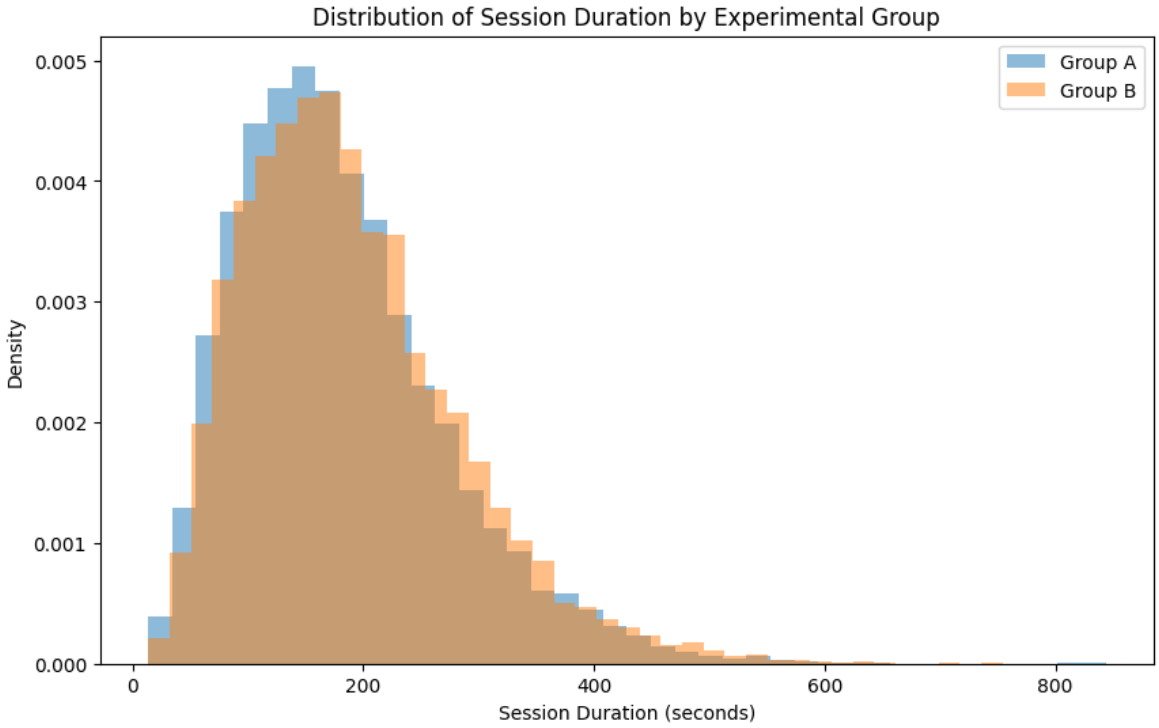


Figure 44. Distribution of Session Duration by Experimental Group

The histograms overlap considerably. Therefore, visual inspection alone is insufficient to determine whether the difference between the groups is statistically significant.

5. A box plot can also be used to compare the central values and variability of the groups.

Listing 3.5. Box Plot of Session Duration

```
plt.figure(figsize=(8, 6))

plt.boxplot(
    [
        session_group_a,
        session_group_b
    ],
    labels=[
        "Group A",
        "Group B"
```

```
    ],
    showfliers=False
)

plt.title(
    "Box Plot of Session Duration by Experimental Group"
)
plt.xlabel("Experimental Group")
plt.ylabel("Session Duration (seconds)")
plt.show()
```

Output:

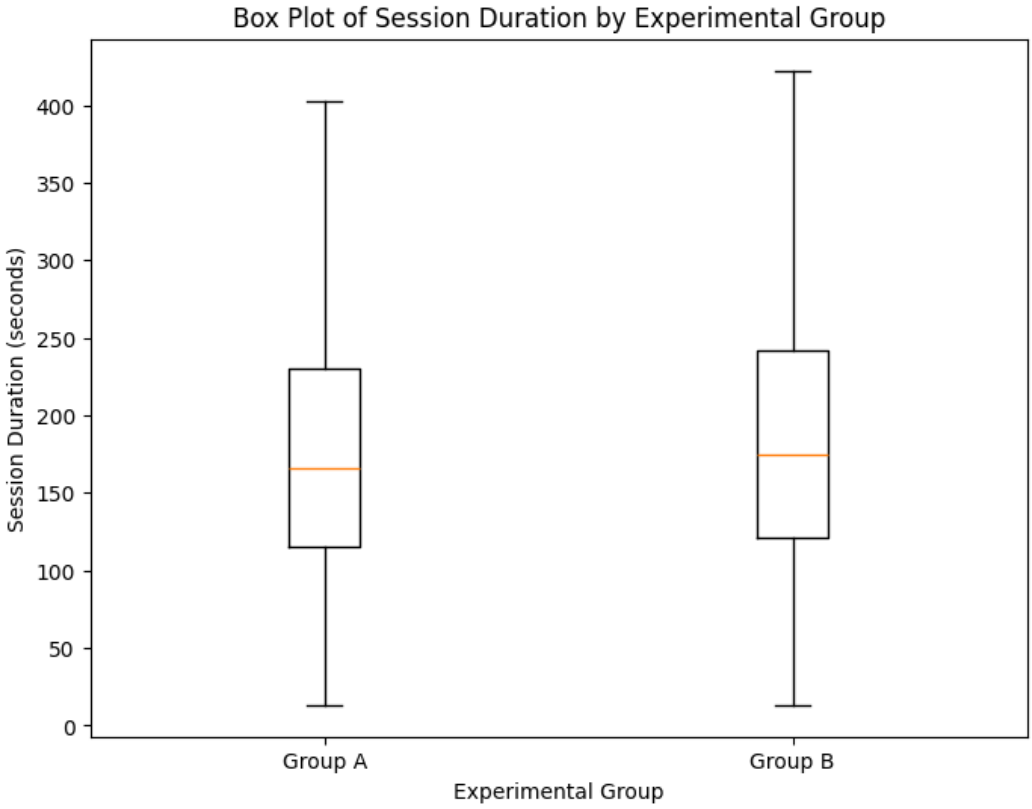


Figure 45. Box Plot of Session Duration by Experimental Group

The median session duration is higher in group B. Both groups have considerable variability and right-skewed distributions.

6. One of the main ideas of Big Data analysis is that it may not always be necessary to process the entire dataset.

A properly selected random sample may provide sufficiently accurate estimates while reducing:

- execution time;
- memory consumption;
- computational cost;
- network traffic;
- storage requirements.

However, estimates based on small samples may be unstable.

Let us extract samples of different sizes from each experimental group and calculate the corresponding estimates [6, 15, 20].

Listing 3.6. Comparison of Estimates Obtained from Samples of Different Sizes

```
sample_sizes = [
    100,
    500,
    2000,
    5000,
    10000
]

group_a = df[df["group"] == "A"]
group_b = df[df["group"] == "B"]

sampling_results = []

for sample_size in sample_sizes:
    sample_a = group_a.sample(
        n=sample_size,
        random_state=sample_size
    )

    sample_b = group_b.sample(
        n=sample_size,
        random_state=sample_size
    )

    sampling_results.append({
        "sample_size_per_group": sample_size,
        "mean_session_A":
            sample_a["session_duration"].mean(),
        "mean_session_B":
            sample_b["session_duration"].mean(),
        "conversion_A":
            sample_a["converted"].mean(),
        "conversion_B":
            sample_b["converted"].mean()
    })

sampling_table = pd.DataFrame(
    sampling_results
)

print(sampling_table.round(4))
```

Output:

```
   sample_size_per_group  mean_session_A  mean_session_B  conversion_A  \
0                    100        199.4830        192.4100        0.1800
1                    500        183.4474        181.7218        0.1160
2                   2000        181.3903        190.2503        0.1210
3                   5000        182.3420        189.5766        0.1204
4                  10000        180.8152        189.7992        0.1188

   conversion_B
0        0.1600
1        0.1100
2        0.1200
```

```
3        0.1298
4        0.1289
```

For samples containing only 100 or 500 users per group, the estimated results differ considerably from the full-dataset values. In the sample containing 100 users, the estimated session duration and conversion rate are higher in group A, even though the complete dataset shows better values for group B. As the sample size increases, the estimates become more stable and approach the values calculated from the complete dataset.

This example demonstrates that small samples may produce misleading results because of random sampling variability. It also shows that a large amount of data does not automatically guarantee a correct conclusion. The sample must be selected randomly and must be sufficiently large.

7. Before selecting a statistical test, it is useful to examine the shape of the analysed distributions. The D'Agostino–Pearson test can be applied to session duration in each group [5].

Listing 3.7. Assessment of Normality for Session Duration

```
from scipy.stats import (
    skew,
    kurtosis,
    normaltest
)

for group_name, group_data in [
    ("A", session_group_a),
    ("B", session_group_b)
]:
    skewness = skew(group_data)

    excess_kurtosis = kurtosis(
        group_data
    )

    statistic, p_value = normaltest(
        group_data
    )

    print("Group:", group_name)
    print(
        "Skewness:",
        round(skewness, 4)
    )
    print(
        "Excess kurtosis:",
        round(excess_kurtosis, 4)
    )
    print(
        "D'Agostino-Pearson statistic:",
        round(statistic, 4)
    )
    print("p-value:", p_value)
    print()
```

Output:

```
Group: A
Skewness: 1.0112
Excess kurtosis: 1.5918
D'Agostino–Pearson statistic: 1561.8422
p-value: 0.0

Group: B
Skewness: 0.9503
Excess kurtosis: 1.2596
D'Agostino–Pearson statistic: 1362.8873
p-value: 1.1292296760748302e-296
```

Both groups have positive skewness, which indicates long right tails. The positive excess kurtosis indicates that the distributions have heavier tails than a normal distribution. The *p*-values are considerably smaller than 0.05. Therefore, the hypothesis of normality is rejected for both groups. This result agrees with the histograms. Session duration does not follow a normal distribution.

However, each group contains 10000 independent observations. With such large samples, the sampling distributions of the sample means are approximately normal according to the central limit theorem. Therefore, a two-sample *t*-test can still be used to compare the group means. A nonparametric test will also be applied as an additional check.

8. **Welch's two-sample *t*-test** is a statistical test used to compare the means of two independent groups. Welch's test does not require the population variances to be equal [3, 6, 15, 21].

It is generally preferred when:

- the groups are independent;
- the analysed variable is quantitative;
- the group variances may differ;
- the sample sizes are different or large;
- the main objective is to compare population means.

The hypotheses are:

- $H_0: \mu A = \mu B$
- $H_1: \mu A \neq \mu B$

We select the significance level $\alpha = 0.05$.

The decision rule is:

- if *p*-value > 0.05, there is insufficient evidence to reject H_0;
- if *p*-value ≤ 0.05, H_0 is rejected.

Listing 3.8. Welch's t-Test for Session Duration

```
from scipy.stats import ttest_ind
```

```
t_statistic, p_value = ttest_ind(
    session_group_b,
    session_group_a,
    equal_var=False
)

print(
    "Mean session duration in Group A:",
    round(session_group_a.mean(), 4)
)

print(
    "Mean session duration in Group B:",
    round(session_group_b.mean(), 4)
)

print(
    "Difference B - A:",
    round(
        session_group_b.mean()
        - session_group_a.mean(),
        4
    )
)

print(
    "t-statistic:",
    round(t_statistic, 4)
)

print("p-value:", p_value)

if p_value <= 0.05:
    print(
        "The difference between the group means "
        "is statistically significant."
    )
else:
    print(
        "There is insufficient evidence "
        "of a difference between the group means."
    )
```

Output:

```
Mean session duration in Group A: 180.8152
Mean session duration in Group B: 189.7992
Difference B - A: 8.984
t-statistic: 6.9018
p-value: 5.289004722741679e-12
The difference between the group means is statistically significant.
```

The obtained *p*-value is considerably smaller than 0.05. Therefore, the null hypothesis is rejected. The data provide evidence that the mean session durations in groups A and B are different.

The positive *t*-statistic indicates that the mean session duration in group B is higher than the mean session duration in group A.

The observed increase is approximately 8.98 seconds. However, statistical significance does not show whether the difference is large enough to be important in practice. A confidence interval and an effect size should also be calculated.

9. **A confidence interval** provides a range of plausible values for the difference between the population means. The estimated difference is: *difference = meanB – meanA*. For Welch's method standard error = *sqrt(sA² / nA + sB² / nB)/* The confidence interval is *difference ± t* · standard error* [3, 6, 15, 20, 21].

Listing 3.9. Confidence Interval for the Difference in Mean Session Duration

```
from scipy.stats import t
n_a = len(session_group_a)
n_b = len(session_group_b)

mean_a = session_group_a.mean()
mean_b = session_group_b.mean()
std_a = session_group_a.std(ddof=1)
std_b = session_group_b.std(ddof=1)

mean_difference = mean_b - mean_a

standard_error = np.sqrt(
    std_a ** 2 / n_a
    + std_b ** 2 / n_b
)

degrees_of_freedom = (
    (
        std_a ** 2 / n_a
        + std_b ** 2 / n_b
    ) ** 2
    /
    (
        (std_a ** 2 / n_a) ** 2
        / (n_a - 1)
        +
        (std_b ** 2 / n_b) ** 2
        / (n_b - 1)
    )
)

critical_value = t.ppf(
    0.975,
    df=degrees_of_freedom
)

margin_of_error = (
    critical_value
    * standard_error
)

lower_limit = (
    mean_difference
    - margin_of_error
)

upper_limit = (
    mean_difference
```

```
    + margin_of_error
)

print(
    "Mean difference:",
    round(mean_difference, 4)
)

print(
    "Standard error:",
    round(standard_error, 4)
)

print(
    "Degrees of freedom:",
    round(degrees_of_freedom, 4)
)

print(
    "95% confidence interval:",
    round(lower_limit, 4),
    "to",
    round(upper_limit, 4)
)
```

Output:

```
Mean difference: 8.984
Standard error: 1.3017
Degrees of freedom: 19979.7102
95% confidence interval: 6.4326 to 11.5354
```

The interval does not contain zero. Therefore, the result agrees with the Welch t-test. The data indicate that the new recommendation algorithm increases mean session duration by approximately 6.43 to 11.54 seconds.

10. An **effect size** is a numerical measure of the magnitude of a difference or relationship. A *p*-value indicates whether an observed difference is statistically significant, whereas an effect size indicates how large the difference is.

For two sample means, Cohen's d can be calculated as:

$$Cohen's\ d = (meanB - meanA) / pooled\ standard\ deviation$$

A commonly used approximate interpretation is:

- $|d| \approx 0.2$ small effect
- $|d| \approx 0.5$ moderate effect
- $|d| \approx 0.8$ large effect

These boundaries are general guidelines and should not replace domain-specific interpretation [4].

Listing 3.10. Calculation of Cohen's d

```
pooled_standard_deviation = np.sqrt(
    (
        (n_a - 1) * std_a ** 2
        + (n_b - 1) * std_b ** 2
    )
```

```
    /
    (
        n_a + n_b - 2
    )
)

cohens_d = (
    mean_b - mean_a
) / pooled_standard_deviation

print(
    "Pooled standard deviation:",
    round(
        pooled_standard_deviation,
        4
    )
)

print(
    "Cohen's d:",
    round(cohens_d, 4)
)
```

Output:

```
Pooled standard deviation: 92.0436
Cohen's d: 0.0976
```

The value of Cohen's d is approximately 0.098. This is a very small standardized effect. The difference is statistically significant because the dataset is large, but the increase in session duration is small relative to the variability of individual user sessions. This illustrates an important principle: **a statistically significant result is not necessarily practically important.**

11. **The Mann–Whitney U test** is a nonparametric test used to compare two independent distributions. It does not require the analysed variable to follow a normal distribution [12].

The hypotheses are:

- H_0: the two groups have the same distribution.
- H_1: the two groups have different distributions.

Listing 3.11. Mann–Whitney U Test for Session Duration

```
from scipy.stats import mannwhitneyu

u_statistic, p_value = mannwhitneyu(
    session_group_b,
    session_group_a,
    alternative="two-sided"
)

print(
    "Mann-Whitney U statistic:",
    u_statistic
)

print("p-value:", p_value)
```

```
if p_value <= 0.05:
    print(
        "The session-duration distributions "
        "are statistically different."
    )
else:
    print(
        "There is insufficient evidence "
        "of a difference between the distributions."
    )
```

Output:

```
Mann–Whitney U statistic: 52862153.0
p-value: 2.3723285890032035e-12
The session-duration distributions are statistically different.
```

The Mann–Whitney test also produces a p-value considerably smaller than 0.05. Therefore, the conclusion agrees with the Welch t-test. The session-duration distributions in groups A and B differ statistically. Using both tests increases confidence that the observed result is not caused only by the non-normal shape of the data.

12. A **conversion rate** is the proportion of users who complete a target action. In this experiment, the target action is a purchase. Let us calculate the conversion rates and the uplift produced by the new algorithm [11].

Listing 3.12. Calculation of Conversion Rates and Uplift

```
converted_a_count = group_a[
    "converted"
].sum()

converted_b_count = group_b[
    "converted"
].sum()

n_a = len(group_a)
n_b = len(group_b)

conversion_a = (
    converted_a_count / n_a
)

conversion_b = (
    converted_b_count / n_b
)

absolute_uplift = (
    conversion_b - conversion_a
)

relative_uplift = (
    absolute_uplift
    / conversion_a
    * 100
)
```

```
print(
    "Converted users in Group A:",
    converted_a_count
)

print(
    "Converted users in Group B:",
    converted_b_count
)

print(
    "Conversion rate in Group A:",
    round(conversion_a * 100, 2),
    "%"
)

print(
    "Conversion rate in Group B:",
    round(conversion_b * 100, 2),
    "%"
)

print(
    "Absolute uplift:",
    round(absolute_uplift * 100, 2),
    "percentage points"
)

print(
    "Relative uplift:",
    round(relative_uplift, 2),
    "%"
)
```

Output:

```
Converted users in Group A: 1188
Converted users in Group B: 1289
Conversion rate in Group A: 11.88 %
Conversion rate in Group B: 12.89 %
Absolute uplift: 1.01 percentage points
Relative uplift: 8.5 %
```

The absolute uplift is 12.89% − 11.88% = 1.01 percentage points. The relative uplift is 1.01 / 11.88 · 100% = 8.50%. The absolute and relative changes must not be confused.

An increase from 11.88% to 12.89% is:

- an increase of 1.01 percentage points;
- a relative increase of approximately 8.50%.

A statistical test is required to determine whether the observed difference could reasonably be explained by random variation.

13. The two-proportion z-test is used to compare two independent proportions [3, 6, 15, 20]. The hypotheses are:

- $H_0: pA = pB$
- $H_1: pA \neq pB$

Listing 3.13. Two-Proportion z-Test for Conversion Rates

```
from scipy.stats import norm

pooled_conversion = (
    converted_a_count
    + converted_b_count
) / (
    n_a + n_b
)

pooled_standard_error = np.sqrt(
    pooled_conversion
    * (1 - pooled_conversion)
    * (
        1 / n_a
        + 1 / n_b
    )
)

z_statistic = (
    conversion_b
    - conversion_a
) / pooled_standard_error

p_value = 2 * (
    1
    - norm.cdf(
        abs(z_statistic)
    )
)

print(
    "z-statistic:",
    round(z_statistic, 4)
)

print(
    "p-value:",
    round(p_value, 4)
)

if p_value <= 0.05:
    print(
        "The difference in conversion rates "
        "is statistically significant."
    )
else:
    print(
        "There is insufficient evidence "
        "of a difference in conversion rates."
    )
```

Output:

```
z-statistic: 2.168
p-value: 0.0302
The difference in conversion rates is statistically significant.
```

Since 0.0302 < 0.05 the null hypothesis is rejected. The data provide evidence that the conversion rates in groups A and B are different. The conversion rate is higher in group B. However, the *p*-value is relatively close to 0.05. Therefore, it is

especially useful to calculate a confidence interval for the difference in conversion rates.

14. Confidence Interval for the Conversion Uplift [3, 6, 15, 20]

Listing 3.14. Confidence Interval for the Conversion Uplift

```
conversion_difference = (
    conversion_b
    - conversion_a
)

standard_error_difference = np.sqrt(
    conversion_a
    * (1 - conversion_a)
    / n_a
    +
    conversion_b
    * (1 - conversion_b)
    / n_b
)

critical_value = norm.ppf(0.975)

lower_limit = (
    conversion_difference
    - critical_value
    * standard_error_difference
)

upper_limit = (
    conversion_difference
    + critical_value
    * standard_error_difference
)

print(
    "Conversion difference:",
    round(
        conversion_difference
        * 100,
        4
    ),
    "percentage points"
)

print(
    "95% confidence interval:",
    round(lower_limit * 100, 4),
    "to",
    round(upper_limit * 100, 4),
    "percentage points"
)
```

Output:

```
Conversion difference: 1.01 percentage points
95% confidence interval: 0.097 to 1.923 percentage points
```

Although the estimated uplift is positive, the lower confidence limit is close to zero. Therefore, additional experiment duration or a larger sample may be useful before a full deployment.

15. A product improvement should not be evaluated using business metrics alone. A new algorithm may improve user engagement but also increase computational load. In this experiment, the response time of group B is higher. Suppose that the engineering team has determined that an increase of more than 20 milliseconds would be practically unacceptable. Let us compare the mean response times using Welch's t-test and construct a confidence interval.

Listing 3.15. Statistical Analysis of Server Response Time

```
response_group_a = df.loc[
    df["group"] == "A",
    "response_time"
]

response_group_b = df.loc[
    df["group"] == "B",
    "response_time"
]

response_t, response_p = ttest_ind(
    response_group_b,
    response_group_a,
    equal_var=False
)

response_mean_a = (
    response_group_a.mean()
)

response_mean_b = (
    response_group_b.mean()
)

response_std_a = (
    response_group_a.std(ddof=1)
)

response_std_b = (
    response_group_b.std(ddof=1)
)

response_difference = (
    response_mean_b
    - response_mean_a
)

response_standard_error = np.sqrt(
    response_std_a ** 2
    / len(response_group_a)
    +
    response_std_b ** 2
    / len(response_group_b)
)
```

```
response_df = (
    (
        response_std_a ** 2
        / len(response_group_a)
        +
        response_std_b ** 2
        / len(response_group_b)
    ) ** 2
    /
    (
        (
            response_std_a ** 2
            / len(response_group_a)
        ) ** 2
        / (
            len(response_group_a) - 1
        )
        +
        (
            response_std_b ** 2
            / len(response_group_b)
        ) ** 2
        / (
            len(response_group_b) - 1
        )
    )
)

response_critical_value = t.ppf(
    0.975,
    df=response_df
)

response_lower = (
    response_difference
    - response_critical_value
    * response_standard_error
)

response_upper = (
    response_difference
    + response_critical_value
    * response_standard_error
)

print(
    "Mean response time in Group A:",
    round(response_mean_a, 2)
)

print(
    "Mean response time in Group B:",
    round(response_mean_b, 2)
)

print(
    "Difference B - A:",
    round(response_difference, 2)
)

print(
```

```
    "t-statistic:",
    round(response_t, 4)
)

print(
    "p-value:",
    response_p
)

print(
    "95% confidence interval:",
    round(response_lower, 2),
    "to",
    round(response_upper, 2)
)
```

Output:

```
Mean response time in Group A: 220.33
Mean response time in Group B: 235.21
Difference B - A: 14.88
t-statistic: 28.9673
p-value: 9.7104855257911e-181
95% confidence interval: 13.87 to 15.89
```

The increase in server response time is statistically significant. The estimated increase is approximately 14.88 milliseconds. The 95% confidence interval is: [13.87, 15.89] milliseconds. The entire confidence interval is below the assumed unacceptable threshold of 20 milliseconds.

Therefore, the response-time increase is statistically detectable but remains within the engineering tolerance defined for this experiment.

16. The **bootstrap** is a resampling method used to estimate the sampling distribution of a statistic [7, 8].

The main procedure is:

1. Draw a sample from the observed data with replacement.
2. Calculate the statistic of interest.
3. Repeat the procedure many times.
4. Analyse the distribution of the obtained statistics.
5. Use percentiles of the bootstrap distribution to construct a confidence interval.

The bootstrap is useful when:

- an analytical confidence interval is difficult to derive;
- the data are not normally distributed;
- the statistic has a complex form;
- the sample is sufficiently representative of the population.

Let us construct a bootstrap confidence interval for the difference in mean session duration.

Listing 3.16. Bootstrap Confidence Interval for the Difference in Mean Session Duration

```
bootstrap_rng = np.random.default_rng(123)
number_of_bootstrap_samples = 2000

bootstrap_differences = np.empty(
    number_of_bootstrap_samples
)

session_a_array = (
    session_group_a.to_numpy()
)

session_b_array = (
    session_group_b.to_numpy()
)

for i in range(
    number_of_bootstrap_samples
):
    bootstrap_sample_a = (
        bootstrap_rng.choice(
            session_a_array,
            size=len(session_a_array),
            replace=True
        )
    )

    bootstrap_sample_b = (
        bootstrap_rng.choice(
            session_b_array,
            size=len(session_b_array),
            replace=True
        )
    )

    bootstrap_differences[i] = (
        bootstrap_sample_b.mean()
        - bootstrap_sample_a.mean()
    )

bootstrap_lower = np.percentile(
    bootstrap_differences,
    2.5
)

bootstrap_upper = np.percentile(
    bootstrap_differences,
    97.5
)

print(
    "Observed mean difference:",
    round(
        session_group_b.mean()
        - session_group_a.mean(),
        4
    )
)

print(
    "Mean bootstrap difference:",
    round(
        bootstrap_differences.mean(),
```

```
        4
    )
)

print(
    "95% bootstrap confidence interval:",
    round(bootstrap_lower, 4),
    "to",
    round(bootstrap_upper, 4)
)
```

Output:

```
Observed mean difference: 8.984
Mean bootstrap difference: 8.998
95% bootstrap confidence interval: 6.496 to 11.468
```

The bootstrap confidence interval is [6.50, 11.47] seconds. This interval is very close to the analytical Welch confidence interval [6.43, 11.54] seconds. Both methods indicate that the mean session duration is higher in group B.

Listing 3.17. Visualization of the Bootstrap Distribution

```
plt.figure(figsize=(10, 6))

plt.hist(
    bootstrap_differences,
    bins=40,
    edgecolor="black"
)

plt.axvline(
    bootstrap_lower,
    linestyle="--",
    label="Lower Limit"
)

plt.axvline(
    bootstrap_upper,
    linestyle="--",
    label="Upper Limit"
)

plt.axvline(
    bootstrap_differences.mean(),
    linewidth=2,
    label="Mean Difference"
)

plt.title(
    "Bootstrap Distribution of the Difference in Mean Session Duration"
)

plt.xlabel(
    "Difference in Mean Session Duration, B - A"
)

plt.ylabel("Frequency")
plt.legend()
plt.show()
```

Output:

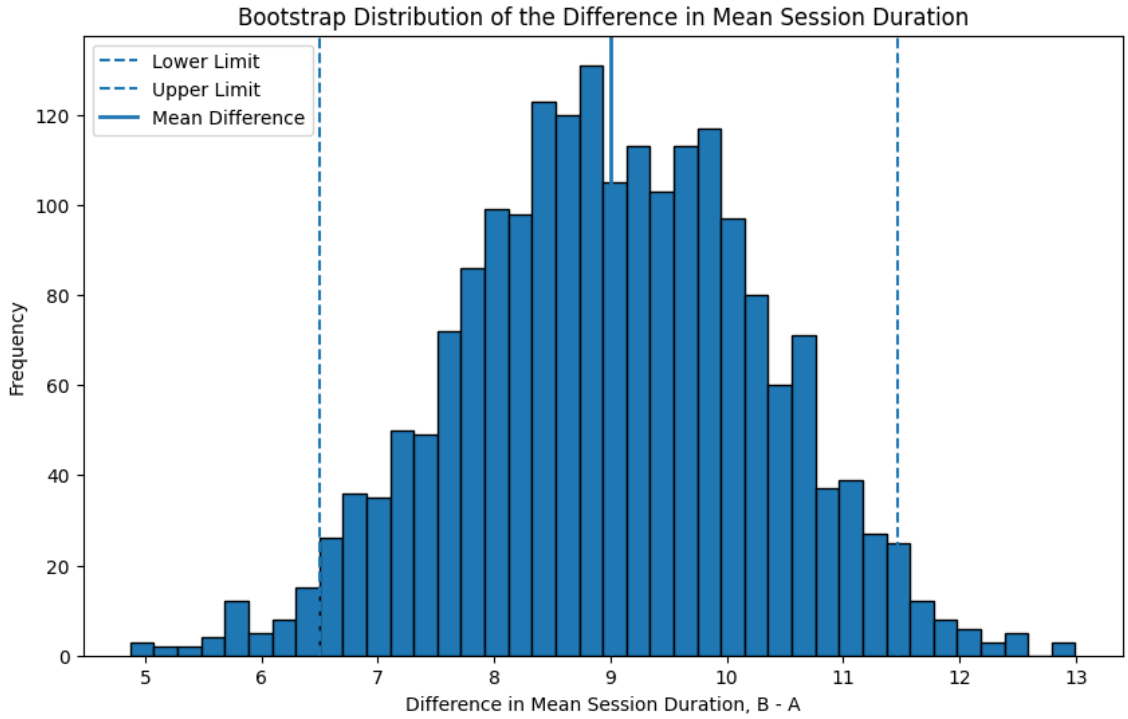


Figure 46. Bootstrap Distribution of the Difference in Mean Session Duration

The bootstrap distribution is centred near the observed difference of approximately nine seconds. The entire 95% bootstrap interval is greater than zero, which supports the conclusion that group B has a higher mean session duration [4, 11].

Statistical significance and practical significance are different concepts. A statistically significant result indicates that the observed difference would be unlikely under the null hypothesis. A practically significant result indicates that the difference is large enough to influence a business, engineering, or product decision.

In this experiment:

Session duration

- Difference: +8.98 seconds
- *p*-value: considerably below 0.05
- Cohen's d: 0.0976

The difference is statistically significant, but the standardized effect is very small.

Conversion rate

- Group A: 11.88%
- Group B: 12.89%
- Absolute uplift: 1.01 percentage points
- Relative uplift: 8.50%
- *p*-value: 0.0302

The conversion improvement is statistically significant. Its business importance depends on:

- the number of users;
- the profit obtained from each conversion;
- the cost of operating the new algorithm;
- the expected long-term effect.

Server response time

- Increase: 14.88 milliseconds
- 95% confidence interval: [13.87, 15.89]

The increase is statistically significant but remains below the assumed practical limit of 20 milliseconds.

A large dataset may produce a very small p-value even when the actual effect is small. Therefore, an engineering decision should not be based on p-values alone.

The new recommendation algorithm produces the following results:

- mean session duration increases;
- the number of viewed pages increases;
- the conversion rate increases;
- average order value increases;
- server response time also increases.

The session-duration effect is statistically significant but small according to Cohen's *d*.

The conversion increase is statistically significant at the 5% level, but the lower confidence limit is close to zero.

The server response-time increase is statistically significant, but its confidence interval remains below the assumed unacceptable threshold.

Based on the obtained results, a reasonable engineering decision would be to deploy the new algorithm gradually rather than immediately replacing the old algorithm for all users.

For example, the company could:

- increase the experimental traffic from 50% to 70%;
- monitor response time and server load;
- continue collecting conversion data;
- analyse results separately for mobile and desktop users;
- investigate whether the effect differs between new and returning users;
- stop the rollout if technical metrics exceed predefined limits.

This approach takes into account both statistical evidence and engineering risk.

Conclusion

A synthetic A/B testing dataset containing 20000 users was analysed in this laboratory work. The users were divided equally between the control group A and the experimental group B. The analysis showed that the new recommendation algorithm increased mean session duration from approximately 180.82 to 189.80

seconds. Welch's two-sample t-test and the Mann–Whitney U test both detected a statistically significant difference between the groups.

The 95% confidence interval indicated that the increase in mean session duration was between approximately 6.43 and 11.54 seconds. However, Cohen's d was only 0.098, which means that the standardized effect was very small.

The conversion rate increased from 11.88% to 12.89%. The absolute uplift was 1.01 percentage points, and the relative uplift was approximately 8.50%. The two-proportion z-test showed that this difference was statistically significant at the 5% level.

The new algorithm also increased average server response time by approximately 14.88 milliseconds. This increase was statistically significant but remained below the assumed practical limit of 20 milliseconds.

The bootstrap confidence interval agreed closely with the analytical confidence interval. The results demonstrate that statistical significance, effect size, confidence intervals, technical limitations, and business importance should all be considered when making decisions based on A/B testing and large-scale user data.

Tasks for Independent Work.

A/B Testing and Statistical Analysis of an Adaptive Code-Hint Algorithm.

An online platform for learning programming provides students with exercises in Python, Java, algorithms, and data structures. The platform is testing a new adaptive code-hint algorithm. The algorithm analyses a student's previous answers and selects hints that are expected to help the student complete programming exercises more effectively.

Users are randomly divided into two groups:

- Group A — the current code-hint algorithm
- Group B — the new adaptive code-hint algorithm

The company wants to determine whether the new algorithm:

- increases the duration of learning sessions;
- increases the number of completed programming tasks;
- increases the probability that a user purchases a paid subscription;
- increases subscription revenue;
- does not cause an unacceptable increase in server response time.

The main research question is:

Does the new adaptive code-hint algorithm improve user engagement and subscription conversion without causing an unacceptable decrease in system performance?

The dataset contains information about 24,000 users:

- 12,000 users in Group A

- 12,000 users in Group B

The dataset contains the following variables:

- user_id (unique user identifier)
- group (experimental group A or B)
- session_duration (duration of the learning session in seconds)
- tasks_completed (number of programming tasks completed during the session)
- subscribed (whether the user purchased a paid subscription)
- subscription_value (value of the purchased subscription)
- response_time (average server response time in milliseconds)

The variable subscribed is binary:

- 0 — the user did not purchase a subscription
- 1 — the user purchased a subscription

For users who did not purchase a subscription, subscription_value is equal to zero.

1. Generate the A/B Testing Dataset.

Use the following program code to generate the dataset:

```
import numpy as np
import pandas as pd
rng = np.random.default_rng(84)
n_per_group = 12000

# Generate session duration in seconds
session_a = rng.gamma(
    shape=3.8,
    scale=42.0,
    size=n_per_group
)
session_b = rng.gamma(
    shape=4.0,
    scale=42.0,
    size=n_per_group
)

# Generate the number of completed tasks
tasks_a = rng.poisson(
    lam=5.8,
    size=n_per_group
)
tasks_b = rng.poisson(
    lam=6.3,
    size=n_per_group
)
# Generate subscription indicators
subscribed_a = rng.binomial(
    n=1,
    p=0.095,
    size=n_per_group
)
subscribed_b = rng.binomial(
    n=1,
    p=0.106,
    size=n_per_group
```

```
)
# Generate subscription values
subscription_a = np.where(
    subscribed_a == 1,
    rng.lognormal(
        mean=np.log(49),
        sigma=0.40,
        size=n_per_group
    ),
    0
)
subscription_b = np.where(
    subscribed_b == 1,
    rng.lognormal(
        mean=np.log(52),
        sigma=0.40,
        size=n_per_group
    ),
    0
)
# Generate server response time in milliseconds
response_a = np.clip(
    rng.normal(
        loc=205,
        scale=30,
        size=n_per_group
    ),
    70,
    None
)
response_b = np.clip(
    rng.normal(
        loc=217,
        scale=33,
        size=n_per_group
    ),
    70,
    None
)
# Create a DataFrame
df = pd.DataFrame({
    "user_id": np.arange(
        1,
        2 * n_per_group + 1
    ),
    "group": np.repeat(
        ["A", "B"],
        n_per_group
    ),
    "session_duration": np.concatenate(
        [session_a, session_b]
    ).round(1),
    "tasks_completed": np.concatenate(
        [tasks_a, tasks_b]
    ),
    "subscribed": np.concatenate(
        [subscribed_a, subscribed_b]
    ),
    "subscription_value": np.concatenate(
        [subscription_a, subscription_b]
    ).round(2),
    "response_time": np.concatenate(
```

```
        [response_a, response_b]
    ).round(1)
})
print(df.head())
print("Dataset shape:", df.shape)
```

Requirements

1. Display the first five rows of the DataFrame.
2. Display the number of rows and columns.
3. Explain the purpose of:
    - np.random.default_rng();
    - rng.gamma();
    - rng.poisson();
    - rng.binomial();
    - rng.lognormal();
    - np.where();
    - np.clip();
    - np.concatenate();
    - np.repeat().
4. Explain why different probability distributions are used for different variables.
5. Explain why a random seed is required.

2. Assess the Quality of the Dataset.

Check the dataset for:

- missing values;
- duplicate user_id values;
- non-positive session durations;
- negative numbers of completed tasks;
- incorrect subscription indicators;
- negative subscription values;
- non-positive response times;
- inconsistency between subscribed and subscription_value.

Questions

1. Does the dataset contain missing values?
2. Are all user identifiers unique?
3. Are all quantitative measurements valid?
4. Are the subscription variables logically consistent?
5. Why should data quality be checked before statistical testing?
6. How could invalid observations affect sample means, conversion rates, and p-values?

3. Calculate Descriptive Statistics by Experimental Group

Calculate the following statistics separately for groups A and B:

- number of users;
- mean session duration;
- median session duration;
- standard deviation of session duration;
- mean number of completed tasks;
- median number of completed tasks;
- subscription conversion rate;
- mean server response time;
- standard deviation of server response time;
- mean subscription value among subscribed users;
- mean subscription revenue per user.

Compare groups A and B and answer the following questions:

1. Which group has the greater mean session duration?
2. Which group has the greater median session duration?
3. Which group completes more programming tasks on average?
4. Which group has the greater subscription conversion rate?
5. Which group has the greater mean subscription value?
6. Which group has the greater server response time?
7. Do the descriptive results suggest that the new algorithm is better?
8. Why are descriptive statistics insufficient for making the final conclusion?

4. Visualize Session Duration.

Construct overlapping histograms of session_duration for groups A and B.

Describe:

1. the general shape of both distributions;
2. whether the distributions are symmetric;
3. whether long right tails are visible;
4. whether one distribution appears shifted relative to the other;
5. whether the distributions overlap;
6. whether the histogram alone proves that the group means are different.

5. Construct Box Plots.

Construct box plots for session duration in groups A and B.

Questions

1. Which group has the higher median?
2. Which group has the larger interquartile range?
3. Are potential outliers present?
4. Does showfliers=False remove observations from the dataset?
5. Why can a box plot be more informative than a comparison of means alone?

6. Study the Influence of Sample Size.

Select random samples of the following sizes from each experimental group:

- 100
- 500
- 2000
- 6000
- 12000

For each sample size, calculate:

- mean session duration in group A;
- mean session duration in group B;
- difference between the two sample means;
- subscription conversion rate in group A;
- subscription conversion rate in group B;
- difference between the two conversion rates.

Explain:

1. how the estimates change when the sample size increases;
2. why estimates based on 100 users may differ considerably from the complete-dataset results;
3. whether a small sample can indicate the wrong direction of the effect;
4. why a large dataset does not remove the need for random sampling;
5. how sampling can reduce computational costs in Big Data systems.

7. Assess the Distribution of Session Duration.

Calculate separately for each group:

- skewness;
- excess kurtosis;
- D'Agostino–Pearson statistic;
- p-value.

Questions

1. What does positive skewness indicate?
2. What does negative skewness indicate?
3. What does positive excess kurtosis indicate?
4. What does excess kurtosis close to zero indicate?
5. Should the hypothesis of normality be rejected?
6. Does the conclusion agree with the histogram?
7. Why can formal normality tests detect very small deviations in large datasets?
8. Why may a t-test still be used to compare means when both samples are large?

8. Apply Welch's Two-Sample t-Test.

Use Welch's two-sample t-test to compare mean session duration in groups A and B.

Questions

1. Why is Welch's test used instead of the classical equal-variance t-test?

2. What does the sign of the t-statistic indicate?
3. Does the sign determine statistical significance?
4. Is the difference between the group means statistically significant?
5. Does statistical significance show that the difference is important in practice?

9. Construct a 95% Confidence Interval for the Difference Between Means.

Construct a 95% confidence interval for $\mu B - \mu A$.

Calculate:

- sample sizes;
- group means;
- group standard deviations;
- difference between sample means;
- standard error of the difference;
- Welch–Satterthwaite degrees of freedom;
- critical t-value;
- margin of error;
- lower confidence limit;
- upper confidence limit.

Interpretation

1. Does the confidence interval contain zero?
2. What does an interval entirely above zero indicate?
3. Is the confidence-interval conclusion consistent with Welch's t-test?
4. What range of increases in mean session duration is plausible?
5. Why is an interval estimate more informative than a p-value alone?

10. Calculate Cohen's d.

Calculate Cohen's d for the difference in session duration.

Questions

1. Is the standardized effect small, moderate, or large?
2. Can a very small effect have a very small p-value?
3. Why does sample size influence statistical significance?
4. Why should practical importance be evaluated using domain knowledge?
5. Could a small effect still be valuable for a platform with millions of users?

11. Apply the Mann–Whitney U Test.

Apply the Mann–Whitney U test to compare the session-duration distributions.

Questions

1. Why is the Mann–Whitney test called nonparametric?
2. Does it require a normal distribution?
3. Is it strictly a test of equality of means?
4. Does its conclusion agree with Welch's t-test?
5. Why is it useful to apply both a parametric and a nonparametric test?

12. Calculate Subscription Conversion Rates

Calculate:

- number of subscribed users in group A;
- number of subscribed users in group B;
- subscription conversion rate in group A;
- subscription conversion rate in group B;
- absolute uplift;
- relative uplift.

Explain the difference between:

- percentage;
- percentage point;
- absolute uplift;
- relative uplift.

13. Apply a Two-Proportion z-Test.

Use a two-proportion z-test to determine whether subscription conversion rates differ between groups A and B.

Questions

1. Why is a proportion test used instead of the ordinary t-test?
2. Why is the conversion indicator considered a binary variable?
3. Should the null hypothesis be rejected?
4. Is the conversion rate higher in group B?
5. Does a statistically significant conversion increase automatically justify deployment?

14. Construct a Confidence Interval for the Difference in Subscription Conversion Rates.

Construct a 95% confidence interval for $pB - pA$.

Interpretation

1. Does the interval contain zero?
2. Is the conclusion consistent with the two-proportion z-test?
3. What is the smallest plausible conversion improvement?
4. What is the largest plausible conversion improvement?
5. Is the lower limit sufficiently large to be important for the company?
6. Would a longer experiment provide a narrower interval?

15. Analyse Server Response Time.

Assume that an increase in mean response time greater than 18 milliseconds is considered operationally unacceptable. Compare server response times in groups A and B.

Calculate:

- mean response time in group A;

- mean response time in group B;
- difference meanB − meanA;
- Welch t-statistic;
- p-value;
- 95% confidence interval for the difference;
- Cohen's d.

Questions

1. Is the increase statistically significant?
2. Is the observed increase greater than 18 milliseconds?
3. Does the 95% confidence interval include values greater than 18 milliseconds?
4. Is the increase within the engineering tolerance?
5. Can a technically small increase be statistically significant in a large dataset?
6. Should system-performance metrics be analysed together with business metrics?

16. Construct a Bootstrap Confidence Interval.

Construct a bootstrap confidence interval for the difference in mean session duration. Use at least 2000 bootstrap samples.

For every bootstrap iteration:

1. sample observations from group A with replacement;
2. sample observations from group B with replacement;
3. calculate the difference between the resampled means;
4. store the obtained difference.

Calculate:

- observed difference between means;
- mean bootstrap difference;
- 2.5th percentile;
- 97.5th percentile.

Construct a histogram of the bootstrap differences.

Questions

1. Is the bootstrap distribution centred near the observed difference?
2. Does the bootstrap confidence interval contain zero?
3. Is the bootstrap interval similar to the analytical Welch interval?
4. Why is sampling performed with replacement?
5. Why is the bootstrap method useful for non-normal data?
6. What happens to the stability of the interval when the number of bootstrap iterations increases?

17. Compare Statistical and Practical Significance

Answer the following questions:

1. Which differences are statistically significant?
2. Which effects are large enough to be practically important?
3. Can a statistically significant effect be practically negligible?
4. Can a practically important effect fail to reach statistical significance?
5. Which metric provides the strongest evidence in favour of the new algorithm?
6. Which metric creates the greatest engineering risk?
7. Why should the decision not be based on one p-value alone?

18. Make the Final Engineering Decision.

Based on all obtained results, select one of the following decisions:

1. Deploy the new algorithm for all users.
2. Reject the new algorithm.
3. Continue the experiment and collect more data.
4. Deploy the new algorithm gradually.
5. Modify the algorithm and repeat the experiment.

The decision must be justified using:

- descriptive statistics;
- Welch's t-test;
- the confidence interval for the difference between means;
- Cohen's d;
- the Mann–Whitney U test;
- conversion-rate analysis;
- the two-proportion z-test;
- the confidence interval for conversion uplift;
- the server response-time analysis;
- the bootstrap confidence interval;
- statistical significance;
- practical significance;
- engineering risk.

The final decision must not be based only on whether p-value < 0.05.

BIBLIOGRAPHY

1. Bertsekas, D. P. Introduction to Probability / D. P. Bertsekas, J. N. Tsitsiklis. – 1st ed. – Cambridge, MA : Athena Scientific, 2002. – 430 p. – ISBN 978-1-886529-40-3.
2. Blitzstein, J. K. Introduction to Probability / J. K. Blitzstein, J. Hwang. – 2nd ed. – Boca Raton : Chapman & Hall/CRC, 2019. – 634 p. – ISBN 978-1-138-36991-7.
3. Casella, G. Statistical Inference / G. Casella, R. L. Berger. – 2nd ed. – Pacific Grove : Duxbury, 2002. – 660 p. – ISBN 978-0-534-24312-8.
4. Cohen, J. Statistical Power Analysis for the Behavioral Sciences / J. Cohen. – 2nd ed. – Hillsdale : Lawrence Erlbaum Associates, 1988. – 567 p. – ISBN 978-0-8058-0283-2.
5. D'Agostino, R. B. Tests for Departure from Normality. Empirical Results for the Distributions of b^2 and $\sqrt{b_1}$ / R. B. D'Agostino, E. S. Pearson. // Biometrika. – 1973. – Vol. 60, no. 3. – P. 613–622. – DOI: 10.1093/biomet/60.3.613.
6. Devore, J. L. Probability and Statistics for Engineering and the Sciences / J. L. Devore. – 9th ed. – Boston : Cengage Learning, 2016. – 768 p. – ISBN 978-1-305-25180-9.
7. Efron, B. An Introduction to the Bootstrap / B. Efron, R. J. Tibshirani. – New York : Chapman & Hall, 1993. – 436 p. – (Monographs on Statistics and Applied Probability ; 57). – ISBN 978-0-412-04231-7.
8. Efron, B. Bootstrap Methods: Another Look at the Jackknife / B. Efron.// The Annals of Statistics. – 1979. – Vol. 7, no. 1. – P. 1–26. – DOI: 10.1214/aos/1176344552.
9. Halmos, P. R. Naive Set Theory / P. R. Halmos. – New York : Springer-Verlag, 1974. – VII, 104 p. – (Undergraduate Texts in Mathematics). – ISBN 978-0-387-90092-6. – DOI: 10.1007/978-1-4757-1645-0.
10. Hunter, J. D. Matplotlib: A 2D Graphics Environment / J. D. Hunter. // Computing in Science & Engineering. – 2007. – Vol. 9, no. 3. – P. 90–95. – DOI: 10.1109/MCSE.2007.55.
11. Kohavi, R. Trustworthy Online Controlled Experiments : A Practical Guide to A/B Testing / R. Kohavi, D. Tang, Y. Xu. – Cambridge : Cambridge University Press, 2020. – 288 p. – ISBN 978-1-108-72426-5. – DOI: 10.1017/9781108653985.
12. Mann, H. B. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other / H. B. Mann, D. R. Whitney. // The Annals of Mathematical Statistics. – 1947. – Vol. 18, no. 1. – P. 50–60. – DOI: 10.1214/aoms/1177730491.
13. McKinney, W. Data Structures for Statistical Computing in Python / W. McKinney. // Proceedings of the 9th Python in Science Conference / edited by S.

van der Walt, J. Millman. – Austin : SciPy, 2010. – P. 56–61. – DOI: 10.25080/Majora-92bf1922-00a.

14. NIST/SEMATECH e-Handbook of Statistical Methods / National Institute of Standards and Technology, SEMATECH. – Gaithersburg, 2002– . – URL: https://www.itl.nist.gov/div898/handbook/ (date of access: 31.07.2026). – DOI: 10.18434/M32189.

15. Rice, J. A. Mathematical Statistics and Data Analysis / J. A. Rice. – 3rd ed. – Belmont : Thomson Brooks/Cole, 2007. – 688 p. – ISBN 978-0-534-39942-9.

16. Rosen, K. H. Discrete Mathematics and Its Applications / K. H. Rosen. – 8th ed. – New York : McGraw-Hill Education, 2019. – 942 p. – ISBN 978-1-259-67651-2.

17. Ross, S. M. A First Course in Probability / S. M. Ross. – 10th ed. – Boston : Pearson, 2019. – 505 p. – ISBN 978-0-13-475311-9.

18. Shapiro, S. S. An Analysis of Variance Test for Normality (Complete Samples) / S. S. Shapiro, M. B. Wilk. // Biometrika. – 1965. – Vol. 52, no. 3–4. – P. 591–611. – DOI: 10.1093/biomet/52.3-4.591.

19. Student. The Probable Error of a Mean / Student. // Biometrika. – 1908. – Vol. 6, no. 1. – P. 1–25. – DOI: 10.1093/biomet/6.1.1.

20. Wasserman, L. All of Statistics : A Concise Course in Statistical Inference / L. Wasserman. – New York : Springer, 2004. – XX, 442 p. – (Springer Texts in Statistics). – ISBN 978-0-387-40272-7. – DOI: 10.1007/978-0-387-21736-9.

21. Welch, B. L. The Generalisation of “Student’s” Problem When Several Different Population Variances Are Involved / B. L. Welch. // Biometrika. – 1947. – Vol. 34, no. 1–2. – P. 28–35. – DOI: 10.1093/biomet/34.1-2.28.

APPENDIX I.

Table of Values of the Standard Normal Density Function $\varphi(x)$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3925	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3725	0.3712	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.3467	0.3448	0.3429	0.3410	0.3391	0.3372	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2897	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1.0	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1456	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.1040	0.1023	0.1006	0.0989	0.0973	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551
2.0	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0194	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0088	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.0061
2.9	0.0060	0.0058	0.0056	0.0055	0.0053	0.0051	0.0050	0.0048	0.0047	0.0046
3.0	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006
3.6	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004
3.7	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003
3.8	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002
3.9	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
4.0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

APPENDIX II.

The table of values of the Laplace function $\Phi(x)$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
4.0	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Study guide

I.A. Zemliakova

Probability Theory and Mathematical Statistics

Заказ № ____ от _____ г.
Усл. печ. лист. ___. Уч. изд. л. ____.
Отдел полиграфической, корпоративной и сувенирной продукции
Издательско-полиграфического комплекса КИБИ МЕДИА ЦЕНТРА ЮФУ.
344090, г. Ростов-на-Дону, пр. Стачки, 200/1, тел (863) 243-41-66.