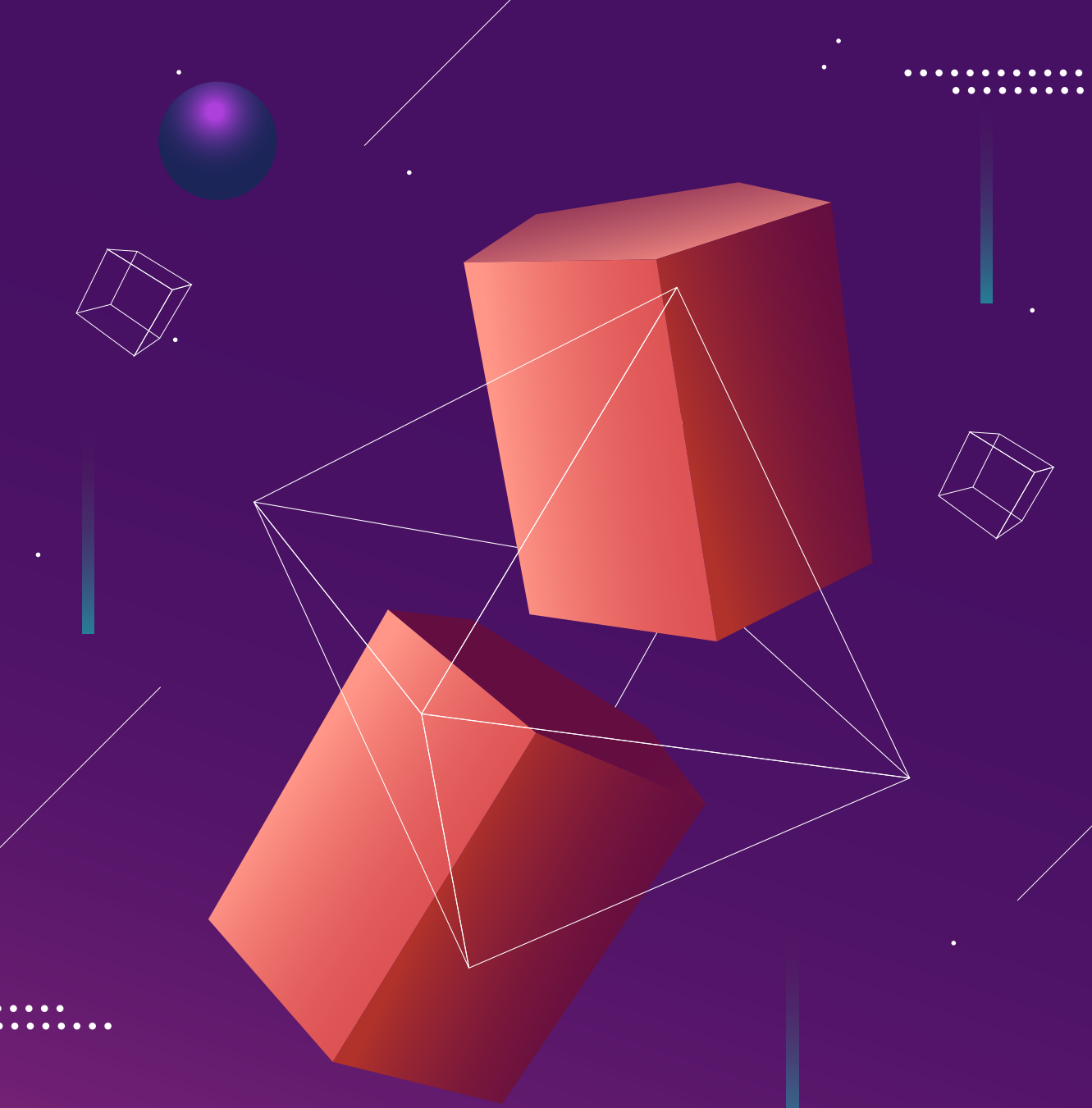




Mathematical Statistics with examples in R

Aneta Velkoska, Ninoslav Marina,
Atanas Hristov, Ljubinka Sandjakoska

Mathematical Statistics with examples in R

Ohrid 2025



**UNIVERSITY OF INFORMATION SCIENCE AND
TECHNOLOGY “ST. PAUL THE APOSTLE” OHRID**

**FACULTY OF COMMUNICATION NETWORKS AND
SECURITY**

MATHEMATICAL STATISTICS WITH EXAMPLES IN R

Aneta Velkoska, Ninoslav Marina,
Atanas Hristov, Ljubinka Sandjakoska

Ohrid 2025

Authors:

Assoc. Prof. Aneta Velkoska, PhD

Prof. Ninoslav Marina, PhD

Assoc. Prof. Atanas Hristov, PhD

Ass. Prof. Ljubinka Sandjakoska, PhD

Title of the issue:

Mathematical Statistics with examples in R

Editor of publishing activity of UINT:

Prof. Ivan Bimbilovski, PhD

Editor of the issue:

Assoc. Prof. Aneta Velkoska, PhD

Reviewers:

Prof. Zoran Misajleski, PhD

Prof. Daniel Velinov, PhD

Cover design:

Jona Nasufi

Technical processing:

Assoc. Prof. Atanas Hristov, PhD

Proofreading in English language:

Assis. Prof. Lela Ivanovska, PhD

Publisher:

University of Information Science and Technology “St. Paul the Apostle”

Place and year of issue:

Ohrid, 2025

CONTENTS

1.	INTRODUCTION TO STATISTICS	1
1.1.	IMPORTANCE OF STATISTICS	1
1.2.	POPULATION, SAMPLE, RANDOM SAMPLING	3
1.3.	VARIABLES. TYPES OF VARIABLES	5
2.	DESCRIPTIVE STATISTICS	8
2.1.	ORGANIZING QUALITATIVE DATA	8
2.2.	ORGANIZING QUANTITATIVE DATA	10
2.3.	NUMERICAL DESCRIPTIVE STATISTICS	25
	2.3.1. MEASURES OF LOCATION	26
	2.3.2. MEASURES OF DISPERSION	30
3.	SAMPLE SPACE AND PROBABILITY	38
3.1.	RANDOM EXPERIMENT AND SAMPLE SPACE	38
3.2.	TECHNIQUES OF COUNTING	44
	3.2.1. PERMUTATIONS	45
	3.2.2. COMBINATIONS	47
3.3.	PROBABILITY	49
	3.3.1. OTHER PROBABILITIES	55
4.	RANDOM VARIABLES	63
4.1.	DISCRETE RANDOM VARIABLES	65
4.2.	CONTINUOUS RANDOM VARIABLES	76
4.3.	SPECIAL DISCRETE RANDOM VARIABLES	84
	4.3.1. BERNOULLI DISTRIBUTION	84
	4.3.2. BINOMIAL DISTRIBUTION	85
	4.3.3. GEOMETRIC DISTRIBUTION	86
	4.3.4. POISSON DISTRIBUTION	88
4.4.	SPECIAL CONTINUOUS RANDOM VARIABLES	92
	4.4.1. UNIFORM DISTRIBUTION ON AN INTERVAL	92

4.4.2. EXPONENTIAL DISTRIBUTION	94
4.4.3. GAUSSIAN OR NORMAL DISTRIBUTION	96
4.4.4. CHI-SQUARE DISTRIBUTION	101
4.4.5. STUDENT'S DISTRIBUTION	104
5. THE CENTRAL LIMIT THEOREM	107
5.1. THE LAW OF LARGE NUMBERS	107
5.2. CENTRAL LIMIT THEOREM	111
6. POINT ESTIMATION. CONFIDENCE INTERVAL	119
6.1. POINT ESTIMATE	119
6.2. MAXIMUM LIKELIHOOD ESTIMATION	121
6.3. METHOD OF MOMENTS	132
6.4. CONFIDENCE INTERVAL OF THE MEAN OF THE POPULATION	134
6.4.1. CONFIDENCE INTERVAL OF THE MEAN OF THE POPULATION μ WHILE THE VARIANCE OF THE POPULATION σ^2 IS KNOWN	134
6.4.2. CONFIDENCE INTERVAL OF THE MEAN μ OF THE NORMALLY DISTRIBUTED POPULATION WHILE THE VARIANCE OF THE POPULATION σ^2 IS UNKNOWN	141
6.5. CONFIDENCE INTERVAL FOR THE POPULATION STANDARD DEVIATION	146
6.5.1. CONFIDENCE INTERVAL FOR THE STANDARD DEVIATION OF NORMALLY DISTRIBUTED POPULATION WITH KNOWN MEAN	146
6.5.2. CONFIDENCE INTERVAL FOR THE STANDARD DEVIATION OF NORMALLY DISTRIBUTED POPULATION WITH UNKNOWN MEAN	149
6.6. CONFIDENCE INTERVAL FOR THE POPULATION PROPORTION	150
7. HYPOTHESIS TESTING	154

7.1.	ONE SAMPLE HYPOTHESIS TESTING	161
	7.1.1. TESTS ABOUT POPULATION MEAN	161
	7.1.2. TESTS FOR THE POPULATION PROPORTION	178
	7.1.3. TESTS FOR THE POPULATION VARIANCE	183
7.2.	TWO SAMPLE HYPOTHESIS TESTING	188
	7.2.1. TESTS FOR EQUALITY OF POPULATION MEANS WITH INDEPENDENT SAMPLES	188
	7.2.2. COMPARING POPULATION MEANS WITH DEPENDENT SAMPLES	207
8.	STATISTICAL INFERENCE IN SIMPLE LINEAR CORRELATION AND REGRESSION. CHI-SQUARE TEST	221
8.1.	SIMPLE LINEAR CORRELATION AND REGRESSION	221
	8.1.1. DETERMINISTIC AND RANDOM RELATIONSHIP	222
	8.1.2. GOALS OF THE CORRELATION AND REGRESSION	222
8.2.	CORRELATION ANALYSIS	223
	8.2.1. SAMPLE CORRELATION COEFFICIENT	224
	8.2.2. HYPOTHESIS TESTING FOR THE SAMPLE CORRELATION COEFFICIENT	228
8.3.	REGRESSION ANALYSIS	231
	8.3.1. POINT ESTIMATES OF SIMPLE LINEAR REGRESSION PARAMETERS AND LINEAR REGRESSION LINES	234
8.4.	CHI square TEST OF INDEPENDENCE	243
9.	PRACTICAL GUIDE TO STATISTICS IN R	249
9.1.	INTRODUCTION TO R ESSENTIALS AND ENVIRONMENT	249
	9.1.1. WHAT IS R AND WHY USE IT?	249
	9.1.2. INSTALLING R and Rstudio	250
	9.1.3. BASICS OF THE R ENVIRONMENT	251

9.1.4.	FIRST STEPS WITH R: SIMPLE COMMANDS	252
9.1.5.	R Packages: EXTENDING FUNCTIONALITY	255
10.	PROBABILITY DISTRIBUTIONS IN R	260
10.1.	BASIC R FUNCTIONS FOR DISTRIBUTIONS	260
10.1.1.	PROBABILITY DENSITY/MASS FUNCTION EXERCISES	260
10.1.2.	CUMULATIVE DISTRIBUTION FUNCTION EXERCISES	262
10.1.3.	QUANTILE FUNCTION (INVERSE CDF) EXERCISES	263
10.1.4.	RANDOM SAMPLING EXERCISES	264
10.2.	DISCRETE DISTRIBUTION IN R	266
10.2.1.	BERNOULLI DISTRIBUTION: <code>rbinom()</code> , <code>dbinom()</code>	266
10.2.2.	BINOMIAL DISTRIBUTION EXAMPLES	271
10.2.3.	POISSON DISTRIBUTION: <code>rpois()</code> and real-world examples	273
10.2.4.	GEOMETRIC DISTRIBUTION: MODELLING THE NUMBER OF TRIALS UNTIL FIRST SUCCESS	276
10.3.	CONTINUOUS DISTRIBUTIONS IN R	278
10.3.1.	UNIFORM DISTRIBUTION: <code>runif()</code>	278
10.3.2.	NORMAL DISTRIBUTION: <code>rnorm()</code> , <code>dnorm()</code> — WITH VISUALIZATION EXAMPLES (<code>hist()</code> , <code>curve()</code>)	280
10.3.3.	EXPONENTIAL DISTRIBUTION: MODELING WAITING TIMES (<code>rexp()</code>)	282
10.3.4.	GAMMA AND BETA DISTRIBUTIONS: USAGE IN BAYESIAN STATISTICS	283
10.4.	FITTING DISTRIBUTION TO DATA	287
10.4.1.	USING <code>fitdistr()</code> from the MASS package	287

10.4.2.	GOODNESS-OF-FIT TESTS (KOLMOGOROV-SMIRNOV TEST WITH <code>ks.test()</code>)	289
10.4.3.	COMPARING EMPIRICAL VS. THEORETICAL DISTRIBUTIONS	290
10.5.	SIMULATION AND RANDOM SAMPLING	293
10.5.1.	GENERATING RANDOM SAMPLES FROM DISTRIBUTIONS	293
10.5.2.	SIMULATING DICE ROLLS AND COIN FLIPS	295
11.	DESCRIPTIVE STATISTICS IN R	297
11.1.	Data Preparation in R	297
11.1.1.	IMPORTING DATA: <code>read.csv()</code> , <code>read.table()</code> , <code>read_excel()</code>	297
11.1.2.	CHECKING STRUCTURE: <code>str()</code> , <code>summary()</code> , <code>head()</code> , <code>tail()</code>	298
11.1.3.	HANDLING MISSING VALUES: <code>na.omit()</code> , <code>is.na()</code> , <code>complete.cases()</code>	299
11.2.	MEASURES OF CENTRAL TENDENCY	300
11.3.	MEASURES OF DISPERSION	305
11.4.	SHAPE OF DISTRIBUTIONS	309
11.5.	FIVE-NUMBER SUMMARY AND BOXPLOTS	312
11.6.	SUMMARIZING MULTIPLE VARIABLES	315
	REFERENCES	319

PREFACE

The textbook *Mathematical Statistics with Examples in R* encompasses the material specified in the syllabus for the Probability and Statistics course for undergraduate studies, as well as portions of the curriculum from the Data-Driven Decision-Making course for graduate studies at the University of Information Science and Technology "St. Apostle Paul" in Ohrid.

This textbook is conceptually divided into two parts. The first part lays out the theoretical bases, addressing standard topics within the field of statistics. The second segment comprises exercises that act as a practical resource for applying statistical methods through R - a robust open-source programming language and software environment specifically designed for statistical analysis, data processing, and graphical representation.

Chapter one, titled *Introduction to Statistics*, introduces the essential principles of statistics as a discipline involving data collection, processing, and quantitative analysis. Chapter two, which focuses on *Descriptive Statistics*, includes aspects of data gathering, organization, numerical characteristics of quantitative data, and methods for graphical representation. Chapter three defines sample spaces, events, probability, conditional probability, and the law of total probability. Chapter four examines random variables, featuring both key discrete and continuous random variables. In chapter five, the Central Limit Theorem is considered as a concept that links the probability theory and mathematical statistics. The statistical estimation methods for population parameters as point estimation and confidence intervals are given in Chapter. In chapter seven, Hypothesis testing, detailing types of errors in hypothesis testing and total error are introduced in Chapter seven. Chapter eight emphasizes regression analysis and the correlation between sample attributes with two-variable values, as well as the chi-square test for assessing the independence of categorical variables.

The second portion of this textbook can be described as a practical reference for executing statistical methods using R and is organized into three chapters: chapters nine, ten, and eleven.

Chapter nine serves as an *Introduction to R*, considering the fundamentals and the R working environment. In chapter ten, titled *Probability Distributions in R*, there are five sections. The first section details basic R functions for distributions, accompanied by resolved examples relating to probability mass functions (PMF), probability density functions (PDF), cumulative distribution functions (CDF), quantile functions (inverse CDF), and random number generation. The second section, *Discrete Distributions in R*, features resolved examples in R concerning the Bernoulli distribution, Binomial distribution, Poisson distribution, and Geometric distribution. The third section, *Continuous Distributions in R*, includes resolved examples in R for the Uniform

distribution, Normal distribution, Exponential distribution, as well as Gamma and Beta distributions. This chapter also contains exercises on fitting distributions to datasets, simulation, and random sampling. Chapter eleven, titled Descriptive Statistics, encompasses data preparation in R, resolved tasks focusing on measures of central tendency, measures of dispersion, distribution shape analysis, five-number summaries, boxplots, and the summarization of multiple variables.

This textbook addresses the fundamental concepts in probability theory and conducts a mathematical examination of statistical data. Simultaneously, it illustrates the use of the R programming language in tackling issues related to probability theory and mathematical statistics.

1. INTRODUCTION TO STATISTICS

1.1. IMPORTANCE OF STATISTICS

In today's world, it's widely accepted that to understand something, you must first gather data. Statistics is the discipline of extracting knowledge from data. It involves collecting, describing, and analyzing data, often leading to insightful conclusions. Studying statistics is excellent preparation for a variety of careers, as it equips you with the tools to respond thoughtfully to the information you encounter in everyday life.

Let's consider some statements as statistical facts:

In a group of just 30 people, there's a 50% chance that two people share the same birthday. With 100 people, the probability jumps to 99%.

Globally, about 107 boys are born for every 100 girls, but men typically live less than women, leading to more women in older age groups.

The richest 1% of the global population possesses about 45% of the world's wealth.

By Pareto Principle in many cases, approximately 80% of outcomes come from 20% of causes. For example, 80% of sales often come from 20% of customers.

Probability of Coin Toss: Flipping a coin is not truly a 50/50 chance; slight biases can exist based on weight distribution or how it is flipped.

The world average life expectancy is around 70 years, but it can vary significantly by country due to factors like healthcare and lifestyle.

More data have been generated in the past five years than in all human history before that. We are not saying that each one of these claims is true!

While not all of these statements may be entirely true, they offer an intriguing glimpse into how statistics can reveal patterns, probabilities, and inequalities in the world around us. These statements underline the power of data to reveal surprising insights, challenge assumptions, and spark interest in the systems and phenomena that shape our lives.

It's important to proceed with statistical statements thoroughly since their exactness depends on the accuracy of the data and the background of analysis including the methods used. Falsified or lousy-sourced statistics can lead to misinterpretations, which is why determining the reliability of data and understanding its limits is essential.

Statistics goes beyond numbers; it's about uncovering the stories they reveal. In business, science, health, and social fields, these facts highlight the crucial role statistics plays in guiding our data-driven world.

What is Statistics?

Statistics is the science of collecting, organizing, analyzing, and interpreting data to make decisions in the presence of uncertainty.

Therefore, Statistics is divided into two areas: Descriptive Statistics and Statistical Inference.

Descriptive statistics is the branch of statistics focused on describing and summarizing data.

Statistical inference uses information from a sample to draw conclusions about the population from which the sample was taken. To make accurate conclusions, it's important to account for the possibility of chance.

The fact that a coin lands tails eight times out of ten flips does not necessarily indicate that the coin is more likely to land on tails than heads in future flips. It could be a perfectly ordinary coin that, by chance, just happened to land on tails eight times out of the total of ten flips. But, if it landed on tails 98 times out of 100 flips, then we would be much more certain that the coin is not fair.

Statistical inference starts with the assumption that key aspects of the phenomenon being studied can be described in terms of probabilities. It then draws conclusions by using data to make inferences about these probabilities. Therefore, understanding statistical inference requires a foundational knowledge of probability theory, which we will cover in the third chapter.

1.2. POPULATION, SAMPLE, RANDOM SAMPLING

A population refers to the complete set of all measurements, observations, or entities of interest to the investigator. It represents the entire group about which we seek to gain knowledge or make inferences. This group could consist of people, animals, plants, objects, events, or anything else that is relevant to the study. Essentially, a population is any entire collection from which data can be gathered. The common characteristic being observed within the population is referred to as a random variable.

For instance, if a researcher is studying the average weight of adults in a country, the population would include all adults in that country and the random variable is their weight. Similarly, in a study about the growth rates of a specific type of vegetable, the population would encompass all such vegetables, regardless of their location and the random variable will be their growth rates.

However, studying the entire population is often impractical due to limitations such as time, cost, and accessibility. To address this, researchers rely on a sample, a subset of the population selected for analysis, or a subset of a population that is chosen to derive conclusions. A sample is designed to be representative of the population so that conclusions drawn from it can be generalized to the larger group with a reasonable degree of confidence.

The process of selecting a representative sample is critical in ensuring that the study's results are valid and unbiased. This involves careful planning and the use of appropriate sampling techniques, such as simple random sampling or stratified sampling, to capture the diversity and characteristics of the population.

Simple random sampling is a method of sampling where, for each selection from the population, every remaining element (sampling unit) has an equal chance of being included in the sample. Additionally, every possible group of the chosen size has the same likelihood of being selected.

Here are two similar, nonmathematical definitions commonly found in general statistics books:

A sample of n elements from a population is a simple random sample if it is true that all possible samples of size n from the population had the same probability of being selected.

A sample of n elements from a population is a simple random sample if it is true that all elements in the population had equal and independent likelihoods of being selected.

These nonmathematical definitions are attempts to interpret the same fundamental mathematical definition that says:

If we consider taking a sample of size n from a population as a statistical experiment with n trials, then the outcome of each trial is an observation of the population random variable X . Because each of the n observations are randomly determined, the n observations correspond to a set of n random variables $X_1, X_2, \dots, X_n$. These n random variables are said to form a simple random sample when two conditions are met:

(1) the n random variables are independent, and

(2) the n random variables all have the same probability distribution as the population random variable X .

If these two conditions are met, then the set of n variables $X_1, X_2, \dots, X_n$, is called a simple random sample, having specific values $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$.

To make exact probability decisions using inferential methods, two assumptions must be satisfied. One assumption of the sampling technique that underlies all inferential methods is the assumption of independence: the selection of an element from a population for one sample does not influence the probability of the selection of any other element. A second assumption is that a simple random sample was taken from the population.

When a simple random sample has been taken, presumably both assumptions have automatically been satisfied. Independence is required for a mathematical definition of a simple random sample and this same condition of independence is stated in the second nonmathematical definition of a simple random sample.

Simple random sampling from an infinitely large population is the ideal sampling situation because it satisfies all the assumptions of the sampling technique. One can assume that: (1) the distribution of the

population remains constant over sampling whether this sampling is with or without replacement, (2) on every selection from this constant population any remaining element of the population has an equal likelihood of being selected, (3) all possible combinations of elements in the population equally likely of being selected, and (4) all elements in the population have an equal and independent probability of being selected.

Simple random sampling from a finite population with replacement is treated as though it were simple random sampling from an infinite population. Since elements are replaced after each selection, the population distribution remains constant, and it can be assumed that all conditions for sampling from an infinite population are met.

Simple random sampling from a finite population without replacement breaks the assumption of independence because the removal of one element changes the probability of all remaining elements. But inferential approaches are resilient to this infraction, meaning if the finite population has a size N that is large compared to the sample size n , then the violation is not a problem. Some statistics books specify that sampling without replacement is not a problem unless the size of the sample n is more than 5% of the size of the population N .

1.3. VARIABLES. TYPES OF VARIABLES

A **variable** is a characteristic that changes or varies over time and/or for different individuals or objects under consideration.

Examples: hair color, white blood cell count, students' height, and time to failure of a computer component.

An **experimental unit** is the individual or object on which a variable is measured, a **measurement** results when a variable is measured on an experimental unit. A set of measurements, called data, can be either a sample or a population.

For example: If the variable is hair color, the experimental unit is a person and typical measurements are brown, black, blonde, etc. If the variable is time until a light bulb burns out, the experimental unit is a light bulb and typical measurements are 1200 hours, 1235.5 hours, etc.

Respectively, to the number of variables that are measured on a single experimental unit we consider **univariate data** when one variable is measured on a single experimental unit, **bivariate data** when two variables are measured on a single experimental unit, and **multivariate data** when more than two variables are measured on a single experimental unit.

Variables are classified into different types based on the kind of data they represent and the way they are measured.

Based on Data Type there are two types of variables:

a) **Quantitative Variables (Numerical)**. These represent numerical values and allow mathematical operations.

They can be divided into:

- Discrete, when the variable takes specific, separate values (often whole numbers). Examples include the number of students in a class, the number of cars in a parking lot, and
- Continuous, when the variable can take any value within a range, including decimals. Examples include the height, weight, temperature, and time.

b) **Qualitative Variables (Categorical)**. These describe categories or labels without a numerical meaning.

We make a difference between

- Nominal qualitative variables are the categories that have no natural order. Examples: gender (male, female), eye color, type of car, and
- Ordinal qualitative variables are the categories with a meaningful order or ranking, but differences between ranks are not measurable. Examples: education level (high school, college, graduate), customer satisfaction (low, medium, high).

Based on the Role in Research there are three types of variables:

a) Independent Variables, which include manipulated or controlled variables in a study to observe their effect on other variables. Examples are temperature of boiling, and time spent studying.

b) Dependent Variables are measured or observed in response to changes in the independent variable. Examples: cooking time, and test scores.

c) Confounding Variables, which are the variables that may influence both the independent and dependent variables, potentially distorting the results. For example, in a study on exercise and weight loss, diet could be a confounding variable.

Based on the Measurement Scale there are four types of variables:

a) Nominal Scale, categorical data without order. Examples include hair color, city names, and marital status.

b) Ordinal Scale, ordered categories, but differences between values are not defined. Examples are movie ratings (1 to 5 stars), and socioeconomic status (low, middle, high).

c) Interval Scale, numerical data with equal intervals between values, but no true zero point. Examples include the temperature in Celsius or Fahrenheit and calendar years.

d) Ratio Scale, numerical data with equal intervals and a true zero point, allowing meaningful comparisons. Examples are height, weight, income, and age.

2. DESCRIPTIVE STATISTICS

2.1. ORGANIZING QUALITATIVE DATA

Let us consider a simple random sample $X_1, X_2, \dots, X_n$, based on observation of the population random variable X having specific qualitative values $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ with frequencies $f_1, f_2, \dots, f_n$, respectively. So, x_1 occurs f_1 times, x_2 - f_2 times, ..., and x_n - f_n times.

The relative frequency of each different value $x_i, i = 1, 2, \dots, n$ is the ratio $\frac{f_i}{n}, i = 1, 2, \dots, n$.

The sequences of ordered pairs $(x_i, f_i), i = 1, 2, \dots, n$, $\left(x_i, \frac{f_i}{n}\right), i = 1, 2, \dots, n$, are called the frequency distribution and relative frequency distribution, respectively of the sample $X_1, X_2, \dots, X_n$ for the variable X with a set of observation values $x_1, x_2, \dots, x_n$.

A dataset with qualitative values can be presented in a summary table for the frequency, relative frequency, cumulative frequency distribution, and cumulative relative frequency distribution.

Data from a frequency table can be graphically represented by a line graph that plots the qualitative values on the horizontal axis and indicates their frequencies by the heights of vertical lines. When the lines in a line graph are replaced with rectangular bars of added thickness, the graph is called a bar graph.

Example 2.1.1. By the Global Cancer Observatory report of the World Health Organization in 2022, the total number of new cancer cases is 4 259, with 873 lung cancer, 790 prostate cancer, 530 colorectum cancer, 250 stomach cancer, 234 bladder cancer, and 1582 other types of cancer. Find the summary table of the data and sketch the relative frequency bar graph.

Solution. The summary table is presented in Table 2.1.1.

Table 2.11.1. Summary table for new cancer cases

i	Cancer	Frequency f_i	Relative frequency	Percent %
1	Lung	873	$\frac{873}{4259} = 0.21$	21
2	Prostate	790	$\frac{790}{4259} = 0.19$	19
3	Colorectum	530	$\frac{530}{4259} = 0.12$	12
4	Stomach	250	$\frac{250}{4259} = 0.06$	6
5	Bladder	234	$\frac{234}{4259} = 0.05$	5
6	Other	1582	$\frac{1582}{4259} = 0.37$	37
Σ		4259	1	

The relative frequency bar graph is given in Figure 2.1.

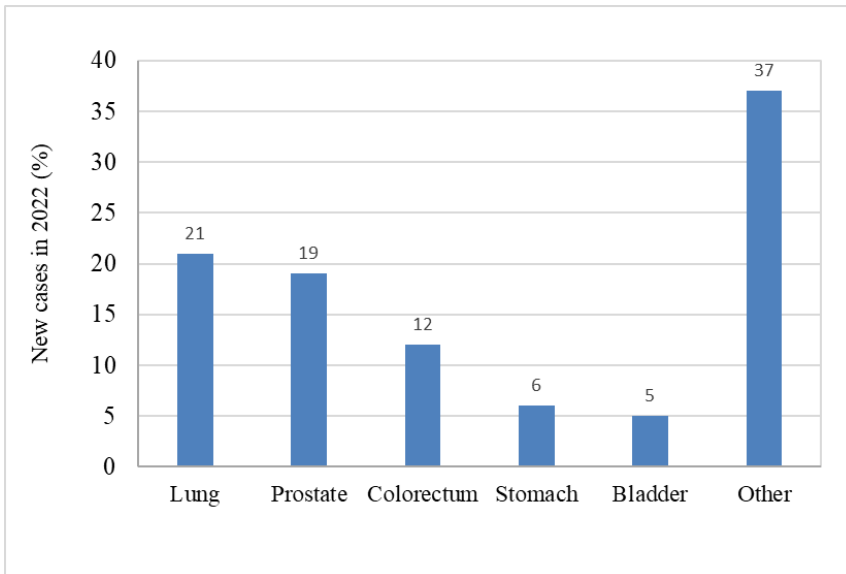


Figure 2.1. The relative frequency bar graph for the cancer new cases

A pie chart is also used to graphically display qualitative data. To construct a pie chart, a circle is divided into portions that represent the relative frequencies or percentages belonging to different values of the qualitative data.

Example 2.2. Graph a pie chart for the dataset in Example 2.1.

Solution. The pie chart is given in Figure 2.2.

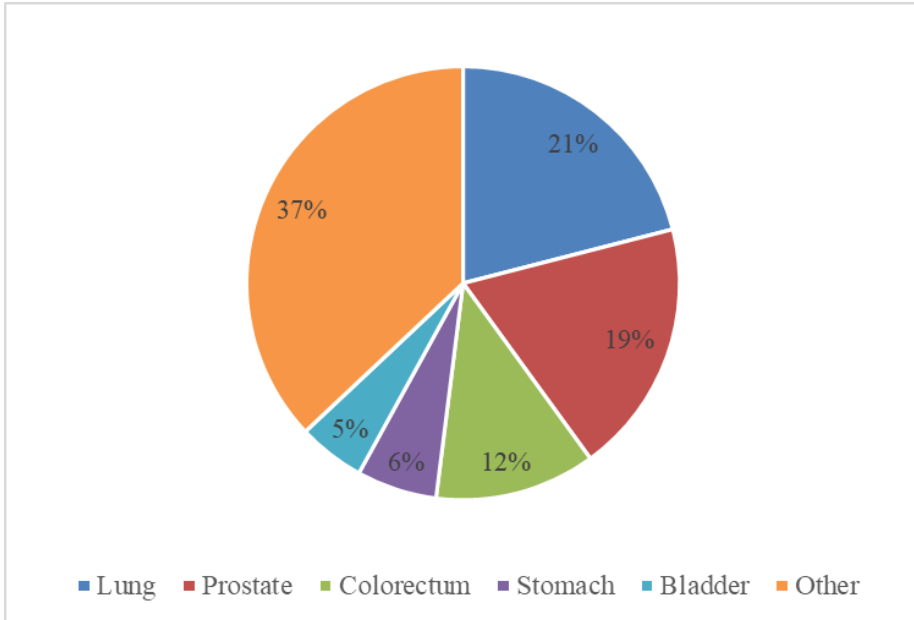


Figure 2.2. The relative frequency pie chart for the cancer new cases

2.2. ORGANIZING QUANTITATIVE DATA

Let us consider a simple random sample $X_1, X_2, \dots, X_n$, based on observation of the population random variable X having specific values $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$, and let X takes k different values, $x_1 \leq x_2 \leq \dots \leq x_k$, increasingly ordered with frequencies $f_1, f_2, \dots, f_k$, respectively. So, x_1 occurs f_1 times, x_2 - f_2 times, ..., and x_k - f_k times.

The relative frequency of each different value $x_i, i = 1, 2, \dots, k$ is the ratio $\frac{f_i}{n}, i = 1, 2, \dots, k$.

The sequences of ordered pairs $(x_i, f_i), i = 1, 2, \dots, k$, $\left(x_i, \frac{f_i}{n}\right), i = 1, 2, \dots, k$, are called the frequency distribution and relative frequency distribution, respectively, of the sample $X_1, X_2, \dots, X_n$ for the variable X with a set of observation values $x_1, x_2, \dots, x_n$.

A dataset having a relatively small number of distinct values can be presented in a summary table for the frequency, relative frequency, cumulative frequency distribution, and cumulative relative frequency distribution. This type of data is called ungrouped data.

Example 2.3. The observed daily temperature was measured at 6:00 am during April, 2024. Twenty days are randomly selected, and the temperature measured in Celsius is recorded as follows:

10, 7, 11, 6, 7, 9, 8, 8, 9, 10, 7, 7, 10, 8, 7, 10, 11, 8, 9, 7.

a) Create a summary table for the frequency and relative frequency, cumulative frequency distribution, and cumulative relative frequency distribution for the given dataset.

b) Construct a frequency line graph and relative frequency bar graph.

Solution. a) First, the given dataset of size 20 contains 6 distinct values arranged in increasing order: 6, 7, 8, 9, 10, 11. The value 6 appears once, the value 7 appears six times, the value 8 appears four times, the value 9 appears three times, the value 10 appears four times, and the value 11 appears twice. Therefore, the requested summary table is shown in Table 2.1.

Table 2.1. Summary table for temperature dataset

i	Temp. x_i	Frequency f_i	Relative frequency
1	6	1	$\frac{1}{20} = 0.05$
2	7	6	$\frac{6}{20} = 0.3$

DESCRIPTIVE STATISTICS

3	8	4	$\frac{4}{20} = 0.2$
4	9	3	$\frac{3}{20} = 0.15$
5	10	4	$\frac{4}{20} = 0.2$
6	11	2	$\frac{2}{20} = 0.1$
Σ		20	1

b) The frequency line graph is given in Figure 2.3.

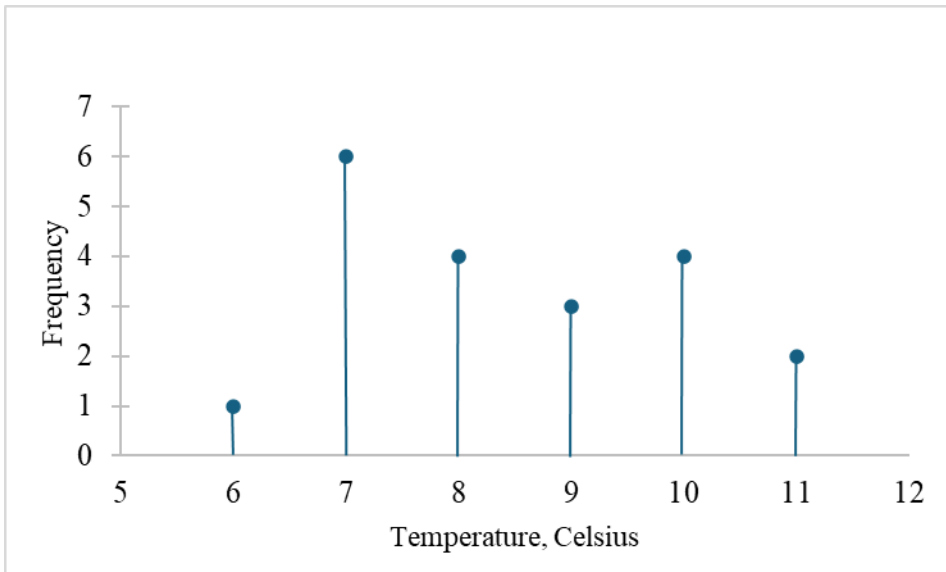


Figure 2.3. Temperature frequency line graph

and the relative frequency bar graph is given in Figure 2.4.

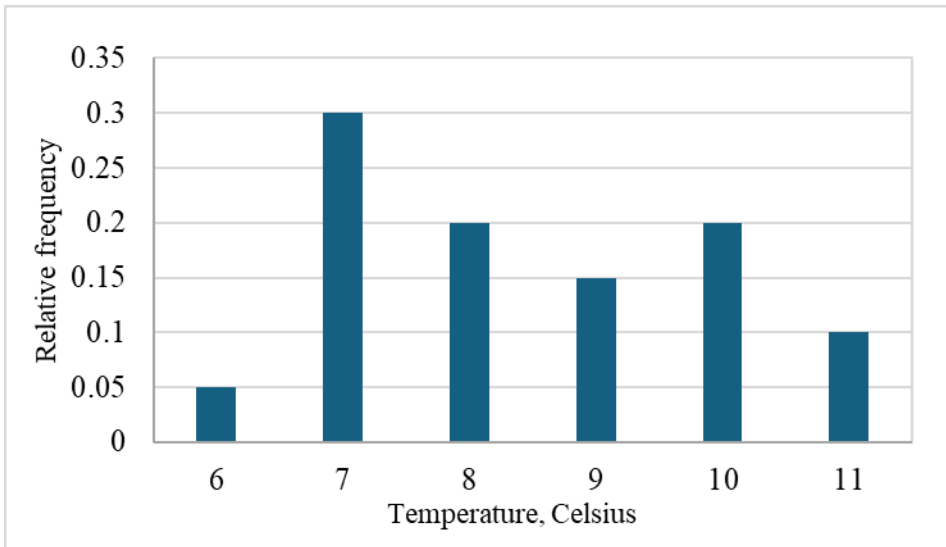


Figure 22.2.4. Temperature relative frequency bar graph

Another type of graph used to represent a frequency or relative frequency distribution is the frequency or relative frequency polygon. This graph plots the frequencies or the relative frequencies of the different data values on the vertical axis and then connects the plotted points with straight lines. Both charts allow us to quickly understand the distribution of values in our dataset.

Example 2.4. Construct frequency and relative frequency polygons for the dataset from Example 2.3.

Solution. Temperature frequency and relative frequency polygons are given in Figure 2.5. and Figure 2.6., respectively.

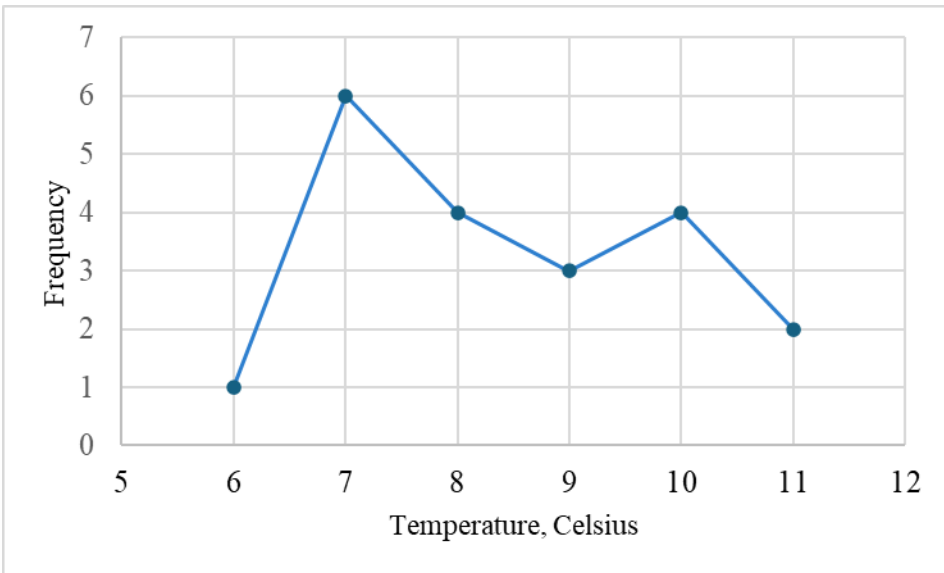


Figure 2.5. Temperature frequency polygon

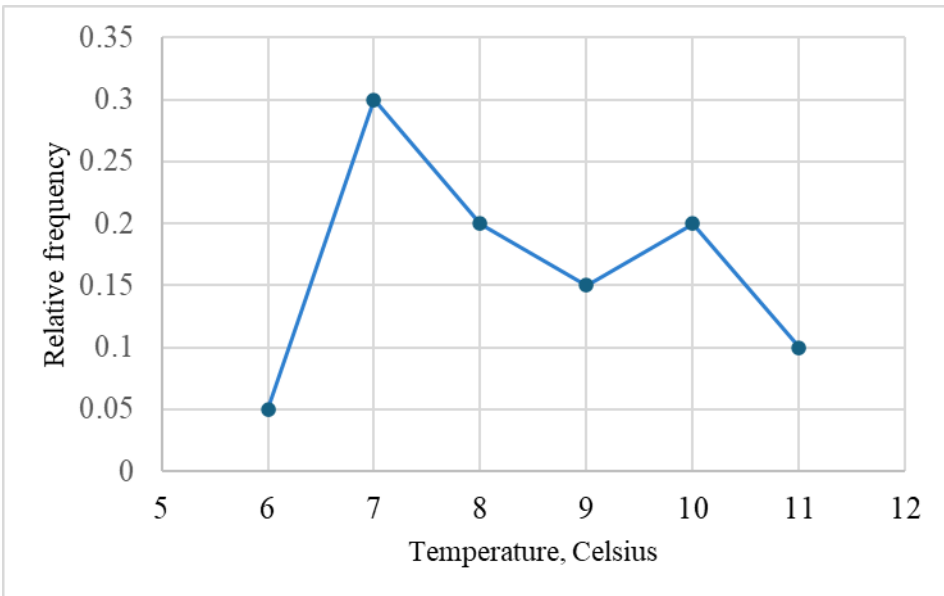


Figure 2.6. Temperature relative frequency polygon

Another type of distribution is the cumulative frequency distribution. This distribution is the total of frequencies, where the frequency of the first-class interval is added to the frequency of the

second-class interval and then the sum is added to the frequency of the third-class interval, and so on. Generally, the cumulative frequency distribution is used to identify the number of values located above or below the special frequency of the dataset.

The cumulative frequency distribution is classified into two different types: less than cumulative frequency and more/greater than cumulative frequency.

The less than cumulative frequency distribution is created by successively adding the frequencies of all the previous classes to the frequency of the current class. In this type of distribution, the cumulative frequency starts from the lowest class and progresses to the highest.

The greater than or more than cumulative frequency is calculated by determining the cumulative total frequencies, starting from the highest class and moving towards the lowest class.

We define less than cumulative frequency distribution, which gives the total number of values that fall below specific values from the sample, i.e.

$F_n(x) = n(x)$, for any $x \in \mathbb{R}$, where $n(x)$ is the number of values from the sample less than the number x , called cumulative frequency.

Therefore,

if $x < x_1$ then $n(x) = 0$, so $F_n(x) = 0$,

if $x_1 \leq x < x_2$ then $n(x) = f_1$, so $F_n(x) = f_1$,

if $x_2 \leq x < x_3$ then $n(x) = f_1 + f_2$, so $F_n(x) = f_1 + f_2$,

.....

If $x_r \leq x < x_{r+1}$ then $n(x) = \sum_{i=1}^r f_i$, so $F_n(x) = \sum_{i=1}^r f_i$

.....

if $x \geq x_k$ then $n(x) = \sum_{i=1}^k n_i = n$, so $F_n(x) = n$

We define less than cumulative relative frequency distribution which is obtained by dividing a cumulative frequency by the total number of observations in the dataset, i.e.

$F(x): \mathbb{R} \rightarrow [0,1]$, defined by

$F(x) = \frac{n(x)}{n}$, for any $x \in \mathbb{R}$, where $n(x)$ is cumulative frequency of $x \in \mathbb{R}$ or

if $x < x_1$ then $n(x) = 0$, so $F(x) = 0$,

if $x_1 \leq x < x_2$ then $n(x) = f_1$, so $F(x) = \frac{f_1}{n}$,

if $x_2 \leq x < x_3$ then $n(x) = f_1 + f_2$, so $F(x) = \frac{f_1 + f_2}{n}$,

.....

If $x_r \leq x < x_{r+1}$ then $n(x) = \sum_{i=1}^r f_i$, so $F(x) = \frac{\sum_{i=1}^r f_i}{n}$

.....

if $x \geq x_k$ then $n(x) = \sum_{i=1}^k n_i = n$, so $F(x) = 1$.

Graphical representations of less than cumulative frequency distribution and less than cumulative relative frequency distribution are plotted using a less than ogive.

The less than cumulative frequency ogive is sketched as follows:

First, we mark the upper limit on the horizontal axis, we mark the cumulative frequency on the vertical axis, and then we plot the points (x, y) in the coordinate plane where x represents the upper limit value and y represents the cumulative frequency. Finally, we join the points and draw a line or a smooth curve, called less than cumulative frequency ogive.

Example 2.5. Make a summary table for the less than cumulative frequency distribution, and less than cumulative relative frequency

distribution for the dataset from Example 2.3. Then construct their ogives.

Solution. The requested summary table is presented in Table 2.2.

Table 2.2. Summary table for temperature dataset

i	Class interval	Less than	$F_n(x)$	$F(x)$
1	$x < 6$	5.5	0	0
2	$6 \leq x < 7$	6.5	1	0.05
3	$7 \leq x < 8$	7.5	7	0.35
4	$8 \leq x < 9$	8.5	11	0.55
5	$9 \leq x < 10$	9.5	14	0.7
6	$10 \leq x < 11$	10.5	18	0.9
Σ	$x \geq 11$	11.5	20	1

Less than cumulative frequency ogive and cumulative relative frequency ogive are given in Figure 2.7. and Figure 2.8., respectively.

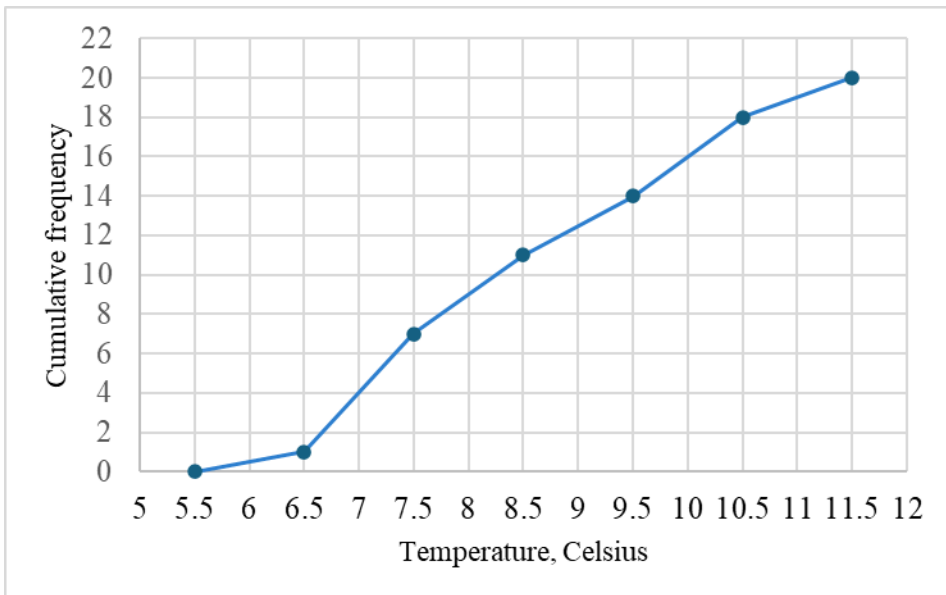


Figure 2.7. Less than cumulative frequency ogive

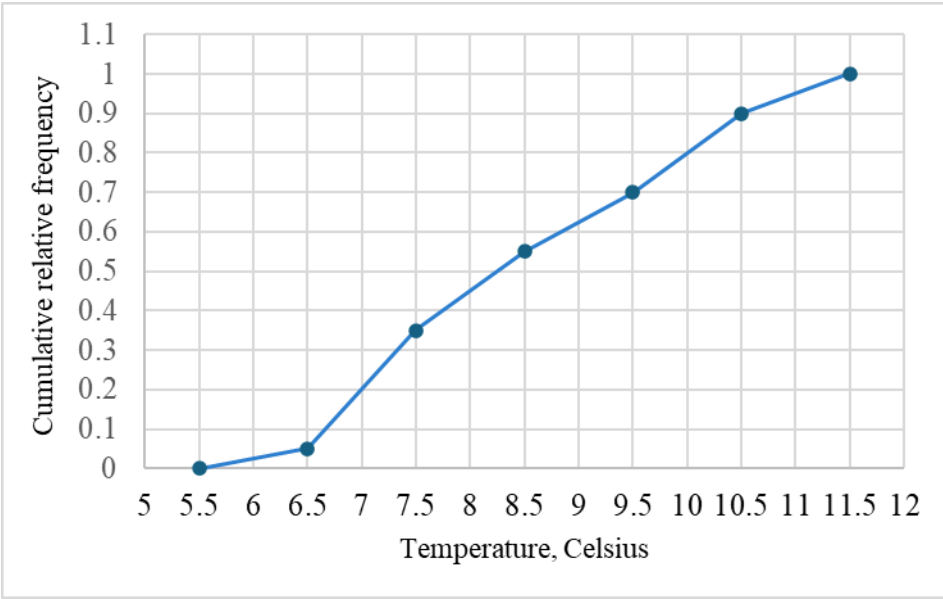


Figure 2.8. Less than cumulative relative frequency ogive

When the sample size is large, organizing the data as above becomes monotonous and time-consuming. In this case, the raw data are grouped in different class intervals, and it is said to be grouped data.

Let us consider the sample given with the dataset of n observation values $x_1, x_2, \dots, x_n$ arranged in increasing order. We define values $a < x_1$ and $b > x_n$, taken as the nearest values to the values x_1 and x_n . Then the interval $[a, b]$ is split in n class-intervals $[a, a_1)$, $[a_1, a_2)$, ..., $[a_{n-2}, a_{n-1})$, $[a_{n-1}, b]$. Usually, the class intervals are chosen to be of equal length. Let's assume there are k different class-intervals:

$$[a, a_1), [a_1, a_2), \dots, [a_{k-2}, a_{k-1}), [a_{k-1}, b], \text{ where} \\ a < a_1 < a_2 < \dots < a_{k-1} < b,$$

with frequencies $f_1, f_2, \dots, f_k$, respectively. So, in the first class intervals are contained f_1 values from the dataset, in the second class interval, are f_2 values from the dataset, ..., and in the k -th class interval are f_k values from the dataset.

The middle points $x_i = \frac{a_{i-1} + a_i}{2}, i = 1, \dots, k$ of each class interval are called classmarks.

The sequences of ordered pairs $(x_i, f_i), i = 1, 2, \dots, k$, and $\left(x_i, \frac{f_i}{n}\right), i = 1, 2, \dots, k$, are called the grouped frequency distribution and grouped relative frequency distribution, respectively of the sample $X_1, X_2, \dots, X_n$ for the variable X with a set of observation values $x_1, x_2, \dots, x_n$. But in most of the literature the word “grouped” is often omitted. Here, we will use the term "grouped" to highlight the distinction between the two types of dataset organization: ungrouped and grouped.

The summary table of the grouped dataset is given in Table 2.3.

Table 2.3. Summary table for grouped data

i	Class interval	Classmark x_i	Frequency f_i	Relative frequency
1	$[a, a_1)$	x_1	f_1	$\frac{f_1}{n}$
2	$[a_1, a_2)$	x_2	f_2	$\frac{f_2}{n}$
...	...	...	...	...
k	$[a_{k-1}, b]$	x_k	f_k	$\frac{f_k}{n}$
Σ			n	1

The graphical representation of a grouped dataset is called a **histogram**. The histogram represents a frequency distribution using rectangles where the widths correspond to class intervals, and the areas are proportional to the corresponding frequencies. The height of each rectangle represents the average frequency density for the interval. The intervals are placed next to each other to show that the data represented by the histogram, while exclusive, is also continuous and adjacent. For example, in a histogram, it is possible to have two connected intervals of 5.5-9.5 and 9.5-13.5, but not two connected intervals of 5.5-9.5 and 13.5-

17.5. The class intervals with no values from the dataset are represented as empty and not skipped.

What would be the best number of class intervals? There is no “best” number of class intervals, and different class interval widths can reveal different data properties.

By using wider class intervals where the density of the underlying data points is low, the noise due to sampling randomness is reduced. Conversely, using narrower class intervals where the density is high provides more accuracy in density estimation..

Thus, varying the class interval width within a histogram can be beneficial. Nonetheless, equal-width width class intervals are widely used.

Some mathematicians have attempted to determine an optimal number of class intervals, but these methods make strong assumptions about the shape of the distribution. Depending on the actual data distribution and the goals of the statistical analysis, different class interval widths may be more suitable. Therefore, experimentation is often necessary to determine the optimal class interval width.

The number of class interval k can be assigned directly or can be calculated from a suggested class interval width w_c as:

$$k = \left\lceil \frac{b-a}{w_c} \right\rceil.$$

The brackets indicate the ceiling function, which takes as input a real number x , and gives as output the least integer greater than or equal to x . So, for an example $\lceil 2.3 \rceil = 3$.

In many software statistical packages, a square-root choice is implemented so-called Square-root choice, which takes the square root of the sample size and rounds to the next integer, i.e.,

$$k = \left\lceil \sqrt{n} \right\rceil.$$

Example 2.6. 50 visitors were randomly selected at a shopping center. The time they spend in the store, expressed in minutes, is:

12,10,15,23,11,57,34,23,45,55,27,35,13,10,45,44,54,

31,33,34,37,16,43,42,11,16,18,17,22,24,45,48,52,56,
51,50,22,14,40,35,45,43,33,33,18,19,32,36,48,12.

After groping the data into 6 class intervals with equal width:

a) Make a summary table for the grouped frequency, grouped relative frequency distributions, and “less than” grouped cumulative frequency distribution for the given data.

b) Construct a frequency histogram and a “less than” grouped cumulative relative frequency ogive for the given data.

Solution. a) Since the largest value of the sample is 57 and the smallest value is 10, the upper-class boundary is 57.5 and the lower-class boundary is 9.5. The width of each class interval is $w_c = \frac{57.5 - 9.5}{6} = 8$.

The requested summary table is presented in Table 1.4.4.

Table 2.4. Summary table for time spent in the store

Class interval	Classmark x_i	f_i	$\frac{f_i}{n}$	less than	$F(x)$
9.5 – 17.5	13.5	12	0.24	9.5	0
17.5 – 25.5	21.5	8	0.16	17.5	0.24
25.5 – 33.5	29.5	6	0.12	25.5	0.4
33.5 – 41.5	37.5	7	0.14	33.5	0.52
41.5 – 49.5	45.5	10	0.2	41.5	0.66
49.5 – 57.5	53.5	7	0.14	49.5	0.86
Σ		50	1	57.5	1

b) Frequency histogram is given in Figure 2.9.

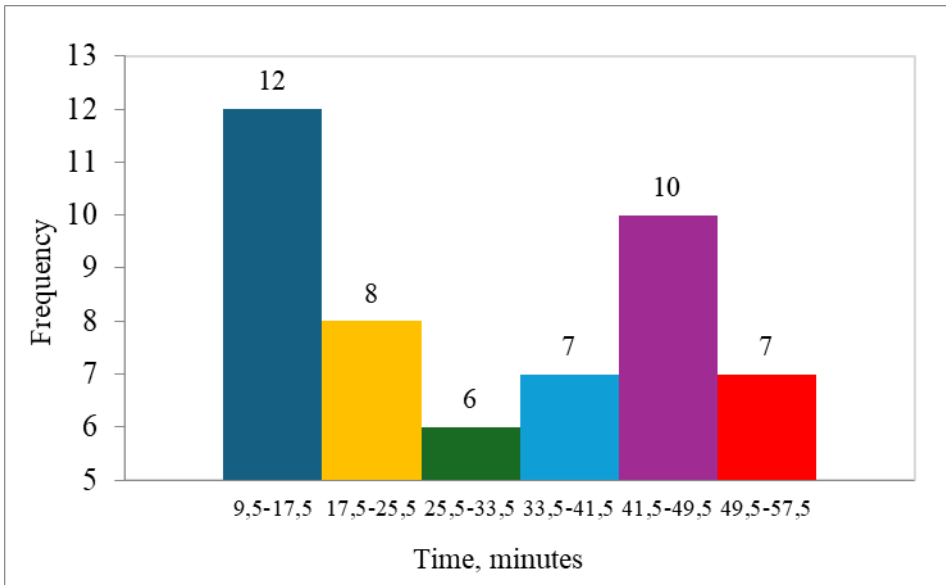


Figure 2.9. Frequency histogram for time spent in the store

and the less than cumulative relative frequency ogive is given in Figure 2.10.

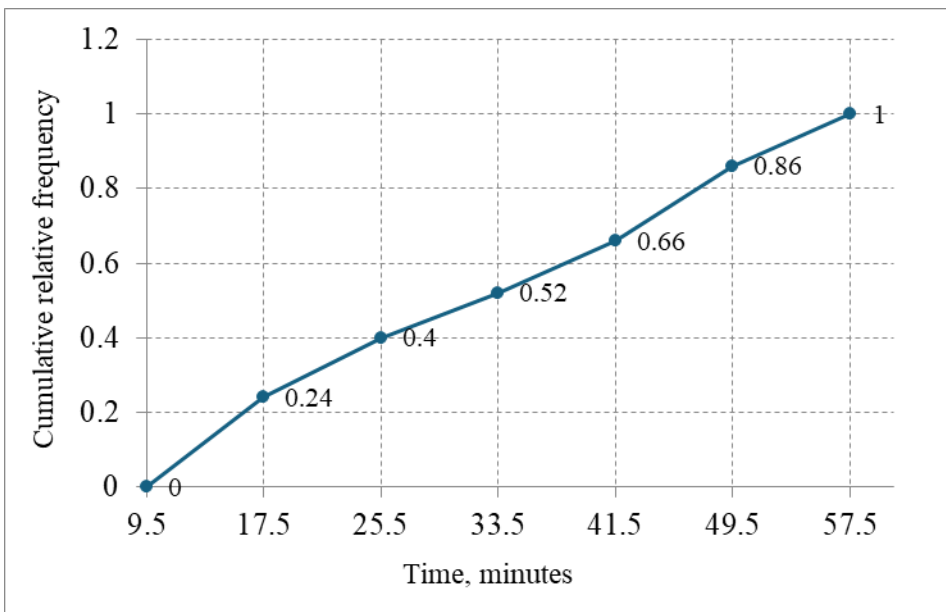


Figure 2.10. Cumulative frequency ogive for time spent in the store

Example 2.7. In one high school, the results of the final exam in mathematics are analyzed. It is considered a sample of the results of 50 students (the exam is scored from 0 to 20 points):

12, 15, 10, 7, 13, 16, 17, 16, 7, 11, 15, 13, 11, 5, 13, 12, 9, 11, 9, 12, 20, 20, 10, 13, 6, 12, 15, 10, 9, 13, 16, 17, 6, 9, 10, 15, 13, 6, 5, 13, 12, 9, 20, 7, 12, 15, 15, 20, 13, 15.

After grouping the data into 8 class intervals with equal width:

a) Make a summary table for the grouped frequency, grouped relative frequency, and the less than grouped cumulative frequency distribution for the given data.

b) Construct a frequency histogram and a “less than” grouped cumulative relative frequency ogive for the given data.

Solution. a) Since the largest value of the sample is 20 and the smallest value is 5, the upper-class boundary is 20.5 and the lower-class boundary is 4.5. The width of each class interval is $w_c = \frac{20.5 - 4.5}{8} = 2$.

The requested summary table is presented in Table 2.5.

Table 2.5. Summary table for final exam results

Class interval	Classmark x_i	f_i	$\frac{f_i}{n}$	less than	$F(x)$
4.5 – 6.5	5.5	5	0.1	4.5	0
6.5 – 8.5	7.5	3	0.06	6.5	0.1
8.5 – 10.5	9.5	9	0.18	8.5	0.16
10.5 – 12.5	11.5	9	0.18	10.5	0.34
12.5 – 14.5	13.5	8	0.16	12.5	0.52
14.5 – 16.5	15.5	10	0.2	14.5	0.68
16.5-18.5	17.5	2	0.04	16.5	0.88
18.5-20.5	19.5	4	0.08	18.5	0.92
Σ		50	1	20.5	1

b) Frequency histogram

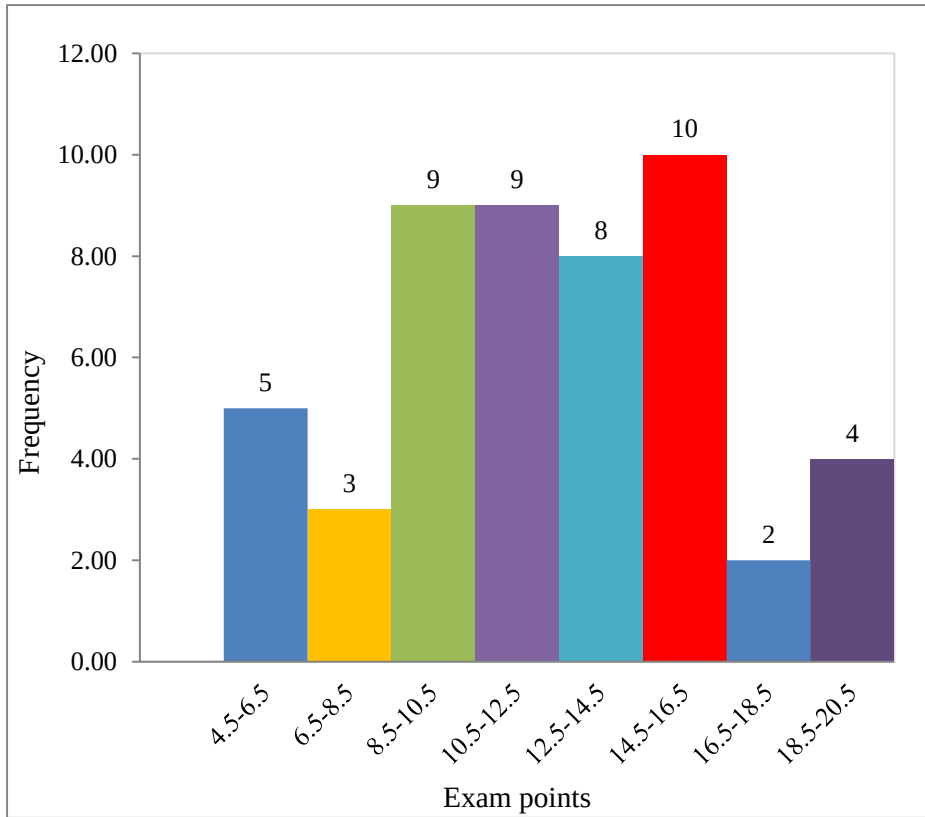


Figure 2.11. Frequency histogram of final exam points

and less than cumulative relative frequency ogive

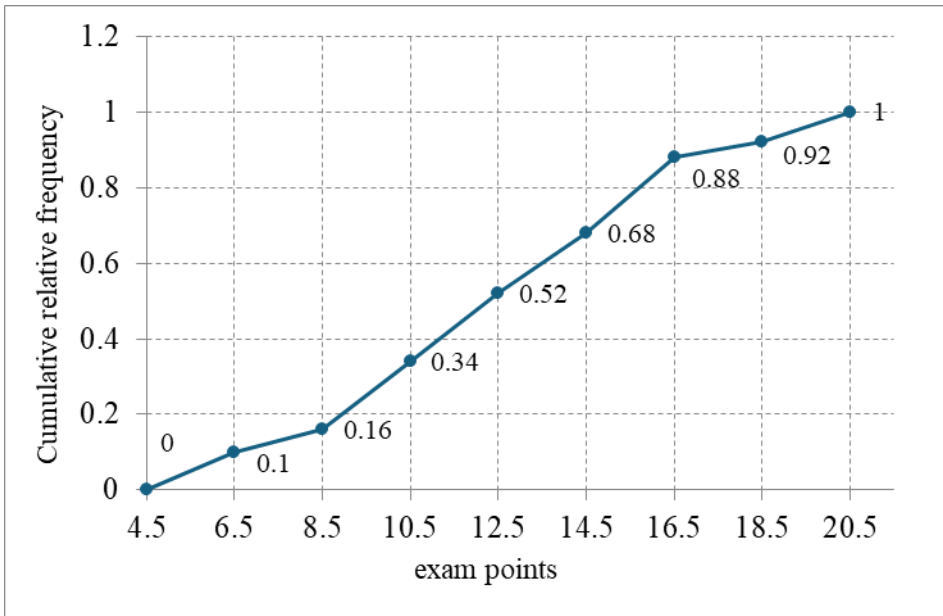


Figure 2.12. Cumulative relative frequency ogive of final exam points

2.3. NUMERICAL DESCRIPTIVE STATISTICS

Although distributions provide useful descriptions of data, they still contain too much detail for some purposes. Assume, for example, that we have collected data on tree dimensions from several forests to detect possible differences in the distribution of tree sizes among these forests. While histograms of the distributions would provide some indication of such differences, they would not offer specific measures of these differences that could be used for quantitative comparisons

Numerical measures that provide descriptions of the characteristics of the distributions, which can then be used to provide more readily interpretable information on such differences, are needed. The two most important aspects are the **location** and the **dispersion** of the data. In other words, we need to find a number that indicates where the observations are on the measurement scale and another to indicate how widely the observations vary.

2.3.1. MEASURES OF LOCATION

The most useful single characteristic of a distribution is some typical, average, or representative value that describes the set of values. Such a value is referred to as a descriptor of **location** or **central tendency**. The most frequently used measure of location is the **arithmetic mean**, usually referred to simply as the mean.

Definition 2.3.1. The **sample mean** is the sum of all the observed values divided by the number of values.

Let $(x_i, f_i), i = 1, 2, \dots, k$ is a frequency distribution of the sample $X_1, X_2, \dots, X_n$ for the variable X with a set of observation values $x_1, x_2, \dots, x_n$, then the **arithmetic mean** of the sample, or **sample mean** is obtained by the formula:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i \cdot x_i.$$

Example 2.3.1. Using the previous formula, calculate the mean of the sample in Table 2.3.1:

Table 2.3.1. Dataset of the products' length

Length (cm), x_i	f_i	$f_i \cdot x_i$	$f_i \cdot x_i^2$
1.2	2	2.4	2.88
1.3	7	9.1	11.83
1.4	10	14.0	19.6
1.5	12	18.0	27
1.6	10	16.0	25.6
1.7	7	11.9	20.23
1.8	2	3.6	6.48

Σ	50	75.0 cm	113.62
$\bar{x} = \frac{1}{n} \sum f_i \cdot x_i = \frac{75.0 \text{ cm}}{50} = 1.50 \text{ cm}$			

For determining the **approximate sample mean** from a grouped frequency distribution, first we obtain the class marks for the class intervals and then we use the previous formula for the sample mean.

Example 2.3.2. Calculate the mean of the sample with grouped frequency distribution in Table 2.3.2.:

Table 1.5.2. Grouped dataset

Class interval	Classmark x_i	f_i	$f_i x_i$	$f_i x_i^2$
35.5 – 39.5	37.5	11	412.5	15468.75
39.5 – 43.5	41.5	5	207.5	8611.25
43.5 – 47.5	45.5	7	318.5	14491.75
47.5 – 51.5	49.5	8	396	19602
51.5 – 55.5	53.5	7	374.5	20035.75
55.5 – 59.5	57.5	2	115	6612.5
Σ		40	1824	84822
$\bar{x} = \frac{1}{n} \sum f_i \cdot x_i = \frac{1824}{40} = 45.6$				

Another useful measure of location is the median.

Definition 2.3.2. The **median** of a set of observed values is defined as the middle value when the measurements are arranged in ascending order. In other words, 50% of the measurements lie above the median, and 50% fall below it.

The precise definition of the median depends on whether the number of observations is odd or even as follows:

If there is an odd number of values in an arrayed dataset with an ungrouped frequency distribution, then the median is the middle value of the array; if there is an even number of values in an array, then the median is the arithmetic mean of the two middle values.

Example 2.3.3. Since in Table 2.3.1. are 7 values of observation, the median of the sample is the fourth value, i.e., $Median = 1,5$.

For determining the **approximate median** from a grouped frequency distribution first, we determine the median class (if there is an odd number of classes, then the median class is the middle class; if there is an even number of classes, then the median class is one of two classes in the middle by which the “less than” cumulative relative frequency is closer to the 0.5) and then we use following median-locating formula for grouped frequency distributions (which can be used only when the grouped distribution has equal class widths):

$$Median \approx b_c + \frac{\frac{n}{2} - \sum f_1}{f_{me}} \cdot w_c$$

where b_c is the lower boundary of the median class, n is the size of the sample, $\sum f_1$ is the cumulative frequency from classes that precede the median class, f_{me} is the frequency of the median and w_c is the class width.

Example 2.3.4. Since there is an even number of classes in Table 1.5.2. then the median class is one of two classes in the middle, meaning **43.5 – 47.5** with “less than” cumulative relative frequency 0.4 or **47.5 – 51.5** with “less than” cumulative relative frequency 0.57; and because the “less than” cumulative relative frequency 0.575 is more closer to the 0.5 than “less than” cumulative relative frequency 0.4; the median class of the sample in Table 1.5.2. is the class interval **47.5 -51.57**. Therefore $b_c = 47.5$, $n = 40$, $\sum f_1 = 11 + 5 + 7 = 23$, $f_{me} = 8$ and $w_c = 4$. So by the previous formula the approximate median of the sample is:

$$Median \approx 47.5 + \frac{\frac{40}{2} - 23}{8} \cdot 4 = 46.$$

Now that we have two measures of location, it is logical to ask which one is better and which one should we use?

Note that the mean is calculated using the value of each observation, so all the information available from the data is utilized. This is not the case for the median. For the median we only need to know where the “middle” of the data is. Therefore, the mean is the more useful measure and, in most cases, the mean will give a better measure of the location of the data. However, for symmetric or nearly symmetric distributions, the mean and median will be the same or nearly the same, while for skewed distributions the value of the mean will tend to be “pulled” toward the long tail. This phenomenon can be explained by the fact that the mean can be interpreted as the center of gravity of the distribution. That is, if the observations are viewed as weights placed on a plane, then the mean is the position at which the weights on each side balance. It is a well-known fact of physics that weights placed further from the center of gravity exert a larger degree of influence (also called leverage); hence the mean must shift toward those weights to achieve balance. However, the median assigns equal weights to all observations regardless of their actual values; hence the extreme values have no special leverage.

Therefore, **in general, we will use the mean as the primary measure of location unless the distribution of the variable is skewed.**

Definition 2.3.3. The **mode** is the most frequently occurring value.

This measure may not be unique in that two (or more) values may occur with the same greatest frequency. Also, the mode may not be defined if all values occur only once, which usually happens with continuous numeric variables.

In an arrayed dataset with an ungrouped frequency distribution, if two consecutive values have the same frequency, which is greater than the frequency of any other value, the mode is considered to be the arithmetic mean of these consecutive values. If there are two nonconsecutive values in the dataset with the same highest frequency, then both values are referred to as modes.

Example 2.3.5. Since in Table 2.3.1. there are two nonconsecutive values that have the same and the greatest frequency (10), then both values are modes. Therefore $Mode=1.4$ and $Mode=1.6$.

For determining the **approximate mode** from a grouped frequency distribution, first we determine the modal class (the class with the highest frequency), and then we use the following mode-locating formula for grouped frequency distributions (which can be used only when the grouped distribution has equal class widths):

$$Mode \approx b_c + \frac{f_2 - f_1}{(f_2 - f_1) + (f_2 - f_3)} \cdot w_c$$

where b_c is the lower boundary of the modal class, f_1, f_2, f_3 are the frequency of the class that precedes the modal class, the modal class, and the class that follows the modal class in the distribution, respectively, and w_c is the class width.

Example 2.3.6. The modal class for the sample is in Table 2.3.2. is class **35.5 – 39.5** since it has the highest frequency. Therefore $b_c = 35.5$, $f_1 = 0$, $f_2 = 11$, $f_3 = 5$, and $w_c = 4$. So by the previous formula, the approximate mode of the sample is:

$$Mode \approx 35.5 + \frac{11 - 0}{(11 - 0) + (11 - 5)} \cdot 4 = 38.09.$$

2.3.2. MEASURES OF DISPERSION

Summarizing a dataset may help us understand the data, especially large ones. Although location is generally considered the most important single characteristic of a distribution, the spread measures summarise the data to show how scattered the values are and how far they are from the mean.

Definition 2.3.4: The **range** is the difference between the smallest value and the largest value in a dataset.

Example 2.3.7. Find the range of the dataset: 5, 4, 5, 5, 6, 6, 6, 6, 7, 8, 7, 8, 3, 4.

Solution. After ordering the dataset into increasing array: 3, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, the range is 5, the difference between the highest value 8 and the lowest value 3, i.e., $Range = 8 - 3 = 5$.

Quartiles divide an ordered dataset into four equal parts and refer to the values of the point between the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).

Definition 2.3.5: The **lower quartile** (Q_1) is the value which is larger than 25% and less than 75% of the ordered n measurements $x_1, x_2, \dots, x_n$, and the **position of** the lower quartile Q_1 is $0.25(n+1)$ and if the position is not integer, i.e., $0.25(n+1) = i.dd$ then

$$Q_1 = x_i + 0.25(x_{i+1} - x_i).$$

Definition 2.3.6: The **upper quartile** (Q_3) is the value which is larger than 75% and less than 25% of the ordered n measurements $x_1, x_2, \dots, x_n$, and the **position of** the upper quartile Q_1 is $0.25(n+1)$ and if the position is not integer, i.e., $0.75(n+1) = i.dd$ then

$$Q_3 = x_i + 0.75(x_{i+1} - x_i).$$

Note: The lower quartile (Q_1) is also called the 25th percentile, the second quartile (Q_2) is the middle of the data set, i.e., the median and is also called the 50th percentile, and the upper quartile (Q_3) is also called the 75th percentile.

Definition 2.3.7: The range of the “middle 50%” of the ordered measurements is the **interquartile range**, i.e.,

$$IQR = Q_3 - Q_1.$$

Example 2.3.8. Identify the first and third quartiles for the following dataset: 11, 14, 9, 3, 27, 26, 7, 8, 19, 10, 12, 6. Then calculate the interquartile range.

Solution. First, we arrange the values in increasing order: 3, 6, 7, 8, 9, 10, 11, 12, 14, 19, 26, 27.

The size of the dataset is $n=12$, so the position of the lower quartile Q_1 is $0.25(n+1)=0.25 \cdot 13=3.25$ and since the position is not integer, and $x_3=7, x_4=8$ then the lower quartile is

$$Q_1 = 7 + 0.25(8 - 7) = 7.25.$$

The position of the upper quartile Q_3 is

$$0.75(n+1) = 0.75 \cdot 13 = 9.75,$$

and since the position is not an integer, and $x_9=14, x_{10}=19$ then the upper quartile is

$$Q_3 = 14 + 0.75(19 - 14) = 17.75.$$

The interquartile range is $IQR = Q_3 - Q_1 = 17.75 - 7.25 = 10.5$.

Definition 2.3.8: The **approximate lower quartile** for the grouped dataset is obtained by the following formula

$$Q_1 \approx b_c + \frac{\frac{n}{4} - \sum f_1}{f_{lq}} \cdot w_c$$

where b_c is the lower boundary of the lower quartile class, i.e., the class interval that contains the lower quartile value, n is the size of the sample, $\sum f_1$ is the cumulative frequency from classes that precede the lower quartile class, f_{lq} is the frequency of the lower quartile class and w_c is the class width.

Definition 2.3.9: The **approximate upper quartile** for the grouped dataset is obtained by the following formula

$$Q_3 \approx b_c + \frac{\frac{3n}{4} - \sum f_1}{f_{uq}} \cdot w_c$$

where b_c is the lower boundary of the upper quartile class, i.e., the class interval that contains the upper quartile value, n is the size of the sample, $\sum f_1$ is the cumulative frequency from classes that precede

the upper quartile class, f_{uq} is the frequency of the upper quartile class and w_c is the class width.

Definition 2.3.10. The **sample variance**, denoted by s^2 , of a set of n observed values having a mean $\bar{x}$ is the sum of the squared deviations divided by $n - 1$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{or} \quad s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right).$$

Let $(x_i, f_i), i = 1, 2, \dots, k$ is a frequency distribution of the sample $X_1, X_2, \dots, X_n$ for the variable X with a set of observation values $x_1, x_2, \dots, x_n$, then the

sample variance of the sample is obtained by the formula:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (x_i - \bar{x})^2 \quad \text{or} \quad s^2 = \frac{1}{n-1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n\bar{x}^2 \right).$$

Definition 2.3.1. The **sample standard deviation** s , of a set of observed values is defined to be the positive square root of the variance.

Example 2.3.9. The sample variance for the data in Table 2.3.2 is:

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n\bar{x}^2 \right) = \frac{1}{49} (113.62 - 50 \cdot 1.5^2) = 0.023$$

and standard deviation $s = \sqrt{0.023} = 0.15$.

For determining the **approximate variance** and **approximate standard deviation** from a grouped frequency distribution, first we obtain the classmarks and then we use the previous formula for the sample variance and sample standard deviation.

Example 2.3.10. The sample variance for the data in Table 2.3.2. is:

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n\bar{x}^2 \right) = \frac{1}{39} (84822 - 40 \cdot 45.6^2) = 42.25$$

and standard deviation $s = \sqrt{42.25} = 6.5$.

Example 2.3.11. The observed daily temperature was measured at 6:00 am during April 2024. For randomly selected 20 days, the temperature measured in Celsius is:

10, 7, 11, 6, 7, 9, 8, 8, 9, 10, 7, 7, 10, 8, 7, 10, 11, 8, 9, 7.

Find the sample mean, mode, median, lower and upper quartiles, interquartile range, sample variance, and sample standard deviation for the given dataset.

Solution. First, we determine the summary Table 2.3.3. for the given dataset:

Table 2.3.3. Summary table for temperature dataset

i	x_i	f_i	$\frac{f_i}{n}$	$f_i \cdot x_i$	$f_i \cdot x_i^2$
1	6	1	$\frac{1}{20} = 0.05$	6	36
2	7	6	$\frac{6}{20} = 0.3$	42	294
3	8	4	$\frac{4}{20} = 0.2$	32	256
4	9	3	$\frac{3}{20} = 0.15$	27	243
5	10	4	$\frac{4}{20} = 0.2$	40	400
6	11	2	$\frac{2}{20} = 0.1$	22	242
Σ		20		169	1471

The sample mean is: $\bar{x} = \frac{1}{n} \sum f_i \cdot x_i = \frac{169}{20} = 8.45$.

Since the most frequently occurring value in Table 2.3.3. is 7, with frequency 6, the mode is $Mode = 7$.

Since in Table 2.3.3. are 20 values of observation, the median of the sample is the arithmetic mean of the two middle values, i.e., the 10-th and 11-th value, or $Median = \frac{8+8}{2} = 8$.

The size of the dataset is $n = 20$, so the position of the lower quartile Q_1 is $0.25(n+1) = 0.25 \cdot 21 = 5.25$ and since the position is not an integer, and $x_5 = 7, x_6 = 7$ then the lower quartile is

$$Q_1 = 7 + 0.25(7 - 7) = 7.$$

The position of the upper quartile Q_3 is

$$0.75(n+1) = 0.75 \cdot 21 = 15.75,$$

and since the position is not an integer, and $x_{15} = 10, x_{16} = 10$ then the upper quartile is

$$Q_3 = 10 + 0.75(10 - 10) = 10.$$

The interquartile range is $IQR = Q_3 - Q_1 = 10 - 7 = 3$.

The sample variance for the data in Table 2.3.3. is:

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n\bar{x}^2 \right) = \frac{1}{19} (1471 - 20 \cdot 8.45^2) = 2.26$$

and standard deviation $s = \sqrt{2.26} = 1.503$.

Example 2.3.12. 50 visitors were randomly selected at a shopping center. The time they spend in the store, expressed in minutes, is:

12,10,15,23,11,57,34,23,45,55,27,35,13,10,45,44,54,

31,33,34,37,16,43,42,11,16,18,17,22,24,45,48,52,56,

51,50,22,14,40,35,45,43,33,33,18,19,32,36,48,12.

After grouping the data into 6 class intervals with equal width, find the approximate arithmetic mean, approximate mode, approximate median, approximate lower and upper quartiles, approximate variance and approximate standard deviation for the sample.

Solution. The largest value of the dataset is 57 and the smallest value is 10, so the upper-class boundary is 57.5 and the lower-class boundary is 9.5. The width of each class interval is $w_c = \frac{57.5 - 9.5}{6} = 8$.

The requested summary table is presented in Table 2.3.4.

Table 2.3.4. Summary table for the store grouped data

Class interval	Classmark x_i	f_i	$\frac{f_i}{n}$	less than	$F(x)$	$f_i \cdot x_i$	$f_i \cdot x_i^2$
9.5 – 17.5	13.5	12	0.24	9.5	0	162	2187
17.5 – 25.5	21.5	8	0.16	17.5	0.24	172	3698
25.5 – 33.5	29.5	6	0.12	25.5	0.4	177	5221.5
33.5 – 41.5	37.5	7	0.14	33.5	0.52	262.5	9843.75
41.5 – 49.5	45.5	10	0.2	41.5	0.66	455	20702.5
49.5 – 57.5	53.5	7	0.14	49.5	0.86	374.5	20035.75
Σ		50	1	57.5	1	1603	61688.5

The sample mean is: $\bar{x} = \frac{1}{n} \sum f_i \cdot x_i = \frac{1603}{50} = 32.06$.

Since the class with the highest frequency 12, is 9.5-17.5, then $b_c = 9.5$, $f_1 = 0$, $f_2 = 12$, $f_3 = 8$.

Therefore, the approximate mode is

$$Mode = 9.5 + \frac{12 - 0}{(12 - 0) + (12 - 8)} \cdot 8 = 15.5$$

There are two classes in the middle: 25.5-33.5 and 33.5-41.5, but within the class interval 25.5-33.5 contains 26 values, nearly half of the size of the dataset. Therefore, the median class is the class interval 25.5-33.5, and $b_c = 25.5$, $\sum f_1 = 12 + 8 = 20$, $f_{me} = 6$, so

$$Median = 25.5 + \frac{\frac{50}{2} - 20}{6} \cdot 8 = 32.17.$$

The size of the dataset is $n=50$, so the position of the lower quartile Q_1 is $0.25(n+1)=0.25 \cdot 51=12.75$ but within the class interval 17.5-25.5 are contained 12 values, nearly lower quartile position, so the lower quartile class interval is 17.5 – 25.5. Therefore, $b_c=17.5$, $\sum f_1=12$, $f_{lq}=8$ and $w_c=8$, and

$$Q_1 \approx 17.5 + \frac{12.5-12}{8} \cdot 8 = 18.$$

The size of the dataset is $n=50$, so the position of the upper quartile Q_3 is $0.75(n+1)=0.75 \cdot 51=38.25$ However, within the class interval, 41.5 – 49.5 contains 43 values, nearly in the upper quartile position, so the upper quartile class interval is 41.5 – 49.5. Therefore, $b_c=41.5$, $\sum f_1=33$, $f_{uq}=10$ and $w_c=8$, and

$$Q_1 \approx 41.5 + \frac{37.5-33}{10} \cdot 8 = 45.1.$$

The sample variance for the data in Table 1.5.4. is:

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n\bar{x}^2 \right) = \frac{1}{49} (61688.5 - 50 \cdot 32.06^2) = 210.13$$

and standard deviation $s = \sqrt{210.13} = 14.5$.

3. SAMPLE SPACE AND PROBABILITY

3.1. RANDOM EXPERIMENT AND SAMPLE SPACE

In the study of the theory of probability, an experiment is any process of observation. The results of the observation are called the outcomes of the experiment. A random experiment (phenomenon) is a situation in which we know what outcomes could happen, but we don't know which particular outcome did or will happen. Examples of a random experiment: are the roll of a die, the toss of a coin, drawing a card from a deck, and selecting a message signal for transmission from several messages.

In general, each occasion upon which we observe a random phenomenon is called a trial. At each trial, we note the value of the random phenomenon and call it an outcome.

Definition 3.1.1. The set of all possible outcomes of a random experiment is called a sample space (or universal set), and it is denoted by S .

An element in S is called a sample point. Each outcome of a random experiment corresponds to a sample point.

When we combine outcomes, the resulting combination is an event.

Definition 3.1.2. Event is any subset of the sample space S .

Let A be an event, a subset of S . We say the event A occurs if the outcome of the experiment is in the set A .

An event that cannot occur is called an impossible event.

An event that must occur is called a certain event.

Definition 3.1.3.

Containment: The event A is contained in the event B , i.e. $A \subseteq B$.

Equality: Two events are equal if $A \subseteq B$ and $B \subseteq A$.

Not A: Event “not A ” or opposite event of A consists of all outcomes not in A , denoted by $\bar{A} = S / A$.

Union: Union (sum) of two events A and B is the event that consists of all outcomes that belong to either event A or event B or both, denoted by $A \cup B$ or $A + B$.

Intersection: Intersection, or product, or joint event of two events A and B is the event that consists of all outcomes when both event A and event B occur, denoted by $A \cap B$ or AB .

Definition 3.1.4. Two events A and B are disjoint or mutually exclusive if $A \cap B = \emptyset$.

Definition 3.1.5. The events $A_1, A_2, \dots, A_n$ are pairwise mutually exclusive if $A_i \cap A_j = \emptyset$ for all $i \neq j, i, j \in \{1, \dots, n\}$.

Definition 3.1.6. The events $A_1, A_2, \dots, A_n$ are called mutually exclusive and exhaustive if $A_i \cap A_j = \emptyset$ for all $i \neq j, i, j \in \{1, \dots, n\}$ and

$$S = \bigcup_{i=1}^n A_i.$$

Example 3.1.1. Consider a random experiment of tossing a coin four times.

(a) Find the sample space S_1 , if we wish to observe the exact sequences of heads and tails obtained.

(b) Find the sample space S_2 , if we wish to observe the number of heads in the four tosses.

Solution. (a) The sampling space S_1 , is given by

$$S_1 = \left\{ \begin{array}{l} HHHH, HHHT, HHTH, HTHH, THHH, HHTT, HTTH, TTHH, \\ THHT, HTHT, THTH, HTTT, THTT, TTHT, TTTH, TTTT \end{array} \right\}$$

where, for example, $HTHT$ indicates a head on the first and third throws and a tail on the second and the fourth throws.

There are 16 sample points in S_1 .

(b) The sampling space S_2 , is given by $S_2 = \{0, 1, 2, 3, 4\}$

where, for example, outcome 3 indicates that three heads were obtained in the four tosses. The sample space S_2 contains five sample points.

Example 3.1.2. Consider an experiment of drawing two cards at random from a bag containing three cards marked with the integers 1 through 3.

(a) Find the sample space S_1 of the experiment if the first card is replaced before the second is drawn.

(b) Find the sample space S_2 of the experiment if the first card is not replaced.

Solution. (a) The sample space S_1 contains 9 ordered pairs $S_1 = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\}$ where in each pair the first number indicates the first number drawn.

(b) The sample space S_2 contains 6 ordered pairs

$$S_2 = \{(1, 2), (1, 3), (2, 1), (2, 3), (3, 1), (3, 2)\}$$

where in each pair the first number indicates the first number drawn.

Example 3.1.3. An experiment consists of rolling a die until a 5 is obtained.

(a) Find the sample space S_1 if we are interested in all possibilities.

(b) Find the sample space S_2 if we are interested in the number of throws needed to get a 5.

Solution. (a) The sample space S_1 would be

$$S_1 = \left\{ \begin{array}{l} 5, \\ 15, 25, 35, 45, 65 \\ 115, 125, 135, 145, 165, 215, 225, 235, 245, 265, 315, 325, 335, 345, 365, \dots \end{array} \right\}$$

where the first line indicates that a 5 is obtained in one throw, the second line indicates that a 5 is obtained in the second throw, the third line indicates that 5 is obtained in the third throw and etc.

(b) In this case, the sample space is $S_2 = \{i : i \geq 1\} = \{1, 2, 3, 4, 5, 6, \dots\}$

where i is an integer representing the number of throws needed to get a 5.

Example 3.1.4. Find the sample space for the experiment consisting of measurement of the length of time it takes for a light bulb to burn out, when its maximum “life length” of time is 10 000 hours.

Solution. A suitable sample space for this experiment would be $S = \{t \in \mathbb{R} : 0 \leq t \leq 10000\} = [0, 10000]$.

Example 3.1.5. An experiment consists of tossing two dice.

- (a) Find the sample space S .
- (b) Find the event A that the sum of the dots on the dice equals 4.
- (c) Find the event B that the sum of the dots on the dice is greater than 9.
- (d) Find the event C that the sum of the dots on the dice is less than 2.

Solution. (a) For this experiment, the sample space contains 36 ordered pairs

$$S = \left\{ \begin{array}{l} (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), \\ (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), \\ (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \end{array} \right\}$$

or $S = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$ where in each pair i represents the number of dots appearing on one die and j represents the number of dots appearing on the other die.

(b) The event A consists of 3 points $A = \{(1, 3), (2, 2), (3, 1)\}$.

(c) The event B consists of 6 points

$$B = \{(4, 6), (5, 5), (5, 6), (6, 4), (6, 5), (6, 6)\}$$

(d) The event C is an impossible event, that is, $C = \emptyset$.

Example 3.1.6. State every possible event in the sample space $S = \{a, b, c\}$.

Solution. By definition, an event is any subset of the sample space S . From Discrete Mathematics, we know that a set of n elements has 2^n subsets. Since the given sample space contains three elements, there are $2^3 = 8$ possible events:

$$A_1 = \emptyset, A_2 = \{a\}, A_3 = \{b\}, A_4 = \{c\}, A_5 = \{a, b\}, A_6 = \{a, c\}, A_7 = \{b, c\}, \\ A_8 = S = \{a, b, c\}$$

Example 3.1.7. Consider the experiment of tossing a coin repeatedly and of counting the number of tosses required until the first head appears. We define the events

$A = \{a : a \text{ is even}\}$, $B = \{a : 4 < a < 8\}$, $C = \{a : 2 < a \leq 11\}$, where a is the number of tosses required until the first H (head) appears. Determine the events:

$$\bar{A}, \bar{B}, \bar{C}, A \cup B, B \cup C, A \cap B, A \cap C, B \cap C \text{ and } \bar{A} \cap B.$$

Solution. Since,

$$A = \{a : a \text{ is even}\} = \{2, 4, 6, \dots\},$$

$$B = \{a : 4 < a < 8\} = \{5, 6, 7\},$$

$$C = \{a : 2 < a \leq 11\} = \{3, 4, 5, \dots, 11\},$$

so

$$\bar{A} = \{a : a \text{ is odd}\} = \{1, 3, 5, 7, \dots\},$$

$$\bar{B} = \{a : 1 \leq a \leq 4 \text{ or } a \geq 8\},$$

$$\bar{C} = \{a : a = 1, 2 \text{ or } a > 11\},$$

$$A \cup B = \{a : a = 5, 7 \text{ or } a \text{ is even}\},$$

$$B \cup C = C,$$

$$A \cap B = \{6\},$$

$$A \cap C = \{4, 6, 8, 10\},$$

$$B \cap C = B$$

Example 3.1.8. Verify the distributive law

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

Solution. Let $s \in A \cap (B \cup C)$. Then $s \in A$ and $s \in B \cup C$. This means $s \in A$ and either $s \in B$ or $s \in C$, so either $s \in A$ and $s \in B$ or $s \in A$ or $s \in C$. That is $s \in (A \cap B) \cup (A \cap C)$. Therefore, $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$.

Next, let $s \in (A \cap B) \cup (A \cap C)$. Then $s \in A$ and $s \in B$ or $s \in A$ or $s \in C$. Thus $s \in A$ and $s \in B \cup C$, that is $s \in A \cap (B \cup C)$. Therefore, $(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C)$. Thus, by the definition of equality, we have $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Example 3.1.9. Let A and B be arbitrary events. Show that $A \subseteq B$ if and only if $A \cap B = A$.

Solution: "If" part:

We show that if $A \cap B = A$, then $A \subseteq B$. Let $s \in A$. Then $s \in A \cap B$, since $A \cap B = A$.

Then by the definition of intersection, $s \in B$. Therefore, $A \subseteq B$.

"Only if" part :

We show that if $A \subseteq B$, then $A \cap B = A$. **Note** that from the definition of the intersection,

$A \cap B \subseteq A$. Suppose $s \in A$. Since $A \subseteq B$, then $s \in B$. So $s \in A$ and $s \in B$; that is, $s \in A \cap B$. Therefore, it follows that $A \subseteq A \cap B$. Hence, $A = A \cap B$. This completes the proof.

3.2. TECHNIQUES OF COUNTING

Proper determination of the total number of elementary events of a random experiment is very important in calculating the probability of events. The procedure by listing all the elementary events of random experiments is not always suitable for determining the number its possible outcomes. For example, for the experiment of tossing 10 coins, listing all the elementary events as 10-tuples is a long process and unreliable way to determine the sample set. There are some techniques for determining, without direct enumeration, the number of possible outcomes of a particular event or the number of elements in a set. Such sophisticated counting is sometimes called *combinatorial analysis*. It includes the study of permutations and combinations.

Sum Rule Principle:

Suppose some event E can occur in m ways and a second event F can occur in n ways, and suppose both events cannot occur simultaneously. Then E or F can occur in $m + n$ ways.

Product Rule Principle:

Suppose there are m independent events and each of them can occur in r_i , $i = 1, \dots, m$ ways.

The combinations of these events can occur in $r_1 \cdot r_2 \cdot \dots \cdot r_m$ ways.

For example, let consider m sets, such that the first set contains r_1 elements, the second set contains r_2 elements, ..., and the m set contains r_m elements. The sample set of the experiment of selecting one element from each set contains $r_1 \cdot r_2 \cdot \dots \cdot r_m$ elementary events.

Example 3.2.1. In a women's clothing department there are 3 different models of trousers, 5 models of jackets and 10 different designs of shirts. Based on the offer, a buyer can buy $3 \cdot 5 \cdot 10 = 150$ different combinations of these three products. Therefore, for the experiment of purchasing all three products simultaneously, each combination is one of the $n(S) = 150$ sample points.

3.2.1. PERMUTATIONS

Any arrangement of a set of n objects in a given order is called a *permutation* of the objects (taken all at a time).

Any arrangement of any $r \leq n$ of these objects in a given order is called an “ r -permutation” or “a permutation of the n objects taken r at a time.”

Consider, for example, the set of letters A, B, C, D . Then:

(i) $BDCA, DCBA$, and $ACDB$ are permutations of the four letters (taken all at a time).

(ii) BAD, ACB, DBC are permutations of the four letters taken three at a time.

(iii) AD, BC, CA are permutations of the four letters taken two at a time.

We usually are interested in the number of such permutations without listing them.

The number of permutations of n objects taken all at a time will be denoted by $P(n)$, and is equal to

$$P(n) = n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 2 \cdot 1,$$

where $n!$ is read “ n factorial”, by definition $0! = 1$.

The number of permutations of n objects taken r at a time will be denoted by

$P(n, r)$ (other texts may use ${}_nP_r$, $P_{n,r}$ or $(n)_r$, and is equal to

$$P(n, r) = \frac{n!}{(n-r)!} = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-r+1)$$

Example 3.2.2. There are $P(5) = 5! = 120$ different five-digit numbers formed of the digits 1, 3, 5, 7 and 9 without repetition.

Permutations with Repetitions

Frequently we want to know the number of permutations of a multiset, that is, a set of objects some of which are alike. We will let $P(n; n_1, n_2, \dots, n_m)$

denote the number of permutations of n objects of which n_1 are alike, n_2 are alike, $\dots$, n_m are alike and is equal to

$$P(n; n_1, n_2, \dots, n_m) = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_m!}$$

Example 3.2.3. Find the number m of eleven-letter words that can be formed using the letters of the word “MATHEMATICS.”

Solution:

We seek the number of permutations of 11 objects of which two M’s are alike, two A’s are alike and two T’s are alike, so it’s equal to

$$P(11; 2, 2, 2) = \frac{11!}{2! \cdot 2! \cdot 2!} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2!}{2 \cdot 2 \cdot 2!} = 4\,989\,600$$

Ordered Samples

Many problems are concerned with choosing an element from a set S , say, with n elements. When we choose one element after another, say, r times, we call the choice an *ordered sample* of size r . We consider two cases.

(1) Sampling with replacement

Here the element is replaced in the set S before the next element is chosen. Thus, each time there are n ways to choose an element (repetitions are allowed). The Product rule tells us that the number of such samples is:

$$\underbrace{n \cdot n \cdot n \cdot \dots \cdot n}_{r \text{ factors}} = n^r$$

(2) Sampling without replacement

Here the element is not replaced in the set S before the next element is chosen. Thus, there is no repetition in the ordered sample. Such a sample is simply an r -permutation. Thus the number of such samples is:

$$P(n, r) = \frac{n!}{(n-r)!} = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-r+1)$$

Example 3.2.4. Three cards are chosen one after the other from a 52-card deck. Find the number m of ways this can be done: (a) with replacement; (b) without replacement.

Solution.

(a) Each card can be chosen in 52 ways. Thus $m = 52 \cdot 52 \cdot 52 = 140\,608$.

(b) Here there is no replacement. Thus the first card can be chosen in 52 ways, the second in 51 ways, and the third in 50 ways. Therefore $m = P(52, 3) = 52 \cdot 51 \cdot 50 = 132\,600$.

Example 3.2.5. There are $6^3 = 216$ different three-digits numbers formed of the digits 1, 2, 3, 4, 5, 6 with repetition.

3.2.2. COMBINATIONS

Let S be a set with n elements. A *combination* of these n elements taken r at a time is any selection of r of the elements where order does not count. Such a selection is called an *r -combination*; it is simply a

subset of S with r elements. The number of such combinations is denoted by $C(n, r)$ (other texts may use ${}_nC_r$, $C_{n,r}$, C_r^n or $\binom{n}{r}$) and is equal to:

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

Example 3.2.6. List all combinations of 4 objects, A, B, C, D , taken 3 at a time.

Solution. All combination of three objects are:

ABC, ABD, ACD, BCD

Example 3.2.7. In a box are 4 white balls, 5 blue balls and 7 red balls. There are $\binom{4}{2} \cdot \binom{5}{3} \cdot \binom{7}{4} = 2100$ ways of choosing 2 white, 3 blue and 4 red balls at a time.

COMBINATIONS WITH REPETITIONS

A **r -combination with repetitions**, or **r -multicombination**, or **multiset** of size r from a set S of n elements is given by a sequence of r not necessarily distinct elements of S , where order is not taken into account: two sequences of which one can be obtained from the other by permuting the terms define the same multiset. The number of such r -combinations of n with repetitions is denoted by $\bar{C}(n, r)$ (other texts may use ${}_n\bar{C}_r$, $\bar{C}_{n,r}$ or $\bar{C}_r^n$) and is equal to:

$$\bar{C}(n, r) = C(n+r-1, r)$$

Example 3.2.8. A bakery makes only 4 kinds of cookies: apple, banana, strawberry, and carrot. Find the number of ways a person can buy 8 of the cookies.

Solution. Because the order does not count, every choice of buying 8 of the 4 kinds of cookies is an 8-combination of 4 with

repetition, so there are $\overline{C}(4,8) = C(4+8-1,8) = C(11,8) = 165$ ways of buying the cookies.

3.3. PROBABILITY

The probability of an event reports the likelihood of the event's occurrence.

In a given experiment, it is necessary to assign to each event A in the sample space S a number $P(A)$ that indicates the probability that A will occur. In order to satisfy the mathematical definition of probability, the number $P(A)$ that is assigned must satisfy three specific axioms. These axioms ensure that the number $P(A)$ will have certain properties that we intuitively expect a probability to have.

.

Axiom 1: For any event A , $P(A) \geq 0$ (a negative probability does not make sense).

Axiom 2: If S is the sample space for an experiment, $P(S) = 1$ (probabilities are normalized so that the maximum value is unity).

Axiom 3: If $A_1, A_2, \dots$ are pairwise mutually exclusive events, then $P\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} P(A_i)$

Theorem 3.3.1. Let A and B be any two events, then:

1. $P(\emptyset) = 0$
2. $P(A) \leq 1$
3. Complement rule: $P(\overline{A}) = 1 - P(A)$
4. $P(B \cap \overline{A}) = P(B) - P(A \cap B)$
5. General addition rule:

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

6. If $A \subseteq B$ then $P(A) \leq P(B)$

7. Bonferroni's inequality

$$P(A \cap B) \geq P(A) + P(B) - 1.$$

How are probabilities assigned to particular events?

Mathematically, any assignment that satisfies the given axioms is acceptable. Practically, the probability assignment actually represents the likelihood of occurrence of that event.

Two techniques are typically used for this purpose:

- Classical approach
- Relative frequency approach.

CLASSICAL APPROACH

Specify all the outcomes of the experiment in terms of some fundamental outcomes - the most basic events that cannot be decomposed into simpler events. Quite often we can justify assigning equal probabilities to all possible outcomes in an experiment (they are all equally likely).

Suppose an experiment S has $n(S)$ possible outcomes, all equally likely. An event A that can occur in $n(A)$ ways has probability of occurring: $P(A) = \frac{n(A)}{n(S)}$.

Consider an experiment of measuring the height of an arbitrarily chosen student in a class and rounding that measurement to the nearest inch. The possible outcomes of this experiment would consist of all the heights of the students in your class. It would not be reasonable to assign an equal probability to each height. The heights corresponding to very tall or very short students are expected to be less probable than the heights corresponding to a medium height.

Therefore, how do we assign probabilities to these events?

The problems associated with the classical approach to assigning probabilities can be overcome by using the relative frequency approach.

RELATIVE FREQUENCY APPROACH

The relative frequency approach requires that the experiment be repeatable.

The probability of an event A , can be assigned by repeating the experiment a large number of times and counting how many times the event A occurs.

If n is the number of times the experiment is repeated and n_A is the number of times the event A is observed, then the probability of the event A can be assigned according to

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}.$$

Because this definition is based on repeatedly observing the event's outcome, this definition of probability is often called empirical probability. Many random phenomena in which we might be interested are not repeatable. The situation may occur only once, and hence we cannot assign the probability according to the relative frequency approach.

Example 3.3.1. Let $P(A)=0.9$ and $P(B)=0.8$. Show that $P(A \cap B) \geq 0.7$

Solution.

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

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.9 + 0.8 - P(A \cup B) \geq 1.7 - 1 = 0.7.$$

Example 3.3.2. Show that $P(A) = P(A \cap B) + P(A \cap \bar{B})$.

Solution. Since

$$\begin{aligned} (A \cap B) \cup (A \cap \bar{B}) &= (A \cup (A \cap \bar{B})) \cap (B \cup (A \cap \bar{B})) = \\ &= A \cap ((B \cup A) \cap (B \cup \bar{B})) = A \cap ((B \cup A) \cap U) = A \cap (B \cup A) = A \end{aligned}$$

and $(A \cap B) \cap (A \cap \bar{B}) = A \cap B \cap \bar{B} = \emptyset$ then is true that:

$$P(A) = P((A \cap B) \cup (A \cap \bar{B})) = P(A \cap B) + P(A \cap \bar{B}).$$

Example 3.3.3. Find $P(B)$, if $P(A \cap B) = 0.36$ and $P(B \cap \bar{A}) = 0.23$.

Solution. From the second problem

$$P(B) = P(B \cap A) + P(B \cap \bar{A}) = 0.36 + 0.23 = 0.59.$$

Example 3.3.4. Given that $P(A) = 0.9$, $P(B) = 0.8$, $P(A \cap B) = 0.75$. Find $P(A \cup B)$, $P(A \cap \bar{B})$, $P(\bar{A} \cap \bar{B})$.

Solution.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.9 + 0.8 - 0.75 = 0.95,$$

$$P(A \cap \bar{B}) = P(A) - P(A \cap B) = 0.9 - 0.75 = 0.15$$

$$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.95 = 0.05$$

Example 3.3.5. Let roll a fair dice once. What is the probability that:

- a) number of dots that occurs are divisible by 3?
- b) the number of dots that occurs divisible by 2?

Solution. A: number of dots that occurs are divisible by 3,

$$A = \{3, 6\}, \quad S = \{1, 2, 3, 4, 5, 6\}.$$

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

B: number of dots that occurs are divisible by 2,

$$B = \{2, 4, 6\}, \quad S = \{1, 2, 3, 4, 5, 6\}.$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

Example 3.3.6. A fair dice is rolled twice. What is the probability that the sum of the occurred dots from each roll is equal to 8?

Solution.

A: the sum of the occurred dots from each roll is equal to 8,

$$S = \{(x, y) : x, y \in \{1, 2, 3, 4, 5, 6\}\},$$

$$A = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{5}{6^2} = \frac{5}{36}.$$

Example 3.3.7. In a box are 4 white, 8 red, and 9 yellow balls in the form of equal balls. Suddenly, without seeing, are drawn three balls. What is the probability that two of them are white and one red?

Solution.

A: two of three drawn balls are white and one red,

$$P(A) = \frac{n(A)}{n(S)} = \frac{\binom{4}{2} \binom{8}{1}}{\binom{21}{3}}$$

Example 3.3.8. Let's make a sequence from 5 white, 3 red, and 2 black balls. What is the probability that the first three balls from the sequence will be white?

Solution.

A: The first three balls from the sequence are white balls

$$P(A) = \frac{n(A)}{n(S)} = \frac{P(7; 2, 3, 2)}{P(10; 5, 3, 2)} = \frac{\frac{7!}{2! \cdot 3! \cdot 2!}}{\frac{10!}{5! \cdot 3! \cdot 2!}} = \frac{7! \cdot 5! \cdot 3! \cdot 2!}{10! \cdot 2! \cdot 3! \cdot 2!} = \frac{1}{12}.$$

Example 3.3.9. In a class of 32 students, none of them was born on 29 February. Calculate the probability of the event: at least two students were born on the same day.

Solution.

A : At least two students were born on the same day, $\bar{A}$: each student was born on a different day

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{P(365, 32)}{365^2} = 1 - \frac{\frac{365!}{(365-32)!}}{365^2}.$$

Example 3.3.10. There are 14 boys in one class of 30 students, and eight students were randomly selected for a team. Find the probability of these events, the team has:

- 1) just 4 girls, 2) at least 5 girls, 3) only boys.

Solution. There are 14 boys and $30-14=16$ girls.

A : the team has just 4 girls

$$P(A) = \frac{n(A)}{n(S)} = \frac{\binom{14}{4} \binom{16}{4}}{\binom{30}{8}}$$

B : the team has at least 5 girls

$$P(B) = \frac{n(B)}{n(S)} = \frac{\binom{16}{5}\binom{14}{3} + \binom{16}{6}\binom{14}{2} + \binom{16}{7}\binom{14}{1} + \binom{16}{8}}{\binom{30}{8}}$$

C : the team has only boys

$$P(C) = \frac{n(C)}{n(S)} = \frac{\binom{14}{8}}{\binom{30}{8}}.$$

Example 3.3.11. Five families, each consisting of a father, mother, and three children, were gathered in the same location. Arbitrary selections were made: one father, one mother, and one child. What is the probability that each of them belongs to the same family?

Solution. There are five fathers, five mothers, and 15 children.

A : The father, mother, and child are members of the same family

$$P(A) = \frac{n(A)}{n(S)} = \frac{\begin{matrix} \text{ways of choosing the family} \\ 5 \end{matrix} \cdot \begin{matrix} \text{ways of choosing a child from one family} \\ 3 \end{matrix}}{5 \cdot 5 \cdot 15} = \frac{1}{25}.$$

3.3.1. OTHER PROBABILITIES

Definition 3.4.1. The *conditional probability* of an event A given event B , denoted by $P(A \setminus B)$ is defined as:

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

where $P(A \cap B)$ is the joint probability of A and B .

Similarly,

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

is the conditional probability of an event B given event A .

Definition 3.4.2. Joint probability of events:

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

Bayes's Rule:

$$P(A \setminus B) = \frac{P(B \setminus A)P(A)}{P(B)}.$$

Definition 3.4.3. Let the the events $A_1, A_2, \dots, A_n$ are called *mutually exclusive and exhaustive* and B be any event in S . Then

$$P(B) = \sum_{i=1}^n P(B \cap A_i) = \sum_{i=1}^n P(B \setminus A_i)P(A_i)$$

which is known as the *total probability* of event B .

Bayes' theorem. Posteriori probability of any event $A_1, A_2, \dots, A_n$ in Definition 3.4.3 is calculated by:

$$P(A_i \setminus B) = \frac{P(B \setminus A_i)P(A_i)}{\sum_{i=1}^n P(B \setminus A_i)P(A_i)}$$

Definition 3.4.4. Two events A and B are said to be (*statistically*) *independent* if and only if $P(A \cap B) = P(A)P(B)$.

Note. It follows immediately that if A and B are independent, then

$$P(A \setminus B) = P(A) \quad \text{and} \quad P(B \setminus A) = P(B)$$

To distinguish between the mutual exclusiveness (or disjointness) and independence of a collection of events we summarize as follows:

Example 3.4.1. A fair coin is tossed three times. If at the first two tosses appeared same side, find the probability of the event that appeared at least one head.

Solution.

A : at the first two tosses appeared same side,

B : appeared at least one head

$$S = \{(x, y, z) \mid x, y, z \in \{F, H\}\}, n(S) = 2^3 = 8$$

$$A = \{(H, H, H), (H, H, F), (F, F, H), (F, F, F)\}, P(A) = \frac{4}{8}$$

$$A \cap B = \{(H, H, H), (H, H, F), (F, F, H)\}, P(A \cap B) = \frac{3}{8}$$

$$P(B \setminus A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{3}{8}}{\frac{4}{8}} = \frac{3}{4}.$$

Example 3.4.2. A fair dice is thrown twice. If at the both throws occurs dots with the same parity, find the probability of the event that the sum of the dots is less than 5.

Solution.

A : at the both throws occur dots with same parity,

B : sum of the dots is less than 5

$$S = \{(x, y) \mid x, y \in \{1, 2, \dots, 6\}\}, n(S) = 6^2 = 36$$

$$A = \{(x, y) \mid x, y \in \{1, 3, 5\}\} \cup \{(x, y) \mid x, y \in \{2, 4, 6\}\}, P(A) = \frac{2 \cdot 3^2}{36} = \frac{18}{36}$$

$$A \cap B = \{(1, 1), (1, 3), (3, 1), (2, 2)\}, P(A \cap B) = \frac{4}{36}$$

$$P(B \setminus A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{4}{36}}{\frac{18}{36}} = \frac{4}{18} = \frac{2}{9}.$$

Example 3.4.3. A fair dice is rolled twice. If at the first roll appeared even number, find the probability of the event that the sum was less than 9.

Solution.

A : at the first roll occurs even number,

B : sum of the dots was less than 9

$$S = \{(x, y) \mid x, y \in \{1, 2, \dots, 6\}\}, n(S) = 6^2 = 36$$

$$A = \{(x, y) \mid x \in \{2, 4, 6\}, y \in \{1, 2, 3, 4, 5, 6\}\}, P(A) = \frac{3 \cdot 6}{36} = \frac{18}{36}$$

$$A \cap B = \left\{ \begin{array}{l} (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (4,1), (4,2), (4,3), (4,4), \\ (6,1), (6,2) \end{array} \right\}, P(A \cap B) = \frac{12}{36}$$

$$P(B \setminus A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{12}{36}}{\frac{18}{36}} = \frac{12}{18} = \frac{2}{3}$$

Example 3.4.4. A fair dice is thrown twice. If at least once occur less than 3 dots, find the probability of the event that at least once occur not less than 3 dots.

Solution.

A : at least once occur less than 3 dots,

B : at least once occur not less than 3 dots

$$S = \{(x, y) \mid x, y \in \{1, 2, \dots, 6\}\}, n(S) = 6^2 = 36$$

$$A = \left\{ \begin{array}{l} (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), \\ (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), \\ (3,1), (3,2), (4,1), (4,2), (5,1), (5,2), (6,1), (6,2) \end{array} \right\}, P(A) = \frac{20}{36} = \frac{5}{9}$$

$$A \cap B = \left\{ (1,3), (1,4), (1,5), (1,6), \right. \\ \left. (2,3), (2,4), (2,5), (2,6), \right. \\ \left. (3,1), (3,2), (4,1), (4,2), (5,1), (5,2), (6,1), (6,2) \right\}, P(A \cap B) = \frac{16}{36}$$

$$P(B \setminus A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{16}{36}}{\frac{18}{36}} = \frac{16}{18} = \frac{8}{9}$$

Example 3.4.5. A fair dice is thrown three times. If the sum is greater than 15, find the probability of the event that at least once appear 6 dots.

Solution.

A : sum of the dots is greater than 15,

B : at least once appear 6 dots

$$S = \{(x, y, z) \mid x, y, z \in \{1, 2, \dots, 6\}\}, n(S) = 6^3 = 216$$

$$A = \{(4, 6, 6), (6, 4, 6), (6, 6, 4), (5, 5, 6), (5, 6, 5), (6, 5, 5), \\ (6, 6, 5), (6, 5, 6), (5, 6, 6), (6, 6, 6)\}, P(A) = \frac{10}{216}$$

$$A \cap B = A, P(A \cap B) = \frac{10}{216}$$

$$P(B \setminus A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{10}{216}}{\frac{10}{216}} = 1$$

Example 3.4.6. The box contains 4 white balls and 4 black balls. Three balls are drawn at random. Calculate the probability that all the balls are white, given that at least one of the selected balls is white.

Solution.

A : at least one of the selected balls is white,

B : all selected balls are white

$$P(\bar{A}) = \frac{\binom{4}{3}}{\binom{8}{3}} = \frac{\frac{4!}{3!1!}}{\frac{8!}{3!5!}} = \frac{4}{\frac{2 \cdot 8 \cdot 7 \cdot 6 \cdot 5!}{3 \cdot 2 \cdot 1 \cdot 3!}} = \frac{1}{14}, \quad P(A) = 1 - P(\bar{A}) = 1 - \frac{1}{14} = \frac{13}{14},$$

$$P(A \cap B) = P(B) = \frac{\binom{4}{3}}{\binom{8}{3}} = \frac{1}{14}$$

$$P(B \setminus A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{14}}{\frac{13}{14}} = \frac{1}{13}$$

Example 3.4.7. Each of the two boxes contains 5 white balls and 6 black balls. One ball is drawn from the first box and placed into the second. Then, a ball is randomly drawn from the second box and placed into the first. What is the probability that the ball from the second box is black?

Solution.

A_1 : in the second box are 5 white and 7 black balls,

A_2 : in the second box are 6 white and 6 black balls,

B : the ball from the second box is black

$$P(B) = P(B/A_1)P(A_1) + P(B/A_2)P(A_2) = \frac{7}{12} \cdot \frac{6}{11} + \frac{6}{12} \cdot \frac{5}{11} = \frac{12 \cdot 6}{12 \cdot 11} = \frac{6}{11}.$$

Example 3.4.8. One store has seven TVs on one shelf and five on the other, with one defective TV on each shelf. A TV from the first shelf is placed on the second shelf, and a customer buys one TV from the second shelf. What is the probability that the TV purchased is defective?

Solution.

A_1 : The second shelf contains two defective TVs,

A_2 : The second shelf contains one defective TV,

B : The customer purchased a defective TV

$$P(B) = P(B/A_1)P(A_1) + P(B/A_2)P(A_2) = \frac{2}{6} \cdot \frac{1}{7} + \frac{1}{6} \cdot \frac{6}{7} = \frac{8}{3 \cdot 7} = \frac{4}{21}.$$

Example 3.4.9. Three archers hit at target with a probability $p=0,2$; $q=0,4$; $r=0,8$, respectively. Randomly is chosen one archer who shoots at goal. What is the probability that the target will be hit?

Solution.

A_1 : The first archer shoots,

A_2 : The second archer shoots,

A_3 : The third archer shoots

B : The target is hit

$$\begin{aligned} P(B) &= P(B/A_1)P(A_1) + P(B/A_2)P(A_2) + P(B/A_3)P(A_3) = \\ &= 0.2 \cdot \frac{1}{3} + 0.4 \cdot \frac{1}{3} + 0.8 \cdot \frac{1}{3} = 1.4 \cdot \frac{1}{3} = 0.47 \end{aligned}$$

Example 3.4.10. 25% men and 30% women are left handed. From a group of people having an equal number of men and women randomly is selected one person.

What is the probability that the selected person is left handed?

If the selected person is left handed, how likely is it a man?

Solution.

A_1 : The selected person is a woman,

A_2 : The selected person is a man,

B : The selected person is left handed

$$P(B) = P(B/A_1)P(A_1) + P(B/A_2)P(A_2) = 0.3 \cdot \frac{1}{2} + 0.25 \cdot \frac{1}{2} = 0.55 \cdot \frac{1}{2} = 0.275$$

$$P(A_2/B) = \frac{P(B/A_2)P(A_2)}{P(B)} = \frac{0.25 \cdot \frac{1}{2}}{0.55 \cdot \frac{1}{2}} = 0.45.$$

4. RANDOM VARIABLES

Sometimes the sample space is too large, but by focusing on a specific question of interest, we can reduce it. For example, consider an opinion poll where we ask 50 people whether they agree or disagree with a certain issue. We can assign "1" for those who agree and "0" for those who disagree.

The sample space S of this experiment is composed of 2^{50} ordered strings of 1s and 0s of length 50. But in this experiment the usual question of interest is the number of people who agree.

Therefore, we could define a variable $X : S \rightarrow \mathbb{R}$ such that for any outcome $s \in S$ it assigns an integer $X(s)$ which is the number of 1s in that outcome. So, $X(s) \in \{0, 1, \dots, 50\}$.

Since this variable is defined on the sample space and the probability P could induce a probability on the set of values of the variable X , this variable is called a random variable. A formal mathematical definition of a random variable is given in Definition 4.1.

Definition 4.1. The function $X : S \rightarrow \mathbb{R}$ such that for any $x \in \mathbb{R}$ the set $\{s \mid X(s) \leq x\}$ is an event is called a random variable.

We denote an event $(X \leq x) = \{s \mid X(s) \leq x\}$.

The set of all values of the random variable is called a range of the random variable and it is denoted by $R_X = \{X(s) \mid s \in S\}$.

Since for the event $X \leq x$ is defined probability $P(X \leq x)$, this probability induces the following probability $P_X : R_X \rightarrow [0, 1]$ defined by

$$P_X(x) = P(X \leq x), \quad x \in R_X,$$

and it is called the probability distribution of the random variable X .

Next, we drop the index X and the following function is well defined.

Definition 4.2. The function $F_X : \mathbb{R} \rightarrow \mathbb{R}$ defined by:

$$F_X(x) = P(X < x),$$

is called the cumulative distribution function (cdf) of the random variable X .

Theorem 4.1. Let F is cdf of X . Then:

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

Proof. Since $(-\infty, b] = (-\infty, a] \cup (a, b]$, and since the events on the right side are disjoint events:

$$P(X \leq b) = P(X \leq a) + P(X \in (a, b]) \quad (4.1)$$

If we plugin in (4.1) $F(b) = P(X \leq b)$ and $F(a) = P(X \leq a)$

we obtain that

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

so,

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

Theorem 4.2. Let the function F is cdf of the random variable X , then F is monotonically nondecreasing, i.e. for any $x_1, x_2 \in \mathbb{R}$ such that

$$x_1 < x_2 \text{ follows } F(x_1) \leq F(x_2).$$

Proof. Let $x_1 < x_2$, then $\{X \leq x_1\} \subseteq \{X \leq x_2\}$. So, by Theorem 3.3.1. $P(X \leq x_1) \leq P(X \leq x_2)$ and by the definition of cdf

$$F(x_1) \leq F(x_2).$$

Theorem 4.3. Let the function F is cdf of the random variable X , then $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow +\infty} F(x) = 1$.

Theorem 4.4. Any cdf F is continuous from the right, i.e., $\lim_{x \rightarrow x_0^-} F(x) = F(x_0)$, for all $x_0 \in \mathbb{R}$.

Let $A, B \subseteq \mathbb{R}$.

Definition 4.3. Two random variables X and Y are independent if for any $A, B \subseteq \mathbb{R}$ is valid that:

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Random variables are categorized into discrete and continuous variables. The two categories differ primarily by the nature of possible values that can be assumed by the variables. In addition, the nature of the random variable implies the particular method of deriving a probability distribution function.

A random variable can be a discrete random variable if it has a finite range or, at most, a countably infinite range. It can be a continuous random variable if it can take any value within an interval.

4.1. DISCRETE RANDOM VARIABLES

Let X is a discrete random variable. Then its range is either a finite set $R_X = \{x_1, x_2, \dots, x_n\}$ or countably infinite set $R_X = \{x_1, x_2, \dots\}$.

Obviously, for any $x \in \mathbb{R} \setminus R_X$, $P(X = x) = 0$.

Therefore, we define:

Definition 4.1.1. The probability mass function (pmf) of the random variable X is the function $P_X : \mathbb{R} \rightarrow \mathbb{R}$ defined by:

$$P_X(x) = P(X = x) \text{ for any } x \in \mathbb{R}.$$

Note, $\sum_{x \in R_X} P_X(x) = \sum_{x \in R_X} P(X = x) = 1.$

We could drop the index X , and then the pmf of the random variable X is defined by:

$$P(x) = P(X = x) \text{ for any } x \in \mathbb{R}.$$

If $P: \mathbb{R} \rightarrow \mathbb{R}$ is pmf of a discrete random variable X with range $R_X = \{x_1, x_2, \dots\}$. We denote $p_i = P(X = x_i)$ for all $i = 1, 2, \dots$

then $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^{\infty} p_i = 1,$

Theorem 4.1.1. Let $P: \mathbb{R} \rightarrow \mathbb{R}$ is pmf of the random variable X then

with range $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^{\infty} p_i = 1,$

- 1) $0 \leq P(x) \leq 1$ for any $x \in \mathbb{R}.$
- 2) $\sum_{x \in R_X} P(x) = 1.$
- 3) $P(X \in A) = \sum_{x \in A} P(x),$ for any set $A \subseteq R_X.$

Proof. Since the first and the third statements are obvious we will prove the second statement i.e., since X is a discrete random variable then its range is finite set $R_X = \{x_1, x_2, \dots, x_n\}$ or countably infinite set $R_X = \{x_1, x_2, \dots\}$. Then,

$$\sum_{x \in R_X} P(x) = \sum_{i=1}^n P(X = x_i) \text{ or } \sum_{x \in R_X} P(x) = \sum_{i=1}^{\infty} P(X = x_i), \text{ respectively.}$$

Since $\{X = x_1\}, \{X = x_2\}, \dots, \{X = x_n\}$ or $\{X = x_1\}, \{X = x_2\}, \dots$ are pairwise mutually exclusive and exhaustive events, these are valid

$$\sum_{i=1}^n P(X = x_i) = P\left(\bigcup_{i=1}^n \{X = x_i\}\right) = P(S) = 1$$

and

$$\sum_{i=1}^n P(X = x_i) = P\left(\bigcup_{i=1}^n \{X = x_i\}\right) = P(S) = 1.$$

Therefore, in any case $\sum_{x \in R_X} P(x) = 1$.

If the pmf is determined, we could obtain the cdf, by the following procedure.

Let's observe a discrete random variable with a finite range R_X , i.e., $R_X = \{x_1, x_2, \dots, x_n\}$.

Namely, let's assume that the range of the random variable X is the set $R_X = \{x_1, x_2, \dots, x_n\}$ where the values $x_1, x_2, \dots, x_n$ are given in increasing order i.e., $x_1 < x_2 < \dots < x_n$ with probabilities p_i , $i = 1, 2, \dots, n$, respectively.

Notice that,

if $x < x_1$ then $F_X(x) = P(X \leq x) = P(\emptyset) = 0$;

if $x_1 \leq x < x_2$ then $F_X(x) = P(X \leq x) = P(x = x_1) = p_1$;

if $x_2 \leq x < x_3$ then

$$F_X(x) = P(X \leq x) = P(X = x_1 \cup X = x_2) = P(X = x_1) + P(X = x_2) = p_1 + p_2$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

if $x_{n-1} \leq x < x_n$ then

$$F_X(x) = P(X \leq x) = P\left(x \in \bigcup_{i=1}^{n-1} (X = x_i)\right) = \sum_{i=1}^{n-1} P(X = x_i) = p_1 + p_2 + \dots + p_{n-1}$$

and if $x \geq x_n$ then

$$F_X(x) = P(X \leq x) = P\left(x \in \bigcup_{i=1}^n (X = x_i)\right) = \sum_{i=1}^n P(X = x_i) = p_1 + \dots + p_n.$$

Therefore, the cdf of the random variable X given by its pmf defined as in Definition 4.1.1 is the function:

$$F(x) = \begin{cases} 0, & x < x_1 \\ p_1, & x_1 \leq x < x_2 \\ p_1 + p_2, & x_2 \leq x < x_3 \\ \vdots & \\ p_1 + p_2 + \dots + p_{n-1}, & x_{n-1} \leq x < x_n \\ 1, & x \geq x_n \end{cases}.$$

The opposite is also true, i.e., if the cdf is given, then the pmf can be uniquely determined.

Let F of a discrete random variable with a range

$$R_X = \{x_1, x_2, \dots, x_n\} \text{ or } R_X = \{x_1, x_2, \dots\}.$$

Then for any $i \geq 1$ is true $\{X \leq x_i\} = \{X \leq x_{i-1}\} \cup \{X = x_i\}$, and the events $\{X \leq x_{i-1}\}$ and $\{X = x_i\}$ are disjoint. So,

$$P(X \leq x_i) = P(\{X \leq x_{i-1}\} \cup \{X = x_i\}) = P(X \leq x_{i-1}) + P(X = x_i)$$

$$F(x_i) = F(x_{i-1}) + P(X = x_{i+1})$$

$$P(X = x_i) = F(x_i) - F(x_{i-1})$$

Now, we will define two so-called numerical characteristics, i.e. parameters of discrete random variables.

Definition 4.1.2. The mean or mathematical expectation of a discrete random variable X is the number

$$E(X) = \sum_{x \in R_x} x \cdot P(X = x),$$

The mean of a discrete random variable X is a weighted average of the possible values that the random variable can take. The common symbol for the mean is the Greek letter μ .

Theorem 4.1.1. (Properties of the mean)

1. If $X_1, X_2, \dots, X_n$ are random variables, then

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i)$$

2. $E(a) = a$ for any $a \in \mathbb{R}$.
3. If $a, b \in \mathbb{R}$ then $E(aX + b) = aE(X) + b$.
4. If the random variables X and Y are independent then $E(XY) = E(X) \cdot E(Y)$.

Definition 4.1.3. The variance of the random variable X is the number

$$\text{Var}(X) = E\left((X - E(X))^2\right) \quad (4.2)$$

Definition 4.1.4. The standard deviation of the random variable X is the number $\sqrt{\text{Var}(X)}$.

By the Theorem 4.1.1.

$$\begin{aligned} \text{Var}(X) &= E\left((X - E(X))^2\right) = E\left(X^2 - 2X \cdot E(X) + (E(X))^2\right) = \\ &= E(X^2) - E(2E(X) \cdot X) + E\left((E(X))^2\right) = \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 = \\ &= E(X^2) - 2(E(X))^2 + (E(X))^2 = \\ &= E(X^2) - (E(X))^2 \end{aligned} \quad (4.3)$$

Note. Equation (4.3) is more convenient to use for calculating the variance of the random variable defined by the equation (4.2).

In addition, if X is a discrete random variable, then

$$E(X^2) = \sum_{x \in R_x} x^2 \cdot P(X = x).$$

Theorem 4.1.2. (Properties of the variance)

1. $Var(X) \geq 0$ for any random variable X
2. $Var(aX) = a^2 Var(X)$ for any $a \in \mathbb{R}$ and any random variable X
3. $Var(a + X) = Var(X)$ for any $a \in \mathbb{R}$ and any random variable X
4. If X and Y are independent random variables then

$$Var(X + Y) = Var(X) + Var(Y)$$

5. If $X_1, X_2, \dots, X_n$ are pairwise independent random variables, then

$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i).$$

Example 4.1.1. Consider the experiment of throwing a fair die. Let X be the r.v which assigns 1 if the number that appears is even and 0 if the number that appears is odd.

- (a) What is the range of X ?
- (b) Find pmf and cdf of the r.v X ?
- (c) Determine the probability of the event that the number appears is odd.
- (d) Find the mean and variance of the r.v X .

Solution.

(a) The range of X is $R_X = \{0, 1\}$.

(b) $(X = 0) = \{1, 3, 5\}$. Thus, $P(X = 0) = \frac{3}{6} = \frac{1}{2}$.

$(X = 1) = \{2, 4, 6\}$. Thus, $P(X = 1) = \frac{3}{6} = \frac{1}{2}$.

The pmf of the r.v X is:

X	0	1
$P(X)$	$\frac{1}{2}$	$\frac{1}{2}$

The cdf of the r.v X is:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & 0 \leq x < 1. \\ 1, & x \geq 1 \end{cases}$$

(c) A : the number that appears is odd. $P(A) = P(X = 0) = \frac{1}{2}$

(d) The mean of the r.v X is $E(X) = \sum_{x=0}^1 x \cdot P(X = x)$,

$$E(X) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$$

The variance of the r.v X is $Var(X) = E(X^2) - (E(X))^2$,

where $E(X) = \sum_{x=0}^3 x^2 \cdot P(X = x)$.

Since, $E(X^2) = 0^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{1}{2} = \frac{1}{2}$,

$$Var(X) = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

Example 4.1.2. Consider the experiment of tossing a coin three times. Let X be the r.v giving the number of heads obtained. We assume that the tosses are independent and the probability of a head is $\frac{1}{3}$.

- (a) What is the range of X ?
- (b) Find pmf and cdf of the r.v X .
- (c) Find the probability of the event that at least one tail is obtained.
- (d) Find the mean and variance of the r.v X .

Solution.

(a) The range of X is $R_X = \{0, 1, 2, 3\}$.

(b) H : In one toss a head is obtained, $P(H) = \frac{1}{3}$.

T : In one toss a tail is obtained, $P(T) = P(\overline{H}) = \frac{2}{3}$

$$P(X = 0) = \left(\frac{2}{3}\right)^3 = \frac{8}{27},$$

$$P(X=1) = \binom{3}{1} \frac{1}{3} \cdot \left(\frac{2}{3}\right)^2 = 3 \cdot \frac{4}{27} = \frac{12}{27},$$

$$P(X=2) = \binom{3}{2} \left(\frac{1}{3}\right)^2 \cdot \frac{2}{3} = 3 \cdot \frac{2}{27} = \frac{6}{27}$$

$$P(X=3) = \left(\frac{1}{3}\right)^3 = \frac{1}{27}$$

The pmf of the r.v X is:

X	0	1	2	3
$P(X)$	$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$

The cdf of the r.v X is:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{8}{27}, & 0 \leq x < 1 \\ \frac{20}{27}, & 1 \leq x < 2 \\ \frac{26}{27}, & 2 \leq x < 3 \\ 1, & x \geq 3 \end{cases}$$

(c) A : Tail is obtained at least once,

$\bar{A}$: Tail is not obtained in each toss.

$$P(\bar{A}) = P(X=3) = \frac{1}{27}, \quad P(A) = 1 - P(\bar{A}) = 1 - \frac{1}{27} = \frac{26}{27}.$$

(d) The mean of the r.v X is $E(X) = \sum_{x=0}^3 x \cdot P(X = x)$,

$$E(X) = 0 \cdot \frac{8}{27} + 1 \cdot \frac{12}{27} + 2 \cdot \frac{6}{27} + 3 \cdot \frac{1}{27} = \frac{12+12+3}{27} = 1$$

The variance of the r.v X is $Var(X) = E(X^2) - (E(X))^2$, where

$$E(X) = \sum_{x=0}^3 x^2 \cdot P(X = x).$$

Since, $E(X^2) = 0^2 \cdot \frac{8}{27} + 1^2 \cdot \frac{12}{27} + 2^2 \cdot \frac{6}{27} + 3^2 \cdot \frac{1}{27} = \frac{12+24+9}{27} = \frac{45}{27}$,

$$Var(X) = \frac{45}{27} - 1^2 = \frac{18}{27} = \frac{2}{3}.$$

Example 4.1.3. Consider the experiment of tossing a fair coin repeatedly until the first head appears, but a maximum of five times. Let the r.v. X denotes the number of tosses.

- (a) Find the pmf $p_X(x)$ and the cdf $F_X(x)$ of X ,
- (b) Find $P(1 < X \leq 4)$,
- (c) Find the mean and variance of the r.v X .

Solution.

(a) The range of X is $R_X = \{1, 2, 3, 4, 5\}$.

$$P(X = 1) = \frac{1}{2}, \quad P(X = 2) = \left(\frac{1}{2}\right)^2 = \frac{1}{4}, \quad P(X = 3) = \left(\frac{1}{2}\right)^3 = \frac{1}{8},$$

$$P(X = 4) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}, \quad P(X = 5) = \left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^5 = \frac{2}{32} = \frac{1}{16}$$

So, the pmf of the r.v X is:

X	1	2	3	4	5
$P(X)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$

The cdf of the r.v X is:

$$F_X(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{2}, & 1 \leq x < 2 \\ \frac{3}{4}, & 2 \leq x < 3 \\ \frac{7}{8}, & 3 \leq x < 4 \\ \frac{15}{16}, & 4 \leq x < 5 \\ 1, & x \geq 5 \end{cases}$$

$$(b) \quad P(1 < X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4) = \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{7}{16}$$

$$(c) \quad \text{The mean of the r.v } X \text{ is } E(X) = \sum_{x=1}^5 x \cdot P(X = x),$$

$$E(X) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + 5 \cdot \frac{1}{16} = \frac{8+8+6+4+5}{16} = \frac{31}{16}$$

The variance of the r.v X is $Var(X) = E(X^2) - (E(X))^2$, where

$$E(X) = \sum_{x=1}^5 x^2 \cdot P(X=x).$$

Since,

$$E(X^2) = 1^2 \cdot \frac{1}{2} + 2^2 \cdot \frac{1}{4} + 3^2 \cdot \frac{1}{8} + 4^2 \cdot \frac{1}{16} + 5^2 \cdot \frac{1}{16} = \frac{8+16+18+16+25}{16} = \frac{83}{16}$$

So, the variance is

$$Var(X) = \frac{83}{16} - \left(\frac{31}{16}\right)^2 = \frac{1328}{256} - \frac{961}{256} = \frac{367}{256} = 1.43.$$

4.2. CONTINUOUS RANDOM VARIABLES

Definition 4.2.1. If there exists an integrable function $p(x)$ on (a, b) such that for the cdf function of X is valid:

$$F(x) = \int_{-\infty}^x p(t)dt \text{ for any } x \in (a, b)$$

then the random variable тогаш велиме дека случајната променлива X is called a continuous random variable.

In this case, the function $p(x)$ from the Definition 4.2.1. is called the probability density function (pdf) of X .

Theorem 4.2.1. Let X is a continuous random variable with pdf $p(x)$. Then:

- 1) $p(x) \geq 0$ for any $x \in \mathbb{R}$;

$$2) \int_{-\infty}^{+\infty} p(x) dx = 1;$$

$$3) P(a < X < b) = \int_a^b p(x) dx.$$

Definition 4.2.2. The mean or mathematical expectation of the continuous random variable X is the number

$$E(X) = \int_{-\infty}^{+\infty} x \cdot p(x) dx.$$

If X is a continuous random variable, then

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 \cdot p(x) dx$$

and then we could obtain the variance of the continuous random variable X by

$$\text{Var}(X) = E(X^2) - (E(X))^2$$

Example 4.2.1. Let X be a continuous r.v. X with pdf

$$p_X(x) = \begin{cases} ax^2, & x \in [-1, 1] \\ 0, & x \notin [-1, 1] \end{cases}$$

where a is a constant.

- (a) Determine the value of a ,
- (b) Find the corresponding cdf $F_X(x)$,
- (c) Find $P(0 \leq X < 0.5)$,
- (d) Find the mean and variance of the r.v. X .

Solution.

- (a) Since the function $p_X(x)$ is pdf of the r.v. X is true that

$$\int_{-\infty}^{+\infty} p_X(x) dx = 1.$$

Therefore,

$$\begin{aligned} \int_{-\infty}^{+\infty} p_X(x) dx = 1 &\Leftrightarrow \int_{-\infty}^{-1} 0 dx + \int_{-1}^1 ax^2 dx + \int_1^{+\infty} 0 dx = 1 \Leftrightarrow a \frac{x^3}{3} \Big|_{-1}^1 = 1 \Leftrightarrow \\ &\Leftrightarrow \frac{a}{3} (1^3 - (-1)^3) = 1 \Leftrightarrow \frac{2}{3} a = 1 \end{aligned}$$

and $a = \frac{3}{2}$.

(b) Since $a = \frac{3}{2}$, the pdf of r.v X is $p_X(x) = \begin{cases} \frac{3}{2}x^2, & x \in [-1, 1] \\ 0, & x \notin [-1, 1] \end{cases}$.

So, if $x < -1$ then $F_X(x) = P(X \leq x) = \int_{-\infty}^x p_X(t) dt = \int_{-\infty}^x 0 dt = 0$,

if $-1 \leq x < 1$ then

$$\begin{aligned} F_X(x) &= P(X \leq x) = \int_{-\infty}^x p_X(t) dt = \int_{-\infty}^{-1} 0 dt + \int_{-1}^x \frac{3}{2} t^2 dt = \frac{3}{2} \frac{t^3}{3} \Big|_{-1}^x = \frac{1}{2} (x^3 - (-1)^3) = \\ &= \frac{1}{2} (x^3 + 1) \end{aligned}$$

if $x \geq 1$ then

$$\begin{aligned} F_X(x) &= P(X \leq x) = \int_{-\infty}^x p_X(t) dt = \int_{-\infty}^{-1} 0 dt + \int_{-1}^1 \frac{3}{2} t^2 dt + \int_1^x 0 dt = \frac{3}{2} \frac{t^3}{3} \Big|_{-1}^1 = \\ &= \frac{1}{2} (1^3 - (-1)^3) = \frac{1}{2} (1 + 1) = 1 \end{aligned}$$

Therefore the cdf of r.v X is:

$$F_X(x) = \begin{cases} 0, & x < -1 \\ \frac{1}{2}(x^3 + 1), & -1 \leq x < 1. \\ 1, & x \geq 1 \end{cases}$$

(c)

$$\begin{aligned} P(0 \leq X < 0.5) &= \int_0^{0.5} p_X(x) dx = \int_0^{0.5} \frac{3}{2} x^2 dx = \frac{3}{2} \frac{x^3}{3} \Big|_0^{0.5} = \frac{1}{2} (0.5^3 - 0^3) = \\ &= \frac{1}{2} \cdot 0.125 = 0.0625 \end{aligned}$$

(d) The mean of the r.v X is

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x \cdot p_X(x) dx = \int_{-\infty}^{-1} x \cdot 0 dx + \int_{-1}^1 x \cdot \frac{3}{2} x^2 dx + \int_1^{+\infty} x \cdot 0 dx = \\ &= \int_{-1}^1 \frac{3}{2} x^3 dx = \frac{3}{2} \frac{x^4}{4} \Big|_{-1}^1 = \frac{3}{8} (1^4 - (-1)^4) = 0 \end{aligned}$$

The variance of the r.v X is $Var(X) = E(X^2) - (E(X))^2$, where

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 \cdot p_X(x) dx.$$

Since

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{+\infty} x^2 \cdot p_X(x) dx = \int_{-\infty}^{-1} x^2 \cdot 0 dx + \int_{-1}^1 x^2 \cdot \frac{3}{2} x^2 dx + \int_1^{+\infty} x^2 \cdot 0 dx = \\ &= \int_{-1}^1 \frac{3}{2} x^4 dx = \frac{3}{2} \frac{x^5}{5} \Big|_{-1}^1 = \frac{3}{10} (1^5 - (-1)^5) = \frac{3}{5} \end{aligned}$$

the variance is $Var(X) = \frac{3}{5} - 0^2 = 0.6$.

Example 4.2.2. Let X be a continuous r.v. X with pdf

$$p_X(x) = \begin{cases} a(6+x), & x \in [0, 3] \\ a(6-x), & x \in [3, 6] \\ 0, & x \notin [0, 6] \end{cases}$$

where a is a constant.

- (a) Determine the value of a ,
- (b) Find the corresponding cdf $F_X(x)$,
- (c) Find $P(2 \leq X < 5)$,
- (d) Find the mean and variance of the r.v. X .

Solution.

(a) Since the function $p_X(x)$ is pdf of the r.v. X is true that

$$\int_{-\infty}^{+\infty} p_X(x) dx = 1.$$

Therefore,

$$\begin{aligned} \int_{-\infty}^{+\infty} p_X(x) dx = 1 &\Leftrightarrow \int_{-\infty}^0 0 dx + \int_0^3 a(6+x) dx + \int_3^6 a(6-x) dx + \int_6^{+\infty} 0 dx = 1 \Leftrightarrow \\ &\Leftrightarrow 6a \int_0^3 dx + a \int_0^3 x dx + 6a \int_3^6 dx - a \int_3^6 x dx = 1 \Leftrightarrow \left(6ax + a \frac{x^2}{2} \right) \Big|_0^3 + \left(6ax - a \frac{x^2}{2} \right) \Big|_3^6 = 1 \Leftrightarrow \\ &\Leftrightarrow \left(\left(6a \cdot 3 + a \frac{3^2}{2} \right) - \left(6a \cdot 0 + a \frac{0^2}{2} \right) \right) + \left(\left(6a \cdot 6 - a \frac{6^2}{2} \right) - \left(6a \cdot 3 - a \frac{3^2}{2} \right) \right) = 1 \Leftrightarrow \end{aligned}$$

$$\Leftrightarrow 18a + a\frac{9}{2} + \left((36a - 18a) - \left(18a - a\frac{9}{2} \right) \right) = 1 \Leftrightarrow 18a + a\frac{9}{2} + 18a - 18a + a\frac{9}{2} = 1 \Leftrightarrow \\ \Leftrightarrow 27a = 1$$

and $a = \frac{1}{27}$.

(b) Since $a = \frac{1}{27}$, the pdf of r.v X is

$$p_X(x) = \begin{cases} \frac{1}{27}(6+x), & x \in [0, 3] \\ \frac{1}{27}(6-x), & x \in [3, 6] \\ 0, & x \notin [0, 6] \end{cases}$$

So, if $x < 0$ then $F_X(x) = P(X \leq x) = \int_{-\infty}^x p_X(t) dt = \int_{-\infty}^x 0 dt = 0,$

if $0 \leq x < 3$ then

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x p_X(t) dt = \int_{-\infty}^0 0 dt + \int_0^x \frac{1}{27}(6+t) dt = \frac{1}{27} \left(6t + \frac{t^2}{2} \right) \Big|_0^x = \\ = \frac{1}{27} \left(6x + \frac{x^2}{2} \right)$$

if $3 \leq x < 6$ then

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x p_X(t) dt = \int_{-\infty}^0 0 dt + \int_0^3 \frac{1}{27}(6+t) dt + \int_3^x \frac{1}{27}(6-t) dt = \\ = \frac{1}{27} \left(6t + \frac{t^2}{2} \right) \Big|_0^3 + \frac{1}{27} \left(6t - \frac{t^2}{2} \right) \Big|_3^x = \\ = \frac{1}{27} \left(6 \cdot 3 + \frac{3^2}{2} \right) + \frac{1}{27} \left(\left(6x - \frac{x^2}{2} \right) - \left(6 \cdot 3 - \frac{3^2}{2} \right) \right) =$$

$$\begin{aligned}
 &= \frac{1}{27} \left(6 \cdot 3 + \frac{3^2}{2} \right) - \frac{1}{27} \left(6 \cdot 3 - \frac{3^2}{2} \right) + \frac{1}{27} \left(6x - \frac{x^2}{2} \right) = \\
 &= \frac{9}{27} + \frac{1}{27} \left(6x - \frac{x^2}{2} \right) = \frac{1}{27} \left(9 + 6x - \frac{x^2}{2} \right)
 \end{aligned}$$

if $x \geq 6$ then

$$\begin{aligned}
 F_X(x) &= P(X \leq x) = \int_{-\infty}^x p_X(t) dt = \int_{-\infty}^0 0 dt + \int_0^3 \frac{1}{27} (6+t) dt + \int_3^6 \frac{1}{27} (6-t) dt + \int_6^x 0 dt = \\
 &= \frac{1}{27} \left(6t + \frac{t^2}{2} \right) \Big|_0^3 + \frac{1}{27} \left(6t - \frac{t^2}{2} \right) \Big|_3^6 = \\
 &= \frac{1}{27} \left(6 \cdot 3 + \frac{3^2}{2} \right) + \frac{1}{27} \left(\left(6 \cdot 6 - \frac{6^2}{2} \right) - \left(6 \cdot 3 - \frac{3^2}{2} \right) \right) = \\
 &= \frac{1}{27} \left(6 \cdot 3 + \frac{3^2}{2} \right) - \frac{1}{27} \left(6 \cdot 3 - \frac{3^2}{2} \right) + \frac{1}{27} \left(36 - \frac{36}{2} \right) = \frac{9}{27} + \frac{18}{27} = 1
 \end{aligned}$$

Therefore the cdf of r.v X is:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{27} \left(6x + \frac{x^2}{2} \right), & 0 \leq x < 3 \\ \frac{1}{27} \left(9 + 6x - \frac{x^2}{2} \right), & 3 \leq x < 6 \\ 1, & x \geq 6 \end{cases} .$$

(c)

$$\begin{aligned}
 P(2 \leq X < 5) &= \int_2^5 p_X(x) dx = \int_2^3 \frac{1}{27} (6+t) dt + \int_3^5 \frac{1}{27} (6-t) dt = \\
 &= \frac{1}{27} \left(6t + \frac{t^2}{2} \right) \Big|_2^3 + \frac{1}{27} \left(6t - \frac{t^2}{2} \right) \Big|_3^5 =
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{27} \left(\left(6 \cdot 3 + \frac{3^2}{2} \right) - \left(6 \cdot 2 + \frac{2^2}{2} \right) \right) + \frac{1}{27} \left(\left(6 \cdot 5 - \frac{5^2}{2} \right) - \left(6 \cdot 3 - \frac{3^2}{2} \right) \right) = \\
 &= \frac{1}{27} \left(6 \cdot 3 + \frac{3^2}{2} \right) - \frac{1}{27} \left(6 \cdot 3 - \frac{3^2}{2} \right) - \frac{14}{27} + \frac{1}{27} \left(30 - \frac{25}{2} \right) = \\
 &= \frac{9}{27} - \frac{14}{27} + \frac{35}{54} = \frac{25}{54}
 \end{aligned}$$

(d) The mean of the r.v X is

$$\begin{aligned}
 E(X) &= \int_{-\infty}^{+\infty} x \cdot p_X(x) dx = \int_{-\infty}^0 0 dt + \int_0^3 \frac{1}{27} x(6+x) dx + \int_3^6 \frac{1}{27} x(6-x) dx + \int_6^{+\infty} 0 dt = \\
 &= \frac{1}{27} \left(6 \frac{x^2}{2} + \frac{x^3}{3} \right) \Big|_0^3 + \frac{1}{27} \left(6 \frac{x^2}{2} - \frac{x^3}{3} \right) \Big|_3^6 = \\
 &= \frac{1}{27} \left(3 \cdot 3^2 + \frac{3^3}{3} \right) + \frac{1}{27} \left(\left(3 \cdot 6^2 - \frac{6^3}{3} \right) - \left(3 \cdot 3^2 - \frac{3^3}{3} \right) \right) = \\
 &= \frac{1}{27} (27 + 9) - \frac{1}{27} (27 - 9) + \frac{1}{27} (108 - 72) = \frac{18}{27} + \frac{36}{27} = \frac{54}{27} = 2
 \end{aligned}$$

The variance of the r.v X is $Var(X) = E(X^2) - (E(X))^2$, where

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 \cdot p_X(x) dx.$$

Since

$$\begin{aligned}
 E(X^2) &= \\
 &= \int_{-\infty}^{+\infty} x^2 \cdot p_X(x) dx = \int_{-\infty}^0 0 dt + \int_0^3 \frac{1}{27} x^2 (6+x) dx + \int_3^6 \frac{1}{27} x^2 (6-x) dx + \int_6^{+\infty} 0 dt =
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{27} \left(6 \frac{x^3}{3} + \frac{x^4}{4} \right) \Bigg|_0^3 + \frac{1}{27} \left(6 \frac{x^3}{3} - \frac{x^4}{4} \right) \Bigg|_3^6 = \\
 &= \frac{1}{27} \left(2 \cdot 3^3 + \frac{3^4}{4} \right) + \frac{1}{27} \left(\left(2 \cdot 6^3 - \frac{6^4}{4} \right) - \left(2 \cdot 3^3 - \frac{3^4}{4} \right) \right) = \\
 &= \frac{1}{27} \left(18 + \frac{81}{4} \right) - \frac{1}{27} \left(18 - \frac{81}{4} \right) + \frac{1}{27} (432 - 324) = \frac{81}{54} + 4 = \frac{3}{2} + 4 = \frac{11}{2}
 \end{aligned}$$

the variance is $\text{Var}(X) = \frac{11}{2} - 2^2 = \frac{3}{2}$.

4.3. SPECIAL DISCRETE RANDOM VARIABLES

4.3.1. BERNOULLI DISTRIBUTION

The simplest type of experiment has only two possible outcomes, call them success and failure. Let the $P(\text{success}) = p$. If a random variable X takes value 0 if failure, and 1 if success, this random variable is called Bernoulli distribution with parameter p and is denoted by $\mathcal{B}(p)$.

The pmf of $\mathcal{B}(p)$ is:

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

Example 4.3.1. The treatment given to a particular patient in a clinical trial can either succeed or fail. Let $X=0$ if the treatment fails, and let $X=1$ if the treatment succeeds. This is a Bernoulli distribution given by the pmf

$$P(X=1) = p, P(X=0) = 1-p$$

The mean of Bernoulli distribution is:

$$E(X) = 0 \cdot (1-p) + 1 \cdot p = p.$$

For the variance first we obtain:

$$E(X^2) = 0^2 \cdot (1-p) + 1^2 \cdot p = p$$

Therefore, the variance of Bernoulli distribution is::

$$\text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1-p).$$

4.3.2. BINOMIAL DISTRIBUTION

The random variable X that counts the number of successes if we repeat Bernoulli's experiments n times is called Binomial distribution with parameters n and p , denoted by $\mathcal{B}(n, p)$.

Since, $P(\text{success}) = p$ and $q = 1-p$, then the probability of k successes in n independent Bernoulli trials is:

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n.$$

The above expression gives the pmf of $\mathcal{B}(n, p)$.

Now, let's calculate the mean of Binomial distribution with parameters n and p directly.

$$\begin{aligned} E(X) &= \sum_{k=0}^n k \cdot P(X = k) = \sum_{k=0}^n k \cdot \binom{n}{k} p^k q^{n-k} = \sum_{k=1}^n k \cdot \binom{n}{k} p^k q^{n-k} = \\ &= \sum_{k=1}^n k \cdot \frac{n!}{k!(n-k)!} p^k q^{n-k} = \sum_{k=1}^n \frac{n!}{(k-1)!(n-k)!} p^k q^{n-k} = \\ &= p \cdot \sum_{k=1}^n \frac{n \cdot (n-1)!}{(k-1)!(n-(k-1)-1)!} p^{k-1} q^{n-k} = \\ &= np \cdot \sum_{k=1}^n \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} q^{(n-1)-(k-1)} = \\ &= np \cdot \sum_{i=0}^{n-1} \frac{(n-1)!}{i!((n-1)-i)!} p^i q^{(n-1)-i} = np \cdot (p+q)^{n-1} = np \cdot 1 = np \end{aligned}$$

Since the calculations of the variance are more complicated we could consider the Binomial distribution as a sum of n independent Bernoulli random variables X_i , $i = 1, 2, \dots, n$, i.e.

$$X = X_1 + X_2 + \dots + X_n.$$

Since,

$$E(X)_i = p \text{ and } Var(X_i) = pq, \quad i = 1, 2, \dots, n,$$

by Theorem 4.1.2.

$$Var(X) = Var(X_1 + X_2 + \dots + X_n) = Var(X_1) + \dots + Var(X_n) = npq.$$

4.3.3. GEOMETRIC DISTRIBUTION

Let us consider an experiment in which the probability of occurrence of a particular event A is $P(A) = p$ and $q = 1 - p$. Then we obtain this experiment independently until the event A has occurred.

Let A_i is the event: in i -th experiment the event A occurs, for all $i = 1, 2, \dots$.

The random variable X that counts how many times we repeat the experiment until the event A occurs is called geometric distribution with parameter p and its pmf is given by:

$$P(X = k) = P(\bar{A}_1 \bar{A}_2 \dots \bar{A}_{k-1} A_k) = P(\bar{A}_1) P(\bar{A}_2) \dots P(\bar{A}_{k-1}) P(A_k) = q^{k-1} p$$

for any integer $k \geq 1$,

By the properties of the geometric series:

$$\sum_{k=1}^{\infty} q^{k-1} p = p \sum_{i=0}^{\infty} q^i = p \cdot \frac{1}{1-q} = p \cdot \frac{1}{p} = 1,$$

and

$$E(X) = \sum_{k=1}^{+\infty} k \cdot q^{k-1} p = p(1 + 2q + 3q^2 + \dots) = p(1 + 2q + 3q^2 + \dots) =$$

$$\begin{aligned}
 &= p \frac{d}{dq} (q + q^2 + q^3 + \dots) = p \frac{d}{dq} (q(1 + q + q^2 + \dots)) \cdot p \frac{d}{dq} \left(q \cdot \frac{1}{1-q} \right) = \\
 &= p \cdot \frac{q'(1-q) - q(1-q)'}{(1-q)^2} = p \cdot \frac{(1-q) + q}{p^2} = \frac{1}{p}
 \end{aligned}$$

Since,

$$\begin{aligned}
 \text{Var}(X) &= E(X^2) - (E(X))^2 = E(X^2) - E(X) + E(X) - (E(X))^2 = \\
 &= E(X(X-1)) + E(X) - (E(X))^2 = E(X(X-1)) + \frac{1}{p} - \frac{1}{p^2}
 \end{aligned}$$

and

$$\begin{aligned}
 E(X(X-1)) &= \sum_{k=1}^{+\infty} k(k-1) \cdot q^{k-1} p = pq \sum_{k=1}^{+\infty} k(k-1) \cdot q^{k-2} = pq \sum_{k=1}^{+\infty} \frac{d^2}{dq^2} (q^k) = \\
 &= pq \frac{d^2}{dq^2} \left(\sum_{k=1}^{+\infty} q^k \right) = pq \frac{d^2}{dq^2} \left(q \sum_{k=1}^{+\infty} q^{k-1} \right) = pq \frac{d^2}{dq^2} \left(q \sum_{i=0}^{+\infty} q^i \right) = \\
 &= pq \frac{d^2}{dq^2} \left(\frac{q}{1-q} \right) = pq \frac{d}{dq} \left(\frac{d}{dq} \left(\frac{q}{1-q} \right) \right)' = pq \left(\frac{1}{(1-q)^2} \right)' = \\
 &= pq \frac{d}{dq} \left((1-q)^{-2} \right) = pq \cdot (-2) \frac{-1}{(1-q)^3} = pq \cdot \frac{2}{p^3} = \frac{2q}{p^2}
 \end{aligned}$$

the variance of geometric series is

$$\text{Var}(X) = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q + p - 1}{p^2} = \frac{2q - (1-p)}{p^2} = \frac{2q - q}{p^2} = \frac{q}{p^2}.$$

4.3.4. POISSON DISTRIBUTION

Poisson distribution is a random variable that gives the probability of an event occurring a certain number of times (k) within a unit interval of time or space if the probability of occurring the event is the parameter λ and is denoted by $\mathcal{P}(\lambda)$.

The probability mass function of the Poisson distribution is given by:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, 2, \dots$$

By the series expansion of the function e^x i.e., $e^x = \sum_{k=0}^{+\infty} \frac{x^k}{k!}$,

we notice that:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

The mean of the Poisson distribution is:

$$\begin{aligned} E(X) &= \lambda e^{-\lambda} + 2\lambda^2 \frac{e^{-\lambda}}{2!} + \dots + n \frac{\lambda^n}{n!} e^{-\lambda} = \lambda e^{-\lambda} \left(1 + \lambda + \dots + \frac{\lambda^{n-1}}{(n-1)!} + \frac{\lambda^n}{n!} \dots \right) = \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned}$$

Since,

$$\begin{aligned} E(X^2) &= \lambda e^{-\lambda} + 2^2 \lambda^2 \frac{e^{-\lambda}}{2!} + \dots + n^2 \frac{\lambda^n}{n!} e^{-\lambda} + \dots = \\ &= \lambda e^{-\lambda} \left(1 + 2\lambda + \dots + n \frac{\lambda^{n-1}}{(n-1)!} + (n+1) \frac{\lambda^n}{n!} + \dots \right) = \\ &= \lambda e^{-\lambda} \left(\left(1 + \lambda + \dots + \frac{\lambda^n}{n!} + \dots \right) + \lambda + \dots + n \frac{\lambda^n}{n!} + \dots \right) = \\ &= \lambda e^{-\lambda} \left(e^{\lambda} + \lambda \left(1 + \lambda + \dots + \frac{\lambda^{n-1}}{(n-1)!} + \frac{\lambda^n}{n!} + \dots \right) \right) = \end{aligned}$$

$$= \lambda e^{-\lambda} (e^{\lambda} + \lambda e^{\lambda}) = \lambda + \lambda^2.$$

the variance of the random variable is:

$$\text{Var}(X) = E(X^2) - (EX)^2 = \lambda + \lambda^2 - \lambda^2 = \lambda.$$

So, the Poisson distribution has the same mean and variance.

Next theorem gives an approximation of Binomial distribution as the number of the trials n large number and the probability of success is a very small.

Theorem 4.3.1. (Poisson Theorem) Let n approach infinity and p approach zero in such a way that $\lim_{n \rightarrow \infty} np = \alpha$. Then:

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\alpha^k}{k!} e^{-\alpha}.$$

Example 4.3.1. A software manufacturer is aware that one out of 10 software games that the company markets will be financially successful. The manufacturer selects 10 new games to market. What is the probability that exactly one game will be a financial success? What is the probability that at least two games will be successful?

Solution.

Let

$$p = P(\text{success}) = \frac{1}{10}$$

$$P(\text{one success}) = \binom{10}{1} p(1-p)^9 = 0.3874$$

$$\begin{aligned} P(\text{at least 2 success}) &= 1 - P(\text{less than 2 success}) = \\ &= 1 - P(\text{no 2 success}) - P(1 \text{ success}) = \end{aligned}$$

$$= 1 - \binom{10}{0} p^0 (1-p)^{10} - \binom{10}{1} p (1-p)^9 = 1 - 0.3487 - 0.3487 = 0.2639.$$

Example 4.3.2. A fair die is rolled repeatedly five times. Find the probability of the event that a 6 occurred at least once.

Solution.

Let

$$p = P(\text{success}) = P(\text{a "6" occurs}) = \frac{1}{6},$$

$$q = 1 - p = P(\text{failure}) = P(\text{a "6" doesn't occur}) = \frac{5}{6}$$

$$\begin{aligned} P(\text{in five trials at least one success}) &= 1 - P(\text{no success}) = \\ &= 1 - \binom{5}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^5 = 1 - \left(\frac{5}{6}\right)^5 \end{aligned}$$

Example 4.3.3. A test consists of 10 multiple-choice questions with five choices for each question. As an experiment, you GUESS the answer to each question without reading them. What is the probability of getting exactly 6 questions correct on this test?

Solution.

Let

$$p = P(\text{success}) = P(\text{guessing the correct answer}) = \frac{1}{5},$$

$$q = 1 - p = P(\text{failure}) = P(\text{guessing the incorrect answer}) = \frac{4}{5}$$

$$P(6 \text{ successes}) = \binom{10}{6} p^6 q^{10-6} = \binom{10}{6} \left(\frac{1}{5}\right)^6 \left(\frac{4}{5}\right)^4$$

Example 4.3.4. At a certain intersection, the light for eastbound traffic is red for 15 seconds, yellow for 5 seconds, and green for 30 seconds. Determine the probability that exactly three out of the next eight

eastbound cars arriving randomly at the light will be stopped by a red light.

Solution. Let

$$p = P(\text{success}) = P(\text{red light}) = \frac{15}{50} = \frac{3}{10},$$

$$q = 1 - p = P(\text{not red light}) = \frac{7}{10}$$

$$P(3 \text{ of } 8 \text{ cars will stop}) = \binom{8}{3} p^3 q^{8-3} = \binom{8}{3} \left(\frac{3}{10}\right)^3 \left(\frac{7}{10}\right)^5$$

Example 4.3.5. In a certain lottery, six numbers are randomly chosen from the set $\{0,1,2,\dots,49\}$ (without replacement). To win the lottery, a player must guess correctly all six numbers, but it is not necessary to specify in which order the numbers are selected.

a) What is the probability of winning the lottery with only one ticket?

b) Suppose 6 million lottery tickets are sold in a given week, and each player independently chooses any of the possible number combinations with equal likelihood. What is the probability that exactly four players correctly select the winning combination? Use the Poisson approximation to calculate this probability.

c) Assuming again that 6 million tickets are sold, what is the most likely number of winning tickets? Use the Poisson approximation.

Solution.

a) Since

$$p = P(\text{win with one ticket}) = \frac{1}{\binom{50}{6}} = 6.29 \cdot 10^{-8}$$

b) Here, for the Poisson approximation,

$$\alpha = np = \frac{1}{\binom{50}{6}} \cdot 6 \cdot 10^6 = 0.3774.$$

Hence,

$$P(4) = \frac{\alpha^4}{4!} e^{-\alpha} = \frac{0.3774^4}{24} e^{-0.3774} = 5.7955 \cdot 10^{-4}.$$

c) The probability that n people win is

$$P(n) = \frac{\alpha^n}{n!} e^{-\alpha} = \frac{0.3774^n}{n!} e^{-0.3774}$$

When $n = 0$,

$$P(\text{no winners}) = P(0) = \frac{\alpha^0}{0!} e^{-\alpha} = \frac{0.3774^0}{0!} e^{-0.3774} = 0.6799$$

Since this probability is greater than 0.5, it is impossible for any other number of winners to be more probable. Hence the most probable number of winning tickets is zero.

4.4. SPECIAL CONTINUOUS RANDOM VARIABLES

4.4.1. UNIFORM DISTRIBUTION ON AN INTERVAL

A uniform distribution explains an experiment where there is an arbitrary outcome that lies between certain bounds. The bounds are defined by the parameters a and b , which are the minimum and maximum values. The interval can either be closed or open.

Therefore, the distribution is abbreviated as $\mathcal{U}(a, b)$ or $\mathcal{U}[a, b]$.

The random variable X has a uniform distribution on an interval (a, b) , if its probability density function is:

$$p(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b) \\ 0, & x \notin (a, b) \end{cases}.$$

Let's obtain the cdf of the uniform distribution X i.e.,

If $x \leq a$

$$F(x) = \int_{-\infty}^x p(t) dt = \int_{-\infty}^x 0 dt = 0;$$

If $a < x < b$

$$F(x) = \int_{-\infty}^x p(t) dt = \int_{-\infty}^a 0 dt + \int_a^x \frac{1}{b-a} dt = \frac{1}{b-a} \cdot t \Big|_a^x = \frac{x-a}{b-a};$$

If $x \geq b$,

$$\begin{aligned} F(x) &= \int_{-\infty}^x p(t) dt = \int_{-\infty}^a 0 dt + \int_a^b p(t) dt + \int_b^x 0 dt = \int_a^b \frac{1}{b-a} dt = \frac{1}{b-a} \cdot t \Big|_a^b = \\ &= \frac{b-a}{b-a} = 1 \end{aligned}$$

so the cdf of uniform distribution function on an interval (a, b) is defined by:

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

The mean of uniform distribution on an open interval (a, b) is,

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} xp(x) dx = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \\ &= \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2} \\ \frac{b^2 - a^2}{2(b-a)} &= \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}. \end{aligned}$$

Since

$$\begin{aligned} EX^2 &= \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \frac{x^3}{3} \Big|_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{(b-a)(b^2 + ab + a^2)}{3(b-a)} = \\ &= \frac{a^2 + ab + b^2}{3} \end{aligned}$$

the variance of a uniform distribution on (a, b) is,

$$\begin{aligned} \text{Var}(X) &= E(X^2) - (EX)^2 = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \\ &= \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{a^2 - 2ab + b^2}{12} = \frac{(a-b)^2}{12}. \end{aligned}$$

4.4.2. EXPONENTIAL DISTRIBUTION

The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events. Namely, The exponential distribution is the probability distribution of the distance between events in a Poisson in which events occur continuously and independently at a constant average rate λ .

The pdf of an exponential distribution X with parameter λ denoted by $\mathcal{E}(\lambda)$ is:

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

Then since

for $x \leq 0$, $F(x) = 0$ and

$$\text{for } x > 0, F(x) = \int_0^x \lambda e^{-\lambda t} dt = \lambda \frac{e^{-\lambda t}}{-\lambda} \Big|_0^x = e^{-\lambda t} \Big|_x^0 = 1 - e^{-\lambda x}.$$

The cdf of the exponential distribution is

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda x}, & x > 0 \end{cases}.$$

The mean of an exponential distribution with parameter λ , is

$$E(X) = \int_{-\infty}^{+\infty} xp(x)dx = \int_0^{+\infty} \lambda xe^{-\lambda x} dx = \lambda \lim_{x \rightarrow +\infty} \int_0^x te^{-\lambda t} dt,$$

since:

$$\begin{aligned} \lambda \int te^{-\lambda t} dt &= \left(\begin{array}{ll} u = t & dv = e^{-\lambda t} dt \\ du = dt & v = \frac{e^{-\lambda t}}{-\lambda} \end{array} \right) = \lambda \left(-\frac{te^{-\lambda t}}{\lambda} + \frac{1}{\lambda} \int e^{-\lambda t} dt \right) = \\ &= -te^{-\lambda t} + \frac{e^{-\lambda t}}{-\lambda}, \end{aligned}$$

then

$$\begin{aligned} E(X) &= \lim_{x \rightarrow +\infty} \lambda \int_0^x te^{-\lambda t} dt = \lim_{x \rightarrow +\infty} \left(-te^{-\lambda t} + \frac{e^{-\lambda t}}{-\lambda} \right) \Big|_0^x = \\ &= \lim_{x \rightarrow +\infty} \left(-xe^{-\lambda x} - \frac{e^{-\lambda x}}{\lambda} + \frac{1}{\lambda} \right) = \lim_{x \rightarrow +\infty} (-xe^{-\lambda x}) - \lim_{x \rightarrow +\infty} \frac{e^{-\lambda x}}{\lambda} + \frac{1}{\lambda} = \\ &= \lim_{x \rightarrow +\infty} \frac{-x}{e^{\lambda x}} - \lim_{x \rightarrow +\infty} \frac{1}{\lambda e^{\lambda x}} + \frac{1}{\lambda} = \lim_{x \rightarrow +\infty} \frac{-(x)'}{(e^{\lambda x})'} + \frac{1}{\lambda} = \lim_{x \rightarrow +\infty} \frac{-1}{\lambda e^{\lambda x}} + \frac{1}{\lambda} = \frac{1}{\lambda} \end{aligned}$$

Now since,

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 p(x) dx = \int_0^{+\infty} \lambda x^2 e^{-\lambda x} dx = \lambda \lim_{x \rightarrow +\infty} \int_0^x t^2 e^{-\lambda t} dt,$$

where

$$\begin{aligned} \lambda \int t^2 e^{-\lambda t} dt &= \left(\begin{array}{ll} u = t^2 & dv = e^{-\lambda t} dt \\ du = 2t dt & v = \frac{e^{-\lambda t}}{-\lambda} \end{array} \right) = \lambda \left(-\frac{t^2 e^{-\lambda t}}{\lambda} + \frac{2}{\lambda} \int te^{-\lambda t} dt \right) = \\ &= -t^2 e^{-\lambda t} + 2 \left(-\frac{te^{-\lambda t}}{\lambda} + \frac{1}{\lambda} \int e^{-\lambda t} dt \right) = -t^2 e^{-\lambda t} - \frac{2te^{-\lambda t}}{\lambda} - \frac{2e^{-\lambda t}}{\lambda^2}, \end{aligned}$$

$$\begin{aligned}
 E(X^2) &= \lim_{x \rightarrow +\infty} \lambda \int_0^x t^2 e^{-\lambda t} dt = \lim_{x \rightarrow +\infty} \left(-t^2 e^{-\lambda t} - \frac{2te^{-\lambda t}}{\lambda} - \frac{2e^{-\lambda t}}{\lambda^2} \right) \Big|_0^x = \\
 &= \lim_{x \rightarrow +\infty} \left(-x^2 e^{-\lambda x} - \frac{2xe^{-\lambda x}}{\lambda} - \frac{2(e^{-\lambda x} - 1)}{\lambda^2} \right) = \\
 &= \lim_{x \rightarrow +\infty} (-x^2 e^{-\lambda x}) - \lim_{x \rightarrow +\infty} \left(\frac{2xe^{-\lambda x}}{\lambda} \right) - \lim_{x \rightarrow +\infty} \left(\frac{2e^{-\lambda x}}{\lambda^2} \right) + \frac{2}{\lambda^2} = \\
 &= \lim_{x \rightarrow +\infty} \left(-\frac{x^2}{e^{\lambda x}} \right) - \frac{2}{\lambda} \lim_{x \rightarrow +\infty} \left(\frac{x}{e^{\lambda x}} \right) - \frac{2}{\lambda^2} \lim_{x \rightarrow +\infty} \left(\frac{1}{e^{\lambda x}} \right) + \frac{2}{\lambda^2} = \\
 &= \lim_{x \rightarrow +\infty} \left(-\frac{x^2}{e^{\lambda x}} \right) - \frac{2}{\lambda} \lim_{x \rightarrow +\infty} \left(\frac{x}{e^{\lambda x}} \right) + \frac{2}{\lambda^2}
 \end{aligned}$$

By the L'Hôpital's rule the limits from above when $x \rightarrow +\infty$ are 0.

So, $E(X^2) = \frac{2}{\lambda^2}$, and therefore,

$$Var(X) = E(X^2) - (EX)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

4.4.3. GAUSSIAN OR NORMAL DISTRIBUTION

The random variable X has a normal distribution with parameters μ and σ^2 , denoted by $N(\mu, \sigma^2)$, if its pdf is:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \mu \in \mathbb{R}, \quad \sigma > 0.$$

Note, μ is the mean and σ^2 is the variance of this distribution.

The pdf of the normal distribution with the mean μ and variance σ^2 is symmetric is “bell shaped” symmetric to the line $x = \mu$.

The normal distribution with mean 0 and variance 1 has pdf $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2}}$ and it is called standard normal distribution abbreviated by $N(0,1)$, see Fig. 4.1.

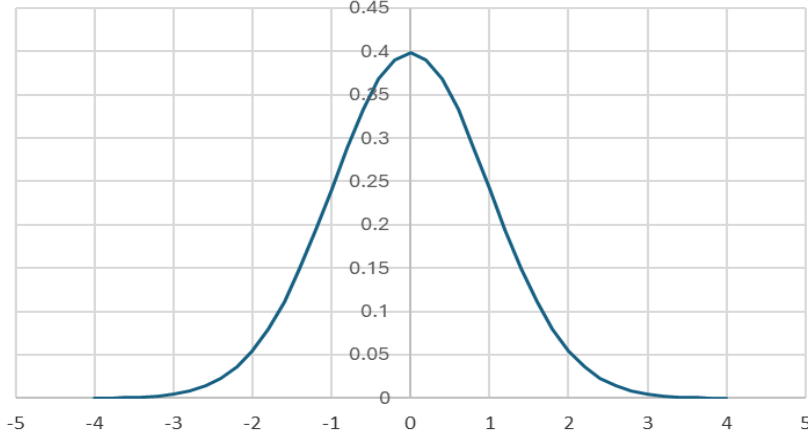


Figure 4.1. Standard normal distribution

The pdf of the standard normal distribution is often denoted with the Greek letter Φ . The antiderivative of the pdf of a standard normal distribution doesn't exist, but the value

$$\Phi(z) = \int_{-\infty}^z p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2}} dx$$

which gives the probability of the event $\{Z \leq z\}$ can be approximated very accurately by a variety of methods, such as numerical integration, Taylor series, or asymptotic series. These values for $z \geq 0$ are included in a special table known as the Table for the Φ function and given below.

If $z < 0$ then $\Phi(z) = 1 - \Phi(-z)$.

If the random variable X has a normal distribution with parameters μ and σ^2 , then the random variable

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution.

Therefore, we could consider that any normal distribution is obtained by the Table for the Φ function.

Theorem 4.4.1. (Normal approximation to the binomial distribution) If X is a Binomial distribution with parameters n and p i.e., $\mathcal{B}(n, p)$ then if $np \geq 10$ and $nq \geq 10$, when n goes to infinite X has a normal distribution with mean np and variance npq i.e.,

for any $a, b, k \in \mathbb{R}$

$$P(a \leq X \leq b) \approx P\left(\frac{a - np}{\sqrt{npq}} \leq \frac{X - np}{\sqrt{npq}} \leq \frac{b - np}{\sqrt{npq}}\right) = \Phi\left(\frac{b - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - np}{\sqrt{npq}}\right)$$

$$\text{and } P(X \leq k) \approx P\left(\frac{X - np}{\sqrt{npq}} \leq \frac{k - np}{\sqrt{npq}}\right) = \Phi\left(\frac{k - np}{\sqrt{npq}}\right).$$

Example 4.4.1. A fair coin is tossed 900 times. Find the probability of the event that the number of heads is between 420 and 465.

Solution.

$$P(\text{head at one trial}) = p = \frac{1}{2},$$

Since $np = nq = 900 \cdot \frac{1}{2} = 450 \geq 10$, then

$$\begin{aligned} P(420 \leq k \leq 465) &= \Phi\left(\frac{465 - 420}{\sqrt{225}}\right) - \Phi\left(\frac{465 - 450}{\sqrt{225}}\right) = \Phi(1) - \Phi(-2) = \\ &= \Phi(1) - (1 - \Phi(2)) = 0.8413 + 0.9772 - 1 = 0.8185. \end{aligned}$$

Example 4.4.2. If $P(\text{success}) = 0.6$ and k is the number of successes in n trials.

- a) Show that $P(550 \leq k \leq 650) = 0.999$ for $n = 1000$.
- b) Find n such that $P(0.59n \leq k \leq 0.61n) = 0.95$.

Solution.

- a) Since $np = 1000 \cdot 0.6 = 600 \geq 10$, $nq = 1000 \cdot 0.4 = 400 \geq 10$ by Theorem 4.4.1.

$$\begin{aligned}
 P(550 \leq k \leq 650) &= P\left(\frac{550 - np}{\sqrt{npq}} \leq \frac{k - np}{\sqrt{npq}} \leq \frac{650 - np}{\sqrt{npq}}\right) = \\
 &= P\left(\frac{550 - 600}{\sqrt{240}} \leq \frac{k - np}{\sqrt{npq}} \leq \frac{650 - 600}{\sqrt{240}}\right) = \Phi(3.23) - \Phi(-3.23) = \\
 &= \Phi(3.23) - (1 - \Phi(3.23)) = 2 \cdot \Phi(3.23) - 1 = 2 \cdot 0.9994 - 1 = 0.999
 \end{aligned}$$

b) Because

$$\begin{aligned}
 0.95 &= P(0.59n \leq k \leq 0.61n) = P\left(\frac{0.59n - np}{\sqrt{npq}} \leq \frac{k - np}{\sqrt{npq}} \leq \frac{0.61n - np}{\sqrt{npq}}\right) = \\
 &= \Phi\left(\frac{0.61n - np}{\sqrt{n \cdot 0.6 \cdot 0.4}}\right) - \Phi\left(\frac{0.59n - np}{\sqrt{n \cdot 0.6 \cdot 0.4}}\right) = \\
 &= \Phi\left(\frac{0.61n - 0.6n}{\sqrt{0.24n}}\right) - \Phi\left(\frac{0.59n - 0.6n}{\sqrt{0.24n}}\right) = \\
 &= \Phi\left(\frac{0.01n}{\sqrt{0.24n}}\right) - \Phi\left(\frac{-0.01n}{\sqrt{0.24n}}\right) = \Phi\left(\frac{0.01n}{\sqrt{0.24n}}\right) - \left(1 - \Phi\left(\frac{0.01n}{\sqrt{0.24n}}\right)\right) = \\
 &= 2\Phi\left(\frac{0.01n}{\sqrt{0.24n}}\right) - 1,
 \end{aligned}$$

we can find the number of trial by solving this equation

$$2\Phi\left(\frac{0.01n}{\sqrt{0.24n}}\right) - 1 = 0.95.$$

$$2\Phi\left(\frac{0.01n}{\sqrt{0.24n}}\right) = 1.95/2 \Leftrightarrow \Phi\left(\frac{0.01n}{\sqrt{0.24n}}\right) = 0.975 \Leftrightarrow \Phi\left(\frac{0.01n}{\sqrt{0.24n}}\right) = 0.975$$

$$\Phi\left(\frac{0.01n}{\sqrt{0.24n}}\right) = \Phi(1.96) \Leftrightarrow \frac{0.01n}{\sqrt{0.24n}} = 1.96 \Leftrightarrow 0.02\sqrt{n} = 1.96 \Leftrightarrow n = 9604$$

RANDOM VARIABLES

Table of normal distribution $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{x^2}{2}} dx$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.504	0.508	0.512	0.516	0.519	0.523	0.527	0.531	0.535
0.1	0.539	0.543	0.547	0.551	0.555	0.559	0.563	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.594	0.598	0.602	0.606	0.61	0.614
0.3	0.617	0.621	0.625	0.629	0.633	0.636	0.64	0.644	0.648	0.651
0.4	0.655	0.659	0.662	0.666	0.67	0.673	0.677	0.68	0.684	0.687
0.5	0.691	0.695	0.698	0.701	0.705	0.708	0.712	0.715	0.719	0.722
0.6	0.725	0.729	0.732	0.735	0.738	0.742	0.745	0.748	0.751	0.754
0.7	0.758	0.761	0.764	0.767	0.77	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.793	0.796	0.799	0.802	0.805	0.807	0.81	0.813
0.9	0.815	0.818	0.821	0.823	0.826	0.828	0.831	0.834	0.836	0.838
1	0.841	0.843	0.846	0.848	0.85	0.853	0.855	0.857	0.859	0.862
1.1	0.864	0.866	0.868	0.87	0.872	0.874	0.877	0.879	0.881	0.883
1.2	0.884	0.886	0.888	0.89	0.892	0.894	0.896	0.898	0.899	0.901
1.3	0.903	0.904	0.906	0.908	0.909	0.911	0.913	0.914	0.916	0.917
1.4	0.919	0.92	0.922	0.923	0.925	0.926	0.927	0.929	0.93	0.931
1.5	0.933	0.934	0.935	0.937	0.938	0.939	0.94	0.941	0.942	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.95	0.951	0.952	0.953	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.959	0.96	0.961	0.962	0.963
1.8	0.964	0.964	0.965	0.966	0.967	0.967	0.968	0.969	0.969	0.97
1.9	0.971	0.971	0.972	0.973	0.973	0.974	0.975	0.975	0.976	0.976
2	0.977	0.977	0.978	0.978	0.979	0.979	0.98	0.98	0.981	0.981
2.1	0.982	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985
2.2	0.986	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989
2.3	0.989	0.989	0.989	0.99	0.99	0.99	0.99	0.991	0.991	0.991
2.4	0.991	0.992	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993
2.5	0.993	0.994	0.994	0.994	0.994	0.994	0.994	0.994	0.995	0.995
2.6	0.995	0.995	0.995	0.995	0.995	0.996	0.996	0.996	0.996	0.996
2.7	0.996	0.996	0.996	0.996	0.996	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
3	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999
3.1	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
3.2	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999

4.4.4. CHI-SQUARE DISTRIBUTION

Let's assume that the random variable X has a standard normal distribution and $Y = X^2$. This variable is called a chi-square random variable.

Moreover, if $X_1, X_2, \dots, X_n$ are random variables with standard normal distribution, the sum of their squares $C = X_1^2 + X_2^2 + \dots + X_n^2$ defines a random variable that has a chi-square distribution. This random variable depends on n parameters and therefore we say that the random variable has a distribution with n degrees of freedom. The degree of freedom is the number of linearly independent random variables $X_1, X_2, \dots, X_n$ in the random variable C and if at least one relationship between the variables is known, then the random variable C will have $n-1$ degrees of freedom.

The probability density function of a chi-square distribution X with n degrees of freedom is defined by:

$$p(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, & x > 0, \end{cases}$$

where $\Gamma\left(\frac{n}{2}\right)$ is gamma function, where gamma function is defined by $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$.

Note: $\Gamma(\alpha) = (\alpha-1)!$, $\alpha \in \mathbb{N}$.

By the definition the pdf of chi-square is defined only for positive values and for negative values is zero. For small degrees of freedom, it is asymmetric, as it increases, it becomes increasingly symmetric, and for enough large degrees of freedom ($n \geq 30$) it asymptotically approaches the normal distribution.

RANDOM VARIABLES

The mean of chi-square distribution with n degrees of freedom is n and its variance is $2n$.

Some important statistics have chi-square distribution and therefore, we need to use a table for this distribution designed as: for a given degree of freedom, usually 1,2,...,30, and so-called p value the value $\chi^2_{n,p}$ is the number such that the probability of the event $\chi^2 < \chi^2_{n,p}$ is $1-p$. One such table is given in the appendix for a maximum of 30 degrees of freedom.

df	Chi-square value									
1	0.00	0.00	0.02	0.45	1.64	2.71	3.84	5.02	6.63	10.83
2	0.05	0.10	0.21	1.39	3.22	4.61	5.99	7.38	9.21	13.82
3	0.22	0.35	0.58	2.37	4.64	6.25	7.81	9.35	11.34	16.27
4	0.48	0.71	1.06	3.36	5.99	7.78	9.49	11.14	13.28	18.47
5	0.83	1.15	1.61	4.35	7.29	9.24	11.07	12.83	15.09	20.52
6	1.24	1.64	2.20	5.35	8.56	10.64	12.59	14.45	16.81	22.46
7	1.69	2.17	2.83	6.35	9.80	12.02	14.07	16.01	18.48	24.32
8	2.18	2.73	3.49	7.34	11.03	13.36	15.51	17.53	20.09	26.12
9	2.70	3.33	4.17	8.34	12.24	14.68	16.92	19.02	21.67	27.88
10	3.25	3.94	4.87	9.34	13.44	15.99	18.31	20.48	23.21	29.59
11	3.82	4.57	5.58	10.34	14.63	17.28	19.68	21.92	24.72	31.26
12	4.40	5.23	6.30	11.34	15.81	18.55	21.03	23.34	26.22	32.91
13	5.01	5.89	7.04	12.34	16.98	19.81	22.36	24.74	27.69	34.53
14	5.63	6.57	7.79	13.34	18.15	21.06	23.68	26.12	29.14	36.12
15	6.26	7.26	8.55	14.34	19.31	22.31	25.00	27.49	30.58	37.70
16	6.91	7.96	9.31	15.34	20.47	23.54	26.30	28.85	32.00	39.25
17	7.56	8.67	10.09	16.34	21.61	24.77	27.59	30.19	33.41	40.79

RANDOM VARIABLES

18	8.23	9.39	10.86	17.34	22.76	25.99	28.87	31.53	34.81	42.31
19	8.91	10.12	11.65	18.34	23.90	27.20	30.14	32.85	36.19	43.82
20	9.59	10.85	12.44	19.34	25.04	28.41	31.41	34.17	37.57	45.31
21	10.28	11.59	13.24	20.34	26.17	29.62	32.67	35.48	38.93	46.80
22	10.98	12.34	14.04	21.34	27.30	30.81	33.92	36.78	40.29	48.27
23	11.69	13.09	14.85	22.34	28.43	32.01	35.17	38.08	41.64	49.73
24	12.40	13.85	15.66	23.34	29.55	33.20	36.42	39.36	42.98	51.18
25	13.12	14.61	16.47	24.34	30.68	34.38	37.65	40.65	44.31	52.62
26	13.84	15.38	17.29	25.34	31.79	35.56	38.89	41.92	45.64	54.05
27	14.57	16.15	18.11	26.34	32.91	36.74	40.11	43.19	46.96	55.48
28	15.31	16.93	18.94	27.34	34.03	37.92	41.34	44.46	48.28	56.89
29	16.05	17.71	19.77	28.34	35.14	39.09	42.56	45.72	49.59	58.30
30	16.79	18.49	20.60	29.34	36.25	40.26	43.77	46.98	50.89	59.70
p	0.975	0.95	0.9	0.5	0.2	0.1	0.05	0.025	0.01	0.001

4.4.5. STUDENT'S DISTRIBUTION

If a random variable Y has a standard normal distribution and C has a chi-square distribution with k degrees of freedom, then the random variable $T = \frac{Y}{\sqrt{\frac{C}{k}}}$ has so-called Student's t -distribution with k degrees of freedom.

This distribution is calculated as:

Let $X_1, X_2, \dots, X_n$ are i.i.d. with normal $N(\mu, \sigma^2)$ distribution with parameters μ and σ^2 , and $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ is the sample mean.

The unbiased estimator for the variance of the sample is the random variable defined by $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Then the random variable $C = (n-1) \frac{S^2}{\sigma^2}$ has a chi-square distribution with $k = n-1$ degrees of freedom.

Random variable $Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n}$ has standard normal distribution since $\bar{X}$ has a normal distribution with parameters μ and $\frac{\sigma^2}{n}$.

Then the random variable

$$T = \frac{Z}{\sqrt{\frac{C}{k}}} = \frac{(\bar{X} - \mu) \sqrt{n}}{S}$$

has Student's distribution defined as above.

The mean of Student's distribution with k degrees of freedom is zero and its variance for $k > 2$ degrees of freedom is $\frac{k}{k-2}$.

Student's t -distribution is a family of distributions, each of which is specified by the parameter k degrees of freedom. The density plots of the t -distributions are symmetric about zero, just like the density plot of the standard normal distribution. However, the standard deviation of each t -distribution is greater than one and depends on the value of the number of degrees of freedom k , while the standard deviation of the standard normal distribution is equal to one. As the degrees of freedom increase, the t -distribution approaches the normal distribution. Above 30 degrees of freedom, the t -distribution and the normal distribution become similar. For these reasons, for samples larger than 30, the statistic determined by the standard normal distribution is used instead of the statistic determined by the t -distribution.

A lot of statistics have the Student's distribution. Therefore, the next table for Student's t -distribution of one-tailed and two-tailed statistics for the corresponding degrees of freedom is essential.

RANDOM VARIABLES

<i>t</i> - distribution	Level of significance						
	α level of significance two-tailed						
	0.2	0.1	0.05	0.02	0.01	0.002	0.001
Degrees of freedom	α level of significance one-tailed						
	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	636.619	3.078
2	1.886	2.920	4.303	6.965	9.925	31.599	1.886
3	1.638	2.353	3.182	4.541	5.841	12.924	1.638
4	1.533	2.132	2.776	3.747	4.604	8.610	1.533
5	1.476	2.015	2.571	3.365	4.032	6.869	1.476
6	1.440	1.943	2.447	3.143	3.707	5.959	1.440
7	1.415	1.895	2.365	2.998	3.499	5.408	1.415
8	1.397	1.860	2.306	2.896	3.355	5.041	1.397
9	1.383	1.833	2.262	2.821	3.250	4.781	1.383
10	1.372	1.812	2.228	2.764	3.169	4.587	1.372
11	1.363	1.796	2.201	2.718	3.106	4.437	1.363
12	1.356	1.782	2.179	2.681	3.055	4.318	1.356
13	1.350	1.771	2.160	2.650	3.012	4.221	1.350
14	1.345	1.761	2.145	2.624	2.977	4.140	1.345
15	1.341	1.753	2.131	2.602	2.947	4.073	1.341
16	1.337	1.746	2.120	2.583	2.921	4.015	1.337
17	1.333	1.740	2.110	2.567	2.898	3.965	1.333
18	1.330	1.734	2.101	2.552	2.878	3.922	1.330
19	1.328	1.729	2.093	2.539	2.861	3.883	1.328
20	1.325	1.725	2.086	2.528	2.845	3.850	1.325
21	1.323	1.721	2.080	2.518	2.831	3.819	1.323
22	1.321	1.717	2.074	2.508	2.819	3.792	1.321
23	1.319	1.714	2.069	2.500	2.807	3.768	1.319
24	1.318	1.711	2.064	2.492	2.797	3.745	1.318
25	1.316	1.708	2.060	2.485	2.787	3.725	1.316
26	1.315	1.706	2.056	2.479	2.779	3.707	1.315
27	1.314	1.703	2.052	2.473	2.771	3.690	1.314
28	1.313	1.701	2.048	2.467	2.763	3.674	1.313
29	1.311	1.699	2.045	2.462	2.756	3.659	1.311
30	1.310	1.697	2.042	2.457	2.750	3.646	1.310

5. THE CENTRAL LIMIT THEOREM

5.1. THE LAW OF LARGE NUMBERS

This section will start with the presentation of two simple and general results, known as the Markov inequality and the Chebyshev inequality.

The Markov inequality is related to the claim made about how the mean of a distribution can be affected by moving a small amount of probability to an arbitrarily large value. The Markov inequality puts a bound on how much probability can be at arbitrarily large values once the mean is specified.

Theorem 5.1.1. (Markov Inequality) Let X is non-negative random variable. Then for any real number $t > 0$ the inequality holds:

$$P\{X \geq t\} \leq \frac{E(X)}{t}.$$

Proof. First, let's X is non-negative random variable with range R_X and probability mass function given by

$$p_n = P\{X = x_n\}, x_n \in R_X.$$

Since X is non-negative, then $x_n \geq 0$, for all n . Therefore,

$$E(X) = \sum_{x_n \in R_X} x_n p_n \geq \sum_{x_n, x_n \geq t} x_n p_n \geq \sum_{x_n, x_n \geq t} t \cdot p_n = t \cdot \sum_{x_n, x_n \geq t} p_n = t \cdot P\{X \geq t\}.$$

If X is a non-negative continuous random variable given by its probability density function $p(x)$, then:

$$E(X) = \int_0^{+\infty} xp(x)dx \geq \int_t^{+\infty} xp(x)dx \geq \int_t^{+\infty} tp(x)dx = t \int_t^{+\infty} p(x)dx = t \cdot P\{X \geq t\}$$

In any case,

$$P\{X \geq t\} \leq \frac{E(X)}{t}.$$

Theorem 5.1.2. (Chebyshev Inequality) For any random variable X with finite variance $Var(X)$ and for any real number $\varepsilon > 0$ the following inequality holds:

$$P\{|X - E(X)| \geq \varepsilon\} \leq \frac{Var(X)}{\varepsilon^2}.$$

Proof. By Markov inequality for non-negative random variable

$$\begin{aligned} P\{|X - E(X)| \geq \varepsilon\} &= P\{(X - E(X))^2 \geq \varepsilon^2\} \\ &\leq \frac{E(X - EX)^2}{\varepsilon^2} = \frac{Var(X)}{\varepsilon^2} \end{aligned}$$

Let $X_1, X_2, \dots$ is a sequence of independent random variables. We consider the sequence of arithmetic means:

$$X_1, \bar{X}_2 = \frac{X_1 + X_2}{2}, \bar{X}_3 = \frac{X_1 + X_2 + X_3}{3}, \dots, \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \dots$$

Theorem 5.1.3. Let $X_1, X_2, \dots$ is a sequence of independent and identically distributed random variables with parameters, mean μ and variance σ^2 , then

$$E(\bar{X}_n) = \mu \text{ and } Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Proof. By the properties of the mean of a random variable

$$E(\bar{X}_n) = E\left(\frac{1}{n} \sum_{i=1}^n X_n\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_n\right) = \frac{1}{n} \sum_{i=1}^n E(X_n) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \cdot n\mu = \mu$$

By the properties of the variance

$$\begin{aligned} Var(\bar{X}_n) &= Var\left(\frac{1}{n} \sum_{i=1}^n X_n\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^n X_n\right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_n) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \\ &= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Definition 5.1.4. Let $X_1, X_2, \dots$ is a sequence of random variables and L is a real number. This sequence converges to L in probability, if for any real number $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P\{|X_n - L| < \varepsilon\} = 1.$$

This property is denoted by:

$$X_n \xrightarrow{P} L, n \rightarrow \infty$$

and we read X_n converges to L in probability.

Let $X_1, X_2, \dots$ is a sequence of independent random variables, we want to check if this sequence converges to $\frac{1}{n} \sum_{i=1}^n E(X_i)$ in probability, i.e. we check if

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \frac{1}{n} \sum_{i=1}^n E(X_i).$$

If this is true, we say that for the sequence of independent random variables $X_1, X_2, \dots$ is valid the weak law of large numbers.

Theorem 5.1.5. (Law of Large Numbers) Let $X_1, X_2, \dots$ independent random variables with finite variances bounded by the real number c . Then for any real number $\varepsilon > 0$ is valid the weak law of large numbers, i.e.

$$\lim_{n \rightarrow +\infty} P\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n EX_i\right| < \varepsilon\right\} = 1$$

Proof. Let $Y_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then $E(Y_n) = \frac{1}{n} \sum_{i=1}^n E(X_i)$. Since the random variables $X_i, i=1, 2, \dots$ are independent and their variances are bounded by constant number c , then

$$\text{Var}(Y_n) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \leq \frac{1}{n^2} \sum_{i=1}^n c = \frac{1}{n^2} \cdot nc = \frac{c}{n}$$

By Theorem 5.1.2.

$$\begin{aligned} P\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i)\right| < \varepsilon\right\} &= P\{|Y_n - E(Y_n)| < \varepsilon\} = \\ &= 1 - P\{|Y_n - E(Y_n)| \geq \varepsilon\} \geq 1 - \frac{DY_n}{\varepsilon^2} \geq 1 - \frac{c}{n\varepsilon^2} \rightarrow 1, n \rightarrow \infty \end{aligned}$$

$$\text{Therefore, } \lim_{n \rightarrow +\infty} P\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i)\right| < \varepsilon\right\} = 1.$$

Theorem 5.1.6. Let $X_1, X_2, \dots$ is a sequence of independent and identically distributed random variables with parameters, mean μ and variance σ^2 . Then for this sequence the weak law of large numbers is valid, i.e.

$$\lim_{n \rightarrow +\infty} P\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| < \varepsilon\right\} = 1$$

Proof. By Theorem 5.1.3.

$$\frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu$$

So the statement follows directly by Theorem 5.1.5.

The last theorem has great importance in statistics. Namely, often when the distribution of a random variable, or some quantitative characteristic of a given population, is not known, according to the

previous theorem $\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu = E(X)$, i.e. the point estimates for the realization $x_1, x_2, \dots, x_n$ and the sample mean is estimated by $m = \frac{1}{n} \sum_{i=1}^n x_i$, where $x_1, x_2, \dots, x_n$ are the measured values of the random variable from which the sample is taken. Let's notice that as n is larger, then the point estimate is more precise.

5.2. CENTRAL LIMIT THEOREM

Let $X_1, X_2, \dots$ is a sequence of independent and identically distributed random variables with parameters, mean μ and variance σ^2 . Then by Theorem 5.1.3. the descriptive statistics for the random variable $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ are:

$$E(\bar{X}_n) = \mu \text{ and } Var(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Central Limit Theorem determines the distribution of $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, for enough large n .

Theorem 5.2.1. (Central Limit Theorem, CLT) Let $X_1, X_2, \dots$ is a sequence of independent and identically distributed random variables with parameters, mean μ and variance σ^2 . Then

$$P\left\{\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x\right\} \xrightarrow{P} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt = \Phi(x), n \rightarrow \infty$$

Proof. Let $Z_n = \sum_{i=1}^n X_i$. Then

$$\bar{Z}_n = \frac{\sum_{i=1}^n X_i - E\left(\sum_{i=1}^n X_i\right)}{\sqrt{\text{Var}\left(\sum_{i=1}^n X_i\right)}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma}.$$

let $Y_i = \frac{X_i - \mu}{\sigma}$, for $i=1, 2, \dots$. Then

$$EY_i = E\left(\frac{X_i - \mu}{\sigma}\right) = \frac{EX_i - \mu}{\sigma} = 0, \quad DY_i = D\left(\frac{X_i - \mu}{\sigma}\right) = \frac{DX_i}{\sigma^2} = 1$$

and Y_i are independent random variables as functions of X_i . Let $\varphi_{Y_i}(t) = E(e^{itY_i})$ is the characteristic function of Y_i , for any $i=1, 2, \dots, n$. By the Theorem that gives the relationship between the derivatives of the and the k -th moment $E(Y_i^k)$, $k=1, 2, 3, \dots$ of the random variable Y_i , for any $i=1, 2, \dots, n$, i.e. $\varphi_{Y_i}^{(k)}(0) = i^k EY_i^k$, we have:

$$\varphi_{Y_i}'(0) = i \cdot EY_i = 0, \quad \varphi_{Y_i}''(0) = i^2 EY_i^2 = -DY_i = -1$$

If we write $\varphi_{Y_i}(t)$ for any $i=1, 2, \dots, n$, in Maclaurin series we will have:

$$\varphi_{Y_i}(t) = \varphi_{Y_i}(0) + \varphi_{Y_i}'(0)t + \frac{\varphi_{Y_i}''(0)}{2!}t^2 + o(t^2) \text{ for all } i=1, 2, \dots, n,$$

And by substitution of the values of the previously calculated derivatives we have that

$$\varphi_{Y_i}(t) = 1 - \frac{t^2}{2} + o(t^2) \text{ for any } i=1, 2, \dots, n.$$

Now, since $\bar{Z}_n$ could be written as:

$$\bar{Z}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i = \sum_{i=1}^n \frac{Y_i}{\sqrt{n}}$$

the characteristic function of $\bar{Z}_n$ is:

$$\varphi_{\bar{Z}_n}(t) = E \left(e^{it \sum_{r=1}^n \frac{Y_r}{\sqrt{n}}} \right) = \prod_{r=1}^n e^{it \frac{Y_r}{\sqrt{n}}} = \prod_{r=1}^n \varphi_{Y_r}(t) = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) \right]^n \rightarrow e^{-\frac{t^2}{2}}$$

So, the sequence of characteristic functions $\{\varphi_{\bar{Z}_n}(t)\}$ converges to the function $\varphi_N(t) = e^{-t^2/2}$, which is the characteristic function of standard normal distribution and $\varphi_N(t)$ is continuous in $t=0$. By the inverse limit theorem, follows that the sequence $\{F_{\bar{Z}_n}(x)\}$ converges to standard normal distribution, i.e.

$$\{F_{\bar{Z}_n}(x)\} \rightarrow \Phi(x), n \rightarrow \infty$$

or

$$P\{\bar{Z}_n(x) \leq x\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt, n \rightarrow \infty$$

$$P\left\{ \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x \right\} = P\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \leq x \right\} = P\{\bar{Z}_n(x) \leq x\} \rightarrow$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt, n \rightarrow \infty,$$

i.e.

$$P\left\{ \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x \right\} \xrightarrow{P} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt = \Phi(x), n \rightarrow \infty.$$

According to the Central Limit Theorem, we conclude that the distribution of a normalized sum of independent and identically

distributed random variables with an arbitrary distribution tends to standard normal distribution, for enough large n , larger than 30.

Let's notice that if $X_1, X_2, \dots$ is a sequence of independent identically distributed random variables with a normal distribution μ and variance σ^2 , then for any n CLT is valid.

The central limit theorem has a great theoretical and practical significance, as it provides a theoretical explanation of the fact that the normal distribution is the most common distribution of real quantities, and in practice, especially in statistics, it allows the distribution of a normalized sum of random variables, even under general conditions, to be approximated by a normalized distribution.

Example 5.2.1. A bank teller serves customers standing in the queue one by one. Suppose that the service time X_i for the customer i has mean $E(X_i) = 2$ (minutes) and $Var(X_i) = 1$. We assume that service times for different bank customers are independent.

Let Y be the total time the bank teller spends serving 50 customers. Find $P(90 < Y < 110)$.

Solution. Since $Y = \sum_{i=1}^{50} X_i$ and X_i are $n = 50 > 30$ independent identically distributed variables with $\mu = 2$ and standard deviation $\sigma = 1$.

By CLT

$$\begin{aligned} P(90 < Y < 110) &= P\left(90 < \sum_{i=1}^{50} X_i < 110\right) = \\ &= P\left(\frac{90 - 50 \cdot 2}{1\sqrt{50}} < \frac{\sum_{i=1}^{50} X_i - n\mu}{\sigma\sqrt{n}} < \frac{110 - 50 \cdot 2}{1\sqrt{50}}\right) = \end{aligned}$$

$$\begin{aligned}
 &= P \left(-1.41 < \frac{\sum_{i=1}^{16} X_i - n\mu}{\sigma\sqrt{n}} < 1.41 \right) = \Phi_0(1.41) - \Phi_0(-1.41) = \\
 &= \Phi_0(1.41) - (1 - \Phi_0(1.41)) = 2\Phi_0(1.41) - 1 = 2 \cdot 0.9207 - 1 = 0.8414.
 \end{aligned}$$

Example 5.2.2. You are a director of transportation safety for the state of New Jersey. You are concerned because the average highway speed of all trucks may exceed the 62 km/h speed limit. There are 120 independent trucks and each truck has a mean speed of 60 km/h and standard deviation of 12.5 km/h, find the probability of the average of the speed greater than or equal to 62 km/h.

Define the event that the mean speed of the 120 trucks exceeds the 62 km/h speed limit. Is it possible to calculate the exact probability of the event that the mean speed of the 120 trucks exceeds the 62 km/h speed limit? Explain if “YES” or “NOT”. Find a good approximation of the probability of the event that the mean speed of the 120 trucks exceeds the 62 km/h speed limit. Justify the use of the CLT and calculate the probability.

Solution. We consider $n=120$ independent identically distributed variables with $\mu=60$ and standard deviation $\sigma=12.5$. The event that the mean speed of the 120 trucks exceeds the 62 km/h speed limit is defined by $\bar{X}_{120} > 62$. It is not possible to calculate the exact probability of the event that the mean speed of the 120 trucks exceeds the 62 km/h speed limit since the distribution of the speed of the trucks is unknown.

Since the size $n=120 > 30$ and the speed of the trucks is independently identically distributed by CLT, the approximated probability of the event that the mean speed of the 120 trucks exceeds the 62 km/h speed limit could be calculated.

$$\begin{aligned}
 P(\bar{X}_{120} > 62) &= 1 - P(\bar{X}_{120} \leq 62) = 1 - \Pr\left(\frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{62 - 60}{\frac{12.5}{\sqrt{120}}}\right) = 1 - \Phi\left(\frac{2}{\frac{12.5}{10.95}}\right) = \\
 &= 1 - \Phi(1.75) = 1 - 0.9599 = 0.0401 = 4\%
 \end{aligned}$$

And so, the probability of the average of the speed greater than or equal to 62 km/h is 0,04.

Example 5.2.3. Consider a production line of a critical mechanical element. The nominal length of the element is 30 mm. There is a standard deviation of 3 mm in the production process. The exact distribution is not known. A sample of 49 different elements is chosen randomly from a very large number of elements, as the daily quality assurance (QA) checkup set. The sample of the 49 elements is measured and if the average of the sample is deviating from the nominal value of 30 mm by more than 1 mm, then the production line is immediately stopped. We define the event of “false alarm”, that is, the production line was stopped due to “QA alarm” of failure in production quality, but the population of elements was in the correct average value of 30 mm. Find the probability of the event that the sample average of 49 elements taken from population of “good” daily production elements, is giving “false alarm”.

Solution. We consider $n = 49$ independent identically distributed random variables with mean $\mu = 30$ and standard deviation $\sigma = 3$.

We need to find the probability of the event of “false alarm”, i.e.

$$P\left(\left|\frac{\sum_{i=1}^{49} X_i}{49} - 30\right| > 1\right).$$

By the CLT

$$P\left(\left|\frac{\sum_{i=1}^{49} X_i}{49} - 30\right| > 1\right) = 1 - P\left(\left|\frac{\sum_{i=1}^{49} X_i}{49} - 30\right| \leq 1\right).$$

Since,

$$\begin{aligned} P\left(\left|\frac{\sum_{i=1}^{49} X_i}{49} - 30\right| \leq 1\right) &= P\left(-1 \leq \frac{\sum_{i=1}^{49} X_i}{49} - 30 \leq 1\right) = P\left(29 \leq \frac{\sum_{i=1}^{49} X_i}{49} \leq 31\right) = \\ &= P\left(\frac{29-30}{\frac{3}{\sqrt{49}}} \leq \frac{\frac{\sum_{i=1}^{49} X_i}{49} - \mu}{\frac{\sigma}{\sqrt{49}}} \leq \frac{31-30}{\frac{3}{\sqrt{49}}}\right) = P\left(\frac{-1}{0.43} \leq \frac{\frac{\sum_{i=1}^{49} X_i}{49} - \mu}{\frac{\sigma}{\sqrt{49}}} \leq \frac{1}{0.43}\right) = \\ &= \Phi(2.32) - \Phi(-2.32) = \Phi(2.32) - (1 - \Phi(2.32)) = 2\Phi(2.32) - 1 \\ P\left(\left|\frac{\sum_{i=1}^{49} X_i}{49} - 30\right| > 2\right) &= 1 - (2\Phi(2.32) - 1) = 2 - 2\Phi(2.32) = 2 - 2 \cdot 0.9898 = \\ &= 0.0204 \end{aligned}$$

Therefore, the probability of the event that the sample average of 49 elements taken from a population of “good” daily production elements, is giving “false alarm” is 0.0204.

Example 5.2.4. In a communication system, each data packet consists of 1000 bits. Due to the noise, each bit may be received in error with probability 0.1. It is assumed bit errors occur independently. Find the probability that there are more than 110 errors in a certain data packet.

Solution. Let us define X_i as the indicator random variable for the i -th bit in the packet. Then $X_i = 1$ if the i -th bit is received in error and $X_i = 0$ otherwise. So, $X_i, i = 1, \dots, 1000$ are independent identically distributed random variables with Bernoulli distribution with parameter $p = 0.1$. The total number of errors is the random variable $Y = \sum_{i=1}^{1000} X_i$. Since $X_i, i = 1, \dots, 1000$ have Bernoulli distribution with parameter $p = 0.1$,

$$E(X_i) = \mu = p = 0.1 \text{ and } \text{Var}(X_i) = \sigma^2 = p(1-p) = 0.09.$$

Using the CLT, we have

$$\begin{aligned} P(Y > 110) &= P\left(\sum_{i=1}^{1000} X_i > 110\right) = 1 - P\left(\sum_{i=1}^{1000} X_i \leq 110\right) = \\ &= 1 - P\left(\frac{\sum_{i=1}^{1000} X_i - n\mu}{\sigma\sqrt{n}} < \frac{110 - 1000 \cdot 0.1}{\sqrt{0.09}\sqrt{1000}}\right) = 1 - P\left(\frac{\sum_{i=1}^{1000} X_i - n\mu}{\sigma\sqrt{n}} < \frac{10}{\sqrt{90}}\right) = \\ &= 1 - \Phi_0(1.05) = 1 - 0.8531 = 0.1469. \end{aligned}$$

The probability that there are more than 110 errors in a certain data packet is 14.69%.

6. POINT ESTIMATION. CONFIDENCE INTERVAL

When there is a lack of available data to predict certain population parameters (such as the mean, variance, or standard deviation), these values are estimated using statistical methods. Because the parameter values are based on sample data, we cannot fully guarantee the precision of the resultant conclusions. Consequently, the conclusions drawn from these estimates are accepted with a confidence level that is somewhat less than 100%. Statistical estimation techniques can be categorized based on different criteria. In this chapter, we will focus on the classification that separates point estimates from confidence intervals.

6.1. POINT ESTIMATE

In the statistical estimation, the conclusion about the unknown value of the parameter of the population X is drawn based on the corresponding statistic (formula) defined on the sample $X_1, X_2, \dots, X_n$ of the population X , then, based on its realized values $x_1, x_2, \dots, x_n$, the numerical (quantitative) value of the statistic is calculated, which serves as an estimate of the population parameter. In the general case, the unknown population parameter is denoted by φ . The statistic chosen based on the sample and used to estimate the parameter is denoted by $\varphi = f(X_1, X_2, \dots, X_n)$ and it is known as the **point estimate** of the parameter, while the realized value of the estimate in the selected sample is called the **point estimated value** of the parameter and is denoted $\eta = f(x_1, x_2, \dots, x_n)$.

Consequently, for a different realization $x_1^*, x_2^*, \dots, x_n^*$, a different numerical value of the estimate is obtained $\eta^* = f(x_1^*, x_2^*, \dots, x_n^*)$.

The point estimate is a function, given by a formula, and the numerical value of that function for the given values in the sample, represents the point estimate of the parameter.

The key characteristics of a statistical estimate, commonly discussed in statistical literature, include unbiasedness, sampling variability, efficiency, and consistency.

A point estimate φ is unbiased if its mathematical expectation is equal to the population parameter φ , i.e., $E(\varphi) = \varphi$, a statistic is biased if the long-term average value of the statistic is not the parameter it is estimating. This feature is necessary, but not sufficient.

For example, the estimator may be unbiased and at the same time have a large dispersion. In that case, the estimated values in different samples significantly deviate from the expected value of the parameter, which may lead to a large random error in the conclusion. If an estimator is unbiased, then its variability will determine its reliability. If an estimator is extremely variable, then the estimates it produces may not on average be as close to the population parameter as a biased estimator with a small variance. An efficient estimate has the lowest standard error compared to all other unbiased estimators.

An estimator is considered consistent if increasing the sample size leads to an estimate with a smaller standard error. The standard deviation of an estimate is called the standard error of that estimate. The larger the standard error, the greater the error in your estimate.

The standard deviation of an estimate is a commonly used index of the error entailed in estimating a population parameter based on the information in a random sample of size n from the entire population.

In general, consistency is not a guarantee that the estimator will be particularly useful because, at least mathematically, there is no guarantee that a consistent estimator will have any desirable properties for any finite sample size. Consistency is a property that involves limits, and mathematics allows values to be arbitrarily far from the limiting value, even over time. However, in practice, this is not usually the case. In reality, consistent estimators tend to converge steadily. As more data is collected, the result becomes less biased, even with practical sample sizes.

Next, we will consider two methods for point estimation of unknown parameters, namely the maximum likelihood method and the method of moments.

Let the distribution of the random variable X is given by its probability density function $p(x; \varphi_1, \varphi_2, \dots, \varphi_k)$ or in a discrete case by its probability mass function $P(x; \varphi_1, \varphi_2, \dots, \varphi_k)$, $x = x_1, x_2, \dots$,

The joint probability density function

or joint probability mass function

of the observations in a random sample with size n , $x_1, x_2, \dots, x_n$ considered a function concerning the unknown parameters, it is called the likelihood function. The point estimators, called the maximum likelihood estimator, i.e. MLE:

are statistics that obtain the maximum value of the likelihood function, meaning they are functions that make the occurrence of the sample most probable. This means that the likelihood function reaches its maximum for the parameter values that satisfy the solutions of the following system of equations:

$$\frac{\partial L}{\partial \varphi_i} = 0, \quad i = 1, 2, \dots, k$$

Since the likelihood function has maximum extrema at the same points as log-likelihood functions

$$\ln L(x_1, x_2, \dots, x_n; \varphi_1, \varphi_2, \dots, \varphi_k) = \sum_{i=1}^n \ln p(x_i; \varphi_1, \varphi_2, \dots, \varphi_k),$$

$$\text{or } \ln L(x_1, x_2, \dots, x_n; \varphi_1, \varphi_2, \dots, \varphi_k) = \sum_{i=1}^n \ln P(x_i; \varphi_1, \varphi_2, \dots, \varphi_k)$$

and the procedure for solving the system of equations

$$\frac{\partial \ln L}{\partial \varphi_i} = 0, \quad i = 1, 2, \dots, k$$

which determines the maximum value of the likelihood function, i.e., the statistics (6.1), is simpler. Therefore, the **required point estimators** are the solutions of this system, and based on them, with the realization of the sample, we obtain the **point estimates** of the parameters in the form:

$$\begin{aligned} \eta_1 &= \varphi_1(x_1, x_2, \dots, x_n) \\ \eta_2 &= \varphi_2(x_1, x_2, \dots, x_n) \\ &\dots\dots\dots \\ \eta_k &= \varphi_k(x_1, x_2, \dots, x_n). \end{aligned}$$

Example 6.2.1. By the maximum likelihood estimation method, find the point estimators for the random variable $X : N(\mu, \sigma^2)$ with Normal distribution:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)}{2\sigma^2}}, \quad x \in \mathbb{R}$$

Then verify if these point estimators are unbiased.

Solution. Likelihood function of the Normal distribution is:

$$L(x_1, x_2, \dots, x_n, \mu, \sigma^2) = \prod_{i=1}^n p(x_i, \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i - \mu)}{2\sigma^2}}.$$

Log-likelihood function is:

$$\begin{aligned} \ln L(x_1, x_2, \dots, x_n, \mu, \sigma^2) &= \sum_{i=1}^n \ln \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i - \mu)}{2\sigma^2}} = \\ &= \sum_{i=1}^n \left(\ln 1 + \ln \left(e^{\frac{-(x_i - \mu)}{2\sigma^2}} \right) - \ln(\sqrt{2\pi}\sigma) \right) = \sum_{i=1}^n \left(\frac{-(x_i - \mu)}{2\sigma^2} - \ln(\sqrt{2\pi}) - \ln \sigma \right) = \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma - \sum_{i=1}^n \frac{(x_i - \mu)}{2\sigma^2} \end{aligned}$$

Therefore, we need to solve the following system of equations:

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln f(x_1, \dots, x_n | \mu, \sigma) &= \left(-\frac{n}{2} \ln(2\pi) \right)' - (n \ln \sigma)' - \left(\frac{\sum_{i=1}^n (x_i - \mu)}{2\sigma^2} \right)' = \\ &= 0 - 0 - \frac{\sum_{i=1}^n ((x_i - \mu))'}{2\sigma^2} = -\frac{\sum_{i=1}^n 2(x_i - \mu)((x_i - \mu))'}{2\sigma^2} = -\frac{\sum_{i=1}^n 2(x_i - \mu)(-1)}{2\sigma^2} = \\ &= \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} = 0 \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \sigma} \ln f(x_1, \dots, x_n | \mu, \sigma) &= \left(-\frac{n}{2} \ln(2\pi) \right)' - (n \ln \sigma)' - \left(\frac{\sum_{i=1}^n (x_i - \mu)}{2\sigma^2} \right)' = \\ &= 0 - \frac{n}{\sigma} - \sum_{i=1}^n \left(\frac{(x_i - \mu)}{2} \right) (\sigma^{-2})' = -\frac{n}{\sigma} - \sum_{i=1}^n \left(\frac{(x_i - \mu)}{2} \right) (-2\sigma^{-3}) = \end{aligned}$$

$$= -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^3} = 0$$

So,

$$\frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} = 0 \Rightarrow \sum_{i=1}^n (x_i - \mu) = 0 \Rightarrow \sum_{i=1}^n x_i = \sum_{i=1}^n \mu \Rightarrow \sum_{i=1}^n x_i = n\mu$$

$$\Rightarrow \tilde{\mu} = \frac{\sum_{i=1}^n x_i}{n}$$

$$-\frac{n}{\sigma} + \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^3} = 0 \Rightarrow -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^3} = 0 \Rightarrow \frac{n}{\sigma} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^3}$$

$$\Rightarrow n\sigma^3 = \sigma \sum_{i=1}^n (x_i - \mu)^2 \Rightarrow n\sigma^3 - \sigma \sum_{i=1}^n (x_i - \mu)^2 = 0$$

$$\Rightarrow \sigma \left(n\sigma^2 - \sum_{i=1}^n (x_i - \mu)^2 \right) = 0 \Rightarrow n\sigma^2 - \sum_{i=1}^n (x_i - \mu)^2 = 0 \Rightarrow \tilde{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

We conclude that the point estimators for the Normal distribution or MLE, known as the sample mean and the sample variance, respectively, are:

$$\tilde{\mu} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}_n$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \tilde{\mu})^2}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}$$

Additionally, the random variables X_i for all $i=1, \dots, n$ have identical normal distribution with parameters $E(X_i) = \mu, Var(X_i) = \sigma^2, i=1, \dots, n$.

Let's notice that the following identity is valid

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu)^2$$

Now let verify that the point estimators from above are unbiased:

$$\text{Since, } E(\tilde{\mu}) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{n\mu}{n} = \mu, \text{ the point}$$

estimator for the mean is unbiased.

But,

$$\begin{aligned} E(\tilde{\sigma}^2) &= E\left(\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}\right) = E\left(\frac{\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X}_n - \mu)^2}{n}\right) = \\ &= \frac{\sum_{i=1}^n E((X_i - \mu)^2) - E(n(\bar{X}_n - \mu)^2)}{n} = \frac{\sum_{i=1}^n V(X_i) - nE(\bar{X}_n - \mu)^2}{n} = \\ &= \frac{n\sigma^2}{n} - V(\bar{X}_n) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \end{aligned}$$

Therefore, this statistic for the variance is biased and we notice that by a small correction we could find an unbiased statistic for the variance given by:

$$\tilde{\sigma}^2 = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^n (X_i - \tilde{\mu})^2}{n} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}.$$

Example 6.2.2. Suppose that X is a discrete random variable with the following probability mass function:

X	0	1	2	3
P(X)	$\frac{2\theta}{3}$	$\frac{\theta}{3}$	$\frac{2(1-\theta)}{3}$	$\frac{(1-\theta)}{3}$

where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations were taken from this distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of the parameter?

Solution. Since the sample is (3,0,2,1,3,2,1,0,2,1), the likelihood function is

$$L(\theta) = \prod_{i=1}^n P(X_i, \theta)$$

Since $n=10$ for the given obseravtions, the likelihood function is

$$\begin{aligned} L(\theta) &= P(X=3) \cdot P(X=0) \cdot P(X=2) \cdot P(X=1) \cdot P(X=3) \cdot \\ &\quad \cdot P(X=2) P(X=1) P(X=0) P(X=2) P(X=1) = \\ &= (P(X=0))^2 (P(X=1))^3 (P(X=2))^3 (P(X=3))^2 = \\ &= \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2. \end{aligned}$$

Let us look at the log-likelihood function:

$$\begin{aligned} \ln L(\theta) &= \ln \left(\left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2 \right) = \\ &= \ln \left(\frac{2\theta}{3}\right)^2 + \ln \left(\frac{\theta}{3}\right)^3 + \ln \left(\frac{2(1-\theta)}{3}\right)^3 + \ln \left(\frac{1-\theta}{3}\right)^2 = \\ &= 2 \ln \left(\frac{2\theta}{3}\right) + 3 \ln \left(\frac{\theta}{3}\right) + 3 \ln \left(\frac{2(1-\theta)}{3}\right) + 2 \ln \left(\frac{1-\theta}{3}\right) = \\ &= 2 \ln \frac{2}{3} + 2 \ln \theta + 3 \ln \frac{1}{3} + 3 \ln \theta + 3 \ln \frac{2}{3} + 3 \ln(1-\theta) + 2 \ln \frac{1}{3} + 2 \ln(1-\theta). \end{aligned}$$

Let the derivative of log-likelihood function with respect to the parameter θ be zero:

$$\begin{aligned}
 \frac{d \ln L(\theta)}{d\theta} &= \\
 &= \left(2 \ln \frac{2}{3} + 2 \ln \theta + 3 \ln \frac{1}{3} + 3 \ln \theta + 3 \ln \frac{2}{3} + 3 \ln (1-\theta) + 2 \ln \frac{1}{3} + 2 \ln (1-\theta) \right)'_{\theta} = \\
 &= \left(2 \ln \frac{2}{3} \right)'_{\theta} + 2 (\ln \theta)'_{\theta} + \left(3 \ln \frac{1}{3} \right)'_{\theta} + 3 (\ln \theta)'_{\theta} + \left(3 \ln \frac{2}{3} \right)'_{\theta} + 3 (\ln (1-\theta))'_{\theta} + \\
 &\quad + \left(2 \ln \frac{1}{3} \right)'_{\theta} + 2 (\ln (1-\theta))'_{\theta} = 0 + \frac{2}{\theta} + 0 + \frac{3}{\theta} + 0 - \frac{3}{1-\theta} + 0 - \frac{3}{1-\theta} = \\
 &= \frac{5}{\theta} - \frac{6}{1-\theta} = \frac{5(1-\theta) - 6\theta}{\theta(1-\theta)} = \frac{5-11\theta}{\theta(1-\theta)} = 0.
 \end{aligned}$$

Therefore, $5-11\theta=0$, then $11\theta=5$, so the maximum point estimates for the parameter of the given distribution is $\theta = 0.5$.

Example 6.2.3. Find Maximum Likelihood Estimator (MLE) for Poisson distributed data with parameter λ :

$$P(x, \lambda) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \dots,$$

Solution. Let there be a Poisson distributed data set $x_1, x_2, \dots, x_n$. Then,

$$P(x_i, \lambda) = P(X_i = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \dots \text{ for all } i = 1, 2, \dots, n.$$

The likelihood function is defined by:

$$L(x_1, x_2, \dots, x_n, \lambda) = \prod_{i=1}^n P(X_i, \lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}.$$

Log-likelihood function is:

$$\begin{aligned}\ln L(x_1, x_2, \dots, x_n, \lambda) &= \sum_{i=1}^n \ln P(X_i, \lambda) = \sum_{i=1}^n \ln \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \\ &= \sum_{i=1}^n (x_i \ln \lambda - \ln(x_i!) - \lambda)\end{aligned}$$

Therefore, the critical point is obtained by solving $\frac{d(\ln L)}{d\lambda} = 0$, i.e.

$$\begin{aligned}\frac{d \ln L(x_1, x_2, \dots, x_n, \lambda)}{d\lambda} &= \frac{d}{d\lambda} \left(\sum_{i=1}^n (x_i \ln \lambda - \ln(x_i!) - \lambda) \right) = \\ &= \sum_{i=1}^n \left(\frac{d}{d\lambda} (x_i \ln \lambda) - \frac{d}{d\lambda} (\ln(x_i!)) - \frac{d}{d\lambda} (\lambda) \right) = \sum_{i=1}^n \left(\frac{x_i}{\lambda} - 1 \right) = \\ &= \frac{1}{\lambda} \sum_{i=1}^n x_i - n = 0.\end{aligned}$$

The candidate for MLE of the parameter λ is:

$$\lambda = \frac{1}{n} \sum_{i=1}^n x_i,$$

Check if the second derivative is negative. Since,

$$\begin{aligned}\frac{d^2 \ln L(x_1, x_2, \dots, x_n, \lambda)}{d\lambda^2} &= \frac{d}{d\lambda} \left(\frac{d \ln L(x_1, x_2, \dots, x_n, \lambda)}{d\lambda} \right) = \frac{d}{d\lambda} \left(\frac{1}{\lambda} \sum_{i=1}^n x_i - n \right) = \\ &= -\frac{1}{\lambda^2} \sum_{i=1}^n x_i,\end{aligned}$$

so

$$\frac{d^2 \ln L(x_1, x_2, \dots, x_n, \lambda)}{d\lambda^2} = -\frac{1}{\left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2} \sum_{i=1}^n x_i = -\frac{n^2}{\sum_{i=1}^n x_i} < 0$$

because $\sum_{i=1}^n x_i \geq 0$, unless $\sum_{i=1}^n x_i = 0$. So, it's a maximum unless $\sum_{i=1}^n x_i = 0$.

Finally, we conclude that the MLE for Poisson-distributed data with parameter λ is:

$$\lambda = \frac{1}{n} \sum_{i=1}^n x_i .$$

Example 6.2.4. The number of phone calls received by an office during the noon hours on ten randomly selected days in 2023 and 2024 is:

4, 0, 5, 7, 4, 2, 5, 5, 3, 5.

Estimate the percentage of days in which the number of phone calls received by an office during the noon hours in 2023 and 2024 is less than 3.

Solution. First, we need to determine which distribution is appropriate for the number of phone calls received by an office during the noon hours. Because the probability of an office receiving a phone call in the noon hour is relatively low, and the number of days being considered is sufficiently large, the Poisson distribution is appropriate for modeling the number of phone calls an office would receive during the noon hour. Since, by Example 6.2.3., the MLE for the parameter of a Poisson distributed data with parameter is $\lambda = \frac{1}{n} \sum_{i=1}^n x_i$, then the point estimate is:

$$\lambda = \frac{4+0+5+7+4+2+5+5+3+5}{10} = 4 .$$

The percentage of days in which the number of phone calls an office received during the noon hours in 2023 and 2024 is less than 3 is

$$P(X < 3) = \sum_{x=0}^2 \frac{4^x}{x!} e^{-4} = 0,018 + 4 \cdot 0,018 + 8 \cdot 0,018 = 0.234 = 23.4\% .$$

Example 6.2.5. Find the Maximum Likelihood Estimator (MLE) for n independent Bernoulli distributed data with a probability of success parameter p .

Solution. Let there be n independent Bernoulli distributed data set with a probability of success parameter p , meaning let's consider the sample $X_1, X_2, \dots, X_n$ where X_i for all $i=1, \dots, n$ has a Bernoulli distribution with parameter p , i.e.

$$P(X_i = x) = p^x (1-p)^{1-x}, x = 0, 1.$$

Then, the maximum likelihood function for this distribution is:

$$L(x_1, x_2, \dots, x_n, p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}.$$

Log-likelihood function is:

$$\begin{aligned} \ln L(x_1, x_2, \dots, x_n, p) &= \ln \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = \sum_{i=1}^n \ln p^{x_i} (1-p)^{1-x_i} = \\ &= \sum_{i=1}^n (x_i \ln p + (1-x_i) \ln (1-p)) = \ln p \sum_{i=1}^n x_i + \left(n - \sum_{i=1}^n x_i \right) \ln (1-p). \end{aligned}$$

Therefore, the critical point is obtained by solving $\frac{d(\ln L)}{d\lambda} = 0$, i.e.

$$\begin{aligned} \frac{d}{dp} \ln L(x_1, \dots, x_n, p) &= \frac{d}{dp} \left(\ln p \sum_{i=1}^n x_i + \left(n - \sum_{i=1}^n x_i \right) \ln (1-p) \right) = \\ &= \frac{d}{dp} \left(\ln p \sum_{i=1}^n x_i \right) + \frac{d}{dp} \left(\left(n - \sum_{i=1}^n x_i \right) \ln (1-p) \right) = \\ &= \frac{1}{p} \cdot \sum_{i=1}^n x_i + \frac{1}{1-p} \cdot \frac{d}{dp} (1-p) \cdot \left(n - \sum_{i=1}^n x_i \right) = \\ &= \frac{1}{p} \cdot \sum_{i=1}^n x_i - \frac{1}{1-p} \cdot \left(n - \sum_{i=1}^n x_i \right) = 0 \end{aligned}$$

By solving the equation from above:

$$\frac{\sum_{i=1}^n X_i}{p} = \frac{\left(n - \sum_{i=1}^n X_i\right)}{1-p} \Rightarrow (1-p) \sum_{i=1}^n X_i = \left(n - \sum_{i=1}^n X_i\right) p$$

$$\sum_{i=1}^n X_i - p \sum_{i=1}^n X_i = np - p \sum_{i=1}^n X_i \Rightarrow np = \sum_{i=1}^n X_i$$

the MLE for the parameter of the parameter p of Bernoulli distribution is:

$$\tilde{p} = \frac{\sum_{i=1}^n X_i}{n}.$$

Example 6.2.6. A sample of 1000 pair of shoes is taken, indpenedently produced, with probability p of defective pair. By testing the sample 33 pairs of shoes are defective. Estimate the probability p .

Solution. First, let notice that this distribution of defective pair is Bernoulli with parameter p . By Example 6.2.5. the MLE for the

parameter p is $\tilde{p} = \frac{\sum_{i=1}^n X_i}{n}$.

Therefore, the point estimates of the parameter p is

$$p = \frac{33}{1000} = 0.033.$$

So, the probability of a defective pair during production is 3.3%.

Namely, if X is a random variable, then the mathematical

If $x_1, x_2, \dots, x_n$ is a sample taken from the random variable X ,

Let the distribution of the random variable X is given by its

In order to find the point estimates $\varphi_1, \varphi_2, \dots, \varphi_k$ based on a sample

$$\eta_1 = \eta_1(X_1, X_2, \dots, X_n)$$

$$\eta_2 = \eta_2(X_1, X_2, \dots, X_n)$$

.....

$$\eta_k = \eta_k(X_1, X_2, \dots, X_n)$$

132

$$\begin{aligned}\eta_1 &= \varphi_1(x_1, x_2, \dots, x_n) \\ \eta_2 &= \varphi_2(x_1, x_2, \dots, x_n) \\ &\dots\dots\dots \\ \eta_k &= \varphi_k(x_1, x_2, \dots, x_n).\end{aligned}$$

are point estimates of the parameters $\varphi_1, \varphi_2, \dots, \varphi_k$.

The point estimators by the method of moments are the solutions of the following system of equations:

$$\begin{aligned}f_1(X; \varphi_1, \varphi_2, \dots, \varphi_k) &= g_1(X_1, X_2, \dots, X_n) \\ f_2(X; \varphi_1, \varphi_2, \dots, \varphi_k) &= g_2(X_1, X_2, \dots, X_n) \\ &\dots\dots\dots \\ f_k(X; \varphi_1, \varphi_2, \dots, \varphi_k) &= g_k(X_1, X_2, \dots, X_n)\end{aligned}$$

where $f_i, i=1, 2, \dots, k$ are the k -th moments of the random variable X , and $g_i, i=1, 2, \dots, k$ are the k -th sample moments.

Example 6.3.1. Find the point estimator of the parameter β of Pareto random variables X with probability density function

$$p(x; \beta) = \frac{\beta}{x^{\beta+1}}, \quad x > 1, \beta > 0.$$

by the method of moments.

Solution. Since we need to obtain the point estimator for only one parameter, we need to solve the equation:

$$\mu = \bar{x}.$$

The mean of the Pareto distribution is

$$\mu = \int_1^{+\infty} x \cdot \frac{\beta}{x^{\beta+1}} dx = \int_1^{+\infty} \frac{\beta}{x^{\beta}} dx = \beta \int_1^{+\infty} x^{-\beta} dx = \beta \cdot \left. \frac{x^{-\beta+1}}{-\beta+1} \right|_1^{+\infty} =$$

$$= \beta \cdot \left(\lim_{x \rightarrow \infty} \frac{x^{-\beta+1}}{-\beta+1} - \frac{1^{-\beta+1}}{-\beta+1} \right) = \frac{\beta}{\beta-1}$$

Therefore,

$$\frac{\beta}{\beta-1} = \bar{x} \Leftrightarrow \beta = (\beta-1)\bar{x} \Leftrightarrow \beta = \beta\bar{x} - \bar{x} \Leftrightarrow \beta\bar{x} - \beta = \bar{x} \Leftrightarrow \beta(\bar{x}-1) = \bar{x}$$

So, the point estimator of the Pareto distribution parameter is

$$\tilde{\beta} = \frac{\bar{x}}{\bar{x}-1}.$$

6.4. CONFIDENCE INTERVAL OF THE MEAN OF THE POPULATION

In addition to a point estimate, it is important to evaluate the extent of the sampling error, which represents the difference between the point estimate and the population parameter. By using the standard error of the sample statistic and its distribution, an interval estimate for the population parameter can be formed. A confidence interval is a particular kind of interval estimate that encompasses a set of values derived from the point estimate, along with a specified percentage that represents the likelihood that the actual parameter value resides within this range. This percentage is referred to as the confidence level β .

We will consider the confidence interval of the mean of the population with different constraints.

6.4.1. CONFIDENCE INTERVAL OF THE MEAN OF THE POPULATION μ WHILE THE VARIANCE OF THE POPULATION σ^2 IS KNOWN

Let us assume that we need to estimate the expected value of a population that follows a normal probability distribution $X : N(\mu, \sigma^2)$, where its variance is known σ^2 .

Since the population has a normal distribution, the sample mean $\bar{X}$ of a sample of size n taken from that population also follows a normal distribution with parameters μ and $\frac{\sigma^2}{n}$, i.e., $\bar{X} : N\left(\mu, \frac{\sigma^2}{n}\right)$.

Therefore, the statistic $Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n}$ follows a standard normal distribution.

If the confidence level is β , then using the normal distribution table, we find a real number z_β such that:

$$P\left(-z_\beta \leq \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \leq z_\beta\right) = \beta,$$

or

$$\begin{aligned} P\left(-z_\beta \leq \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \leq z_\beta\right) &= \Phi_0(z_\beta) - \Phi_0(-z_\beta) = \Phi_0(z_\beta) - (1 - \Phi_0(z_\beta)) = \\ &= 2\Phi_0(z_\beta) - 1 = \beta \end{aligned}$$

$$\text{so, } \Phi_0(z_\beta) = \frac{1 + \beta}{2}.$$

By solving the inequality within the parentheses for $\bar{X}$, we obtain the following equation:

$$P\left(\mu - z_\beta \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_\beta \frac{\sigma}{\sqrt{n}}\right) = \beta.$$

If, for example, the confidence level is $\beta = 0,95$, so

$$\Phi_0(z_\beta) = \frac{1 + 0,95}{2} = 0,975 \text{ and then } z_\beta = 1,96. \text{ Therefore, the equation}$$

$$P\left(\mu - 1,96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

is interpreted as: “in 95% of random samples of size n , their mean отстапува од математ deviates from the population mean μ by at most 1.96 standard errors $\frac{\sigma}{\sqrt{n}}$ ”.

Let us select any of these 95% samples, compute its arithmetic mean, and determine the following interval $\left[\bar{x} - 1,96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1,96 \frac{\sigma}{\sqrt{n}} \right]$ around it. Since $\bar{x}$ deviates from μ by less than $1,96 \frac{\sigma}{\sqrt{n}}$, the computed interval will contain the value of the parameter μ . This means that 95% of such constructed intervals will include the value μ . The arithmetic means $\bar{x}$ in the remaining 5% of the samples will differ from μ by more than $1,96 \frac{\sigma}{\sqrt{n}}$, so the intervals around these values would not contain the parameter μ . This conclusion is more precisely formulated by the following equation:

$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}} \right) = 0,95,$$

and interprets as the probability that the interval around the sample mean will contain the value of the parameter μ is 0.95.

Some other interpretations:

There is a 95% probability that the 95% confidence interval calculated from a given future sample will cover the true value of the population parameter, or

The 95% confidence interval represents values that are not statistically significantly different from the point estimate at the 0.05 level.

This result is used in estimating the unknown mean of the population. The estimation is conducted using a random sample and its arithmetic mean $\bar{x}$, around which the following interval is formed:

$$\left[\bar{x} - z_{\beta} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\beta} \frac{\sigma}{\sqrt{n}} \right],$$

called the **confidence interval of the mean of the population with normal distribution**, while the variance of the population is

known if the sample is taken from an infinitely large population or a finite population of size N with replacement or from a finite population of size N without replacement but the sample size is less than 5% of the population size is the interval.

If the sample is taken from a finite population of size N without replacement, but the sample size is greater than 5% of the population size, the confidence interval is :

$$\left[\bar{x} - z_{1-\alpha} \sigma \sqrt{\frac{N-n}{n(N-1)}}, \bar{x} + z_{1-\alpha} \sigma \sqrt{\frac{N-n}{n(N-1)}} \right].$$

Example 6.4.1. A sample of size $n=25$ is taken from a population with a distribution $N(\mu, 100)$ and its sample mean is $\bar{x}=100$. If the confidence level is $\beta=0,95$, then the confidence interval for the population mean is:

$$\left[100 - 1,96 \frac{10}{\sqrt{25}}, 100 + 1,96 \frac{10}{\sqrt{25}} \right], \text{ i.e., } [96,08; 103,92].$$

There is a 95% probability that the $[96,08; 103,92]$ will cover the true value of the population parameter.

The risk that we made a statistical mistake is called the **significance level** and it is denoted by α , and with the confidence level are added up to 1, so $\alpha=1-\beta$. Significance level is a threshold that determines whether a study result can be considered statistically significant.

How to choose the confidence level?

The mathematical expectation can be evaluated with different confidence levels, not just 0.95. Since the endpoints of the confidence interval depend on the confidence level, changing its value also changes the width of the interval. Specifically, as the confidence level increases,

the confidence interval widens. Thus, if we want to conclude with 100% certainty, the formed interval must be very wide, which means the conclusion becomes less precise. In other words, as the confidence level increases, the precision of the estimate decreases. Sometimes, the interval becomes so wide that the result loses its informativeness and practical usefulness. Therefore, in practice, a 95% confidence level is most commonly used, as it simultaneously provides high confidence and a relatively precise estimate.

Relationship between confidence level and the sample size

Let us assume that in Example 6.4.1. instead of a sample of size $n=25$ a sample of size $n=10000$ is considered, and the arithmetic mean of this sample is equal to the arithmetic mean of the previous one, i.e. $\bar{x}=100$. Then for the same confidence level $\beta=0,95$, the confidence interval is:

$$\left[100 - 1,96 \frac{10}{\sqrt{10000}}; 100 + 1,96 \frac{10}{\sqrt{10000}} \right] = [99,804; 100,196].$$

If we compare it with the previously calculated interval, we observe that for $n=25$ the width of the 95% confidence interval is $103.92 - 116.08 = 7.84$, while for $n=10000$ the width of the 95% confidence interval is $100,196 - 99,804 = 0,392$.

This means that at a given confidence level, we can reduce the confidence interval if we increase the sample size. In other words, for a given confidence level, the precision of the conclusion increases with an increase in the sample size.

Often, both the confidence level and precision are predetermined. Therefore, it is necessary to determine the sample size that satisfies both values. Since increasing the sample size also increases the costs of the research, the smallest sample size that ensures both the predetermined precision and confidence of the estimate can be calculated mathematically.

If a population with known variance σ^2 does not have a normal probability distribution, or the shape of the distribution is unknown, the Central Limit Theorem is used to estimate the population mean. A larger

sample size is selected from the population, and then the statistics $\bar{X}$ has a normal distribution with parameters μ and $\frac{\sigma^2}{n}$. The interval estimation is then identical to the interval estimation of the mean for a population with a normal distribution.

Note: This estimate is only appropriate when the sample size is $n \geq 30$. In the case where the population is not normally distributed and the sample size is less than 30, there is no statistic for calculating a confidence interval.

Example 6.4.2. Assume that we have enough reasons to believe the height of the students in Macedonia has a normal distribution with standard deviation $\sigma = 12cm$. The mean height of 225 randomly selected students is $\bar{x} = 177cm$. Find the confidence interval for the mean height with a confidence level $\beta = 0.99$.

Solution. We are considering a sample of size $n = 225$ that is taken from normally distributed population with unknown size (infinitely large population), standard deviation $\sigma = 12cm$, sample mean $\bar{x} = 177cm$ and confidence level $\beta = 0.99$. Therefore the confidence interval for the mean height of the students is the interval:

$$\left[\bar{x} - z_{\beta} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\beta} \frac{\sigma}{\sqrt{n}} \right].$$

From the table of standard normal distribution we have

$$\Phi_0(2.58) = 0.995 = \frac{1 + 0.99}{2}$$

So, the confidence interval of the population mean is:

$$\left[177 - 2.58 \frac{12}{\sqrt{225}}, 177 + 2.58 \frac{12}{\sqrt{225}} \right] \text{ or } [174.936, 179.064].$$

Example 6.4.3. Data processing time on the orders of the clients through the bank terminal has a normal distribution $N(\mu, 9)$. Of 960

processed orders during one day, a systematic sample is taken up by selecting every fifteenth order and the data set of the time lengths for the processing of the orders expressed in minutes is given in the table:

16.10	11.85	9.37	7.26	10.12	12.53	9.93	12.79	9.23	10.96
12.69	7.91	2.00	9.87	11.88	8.07	11.81	10.86	12.93	12.54
12.70	7.18	9.57	9.88	16.20	4.90	7.44	8.05	8.10	9.93
5.60	7.80	2.00	10.36	12.25	11.89	11.09	2.50	11.72	9.64
4.00	9.72	15.50	9.88	13.55	10.67	10.84	13.15	8.44	8.65
9.70	10.42	10.98	11.09	8.03	10.13	5.30	12.81	8.90	9.08
11.04	6.26	9.54	4.20						

Find the confidence interval for the data processing time on the orders of the clients with a confidence level 0.95 .

Solution. We are considering a random sample of size $n=64$ from normally distributed population with size $N=960$, standard deviation $\sigma=\sqrt{9}=3$, confidence coefficient $1-\alpha=0.95$.

Since the sample is without replacement and its size is $\frac{64}{960}=6.7\%$ greater than 5% of the size of the population the confidence interval for the mean of the population is the interval

$$\left[\bar{x} - z_{1-\alpha} \sigma \sqrt{\frac{N-n}{n(N-1)}}, \bar{x} + z_{1-\alpha} \sigma \sqrt{\frac{N-n}{n(N-1)}} \right]$$

Since $\bar{x} = \frac{1}{64} \sum_{i=1}^{64} x_i = 9.71$ and from the table of standard normal distribution, we have

$$\Phi_0(1.96) = 0.975 = \frac{1 + 0.95}{2}$$

therefore $z_{0.95} = 1.96$.

So, the confidence interval of the population mean is:

$$\left[9.71 - 1.96 \cdot 3 \sqrt{\frac{960 - 64}{64(960 - 1)}}, 9.71 + 1.96 \cdot 3 \sqrt{\frac{960 - 64}{64(960 - 1)}} \right] \text{ or } [9, 10.42].$$

6.4.2. CONFIDENCE INTERVAL OF THE MEAN μ OF THE NORMALLY DISTRIBUTED POPULATION WHILE THE VARIANCE OF THE POPULATION σ^2 IS UNKNOWN

The assumption that when estimating a parameter μ we know the value of the variance σ^2 is usually not met. In practice, it is rare that the population mean is not known, and at the same time its variance is known.

If the variance σ^2 is unknown, then we cannot calculate the standard error $\frac{\sigma}{\sqrt{n}}$, and without it the confidence interval of the population mean cannot be determined. Therefore, it is first necessary to estimate the unknown value σ . When in the formula for the standard error $\frac{\sigma}{\sqrt{n}}$, the population standard deviation will be replaced by its point estimator, that is, by the unbiased statistic

$$S = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)},$$

then the point estimator of the standard error is:

$$\frac{S}{\sqrt{n}}.$$

The realized value of this estimate with the values from the sample under consideration, i.e. the point estimate of the standard error is $\frac{s}{\sqrt{n}}$, where s is the sample standard deviation, i.e.

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

In the case when the dispersion of the population is known, the statistic $Z = \frac{\bar{X} - \mu}{\sigma} \sqrt{n}$ has a $N(0,1)$ distribution, but when the standard deviation of the population σ is replaced by the sample standard deviation, then the following statistic $t = \frac{\bar{X} - \mu}{s} \sqrt{n}$ is obtained and it has Student's distribution or t -distribution.

So when determining the confidence interval of the population mean if its variance is not known, for a sample size n and confidence level β , the statistic $t = \frac{\bar{X} - \mu}{s} \sqrt{n}$ that has t -distribution with $n-1$ degrees of freedom is considered.

For the given confidence level β , the significance level $\alpha = 1 - \beta$ and for this α we obtain a number $t_\alpha(n-1)$, such that:

$$P\left(\left|\frac{\bar{X} - \mu}{s} \sqrt{n}\right| > t_\alpha(n-1)\right) = \alpha, \text{ i.e.}$$

$$P\left(\left|\frac{\bar{X} - \mu}{s} \sqrt{n}\right| \leq t_{1-\beta}(n-1)\right) = 1 - \alpha = \beta.$$

$$\text{Therefore, } P\left(-t_{1-\beta}(n-1) \leq \frac{\bar{X} - \mu}{s} \sqrt{n} \leq t_{1-\beta}(n-1)\right) = \beta.$$

Solving the inequality in the previous equation by μ , where the value of the point estimate $\bar{x}$ of the sample mean statistic $\bar{X}$ is taken, the confidence interval of the mean of the population is obtained in the following form:

$$\left[\bar{x} - \frac{s}{\sqrt{n}} t_{1-\beta}(n-1), \bar{x} + \frac{s}{\sqrt{n}} t_{1-\beta}(n-1) \right],$$

where the number $t_{1-\beta}(n-1)$ is read from the table of Student's distribution two-tailed test.

Note: The confidence interval for the population mean μ when the population variance is unknown, and the population has a normal distribution, is calculated with the previous expression when the sample size is $n < 30$. In the case when the sample size is greater than or equal to 30, since with the increase in the degrees of freedom of the t -distribution, it approaches the normal distribution, then the $N(0,1)$ distribution corresponds to the statistic $\frac{\bar{X} - m}{s} \sqrt{n}$.

Therefore, the confidence interval for the population mean while the population variance is unknown, and the sample size is $n \geq 30$, more precisely is calculated with the expression:

$$\left[\bar{x} - z_{\beta} \frac{s}{\sqrt{n}}, \bar{x} + z_{\beta} \frac{s}{\sqrt{n}} \right],$$

where the real value z_{β} is the number read from the z -distribution table such that $\Phi_0(z_{\beta}) = \frac{1+\beta}{2} = \frac{1+1-\alpha}{2} = 1 - \frac{\alpha}{2}$.

Example 6.4.4. At one gas station in July 2024, 500 customers purchased petrol for their vehicles. Data on sold quantity of petrol per vehicle for a random sample without replacement is given in the following table:

35	14	7	33	9	5	15	23	17	14
38	37	19	7	27	23	34	28	25	11
24	24	36	17	15	18	16	13	35	

Let's assume that the sale is a normally distributed random variable. Find the confidence interval for the average petrol sale per a car, for the confidence level $1 - \alpha = 0.95$.

Solution. For a random sample of size $n = 29 < 30$ that is taken from normally distributed population with unknown standard deviation, the 95% confidence interval for the population mean is the interval

$$\left[\bar{x} - \frac{s}{\sqrt{n}} t_{1-\beta}(n-1), \bar{x} + \frac{s}{\sqrt{n}} t_{1-\beta}(n-1) \right]$$

where $t_{1-\beta}(n-1)$ is the value from the table for the two tailed t -distribution for $1 - \beta = 0.05$ and $n - 1 = 28$ degrees of freedom, i.e. $t_{0.05}(28) = 1.699$.

Since, the sample mean is $\bar{x} = \frac{1}{29} \sum_{i=1}^{29} x_i = \frac{619}{29} = 21.34$, and the sample variance

$$s^2 = \frac{1}{28} \left(\sum_{i=1}^{29} x_i^2 - n\bar{x}^2 \right) = \frac{2800.528}{28} = 100.0188$$

sample standard deviation $s = \sqrt{100.0188} = 10$, so the confidence interval for the population mean is the interval

$$\left[21.34 - \frac{10}{\sqrt{29}} \cdot 1.699, 21.34 + \frac{10}{\sqrt{29}} \cdot 1.699 \right], \text{ or } [18.19, 24.49].$$

Example 6.4.5. Let's consider a sample with size N_1 and 95% confidence interval of the population mean with normal distribution and known variance has width w . What should be the size N_2 of another sample taken from the same population to increase the confidence level to 99%, but the width of this 99% confidence interval is same, w ?

Solution. Since $z_{0.95} = 1.96$, the 95% confidence interval:

$$\left(\bar{x}_{N_1} - \frac{\sigma}{\sqrt{N_1}} z_{0.95}, \bar{x}_{N_1} + \frac{\sigma}{\sqrt{N_1}} z_{0.95} \right) = \left(\bar{x}_{N_1} - 1.96 \frac{\sigma}{\sqrt{N_1}}, \bar{x}_{N_1} + 1.96 \frac{\sigma}{\sqrt{N_1}} \right)$$

Therefore, its width is

$$w = \bar{x}_{N_1} + 1.96 \frac{\sigma}{\sqrt{N_1}} - \left(\bar{x}_{N_1} - 1.96 \frac{\sigma}{\sqrt{N_1}} \right) = 2 \cdot 1.96 \frac{\sigma}{\sqrt{N_1}} = 3.92 \frac{\sigma}{\sqrt{N_1}}$$

Since

$$\Phi(2.57) = 0.9949 < \Phi(z_{0.99}) = \frac{1+0.99}{2} = 0.995 < 0.9951 = \Phi(2.58),$$

so $\Phi(z_{0.99}) = \frac{2.57+2.58}{2} = 2.575$ and the 99% confidence interval is:

$$\left(\bar{x}_{N_2} - z_{0.99} \frac{\sigma}{\sqrt{N_2}}, \bar{x}_{N_2} + z_{0.99} \frac{\sigma}{\sqrt{N_2}} \right) = \left(\bar{x}_{N_2} - 2.575 \cdot \frac{\sigma}{\sqrt{N_2}}, \bar{x}_{N_2} + 2.575 \cdot \frac{\sigma}{\sqrt{N_2}} \right)$$

$$\text{and its width } w = 2 \cdot 2.575 \frac{\sigma}{\sqrt{N_2}} = 5.15 \frac{\sigma}{\sqrt{N_2}}$$

So,

$$3.92 \cdot \frac{\sigma}{\sqrt{N_1}} = 5.15 \frac{\sigma}{\sqrt{N_2}} / : \sigma \Rightarrow$$

$$3.92 \cdot \frac{1}{\sqrt{N_1}} = 5.15 \frac{1}{\sqrt{N_2}} \Rightarrow \sqrt{\frac{N_2}{N_1}} = \frac{5.15}{3.92} \Rightarrow \frac{N_2}{N_1} = \left(\frac{5.15}{3.92} \right)^2 = 1.71$$

This means that the sample size N_2 should be 1.71 times greater than the sample size N_1 .

6.5. CONFIDENCE INTERVAL FOR THE POPULATION STANDARD DEVIATION

Let us assume that we need to estimate the variance and the standard deviation of a population with normal probability distribution with parameters μ and σ^2 . The variance is defined as the expectation of the squared difference of data points from the mean of the data set. The Chi-square distribution is a family of distributions that represents a sum of independent squared standard normal random variables. Since the Chi-square represents squared data, it can be used to find a confidence interval for the standard deviation or variance.

6.5.1. CONFIDENCE INTERVAL FOR THE STANDARD DEVIATION σ OF NORMALLY DISTRIBUTED POPULATION WITH KNOWN MEAN

For a random sample $X_1, X_2, \dots, X_n$ of size n that is taken from a normally distributed population X with known mean μ , and confidence level β ,

the confidence interval for the standard variance σ^2 is the interval:

$$\left[\frac{1}{k_{\frac{\alpha}{2}}(n-1)} \sum_{i=1}^n (x_i - \mu)^2, \frac{1}{k_{1-\frac{\alpha}{2}}(n-1)} \sum_{i=1}^n (x_i - \mu)^2 \right], \text{ and}$$

the confidence interval for the standard deviation σ is the interval:

$$\left[\sqrt{\frac{1}{k_{\frac{\alpha}{2}}(n-1)} \sum_{i=1}^n (x_i - \mu)^2}, \sqrt{\frac{1}{k_{1-\frac{\alpha}{2}}(n-1)} \sum_{i=1}^n (x_i - \mu)^2} \right],$$

where for $n-1$ degrees of freedom the values $k_{\frac{\alpha}{2}}(n-1)$ and $k_{1-\frac{\alpha}{2}}(n)$ are read from the table of the χ^2 - distribution.

Example 6.5.1. In Spiti Valley water sediment in May is a random variable with a normal distribution with a mean 30. In the last twenty years in the area of $1m^2$ are registered with the following sediment quantities (in mm).

32	33	29	27	31	30	28	34	35	33
27	28	28	32	31	30	33	29	28	33

Find the confidence interval for the standard deviation of the water-sediment with 0.90 confidence level.

Solution. We are considering a sample of size $n=20$ that is taken from normally distributed population with known mean $\mu=30$, confidence level $1-\alpha=0.90$. Therefore the confidence interval for the standard deviation σ is the interval:

$$\left[\sqrt{\frac{1}{k_{\frac{\alpha}{2}}(n-1)} \sum_{i=1}^n (x_i - \mu)^2}, \sqrt{\frac{1}{k_{1-\frac{\alpha}{2}}(n-1)} \sum_{i=1}^n (x_i - \mu)^2} \right],$$

where for $\alpha=1-0.90=0.10$ and 19 degrees of freedom from the table for the χ^2 - distribution we have

$$k_{\frac{\alpha}{2}}(n-1) = k_{0.05}(19) = 30.14 \text{ and } k_{1-\frac{\alpha}{2}}(n-1) = k_{0.95}(19) = 10.12 .$$

Since, $x_i - \mu$

i	x_i	$x_i - \mu$	$(x_i - \mu)^2$
-----	-------	-------------	-----------------

POINT ESTIMATION. CONFIDENCE INTERVAL

1	32	2	4
2	33	3	9
3	29	-1	1
4	27	-3	9
5	31	1	1
6	30	0	0
7	28	-2	4
8	34	4	16
9	35	5	25
10	33	3	9
11	27	-3	9
12	28	-2	4
13	28	-2	4
14	32	2	4
15	31	1	1
16	30	0	0
17	33	3	9
18	29	-1	1
19	28	-2	4
20	33	3	9
Σ			123

We have the confidence interval for the standard deviation σ is the interval:

$$\left[\sqrt{\frac{123}{31.14}}, \sqrt{\frac{123}{10.12}} \right], \text{ or } [1.99, 3.49].$$

6.5.2. CONFIDENCE INTERVAL FOR THE STANDARD DEVIATION σ OF NORMALLY DISTRIBUTED POPULATION WITH UNKNOWN MEAN

For a random sample $X_1, X_2, \dots, X_n$ of size n that is taken from a normally distributed population X with unknown mean μ , and confidence level β ,

the confidence interval for the standard variance σ^2 is the interval:

$$\left[\frac{(n-1)s^2}{k_{\frac{\alpha}{2}}(n-1)}, \frac{(n-1)s^2}{k_{1-\frac{\alpha}{2}}(n-1)} \right], \text{ and}$$

the confidence interval for the standard deviation σ is the interval:

$$\left[s \cdot \sqrt{\frac{n-1}{k_{\frac{\alpha}{2}}(n-1)}}, s \cdot \sqrt{\frac{n-1}{k_{1-\frac{\alpha}{2}}(n-1)}} \right],$$

where s^2 is the sample variance and for $n-1$ degrees of freedom the values $k_{\frac{\alpha}{2}}(n-1)$ and $k_{1-\frac{\alpha}{2}}(n)$ are read from the table of the χ^2 -distribution.

Example 6.5.2. Consider the sample of size $n=25$ that is taken from normally distributed population such that its sample variance $s^2=16$. Find the confidence interval for the standard deviation of the population with confidence coefficient $1-\alpha=0.90$.

Solution. We are considering a sample of size $n=25$ that is taken from normally distributed population with unknown mean, but sample standard deviation $s=\sqrt{16}=4$ and confidence level $1-\alpha=0.90$. Therefore the confidence interval for the standard deviation is the interval:

$$\left[s \cdot \sqrt{\frac{n-1}{k_{\frac{\alpha}{2}}(n-1)}}, s \cdot \sqrt{\frac{n-1}{k_{1-\frac{\alpha}{2}}(n-1)}} \right],$$

where for $\alpha=1-0.90=0.10$ and 24 degrees of freedom from the table for the χ^2 -distribution we have

$$k_{\frac{\alpha}{2}}(n-1) = k_{0.05}(24) = 36.42 \text{ and } k_{1-\frac{\alpha}{2}}(n-1) = k_{0.95}(24) = 13.85.$$

So, the confidence interval of the standard deviation of the population is:

$$\left[4 \cdot \sqrt{\frac{24}{36.42}}, 4 \cdot \sqrt{\frac{24}{13.85}} \right] \text{ or } [3.24, 5.27].$$

6.6. CONFIDENCE INTERVAL FOR THE POPULATION PROPORTION

The population proportion parameter is typically denoted as p , represents the probability that a randomly selected individual from the population has a specific characteristic, known as the probability of “success”. It is a fixed but often unknown value that describes the distribution of a binary experiment (e.g., success/failure, yes/no, defective/non-defective) within the entire population.

If we consider a random sample of size n from the binomial population such that the number of successes is S_n , confidence coefficient $1-\alpha$ and number $z_{1-\alpha}$ that is read from the table of standard normal distribution such that $\Phi_0(z_{1-\alpha}) = 1 - \frac{\alpha}{2}$.

The confidence interval for the binomial parameter p :

If the sample is taken from an infinitely large population, or from the finite population of size N with replacement or from finite population of size N without replacement but the sample size is less than 5% of the population size is the interval $[p_1, p_2]$, where p_1, p_2 are solutions of the equation:

$$(n^2 + nz_{1-\alpha}^2)p^2 - (2nS_n + nz_{1-\alpha}^2)p + S_n^2 = 0,$$

If the sample is taken from the finite population of size N without replacement but the sample size is greater than 5% of the population size is the interval $[p_1, p_2]$, where p_1, p_2 are solutions of the equation:

$$\left(n^2 + \frac{n(N-n)}{N-1}z_{1-\alpha}^2\right)p^2 - \left(2nS_n + \frac{n(N-n)}{N-1}z_{1-\alpha}^2\right)p + S_n^2 = 0.$$

Example 6.6.1. One political party is observing the willingness of the electorates in the next local elections to vote for its candidate. The electoral list is 6,000 registered voters and is considered a sample without replacement of 196 voters such that 138 of them stated that they will vote for the candidate of that party at the upcoming elections. Find the confidence interval of the proportion:

$$p = P\{\text{"the voter will vote for the candidate"}\},$$

with confidence coefficient $1 - \alpha = 0.90$.

Solution. From the table of standard normal distribution we have

$$\Phi_0(1.64) < \Phi_0(z_{0.9}) = 1 - \frac{\alpha}{2} = 1 - \frac{0.1}{2} = 0.95 < \Phi_0(1.65).$$

Therefore $z_{0.9} = 1.645$.

We are considering a random sample of size $n = 196$ from the binomial population with size $N = 6000$ such that the success is the

event “the voter will vote for the candidate”, so the number of successes is $S_n = 138$.

Since the size of the sample is $\frac{196}{6000} = 3.2\% < 5\%N$ the confidence interval for the proportion p is the interval $[p_1, p_2]$, where p_1, p_2 are solutions of the equation:

$$(n^2 + nz_{1-\alpha}^2)p^2 - (2nS_n + nz_{1-\alpha}^2)p + S_n^2 = 0.$$

By replacing all the values for the problem we will consider the following equation:

$$(196^2 + 196 \cdot 1.645^2)p^2 - (2 \cdot 196 \cdot 138 + 196 \cdot 1.645^2)p + 138^2 = 0, \text{ or}$$

$$38946.38p^2 - 54626.38p + 19044 = 0.$$

The solutions of the previous equation are:

$$p_{\frac{1}{2}} = \frac{54626.38 \pm \sqrt{(-54626.38)^2 - 4 \cdot 19044 \cdot 38946.38}}{2 \cdot 38946.38} = \text{i.e.,}$$

$$= \frac{54626.38 \pm 4154.75}{77892.76}$$

$$p_1 = 0.648 \text{ and } p_2 = 0.755.$$

Therefore, the confidence interval for the binomial parameter p is the interval $p_2 = [0.648, 0.755]$.

Example 6.6.2. One insurance company has 6 432 motor casco in Ohrid. From this population is taken a sample without replacement of 400 randomly selected clients, and it is determined that 320 of these clients have not participated in traffic accidents. Find the 95% confidence interval of the proportion:

$$p = P\{\text{"the client has participated in a traffic accident"}\},$$

Solution. From the table of standard normal distribution we have $\Phi_0(z_{0.95}) = 1 - \frac{0.05}{2} = 0.975 = \Phi_0(1.96)$, therefore $z_{0.95} = 1.96$.

We are considering a random sample of size $n = 400$ from the population with size $N = 6432$ such that the success is the event “the client has participated in a traffic accident”, so the number of successes is $S_n = 400 - 320 = 80$.

Since the sample is without replacement and its size is $\frac{400}{6432} = 6.2\% > 5\%$, the confidence interval for the proportion p is the interval $[p_1, p_2]$, where p_1, p_2 are solutions of the equation:

$$\left(n^2 + \frac{n(N-n)}{N-1} z_{1-\alpha}^2 \right) p^2 - \left(2nS_n + \frac{n(N-n)}{N-1} z_{1-\alpha}^2 \right) p + S_n^2 = 0.$$

By replacing all the values for the problem we will consider the following equation:

$$\left(400^2 + \frac{400(6432-400)}{6432-1} \cdot 1.96^2 \right) p^2 - \left(2 \cdot 400 \cdot 80 + \frac{n(N-n)}{N-1} \cdot 1.96^2 \right) p + 80^2 = 0,$$

$$\text{or } 161441.3p^2 - 65441.3p + 6400 = 0.$$

The solutions of the previous equation are:

$$p_{\frac{\gamma}{2}} = \frac{65441.3 \pm \sqrt{(-65441.3)^2 - 4 \cdot 6400 \cdot 161441.3}}{2 \cdot 161441.3} = \frac{65441.3 \pm 12233.8}{322882.6},$$

$$p_1 = 0.1648 \text{ and } p_2 = 0.2406.$$

Therefore, the confidence interval for the binomial parameter p is the interval $[0.1648, 0.2406]$.

7. HYPOTHESIS TESTING

In the field of mathematical statistics known as hypothesis testing, data is used to determine whether a specific claim or assumption about a population characteristic can be considered true within a certain margin of error. This process typically involves analyzing the value of a particular population parameter, the characteristics of the statistical distribution, variations of the same parameter across different datasets, relationships between two or more characteristics within a dataset, and commonalities.

For instance, hypothesis testing can determine whether the claim that the mean weight of a particular product differs from the standard values is valid, whether the average age of employees in higher education differs from that in primary education, whether a population conforms to an exponential distribution, the type of correlation between time spent studying and grades achieved in a class, or the opinions of voters on a new tax reform bill to their income levels, and so forth.

For example, hypothesis testing can be used to evaluate whether the mean weight of a product deviates from standard values, whether the average age of employees in higher education differs from that in primary education, whether a population follows an exponential distribution, the nature of the correlation between study time and academic performance, or the relationship between voters' opinions on a tax reform bill and their income levels, among other applications.

The validity of such statements and assumptions can only be determined by thoroughly analyzing all elements of the population. However, this is often impractical due to the population's large size, making it necessary to test statistical hypotheses using a sample dataset. A statistical hypothesis is a well-defined assumption about a population characteristic. The methodological approach used to assess the validity of such claims is known as the hypothesis testing method. Since conclusions about a hypothesis are drawn from a sample dataset, there is always a possibility of error in the final decision.

In this chapter, we will focus exclusively on testing hypotheses concerning specific values of population parameters and the differences in these parameters between two populations. We will examine the basic principles, methods, and categories of errors linked to hypothesis testing,

including Type I and Type II errors. Furthermore, we will cover the significance level, p -values, and decision-making rules that determine whether to accept or reject a hypothesis. By understanding these components, we can make accurate statistical inferences while minimizing the risk of errors.

Hypothesis testing is a formal procedure for investigating our ideas about the world using mathematical statistics. There are five main steps in testing a hypothesis. While specific details may vary, the process will always follow some variation of these steps.

1 step. State the research hypothesis as a null hypothesis and alternate hypothesis.

2 step. Collect data in a way designed to test the hypothesis.

3 step. Recognize and perform an appropriate statistical test.

4 step. Decide whether to reject or fail to reject the null hypothesis.

5 step. Conclude and present the findings based on the results.

After formulating an initial research hypothesis—the assumption to be investigated—it is essential to express it as a null and alternative hypothesis, allowing for mathematical testing.

The alternate hypothesis is usually the initial hypothesis that predicts a relationship between variables. The null hypothesis is a prediction of no relationship between the variables you are interested in.

For example, if we want to test whether there is a relationship between gender and height, based on the knowledge of human physiology, we formulate a hypothesis that men are, on average, taller than women. To test this hypothesis, it will be restated as:

H_0 : Men are, on average, not taller than women.

H_1 : Men are, on average, taller than women.

The null hypothesis (H_0) is a statement about the value of a parameter θ that we attempt to challenge through the testing process. It is assumed to be true until proven otherwise. The null hypothesis can be either simple or composite.

The null hypothesis is simple if it states that the parameter is equal to a single, predefined numerical value. It is written in the following form:

$H_0 : \nu = \nu_0$, which is read as: “the null hypothesis states that the parameter ν is equal to ν_0 ”.

If the null hypothesis includes multiple values, for example $H_0 : \nu \geq \nu_0$, then it is considered composite.

The null hypothesis must contain the equal sign because the distribution of test statistic requires the null hypothesis to be assumed to be true and the value attached to the equal sign is then the value assumed to be true.

Every null hypothesis is accompanied by an alternative hypothesis (H_1). The alternative hypothesis typically includes all values of the parameter ν , that are not covered by the null hypothesis. It represents the statement that we predict and want to prove as true.

The null and alternative hypotheses appear in the following three forms:

$$H_0 : \nu = \nu_0, H_1 : \nu \neq \nu_0$$

$$H_0 : \nu \leq \nu_0, H_1 : \nu > \nu_0$$

$$H_0 : \nu \geq \nu_0, H_1 : \nu < \nu_0.$$

Apart from the structure of the hypotheses, the significance of the content in the null hypothesis compared to the alternative hypothesis plays a crucial role in the test's outcome. More specifically, the assertion we consider correct and seek to validate through our testing is represented as the alternative hypothesis. The reasoning behind this is as follows: in hypothesis testing, we only reject the null hypothesis when the sample data provides strong evidence against it. If the data does not significantly contradict the null hypothesis, we fail to reject it. In other words, rejecting the null hypothesis requires much stronger evidence than is needed to accept it. If we determine, based on the sample, that the null hypothesis should be rejected, we can be quite confident that the alternative hypothesis holds. On the other hand, if the sample supports

the null hypothesis, this does not imply that it is true, but rather that there is not enough evidence to dispute it.

For this reason, the hypothesis we aim to validate is expressed as the alternative hypothesis. If the sample data challenges the null hypothesis, we can be fairly confident in our correctness. Conversely, if the null hypothesis is supported by the information obtained from a random sample, we cannot entirely guarantee its correctness. The results from the sample are insufficient to claim the alternative hypothesis is true; instead, based on those results, we can conclude that the alternative hypothesis is likely accepted as true. Conversely, if the sample data backs the null hypothesis, we determine that the null hypothesis remains unchallenged. Even though we express our belief through the alternative hypothesis, during the testing procedure, we focus solely on the null hypothesis. We start with the assumption that the null hypothesis is valid and attempt to challenge or refute this premise.

To ensure that a statistical test is valid, it is crucial to conduct sampling and gather data in a manner that is tailored to evaluate your hypothesis. If your data is not representative, making statistical inferences about the population of interest becomes impossible. For instance, to examine the differences in average height between males and females, the sample should consist of an equal number of men and women, while also encompassing various socioeconomic backgrounds and other control variables that may affect average height. Additionally, the geographical scope needs to be considered, such as whether the focus is global or limited to a specific country. A potential source for a dataset in this context could be a census dataset, as it contains information from diverse regions and social classes and is accessible to numerous countries worldwide.

The next step in the hypothesis testing process, after precisely formulating the null and alternative hypothesis, is selecting a test statistic and determining its probability distribution.

The test statistic is the criterion based on which the hypothesis test is performed. Most commonly, an unbiased estimate of the parameter ν or some transformation of it is used. For example, if we are testing a hypothesis about the population's mean, the point estimate $\bar{X}$ of the mathematical expectation, or one of its standardized values, such as Z and t is applied as the test criterion.

Since an estimate of a population parameter is a random variable, the test statistic is also a random variable. The decision to reject or not reject the null hypothesis depends on the observed values of the test statistic from a given random sample.

The null hypothesis, as a statement about the value of a population parameter, is either true or false, and the information from the random sample is either consistent with the null hypothesis or contradictory to it. Hence, the sample data may either support the null hypothesis or contradict it. Since the decision to reject or fail to reject the null hypothesis is based on a single random sample, there is a possibility that the conclusion drawn from the test may be incorrect. The following Table 7.1 illustrates the possible test outcomes in relation to their truth.

Table 7.1. Outcomes of testing hypotheses and their real truth/false values

Actual Fact		H_0	
		True	False
Decision	H_0 failed to reject	Correct ($1 - \alpha$)	Type II error (β)
	H_0 reject	Type I error (α)	Correct ($1 - \beta$)

According to Table 7.1, if the null hypothesis is true and the sample information is consistent with it, we do not reject the null hypothesis and the conclusion from the test will be correct. We will also act correctly if we reject the false null hypothesis. However, the null hypothesis may be true, but we still reject it based on the sample data. Then the conclusion is wrong, and the error we have made is called a Type I error. Another error that we can make when inferring is not rejecting the false null hypothesis. This error is called a Type II error. We can only make one error when inferring. Namely, if we decide not to reject the null hypothesis, it is possible to make a Type II error, and if we decide to reject the null hypothesis, then it is possible to make a Type I

error. When testing hypotheses, we strive to determine the probability of the above errors occurring and, if possible, reduce them.

The probability that we reject the true null hypothesis (the probability of making a Type I error) is called **the level of significance of the test, the acceptable risk, false alarm**, and it is denoted by α :

$$\alpha = P(H_0 \text{ reject} \mid H_0 \text{ is true}).$$

The probability that we will not reject the false null hypothesis (the probability of making a Type II error) is called **the risk of a Type II error** and is denoted by β :

$$\beta = P(H_0 \text{ fail to reject} \mid H_0 \text{ is not true}).$$

In the testing hypothesis procedure, the significance level α is chosen in advance, i.e. we choose the probability with which it is possible to reject the true null hypothesis. Namely, by accepting a certain significance level of the test, α , which are most often the levels 0.05 and 0.01, we accept the possibility of making a mistake and rejecting the true null hypothesis. While the probability of making the correct decision, i.e. accepting the true null hypothesis, is called the confidence level, and is $1 - \alpha$.

For a given test statistic and sample size, the choice of the test significance level completely determines the risk of a Type 2 error. The probability of a Type 2 error β can be very large, and since the consequences of accepting a false null hypothesis are often as unfavorable as the consequences of rejecting a true null hypothesis, it is desirable to control the acceptable risk α in addition to controlling the risk of a Type II error β .

Instead of risk of a Type II error, the power of the test is often mentioned, which represents the probability of rejecting a false null hypothesis and is the complementary value of risk of a Type II error β , that is, it is equal to the value $1 - \beta$.

So, the risk of a Type II error β is determined by the significance level of the test. For a given sample size, as the risk of a Type I error α increases, the risk of a Type II error β decreases and vice versa, but these two risks do not add up to 1, that is, they are not complementary. The only way to reduce the probability of one error and not increase the

probability of the other error is to increase the sample size. For a given value of the risk of a Type I error α , as the sample size increases, the risk of a Type II error β decreases, that is, the power of the test increases. So, by using large samples, we increase the probability of rejecting the false null hypothesis.

When the probabilities of all the statements in Table 7.1 are known, we can calculate the total error of the test with the formula:

$$\begin{aligned} \text{Total error} &= \\ &= P(H_0 \text{ rejected} \mid H_0 \text{ is true}) \cdot P(H_0 \text{ is true}) + \\ &+ P(H_0 \text{ fail to reject} \mid H_0 \text{ is not true}) \cdot P(H_0 \text{ is not true}) = \\ &= \alpha \cdot P(H_0 \text{ is true}) + \beta \cdot P(H_0 \text{ is not true}) \end{aligned}$$

When performing the hypothesis testing procedure, different values of the test statistic occur, some of which are less likely and some more likely when the null hypothesis is true. In the third phase of hypothesis testing, these values are separated into the so-called rejection region of the null hypothesis and the non-rejection region (which is conventionally called the acceptance region) of the null hypothesis. The critical region, or rejection region of the null hypothesis, is a subset of unlikely values of the test statistic when the null hypothesis is true. If one of these values is realized, the null hypothesis is rejected. The subset of the remaining possible values of the test statistic is the acceptance region of the null hypothesis; with the realization of one of these values, the null hypothesis is not rejected. The critical region, its size and position, are determined by: the shape of the hypothesis, the test statistic and its distribution, and the level of significance.

7.1. ONE SAMPLE HYPOTHESIS TESTING

7.1.1. TESTS ABOUT POPULATION MEAN

Z – test

The hypothesis testing for the mathematical expectation as mentioned above is based on two statistics Z and t . In this process, the statistics are not always justified, but some assumptions are necessary that relate to the characteristics of the population, which were previously mentioned in the procedure for interval estimation of the mathematical expectation. Namely, if the population has a normal distribution, or the distribution is arbitrary, but the sample is large enough, $n \geq 30$, so that according to the central limit theorem we can assume that its average has a normal distribution, the statistic Z has a standardized normal distribution.

The procedure for testing the hypothesis about population mean by the Z test is:

First, we check the *Conditions for using Z test when σ^2 is known* - a random sample of size n that is taken from a normally distributed population or a random sample of size $n \geq 30$ that is taken from a population with unknown distribution, the sample is taken from an infinitely large population, or the sample is taken from a finite population with replacement or the sample is taken from a finite population without replacement but the sample size n is less than 5% of the population size N .

Then we choose the null and alternative hypotheses

$$\begin{array}{ll} H_0 : \mu = \mu_0, & H_1 : \mu \neq \mu_0 \\ H_0 : \mu \leq \mu_0, & H_1 : \mu > \mu_0 \\ H_0 : \mu \geq \mu_0, & H_1 : \mu < \mu_0 \end{array}$$

where μ_0 is a real number and we use this test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}.$$

We consider three cases:

$$\text{C1) } H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

In this case, first, we calculate a specific value $\bar{z} = \frac{|\bar{x} - \mu_0|}{\sigma} \sqrt{n}$ where $\bar{x}$ is the sample mean, then we determine the critical value $z_{\alpha/2}$ such that $P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}$, i.e., $\Phi_0(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$ and make a decision

- if $\bar{z} \leq z_{\alpha/2}$ the null hypotheses H_0 is not rejected or
- if $\bar{z} > z_{\alpha/2}$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

$$\text{C2) } H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0.$$

In this case, first, we calculate a specific value $\bar{z} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$, where $\bar{x}$ is the sample mean, then we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \geq z_\alpha \right\} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 or
- ako $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C3) } H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0.$$

In this case, first, we calculate a specific value $\bar{z} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$, where $\bar{x}$ is the sample mean, then we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \leq -z_\alpha \right\} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 or

- if $\bar{z} < -z_{\alpha}$ reject H_0 with an acceptable risk of a Type I error less α .

But if the sample is taken from a finite population without replacement but the sample size n is greater than 5% of the population size N we use this statistic $Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{\frac{n(N-1)}{N-n}}$ and the procedure is:

First, we check the *Conditions for using Z test when σ^2 is known* - a random sample of size n that is taken from a normally distributed population or a random sample of size $n \geq 30$ that is taken from a population with unknown distribution, and the sample is taken from a finite population without replacement but the sample size n is greater than 5% of the population size N .

Then we choose the null and alternative hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$$

where μ_0 is a real number and we use this statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{\frac{n(N-1)}{N-n}}$$

We consider three cases:

$$C1) \quad H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

We calculate a specific value $\bar{z} = \frac{|\bar{x} - \mu_0|}{\sigma} \sqrt{\frac{n(N-1)}{N-n}}$, where $\bar{x}$ is the sample mean, then we determine the critical value $z_{\alpha/2}$ such that

$$P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}, \text{ i.e., } \Phi_0 \left(z_{\alpha/2} \right) = 1 - \frac{\alpha}{2} \text{ and make a decision}$$

- if $\bar{z} \leq z_{\alpha/2}$ the null hypotheses H_0 is not rejected or

- if $\bar{z} > z_{\alpha/2}$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

C2) $H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$.

We calculate a specific value $\bar{z} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{\frac{n(N-1)}{N-n}}$, where $\bar{x}$ is the sample mean, then we determine the critical value z_α such that $P_{H_0} \{ \bar{z} \geq z_\alpha \} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 or

- ako $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

C3) $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$.

We calculate a specific value $\bar{z} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{\frac{n(N-1)}{N-n}}$, where $\bar{x}$ is the sample mean, then we determine the critical value z_α such that $P_{H_0} \{ \bar{z} \leq -z_\alpha \} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 or

- if $\bar{z} < -z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

Example 7.1.1. During the test output from 25 shifts, the machine ATV produces on average 625 products per shift with a standard deviation of 30 products. Does the test production confirm the declaration of the supplier that the machine produces more than 620 products per shift, with acceptable risk $\alpha = 0.05$?

Solution. We are considering a sample size $n = 25$ with sample mean 625 that is taken from an infinitely large normally distributed population with known standard deviation $\sigma = 30$ and we want to test if

the average number of products per shift is greater than 620 (alternative hypothesis), so we will use z-test statistic for the following hypotheses:

$$H_0 : \mu = 620, \quad H_1 : \mu > 620 \quad \text{and statistic}$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}.$$

Therefore, we calculate the specific value

$$\bar{z} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} = \frac{625 - 620}{60} \sqrt{25} = 0.42$$

and from the table of the standard normal distribution about the critical value $Z_{0.05}$:

$$\Phi_0(1.64) = 0.94950 < \Phi_0(z_{0.05}) = 1 - 0.05 = 0.95 < 0.95053 = \Phi_0(1.65)$$

$$\text{or } z_{0.05} = \frac{1.64 + 1.65}{2} = 1.645.$$

Since $\bar{z} = 0.42 < 1.645 = z_{0.05}$, the null hypothesis is not rejected. So the test production doesn't confirm the declaration of the supplier that the machine produces more than 620 products per shift.

Example 7.1.2. The management of one company decided to change the program support of desk operations to reduce the average time of processing the orders of the clients and thereby reduce the waiting time for the clients. It has been determined that a change in program support is payable if the average order processing time is less than 30 seconds per order." The new support program was tested on 324 randomly selected orders and noted that the average time of processing one order is 29 seconds. What conclusion can be drawn based on the conducted investigation, if we know that the time of processing an order is a normally distributed random variable with a standard deviation of 5 and acceptable risk $\alpha = 0.02$?

Solution. We are considering a sample size $n = 324$ with sample mean $\bar{x} = 29$ that is taken from an infinitely large normally distributed population with a known standard deviation $\sigma = 5$, and we want to test if the average time of processing the orders is less than 30 seconds per order (alternative hypothesis), so we will use z - test statistic for the following hypotheses:

$$H_0 : \mu = 30 \quad H_1 : \mu < 30 \text{ and test statistic } Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}.$$

The specific value is:

$$\bar{z} = \frac{29 - 30}{5} \sqrt{324} = -3.6$$

and from the table of the standard normal distribution about the critical value $Z_{0.02}$

$$\Phi_0(2.05) = 0.97982 < \Phi_0(z_{0.02}) = 1 - 0.02 = 0.98 < 0.9803 = \Phi_0(2.06)$$

or $z_{0.02} = \frac{2.05 + 2.06}{2} = 2.055.$

Since $\bar{z} = -3.6 < -2.055 = z_{0.02}$ the null hypothesis is rejected, with an acceptable risk of a Type I error less than 0.02. So we have enough reasons to think that the change of program support of the desk operations is payable.

Example 7.1.3. Performed nine measurements of the length of randomly selected units of one type of product and thus obtained the following results:

5.29; 5.31; 5.32; 5.31; 5.30; 5.32; 5.28; 5.33 and 5.33 cm

If the length of the product's units follows a normally distributed random variable with a measurement standard error of 0.02, test the assumption that the true length of the product is 5.3, given an acceptable level of risk $\alpha = 0.05$.

Solution: We are considering a sample of size $n = 9$ that is taken from an infinitely large normally distributed population with known standard deviation $\sigma = 0.02$, and we want to test that the exact value of the length of the units of the product is 5.3, so we will use z-test statistic for the following hypotheses:

$$H_0 : \mu = 5.3 \quad H_1 : \mu \neq 5.3 \text{ and test statistic } Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}.$$

The sample mean is

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^9 x_i = \\ &= \frac{1}{9} (5.29 + 5.31 + 5.32 + 5.31 + 5.30 + 5.32 + 5.28 + 5.33 + 5.33) = 5.31 \end{aligned}$$

Therefore, we calculate the specific value

$$\bar{z} = \frac{|\bar{x} - \mu_0|}{\sigma} \sqrt{n} = \frac{|5.31 - 5.3|}{0.02} \sqrt{9} = 1.5$$

and from the table of the standard normal distribution about the critical value $Z_{0.05}$:

$$\Phi_0(z_{0.05}) = 1 - \frac{0.05}{2} = \Phi_0(1.96) \text{ or } z_{0.05} = 1.96.$$

Since $\bar{z} = 1.5 < 1.96 = z_{0.05}$, the null hypothesis is not rejected. So the test production confirms the exact value of the length of the product is 5.3.

Example 7.1.4. The coffee “Aroma” is distributed in packages of 250 grams. Let's assume that the weight of the packaging is a normally distributed random variable with a variance 36. A series of 10,000 packs is checked by a sample of 625 packages, taken without replacement. The average weight of the sample's packs is 247.5 gr. Can we conclude from the results of the observation that the average weight of packaging is 250 grams with acceptable risk $\alpha = 0.01$?

Solution: We are considering a sample size $n = 625$ with sample mean $\bar{x} = 247.5$ that is taken from finite normally distributed population with size $N = 10,000$ without replacement, known standard deviation $\sigma = \sqrt{36} = 6$, and we want to test if the average weight of packaging is 250 grams. Since the sample size is $\frac{625}{10000} = 6.25\%$ greater than 5% of the population size we will use z-test statistic for the following hypotheses:

$$H_0 : \mu = 250 \quad H_1 : \mu \neq 250 \text{ and test statistic } Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{\frac{n(N-1)}{N-n}}.$$

Therefore, we calculate the specific value

$$\bar{z} = \frac{|\bar{x} - \mu_0|}{\sigma} \sqrt{\frac{n(N-1)}{N-n}} = \frac{|247.5 - 250|}{6} \sqrt{\frac{625(10000-1)}{10000-625}} = 10.76$$

and from the table of the standard normal distribution about the critical value $Z_{0.01}$:

$$\Phi_0(z_{0.01}) = 1 - \frac{0.01}{2} = 0.995 = \Phi_0(2.58) \text{ or } z_{0.01} = 2.58.$$

Since $\bar{z} = 10.76 > 2.58 = z_{0.01}$, the null hypothesis is rejected, with acceptable risk of a Type I error less than 0.01. So we don't have enough reasons to think that average weight of packaging is 250 grams.

Student's t - test. The above assumptions, which are necessary for the Z - test to be used in testing, are often not met in practice, primarily because the population variance is not known. Then the standard error $\frac{\sigma}{\sqrt{n}}$ cannot be calculated, and therefore the value of the Z - test statistic cannot be calculated. Therefore, the standard deviation σ is first estimated, using the standard deviation, i.e., the sample deviation S , and then we obtain a new test statistic $t_{n-1} = \frac{\bar{X} - \mu_0}{S} \sqrt{n}$, the

so-called t - test statistic, for which the sample has a size $n < 30$, which has a Student's distribution with $n-1$ degrees of freedom.

The procedure for testing the hypotheses for the population mean using the t - test is:

First, we check the *Conditions for using t -test* - a random sample of size n that is taken from a normally distributed population and unknown population variance and with a random sample of size $n < 30$.

Then we choose the null and alternative hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$$

where μ_0 is a real number.

If the sample is taken from an infinitely large population, from a finite population with replacement, or from a finite population without replacement but the sample size n is less than 5% of the population size N , we use this test statistic

$$t_{n-1} = \frac{\bar{X} - \mu_0}{S} \sqrt{n} .$$

We consider three cases:

$$C1) \ H_0 : \mu = \mu_0 , \ H_1 : \mu \neq \mu_0 .$$

In this case, first, we calculate a specific value $\bar{t}_{n-1} = \frac{|\bar{x} - \mu_0|}{s} \sqrt{n}$ for the test, where $\bar{x}$ is the sample mean and s is the sample variance, then we determine the critical value $t_\alpha (n-1)$ from Student's distribution table, two-tailed test, and make a decision

- if $\bar{t}_{n-1} \leq t_\alpha (n-1)$ the null hypotheses H_0 is not rejected, or

- if $\bar{t}_{n-1} > t_\alpha (n-1)$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

$$C2) H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0.$$

In this case, first, we calculate a specific value $\bar{t}_{n-1} = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$, where $\bar{x}$ is the sample mean and s is the sample variance, then we determine the critical value $t_\alpha(n-1)$ from the Student's distribution table, one-tailed test and make a decision:

- if $\bar{t}_{n-1} \leq t_\alpha(n-1)$ the null hypotheses H_0 is not rejected, or
- if $\bar{t}_{n-1} > t_\alpha(n-1)$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

$$C3) H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0.$$

In this case, first, we calculate a specific value $\bar{t}_{n-1} = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$, where $\bar{x}$ is the sample mean and s is the sample variance, then we determine the critical value $t_\alpha(n-1)$ from the Student's distribution table, one-tailed test, and make a decision:

- if $\bar{t}_{n-1} \geq -t_\alpha(n-1)$ fail to reject H_0 , or
- if $\bar{t}_{n-1} < -t_\alpha(n-1)$ reject H_0 with an acceptable risk of a Type I error less α .

Example 7.1.5. Performed twenty measurements of the weight of randomly selected units of one type of product and thus obtained the following results in grams:

97, 99, 98, 96, 98, 101, 98, 99, 97, 95, 98, 101, 97, 99, 98, 100, 98, 96,
97, 98.

Let's assume that the weight of the product is a normally distributed random variable. Can we conclude from the results of the observation that the average weight of the products is 100 grams with acceptable risk $\alpha = 0.01$?

Solution:

We are considering a sample of size $n=20$ that is taken from infinitely large normally distributed population with unknown standard deviation, and we want to test that the average weight of the units of the products is 100 grams, so we will use t - test statistic for the following hypotheses:

$$H_0 : \mu = 100 \quad H_1 : \mu \neq 100 \quad \text{and test statistic } t_{n-1} = \frac{\bar{X} - \mu_0}{S} \sqrt{n}$$

The sample mean is

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^{20} x_i = \\ &= \frac{1}{20} (97 + 99 + 98 + 96 + 98 + 101 + 98 + 99 + 97 + 95 + 98 + 101 + . \\ &\quad + 97 + 99 + 98 + 100 + 98 + 69 + 67 + 98) = 98 \end{aligned}$$

and the sample variance

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n\bar{x}^2 \right) = \\ &= \frac{1}{20-1} (4 \cdot 97^2 + 3 \cdot 99^2 + 7 \cdot 98^2 + 2 \cdot 96^2 + 2 \cdot 101^2 + 95^2 + 100^2 - 20 \cdot 98^2) = \\ &= \frac{1}{19} \cdot 46 = 2.42 \end{aligned}$$

and sample standard deviation $s = \sqrt{2.42} = 1.56$.

Therefore, we calculate the specific value

$$\bar{t}_{19} = \frac{|\bar{x} - \mu_0|}{s} \sqrt{n} = \frac{|98 - 100|}{1.56} \sqrt{20} = 5.73$$

and from the table of t – distribution two-tailed test for 19 degrees of freedom the critical value is $t_{\alpha}(n-1) = t_{0.01}(19) = 2.861$.

Since $\bar{t}_{19} = 5.73 > t_{0.01}(19) = 2.861$ null hypotheses H_0 is rejected with acceptable risk of a Type I error less than 0.01, so we conclude that

average weight of the products is 100 grams with acceptable risk of a Type I error less than 0.01.

Example 7.1.6. Due to the increase in turnover, leadership chain supermarkets Zeno decided to allow payment by credit cards. The management assumes that after the introduction of this novelty average monthly circulation at the supermarket will be greater than 300 000 euros. The payment was introduced experimentally in 15 randomly selected from total of 350 stores in this chain of supermarkets. The sample mean of the turnover is 316 345 euros, with sample standard deviation 4 812 euros. To what conclusion can be reached on the basis of the conducted observation, if we assume that the turnover is normally distributed random variable with acceptable risk $\alpha = 0.05$?

Solution: We are considering a sample of size $n = 15 < 30$ with sample mean $\bar{x} = 316345$ and sample variance $s = 4812$ that is taken from normally distributed population with size $N = 350$ without replacement ($\frac{15}{350} = 4.3\% < 5\%N$), unknown population standard deviation, and we want to test if the average turnover is greater than 300 000 euros (alternative hypothesis), so we will use t -test statistic for the following hypotheses:

$$H_0 : \mu = 300000 \quad H_1 : \mu > 300000 \text{ and statistic } t_{n-1} = \frac{\bar{X} - \mu_0}{S} \sqrt{n}.$$

Therefore, we calculate the specific value

$$\bar{t}_{14} = \frac{\bar{x} - \mu_0}{s} \sqrt{n} = \frac{316345 - 300000}{4812} \sqrt{15} = 13.15$$

and from the table of t – distribution one-tailed test for 14 degrees of freedom the critical value is $t_{\alpha}(n-1) = t_{0.05}(14) = 1.761$.

Since $\bar{t}_{14} = 13.15 > t_{0.05}(14) = 1.761$ null hypotheses H_0 is rejected with an acceptable risk of a Type I error of less than 0.05, so we conclude that allowing payment by credit cards will increase the turnover with an acceptable risk of a Type I error of less than 0.05.

As the number of samples increases, the Student's t -distribution approximates to the normal distribution. Thus, in practice, it is commonly

accepted that when a statistical population exhibits a normal distribution with unknown variance, and the sample size is sufficiently large, usually $n > 30$, hypothesis testing can utilize the z -test instead of the t -test.

The procedure for testing the hypothesis about the population mean when variance is unknown with large samples is:

First, we check the *Conditions for Z - test* - a random sample of size n that is taken from a population with unknown variance and with a random sample of size $n > 30$, the sample is taken from an infinitely large population, if the sample is taken from the finite population with replacement, or if the sample is taken from a finite population without replacement but the sample size n is less than 5% of the population size N .

Then we choose the null and alternative hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$$

where μ_0 is a real number and we use this statistic

$$Z = \frac{\bar{X} - \mu_0}{S} \sqrt{n}.$$

Three cases:

$$C1) \ H_0 : \mu = \mu_0, \ H_1 : \mu \neq \mu_0.$$

First, we calculate a specific value $\bar{z} = \frac{|\bar{x} - \mu_0|}{s} \sqrt{n}$, where $\bar{x}$ is the sample mean and s is the sample variance, then we determine the critical value $z_{\alpha/2}$ such that $P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}$, i.e., $\Phi_0 \left(z_{\alpha/2} \right) = 1 - \frac{\alpha}{2}$ and make a decision:

- if $\bar{z} \leq z_{\alpha/2}$ the null hypotheses H_0 is not rejected, or

- if $\bar{z} > z_{\alpha/2}$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

C2) $H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0 :$

First, we calculate a specific value $\bar{z} = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$, where $\bar{x}$ is the sample mean and s is the sample variance, then we determine the critical value z_α such that $P_{H_0} \{ \bar{z} \geq z_\alpha \} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 ,

-ako $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

C3) $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0 :$

First we calculate the specific value $\bar{z} = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$, where $\bar{x}$ is the sample mean and s is the sample variance, then we determine the critical value z_α such that $P_{H_0} \{ \bar{z} \leq -z_\alpha \} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 , or

- if $\bar{z} < -z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

But if the sample is taken from a finite population without replacement but the sample size n is greater than 5% of the population size N we use the statistic $Z = \frac{\bar{X} - \mu_0}{S} \sqrt{\frac{n(N-1)}{N-n}}$, and the procedure for the test is:

First, we check the *Conditions for Z - test* - a random sample of size n that is taken from a population with unknown variance and with a random sample of size $n > 30$, the sample is taken from a finite

population without replacement but the sample size n is greater than 5% of the population size N .

Then we choose the null and alternative hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$$

where μ_0 is a real number and we use this statistic

$$Z = \frac{\bar{X} - \mu_0}{S} \sqrt{\frac{n(N-1)}{N-n}}.$$

Three cases:

$$C1) \ H_0 : \mu = \mu_0, \ H_1 : \mu \neq \mu_0.$$

First, we calculate a specific value $\bar{z} = \frac{|\bar{x} - \mu_0|}{s} \sqrt{\frac{n(N-1)}{N-n}}$, where $\bar{x}$ is the sample mean and s is the sample variance, then we determine the critical value $z_{\alpha/2}$ such that $P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}$, i.e., $\Phi_0(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$ and make a decision:

- if $\bar{z} \leq z_{\alpha/2}$ the null hypotheses H_0 is not rejected, or

- if $\bar{z} > z_{\alpha/2}$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

$$C2) \ H_0 : \mu \leq \mu_0, \ H_1 : \mu > \mu_0:$$

First, we calculate a specific value $\bar{z} = \frac{\bar{x} - \mu_0}{s} \sqrt{\frac{n(N-1)}{N-n}}$, where $\bar{x}$ is the sample mean and s is the sample variance, then we determine critical value z_α such that $P_{H_0} \left\{ \bar{z} \geq z_\alpha \right\} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 ,

-ako $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

C3) $H_0: \mu \geq \mu_0, H_1: \mu < \mu_0$:

First we calculate the specific value or $\bar{z} = \frac{\bar{x} - \mu_0}{s} \sqrt{\frac{n(N-1)}{N-n}}$,

where $\bar{x}$ is the sample mean and s is the sample variance, then we determine the critical value z_α such that $P_{H_0}\{\bar{z} \leq -z_\alpha\} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 , or

- if $\bar{z} < -z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

Example 7.1.7. Department of Market Research follows the average monthly sales of chocolate “Milka” in 2856 retail stores. The stores are of approximately equal size and structure of the sale. Under the assumptions of the sales department, the sale of chocolate in retail is a random variable with approximately normal distribution, with a mean 300 kg and an unknown standard deviation. To verify the assumption, it is selected a random sample of 40 shops and the sale of this chocolate (in kg) found in the month of September 2024 is given in the following table:

289	291	296	301	305	311	318	292	299	302
305	313	292	299	302	305	314	292	303	305
305	308	309	309	308	313	298	304	306	307
298	299	301	303	307	318	305	308	309	306

To what conclusion can be reached based on the conducted observation if the acceptable risk is $\alpha = 0.001$?

Solution: We are considering a sample of size $n = 40 > 30$ that is taken from normally distributed population with size $N = 2856$ without replacement ($\frac{40}{2856} = 1.4\% < 5\%$ of N), unknown standard deviation, and we want to test if the average sale is equal to 300 kg, so we will use z- test statistic for the following hypotheses:

$$H_0 : \mu = 300, \quad H_1 : \mu \neq 300 \text{ and statistic } Z = \frac{\bar{X} - m_0}{S} \sqrt{n}.$$

The sample mean is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{40} (289 + 291 + 296 + \dots + 308 + 390 + 306) = 303.875$$

and the sample variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n\bar{x}^2 \right) = \\ &= \frac{1}{40-1} \left((289^2 + 291^2 + 296^2 + \dots + 308^2 + 309^2 + 306^2) - \right. \\ &\quad \left. - 40 \cdot 303.875^2 \right) = 49.292 \end{aligned}$$

and sample standard deviation $s = \sqrt{49.292} = 7.02$.

Therefore, we calculate the specific value

$$\bar{z} = \frac{|\bar{x} - \mu_0|}{s} \sqrt{n} = \frac{|303.875 - 300|}{7.02} \sqrt{40} = 3.49$$

and from the table of the standard normal distribution about the critical value $Z_{0.001}$:

$$\Phi_0(z_{0.001}) = 1 - 0.001 = 0.999 = \Phi_0(3.08) \text{ or } z_{0.001} = 3.08.$$

Since $\bar{z} = 3.49 > 3.8 = z_{0.001}$, the null hypothesis is rejected with acceptable risk of a Type I error less than 0.001, so we conclude that

average sale is different than 300 kg with acceptable risk of a Type I error less than 0.01.

7.1.2. TESTS FOR THE POPULATION PROPORTION

Hypothesis testing for population proportion using small samples is performed with its unbiased estimator that has a binomial distribution (in samples with repetition), that is, $P = \frac{S_n}{n}$, where S_n , denotes the number of randomly selected individuals from the population having a specific characteristic. However, if we have large samples, the previous estimator as a test statistic is standardized to the variable $Z = \frac{S_n - np_0}{\sqrt{np_0q_0}}$, where p_0 is the probability of “success”, and $q_0 = 1 - p_0$.

The procedure for hypothesis testing for the population proportion by large sample using the Z test is:

First, we check the *conditions for Z test* - the probability of “success” is the same constant in each trial, $np_0 \geq 5$, $nq_0 \geq 5$ and $n \geq 30$, the sample is taken from an infinitely large population, or the sample is taken from the finite population with replacement, or the sample is taken from a finite population without replacement but the sample size n is less than 5% of the population size N .

Then we choose the null and alternative hypotheses:

$$H_0 : p = p_0, \quad H_1 : p \neq p_0$$

$$H_0 : p \leq p_0, \quad H_1 : p > p_0$$

$$H_0 : p \geq p_0, \quad H_1 : p < p_0$$

where p_0 is a given number and we use test statistic

$$Z = \frac{S_n - np_0}{\sqrt{np_0q_0}}.$$

There are three cases:

C1) $H_0 : p = p_0$, $H_1 : p \neq p_0$.

First, we calculate the specific value $\bar{z} = \frac{|S_n - np_0|}{\sqrt{np_0q_0}}$, where S_n is the number of “successes”. Then we determine the critical value $z_{\alpha/2}$ such that $P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}$, i.e., $\Phi_0 \left(z_{\alpha/2} \right) = 1 - \frac{\alpha}{2}$ and make a decision:

- if $\bar{z} \leq z_{\alpha/2}$ the null hypotheses H_0 is not rejected, or
- if $\bar{z} > z_{\alpha/2}$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

C2) $H_0 : p \leq p_0$, $H_1 : p > p_0$.

First, we calculate the specific value $\bar{z} = \frac{S_n - np_0}{\sqrt{np_0q_0}}$, where S_n is the number of “successes”. Then we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \geq z_\alpha \right\} = \alpha$, i.e., $\Phi_0 \left(z_\alpha \right) = 1 - \alpha$ and make a decision:

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 , or
- if $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

C3) $H_0 : p \geq p_0$, $H_1 : p < p_0$.

First, we calculate the specific value $\bar{z} = \frac{S_n - np_0}{\sqrt{np_0q_0}}$, where S_n is the number of times that the event A occurs. Then we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \leq -z_\alpha \right\} = \alpha$, i.e., $\Phi_0 \left(z_\alpha \right) = 1 - \alpha$ and make a decision.

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 , or

- if $\bar{z} < -z_{\alpha}$ reject H_0 with an acceptable risk of a Type I error less α .

But if the sample is taken from a finite population without replacement but the sample size n is greater than 5% of the population size N we use the statistic $\bar{z} = \frac{S_n - np_0}{\sqrt{np_0q_0}} \sqrt{\frac{N-1}{N-n}}$, and the procedure for the test is:

First, we check the *conditions for Z test* - the probability of “success” is the same constant in each trial, $np_0 \geq 5$, $nq_0 \geq 5$ and $n \geq 30$, the sample is taken from a finite population without replacement but the sample size n is greater than 5% of the population size N .

Then we choose the null and alternative hypotheses:

$$H_0 : p = p_0, \quad H_1 : p \neq p_0$$

$$H_0 : p \leq p_0, \quad H_1 : p > p_0$$

$$H_0 : p \geq p_0, \quad H_1 : p < p_0$$

where p_0 is a given number, and we use a test statistic

$$Z = \frac{S_n - np_0}{\sqrt{np_0q_0}} \sqrt{\frac{N-1}{N-n}}.$$

There are three cases:

$$C1) \ H_0 : p = p_0, \ H_1 : p \neq p_0.$$

First, we calculate the specific value $\bar{z} = \frac{|S_n - np_0|}{\sqrt{np_0q_0}} \sqrt{\frac{N-1}{N-n}}$, where

S_n is the number of “successes”. Then we determine the critical value $z_{\alpha/2}$ such that $P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}$, i.e., $\Phi_0 \left(z_{\alpha/2} \right) = 1 - \frac{\alpha}{2}$ and make a decision:

- if $\bar{z} \leq z_{\alpha/2}$ the null hypotheses H_0 is not rejected, or

- if $\bar{z} > z_{\alpha/2}$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

C2) $H_0 : p \leq p_0, H_1 : p > p_0$.

First, we calculate the specific value $\bar{z} = \frac{S_n - np_0}{\sqrt{np_0q_0}} \sqrt{\frac{N-1}{N-n}}$, where

S_n is the number of “successes”. Then we determine the critical value z_α such that $P_{H_0} \{ \bar{z} \geq z_\alpha \} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision:

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 , or

- if $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

C3) $H_0 : p \geq p_0, H_1 : p < p_0$.

First, we calculate the specific value $\bar{z} = \frac{S_n - np_0}{\sqrt{np_0q_0}} \sqrt{\frac{N-1}{N-n}}$, where

S_n is the number of times that the event A occurs. Then we determine the critical value z_α such that $P_{H_0} \{ \bar{z} \leq -z_\alpha \} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision.

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 , or

- if $\bar{z} < -z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

Example 7.1.8. It is assumed that $\frac{3}{4}$ of all households in Ohrid have smart TVs. From a randomly selected sample of 1000 households is concluded that 732 households have smart TVs. What conclusion can we make from the observation of acceptable risk $\alpha = 0.05$?

Solution: We are considering a sample of size $n = 1000 > 30$ that is taken from an infinitely large population, and the number of successes is $S_n = 732$. We want to test if 0.75 of households in Ohrid have smart TVs, so we will use z- test statistic for the following hypotheses:

$$H_0 : p = 0.75, \quad H_1 : p \neq 0.75$$

and statistic $Z = \frac{S_n - np_0}{\sqrt{np_0q_0}}.$

Therefore, we calculate the specific value

$$\bar{z} = \frac{|S_n - np_0|}{\sqrt{np_0q_0}} = \frac{|732 - 1000 \cdot 0.75|}{\sqrt{1000 \cdot 0.75 \cdot 0.25}} = 1.315$$

and from the table of standard normal distribution about the critical value we have $\Phi_0(z_{0.05}) = 1 - \frac{0.05}{2} = 0.975 = \Phi_0(1.96)$, or $\Phi_0(z_{0.05}) = 1.96$.

Since $\bar{z} = 1.315 \leq z_{0.05} = 1.96$ the null hypothesis is not rejected, so we don't have enough reasons to say that the total number of households in Ohrid that have smart TVs is different than 0.75.

Example 7.1.9. The financial department of a wholesale company analyzes the reasons for the low inflow of funds. It is assumed that one reason is the irregularity in the payment of due invoices. The Department assumes that more than 75% of the clients pay out of the prescribed period of 60 days. Out of 4 270 unpaid invoices is taken a sample of 235 unpaid invoices, and 198 were found by exceeding the period of 60 days. What can you conclude based on this sample, with acceptable risk $\alpha = 0.02$?

Solution: We are considering a sample of size $n = 235 > 30$ that is taken from population with size $N = 4270$ such that the success is the event “the clients pay out of the prescribed period of 60 days”, so the number of successes is $S_n = 198$.

We want to test if the percentage of the clients that pay out of the prescribed period of 60 days is more than 75% and since the size of the sample is $\frac{235}{4270} = 5.5\% > 5\%N$, so we will use z- test statistic for the following hypotheses:

$$H_0 : p \leq 0.75, \quad H_1 : p > 0.75$$

and statistic $Z = \frac{S_n - np_0}{\sqrt{np_0q_0}} \sqrt{\frac{N-1}{N-n}}.$

Therefore, we calculate the specific value

$$\bar{z} = \frac{S_n - np_0}{\sqrt{np_0q_0}} \sqrt{\frac{N-1}{N-n}} = \frac{198 - 235 \cdot 0.75}{\sqrt{235 \cdot 0.75 \cdot 0.25}} \sqrt{\frac{4270-1}{4270-235}} = 3.4$$

and from the table of standard normal distribution about the critical value we have $\Phi_0(z_{0.02}) = 1 - 0.02 = 0.98 = \Phi_0(2.06)$, or $z_{0.02} = 2.06$.

Since $\bar{z} = 3.4 > z_{0.02} = 2.06$ the null hypothesis is rejected with acceptable risk of a Type I error less than $\alpha = 0.02$, so we have enough reasons to say that the percentage of the clients that pay out of the prescribed period of 60 days is more than 75%.

7.1.3. TESTS FOR THE POPULATION VARIANCE

χ^2 - test. For any experimental work aimed at making improvements, whether in the design, manufacturing process or field performance, there are two ways to make improvements. Move the center of the distribution, or reduce the spread of the data.

The ability to determine changes in variance is the subject of hypothesis testing for variance. We will explore the situation where the variance of the population is known.

For a random sample of $X_1, X_2, \dots, X_n$ measurements drawn from a normal population with mean μ and variance σ^2 , the sample variance

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) \text{ is an unbiased point estimator for the}$$

population variance σ^2 and moreover the statistics $\frac{(n-1)S^2}{\sigma^2}$ follows a Chi-square distribution, with $df = n-1$ degrees of freedom.

Therefore, χ^2 - test is used for hypothesis testing about the population variance and the procedure for this test is:

First, we check the *Conditions for χ^2 - test for the population variance* - the population must be normally distributed.

Then we choose the null and alternative hypotheses:

$$\begin{array}{ll} H_0 : \sigma^2 = \sigma_0^2 & H_1 : \sigma^2 \neq \sigma_0^2 \\ H_0 : \sigma^2 \leq \sigma_0^2 & H_1 : \sigma^2 > \sigma_0^2 , \\ H_0 : \sigma^2 \geq \sigma_0^2 & H_1 : \sigma^2 < \sigma_0^2 \end{array}$$

where σ_0 is a given number, and we use test statistics $k = \frac{(n-1)S^2}{\sigma^2}$.

There are three cases:

$$\text{C1) } H_0 : \sigma^2 = \sigma_0^2 \quad H_1 : \sigma^2 \neq \sigma_0^2 .$$

First, we calculate the specific value $k = \frac{(n-1)s^2}{\sigma_0^2}$ where s^2 is the sample variance. Then we determine two critical values $k_{\frac{\alpha}{2}}(n-1)$ and $k_{1-\frac{\alpha}{2}}(n-1)$ from the table for the χ^2 - distribution and make a decision:

- if $k_{1-\frac{\alpha}{2}}(n-1) \leq k \leq k_{\frac{\alpha}{2}}(n-1)$ null hypotheses H_0 fail to reject,

- if $k < k_{1-\frac{\alpha}{2}}(n-1)$ or $k > k_{\frac{\alpha}{2}}(n-1)$ null hypotheses H_0 is

rejected with an acceptable risk of a Type I error less than α .

$$\text{C2) } H_0 : \sigma^2 \leq \sigma_0^2 \quad H_1 : \sigma^2 > \sigma_0^2 .$$

First, we calculate the specific value $k = \frac{(n-1)s^2}{\sigma_0^2}$ where s^2 is the sample variance. Then we determine the critical value $k_\alpha(n-1)$ from the table for the χ^2 - distribution and make a decision:

- if $k \leq k_\alpha(n-1)$ null hypotheses H_0 fail to reject or
- if $k > k_\alpha(n-1)$ null hypotheses H_0 is rejected with an acceptable risk of a Type I error less than α .

$$\text{C3) } H_0 : \sigma^2 \geq \sigma_0^2 \quad H_1 : \sigma^2 < \sigma_0^2.$$

First, we calculate the specific value $k = \frac{(n-1)s^2}{\sigma_0^2}$ where s^2 is the sample variance. Then we determine the critical value $k_{1-\alpha}(n-1)$ from the table for the χ^2 - distribution and make a decision:

- if $k \geq k_{1-\alpha}(n-1)$ null hypotheses H_0 fail to reject or
- if $k < k_{1-\alpha}(n-1)$ null hypotheses H_0 is rejected with an acceptable risk of a Type I error less than α .

Example 7.1.10. Certain technological process is considered stable if the variance of a certain qualitative property is not greater than 2, otherwise the process is considered unstable. If from the sample with size 25 is obtained sample variance 1.8, what can you conclude about the instability of the technological process with acceptable risk $\alpha = 0.05$?

Solution: We are considering a sample of size $n = 25$ of normally distributed population, the sample variance is $s^2 = 1.8$.

We want to test if the variance of population is greater than 2, so we will use χ^2 - test statistic for the following hypotheses:

$$H_0 : \sigma^2 = 2 \quad H_1 : \sigma^2 > 2$$

and statistic $k = \frac{(n-1)S^2}{\sigma^2}$.

We calculate the specific value $k = \frac{(n-1)s^2}{\sigma_0^2} = \frac{24 \cdot 1.8}{2} = 21.6$ and from the table of χ^2 - distribution with 24 degrees of freedom, the critical value $k_\alpha(n-1) = k_{0.05}(24) = 36.42$.

Since $k = 21.6 < k_{0.05}(24) = 36.42$ the null hypothesis is not rejected. So we don't have enough reasons to consider that the technological process is unstable.

Example 7.1.11. The amount of protein in 250 gr packaging is controlled. For this purpose, a random sample of size 8 and a sample variance of 0.1608 is taken. Would be accepted the assumption that the distribution of packaging according to the protein quantity is normal that is the variance of the amount of protein is less than 0.16, with acceptable risk $\alpha = 0.025$?

Solution: We are considering a sample of size $n = 8$ of normally distributed population, the sample variance is $s^2 = 0.1608$.

We want to test if the variance of population is less than 0.16, so we will use χ^2 - test statistic for the following hypotheses:

$$H_0 : \sigma^2 = 0.16 \quad H_1 : \sigma^2 < 0.16$$

and statistic $k = \frac{(n-1)S^2}{\sigma^2}$.

We calculate the specific value $k = \frac{(n-1)s^2}{\sigma_0^2} = \frac{7 \cdot 0.1608}{0.16} = 7.035$ and from the table of χ^2 - distribution with 7 degrees of freedom, the critical value $k_{1-\alpha}(n-1) = k_{0.975}(7) = 1.69$.

Since $k = 7.035 > k_{0.975}(7) = 1.69$ the null hypothesis is rejected with an acceptable risk of a Type I error less than 0.025. So we have enough reasons to consider that the distribution of packaging according to the protein quantity is normal.

Example 7.1.12. The packaging machine fills packs with sugar with a nominal weight of 1000 grams and a variance of 15 grams. Upgrading the machine is intended the standard deviation to reduce by 20%. After the upgrade of the machine, test machine performance is achieved by using a random sample of 30 bags and sample standard deviation $s=10$. Let's assume that filling the packs is a normally distributed random variable. Does upgrading the machine achieve the desired reduction in variation of the weight of bags of sugar with acceptable risk $\alpha = 0.1$?

Solution: We are considering a sample size $n=30$ of normally distributed population, and the sample variance of the sample is $s^2 = 100$. Before the upgrading the standard deviation of the population is $\sigma = 15$ and after it we intend to reduce the standard deviation by 20% or equal to $\sigma_0 = \frac{80}{100}\sigma = 12$, so we want to test if the variance of population after the upgrading is equal to $\sigma_0^2 = 12^2 = 144$. Therefore we will use χ^2 - test statistic for the following hypotheses:

$$H_0 : \sigma^2 = 144, \quad H_1 : \sigma^2 \neq 144$$

and statistic $k = \frac{(n-1)S^2}{\sigma^2}$.

$$\text{We calculate the specific value } k = \frac{(n-1)s^2}{\sigma_0^2} = \frac{29 \cdot 100}{144} = 20.14$$

and from the table of χ^2 - distribution with 29 degrees of freedom the critical values $k_{\frac{\alpha}{2}}(n-1)$ and $k_{1-\frac{\alpha}{2}}(n-1)$ are:

$$k_{0.05}(29) = 42.56 \text{ and } k_{0.95}(29) = 17.71$$

$$\text{Since } k_{0.95}(29) = 17.71 < k = 20.14 < 42.56 = k_{0.05}(29)$$

the null hypothesis is not rejected, so we don't have enough reasons to say that upgrading the machine achieves the desired reduction in variation of the weight of bags of sugar.

7.2. TWO SAMPLE HYPOTHESIS TESTING

In addition to evaluating hypotheses regarding the value of a single population parameter, statistical testing includes techniques for comparing two datasets to determine if the same parameter differs between them. This process involves assessing whether two statistical sets have an identical parameter value or if a significant difference exists. For example, one might investigate whether customer satisfaction rates for two rival brands are equivalent or if the percentage of students passing an examination differs between two educational institutions. In these situations, the emphasis is placed not on estimating an individual parameter but rather on exploring the relationship of its values across different statistical sets.

7.2.1. TESTS FOR EQUALITY OF POPULATION MEANS WITH INDEPENDENT SAMPLES

Z – test

Let $X_1, X_2, \dots, X_{n_1}$ and $Y_1, Y_2, \dots, Y_{n_2}$ are two independent samples taken from normally distributed populations X with mean μ_1 and variance σ_1^2 and Y with mean μ_2 and variance σ_2^2 , where the variances σ_1^2 and σ_2^2 are known. The sample means $\bar{X}$ and $\bar{Y}$ are the point estimators of the population parameters μ_1 and μ_2 of the populations X and Y , respectively.

$$\text{Then } E(\bar{X}) = \mu_1, E(\bar{Y}) = \mu_2, \text{Var}(\bar{X}) = \frac{\sigma_1^2}{n_1} \text{ and } \text{Var}(\bar{Y}) = \frac{\sigma_2^2}{n_2}.$$

Then the difference of the sample means $\bar{X} - \bar{Y}$ is unbiased point estimator of the difference $\mu_1 - \mu_2$ of the population means, such that:

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2 \text{ and}$$

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + (-1)^2 Var(\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

This difference has a normal distribution, so hypothesis testing is conducted by applying the Z test, using the following test statistic

$$Z = \frac{(\bar{X} - \bar{Y}) - (m_1 - m_2)}{\sqrt{n_2\sigma_1^2 + n_1\sigma_2^2}} \sqrt{n_1n_2}.$$

The procedure for hypothesis testing of the difference of population means by independent samples with Z test is:

First, we check the *conditions for Z - test* – the samples are independent, known variances, the populations have normal distributions or the distribution is unknown but large samples, i.e. $n_1 \geq 30$ and $n_2 \geq 30$, the samples are taken from infinitely large populations, or if the samples are taken from finite populations with replacement, or if the samples are taken from finite populations without replacement but the sample size n_1 and n_2 are less than 5% of the population size N_1 and N_2 .

Then we state the null and alternative hypotheses

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2$$

$$H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2$$

and we use the statistic $Z = \frac{\bar{X} - \bar{Y}}{\sqrt{n_2\sigma_1^2 + n_1\sigma_2^2}} \sqrt{n_1n_2}.$

There are three cases:

C1) $H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2 :$

First, we calculate the specific value $\bar{z} = \frac{|\bar{x} - \bar{y}|}{\sqrt{n_2\sigma_1^2 + n_1\sigma_2^2}} \sqrt{n_1n_2}$,

where $\bar{x}$ and $\bar{y}$ are the sample means, then we determine the critical value $z_{\alpha/2}$ such that $P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}$, i.e., $\Phi_0 \left(z_{\alpha/2} \right) = 1 - \frac{\alpha}{2}$ and make a decision

- if $\bar{z} \leq z_{\alpha/2}$ the null hypotheses H_0 is not rejected or
- if $\bar{z} > z_{\alpha/2}$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

$$\text{C2) } H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

We calculate the specific value $\bar{z} = \frac{\bar{x} - \bar{y}}{\sqrt{n_2\sigma_1^2 + n_1\sigma_2^2}} \sqrt{n_1n_2}$, where

$\bar{x}$ and $\bar{y}$ are the sample means, then we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \geq z_\alpha \right\} = \alpha$, i.e., $\Phi_0 \left(z_\alpha \right) = 1 - \alpha$ and make a decision

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 or
- if $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C3) } H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

We calculate the specific value $\bar{z} = \frac{\bar{x} - \bar{y}}{\sqrt{n_2\sigma_1^2 + n_1\sigma_2^2}} \sqrt{n_1n_2}$, where

$\bar{x}$ and $\bar{y}$ are the sample means, then we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \leq -z_\alpha \right\} = \alpha$, i.e., $\Phi_0 \left(z_\alpha \right) = 1 - \alpha$ and make a decision

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 or
- if $\bar{z} < -z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

If the sample is taken from a finite population without replacement but the sample size n is greater than 5% of the population size N we use the statistic $Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} \cdot \frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2} \cdot \frac{N_2 - n_2}{N_2 - 1}}}$,

and the procedure for the test is:

First, we check the *conditions for Z - test* – the samples are independent, known variances, the populations have normal distributions or the distribution is unknown but large samples, i.e. $n_1 \geq 30$ and $n_2 \geq 30$, the sample is taken from a finite population without replacement but the sample size n is greater than 5% of the populations size N_1 and N_2 .

Then we state the null and alternative hypotheses

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2$$

$$H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2$$

and we use this test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} \cdot \frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2} \cdot \frac{N_2 - n_2}{N_2 - 1}}}.$$

There are three cases:

$$C1) \quad H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2 :$$

$$\text{We calculate the specific value } \bar{z} = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sigma_1^2}{n_1} \cdot \frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2} \cdot \frac{N_2 - n_2}{N_2 - 1}}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, then we determine the critical

value $z_{\alpha/2}$ such that $P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}$, i.e., $\Phi_0 \left(z_{\alpha/2} \right) = 1 - \frac{\alpha}{2}$ and make a decision

- if $\bar{z} \leq z_{\alpha/2}$ the null hypotheses H_0 is not rejected or
- if $\bar{z} > z_{\alpha/2}$ the null hypotheses H_0 is rejected with an acceptable risk of a Type I error less α .

$$\text{C2) } H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

$$\text{We calculate the specific value } \bar{z} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} \cdot \frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2} \cdot \frac{N_2 - n_2}{N_2 - 1}}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, then we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \geq z_\alpha \right\} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 or
- if $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C3) } H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

$$\text{We calculate the specific value } \bar{z} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} \cdot \frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2} \cdot \frac{N_2 - n_2}{N_2 - 1}}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, then we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \leq -z_\alpha \right\} = \alpha$, i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and make a decision

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 or
- if $\bar{z} < -z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

***t* – test**

The above assumptions, which are necessary for the *Z* - test to be used in testing, are often not satisfied in practice, primarily because the population variances σ_1 and σ_2 are not known. In this situation, first, the standard deviations σ_1 and σ_2 are estimated, using the sample variance s_1^2 and s_2^2 . In doing so, we distinguish two cases:

First case, when it is assumed that the variances of the two populations are equal, i.e. $\sigma_1^2 = \sigma_2^2$ and

Second case, when it is assumed that the two population variances are different, i.e. $\sigma_1^2 \neq \sigma_2^2$.

First, we consider the first case, i.e. if $\sigma_1^2 = \sigma_2^2$, and we obtain new statistic for *t* - test:

$$t_{n_1+n_2-2} = \frac{\bar{X} - \bar{Y}}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2} (n_1 + n_2 - 2)},$$

that has Student's distribution with $n_1 + n_2 - 2$ degrees of freedom.

The procedure for hypothesis testing of the difference in population means of two populations with unknown equal variances, when the samples are independent, using the *t* – test is:

We check the *conditions for the t – test* - independent samples, the populations have normal distributions with unknown but equal variances $\sigma_1^2 = \sigma_2^2$, and the samples are small i.e., $n_1 < 30$ and $n_2 < 30$.

Then we choose the null and alternative hypotheses:

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2$$

$$H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2$$

and we use this test statistic

$$t_{n_1+n_2-2} = \frac{\bar{X} - \bar{Y}}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2} (n_1 + n_2 - 2)}.$$

There are three cases:

$$C1) \quad H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

We calculate the specific value

$$\bar{t}_{n_1+n_2-2} = \frac{|\bar{x} - \bar{y}|}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2} (n_1 + n_2 - 2)},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value $t_\alpha (n_1 + n_2 - 2)$ from the Student's distribution table for the two-tailed test and we make a decision:

- if $\bar{t}_{n_1+n_2-2} \leq t_\alpha (n_1 + n_2 - 2)$ fail to reject H_0 or
- if $\bar{t}_{n_1+n_2-2} > t_\alpha (n_1 + n_2 - 2)$ reject H_0 with an acceptable risk of a Type I error less α .

$$C2) \quad H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

We calculate the specific value

$$\bar{t}_{n_1+n_2-2} = \frac{\bar{x} - \bar{y}}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2} (n_1 + n_2 - 2)},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value $t_\alpha (n_1 + n_2 - 2)$ from the Student's distribution table for the one-tailed test and we make a decision:

- if $\bar{t}_{n_1+n_2-2} \leq t_\alpha (n_1 + n_2 - 2)$ fail to reject H_0 or
- if $\bar{t}_{n_1+n_2-2} > t_\alpha (n_1 + n_2 - 2)$ reject H_0 with an acceptable risk of a Type I error less α .

$$C3) \quad H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

We calculate the specific value

$$\bar{t}_{n_1+n_2-2} = \frac{\bar{x} - \bar{y}}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} (n_1 + n_2 - 2),$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value $t_\alpha (n_1 + n_2 - 2)$ from the Student's distribution table for the one-tailed test and we make a decision:

- if $\bar{t}_{n_1+n_2-2} \geq -t_\alpha (n_1 + n_2 - 2)$ fail to reject H_0 or
- if $\bar{t}_{n_1+n_2-2} < -t_\alpha (n_1 + n_2 - 2)$ reject H_0 with an acceptable risk of a Type I error less α .

Example 7.2.1 Is there a significant difference between the examination results in the summer and winter semesters considering the data below:

Summer semester exam points	Winter semester exam points
52	53
61	71
40	38
46	34
50	68
56	68
44	46
47	41
70	38
40	23
65	28
38	
68	

Acceptable risk 0.05 and if we assume that the variances are unknown but equal, and examination results are normally distributed.

Solution. We consider independent samples with size $n_1 = 13 < 30$ и $n_2 = 11 < 30$, from normally distributed populations and unknown but equal variances. We want to test if there is significant difference between the examination results in the summer and winter semesters, so we test the following hypotheses:

$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$ with the statistic:

$$\bar{t}_{n_1+n_2-2} = \frac{|\bar{x} - \bar{y}|}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2} (n_1 + n_2 - 2)}.$$

The sample mean of the summer semester exam is:

$$\bar{x} = \frac{52 + 61 + 40 + 46 + 1... + 65 + 38 + 38}{13} = 52.077$$

and the sample variance of the summer semester exam is:

$$\begin{aligned} s_1^2 &= \frac{1}{n_1 - 1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n \bar{x}^2 \right) = \\ &= \frac{1}{12} (52^2 + 61^2 + 40^2 + 46^2 + 1^2 + ... + 65^2 + 38^2 + 38^2 - 13 \cdot 52.077^2) = \\ &= 121.57 \end{aligned}$$

The sample mean of the winter semester exam is:

$$\bar{y} = \frac{53 + 71 + 38 + 34 + ... + 23 + 28}{11} = 46.18$$

and the sample variance of the winter semester exam is:

$$\begin{aligned} s_2^2 &= \frac{1}{n_2 - 1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n \bar{x}^2 \right) = \\ &= \frac{1}{10} (53^2 + 71^2 + 38^2 + ... + 23^2 + 28^2 - 11 \cdot 46.18^2) = 279.35 \end{aligned}$$

The specific value is:

$$\bar{t}_{22} = \frac{|52.077 - 46.18|}{\sqrt{13 \cdot 121.57^2 + 11 \cdot 279.35^2}} \sqrt{\frac{13 \cdot 11}{13 + 11}} (13 + 11 - 2) = 0.066$$

and the critical values from the Student's distribution table two-tailed test for $\alpha = 0.05$ and $n_1 + n_2 - 2 = 22$ degrees of freedom is $t_{0.05}(22) = 2.074$.

Since $\bar{t}_{22} = 0.066 < 2.074 = t_{0.05}(22)$ we do not reject the null hypothesis, so there is no evidence for significant difference between the examination results in the summer and winter semesters.

In the last test, we assumed that the two population variances are equal, but if they are not i.e., $\sigma_1^2 \neq \sigma_2^2$ when the samples are small we use this t -test statistic:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2}.$$

It has Student's distribution with

$$k = \left[\frac{(n_2 s_1^2 + n_1 s_2^2)^2 (n_1 - 1)(n_2 - 1)}{(n_2 - 1)(n_2 s_1^2)^2 + (n_1 - 1)(n_1 s_2^2)^2} \right] \text{ degrees of freedom.}$$

The procedure for hypothesis testing of the difference in the population means of two populations with unknown different variances, when the samples are independent, using the t -test is:

We check the *conditions for the Z-test* - independent samples, the populations have normal distributions with unknown but different variances i.e., $\sigma_1^2 \neq \sigma_2^2$, and the samples are small i.e., $n_1 < 30$ and $n_2 < 30$.

Then we choose the null and alternative hypotheses:

$$\begin{array}{ll} H_0 : \mu_1 = \mu_2, & H_1 : \mu_1 \neq \mu_2 \\ H_0 : \mu_1 \leq \mu_2, & H_1 : \mu_1 > \mu_2 \\ H_0 : \mu_1 \geq \mu_2, & H_1 : \mu_1 < \mu_2 \end{array}$$

and we use this test statistic

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2} .$$

There are three cases:

$$\text{C1) } H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2 .$$

We calculate the specific value

$$\bar{t} = \frac{|\bar{x} - \bar{y}|}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2} ,$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value $t_\alpha(k)$ from the Student's distribution table for the two-tailed test and degrees of freedom

$$k = \left[\frac{(n_2 s_1^2 + n_1 s_2^2)^2 (n_1 - 1)(n_2 - 1)}{(n_2 - 1)(n_2 s_1^2)^2 + (n_1 - 1)(n_1 s_2^2)^2} \right]$$

and we make a decision:

- if $\bar{t} \leq t_\alpha(k)$ fail to reject H_0 or

- if $\bar{t} > t_\alpha(k)$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C2) } H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2 .$$

We calculate the specific value

$$\bar{t} = \frac{\bar{x} - \bar{y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2} ,$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value $t_\alpha(k)$ from the Student's distribution table for the one-tailed test and degrees of freedom

$$k = \left[\frac{(n_2 s_1^2 + n_1 s_2^2)^2 (n_1 - 1)(n_2 - 1)}{(n_2 - 1)(n_2 s_1^2)^2 + (n_1 - 1)(n_1 s_2^2)^2} \right]$$

and we make a decision:

- if $\bar{t} \leq t_\alpha(k)$ fail to reject H_0 or
- if $\bar{t} > t_\alpha(k)$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C3) } H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

We calculate the specific value

$$\bar{t} = \frac{\bar{x} - \bar{y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value $t_\alpha(k)$ from the Student's distribution table for the one-tailed test and degrees of freedom

$$k = \left[\frac{(n_2 s_1^2 + n_1 s_2^2)^2 (n_1 - 1)(n_2 - 1)}{(n_2 - 1)(n_2 s_1^2)^2 + (n_1 - 1)(n_1 s_2^2)^2} \right]$$

and we make a decision:

- if $\bar{t} \geq -t_\alpha(k)$ fail to reject H_0 or
- if $\bar{t} < -t_\alpha(k)$ reject H_0 with an acceptable risk of a Type I error less α .

Example 7.2.2. Is there a significant difference between the examination results in the summer and winter semesters considering the data below:

Summer semester exam points	Winter semester exam points
52	53
61	71
40	38
46	34
50	68
56	68
44	46
47	41
70	38
40	23
65	28
38	
68	

Acceptable risk 0.05 and if we assume that the variances are unknown and different, and examination results are normally distributed.

Solution. We consider independent samples with size $n_1 = 13 < 30$ и $n_2 = 11 < 30$, from normally distributed populations and unknown but different variances. We want to test if there is significant

difference between the examination results in the summer and winter semesters, so we test the following hypotheses:

$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$ with the statistic:

$$\bar{t} = \frac{\bar{x} - \bar{y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2}.$$

The sample mean of the summer semester exam is:

$$\bar{x} = \frac{52 + 61 + 40 + 46 + 1 + \dots + 65 + 38 + 38}{13} = 52.077$$

and the sample variance of the summer semester exam is:

$$\begin{aligned} s_1^2 &= \frac{1}{n_1 - 1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n \bar{x}^2 \right) = \\ &= \frac{1}{12} (52^2 + 61^2 + 40^2 + 46^2 + 1^2 + \dots + 65^2 + 38^2 + 38^2 - 13 \cdot 52.077^2) = \\ &= 121.57 \end{aligned}$$

The sample mean of the winter semester exam is:

$$\bar{y} = \frac{53 + 71 + 38 + 34 + \dots + 23 + 28}{11} = 46.18$$

and the sample variance of the winter semester exam is:

$$\begin{aligned} s_2^2 &= \frac{1}{n_2 - 1} \left(\sum_{i=1}^k f_i \cdot x_i^2 - n \bar{x}^2 \right) = \\ &= \frac{1}{10} (53^2 + 71^2 + 38^2 + \dots + 23^2 + 28^2 - 11 \cdot 46.18^2) = 279.35 \end{aligned}$$

The specific value is:

$$\bar{t} = \frac{|52.077 - 46.18|}{\sqrt{11 \cdot 121.57^2 + 13 \cdot 279.35^2}} \sqrt{13 \cdot 11} = 0.065$$

and the critical values from the Student's distribution table two-tailed test for $\alpha = 0.05$ and

$$k = \left[\frac{(n_2 s_1^2 + n_1 s_2^2)^2 (n_1 - 1)(n_2 - 1)}{(n_2 - 1)(n_2 s_1^2)^2 + (n_1 - 1)(n_1 s_2^2)^2} \right] =$$

$$= \left[\frac{(11 \cdot 121.57^2 + 13 \cdot 279.35^2)^2 (13 - 1)(11 - 1)}{(11 - 1)(11 \cdot 121.57^2)^2 + (13 - 1)(13 \cdot 279.35^2)^2} \right] = 13$$

degrees of freedom is $t_{0.05}(13) = 1.771$.

Since $\bar{t} = 0.065 < 1.771 = t_{0.05}(13)$ we do not reject the null hypothesis, so there is no evidence for significant difference between the examination results in the summer and winter semesters.

As the sample size increases, the Student's t -distribution approaches the normal distribution. Consequently, for significantly large values, enough $n_1 \geq 30$ and $n_2 \geq 30$ the critical values obtained by t -test are roughly equal to the critical values obtained by the z -test. Thus, it is generally accepted that if the statistical sets are normally distributed with an unknown variance and the sample sizes are quite large, the testing can be conducted using the z -test instead of the t -test.

The procedure for hypothesis testing of the difference between the two populations means, when the population variances are unknown but equal and the samples are large is as follows:

We check the *conditions for the Z-test* - independent samples, the populations have normal distributions with unknown but equal variances $\sigma_1^2 = \sigma_2^2$, and the samples are large i.e., $n_1 \geq 30$ and $n_2 \geq 30$.

Then we choose the null and alternative hypotheses:

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2$$

$$H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2$$

and we use this test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2} (n_1 + n_2 - 2)}.$$

There are three cases:

$$C1) \quad H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

We calculate the specific value

$$\bar{z} = \frac{|\bar{x} - \bar{y}|}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2} (n_1 + n_2 - 2)},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value $z_{\alpha/2}$ such that

$$P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}, \text{ i.e., } \Phi_0(z_{\alpha/2}) = 1 - \frac{\alpha}{2} \text{ and we make a decision:}$$

- if $\bar{z} \leq z_{\alpha/2}$ fail to reject H_0 or

- if $\bar{z} > z_{\alpha/2}$ reject H_0 with an acceptable risk of a Type I error

less α .

$$C2) \quad H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

We calculate the specific value

$$\bar{z} = \frac{\bar{x} - \bar{y}}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2} (n_1 + n_2 - 2)},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \geq z_\alpha \right\} = 1 - \alpha$ i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and we make a decision:

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 or

- if $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error

less α .

$$C3) \quad H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

We calculate the specific value

$$\bar{z} = \frac{\bar{x} - \bar{y}}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} (n_1 + n_2 - 2),$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value z_α such that $P_{H_0} \{ \bar{z} \leq -z_\alpha \} = 1 - \alpha$ i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and we make a decision:

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 or
- if $\bar{z} < -z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

The procedure for hypothesis testing of the difference between the two populations means if the population variances are unknown but different, and the samples are large is as follows:

We check the *conditions for the Z – test* - independent samples, the populations have normal distributions with unknown but different variances $\sigma_1^2 \neq \sigma_2^2$, and the samples are large i.e., $n_1 \geq 30$ and $n_2 \geq 30$.

Then we choose the null and alternative hypotheses:

$$\begin{array}{ll} H_0 : \mu_1 = \mu_2, & H_1 : \mu_1 \neq \mu_2 \\ H_0 : \mu_1 \leq \mu_2, & H_1 : \mu_1 > \mu_2 \\ H_0 : \mu_1 \geq \mu_2, & H_1 : \mu_1 < \mu_2 \end{array}$$

and we use this test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2}$$

There are three cases:

$$C1) \quad H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

We calculate the specific value

$$\bar{z} = \frac{|\bar{x} - \bar{y}|}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value $z_{\alpha/2}$ such that

$$P_{H_0} \left\{ \bar{z} \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}, \text{ i.e., } \Phi_0 \left(z_{\alpha/2} \right) = 1 - \frac{\alpha}{2} \text{ and we make a decision:}$$

- if $\bar{z} \leq z_{\alpha/2}$ fail to reject H_0 or

- if $\bar{z} > z_{\alpha/2}$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C2) } H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

We calculate the specific value

$$\bar{z} = \frac{\bar{x} - \bar{y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \geq z_\alpha \right\} = 1 - \alpha$ i.e., $\Phi_0 \left(z_\alpha \right) = 1 - \alpha$ and we make a decision:

- if $\bar{z} \leq z_\alpha$ fail to reject H_0 or

- if $\bar{z} > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C3) } H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

We calculate the specific value

$$\bar{z} = \frac{\bar{x} - \bar{y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_1^2 and s_2^2 are the sample variances and we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z} \leq -z_\alpha \right\} = 1 - \alpha$ i.e., $\Phi_0 \left(z_\alpha \right) = 1 - \alpha$ and we make a decision:

- if $\bar{z} \geq -z_\alpha$ fail to reject H_0 or

- if $\bar{z} < -z_{\alpha}$ reject H_0 with an acceptable risk of a Type I error less α .

Example 7.2.3. The personnel officer of a large corporation claimed that college graduates applying for jobs this year tend to have higher grade point averages than those who applied in the previous year. Samples from the group of applicants gave the following results: the sample mean and the sample standard deviation of the grade point in the current year 2.98 and 0.4, respectively; the sample mean and the sample standard deviation of the grade point in the previous year 2.8 and 0.5, respectively. The samples have size 60 and 52 respectively. Acceptable risk 0.01.

Solution. These are independent samples with size $n_1 = 60 > 30$ и $n_2 = 52 > 30$, sample means $\bar{x} = 2.98$ и $\bar{y} = 2.8$ and sample standard deviations $s_1 = 0.4$ and $s_2 = 0.5$, taken from infinitely large populations with different variances. We aim to test whether college graduates applying for jobs with the firm this year have higher grade point averages than those who applied in the previous year. Therefore, we test the hypotheses:

$$H_0 : \mu_1 \leq \mu_2 \quad H_1 : \mu_1 > \mu_2, \text{ and test statistic is } Z = \frac{\bar{X} - \bar{Y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2}.$$

We calculate the specific value

$$\bar{z} = \frac{\bar{x} - \bar{y}}{\sqrt{n_2 s_1^2 + n_1 s_2^2}} \sqrt{n_1 n_2} = \frac{2.98 - 2.8}{\sqrt{52 \cdot 0.4^2 + 60 \cdot 0.5^2}} \sqrt{60 \cdot 52} = 2.082$$

The critical value $Z_{0,01}$ is calculated as, since

$$\Phi_0(z_{0.05}) = 1 - 0.05 = 0.95 = \Phi_0(1.65) \text{ so } z_{0.05} = 1.65.$$

Since $\bar{z} = 2.082 > 1.65 = z_{0.05}$ we reject the null hypothesis with an acceptable risk of a Type I error less of 0.05, so the college graduates applying for jobs with their firm in the current year tended to have higher grade point averages than those applying in the previous year.

7.2.2. **COMPARING POPULATION MEANS WITH DEPENDENT SAMPLES**

Dependent sample tests, also known as matched-pair tests or paired tests in statistics, are utilized to analyze two related datasets where the observations are interconnected or inherently linked. These tests determine if there is a statistically significant difference between two measurements taken from the same subjects under varying conditions. Typical examples include studies that measure changes before and after an intervention, such as blood pressure readings before and following drug administration, assessing student performance before and after a training initiative, exploring employee productivity before and after the introduction of a new workflow system, evaluating the effects of two different teaching methods on a single group of students, or comparing drivers' reaction times before and after caffeine consumption. Due to the correlation between data points, paired tests such as the paired t -test for means or the Wilcoxon signed-rank test for non-parametric cases—account for this dependency, leading to more precise conclusions compared to independent sample tests.

Let's consider the joint probability distribution $(X_i, Y_i), i = 1, 2, \dots, n$ of a sample taken from two dimensional normal distribution (population) (X, Y) , such μ_1, σ_1^2 and μ_2, σ_2^2 are the mean and variance of X and Y , respectively.

The new random variable

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$$

where $Z_i = X_i - Y_i, i = 1, 2, \dots, n$ taken from $Z = X - Y$ has mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_Z^2}{n}$, with σ_Z^2 being the variance of $Z = X - Y$.

Then, the statistic $\frac{\bar{Z}}{\sigma_z} \sqrt{n}$ has Student's distribution with $n-1$ degrees of freedom if $n < 30$ or standard normal distribution if $n \geq 30$. And if the variance σ_Z^2 is unknown, first, we form the sample variance

S_z of the population variance σ_z , then the statistic $\frac{\bar{Z}}{S_z} \sqrt{n}$ has Student's distribution with $n-1$ degrees of freedom if $n < 30$ or standard normal distribution if $n \geq 30$. These statistics are used for hypothesis testing of the difference between the means of two dependent populations.

The procedure for hypothesis testing of the difference in the population means with dependent small samples by the t -test is:

We check the *conditions for the Z-test* - dependent samples, the populations X and Y have normal distributions, the variance σ_z^2 of $Z = X - Y$ is known, and the samples are small i.e., $n < 30$.

Then we choose the null and alternative hypotheses:

$$\begin{array}{ll} H_0 : \mu_1 = \mu_2, & H_1 : \mu_1 \neq \mu_2 \\ H_0 : \mu_1 \leq \mu_2, & H_1 : \mu_1 > \mu_2 \\ H_0 : \mu_1 \geq \mu_2, & H_1 : \mu_1 < \mu_2 \end{array}$$

and we use this test statistic

$$t = \frac{\bar{Z}}{\sigma_z} \sqrt{n}.$$

There are three cases:

$$\text{C1) } H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

We calculate the specific value

$$\bar{t} = \frac{|\bar{Z}|}{\sigma_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean and we determine the critical value $t_\alpha(n-1)$ from the Student's distribution table for the two-tailed test and $n-1$ degrees of freedom and we make a decision:

- if $\bar{t} \leq t_{\alpha}(k)$ fail to reject H_0 or
- if $\bar{t} > t_{\alpha}(k)$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C2) } H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

We calculate the specific value

$$\bar{t} = \frac{\bar{Z}}{\sigma_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean and we determine the critical value $t_{\alpha}(n-1)$ from the Student's distribution table for the one-tailed test and $n-1$ degrees of freedom and we make a decision:

- if $\bar{t} \leq t_{\alpha}(k)$ fail to reject H_0 or
- if $\bar{t} > t_{\alpha}(k)$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C3) } H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

We calculate the specific value

$$\bar{t} = \frac{\bar{Z}}{\sigma_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean and we determine the critical value $t_{\alpha}(n-1)$ from the Student's distribution table for the one-tailed test and $n-1$ degrees of freedom and we make a decision:

- if $\bar{t} \geq -t_{\alpha}(k)$ fail to reject H_0 or
- if $\bar{t} < -t_{\alpha}(k)$ reject H_0 with an acceptable risk of a Type I error less α .

Usually the variance σ_Z^2 is unknown and in this situation the procedure for hypothesis testing of the difference in the population means with dependent small samples by the t -test is:

We check the *conditions for the Z – test* - dependent samples, the populations X and Y have normal distributions, the variance σ_Z^2 of $Z = X - Y$ is unknown, and the sample is small i.e., $n < 30$.

Then we choose the null and alternative hypotheses:

$$\begin{array}{ll} H_0 : \mu_1 = \mu_2, & H_1 : \mu_1 \neq \mu_2 \\ H_0 : \mu_1 \leq \mu_2, & H_1 : \mu_1 > \mu_2 \\ H_0 : \mu_1 \geq \mu_2, & H_1 : \mu_1 < \mu_2 \end{array}$$

and we use this test statistic

$$t = \frac{\bar{Z}}{S_z} \sqrt{n}.$$

where S_z is the sample variance.

There are three cases:

$$C1) \quad H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

We calculate the specific value

$$\bar{t} = \frac{|\bar{Z}|}{S_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean, S_z is the sample variance and we determine the critical value $t_\alpha(n-1)$ from the Student's distribution table for the two-tailed test and $n-1$ degrees of freedom and we make a decision:

- if $\bar{t} \leq t_\alpha(k)$ fail to reject H_0 or

- if $\bar{t} > t_{\alpha}(k)$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C2) } H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

We calculate the specific value

$$\bar{t} = \frac{\bar{Z}}{S_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean, S_z is the sample variance and we determine the critical value $t_{\alpha}(n-1)$ from the Student's distribution table for the one-tailed test and $n-1$ degrees of freedom and we make a decision:

- if $\bar{t} \leq t_{\alpha}(k)$ fail to reject H_0 or

- if $\bar{t} > t_{\alpha}(k)$ reject H_0 with an acceptable risk of a Type I error less α .

$$\text{C3) } H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

We calculate the specific value

$$\bar{t} = \frac{\bar{Z}}{S_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean, S_z is the sample variance and we determine the critical value $t_{\alpha}(n-1)$ from the Student's distribution table for the one-tailed test and $n-1$ degrees of freedom and we make a decision:

- if $\bar{t} \geq -t_{\alpha}(k)$ fail to reject H_0 or

- if $\bar{t} < -t_{\alpha}(k)$ reject H_0 with an acceptable risk of a Type I error less α .

Example 7.2.4. Suppose a sample of $n = 20$ students were given a diagnostic test before studying a special course and then again after completing the course.

The following results were obtained

Student	Pre-course score	Post-course score	Difference
1	18	22	-4
2	21	25	-4
3	16	17	-1
4	22	24	-2
5	19	16	+3
6	24	29	-5
7	17	20	-3
8	21	23	-2
9	23	19	+4
10	18	20	-2
11	14	15	-1
12	16	15	+1
13	16	18	-2
14	19	26	-7
15	18	18	0
16	20	24	-4
17	12	18	-6
18	22	25	-3
19	15	19	-4
20	17	16	+1

We want to determine whether teaching generally leads to improvements in students' knowledge and skills, as measured by test scores. Using our sample of students, we can assess the overall impact of the course, $\alpha = 0.05$

Solution: Let the random variable X is the pre-course scores and Y is the post-course scores, with means μ_1 and μ_2 , respectively. We consider dependent samples with size $n = 20 < 30$. Since we aim to determine whether teaching generally improves students' knowledge and skills, we consider the following hypotheses:

$$H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2$$

and we use this test statistic

$$t = \frac{\bar{Z}}{S_z} \sqrt{n}.$$

where S_z is the sample variance.

Therefore, we calculate the specific value $\bar{t} = \frac{\bar{z}}{s_z} \sqrt{n}$, where

$$\begin{aligned} \bar{z} &= \frac{-7 + (-6) + (-5) + 4 \cdot (-4) + 2 \cdot (-3) + 4 \cdot (-2) + 2 \cdot (-1) + 2 \cdot 1 + 0 + 3 + 4}{20} = \\ &= -2.05 \end{aligned}$$

$$\begin{aligned} s_z^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n (x_i - y_i)^2 - n\bar{z}_n^2 \right) = \\ &= \frac{1}{19} \left(\left((-7)^2 + (-6)^2 + (-5)^2 + 4 \cdot (-4)^2 + 2 \cdot (-3)^2 + \right. \right. \\ &\quad \left. \left. 4 \cdot (-2)^2 + 2 \cdot (-1)^2 + 2 \cdot 1^2 + 0^2 + 3^2 + 4^2 \right) - \right. \\ &\quad \left. - 20 \cdot (-2.05)^2 \right) = \frac{1}{19} (237 - 84.05) = 8.05 \end{aligned}$$

$$s_z = \sqrt{8.05} = 2.837$$

$$\text{so, } \bar{t} = \frac{-2.05}{2.837} \sqrt{20} = -3.23$$

The critical value from the Student's distribution table for the one-tailed test for 19 degrees of freedom is $t_{\alpha}(n-1) = t_{0.05}(19) = 1.729$.

Since $\bar{t} = -3.23 < -1.729 = -t_{0.05}(19)$ the null hypothesis is rejected with an acceptable risk of a Type I error less, so in general, if the teaching leads to improvements in students' knowledge/skills.

Example 7.2.5. A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked four of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight each participant could lift was measured again. The data are as follows:

Weight (in pounds)	Player 1	Player 2	Player 3	Player 4
Amount of weight lifted prior to the class	230	206	308	358
Amount of weight lifted after the class	241	296	300	350

The coach wants to know if the strength development class makes his players stronger, on average. $\alpha = 0.05$

Solution: Let the random variable X is the amount of weight lifted after the class and Y is the amount of weight lifted prior to the class, with means μ_1 and μ_2 , respectively. The differences are: +11, +90, -8 and -8. Assume the differences have a normal distribution. We consider dependent samples with size $n = 4 < 30$ and since the coach wants to know if the strength development class makes his players stronger we consider the following hypotheses:

$$H_0 : \mu_1 < \mu_2, \quad H_1 : \mu_1 \geq \mu_2$$

and we use this test statistic

$$t = \frac{\bar{Z}}{S_z} \sqrt{n}.$$

where S_z is the sample variance.

Therefore, we calculate the specific value $\bar{t} = \frac{\bar{Z}}{s_z} \sqrt{n}$, where

$$\bar{Z} = \frac{11+90-8-8}{4} = 21.25$$

$$\begin{aligned} s_z^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n (x_i - y_i)^2 - n\bar{Z}^2 \right) = \\ &= \frac{1}{3} \left((90^2 + 11^2 + 2 \cdot (-8)^2) - 4 \cdot 21.25^2 \right) = \frac{1}{3} (8349 - 1806.25) = 2180.917 \end{aligned}$$

$$s_z = \sqrt{2180.917} = 46.7$$

$$\text{so, } \bar{t} = \frac{21.25}{46.7} \sqrt{4} = 0.91$$

The critical value from the Student's distribution table for the one-tailed test for 3 degrees of freedom is $t_{\alpha}(n-1) = t_{0.05}(3) = 2.353$.

Since $\bar{t} = 0.91 < 2.353 = t_{0.05}(3)$ the null hypothesis is not rejected, so there is no evidence that the strength development class makes the players stronger, on average.

When the sample size is large, the Student's distribution approximates the normal distribution. Therefore, for hypothesis testing with a large sample, we will use a statistic based on the normal distribution.

The procedure for hypothesis testing of the difference in the population means with dependent large samples by the Z -test is:

We check the *conditions for the Z – test* - dependent samples, the populations X and Y have normal distributions, the variance σ_Z^2 of $Z = X - Y$ is known, and the samples are large i.e., $n \geq 30$.

Then we choose the null and alternative hypotheses:

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2$$

$$H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2$$

and we use this test statistic

$$Z_s = \frac{\bar{Z}}{\sigma_z} \sqrt{n}.$$

There are three cases:

$$C1) \quad H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

We calculate the specific value

$$\bar{z}_s = \frac{|\bar{Z}|}{\sigma_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean, and we determine the critical value $z_{\alpha/2}$

such that $P_{H_0} \left\{ \bar{z}_s \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}$, i.e, $\Phi_0 \left(z_{\alpha/2} \right) = 1 - \frac{\alpha}{2}$ and we make a decision:

- if $\bar{z}_s \leq z_{\alpha/2}$ fail to reject H_0 or

- if $\bar{z}_s > z_{\alpha/2}$ reject H_0 with an acceptable risk of a Type I error

less α .

$$C2) \quad H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

We calculate the specific value

$$\bar{z}_s = \frac{\bar{Z}}{\sigma_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean, and we determine the critical value z_α such that $P_{H_0} \{ \bar{Z}_s \geq z_\alpha \} = 1 - \alpha$ i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and we make a decision:

- if $\bar{Z}_s \leq z_\alpha$ fail to reject H_0 or
- if $\bar{Z}_s > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

$$C3) \ H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

We calculate the specific value

$$\bar{Z}_s = \frac{\bar{Z}}{\sigma_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean, and we determine the critical value z_α such that $P_{H_0} \{ \bar{Z}_s \geq z_\alpha \} = 1 - \alpha$ i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and we make a decision:

- if $\bar{Z}_s \geq -z_\alpha$ fail to reject H_0 or
- if $\bar{Z}_s < -z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

If the σ_z^2 is unknown, we use its estimator for the sample variance S_z^2 , and the procedure for hypothesis testing of the difference in the population means with dependent large samples by the Z -test is:

We check the *conditions for the Z -test* - dependent samples, the populations X and Y have normal distributions, the variance σ_z^2 of $Z = X - Y$ is unknown, and the samples are large i.e., $n \geq 30$.

Then we choose the null and alternative hypotheses:

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2$$

$$H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2$$

and we use this test statistic

$$Z_s = \frac{\bar{Z}}{S_z} \sqrt{n}.$$

There are three cases:

$$C1) \quad H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2.$$

We calculate the specific value

$$\bar{z}_s = \frac{|\bar{Z}|}{S_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean, and S_z is the sample variance and we determine the critical value $z_{\alpha/2}$ such that $P_{H_0} \left\{ \bar{z}_s \geq z_{\alpha/2} \right\} = \frac{\alpha}{2}$, i.e.,

$$\Phi_0 \left(z_{\alpha/2} \right) = 1 - \frac{\alpha}{2} \text{ and we make a decision:}$$

- if $\bar{z}_s \leq z_{\alpha/2}$ fail to reject H_0 or

- if $\bar{z}_s > z_{\alpha/2}$ reject H_0 with an acceptable risk of a Type I error less α .

$$C2) \quad H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2.$$

We calculate the specific value

$$\bar{z}_s = \frac{\bar{Z}}{S_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean, and S_z is the sample variance and we determine the critical value z_α such that $P_{H_0} \left\{ \bar{z}_s \geq z_\alpha \right\} = 1 - \alpha$ i.e., $\Phi_0 \left(z_\alpha \right) = 1 - \alpha$ and we make a decision:

- if $\bar{z}_s \leq z_\alpha$ fail to reject H_0 or

- if $\bar{z}_s > z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

$$C3) H_0 : \mu_1 \geq \mu_2, \quad H_1 : \mu_1 < \mu_2.$$

We calculate the specific value

$$\bar{z}_s = \frac{\bar{Z}}{S_z} \sqrt{n},$$

where $\bar{Z}$ is the sample mean and S_z is the sample variance and we determine the critical value z_α such that $P_{H_0} \{ \bar{z}_s \geq z_\alpha \} = 1 - \alpha$ i.e., $\Phi_0(z_\alpha) = 1 - \alpha$ and we make a decision:

- if $\bar{z}_s \geq -z_\alpha$ fail to reject H_0 or

- if $\bar{z}_s < -z_\alpha$ reject H_0 with an acceptable risk of a Type I error less α .

Example 7.2.6. Consider a sample of size 100 and test whether the mean intelligence quotient increases by comparing the scores before training (IQ1) and after training (IQ2). It is known that the standard deviation of the difference is 1.40 and the sample mean of the difference is -3. Acceptable risk $\alpha = 0.01$.

Solution. Let the random variable X is the IQ1 and Y is the IQ2, with means μ_1 and μ_2 , respectively. We consider a sample with a size $n = 100 > 30$, the sample mean of the difference $\bar{z} = -3$, the standard deviation of the difference $\sigma_z = 1.40$. We want to know if the mean intelligence quotient increases comparing before training (IQ1) and after training (IQ2), so we use the z - test for the following hypothesis:

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 < \mu_2, \text{ with the statistic } Z_s = \frac{\bar{Z}}{S_z} \sqrt{n}.$$

We calculate the specific value

$$\bar{z}_s = \frac{\bar{z}}{\sigma_z} \sqrt{n} = \frac{-3}{1.4} \sqrt{100} = -21.42,$$

and we determine the critical value $z_{0.01}$ such that

$$\Phi_0(2.32) = 0.9898 < \Phi_0(z_{0.01}) = 1 - 0.01 = 0.99 < 0.9901 = \Phi_0(2.33),$$

$$\text{so } z_{0.01} = \frac{2.32 + 2.33}{2} = 2.325.$$

Since $\bar{z}_s = -21.42 < -2.325 = -z_{0.01}$ the null hypothesis is rejected with an acceptable risk of a Type I error less 0.01, so the mean intelligence quotient increases by the training.

8. STATISTICAL INFERENCE IN SIMPLE LINEAR CORRELATION AND REGRESSION. CHI- SQUARE TEST

8.1. SIMPLE LINEAR CORRELATION AND REGRESSION

Correlation and regression are the two most commonly used techniques for investigating the relationship between quantitative variables. Regression refers to linear regression. Correlation measures the strength of the relationship between variables, while linear regression uses an equation to describe this relationship. Simple correlation and regression are used to define some form of association between quantitative variables that are assumed to have a linear relationship. Here, we will drop the word simple in the next discussion, although we will continue to explain the linear relationship between two quantitative variables. For instance, if a person is driving an expensive car, it is often assumed that they must be financially well-off. Correlation and regression are used to numerically quantify this relationship. In the previous chapters, we became familiar with the methods of inferential statistics, according to which conclusions are drawn about unknown population characteristics based on sample information (through estimation or hypothesis testing). There are hypothesis tests that examine the relationship between two variables, which were not mentioned due to the time constraints for mastering the basic methods of statistical inference. However, such an analysis does not allow for the evaluation or prediction of the values of one phenomenon based on the known values of another phenomenon. The statistical methods that enable this will be discussed in this chapter. Therefore, our focus will be on exploring the relationships and influences between the two phenomena. As before, the conclusions will again be based on the examination of a sample. However, now, based on the sample, we will investigate how the variability of one phenomenon is related to the variability of another phenomenon. The goal of our research will not only concern the sample but primarily the population from which the sample was drawn.

8.1.1. DETERMINISTIC AND RANDOM RELATIONSHIP

A deterministic relationship refers to a scenario that is completely objective without any chance of variance or randomness entering the equation. Such cases ensure that the dependent variable can always be accurately predicted, as long as the independent variable is known. A classic example is Ohm's Law in physics, which shows how voltage can always be determined given a known electric current and resistance. Similarly, the geometric definition of a circle defines each point about the circle's radius, asserting that its value is not random. On the contrary, a random relationship introduces some value of unpredictability. In the area of random relationships, although the two variables remain correlative, one is unable to predict with certainty the outcome due to the influence of other factors that lie beyond control.

A clear example of this is the correlation between the number of hours the student studies and their exam score. While studying more typically leads to better performance, factors such as stress, sleep quality, and test difficulty also contribute to variability in the outcome. Similarly, stock market prices fluctuate due to various unpredictable factors, making precise predictions impossible despite observable trends.

The branch of mathematics that deals with random relationships is probability and statistics. Within this field, several specific areas focus on understanding and modeling randomness. In this chapter we will use regression analysis to obtain the relationship between two quantitative variables.

8.1.2. GOALS OF THE CORRELATION AND REGRESSION

When investigating quantitative relationships between two or more phenomena in statistics, the methods of regression and correlation analysis are used. These methods, although closely related and complementary, have significant differences.

In regression analysis, it is essential to identify beforehand which variable will be independent and which will be dependent. This is based

on previous theoretical or empirical knowledge or assumptions about the nature of the phenomena being analyzed. In correlation analysis, when examining two phenomena, it is insignificant which is labelled as the dependent or independent variable - the result remains the same. However, when examining the correlation between three or more phenomena, one of them should be defined as dependent, while the others take on the role of independent variables.

The goal of regression analysis is to determine the type of relationship, i.e. the form of dependence between two phenomena under consideration, which is achieved through an appropriate regression model. A regression model is a statistical model that, through mathematical formulas and appropriate assumptions, describes the quantitative dependence between the variability of the analyzed phenomena. Due to the stochastic nature of these relationships, the regression model shows the average consistency of the variability of the phenomena. However, the regression model is not the ultimate goal, but a tool for assessing and predicting the behavior of the dependent variable at a certain value of the independent variable.

The goal of correlation analysis is to determine whether there is a match between the variability of the phenomena and, if so, how strong the relationship is. It is important to note that neither regression nor correlation analysis can determine the cause-and-effect relationship between the phenomena - whether one is a cause and the other a consequence. This is determined by other methods of quantitative analysis.

When examining the relationships between two variables, the methods of simple (linear or nonlinear) regression and correlation analysis are used. We will limit ourselves to linear relationships, as they are easier to apply and more often used in practice.

8.2. CORRELATION ANALYSIS

Correlation analysis is closely related to regression analysis. The purpose of correlation analysis is to determine whether there is a quantitative agreement (correlation relationship) between the variability of the phenomena under consideration and, if so, to what extent. If two

phenomena are considered, then we are talking about simple correlation. With simple correlation, we can examine whether there is a linear or curvilinear correlation between the phenomena. We will consider only simple linear correlation, but we will drop the word simple.

8.2.1. SAMPLE CORRELATION COEFFICIENT

Unlike regression analysis, in simple linear correlation, the two phenomena under consideration are considered random variables. Here, no distinction is made between dependent and independent variables - both phenomena are given equal status. So, it is completely irrelevant which phenomenon we will label with X , and which with Y , because identical results are obtained.

Namely, if the characteristics X and Y the following sample $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n)$ is considered, whose realization is $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$, then the number:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

is called the **sample covariance**.

To calculate the sample covariance more effectively, the previous formula is modified, resulting in the following formula for sample covariance:

$$\begin{aligned} s_{xy} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^n (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \cdot \bar{y}) = \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y} \right). \end{aligned}$$

As a measure of the strength of a simple linear correlation relationship in the sample, a relative measure is used, which is called the sample correlation coefficient. This coefficient shows the degree of linear quantitative agreement between two phenomena. It is denoted by r and is calculated by the formula:

$$r = \frac{s_{xy}}{s_x s_y},$$

where s_{xy} is a sample covariance and s_x^2, s_y^2 are sample variances of X and Y , respectively.

The coefficient of simple linear correlation as a relative measure takes values from -1 to +1. If it takes positive values, the correlation between the phenomena is direct or directly proportional. In the case when $-1 \leq r < 0$ the relationship is inverse or inversely proportional.

If $r = \pm 1$, then there is a functional relationship between the phenomena under consideration (all empirical points lie on one straight line) and we say that it is a perfect correlation. So, the closer the coefficient of simple linear correlation in absolute value is to 1, the stronger the correlation relationship between the phenomena. On the contrary, the closer the coefficient of simple linear correlation is to 0, the weaker the correlation relationship between the phenomena.

In the extreme case, when the coefficient of simple linear correlation is equal to zero, it is concluded that there is no linear correlation between the phenomena. However, it is wrong to conclude that there is no relationship between the phenomena. In that case, there may be some form of curvilinear arrangement between the phenomena or there may be no quantitative relationship at all; but based on the zero correlation coefficient we are not able to conclude what is true.

Example 8.2.1. Consider the following data for the joint distribution of bivariate random variables (X, Y) :

(3,4)	(3,4)	(4,3)	(4,2)	(2,3)	(3,4)	(3,2)	(4,2)	(6,1)	(2,5)
(5,2)	(3,3)	(4,3)	(5,1)	(2,4)	(6,3)	(3,4)	(5,1)	(4,3)	(5,2)
(5,3)	(1,6)	(2,4)	(3,3)	(6,2)	(4,3)	(2,2)	(3,2)	(5,3)	(5,2)
(4,3)	(5,1)	(4,2)	(3,2)	(5,3)	(6,2)	(4,4)	(3,3)	(5,2)	(3,3)
(2,5)	(3,3)	(4,3)	(5,1)	(2,4)	(6,3)	(3,4)	(5,1)	(4,3)	(4,4)

Find the correlation coefficient.

STATISTICAL INFERENCE IN SIMPLE LINEAR CORRELATION
AND REGRESSION. CHI-SQUARE TEST

Solution. First, we should find the joint frequency distribution, therefore we make the following table:

y_j/x_i	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

Then we form the following table:

STATISTICAL INFERENCE IN SIMPLE LINEAR CORRELATION
AND REGRESSION. CHI-SQUARE TEST

$y_j \setminus x_i$	1	2	3	4	5	6	f_j	$f_j y_j$	$f_j y_j^2$	$\sum_i f_{ij} x_i$	$y_i \sum_i f_{ij} x_i$
1					5	1	6	6	6	31	31
2		1	3	3	4	2	13	26	52	55	110
3			5	7	3	2	18	54	162	70	210
4		3	5	2			10	40	160	29	116
5		2					2	10	50	4	20
6	1						1	6	36	1	6
f_i	1	6	13	13	11	6	Σ	142	466	190	493
$f_i x_i$	1	12	39	52	55	36	195				
$f_i x_i^2$	1	24	117	208	275	216	841				
$\sum_j f_{ij} y_j$	6	24	41	35	22	11	139				
$x_i \sum_j f_{ij} y_j$	6	48	123	140	110	66	493				

From the table, we are determining the following values:

$$\bar{x} = \frac{195}{50} = 3.9, \bar{y} = \frac{142}{50} = 2.84, s_x^2 = \frac{1}{49} (841 - 50 \cdot 3.9^2) = 1.64$$

$$s_y^2 = \frac{1}{49} (466 - 50 \cdot 2.84^2) = 1.28, s_{xy} = \frac{1}{49} (493 - 50 \cdot 3.9 \cdot 2.84) = -1.24$$

So, the correlation coefficient is

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-1.24}{\sqrt{1.64} \cdot \sqrt{1.28}} = -0.86.$$

8.2.2. HYPOTHESIS TESTING FOR THE SAMPLE CORRELATION COEFFICIENT

Since both X_i and Y_i $i=1,2,\dots,n$ are random variables, we conclude that the population correlation coefficient, denoted by ρ , is also a random variable. It turns out that the probability function of the random variable ρ depends only on the parameters r and n . However, the value of the sample correlation coefficient will not coincide with the value of the correlation coefficient of the two-dimensional random variable from which the sample is formed.

The population correlation coefficient indicates the strength of the linear relationship between the two variables being examined within the population. To compute its value accurately, it is essential to have data from the entire population. However, since we typically deal with a sample in practice, the actual value of the population correlation coefficient cannot be determined. Therefore, the question arises: how can a reliable conclusion about the presence of correlation in the population be drawn based on the sample correlation coefficient?

It can be proven that, if the population from which the sample is drawn follows a normal distribution, the sample correlation coefficient provides the most reliable estimate of the correlation coefficient ρ between the random variables of the random variables X and Y and the statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ has Student's distribution with $df = n-2$ degrees of freedom.

Next, we present the procedure for testing hypotheses for the correlation coefficient:

First, we check the *Conditions for applying t- test* - random variables have normal distributions.

Then we choose the null and alternative hypotheses:

$$H_0 : \rho(X, Y) = 0, \quad H_1 : \rho(X, Y) \neq 0$$

and we the statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

Therefore, we calculate the specific value $\bar{t} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$ where r is the sample correlation coefficient, then we determine the critical value $t_\alpha(n-2)$ from the Student's distribution for the two-tailed test and we make a decision:

-if $\bar{t} \leq t_\alpha(n-2)$ fail to reject H_0 or

-if $\bar{t} > t_\alpha(n-2)$ reject H_0 with an acceptable risk of Type I error less than α .

Example 8.2.2. Considering a sample with a size $n=25$ and sample correlation coefficient $r_{xy}=0.59$ test the hypothesis that the population correlation coefficient of X and Y is equal to zero, with an acceptable risk of Type I error $\alpha=0.01$.

Solution. From the table of Student's t - distribution:

$$t_{0.01}(25-2) = 2.807$$

and for the sample:

$$\bar{t} = \frac{0.59}{\sqrt{1-0.59^2}} \sqrt{25-2} = 3.5.$$

Since:

$$\bar{t} = 3.5 > 2.807 = t_{0.01}(23),$$

the null hypothesis is rejected, with an acceptable risk of a Type I error less than $\alpha=0,01$.

Let us assume that the correlation coefficient of the bivariate distribution is nonzero. In this case, the prior statistic doesn't have t - distribution. However, for $\rho \neq 0$ the statistic:

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r},$$

when the sample size increases, approaches a normal distribution with the parameters:

$$m_z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} \quad \text{and} \quad \sigma_z^2 = \frac{1}{n-3}.$$

Then the statistics $u = \frac{Z - m_z}{\sigma_z}$ has $N(0,1)$ distribution.

So, the procedure for this test is:

We check the *Conditions for applying Z - test* - random variables have a normal distribution, the correlation coefficient is different than zero and the sample size is $n \geq 30$,

Then we choose the null and alternative hypotheses:

$$H_0 : \rho(X, Y) = \rho_0 \neq 0, \quad H_1 : \rho(X, Y) = \rho_1 \neq \rho_0$$

where ρ_0 is a number.

Next, we calculate the specific value $\bar{u} = \frac{|z - m_z|}{\sigma_z}$,

where $z = \frac{1}{2} \ln \frac{1+r}{1-r}$, r is the sample correlation coefficient, $\sigma_z^2 = \frac{1}{n-3}$ and

$$m_z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}.$$

Then we determine the critical value u_α such that $P_{H_0} \{ \bar{u} \geq u_{1-\alpha} \} = \alpha$, i.e.

$\Phi_0(u_\alpha) = 1 - \frac{\alpha}{2}$ and make a decision:

- if $\bar{u} \leq u_\alpha$ fail to reject H_0 , or

- if $\bar{u} > u_\alpha$ reject H_0 with acceptable risk of Type I error less than

α .

Example 8.2.3. Considering a sample with a size $n=30$ and sample correlation coefficient $r=0.52$, test the hypothesis that the population correlation coefficient of X and Y is equal to $\rho_0=0.50$, with a level of significance $\alpha=0.01$.

Solution. We calculate:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln \frac{1+0.52}{1-0.52} = 0.58,$$

$$\sigma_z = \frac{1}{n-3} = 0.037$$

and

$$\begin{aligned} m_z &= \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} = \frac{1}{2} \ln \frac{1+0.5}{1-0.5} + \frac{0.5}{2(30-1)} = \\ &= 0.55 + 0.0086 = 0.5586. \end{aligned}$$

Therefore:

$$\bar{u} = \left| \frac{z - m_z}{\sigma_z} \right| = \left| \frac{0.58 - 0.5586}{0.037} \right| = |0.578| = 0.578$$

Од таблицата на нормалната распределба за нивото на значајност $\alpha=0.01$ наоѓаме $\Phi_0(u_\alpha) = 1 - \frac{0.01}{2} = 0.995 \Rightarrow u_{0.01} = 2.585$ and since:

$$u_{0.01} = 2.585 > 0.578 = \bar{u}$$

We fail to reject the null hypothesis, i.e., we do not have enough reason to say that $r=0.52$ is statistically significantly different than the population correlation coefficient $\rho_0=0.50$.

8.3. REGRESSION ANALYSIS

Suppose that each of n patients are treated with the same amount of two different drugs that can affect blood pressure. Namely, each patient is initially treated with a standard drug A, and their change in blood pressure is then measured. After the effect of the drug wears off, the patient is treated with an equal amount of a new drug B, and their

change in blood pressure is measured again. These changes in blood pressure will be called the reaction of the patient to each drug. For $i = 1, 2, \dots, n$, we shall let x_i denote the reaction, measured in appropriate units, of the i -th patient to drug A, and we shall let y_i denote her reaction to drug B. Next, we can plot the observed reaction values as points on a graph (x_i, y_i) $i = 1, 2, \dots, n$ in a two-dimensional coordinate system. To find a simple expression for this relationship, we might wish to fit a straight line to the n points plotted in the graph. Although these n points obviously do not lie exactly on a straight line, we might believe that the deviations from such a line are caused by the fact that the observed change in the blood pressure of each patient is affected not only by the two drugs but also by various other factors. In other words, we might believe that if it were possible to control all of these other factors, the observed points would actually lie on a straight line. This analysis is obtained by a regression model.

To illustrate the nature of regression models, let's first consider an example:

Namely, it is known that air temperature depends on the altitude of the place where it is measured and that it decreases with increasing altitude (by air temperature we mean the average annual air temperature at a certain place). If we make measurements at n places of different heights $x_1, x_2, \dots, x_n$, we will obtain the corresponding values of air temperature $Y_1, Y_2, \dots, Y_n$. However, if we repeat these measurements at the same places, then, in general, we will not obtain the same values of air temperature, but we will obtain new values $Y'_1, Y'_2, \dots, Y'_n$.

Therefore, the variables $Y_1, Y_2, \dots, Y_n$ that indicate air temperatures at certain n places can be considered random variables, while we do not consider $x_1, x_2, \dots, x_n$ as random variables, but consider them to be certain constants.

Since Y_i depends on x_i , we could say that the random variable Y (air temperature) depends on x (altitude of a certain place) which is not random, i.e., $Y = Y(x)$ and for any $i = 1, 2, 3, \dots, n$ this is valid $Y_i = Y(x_i)$.

Therefore, a simple random sample should be understood as follows: for each $i = 1, 2, 3, \dots, n$ of the value x_i , we associate the random variable Y_i . A simple random sample is, therefore, a type of sample where $(x_i, Y_i), i = 1, 2, \dots, n$. Furthermore, in the sample the random variables $Y_1, Y_2, \dots, Y_n$ take on values $y_1, y_2, \dots, y_n$ accordingly, so the realization of the sample is $(x_i, y_i), i = 1, 2, \dots, n$.

It is natural to ask the question based on the sample $(x_i, Y_i), i = 1, 2, \dots, n$ and its realization $(x_i, y_i), i = 1, 2, \dots, n$ to find $Y = Y(x)$ and assess the accuracy of this dependence.

The previous example does not lead to the conclusion that the random variable Y is composed of two components, namely a functional component, i.e. a function $f(x)$ that depends on the argument x and a random component, i.e. a random variable ε that reflects some other influences on the random variable Y .

In practice, the relationship between the random variable Y , the functional component $f(x)$, and the random component ε is usually given by one of the functions:

$$Y = f(x) + \varepsilon \text{ or } Y = f(x)\varepsilon.$$

Here, we will consider the function $Y = f(x) + \varepsilon$, where:

$$f(x) = \alpha x + \beta, \alpha, \beta \in \mathbb{R}$$

and ε is a normally distributed random variable with a mean 0 and a variance σ^2 .

That is, the simple linear regression model is given by the formula

$$f(x) = \alpha x + \beta + \varepsilon,$$

where ε is a normally distributed random variable with mean 0 and variance σ^2 . Furthermore, since $\alpha x + \beta$ is not a random variable, from the properties of the mean and variance it immediately follows that:

$$E(Y) = E(\alpha x + \beta + \varepsilon) = \alpha x + \beta + E(\varepsilon) = \alpha x + \beta + 0 = \alpha x + \beta$$

$$\text{Var}(Y) = \text{Var}(\alpha x + \beta + \varepsilon) = \text{Var}(\varepsilon) = \sigma^2.$$

Let us consider a random sample $(x_i, Y_i), i=1, 2, \dots, n$ with realization $(x_i, y_i), i=1, 2, \dots, n$. Then:

$$Y_i = \alpha x_i + \beta + \varepsilon_i, \text{ for } i=1, 2, \dots, n,$$

where $\varepsilon_i, i=1, 2, \dots, n$ are independent random variables normally distributed random variable with mean 0 and variance σ^2 . Therefore, $Y_i, i=1, 2, \dots, n$ are independent random variables that have normal distributions with mean $\alpha x_i + \beta$ and variance σ^2 , respectively.

Hence, it follows that in the case of linear regression, the determination of an analytical expression for the random variable Y is reduced to finding the functional and random component of Y . This means that based on the sample, the parameters α , β , and σ^2 should be estimated. That is, based on the sample $(x_i, Y_i), i=1, 2, \dots, n$, we should find point estimators a , b and s^2 of the parameters α , β , and σ^2 , respectively and later we need to compute the point estimates a' , b' , and s'^2 considering the realization $(x_i, y_i), i=1, 2, \dots, n$ of the sample $(x_i, Y_i), i=1, 2, \dots, n$.

8.3.1. POINT ESTIMATES OF SIMPLE LINEAR REGRESSION PARAMETERS AND LINEAR REGRESSION LINES

Let the realization $(x_i, y_i), i=1, 2, \dots, n$ of the sample $(x_i, Y_i), i=1, 2, \dots, n$, be given. To determine the estimates a and b , of the parameters α and β , respectively, based on the sample, in the linear regression model $Y = \alpha x + \beta + \varepsilon$ we proceed as follows. For each value x_i of the parameter x , let the value $\hat{y}_i = \alpha x_i + \beta$ be calculated. Then, at

the point (x_i, y_i) the deviation of the calculated value $\hat{y}_i$ from the realized value y_i is equal to the realized value ε_i' of the random variable ε_i , that is:

$$\varepsilon_i' = y_i - \hat{y}_i = y_i - \alpha x_i - \beta$$

Hence, the total mean square deviation of the realized values $y_i, i=1,2,...,n$ from the calculated values $\hat{y}_i, i=1,2,...,n$ is given by the unbiased statistic:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \alpha x_i - \beta)^2 \quad (8.1)$$

Furthermore, we will determine the estimates a and b of the parameters α and β , so that the mean square deviation is the smallest, that is, the function $s^2 = \frac{1}{n-1} \sum_i (y_i - \alpha x_i - \beta)^2$ has a minimum extreme

value. Therefore, we find the partial derivatives $\frac{\partial s^2}{\partial \alpha}$ and $\frac{\partial s^2}{\partial \beta}$ and set them equal to zero, which gives us the system of equations:

$$\frac{\partial s^2}{\partial \alpha} = 0, \quad \frac{\partial s^2}{\partial \beta} = 0 \quad (8.2)$$

Since,

$$\begin{aligned} \frac{\partial s^2}{\partial \alpha} &= -\frac{2}{n-1} \sum_{i=1}^n x_i (y_i - \alpha x_i - \beta) = \\ &= -2 \left(\frac{1}{n-1} \sum_{i=1}^n x_i y_i - \alpha \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \beta \frac{1}{n-1} \sum_{i=1}^n x_i \right) \end{aligned}$$

and if we use a plugin the following substitutions

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s_{xy} = \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \cdot \bar{y} \right),$$

$$s_x^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right), \quad s_y^2 = \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 \right)$$

then

$$\frac{\partial s^2}{\partial \alpha} = -2 \left(s_{xy} + \frac{n}{n-1} \bar{x} \cdot \bar{y} - \alpha \left(s_x^2 + \frac{n}{n-1} \bar{x}^2 \right) - \beta \frac{n}{n-1} \bar{x} \right).$$

For the second partial derivative

$$\begin{aligned} \frac{\partial s^2}{\partial \beta} &= -2 \frac{1}{n-1} \sum_{i=1}^n (y_i - \alpha x_i - \beta) = \frac{-2n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n y_i - \alpha \frac{1}{n} \sum_{i=1}^n x_i - \beta \right) = \\ &= -\frac{2n}{n-1} (\bar{y} - \alpha \bar{x} - \beta) \end{aligned}$$

The system of equations (8.2) is equivalent to the system

$$\begin{cases} s_{xy} + \frac{n}{n-1} \bar{x} \cdot \bar{y} - \alpha \left(s_x^2 + \frac{n}{n-1} \bar{x}^2 \right) - \beta \frac{n}{n-1} \bar{x} = 0 \\ \bar{y} - \alpha \bar{x} - \beta = 0 \end{cases}$$

The solution of the last system is $\alpha_0 = \frac{s_{xy}}{s_x^2}$ and $\beta_0 = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}$.

This means that for α_0 and β_0 the function (8.1) reaches a minimum and these are the required estimates a' and b' of the parameters α and β , respectively. Furthermore, if we replace the parameters α and β in the equation $y = \alpha x + \beta$ with the found estimates a' and b' for the linear regression equation of the sample, we find:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x}) \quad (8.3).$$

To determine the minimum value of the mean square deviation, it is sufficient to substitute the found estimates a' and b' of the parameters α and β in equation (8.1). We obtain:

$$\begin{aligned}
 s'^2 &= \frac{1}{n-1} \sum_{i=1}^n \left(y_i - \frac{s_{xy}}{s_x^2} x_i - \left(\bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \right) \right)^2 = \frac{1}{n-1} \sum_{i=1}^n \left(y_i - \bar{y} - \frac{s_{xy}}{s_x^2} (x_i - \bar{x}) \right)^2 = \\
 &= \frac{1}{n-1} \sum_{i=1}^n \left((y_i - \bar{y})^2 - 2 \frac{s_{xy}}{s_x^2} (x_i - \bar{x})(y_i - \bar{y}) + \frac{s_{xy}^2}{s_x^4} (x_i - \bar{x})^2 \right) = \\
 &= \\
 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 - 2 \frac{s_{xy}}{s_x^2} \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \frac{s_{xy}^2}{s_x^4} \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \\
 &= s_y^2 - 2 \frac{s_{xy}^2}{s_x^2} + \frac{s_{xy}^2}{s_x^4} s_x^2 = s_y^2 - \frac{s_{xy}^2}{s_x^2} = s_y^2 (1 - r^2),
 \end{aligned}$$

where $r = \frac{s_{xy}}{s_x s_y}$ is the correlation coefficient.

To the previous discussion $s'^2 = s_y^2 (1 - r^2)$ is the point estimator of the variance σ^2 of the random component ε .

We consider a' , b' and s'^2 as realized values of the statistics a , b and s^2 , respectively, for the sample (x_i, y_i) .

If we use the notations:

$$\bar{Y} = \frac{1}{n} \sum_i Y_i, \quad s_{xY} = \frac{1}{n-1} \left(\sum_{i=1}^n x_i Y_i - n \bar{x} \bar{Y} \right), \quad s_Y^2 = \frac{1}{n-1} \left(\sum_{i=1}^n Y_i^2 - n \bar{Y}^2 \right)$$

and $r = \frac{s_{xY}}{s_x s_Y}$, then for the statistics a , b and s^2 we get:

$$a = \frac{s_{xY}}{s_x^2}, \quad b = \bar{Y} - \frac{s_{xY}}{s_x^2} \bar{x} \quad \text{and} \quad s^2 = s_Y^2 (1 - r^2), \quad \text{respectively.}$$

In addition, for the statistics a and b this is valid:

$$\begin{aligned}
 E(a) &= E\left(\frac{s_{xy}}{s_x^2}\right) = \frac{1}{s_x^2} E(s_{xy}) = \frac{1}{s_x^2} E\left(\frac{1}{n-1} \left(\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{y}\right)\right) = \\
 &= \frac{1}{s_x^2} E\left(\frac{1}{n-1} \sum_{i=1}^n x_i Y_i - \frac{n}{n-1} \bar{x}\bar{y}\right) = \frac{1}{s_x^2} E\left(\frac{1}{n-1} \sum_{i=1}^n x_i Y_i - \frac{n}{n-1} \bar{x} \frac{1}{n} \sum_{i=1}^n Y_i\right) = \\
 &= \frac{1}{s_x^2} E\left(\frac{1}{n-1} \sum_{i=1}^n x_i Y_i - \frac{1}{n-1} \bar{x} \sum_{i=1}^n Y_i\right) = \frac{1}{s_x^2} E\left(\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) Y_i\right) = \\
 &= \frac{1}{s_x^2} \cdot \frac{1}{n-1} E\left(\sum_{i=1}^n (x_i - \bar{x}) Y_i\right) = \frac{1}{s_x^2} \cdot \frac{1}{n-1} E\left(\sum_{i=1}^n (x_i - \bar{x})(\alpha x_i + \beta)\right) = \\
 &= \frac{1}{s_x^2} \cdot \frac{1}{n-1} E\left(\alpha \sum_{i=1}^n (x_i^2 - \bar{x} \cdot x_i)\right) + \frac{1}{s_x^2} \cdot \frac{1}{n-1} \beta E\left(\sum_{i=1}^n (x_i - \bar{x})\right) = \\
 &= \frac{\alpha}{s_x^2} \cdot \frac{1}{n-1} E\left(\sum_{i=1}^n x_i^2 - \bar{x} \cdot \sum_{i=1}^n x_i\right) + \frac{1}{s_x^2} \cdot \frac{1}{n-1} \beta E\left(\sum_{i=1}^n x_i - n\bar{x}\right) = \\
 &= \frac{\alpha}{s_x^2} \cdot \frac{1}{n-1} E\left(\sum_{i=1}^n x_i^2 - \bar{x} \cdot n\bar{x}\right) + \frac{1}{s_x^2} \cdot \frac{1}{n-1} \beta (n\bar{x} - n\bar{x}) = \\
 &= \frac{\alpha}{s_x^2} \cdot \frac{1}{n-1} E\left(\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2\right) + \frac{1}{s_x^2} \cdot \frac{1}{n-1} \beta \cdot 0 = \\
 &= \frac{\alpha}{s_x^2} \cdot s_x^2 + 0 = \alpha
 \end{aligned}$$

and

$$\begin{aligned}
 E(b) &= E(\bar{y} - \alpha\bar{x}) = E(\bar{y}) - \bar{x}E(\alpha) = E\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) - \alpha\bar{x} = \\
 &= \frac{1}{n} \sum_{i=1}^n (\alpha x_i + \beta) - \alpha\bar{x} = \beta,
 \end{aligned}$$

which means a and b are unbiased estimators of α and β .

From the previous discussion we can say that the equation (model) of a simple linear regression is given by the expression (8.3). In

this case, $b = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}$ is called a constant term, and $a = \frac{s_{xy}}{s_x^2}$ is called a regression coefficient.

The regression coefficient gives the change in the regression value for a unit increase in the independent variable X , that is, it indicates how much the value of the dependent variable Y changes on average for a unit change in the independent variable X .

Furthermore, the total mean square deviation of the data from the regression line (8.3) is $s^2 = s_y^2(1-r^2)$ and we call it the variance of the linear regression, and the square root $s = s_y\sqrt{1-r^2}$ is called the standard deviation of the linear regression.

In this case, the standard deviation of linear regression shows the average deviation of the empirical values $y_i, i=1,2,\dots,n$ of the dependent variable Y from the regression values $\hat{y}_i, i=1,2,\dots,n$ in units of measurement of the dependent variable Y .

The ratio $\frac{100s}{\bar{y}} = V_s$, which we call the coefficient of variation of linear regression, gives the relative average deviation in percentage of the empirical values from the regression values.

A specific indicator of the representativeness of linear regression is the coefficient of determination c_d , which is equal to the quotient of the standard deviation of the regression values $\hat{y}_i, i=1,2,\dots,n$ of the sample mean $\bar{y}$ and the sample variance s_y^2 .

Therefore,

$$\begin{aligned} c_d &= \frac{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{s_y^2} = \frac{\frac{1}{n} \sum_{i=1}^n \left(\frac{s_{xy}}{s_x^2} x_i + \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} - \bar{y} \right)^2}{s_y^2} = \frac{\frac{1}{n} \frac{s_{xy}^2}{s_x^4} \sum_{i=1}^n (x_i - \bar{x})^2}{s_y^2} = \\ &= \frac{s_{xy}^2}{s_x^4 s_y^2} s_x^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = r^2. \end{aligned}$$

The coefficient of determination provides insights into the extent to which the independent variable accounts for the overall variability of the dependent variable. Its values range from 0 to 1. Additionally, a coefficient of determination nearing 1 indicates that the regression line closely aligns with the observed data.

Note. With a similar procedure, we can find the other regression line:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2} (y - \bar{y})$$

which gives the dependence of the variable X on the variable Y .

Example 8.3.1. Consider the following data for the joint distribution of bivariate random variables (X, Y) :

(5,3) (1,6) (2,4) (3,3) (6,2) (4,3) (4,2) (3,2) (5,3) (5,2)
 (3,4) (3,4) (4,3) (4,2) (2,3) (3,4) (3,2) (4,2) (6,1) (2,5)
 (2,5) (3,3) (4,3) (5,1) (2,4) (6,3) (3,4) (5,1) (4,3) (4,4)
 (5,2) (3,3) (4,3) (5,1) (2,4) (6,3) (3,4) (5,1) (4,3) (5,2)
 (4,4) (6,1) (4,2) (3,2) (5,3) (6,2) (4,3) (3,3) (5,2) (3,3)

- a) What is the regression line if we consider X for a fixed variable?
- b) Find the regression standard deviation and regression variance coefficient for the regression line from a) and then explain.
- c) How much of the total variability of the dependent variable in terms of the independent variable is explained?

Solution. First, we should find the joint frequency distribution, therefore we make the following table:

STATISTICAL INFERENCE IN SIMPLE LINEAR CORRELATION
AND REGRESSION. CHI-SQUARE TEST

y_j/x_i	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

Then we form the following table:

STATISTICAL INFERENCE IN SIMPLE LINEAR CORRELATION
AND REGRESSION. CHI-SQUARE TEST

$y_j \setminus x_i$	1	2	3	4	5	6	f_j	$f_j y_j$	$f_j y_j^2$	$\sum_i f_{ij} x_i$	$y_i \sum_i f_{ij} x_i$
1					4	2	6	6	6	32	32
2			3	4	4	2	13	26	52	57	114
3		1	5	7	3	2	18	54	162	72	216
4		3	5	2			10	40	160	29	116
5		2					2	10	50	4	20
6	1						1	6	36	1	6
f_i	1	6	13	13	11	6	Σ	142	466	195	504
$f_i x_i$	1	12	39	52	55	36	195				
$f_i x_i^2$	1	24	117	208	275	216	841				
$\sum_j f_{ij} y_j$	6	25	41	37	21	12	142				
$x_i \sum_j f_{ij} y_j$	6	50	123	148	105	72	504				

a) From the table we are determining the following values:

$$\bar{x} = \frac{195}{50} = 3.9, \quad \bar{y} = \frac{142}{50} = 2.84, \quad s_x^2 = \frac{1}{49} (841 - 50 \cdot 3.9^2) = 1.64,$$

$$s_y^2 = \frac{1}{49} (466 - 50 \cdot 2.84^2) = 1.28, \quad s_{xy} = \frac{1}{49} (504 - 50 \cdot 3.9 \cdot 2.84) = -1.02$$

So, the regression line is

$$y - 2.84 = \frac{-1.02}{1.64} (x - 3.9), \text{ or } y = -0.62x + 5.26$$

b) Regression standard deviation is

$$s_1 = s_y \sqrt{1 - r^2} = 1.13 \sqrt{1 - 0.49} = 0.8,$$

regression variance coefficient is:

$$V_{s_1} = \frac{100 \cdot 0.8}{2.84} = 28\%.$$

Therefore, the relative average deviation of the measurements of the dependent variable Y from its regression values is 28%.

- c) Since the determination coefficient is $c_d r = r^2 = 0.49$, 49% of the total variability of the dependent variable Y in terms of the independent variable X is explained.

8.4. CHI SQUARE TEST OF INDEPENDENCE

The chi-square test of independence serves to assess whether a statistically significant connection exists between two categorical variables. The fundamental concept involves comparing the distribution of the actual data with what would be anticipated if the two variables were independent. This test enables us to ascertain if variables are independent or exhibit some form of association. For instance, we can examine if there is a correlation between gender and smoking habits, which means determining if one variable (gender) influences the other (smoking). The test is particularly effective for categorical data (nominal or ordinal) that cannot be analyzed easily through other methods like correlation or regression. The chi-square test utilizes a mathematical framework to contrast observed and expected figures, offering an impartial means to draw conclusions about the relationship. The test can accommodate various table sizes (e.g., 2x2, 3x4, etc.), rendering it appropriate for a broad spectrum of research and data configurations.

Using this test, if the relationship is found to be statistically significant, it means that there is a high probability that one category depends on another category, which can have important implications in various research fields.

Suppose we are analyzing two categorical variables X and Y in a population, assuming that the first variable has r possible outcomes, and

the second has s outcomes. We can summarize the data with a table, known as a contingency table, which consists of r rows and s columns.

The contingency table displays the frequencies of each pair for every combination of the categories of the two variables. Each cell from the total number of cells $r \times s$ represents one combination of the categories of the two variables.

With this test, we test the hypotheses in the following form:

- H_0 : The two variables are considered independent.

- H_1 : The two variables are considered dependent.

Two different experimental situations lead to contingency tables and here we are interested in testing the association between two categorical variables, that is, the two characteristics of the population. In this case, the null hypothesis is a statement of independence between the two variables.

For $i=1$ to r (rows) and $j=1$ to c (columns) we denote by R_i the number of observed values/frequencies in i - the row, by C_j the number of observed values/frequencies in j - th column, O_{ij} the number of observed values/frequency in the field of i - th row and j - th column and by E_{ij} the expected value in the field of i - th row and j - th column if the variables are independent, i.e. the null hypothesis is true. The last values are calculated by

$$E_{ij} = \frac{R_i \cdot C_j}{n}, i=1, \dots, r \text{ and } j=1, \dots, c;$$

The test statistic measures the difference between the observed frequencies and the expected frequencies, assuming independence, and is calculated by the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

If the chi-square statistic is large, this means that the observed frequencies are not close to the frequencies that we would expect if the two variables were independent. Therefore, "large" values of the chi-

square statistic according to the chi-square distribution show evidence against the null hypothesis and support the alternative hypothesis.

If the null hypothesis is true, the probability that the statistic that has a chi-square distribution with $(r-1) \times (c-1)$ degrees of freedom would be as large as or greater than the value of the chi-square test.

Therefore, the p -value for a chi-square test is always the area to the right of the test statistic under the density plot of the chi-square distribution with $(r-1) \times (c-1)$ degrees of freedom, that is $P(X > \chi^2) = p$, where there X has a chi-square distribution with $(r-1) \times (c-1)$ degrees of freedom. To calculate this probability, we need to use the chi-square distribution with $(r-1) \times (c-1)$ degrees of freedom. With the help of some statistical software, the p -value can be obtained from the results, and if it is less than the significance level α , the null hypothesis is rejected. Using the chi-square distribution table (since it will usually not be possible to find the exact p -value in the table), a range for the p -value can be determined.

There is another way to decide whether the null hypothesis is rejected or whether there is not enough statistically significant evidence to reject the null hypothesis, by determining the critical value of the test $\chi^2_{(r-1) \times (c-1), \alpha}$, which is read from the Table of the chi-square distribution for $(r-1) \times (c-1)$ degrees of freedom and probability p equal to the significance level of the test, i.e. $p = \alpha$. If the value of the test statistic is greater than the critical value, then the null hypothesis is rejected, i.e. it is concluded that the two variables are dependent with a probability of error of the first type less than α .

Furthermore, two conditions must be satisfied before the test can be applied, all expected values must be greater than 1 and at least 80% of the expected values must be greater than 5.

Example 8.4.1. A University conducted a survey of its recent graduates to collect demographic and health information for future planning purposes as well as to assess students' satisfaction with their undergraduate experiences. The survey revealed that a substantial proportion of students were not engaging in regular exercise, many felt

STATISTICAL INFERENCE IN SIMPLE LINEAR CORRELATION
AND REGRESSION. CHI-SQUARE TEST

their nutrition was poor, and a substantial number were smoking. Graduates were asked if they were taking on regular exercise and where they lived their senior year. The response options were dormitory, on-campus apartment, off-campus apartment, and at home (i.e., commuted to and from the university). The data are shown below:

	No Regular Exercise	Sporadic Exercise	Regular Exercise
Dormitory	32	30	28
On-Campus Apartment	74	64	42
Off-Campus Apartment	110	25	15
At Home	39	6	5

Based on the data, is there a relationship between exercise and students' living arrangements? Do you think a person's living situation influences their exercise habits? $\alpha = 0.05$.

Solution. We need to test the following hypothesis:

H_0 : There is no relationship between exercise and students' living arrangements.

H_1 : There is a relationship between exercise and students' living arrangements.

First, let's calculate the expected values in each field.

$$E_{11} = \frac{R_1 \cdot C_1}{n} = \frac{90 \cdot 255}{470} = 48.83$$

$$E_{12} = \frac{R_1 \cdot C_2}{n} = \frac{90 \cdot 125}{470} = 23.94$$

$$E_{13} = \frac{R_1 \cdot C_3}{n} = \frac{90 \cdot 90}{470} = 17.23$$

$$E_{21} = \frac{R_2 \cdot C_1}{n} = \frac{180 \cdot 255}{470} = 97.66$$

STATISTICAL INFERENCE IN SIMPLE LINEAR CORRELATION
AND REGRESSION. CHI-SQUARE TEST

$$E_{22} = \frac{R_2 \cdot C_2}{n} = \frac{180 \cdot 125}{470} = 47.88$$

$$E_{23} = \frac{R_2 \cdot C_3}{n} = \frac{180 \cdot 90}{470} = 34.47$$

$$E_{31} = \frac{R_3 \cdot C_1}{n} = \frac{150 \cdot 255}{470} = 81.38$$

$$E_{32} = \frac{R_3 \cdot C_2}{n} = \frac{150 \cdot 125}{470} = 39.89$$

$$E_{33} = \frac{R_3 \cdot C_3}{n} = \frac{150 \cdot 90}{470} = 28.72$$

$$E_{41} = \frac{R_4 \cdot C_1}{n} = \frac{50 \cdot 255}{470} = 27.13$$

$$E_{42} = \frac{R_4 \cdot C_2}{n} = \frac{50 \cdot 125}{470} = 13.3$$

$$E_{43} = \frac{R_4 \cdot C_3}{n} = \frac{50 \cdot 90}{470} = 9.57$$

The necessary conditions are valid and we form the following table:

	No Regular Exercise	Sporadic Exercise	Regular Exercise
Dormitory	O ₁₁ = 32 E ₁₁ = 48.83	O ₁₂ = 30 E ₁₂ = 23.94	O ₁₃ = 28 E ₁₃ = 17.23
On-Campus Apartment	O ₂₁ = 74 E ₂₁ = 97.66	O ₂₂ = 64 E ₂₂ = 47.88	O ₂₃ = 42 E ₂₃ = 34.47
Off-Campus Apartment	O ₃₁ = 110 E ₃₁ = 81.38	O ₃₂ = 25 E ₃₂ = 39.89	O ₃₃ = 15 E ₃₃ = 28.72
At Home	O ₄₁ = 39 E ₄₁ = 27.13	O ₄₂ = 6 E ₄₂ = 13.3	O ₄₃ = 5 E ₄₃ = 9.57

STATISTICAL INFERENCE IN SIMPLE LINEAR CORRELATION
AND REGRESSION. CHI-SQUARE TEST

Then, χ^2 - statistics for this sample is:

$$\chi^2 = \sum_{i=1}^4 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 60.44$$

The critical value of the test from the Table of χ^2 - distribution for $(r-1) \cdot (c-1) = (4-1) \cdot (3-1) = 6$ degrees of freedom and significant level $\alpha = 0.05$ e:

$$\chi^2_{6;0.05} = 12.59$$

Since $\chi^2 = 60.44 > 12.59 = \chi^2_{6;0.05}$ the null hypothesis is rejected, and we conclude that there is a relationship between exercise and students' living arrangements with an acceptable risk of Type I error of less than 0.05.

9. PRACTICAL GUIDE TO STATISTICS IN R

9.1. INTRODUCTION TO R ESSENTIALS AND ENVIRONMENT

9.1.1. WHAT IS R AND WHY USE IT?

R is a powerful open-source programming language and software environment designed specifically for statistical computing, data analysis, and graphical visualization. Developed in the early 1990s by Ross Ihaka and Robert Gentleman at the University of Auckland in New Zealand, R was inspired by the S programming language. Over time, it has become one of the most popular tools for data-driven research and analytics.

One of R's standout features is its open-source nature, which allows anyone to access, modify, and contribute to its development. This has led to the creation of thousands of user-contributed packages available through repositories such as CRAN (Comprehensive R Archive Network). These packages significantly expand R's capabilities, providing tools for everything from advanced machine learning to geospatial analysis and econometrics.

Another key advantage of R lies in its extensive statistical libraries, which support a wide range of techniques including regression, hypothesis testing, time series analysis, and clustering. Its strong visualization capabilities are equally noteworthy, offering users the ability to create high-quality, customizable graphics through base plotting functions and packages.

R has gained immense popularity in both academia and industry due to its flexibility and versatility. Researchers in fields like social sciences, life sciences, and finance rely on R for reproducible and transparent data analysis. At the same time, companies in sectors like healthcare, marketing, and technology increasingly use R to gain insights from large datasets and build predictive models. Its ability to seamlessly integrate with other languages and platforms like Python, SQL, and cloud services further solidifies its role as a vital tool for modern data analytics.

9.1.2. INSTALLING R and Rstudio

To get started with R, you need to install both R and RStudio. R is the core statistical computing environment, while RStudio is an integrated development environment (IDE) that makes working with R more efficient and user-friendly.

Step 1: Installing R

1. Visit the official Comprehensive R Archive Network ([CRAN](https://cran.r-project.org/)).
2. Choose your operating system (Windows, macOS, or Linux).
3. Download and run the appropriate installer for your system.
4. Follow the installation prompts, accepting the default settings unless you have specific requirements.

Once installed, you can access R directly by opening the R console. However, using RStudio simplifies the workflow for coding, data visualization, and project management.

Step 2: Installing RStudio

1. Visit the [RStudio](https://www.rstudio.com/) website.
2. Download the free RStudio Desktop version compatible with your operating system.
3. Run the installer and follow the prompts to complete the installation.
- 4.

Step 3: Verifying the Installation

1. Open RStudio after installation.
2. In the console window, type the command `version` and press Enter.
3. You should see version details confirming that R is correctly installed.

Exploring the RStudio Interface

RStudio provides a clean and intuitive interface divided into four main panels:

- **Source Panel:** For writing and editing R scripts.
- **Console Panel:** Displays outputs and allows interactive command execution.
- **Environment/History Panel:** Shows loaded objects and command history.
- **Plots/Help/Files Panel:** Displays graphics, help documentation, and project files.

With both R and RStudio installed, you're ready to start writing code, analyzing data, and creating visualizations for statistics purposes efficiently.

9.1.3. BASICS OF THE R ENVIRONMENT

The R Console: Running Simple Commands

The R console is an interactive interface where you can execute R commands directly. In RStudio, the console is typically located in the bottom-left panel. When R is ready to accept commands, it displays a `>` prompt. You can type commands after this prompt and press Enter to execute them. For example, typing `2 + 2` and pressing Enter will output `[1] 4`, indicating that R has computed the sum.

The console is ideal for quick calculations or testing small code snippets. However, for more extensive coding tasks, it's beneficial to use R scripts.

R Scripts vs. Interactive Console Usage

An R script is a simple text file that contains a series of R commands. Scripts allow you to write, edit, and save your code for future use, promoting reproducibility and efficient debugging. In RStudio, you can create a new script by navigating to `File > New File > R Script`. It opens the script editor, typically in the top-left panel.

While the console is suitable for immediate command execution, scripts are preferable for:

- Writing and saving complex analyses.
- Organizing code into reusable sections.
- Collaborating with others by sharing script files.

To run code from a script, you can highlight the desired lines and press `Ctrl + Enter` (Windows) or `Cmd + Enter` (Mac). The output will appear in the console.

Workspace and Working Directories

The workspace in R refers to the current environment that includes all the objects (variables, data frames, and functions) you've created during your session. Managing your workspace effectively ensures that your analysis remains organized and reproducible.

Working Directory

The working directory is the default folder where R reads and saves files. You can check your current working directory using the `getwd()` function.

To change the working directory, use the `setwd()` function with the path to your desired folder. Setting the appropriate working directory is crucial for R to locate your data files and to save outputs in the desired location.

Saving and Loading Workspaces

To preserve your work, you can save the entire workspace or specific objects:

- **Save the entire workspace:** This saves all objects in your current session to a file named `.RData` in the working directory.
- **Save specific objects** This allows you to save only selected objects to a file.

9.1.4. FIRST STEPS WITH R: SIMPLE COMMANDS

R can be used as a powerful tool for performing basic arithmetic, managing variables, and working with different data types. In this chapter, we will introduce essential concepts, complete with solved examples ranging from basic to intermediate.

Using R as a Calculator

R supports all standard arithmetic operations, including addition, subtraction, multiplication, and division. You can type commands directly in the R console and press Enter to see the results.

Basic Arithmetic Examples:

```
1  # Simple operations
2  2 + 3  # Addition
3  10 - 4  # Subtraction
4  5 * 3  # Multiplication
5  20 / 4  # Division
6  2^3    # Exponentiation (2 to the power of 3)
7  sqrt(16) # Square root
```

Variables and Assignment

In R, you can store values in variables using the assignment operator `<-` (preferred) or `=`. It is good practice to use `<-` for assignment to distinguish it from comparison operations (`==`).

Examples:

```
1  # Assigning values to variables
2  x <- 10    # Using preferred assignment operator
3  y = 5      # Alternative assignment
4
5  # Performing operations with variables
6  sum <- x + y
7  product <- x * y
8
9  # Displaying the values
10 sum
11 product
12
```

Data Types in R

R supports several data types, including:

- **Numeric:** Numbers (e.g., 10, 3.14)

- **Character:** Text or strings (e.g., "hello", "R language")
- **Logical:** Boolean values (TRUE, FALSE)

```
1  # Numeric example
2  num_val <- 42
3
4  # Character example
5  char_val <- "Hello, R!"
6
7  # Logical example
8  log_val <- TRUE
9
10 # Display the types of each variable
11 class(num_val) # Returns "numeric"
12 class(char_val) # Returns "character"
13 class(log_val) # Returns "logical"
```

Simple Operations on Vectors

A **vector** is a basic data structure in R that holds elements of the same type. You can create vectors using the `c()` function.

```
1  # Create a numeric vector
2  num_vec <- c(1, 2, 3, 4, 5)
3
4  # Access elements using indexing (1-based)
5  num_vec[1] # First element
6  num_vec[3] # Third element
7
8  # Modify an element
9  num_vec[2] <- 10
10
11 # Display the updated vector
12 num_vec
```

Vectorized Operations

```
1  # Adding a constant to all elements
2  num_vec <- c(1, 2, 3, 4, 5)
3  result <- num_vec + 10
4  result # [1] 11 12 13 14 15
5
6  # Element-wise operations
7  vec1 <- c(1, 2, 3)
8  vec2 <- c(4, 5, 6)
9  sum_vec <- vec1 + vec2
10 sum_vec # [1] 5 7 9
```

R allows operations on entire vectors without loops. This is both efficient and concise.

9.1.5. R Packages: EXTENDING FUNCTIONALITY

What are packages?

Packages in R are collections of functions, datasets, and documentation bundled together to extend the capabilities of base R. They allow users to perform specialized tasks, such as advanced data visualization, machine learning, statistical analysis, or working with time-series data. Think of packages as "add-ons" or "plug-ins" that can be installed to solve specific problems without reinventing the wheel.

Key Features of R Packages:

- **Extensibility:** Thousands of packages are available for diverse tasks.
- **Community-Driven:** Many packages are developed and maintained by the R user community.
- **Standard Repository:** Packages are typically hosted on CRAN (Comprehensive R Archive Network).
- **Versioning and Updates:** Packages are regularly updated for improvements and bug fixes.

How to install and load packages (*install.packages()*, *library()*)

To use a package, you need to install it (only once) and then load it every time you start a new R session.

To enhance R's capabilities for statistical analysis, several essential packages can be installed and loaded. These packages offer powerful tools for tasks such as hypothesis testing, regression analysis, time series forecasting, and data visualization.

```
1 install.packages(c("ggplot2", "dplyr", "tidyr", "MASS", "car", "lmtest", "forecast", "psych", "boot"))
```

Brief description of some of the most important packages for statistical analysis:

ggplot2	For creating advanced data visualizations
dplyr	Efficient data manipulation and transformation
tidyr	Helps in reshaping and cleaning data
MASS	Provides functions for statistical methods and datasets
car	Functions for regression diagnostics and hypothesis testing
lmtest	Additional tools for linear models and hypothesis testing
forecast	Time series analysis and forecasting
Psych	Tools for psychological and multivariate statistics
boot	For bootstrapping statistical methods

After installation, load the packages using the `library()` function.

```
1 # Load essential statistical packages
2 library(ggplot2)
3 library(dplyr)
4 library(tidyr)
5 library(MASS)
6 library(car)
7 library(lmtest)
8 library(forecast)
9 library(psych)
10 library(boot)
```

To verify whether packages have been successfully installed, use `installed.packages()`.

Introduction to key package stats

The `stats` package is one of R's core packages and is automatically included in every R installation. It provides essential functions for statistical computations, data modeling, hypothesis testing, and probability distribution analysis. This makes it a vital tool for statisticians and data analysts.

The package covers several key areas of statistics, such as:

- **Descriptive Statistics:** Measures of central tendency (mean, median) and spread (variance, standard deviation).
- **Hypothesis Testing:** T-tests, Chi-squared tests, correlation tests, and more.
- **Probability Distributions:** Functions to generate random samples, compute density, and distribution functions.
- **Linear and Non-linear Models:** Simple and multiple linear regressions, time-series models, and generalized linear models.
- **Time Series Analysis:** Autocorrelation functions and forecasting tools

Examples of using stats package

a) Descriptive statistics example

```
1  # Data
2  data <- c(5, 10, 15, 20, 25)
3
4  # Mean and Median
5  mean(data) # 15
6  median(data) # 15
7
8  # Standard Deviation and Variance
9  sd(data) # 7.905694
10 var(data) # 62.5
```

b) Hypothesis Testing (One-Sample t-Test)

```
1 # Testing if the sample mean is equal to 12
2 t_test <- t.test(data, mu = 12)
3 t_test
4
5 # Explanation: The result of this test provides a p-value
6 # to determine if there is a significant difference between
7 # the sample mean and the hypothesized value mu = 12.
```

c) Correlation and Linear Regression Analysis

```
1 # Example dataset
2 x <- c(1, 2, 3, 4, 5)
3 y <- c(2.3, 2.9, 4.1, 4.8, 5.6)
4
5 # Correlation
6 cor(x, y) # Returns correlation coefficient
7
8 # Fitting a linear regression model
9 model <- lm(y ~ x)
10 summary(model) # Regression output
```

Plotting regression line

```
1 plot(x, y, main = "Linear Regression Example")
2 abline(model, col = "blue")
```

d) Probability Distributions (Normal Distribution)

```
1 # Generate 10 random numbers from a normal distribution with mean 0, sd 1
2 random_samples <- rnorm(10, mean = 0, sd = 1)
3 random_samples
4
5 # Compute density, distribution, and quantile functions
6 dnorm(0) # Density function
7 pnorm(1.96) # Cumulative distribution function
8 qnorm(0.975) # Quantile function for 97.5% percentile
```

e) ANOVA (Analysis of Variance)

```
1  # Data
2  group1 <- c(5, 6, 7, 8, 9)
3  group2 <- c(7, 8, 8, 9, 10)
4  group3 <- c(6, 7, 8, 8, 9)
5
6  # Combine into a data frame
7  data <- data.frame(
8    value = c(group1, group2, group3),
9    group = rep(c("A", "B", "C"), each = 5)
10 )
11
12 # Perform one-way ANOVA
13 anova_result <- aov(value ~ group, data = data)
14 summary(anova_result)
```

f) Time Series Analysis: Autocorrelation and Forecasting

```
1  # Generate time series data
2  time_series_data <- ts(c(100, 102, 105, 110, 120, 130), frequency = 1)
3
4  # Plot the data
5  plot(time_series_data, main = "Simple Time Series Example")
6
7  # Autocorrelation function (ACF)
8  acf(time_series_data)
```

10. PROBABILITY DISTRIBUTIONS IN R

BASIC R FUNCTIONS FOR DISTRIBUTIONS

Since R is a software tool developed for statistical purposes, it provides a powerful set of built-in functions for working with probability distributions. These functions allow users to compute probabilities, generate random samples, and visualize distributions easily. They follow a consistent naming pattern that helps users intuitively apply statistical concepts directly in their code.

In R, most distribution-related functions use a simple prefix system. Table 3.1 gives a summary of the prefixes.

Table 3.1. Prefixes and distributions

Prefix	Distributions	Example
d*	density or probability mass functions (PMF/PDF)	<code>dnorm()</code> for the normal distribution's density.
d* p*	for cumulative distribution functions (CDF)	<code>pnorm()</code> for the cumulative probability of a normal distribution.
q*	quantile functions (inverse CDF)	<code>qnorm()</code> to find quantiles of a normal distribution.
r*	random number generation	<code>rnorm()</code> to generate random samples from a normal distribution.

10.1.1. PROBABILITY DENSITY/MASS FUNCTION EXERCISES

Example 1: Calculate the probability density of a normal distribution with a mean of 50 and a standard deviation of 10 at the point $x = 60$.

```
# Given parameters
mean <- 50
sd <- 10
```



```
x <- 60
# Calculate the probability density
density <- dnorm(x, mean = mean, sd = sd)
print(paste("The probability density at x =", x, "is",
density))
```

Example 2: Calculate and compare the probability densities of a normal distribution with a mean of 100 and a standard deviation of 15 at the points $x = 85, 100$, and 115 . Visualize the results using a plot.

```
# Given parameters
mean <- 100
sd <- 15
x_values <- c(85, 100, 115)
# Calculate the probability densities
densities <- dnorm(x_values, mean = mean, sd = sd)
# Print results
for (i in 1:length(x_values)) {
  print(paste("The probability density at x =",
x_values[i], "is", densities[i]))
}
# Visualization
dev.new() # Opens a new plotting window
plot(x_values, densities, type = "b", col = "blue", pch =
19,
      main = "Probability Densities of N(100, 15^2)",
      xlab = "x values", ylab = "Density")
```

Example 3: Consider a mixture of two normal distributions: $N(50, 10^2)$ with weight 0.6 and $N(70, 15^2)$ with weight 0.4. Calculate the combined probability density at $x = 60$.

```
# Parameters for the first normal distribution
mean1 <- 50
```

```
sd1 <- 10
weight1 <- 0.6
# Parameters for the second normal distribution
mean2 <- 70
sd2 <- 15
weight2 <- 0.4
# Point of interest
x <- 60
# Calculate the density of each distribution at x
density1 <- dnorm(x, mean = mean1, sd = sd1)
density2 <- dnorm(x, mean = mean2, sd = sd2)
# Calculate the combined density
combined_density <- weight1 * density1 + weight2 * density2
# Print result
print(paste("The combined probability density at x =", x, "is",
combined_density))
```

10.1.2. CUMULATIVE DISTRIBUTION FUNCTION EXERCISES

Example 1: Find the cumulative probability $P(X \leq 1.5)$ for a standard normal distribution.

```
# Compute cumulative probability for X <= 1.5 in a
standard normal distribution
probability <- pnorm(1.5)
# Print the result
print(probability)
```

Example 2: Find the cumulative probability $P(X \leq 120)$ for a normal distribution with a mean of 100 and a standard deviation of 15.

```
# Compute cumulative probability for X <= 120 in N(100,
15^2)
```

```
probability <- pnorm(120, mean = 100, sd = 15)
# Print the result
print(probability)
```

Example 3: Calculate the probability $P(X > 130)$ for a normal distribution with a mean of 100 and a standard deviation of 15.

```
# Compute probability for  $P(X > 130)$  in  $N(100, 15^2)$ 
probability <- 1 - pnorm(130, mean = 100, sd = 15)
# Print the result
print(probability)
```

NOTE: `pnorm(130, mean = 100, sd = 15)` computes $P(X \leq 130)$, the cumulative probability up to $X=130$. Since we need $P(X > 130)$, we subtract this value from 1: $P(X > 130) = 1 - P(X \leq 130)$

10.1.3. QUANTILE FUNCTION (INVERSE CDF) EXERCISES

Example 1: Find the 90th percentile of a standard normal distribution.

```
# Compute the 90th percentile of a standard normal
distribution
percentile_90 <- qnorm(0.90)

# Print the result
print(percentile_90)
```

NOTE: `qnorm(0.90)` returns the value of X such that $P(X \leq X_{90}) = 0.90$ in a standard normal distribution (mean = 0, standard deviation = 1). This means that 90% of the values in the standard normal distribution are below this value.

Example 2: Find the 75th percentile of a normal distribution with a mean of 100 and a standard deviation of 15.

```
# Compute the 75th percentile of N(100, 15^2)
percentile_75 <- qnorm(0.75, mean = 100, sd = 15)
# Print the result
print(percentile_75)
```

Example 3: Find the 99th percentile of a normal distribution with a mean of 50 and a standard deviation of 10.

```
# Compute the 99th percentile of N(50, 10^2)
percentile_99 <- qnorm(0.99, mean = 50, sd = 10)

# Print the result
print(percentile_99)
```

10.1.4. RANDOM SAMPLING EXERCISES

Example 1: Generate 5 random samples from a standard normal distribution.

```
# Generate 5 random samples from a standard normal
distribution
samples <- rnorm(5)
# Print the generated samples
print(samples)
```

Example 2: Generate 10 random samples from a normal distribution with mean 100 and standard deviation 15.

```
# Generate 10 random samples from N(100, 15^2)
samples <- rnorm(10, mean = 100, sd = 15)
# Print the generated samples
print(samples)
```

NOTE: The output will differ each time you run the code since the values are randomly generated.

Example 3: Sample 7 values with replacement from a given vector.

```
# Define a vector
my_vector <- c(10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
# Sample 7 values with replacement
sampled_values <- sample(my_vector, size = 7, replace =
TRUE)
# Print the sampled values
print(sampled_values)
```

NOTE: The `replace = TRUE` argument ensures that values can be selected more than once.

Example 4: Perform bootstrap resampling (size = 1000) on a sample dataset.

```
# Load necessary library
library(boot)
# Define a sample dataset
sample_data <- c(5, 10, 15, 20, 25, 30, 35, 40, 45, 50)
# Define bootstrap function (mean as the statistic)
bootstrap_function <- function(data, indices) {
  return(mean(data[indices])) # Compute mean of resampled
data
}
# Perform bootstrap resampling (size = 1000)
set.seed(123) # For reproducibility
bootstrap_result <- boot(data = sample_data, statistic =
bootstrap_function, R = 1000)
# Print bootstrap results
```

```
print(bootstrap_result)
```

NOTES:

- `boot()` from the **boot** package performs bootstrap resampling.
- The function `bootstrap_function()` calculates the mean of resampled data.
- `R = 1000` specifies 1000 bootstrap resamples.
- `set.seed(123)` ensures reproducibility of results.
- The `boot()` function returns the bootstrap estimates of the sample statistic.

DISCRETE DISTRIBUTION IN R

10.2.1. BERNOULLI DISTRIBUTION: `rbinom()`, `dbinom()`

Example 1: A coin is tossed once. Let X be a random variable that takes value 1 if the coin lands on heads (success) and 0 if tails (failure). The probability of heads is 0.5.

- Simulate 10 outcomes using `rbinom()`.
- Find the probability of getting a head using `dbinom()`.
- Simulating 10 outcomes using `rbinom()`:

We will simulate 10 outcomes where each outcome represents a coin toss. The random variable X will take the value 1 for heads and 0 for tails. The `rbinom()` function is used to generate random values from a binomial distribution, which, in the case of a single toss ($n = 1$), is equivalent to a Bernoulli distribution.

```
# Simulate 10 coin tosses (1 trial, probability
of heads = 0.5)

outcomes <- rbinom(10, size = 1, prob = 0.5)

# Print the outcomes (1 = heads, 0 = tails)
print(outcomes)
```

- b) Finding the probability of getting a head using `dbinom()` :

The `dbinom()` function calculates the probability of a specific number of successes (heads) in a Bernoulli trial. Since the coin has a 0.5 probability of landing heads, we can calculate the probability of getting 1 head (success) in a single toss.

```
# Calculate the probability of getting a head (X = 1) in
one toss

prob_head <- dbinom(1, size = 1, prob = 0.5)

# Print the probability

print(prob_head)
```

Example 2: A basketball player has a 70% chance of making a free throw. Let X be a Bernoulli random variable indicating success (1) or failure (0).

- a) Simulate 15 free throws.
- b) Calculate the probability of a miss ($X = 0$).
- a) Simulate 15 free throws using `rbinom()`:

The probability of success (making the free throw) is 0.7, so the probability of failure (missing the free throw) is 0.3. We can use the `rbinom()` function to simulate 15 free throws.

```
# Simulate 15 free throws (1 = success, 0 = failure)

free_throws <- rbinom(15, size = 1, prob = 0.7)

# Print the outcomes (1 = success, 0 = miss)

print(free_throws)
```

- b) Calculate the probability of a miss ($X = 0$) using `dbinom()`:

Since the probability of success (making a free throw) is 0.7, the probability of a miss (failure) is $1 - 0.7 = 0.3$. We can use the `dbinom()` function to calculate the probability of getting a miss ($X = 0$).

```
# Calculate the probability of a miss (X = 0)

prob_miss <- dbinom(0, size = 1, prob = 0.7)

# Print the probability of a miss

print(prob_miss)
```

NOTE: The function `rbinom(15, size = 1, prob = 0.7)` generates 15 random values, each being either 1 (success) or 0 (failure), based on the Bernoulli distribution with a success probability of 0.7. With the code `dbinom(0, size = 1, prob = 0.7)` we compute the probability of getting a miss (failure, $X = 0$) in a single free throw, which will be 0.3 since the probability of a miss is $1 - 0.7$. The `free_throws` vector will show 15 random values representing the outcomes of the 15 simulated free throws (1 = success, 0 = miss). The `prob_miss` will show the probability of a miss, which is 0.3.

Example 3: A website has a 20% chance of a visitor clicking a specific ad. Let X be the Bernoulli random variable (1 for click, 0 for no click).

- a) Simulate 1000 visitors.
- b) Estimate the proportion of clicks.
- c) Compare the empirical and theoretical probabilities.

a) Simulate 1000 visitors using `rbinom()`.

The probability of a visitor clicking the ad is 0.2 (20%), so we can simulate 1000 visitors using the `rbinom()` function.

```
# Simulate 1000 visitors (1 = click, 0 = no click)
visitors <- rbinom(1000, size = 1, prob = 0.2)
# Print the first few simulated outcomes (1 = click, 0 = no click)
head(visitors)
```

b) To estimate the proportion of clicks, we can simply calculate the mean of the simulated outcomes. The mean of a binary random variable will give us the proportion of successes (clicks in this case).

```
# Estimate the proportion of clicks
proportion_clicks <- mean(visitors)
# Print the estimated proportion of clicks
print(proportion_clicks)
```

c) The **theoretical probability** of a click is 0.2, since the problem states that the probability of a visitor clicking the ad is 20%. The

empirical probability is the proportion of clicks we estimated from the simulated data.

```
# Theoretical probability of a click
theoretical_prob <- 0.2

# Compare the empirical and theoretical probabilities
cat("Empirical      Probability      of      Clicks:",
    proportion_clicks, "\n")
cat("Theoretical    Probability      of      Clicks:",
    theoretical_prob, "\n")
```

Explanations:

- `rbinom(1000, size = 1, prob = 0.2)` simulates 1000 visitors, where each visitor can either click the ad (1) or not (0) with a probability of 0.2 for clicking.
- `mean(visitors)` calculates the proportion of clicks by taking the average of the simulated outcomes.

The **theoretical probability** of a click is 0.2 (given in the problem), and we compare this with the **empirical probability** obtained from the simulation. The `proportion_clicks` will give the estimated proportion of clicks based on the simulated data, which should be close to 0.2 (but slightly different due to random variation). The comparison will show how close the empirical probability is to the theoretical probability (which is 0.2).

Example 4: A product has a **30% chance of being defective** during manufacturing. Let X be a Bernoulli random variable:

- $X = 1$ if the product is defective (30% probability)
- $X = 0$ if the product is non-defective (70% probability)

Tasks:

- Simulate 1000 products.
- Calculate empirical probabilities.
- Visualize the distribution of outcomes.
 - Simulate 1000 products using `rbinom()`:

The probability of a product being defective is 0.3, and the probability of it being non-defective is 0.7. We will use the `rbinom()` function to simulate 1000 products.

```
# Simulate 1000 products (1 = defective, 0 = non-
defective)

products <- rbinom(1000, size = 1, prob = 0.3)

# Print the first few simulated outcomes (1 = defective, 0
= non-defective)

head(products)
```

b) Calculate empirical probabilities:

To calculate the empirical probability of a defective product ($X = 1$) and a non-defective product ($X = 0$), we can compute the proportions of 1's and 0's in the simulated data.

```
# Calculate the proportion of defective products (X = 1)
proportion_defective <- mean(products)

# Calculate the proportion of non-defective products (X =
0)

proportion_non_defective <- 1 - proportion_defective

# Print the empirical probabilities

cat("Empirical Probability of Defective Products:",
proportion_defective, "\n")

cat("Empirical Probability of Non-Defective Products:",
proportion_non_defective, "\n")
```

c) Visualize the distribution of outcomes:

To visualize the distribution of outcomes (defective vs non-defective products), we can create a bar plot or a histogram.

```
# Plot the distribution of defective (1) vs non-defective
(0) products

library(ggplot2)

# Create a data frame for ggplot

df <- data.frame(Outcome = factor(products, levels = c(0,
1), labels = c("Non-Defective", "Defective")))

# Create a bar plot
```

```
ggplot(df, aes(x = Outcome)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribution of Defective and Non-
Defective Products",
        x = "Product Condition", y = "Count") +
  theme_minimal()
```

The `proportion_defective` and `proportion_non_defective` will provide the empirical probabilities of defective and non-defective products, respectively. The bar plot will show the number of defective and non-defective products in the simulated data, where you should expect the proportion of defective products to be around 0.3 (but with slight variation due to random sampling).

10.2.2. BINOMIAL DISTRIBUTION EXAMPLES

Example 1: A fair coin is tossed **10 times**. What is the probability of getting exactly **4 heads**?

```
# Parameters
n <- 10 # Number of trials
p <- 0.5 # Probability of success (getting heads)
x <- 4 # Desired number of successes

# Calculate Binomial Probability
probability <- dbinom(x, n, p)
print(probability)
```

Example 2: A factory produces **electronic chips**, and the probability of a defective chip is **0.02**. If a batch of **100** chips is tested, what is the probability that at most **3** are defective?

```
# Parameters
n <- 100 # Number of trials
```

```
p <- 0.02 # Probability of success (defective chip)
x <- 3     # At most 3 defective chips

# Calculate cumulative probability
probability <- pbinom(x, n, p)
print(probability)
```

Example 3: A **basketball player** has a **60%** chance of making a free throw. If they attempt **15** free throws in a game:

- a) What is the probability of making between **8 and 12** shots (inclusive)?
- b) What is the expected number of successful free throws?

Solution.

- a) Probability of making between 8 and 12 shots (inclusive):

This is a binomial distribution problem, where the probability of success (making a free throw) is 0.6, the number of trials (free throws) is 15, and we want to calculate the probability of making between 8 and 12 successful shots (inclusive). We will use the `dbinom()` function to find the probability for each number of successful shots and sum them.

The probability $P(8 \leq X \leq 12)$ is the sum of the probabilities of making 8, 9, 10, 11, or 12 successful free throws.

```
# Probability of making between 8 and 12 successful free
throws
probability_8_to_12 <- sum(dbinom(8:12, size = 15, prob =
0.6))
# Print the probability
print(probability_8_to_12)
```

- b) Expected number of successful free throws:

The expected value $E(X)$ for a binomial distribution is given by:
 $E(X) = n \cdot p$

Where: n is the number of trials (15 free throws) and p is the probability of success (0.6). You can calculate the expected value directly:

```
# Expected number of successful free throws
expected_value <- 15 * 0.6
# Print the expected value
print(expected_value)
```

The `probability_8_to_12` will give the probability of making between 8 and 12 successful free throws in a game. The `expected_value` will give the expected number of successful free throws, which will be $15 \times 0.6 = 9$.

10.2.3. POISSON DISTRIBUTION: `rpois()` and real-world examples

Example 1: A factory produces electronic components. On average, 3 defective components are found in every 100-unit batch. Simulate the number of defective components in 20 batches.

To simulate the number of defective components in 20 batches, we can use the `rpois()` function in R to generate Poisson-distributed random numbers.

```
# Set the rate of defective components per batch (lambda =
3)
lambda <- 3
# Simulate the number of defective components in 20
batches
defective_components <- rpois(20, lambda)
# Print the simulated number of defective components in
each batch
print(defective_components)
```

The output might look something like this (values will vary because of the randomness in simulation):

```
[1] 2 5 3 3 4 2 1 3 3 4 3 2 3 2 2 4 3 2 2 4
```

Example 2: A website receives an average of **12 visitors per hour**. Simulate the number of visitors arriving in **24 hours** and plot the results.

To simulate the number of visitors arriving at a website over a 24-hour period, given an average of 12 visitors per hour, you can use the Poisson distribution. The Poisson distribution is appropriate here since the number of visitors arriving per hour is independent and occurs at a constant rate. The rate λ for the Poisson distribution is 12 visitors per hour, and you want to simulate this for 24 hours.

```
# Set the rate of visitors per hour (lambda = 12)
lambda <- 12

# Simulate the number of visitors arriving in 24 hours
visitors_24_hours <- rpois(24, lambda)

# Print the simulated number of visitors in each hour
print(visitors_24_hours)

# Plot the results (number of visitors each hour)
barplot(visitors_24_hours, main = "Simulated Number of
Visitors in 24 Hours",
        xlab = "Hour", ylab = "Number of Visitors", col =
"skyblue", border = "black")
```

NOTE: The `visitors_24_hours` will contain 24 numbers representing the simulated number of visitors arriving each hour. The bar plot will show the distribution of visitors across the 24 hours, where each bar represents the number of visitors in a specific hour.

Example 3: A hospital emergency room receives an average of **8 patients per hour**. Simulate **7 days (168 hours)** of patient arrivals and analyze the distribution.

To simulate the number of patients arriving at a hospital emergency room over 7 days (168 hours), given an average of 8 patients per hour, you can use the **Poisson distribution**. The rate (λ lambda) for the Poisson distribution is 8 patients per hour, and you need to simulate the arrivals for 168 hours.

Steps:

1. **Simulate patient arrivals:** Use `rpois()` to generate the number of patients arriving each hour for 168 hours.
2. **Analyze the distribution:** You can analyze the distribution by calculating the summary statistics and visualizing it using histograms.

```
# Set the rate of patients per hour (lambda = 8)
lambda <- 8

# Simulate the number of patients arriving in 168 hours (7
days)
patient_arrivals <- rpois(168, lambda)

# Print the first few simulated values (number of patients
per hour)
head(patient_arrivals)

# Calculate summary statistics
summary_stats <- summary(patient_arrivals)
cat("Summary Statistics:\n", summary_stats, "\n")

# Plot the distribution of patient arrivals
hist(patient_arrivals, breaks = 20, main = "Distribution
of Patient Arrivals (168 Hours)",
      xlab = "Number of Patients per Hour", col =
"lightblue", border = "black")

# Additional analysis: Calculate the mean and standard
deviation
mean_patients <- mean(patient_arrivals)
sd_patients <- sd(patient_arrivals)

cat("Mean number of patients per hour:", mean_patients,
"\n")

cat("Standard deviation of patients per hour:",
sd_patients, "\n")
```

NOTES:

Summary Statistics will show the minimum, maximum, median, mean, and quartiles of the number of patients arriving per hour over the 7

days. A histogram will display the distribution of the number of patients arriving each hour. It should roughly follow a Poisson distribution with a mean around 8 patients per hour. The mean will be close to 8, as it was set as the rate for the Poisson distribution, and the standard deviation will provide insight into the variability of the number of patient arrivals per hour.

10.2.4. GEOMETRIC DISTRIBUTION: MODELLING THE NUMBER OF TRIALS UNTIL FIRST SUCCESS

Example 1: A fair coin is flipped until the first **heads** appears. Simulate this process **100 times**, visualize the distribution, and calculate the expected number of flips.

```
# Load necessary library
library(ggplot2)

# Parameters
p <- 0.5 # Probability of heads
n <- 100 # Number of simulations

# Generate random samples
flips_until_heads <- rgeom(n, p) + 1

# Expected number of flips before first heads
expected_flips <- 1 / p
print(paste("Expected number of flips before first
heads:", expected_flips))

# Create dataframe
df <- data.frame(Flips = flips_until_heads)

# Plot histogram
```



```
ggplot(df, aes(x = Flips)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color =
"black") +
  ggtitle("Number of Flips Until First Heads (100
Simulations)") +
  xlab("Number of Flips") + ylab("Frequency") +
  theme_minimal()
```

Example 2: A batch of **light bulbs** is tested for failure. Each bulb has a **0.1 probability** of failing. Simulate how many bulbs need to be tested before finding the first defective one, repeating this **20 times**.

```
# Parameters
p <- 0.1 # Probability of failure
n <- 20  # Simulating 20 batches

# Generate random samples
bulbs_tested <- rgeom(n, p) + 1
print(bulbs_tested)
```

Example 3: A salesperson has a **30%** chance of making a successful sale per call. Simulate how many calls are needed to make the first sale in **50 trials** and visualize the results.

```
# Load library
library(ggplot2)

# Parameters
p <- 0.3 # Probability of success per call
n <- 50  # Simulating 50 trials

# Generate random samples
calls_needed <- rgeom(n, p) + 1
```

```
df <- data.frame(Calls = calls_needed)

# Plot histogram
ggplot(df, aes(x = Calls)) +
  geom_histogram(binwidth = 1, fill = "green", color =
"black") +
  ggtitle("Sales Calls Before First Success") +
  xlab("Number of Calls") + ylab("Frequency") +
  theme_minimal()
```

10.3. CONTINUOUS DISTRIBUTIONS IN R

10.3.1. UNIFORM DISTRIBUTION: `runif()`

Example 1: Generate **10 random numbers** from a **Uniform(0,1)** distribution.

```
# Generate 10 random numbers between 0 and 1
random_numbers <- runif(10)
print(random_numbers)
```

Example 2: Simulate **15 random temperatures** between **20°C** and **35°C**.

```
# Parameters
n <- 15      # Number of random temperatures
min_temp <- 20 # Minimum temperature
max_temp <- 35 # Maximum temperature

# Generate random temperatures
temperatures <- runif(n, min_temp, max_temp)
print(temperatures)
```

Example 3: Generate **500 random values** from a **Uniform(5,15)** distribution and plot a **histogram**.

```
# Load necessary library
library(ggplot2)

# Parameters
n <- 500    # Number of random values
min_val <- 5  # Minimum value
max_val <- 15 # Maximum value

# Generate random values
random_values <- runif(n, min_val, max_val)

# Create dataframe
df <- data.frame(Values = random_values)

# Plot histogram
ggplot(df, aes(x = Values)) +
  geom_histogram(binwidth = 1, fill = "blue",
color = "black", alpha = 0.7) +
  ggtitle("Histogram of Uniform(5,15)
Distribution") +
  xlab("Values") + ylab("Frequency") +
  theme_minimal()
```

10.3.2. NORMAL DISTRIBUTION: `rnorm()`, `dnorm()` — WITH VISUALIZATION EXAMPLES (`hist()`, `curve()`)

Example 1: Generate **10 random numbers** from a **standard normal distribution** ($N(0, 1)$) and display them.

```
# Generate 10 random numbers from N(0,1)
random_numbers <- rnorm(10)
print(random_numbers)
```

Example 2: Generate **1000 random numbers** from $N(50, 10^2)$ (mean = 50, SD = 10) and visualize the distribution with a histogram.

```
# Generate 1000 random numbers from N(50, 10)
n <- 1000
mean_val <- 50
sd_val <- 10

random_values <- rnorm(n, mean_val, sd_val)

# Plot histogram
hist(random_values, breaks = 30, col = "skyblue",
probability = TRUE,
      main = "Histogram of Normal(50, 10) Distribution",
xlab = "Values")

# Add a normal density curve
curve(dnorm(x, mean = mean_val, sd = sd_val), col = "red",
lwd = 2, add = TRUE)
```

Example 3: Generate and visualize three normal distributions with different means and standard deviations on the same plot.

```
# Load necessary library
library(ggplot2)

# Generate x values
x_vals <- seq(-20, 100, length.out = 1000)

# Compute densities for three normal distributions
density1 <- dnorm(x_vals, mean = 50, sd = 10)
density2 <- dnorm(x_vals, mean = 30, sd = 5)
density3 <- dnorm(x_vals, mean = 70, sd = 15)

# Create dataframe
df <- data.frame(x = rep(x_vals, 3),
                  Density = c(density1, density2,
                              density3),
                  Distribution = rep(c("N(50,10)",
                                       "N(30,5)", "N(70,15)"), each = length(x_vals)))

# Plot
ggplot(df, aes(x = x, y = Density, color = Distribution))
+
  geom_line(size = 1.2) +
  ggtitle("Comparison of Different Normal Distributions")
+
  xlab("Values") + ylab("Density") +
  theme_minimal()
```

10.3.3. EXPONENTIAL DISTRIBUTION: MODELING WAITING TIMES (rexp())

Example 1: A call center receives customer calls at an average rate of 3 calls per minute. Simulate the waiting times for 10 calls.

```
# Parameters
lambda <- 3   # Average rate (calls per minute)
n <- 10       # Number of waiting times to generate

# Generate random waiting times
waiting_times <- rexp(n, rate = lambda)
print(waiting_times)
```

Example 2: A bank's ATM serves customers at a rate of 5 customers per hour. Simulate the waiting times for 200 customers and visualize the distribution.

```
# Load necessary library
library(ggplot2)

# Parameters
lambda <- 5   # Service rate (customers per hour)
n <- 200      # Number of customers

# Generate random waiting times
waiting_times <- rexp(n, rate = lambda)

# Create dataframe
df <- data.frame(Waiting_Time = waiting_times)

# Plot histogram
```

```
ggplot(df, aes(x = Waiting_Time)) +
  geom_histogram(binwidth = 0.1, fill = "blue", color =
"black", alpha = 0.7) +
  ggtitle("Exponential Distribution of ATM Waiting Times")
+
  xlab("Waiting Time (hours)") + ylab("Frequency") +
  theme_minimal()
```

Example 3: In a hospital emergency room, patients arrive every 10 minutes on average ($\lambda = 1/10$). Find the probability that a patient waits more than 15 minutes for the next arrival.

```
# Parameter
lambda <- 1/10 # Average arrivals per minute

# Probability that waiting time > 15 minutes
p_wait_15 <- 1 - pexp(15, rate = lambda)

print(paste("Probability of waiting more than 15
minutes:", round(p_wait_15, 4)))
```

10.3.4. GAMMA AND BETA DISTRIBUTIONS: USAGE IN BAYESIAN STATISTICS

Example 1: A coin is suspected to be biased. The prior belief about the probability of heads follows a **Beta(2,2)** distribution. Generate and visualize **1000 samples**.

```
# Load necessary library
library(ggplot2)

# Parameters for Beta prior
```

```
alpha <- 2 # Prior successes
beta <- 2 # Prior failures

# Generate 1000 samples from Beta(2,2)
theta_samples <- rbeta(1000, alpha, beta)

# Create dataframe
df <- data.frame(Theta = theta_samples)

# Plot histogram
ggplot(df, aes(x = Theta)) +
  geom_histogram(binwidth = 0.05, fill = "skyblue", color
= "black", alpha = 0.7) +
  ggtitle("Beta(2,2) Prior Distribution for Coin Bias") +
  xlab("Theta (Probability of Heads)") + ylab("Frequency")
+
  theme_minimal()
```

Example 2: A website tracks the conversion rate of visitors.

- Prior belief: Beta(2,2)
- Observed data: 30 visitors, 12 made a purchase
- Compute and plot the posterior distribution.

```
# Prior parameters
alpha_prior <- 2
beta_prior <- 2

# Observed data
successes <- 12
trials <- 30

# Posterior parameters (Beta update)
```



```
alpha_post <- alpha_prior + successes
beta_post <- beta_prior + (trials - successes)

# Generate samples from posterior Beta distribution
posterior_samples <- rbeta(1000, alpha_post, beta_post)

# Create dataframe
df <- data.frame(Theta = posterior_samples)

# Plot histogram
ggplot(df, aes(x = Theta)) +
  geom_histogram(binwidth = 0.02, fill = "blue", color =
"black", alpha = 0.7) +
  ggtitle("Posterior Distribution of Conversion Rate") +
  xlab("Theta (Conversion Probability)") +
  ylab("Frequency") +
  theme_minimal()
```

- In **Bayesian updating**, a **Beta prior** combines with **Binomial data** to form a **Beta posterior**: $\text{Beta}(\alpha + \text{successes}, \beta + \text{failures})$
- **New parameters**: $\text{Beta}(2 + 12, 2 + (30 - 12)) = \text{Beta}(14, 20)$.
- The **posterior** reflects updated belief about conversion probability.

Example 3: A factory receives defective parts at an average rate.

- **Prior belief:** The defect rate follows $\text{Gamma}(3, 1)$ (mean = 3 defects per hour).
- **Observed data:** 20 defects in 5 hours (Poisson likelihood).
- Compute and visualize the **posterior distribution** for the defect rate λ .

```
# Prior parameters (Gamma prior for Poisson rate)
alpha_prior <- 3
beta_prior <- 1
```

```
# Observed data
total_defects <- 20
total_time <- 5 # in hours

# Posterior parameters (Gamma update)
alpha_post <- alpha_prior + total_defects
beta_post <- beta_prior + total_time

# Generate samples from posterior Gamma distribution
lambda_samples <- rgamma(1000, shape = alpha_post, rate =
beta_post)

# Create dataframe
df <- data.frame(Lambda = lambda_samples)

# Plot density
ggplot(df, aes(x = Lambda)) +
  geom_density(fill = "red", alpha = 0.5) +
  ggtitle("Posterior Distribution of Defect Rate (Gamma)")
+
  xlab("Lambda (Defects per Hour)") + ylab("Density") +
  theme_minimal()
```

- In Bayesian Poisson modeling, the Gamma prior updates with Poisson data to form a Gamma posterior: $\text{Gamma}(\alpha + \sum x_i, \beta + T)$
- Updated parameters: $\text{Gamma}(3+20, 1+5) = \text{Gamma}(23, 6)$.
- The posterior distribution represents our updated belief about the defect rate.

FITTING DISTRIBUTION TO DATA**10.4.1. USING `fitdistr()` from the MASS package**

Example 1: Generate 1000 random numbers from $N(50, 10^2)$, then fit a normal distribution and extract the estimated mean and standard deviation.

```
# Load necessary libraries
library(MASS)

# Generate random data from N(50, 10)
set.seed(123)
data <- rnorm(1000, mean = 50, sd = 10)

# Fit normal distribution
fit <- fitdistr(data, densfun = "normal")

# Print estimated parameters
print(fit)
```

Example 2: Simulate waiting times at a bus stop, fit an Exponential(λ) distribution, and extract the estimated rate λ .

```
# Load necessary libraries
library(MASS)

# Simulate waiting times (Exponential distribution with
rate = 1/5)
set.seed(123)
waiting_times <- rexp(200, rate = 1/5)

# Fit exponential distribution
```

```
fit <- fitdistr(waiting_times, densfun = "exponential")
```

```
# Print estimated rate (lambda)
```

```
print(fit)
```

Example 3: Simulate service times at a call center following a Gamma distribution (`shape = 3`, `rate = 2`), estimate the shape and rate, and overlay the fitted distribution on a histogram.

```
# Load necessary libraries
```

```
library(MASS)
```

```
library(ggplot2)
```

```
# Simulate service times (Gamma distribution)
```

```
set.seed(123)
```

```
service_times <- rgamma(500, shape = 3, rate = 2)
```

```
# Fit Gamma distribution
```

```
fit <- fitdistr(service_times, densfun = "gamma")
```

```
# Extract estimated parameters
```

```
shape_est <- fit$estimate["shape"]
```

```
rate_est <- fit$estimate["rate"]
```

```
# Create a dataframe
```

```
df <- data.frame(Service_Times = service_times)
```

```
# Plot histogram with fitted Gamma curve
```

```
ggplot(df, aes(x = Service_Times)) +
```

```
  geom_histogram(aes(y = ..density..), bins = 30, fill =  
"blue", color = "black", alpha = 0.6) +
```

```
  stat_function(fun = dgamma, args = list(shape =  
shape_est, rate = rate_est),
```

```

        color = "red", size = 1.2) +
  ggtitle("Fitted Gamma Distribution to Service Times") +
  xlab("Service Time (minutes)") + ylab("Density") +
  theme_minimal()

```

10.4.2. GOODNESS-OF-FIT TESTS (KOLMOGOROV-SMIRNOV TEST WITH `ks.test()`)

Example 1: Test if a sample of 1000 random numbers generated from a $N(0,1)$ distribution follows the standard normal distribution.

```

# Generate random sample from N(0,1)
set.seed(123)
sample_data <- rnorm(1000)

# Kolmogorov-Smirnov test for normality (comparing to
N(0,1))
ks_test_result <- ks.test(sample_data, "pnorm", mean = 0,
sd = 1)

# Print test result
print(ks_test_result)

```

Example 2: Test if a sample of 500 random waiting times generated from an Exponential distribution ($\lambda = 1/4$) follows the theoretical exponential distribution with rate = $1/4$.

```

# Generate random waiting times from an
exponential distribution (lambda = 1/4)
set.seed(123)
waiting_times <- rexp(500, rate = 1/4)

# Kolmogorov-Smirnov test for exponentiality

```

```
ks_test_result <- ks.test(waiting_times, "pexp",  
rate = 1/4)
```

```
# Print test result  
print(ks_test_result)
```

Example 3: Test if a sample of 1000 random numbers generated from a $\text{Uniform}(0,1)$ distribution follows a uniform distribution.

```
# Generate random numbers from a uniform distribution  
(U(0,1))  
set.seed(123)  
uniform_data <- runif(1000)
```

```
# Kolmogorov-Smirnov test for uniformity (comparing to  
U(0,1))  
ks_test_result <- ks.test(uniform_data, "punif", min = 0,  
max = 1)
```

```
# Print test result  
print(ks_test_result)
```

10.4.3. COMPARING EMPIRICAL VS. THEORETICAL DISTRIBUTIONS

Example 1: You have a sample of heights of 500 adults. You suspect that the heights follow a Normal distribution. Compare the empirical distribution of the heights to the theoretical Normal distribution.

```
# Generate a sample of heights (normally distributed, mean  
= 170 cm, sd = 10 cm)  
set.seed(123)
```

```

heights <- rnorm(500, mean = 170, sd = 10)

# Plot histogram of empirical distribution
hist(heights, probability = TRUE, main = "Empirical vs
Theoretical Normal Distribution",
      xlab = "Height (cm)", col = "skyblue", border =
"black",
      breaks = 20)

# Overlay theoretical normal distribution curve
curve(dnorm(x, mean = mean(heights), sd = sd(heights)),
      col = "red", lwd = 2, add = TRUE)

# Perform Kolmogorov-Smirnov test for normality
ks_test <- ks.test(heights, "pnorm", mean = mean(heights),
sd = sd(heights))
print(ks_test)

```

Example 2: A store records customer arrival times. You suspect that the arrival times follow an Exponential distribution (common in queueing theory). Compare the empirical distribution of arrival times to the theoretical Exponential distribution.

```

# Simulate customer arrival times with exponential
distribution (lambda = 0.2)
set.seed(123)
arrival_times <- rexp(300, rate = 0.2)

# Plot histogram of empirical distribution
hist(arrival_times, probability = TRUE, main = "Empirical
vs Theoretical Exponential Distribution",
      xlab = "Arrival Time (minutes)", col = "lightgreen",
border = "black",
      breaks = 20)

```

```
# Overlay theoretical exponential distribution curve
curve(dexp(x, rate = 0.2), col = "red", lwd = 2, add =
TRUE)

# Perform Kolmogorov-Smirnov test for exponentiality
ks_test <- ks.test(arrival_times, "pexp", rate = 0.2)
print(ks_test)
```

Example 3: A factory records the number of defects observed during production runs. The number of defects is expected to follow a Poisson distribution. Compare the empirical distribution of defects with the theoretical Poisson distribution.

```
# Simulate number of defects (Poisson distribution with
lambda = 3)
set.seed(123)
defects <- rpois(400, lambda = 3)

# Plot histogram of empirical distribution
hist(defects, probability = TRUE, main = "Empirical vs
Theoretical Poisson Distribution",
      xlab = "Number of Defects", col = "lightcoral",
border = "black",
      breaks = 10)

# Overlay theoretical Poisson distribution curve
max_val <- max(defects)
x_vals <- 0:max_val
pois_vals <- dpois(x_vals, lambda = 3)
lines(x_vals, pois_vals, col = "red", lwd = 2)
```



```
# Perform Chi-square test for goodness of fit (Poisson
distribution)
observed <- table(defects)
expected <- dpois(as.numeric(names(observed)), lambda = 3)
* sum(observed)
chisq_test <- chisq.test(observed, p = expected /
sum(expected), rescale.p = TRUE)
print(chisq_test)
```

10.5. SIMULATION AND RANDOM SAMPLING

10.5.1. GENERATING RANDOM SAMPLES FROM DISTRIBUTIONS

Example 1: Generate 1000 random samples from a normal distribution with mean = 5 and standard deviation = 2.

```
# Generate 1000 random samples from a Normal distribution
N(5, 2)
set.seed(123)
samples_normal <- rnorm(1000, mean = 5, sd = 2)

# Plot histogram of the random samples
hist(samples_normal, breaks = 30, main = "Random Samples
from N(5,2)",
      xlab = "Sample Values", col = "lightblue", border =
"black")
```

Example 2: Generate 500 random samples from an Exponential distribution with rate parameter $\lambda = 0.2$ (mean = 5).

```
# Generate 500 random samples from an Exponential
distribution with rate = 0.2
set.seed(123)
```

```

samples_exponential <- rexp(500, rate = 0.2)

# Plot histogram of the random samples
hist(samples_exponential, breaks = 30, main = "Random
Samples from Exp(0.2)",
      xlab = "Sample Values", col = "lightgreen", border =
"black")

```

Example 3: Simulate 200 random samples from a Beta distribution with shape parameters $\alpha = 2$ and $\beta = 5$. Visualize the generated samples and check how the shape of the distribution changes with different parameters.

```

# Generate 200 random samples from a Beta
distribution with shape parameters alpha = 2 and beta
= 5

set.seed(123)

samples_beta <- rbeta(200, shape1 = 2, shape2 =
5)

# Plot histogram of the random samples
hist(samples_beta, breaks = 20, main = "Random
Samples from Beta(2, 5)",
      xlab = "Sample Values", col = "salmon",
border = "black")

```

Example 4: Simulate the outcomes of 1000 rolls of a fair 6-sided die using a Multinomial distribution. Each roll results in one of the six faces (1, 2, 3, 4, 5, or 6).

```

# Define the probability of each outcome (fair die, so
each face has a 1/6 probability)

probabilities <- rep(1/6, 6)

```

```
# Generate 1000 rolls using a Multinomial distribution
(size = 1000, probabilities for each face)

set.seed(123)

samples_multinomial <- rmultinom(1, size = 1000, prob =
probabilities)

# Convert the result to a more readable format
samples_multinomial <- as.vector(samples_multinomial)

# Plot the frequencies of each outcome (1 to 6)
barplot(samples_multinomial, main = "Results of 1000 Rolls
of a Fair Die",
        xlab = "Outcome (1 to 6)", ylab = "Frequency", col
= "lightblue")
```

10.5.2. SIMULATING DICE ROLLS AND COIN FLIPS

Example 1: Simulate the outcome of 100 coin flips with a fair coin (50% heads, 50% tails).

```
# Simulate 100 coin flips (1 for heads, 0 for tails)

set.seed(123)

coin_flips <- sample(c("Heads", "Tails"), size = 100,
replace = TRUE)

# View the first 10 flips
head(coin_flips, 10)

# Plot the frequencies of heads and tails
table(coin_flips)

barplot(table(coin_flips), main = "Coin Flip Outcomes",
col = "lightblue", ylab = "Frequency")
```

Example 2: Simulate **500 rolls** of a single 6-sided die and plot the distribution of outcomes.

```
# Simulate 500 rolls of a 6-sided die
set.seed(123)
dice_rolls <- sample(1:6, size = 500, replace = TRUE)

# View the first 10 rolls
head(dice_rolls, 10)

# Plot the frequencies of each die face
barplot(table(dice_rolls), main = "Dice Roll Outcomes",
col = "lightgreen", xlab = "Die Face", ylab = "Frequency")
```

Example 3: Simulate 1000 rolls of two 6-sided dice and calculate the sum of the two dice for each roll. Plot the distribution of the sum.

```
# Simulate 1000 rolls of two 6-sided dice
set.seed(123)
dice1 <- sample(1:6, size = 1000, replace = TRUE)
dice2 <- sample(1:6, size = 1000, replace = TRUE)

# Calculate the sum of the two dice for each roll
sum_dice <- dice1 + dice2

# View the first 10 sums
head(sum_dice, 10)

# Plot the distribution of sums
hist(sum_dice, breaks = 11, main = "Sum of Two Dice
Rolls", xlab = "Sum", col = "lightyellow", border = "black")
```

11. DESCRIPTIVE STATISTICS IN R

11.1. DATA PREPARATION IN R

11.1.1. IMPORTING DATA: `read.csv()`, `read.table()`, `read_excel()`

Example 1: Import a CSV file containing sales data.

```
# Importing data from a CSV file
sales_data <- read.csv("sales_data.csv")

# View the first few rows
head(sales_data)
```

Example 2: Import a text file with tab-separated data.

```
# Importing data from a tab-separated file
employee_data <- read.table("employee_data.txt", sep =
"\t", header = TRUE)

# Check the structure of the data
str(employee_data)
```

Example 3: Import an Excel file containing customer information. (Importing Data: Using `read_excel()` from `readxl` Package)

```
# Install and load the readxl package
install.packages("readxl")
library(readxl)

# Import data from an Excel file
customer_data <- read_excel("customer_data.xlsx")
```

```
# View the first few rows  
head(customer_data)
```

11.1.2. CHECKING STRUCTURE: `str()`, `summary()`, `head()`, `tail()`

Example 1: Check the structure of a dataset containing product reviews.

```
# Check the structure of the product reviews dataset  
str(product_reviews)
```

Example 2: Get a summary of a dataset containing student grades.

```
# Get a summary of the student grades dataset  
summary(student_grades)
```

Example 3: Display the first 10 rows of a dataset containing financial transactions.

```
# View the first 10 rows of financial transactions dataset  
head(financial_transactions, 10)
```

Example 4: Display the last 10 rows of a dataset containing sales data.

```
# View the last 10 rows of the sales dataset  
tail(sales_data, 10)
```

**11.1.3. HANDLING MISSING VALUES: `na.omit()`,
`is.na()`, `complete.cases()`**

Example 1: Remove rows with missing values from a dataset containing employee information.

```
# Remove rows with missing values
employee_data_clean <- na.omit(employee_data)

# Check the structure of the cleaned data
str(employee_data_clean)
```

Example 2: Identify missing values in a dataset containing exam scores.

```
# Check for missing values in the dataset
missing_values <- is.na(exam_scores)

# View the rows with missing values
exam_scores[missing_values, ]
```

Example 3: Remove rows with missing values from a dataset and get the cleaned dataset.

```
# Remove rows with missing values using complete.cases
cleaned_data <-
customer_data[complete.cases(customer_data), ]

# Check the structure of the cleaned data
str(cleaned_data)
```

11.2. MEASURES OF CENTRAL TENDENCY

Example 1: Calculate the **average income** of individuals in a dataset.

```
# Example dataset (income data)
income_data <- c(40000, 45000, 60000, 55000, 75000, 80000,
50000, 62000, 70000, 64000)

# Calculate the mean (average income)
average_income <- mean(income_data)

# Display the result
cat("The average income is:", average_income)
```

Example 2: Calculate the **median test score** from a dataset of students' scores.

```
# Example dataset (test scores)
test_scores <- c(85, 90, 75, 92, 88, 84, 95, 78, 89, 91)

# Calculate the median test score
median_score <- median(test_scores)

# Display the result
cat("The median test score is:", median_score)
```

Example 3: Since R lacks a built-in mode function, we can create a **custom function** to calculate the **mode** (most frequent value) in a dataset.

```
# Example dataset (ages of people)
age_data <- c(25, 30, 25, 35, 40, 30, 25, 30, 35, 40, 25)

# Custom function to calculate the mode
```



```
get_mode <- function(x) {  
  uniq_vals <- unique(x)  
  freq_vals <- tabulate(match(x, uniq_vals))  
  mode_val <- uniq_vals[freq_vals == max(freq_vals)]  
  return(mode_val)  
}  
  
# Calculate the mode  
mode_age <- get_mode(age_data)  
  
# Display the result  
cat("The mode age is:", mode_age)
```

Example 4: Combine the **mean**, **median**, and **mode** to analyze central tendency for a dataset of **sales figures**.

```
# Example dataset (sales figures in dollars)  
sales_data <- c(500, 550, 600, 580, 620, 580, 600, 630,  
620, 580)  
  
# Calculate mean, median, and mode  
mean_sales <- mean(sales_data)  
median_sales <- median(sales_data)  
mode_sales <- get_mode(sales_data)  
  
# Display the results  
cat("The mean sales is:", mean_sales, "\n")  
cat("The median sales is:", median_sales, "\n")  
cat("The mode sales is:", mode_sales)
```

Example 5: Calculate the **average price** of cars from a dataset of car prices.

```
# Example dataset (car prices in thousands)
car_prices <- c(25, 30, 35, 40, 28, 33, 36, 31, 29, 38)

# Calculate the mean (average price)
average_price <- mean(car_prices)

# Display the result
cat("The average car price is:", average_price)
```

Example 6: Calculate the **median age** of employees in a company.

```
# Example dataset (ages of employees)
employee_ages <- c(25, 30, 35, 40, 50, 29, 32, 28, 45, 38)

# Calculate the median age
median_age <- median(employee_ages)

# Display the result
cat("The median age of employees is:", median_age)
```

Example 7: Find the **most popular color** chosen by customers for products (mode).

```
# Example dataset (colors chosen by customers)
colors_chosen <- c("Red", "Blue", "Green", "Red", "Blue",
"Red", "Yellow", "Blue", "Red", "Green")

# Custom function to calculate the mode
get_mode <- function(x) {
  uniq_vals <- unique(x)
  freq_vals <- tabulate(match(x, uniq_vals))
  mode_val <- uniq_vals[freq_vals == max(freq_vals)]
}
```

```
    return(mode_val)
  }

  # Calculate the mode (most popular color)
  mode_color <- get_mode(colors_chosen)

  # Display the result
  cat("The most popular color is:", mode_color)
```

Example 8: Compare the **mean** and **median** of a dataset with an **outlier** to demonstrate their difference.

```
# Example dataset with an outlier (house prices in
thousands)
house_prices <- c(150, 160, 170, 180, 200, 250, 300, 350,
5000)

# Calculate mean and median
mean_price <- mean(house_prices)
median_price <- median(house_prices)

# Display the result
cat("The mean house price is:", mean_price, "\n")
cat("The median house price is:", median_price)
```

Example 9: Find the **mode** of a dataset with **multiple modes** (e.g., test scores).

```
# Example dataset (test scores with multiple modes)
test_scores <- c(88, 92, 85, 90, 85, 92, 90, 92, 88, 85)

# Custom function to calculate the mode
get_mode <- function(x) {
  uniq_vals <- unique(x)
```

```
freq_vals <- tabulate(match(x, uniq_vals))
mode_val <- uniq_vals[freq_vals == max(freq_vals)]
return(mode_val)
}
```

```
# Calculate the mode (most frequent scores)
mode_scores <- get_mode(test_scores)

# Display the result
cat("The mode test scores are:", mode_scores)
```

Example 10: Calculate the **mean** and **median** income for different **regions** in a country from a grouped dataset.

```
# Example dataset (incomes by region)
region_data <- data.frame(
  region = c("North", "South", "East", "West"),
  income = c(55000, 60000, 52000, 63000)
)

# Calculate the mean and median income by region
mean_income_region <- mean(region_data$income)
median_income_region <- median(region_data$income)

# Display the result
cat("The mean income by region is:", mean_income_region,
"\n")
cat("The median income by region is:",
median_income_region)
```

Example 11: Calculate and compare the **mean**, **median**, and **mode** for a large dataset (e.g., product ratings).

```
# Example dataset (product ratings out of 5)
```

```
4) ratings_data <- c(5, 4, 3, 4, 5, 3, 5, 4, 3, 5, 4, 4, 3,
# Calculate mean, median, and mode
mean_rating <- mean(ratings_data)
median_rating <- median(ratings_data)
mode_rating <- get_mode(ratings_data)

# Display the result
cat("The mean rating is:", mean_rating, "\n")
cat("The median rating is:", median_rating, "\n")
cat("The mode rating is:", mode_rating)
```

11.3. MEASURES OF DISPERSION

Real-world examples include the following R functions:

- **Range:** `range()`
- **Variance:** `var()`
- **Standard Deviation:** `sd()`
- **Interquartile Range (IQR):** `IQR()`

Example 1: Calculate the **range**, **variance**, **standard deviation**, and **interquartile range (IQR)** for the **salaries** of employees in a company to understand the variability in the salary distribution.

```
# Example dataset (salaries in thousands)
salaries <- c(45, 50, 55, 60, 70, 80, 85, 90, 100, 110)
# Calculate range
salary_range <- range(salaries)
# Calculate variance
salary_variance <- var(salaries)
# Calculate standard deviation
salary_sd <- sd(salaries)
# Calculate interquartile range (IQR)
```

```
salary_iqr <- IQR(salaries)
# Display results
cat("Range of salaries is:", salary_range, "\n")
cat("Variance of salaries is:", salary_variance, "\n")
cat("Standard deviation of salaries is:", salary_sd, "\n")
cat("Interquartile range of salaries is:", salary_iqr,
"\n")
```

Example 2: Analyze the **daily temperatures** over a week by calculating the **range, variance, standard deviation**, and **IQR** to assess the variability in temperature.

```
# Example dataset (daily temperatures in Celsius for a
week)
temperatures <- c(22, 24, 26, 28, 27, 23, 25)
# Calculate range
temp_range <- range(temperatures)
# Calculate variance
temp_variance <- var(temperatures)
# Calculate standard deviation
temp_sd <- sd(temperatures)
# Calculate interquartile range (IQR)
temp_iqr <- IQR(temperatures)
# Display results
cat("Range of temperatures is:", temp_range, "\n")
cat("Variance of temperatures is:", temp_variance, "\n")
cat("Standard deviation of temperatures is:", temp_sd,
"\n")
cat("Interquartile range of temperatures is:", temp_iqr,
"\n")
```

Example 3: Analyze the **stock price fluctuations** of a company over 10 days by calculating the **range, variance, standard deviation**, and **IQR**.

```
# Example dataset (stock prices in dollars for 10 days)
stock_prices <- c(120, 125, 130, 135, 140, 145, 138, 142,
148, 150)
```

```
# Calculate range
stock_price_range <- range(stock_prices)
# Calculate variance
stock_price_variance <- var(stock_prices)
# Calculate standard deviation
stock_price_sd <- sd(stock_prices)
# Calculate interquartile range (IQR)
stock_price_iqr <- IQR(stock_prices)
# Display results
cat("Range of stock prices is:", stock_price_range, "\n")
cat("Variance of stock prices is:", stock_price_variance,
"\n")
cat("Standard deviation of stock prices is:",
stock_price_sd, "\n")
cat("Interquartile range of stock prices is:",
stock_price_iqr, "\n")
```

Example 4: Analyze the distribution of **exam scores** for a group of students by calculating the **range**, **variance**, **standard deviation**, and **IQR** to understand the score variability.

```
# Example dataset (exam scores out of 100)
exam_scores <- c(85, 90, 75, 88, 92, 76, 84, 91,
78, 80)
# Calculate range
score_range <- range(exam_scores)
# Calculate variance
score_variance <- var(exam_scores)
# Calculate standard deviation
score_sd <- sd(exam_scores)
# Calculate interquartile range (IQR)
score_iqr <- IQR(exam_scores)
# Display results
```

```
    cat("Range of exam scores is:", score_range,
"\n")
    cat("Variance of exam scores is:",
score_variance, "\n")
    cat("Standard deviation of exam scores is:",
score_sd, "\n")
    cat("Interquartile range of exam scores is:",
score_iqr, "\n")
```

Example 5: Calculate the **range**, **variance**, **standard deviation**, and **IQR** for **monthly rainfall** in a region over a year to assess variation in rainfall.

```
# Example dataset (monthly rainfall in mm)
rainfall_data <- c(78, 95, 110, 120, 150, 130,
140, 100, 90, 85, 75, 80)
# Calculate range
rainfall_range <- range(rainfall_data)
# Calculate variance
rainfall_variance <- var(rainfall_data)
# Calculate standard deviation
rainfall_sd <- sd(rainfall_data)
# Calculate interquartile range (IQR)
rainfall_iqr <- IQR(rainfall_data)
# Display results
cat("Range of rainfall is:", rainfall_range,
"\n")
cat("Variance of rainfall is:",
rainfall_variance, "\n")
cat("Standard deviation of rainfall is:",
rainfall_sd, "\n")
cat("Interquartile range of rainfall is:",
rainfall_iqr, "\n")
```


Example 6: Assess the **variation in hours worked** by employees over a month by calculating the **range, variance, standard deviation, and IQR**.

```
# Example dataset (hours worked by employees in  
a week)
```

```
hours_worked <- c(40, 42, 38, 44, 45, 40, 41,  
39, 37, 46)
```

```
# Calculate range
```

```
work_hours_range <- range(hours_worked)
```

```
# Calculate variance
```

```
work_hours_variance <- var(hours_worked)
```

```
# Calculate standard deviation
```

```
work_hours_sd <- sd(hours_worked)
```

```
# Calculate interquartile range (IQR)
```

```
work_hours_iqr <- IQR(hours_worked)
```

```
# Display results
```

```
cat("Range of hours worked is:",  
work_hours_range, "\n")
```

```
cat("Variance of hours worked is:",  
work_hours_variance, "\n")
```

```
cat("Standard deviation of hours worked is:",  
work_hours_sd, "\n")
```

```
cat("Interquartile range of hours worked is:",  
work_hours_iqr, "\n")
```

11.4. SHAPE OF DISTRIBUTIONS

Examples in R for analyzing the **shape of distributions** using **skewness, kurtosis, and visualizations** like histograms and density plots:

Example 1: Calculate the **skewness** and **kurtosis** of a dataset with a **right-skewed** distribution, and visualize it with a **histogram** and **density plot**.

```
# Load the e1071 package for skewness and kurtosis functions
```

```
install.packages("e1071")
library(e1071)
# Example dataset (right-skewed data, e.g., income distribution)
income_data <- c(10, 12, 15, 20, 25, 30, 35, 50, 60, 80, 100,
150, 200)
# Calculate skewness
income_skewness <- skewness(income_data)
# Calculate kurtosis
income_kurtosis <- kurtosis(income_data)
# Display results
cat("Skewness of income data:", income_skewness, "\n")
cat("Kurtosis of income data:", income_kurtosis, "\n")
# Visualization
par(mfrow = c(1, 2)) # Plot two plots side by side
# Histogram
hist(income_data, main = "Histogram of Income Data", xlab =
"Income", col = "skyblue", border = "black")
# Density plot
plot(density(income_data), main = "Density Plot of Income Data",
xlab = "Income", col = "red", lwd = 2)
```

Example 2: Calculate the **skewness** and **kurtosis** of a dataset with a **left-skewed** distribution, and visualize it with a **histogram** and **density plot**.

```
# Example dataset (left-skewed data, e.g., test scores)
test_scores <- c(95, 85, 80, 70, 60, 55, 45, 35, 25, 10)
# Calculate skewness
test_scores_skewness <- skewness(test_scores)
# Calculate kurtosis
test_scores_kurtosis <- kurtosis(test_scores)
# Display results
cat("Skewness of test scores:", test_scores_skewness, "\n")
cat("Kurtosis of test scores:", test_scores_kurtosis, "\n")
# Visualization
```

```
par(mfrow = c(1, 2)) # Plot two plots side by side
# Histogram
hist(test_scores, main = "Histogram of Test Scores", xlab =
"Test Scores", col = "lightgreen", border = "black")
# Density plot
plot(density(test_scores), main = "Density Plot of Test Scores",
xlab = "Test Scores", col = "blue", lwd = 2)
```

Example 3: Calculate the **skewness** and **kurtosis** of a **normal distribution**, and visualize it with a **histogram** and **density plot**.

```
# Example dataset (normal distribution, e.g., exam scores)
set.seed(123) # For reproducibility
normal_data <- rnorm(1000, mean = 50, sd = 10) # Generate
normal distribution data
# Calculate skewness
normal_skewness <- skewness(normal_data)
# Calculate kurtosis
normal_kurtosis <- kurtosis(normal_data)
# Display results
cat("Skewness of normal data:", normal_skewness, "\n")
cat("Kurtosis of normal data:", normal_kurtosis, "\n")
# Visualization
par(mfrow = c(1, 2)) # Plot two plots side by side
# Histogram
hist(normal_data, main = "Histogram of Normal Data", xlab =
"Values", col = "lightblue", border = "black", breaks = 30)
# Density plot
plot(density(normal_data), main = "Density Plot of Normal Data",
xlab = "Values", col = "green", lwd = 2)
```

Example 4: Analyze a **bimodal distribution** and calculate the **skewness**, **kurtosis**, and visualize the data with a **histogram** and **density plot**.

```
# Example dataset (bimodal distribution)
set.seed(123)
```

```
bimodal_data <- c(rnorm(500, mean = 20, sd = 3), rnorm(500, mean
= 40, sd = 5))
# Calculate skewness
bimodal_skewness <- skewness(bimodal_data)
# Calculate kurtosis
bimodal_kurtosis <- kurtosis(bimodal_data)
# Display results
cat("Skewness of bimodal data:", bimodal_skewness, "\n")
cat("Kurtosis of bimodal data:", bimodal_kurtosis, "\n")
# Visualization
par(mfrow = c(1, 2)) # Plot two plots side by side

# Histogram
hist(bimodal_data, main = "Histogram of Bimodal Data", xlab =
"Values", col = "orange", border = "black", breaks = 30)
# Density plot
plot(density(bimodal_data), main = "Density Plot of Bimodal
Data", xlab = "Values", col = "purple", lwd = 2)
```

11.5. FIVE-NUMBER SUMMARY AND BOXPLOTS

- **Five-number summary:** `summary()` (min, Q1, median, Q3, max)
- Creating **boxplots**: `boxplot()` to detect outliers
- Analyzing heights or test scores using boxplots

Example 1: A company collects salary data for employees and wants to analyze the five-number summary and detect outliers using a boxplot.

```
# Example dataset: Monthly salaries of employees (in
$1000s)
salaries <- c(30, 32, 35, 28, 40, 45, 50, 55, 60, 80, 100)
# Five-number summary
salary_summary <- summary(salaries)
print(salary_summary)
```

```
# Boxplot visualization
boxplot(salaries, main = "Salary Distribution", ylab =
"Salary ($1000s)", col = "lightblue", border = "blue", notch =
TRUE)

# Add outlier labels
outliers <- boxplot.stats(salaries)$out
if(length(outliers) > 0) {
  text(1, outliers, labels = outliers, pos = 3, col =
"red")
}
```

Example 2: A school wants to analyze student exam **scores** to understand the spread of marks and detect low-performing students.

```
# Example dataset: Exam scores out of 100
exam_scores <- c(55, 60, 65, 70, 75, 80, 85, 90, 95, 98,
100, 30, 35)

# Five-number summary
exam_summary <- summary(exam_scores)
print(exam_summary)

# Boxplot visualization
boxplot(exam_scores, main = "Exam Score Distribution",
ylab = "Scores", col = "lightgreen", border = "darkgreen",
horizontal = TRUE)

# Detect and label outliers
outliers <- boxplot.stats(exam_scores)$out
if(length(outliers) > 0) {
  text(outliers, 1, labels = outliers, pos = 3, col =
"red")}
```

Example 3: A meteorologist collects **daily temperature data** and wants to summarize and visualize the distribution using a **boxplot**.

```
# Example dataset: Daily temperatures recorded in a month
(in Celsius)
temperatures <- c(12, 15, 16, 14, 18, 22, 24, 30, 31, 28,
26, 22, 20, 19, 17, 15, 13, 12, 10, 5, 6)
```

```
# Five-number summary
temp_summary <- summary(temperatures)
print(temp_summary)

# Boxplot visualization
boxplot(temperatures, main = "Daily Temperature
Distribution", ylab = "Temperature (°C)", col = "orange", border
= "red")

# Detect and label outliers
outliers <- boxplot.stats(temperatures)$out
if(length(outliers) > 0) {
  text(1, outliers, labels = outliers, pos = 3, col =
"red")
}
```

Example 4: A real estate analyst wants to examine the **distribution of house prices** and detect expensive outliers.

```
# Example dataset: House prices in $1000s
house_prices <- c(200, 220, 250, 275, 300, 350, 400, 450,
600, 750, 1200)

# Five-number summary
house_summary <- summary(house_prices)
print(house_summary)

# Boxplot visualization
boxplot(house_prices, main = "House Price Distribution",
ylab = "House Price ($1000s)", col = "pink", border = "purple",
notch = TRUE)

# Detect and label outliers
outliers <- boxplot.stats(house_prices)$out
if(length(outliers) > 0) {
  text(1, outliers, labels = outliers, pos = 3, col =
"red")
}
```

11.6. SUMMARIZING MULTIPLE VARIABLES

- **Covariance:** `cov()`
- **Correlation:** `cor()` — Pearson, Spearman methods
- Finding the relationship between study time and test scores

Example 1: A fitness researcher wants to determine how height (cm) and weight (kg) are related using **covariance** and **correlation**.

```
# Height (cm) and Weight (kg) data
height <- c(160, 165, 170, 175, 180, 185, 190)
weight <- c(55, 60, 65, 72, 78, 85, 90)

# Compute covariance
cov_height_weight <- cov(height, weight)
print(paste("Covariance:", cov_height_weight))

# Compute Pearson correlation
cor_height_weight <- cor(height, weight, method =
"pearson")
print(paste("Pearson Correlation:", cor_height_weight))

# Compute Spearman correlation
cor_spearman <- cor(height, weight, method = "spearman")
print(paste("Spearman Correlation:", cor_spearman))

# Note: A high Pearson correlation (near 1) suggests a
strong linear relationship. The Spearman correlation is useful
if the relationship is monotonic but not necessarily linear.
```

Example 2: A teacher wants to examine the relationship between hours studied and test scores using correlation analysis.

```
# Hours studied and test scores
study_time <- c(2, 4, 6, 8, 10, 12, 14, 16)
test_scores <- c(50, 55, 60, 70, 75, 80, 85, 90)

# Compute covariance
cov_study_score <- cov(study_time, test_scores)
print(paste("Covariance:", cov_study_score))
```

```
# Compute Pearson correlation
cor_study_score <- cor(study_time, test_scores, method =
"pearson")
print(paste("Pearson Correlation:", cor_study_score))

# Scatter plot
plot(study_time, test_scores, main = "Study Time vs Test
Scores", xlab = "Study Time (hours)", ylab = "Test Scores", col
= "blue", pch = 16)
abline(lm(test_scores ~ study_time), col = "red")
```

Example 3: An investor wants to analyze the relationship between **two company stock prices** using both **Pearson and Spearman correlations**. (Note: Pearson correlation measures the linear relationship. Spearman correlation measures the rank-based relationship (good for non-linear trends).)

```
# Stock prices of two companies over 8 days
company_A <- c(100, 102, 105, 108, 110, 112, 115, 118)
company_B <- c(50, 53, 54, 58, 60, 65, 70, 75)

# Compute Pearson correlation
cor_pearson <- cor(company_A, company_B, method =
"pearson")
print(paste("Pearson Correlation:", cor_pearson))

# Compute Spearman correlation
cor_spearman <- cor(company_A, company_B, method =
"spearman")
print(paste("Spearman Correlation:", cor_spearman))

# Scatter plot
plot(company_A, company_B, main = "Stock Prices
Relationship", xlab = "Company A Price", ylab = "Company B
Price", col = "blue", pch = 16)
abline(lm(company_B ~ company_A), col = "red")
```

Both companies' stock prices move in the same direction, confirming a positive correlation.

Example 4: A health researcher collects data on **exercise hours, calorie intake, and BMI** to see their relationships. Provide an interpretation.


```
# Simulated dataset
exercise_hours <- c(2, 3, 4, 5, 6, 7, 8, 9)
calories_intake <- c(2500, 2400, 2300, 2200, 2100, 2000,
1900, 1800)
BMI <- c(30, 28, 27, 26, 25, 24, 22, 21)
# Compute covariance matrix
cov_matrix <- cov(data.frame(exercise_hours,
calories_intake, BMI))
print("Covariance Matrix:")
print(cov_matrix)
# Compute correlation matrix
cor_matrix <- cor(data.frame(exercise_hours,
calories_intake, BMI))
print("Correlation Matrix:")
print(cor_matrix)
```

Example 5: An economist wants to study the relationship between a **country's GDP (trillion \$)** and **unemployment**.

```
# GDP (in trillion $) and Unemployment Rate (%)
GDP <- c(2.5, 3.0, 3.2, 3.8, 4.1, 4.5, 5.0, 5.5)
unemployment_rate <- c(8.0, 7.5, 7.2, 6.8, 6.5, 6.0, 5.5,
5.2)
# Compute covariance
cov_gdp_unemp <- cov(GDP, unemployment_rate)
print(paste("Covariance:", cov_gdp_unemp))
# Compute Pearson correlation
cor_gdp_unemp <- cor(GDP, unemployment_rate, method =
"pearson")
print(paste("Pearson Correlation:", cor_gdp_unemp))
# Scatter plot
plot(GDP, unemployment_rate, main = "GDP vs Unemployment
Rate", xlab = "GDP (Trillion $)", ylab = "Unemployment Rate
(%)", col = "blue", pch = 16)
abline(lm(unemployment_rate ~ GDP), col = "red")
```

Interpretation: Negative correlation **indicates** higher GDP leads to lower unemployment.

REFERENCES

- [1] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W., *Robust Estimates of Location: Survey and Advances*, Princeton: Princeton University Press, 1972.
- [2] Belsley, D. A., Kuh, E., and Welsch, R. E., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: John Wiley and Sons, 1980.
- [3] Bickel, P. J., and Doksum, K. A., *Mathematical Statistics: Basic Ideas and Selected Topics*, (Vol. 1, 2nd ed.) Upper Saddle River, NJ: Prentice-Hall, (2000).
- [4] DeGroot M. H. , Schervish Mark. J. , *Probability and Statistics*, Fourth Edition, Pearson Education, Inc., 2012;
- [5] Murray R. Spiegel, Schiller J. J., Alu Srinivasa R., *Probability and Statistics*, Fourth Edition, Mc Graw Hill, 2013.
- [6] Meyer, P. L. *Introductory Probability and Statistical Applications* (2nd ed.), Reading, MA: Addison-Wesley, 1970.
- [7] Miller, I., and Miller, M., John E. Freund's *Mathematical Statistics* (6th ed.), Upper Saddle River, NJ: Prentice-Hall, 1999.
- [8] Mood, A. M., Graybill, F. A., and Boes, D. C., *Introduction to the Theory of Statistics* (3rd ed.). New York: McGraw-Hill, 1974.
- [9] Moore, D. S., and McCabe, G. P. (1999). *Introduction to the Practice of Statistics* (3rd ed.), New York: W. H. Freeman, 1999.
- [10] The R Project: A Brief History and Thoughts About the Future, Ross Ihaka, The University of Auckland, online <https://www.stat.auckland.ac.nz/%7Eihaka/downloads/Otago.pdf>
- [11] cengel.github.io
- [12] pubs.wsb.wisc.edu
- [13] Dalgaard, P., 2008. *Introductory statistics with R*. Springer.
- [14] De Sá, J.P.M., 2007. *Applied statistics using SPSS, Statistica, MatLab and R*. Springer Science & Business Media.
- [15] Dalpiaz, D., 2021. *Applied statistics with R*.
- [16] Rasch, D., Verdooren, R. and Pilz, J., 2019. *Applied statistics: theory and problem solutions with R*. John Wiley & Sons.

REFERENCES

- [17] Gunawan, A., Cheong, M.L.F. and Poh, J., 2018, December. An Essential Applied Statistical Analysis Course using RStudio with Project-Based Learning for Data Science. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (pp. 581-588). IEEE.

CIP- Cataloging of the publication

National and University Library "St. Kliment Ohridski ", Skopje
519.22:004.438R(075.8)(076)

MATHEMATICAL Statistics with Examples in R [electronic source] / [Aneta Velkoska ... et al.]. - Text in PDF format, contains XI, 320 pages. - Ohrid : University of information science and technology "St. Paul the Apostle", 2025

Access method (URL): <https://uist.edu.mk/wp-content/uploads/2025/03/Mathematical-Statistics-with-examples-in-R.pdf>

Title taken from the screen.- Description of the source on 25.03.2025.- Other authors: Ninoslav Marina, Atanas Hristov, Ljubinka Sandjakoska - Bibliography: p. 320

ISBN 978-608-66225-3-4

1. Velkoska, Aneta [author] 2. Ninoslav, Marina [author] 3. Hristov, Atanas [author] 4. Sandjakoska, Ljubinka [author]
a) Mathematical Statistics -- R (programming language) -- University textbooks -- Tutorials --

COBISS.MK-ID 65559045