NICOLAE TESTEMITANU STATE UNIVERSITY OF MEDICINE AND PHARMACY OF THE REPUBLIC OF MOLDOVA

Nicolae Testemitanu Department of Social Medicine and Management

Elena Raevschi, Olga Penina, Galina Obreja

Basic Biostatistics and Research Methodology

Chisinau
Editorial and Printing Centre *Medicina*
2025

CZU[311.2+001.891]:614.2(075.8)
R 19

Approved by the Quality Management Board of *Nicolae Testemitanu* SUMPh, minutes no. 2, 23.10.2024.

Authors:

Elena RAEVSCHI – Habilitated Doctor in Medical Sciences, Professor, *Nicolae Testemitanu* Department of Social Medicine and Management.

Olga PENINA – Habilitated Doctor in Medical Sciences, Associate Professor, *Nicolae Testemitanu* Department of Social Medicine and Management.

Galina OBREJA – Doctor in Medical Sciences, Associate Professor, *Nicolae Testemitanu* Department of Social Medicine and Management.

Reviewers:

Larisa SPINEI – Habilitated Doctor in Medical Sciences, Professor, *Nicolae Testemitanu* Department of Social Medicine and Management.

Alexandru CORLATEANU – Habilitated Doctor in Medical Sciences, Professor, Department of Internal Medicine.

Redactor: În redacția autorilor

This publication is presented in its original authorial form.

The compendium "Basic Biostatistics and Research Methodology" has been developed based on international experience and updated didactic standards. It aligns with the syllabus requirements for the Biostatistics and Scientific Research Methodology course for medical students in integrated higher education and higher licentiate education at the *Nicolae Testemitanu* SUMPh.

DESCRIEREA CIP A CAMEREI NAȚIONALE A CĂRȚII DIN REPUBLICA MOLDOVA
Raevschi, Elena.
Basic Biostatistics and Research Methodology / Elena Raevschi, Olga Penina, Galina Obreja ; *Nicolae Testemitanu* State University of Medicine and Pharmacy of the Republic of Moldova, *Nicolae Testemitanu* Department of Social Medicine and Management. – Chişinău : CEP *Medicina*, 2025. – 176 p. : fig., tab.
Bibliogr.: p. 172-173. – În red. aut. – 200 ex.
ISBN 978-9975-82-403-3.
[311.2+001.891]:614.2(075.8)
R 19

ISBN 978-9975-82-403-3

Contents

Preface **8**
CHAPTER 1. INTRODUCTION TO BASIC BIOSTATISTICS AND MEDICAL RESEARCH METHODOLOGY **9**
CHAPTER 2. DESCRIPTIVE STATISTICS: DATA PRESENTATION **13**
Key Concepts 13
2.1 Basic Concepts 15
2.1.1 Variable 15
2.1.2 Population 16
2.1.3 Sample 17
2.1.4 Parameters and Statistics 17
2.2 Scales of Measurement 18
2.2.1 Nominal Scale 18
2.2.2 Ordinal Scale 19
2.2.3 Numerical Scale 20
2.3 Tables 21
2.3.1 Frequency Distributions 21
2.3.2 Relative Frequency 23
2.4 Graphs 24
2.4.1 Line charts 24
2.4.2 Bar Charts 25
2.4.3 Pie Charts 26
2.4.4 Histograms and Frequency Polygons 27
2.4.5 Box Plots 28
2.4.6 Error Bar Plots 29
2.4.7 Two-Way Scatter Plots 30
Review Exercises 31
Review Questions 32
CHAPTER 3. DESCRIPTIVE STATISTICS: SUMMARIZING NUMERICAL DATA **34**
Key Concepts 34
3.1 Measures of Central Tendency 36
3.1.1 Mean 36
3.1.2 Median 36

3.1.3 Mode....37
3.1.4 Empirical Relationship Between the Measures of Central Tendency....37
3.2 Measures of Variability....39
3.2.1 Range....40
3.2.2 Interquartile Range....40
3.2.3 Variance and Standard Deviation....41
3.2.4 Coefficient of Variation....42
3.3 Normal Distribution and its Properties....44
3.4 Skewness and Kurtosis....45
3.5. Example Calculation....47
Review Exercises:....49
Review Questions....50
CHAPTER 4. DESCRIPTIVE STATISTICS: SUMMARIZING NOMINAL AND ORDINAL DATA....51
Key Concepts....51
4.1 Methods to Describe Categorical Data....52
4.1.1 Proportions and Percentages....52
4.1.2 Ratios....53
4.1.3 Rates....54
4.2 Health Status Indicators in Descriptive Statistics....55
4.2.1 Mortality Rates....55
4.2.2 Morbidity Rates....55
4.3 Standardized Rates: Direct Method of Standardization....56
Review Exercises....61
Review Questions....63
CHAPTER 5. CORRELATION AND REGRESSION....64
Key Concepts....64
5.1 Correlation....65
5.1.1 Types of Correlation Coefficient....65
5.1.2 Coefficient of Determination....69
5.2 Regression: General Approaches....70
5.2.1 Simple Linear Regression....70
5.2.2 Multiple Linear Regression....72
5.2.3 Logistic Regression....72

Review Exercises....74
Review Questions....75
CHAPTER 6. INFERENTIAL STATISTICS: PROBABILITY THEORY AND HYPOTHESES TESTING....76
Key Concepts....76
6.1 Probability Theory....78
6.1.1 General Concepts....78
6.1.2 Law of Large Numbers....78
6.1.3 Central Limit Theorem....79
6.1.4 Using the Standard Error....79
6.2 Sampling....80
6.2.1 Sampling Definition....80
6.2.2 Sampling Methods....80
6.3 Estimation and Hypothesis Testing....81
6.4 Confidence Intervals....82
6.5 Hypothesis Testing: Basic Theoretical Concepts....85
6.5.1 Hypothesis Definition....85
6.5.2 Hypotheses Types....85
6.5.3 Type I and Type II Errors....86
6.5.4 Power of the Study....86
6.5.5 Confidence Level....87
6.5.6 Significance Level....87
6.5.7 p-value....88
6.6 Key Steps in Hypothesis Testing....89
Review Exercises....90
Review Questions....91
CHAPTER 7. HYPOTHESIS TESTING: PARAMETRIC AND NON-PARAMETRIC METHODS....93
Key Concepts....93
7.1 Parametric and Non-Parametric Tests....93
7.2 General Approach to Hypothesis Testing....95
7.2.1 Steps of Hypothesis Testing....95
7.2.2 Hypothesis Testing: Two-tailed, Left-Tailed and Right-Tailed Tests....97
7.3 Parametric Tests....99

7.3.1. One-Sample t-Test 99
7.3.4 Two-Sample Independent t-Test 102
7.3.5 Correlation Coefficient t-test 105
7.4 Non-parametric Tests 108
7.4.1 Chi-Square Test 108
Review Exercises 111
Review Questions 112
CHAPTER 8. INTRODUCTION TO THE RESEARCH METHODOLOGY 113
Key Concepts 113
8.1 Research Definition, Characteristics and Types 114
8.1.1 Definition and Characteristics of Research 114
8.1.2 Random Errors and Systematic Biases in Research 115
8.1.3 Types of Research 117
8.2 The Steps of Research Process 119
8.3 Formulating the Research Problem 120
8.4 Literature Review 122
8.5 Formulation of the Aim and Objectives of the Study 125
8.6 Preparing the Research Design and Collecting the Data 125
8.6.1 Research Design Definition and Steps 125
8.6.3 Determining Sample Design 125
8.6.4 Tool for Data Collection 126
8.6.5 Study Design Classification 129
8.7 Study Design Evidence Strength 132
Review Exercises 134
Review Questions 135
CHAPTER 9. OBSERVATIONAL DESCRIPTIVE STUDIES 137
Key Concepts 137
9.1 Case-series / Case-report study 137
9.2 Cross-sectional study 138
Review Questions 140
CHAPTER 10. OBSERVATIONAL ANALYTICAL STUDIES 141
Key Concepts 141
10.1 Case-control study 142
10.2 Cohort study 145

Review Exercises 150
Review Questions 152
CHAPTER 11. EXPERIMENTAL STUDIES 153
Key Concepts 153
11.1 Classification of Clinical Trials 154
11.2 Controlled Trials with Concurrent (Parallel) Controls 155
11.3 Sequential Controlled Trials 157
11.4 Trials with External Controls 158
11.5 Statistical Analysis of Clinical Trials 160
Review Exercises 163
Review Questions 164
CHAPTER 12. REPORTING THE RESEARCH FINDINGS: GENERAL APPROACHES 165
12.1 Writing a Research Report 165
12.2 Public Presentation of Medical Research Results 166
12.3 Structure and Principles of Graduation Paper Development at Nicolae Testemitanu SUMPh 167
CHAPTER 13. RESEARCH ETHICS INTRODUCTION 168
13.1 Research Ethics Definition and Objectives 168
13.2 Research Ethics Principles 168
CHAPTER 14. PREVENTING PLAGIARISM: KEY PRINCIPLES .. 170
14.1 Plagiarism Meaning and Types 170
14.2 Preventing Plagiarism Techniques 170
BIBLIOGRAPHY 172
APPENDIX A: **Critical values for the "t" distribution 174**
APPENDIX B: **Critical values for the Chi-square distribution. 175**

Preface

Biostatistics and Research Methodology, as a discipline, is an interdisciplinary field that draws on a wide array of knowledge contributions and is crucial for conducting research that adheres to global standards. The compendium “Basic Biostatistics and Research Methodology” aligns with the discipline’s syllabus and offers detailed, well-organized information on every step required to conduct a rigorous scientific study. This guide serves as an introduction to biostatistics and research methodology for health science students, aiming to facilitate the completion of their undergraduate research theses.

This Compendium is published in four languages (Romanian, Russian, English, and French) and features a revised structure, expanded content, and improved discussions on various topics. It includes additional figures and tables to clarify key concepts, particularly in descriptive and inferential statistics. The book provides detailed explanations of hypothesis testing using both parametric and nonparametric tests, as well as comprehensive analyses of correlation and regression. Critical value tables for the t-distribution and chi-square distribution have been added in Appendix A and Appendix B.

Acknowledgements

We extend our heartfelt gratitude to the professors who reviewed the manuscript: Larisa Spinei, Habilitated Doctor in Medical Sciences, Professor at the Nicolae Testemitanu Department of Social Medicine and Health Management, and Alexandru Corlateanu, Habilitated Doctor in Medical Sciences, Professor at the Department of Internal Medicine, both from Nicolae Testemitanu University of Medicine and Pharmacy. Our appreciation also goes to all the professors who contributed to teaching the course and offered invaluable suggestions.

Authors

CHAPTER 1. INTRODUCTION TO BASIC BIOSTATISTICS AND MEDICAL RESEARCH METHODOLOGY

The compendium "Basic Biostatistics and Research Methodology" introduces medical students to the study of statistics applied to medicine and other disciplines in the health field. The main target is to create knowledge about the contemporaneous methods used in practical research. Acquisition of knowledge necessary for the use of modern methods of documentation, assimilation of some theoretical definitions applicable in research and some standards of rules are necessary to highlight research results used in an undergraduate thesis.

The course Biostatistics and Scientific Research Methodology covers the theoretical and practical aspects related to the realization of scientific research and statistical data analysis. The course has content similar to other European universities with up-to-date information, and it equips students with the necessary baggage of knowledge to carry out scientific research in the field of biomedical science. The course presents a predominantly applicative approach to the statistical methods needed to solve practical problems in the biomedical field.

Main Objective of the Course:

To help the students understand the basic concept of Biostatistics and Research Methodology in such a way that they can use it to plan and analyse data in biomedical research.

At the knowing and understanding level:

- To know theoretical concepts of Methodology of Medical Scientific Research.
- Development of clear and continuous thinking, capable to manage and process the data.

- To know the principles, technology, methods and techniques used in Medical Research.
- To understand the correlation among modern methods used in Biostatistics and Medical Research Methodology.
- To identify possibilities of analysis and interpretation and also limits of modern methods used in Scientific Research.

At the application level:

- To analyse definitions, theoretical and practical methods of Methodology of Scientific Research.
- To use statistical methods and techniques in the scientific process.
- To demonstrate the capability of analysis, interpretation and presentation of scientific research results.
- To use base knowledge of biostatistics necessary for understanding its optimal application in getting the right scientific research result.
- To possess special language and terminology specific to scientific style.
- To evaluate the information contained in an article or report of a speciality and to appreciate its relevance.
- To be able to search scientific information using classical methods or computer methods for searching and selection of data.
- To use modern methods of writing and presentation of a scientific proposal and report of final results.

At the integration level:

- To appreciate the theoretic-applicable value of Medical Research Methodology in different disciplines in the health field.
- To assess the place and role of biostatistics and research methodology in the professional medical career;

- To integrate the knowledge in biostatistics and research methodology with clinical disciplines;
- To be able to apply the accumulated knowledge to practical and research activities;
- To be competent in using information critically from scientific publications in own research using new information and communication technologies.

Study outcomes:

- To explain the basic concepts with regard to the organization of scientific research and publication of the results;
- To develop a research project in the biomedical field;
- To present the description of experimental data depending on its nature and to explain correctly the results of the statistical inference;
- To determine the statistical methods for data analysis taking into account the study design characteristics, the scale of measurement, and the number of variables involved;
- To characterize the basic features of the epidemiological study designs (observational and experimental), their advantages and limitations;
- To perform an epidemiological study (observational or experimental) and interpret its results correctly;
- To develop a scientific paper, including the license thesis, and to capitalize on its results;
- To assess the role and importance of biostatistics and the research methodology in the modern context of "evidence-based medicine";
- To have openness to lifelong learning.

Health research is an interdisciplinary field, relying on a broad range of knowledge. Biostatistics and Scientific Research Methodology in medicine are disciplines that integrate and analyse knowledge from fundamental and practical studies. This discipline is crucial for evaluating research activities in line with modern standards. As an integrative field, it connects with other disciplines that use statistics. A solid understanding of high school mathematics and basic biomedicine is essential for grasping this discipline.

CHAPTER 2. DESCRIPTIVE STATISTICS: DATA PRESENTATION

Key Concepts

- *Statistic:* A characteristic or value derived from sample data.
- *Parameter:* A characteristic or value derived from population data.
- *Variable:* a measured characteristic of an observational unit.
- *Quantitative (Numerical) Variable*: A variable for which the assigned values are ordered and meaningful.
- *Qualitative (Categorical) Variable*: A variable for which the assigned values have meaning as nominal references (labels) but not as numerical values.
- *Alternative (Dichotomous) Variable*: A qualitative variable that has only two categories.
- *Non-alternative Variable*: A qualitative variable that has more than two categories.
- *Discrete Variable*: A numerical variable that takes on only whole numbers.
- *Continuous Variable*: A numerical variable that takes on any values on a continuum.
- *Nominal Data*: Data classified into different qualitative categories that can be listed in any order.
- *Ordinal Data*: Data classified into different categories where the order among the categories is meaningful, but there is no information about the quantitative distance between categories.
- *Interval Data:* Data with equal intervals between items but without an absolute zero.
- *Ratio Data:* Data with equal intervals and an absolute zero.
- *Frequency distribution:* A collection of values from a sample on a single variable.

- *Histogram and Frequency Polygon:* Graphs used to display the frequency distribution of numerical data. They inform you about the shape of a frequency distribution.
- *Scatterplot:* A graph used to illustrate the relationship between two numerical variables.
- *Bar Chart and Pie Chart:* Graphs used to display qualitative data.

Descriptive statistics help you understand and summarize large sets of data by using a few key numbers. Think of these numbers as a way to make sense of a lot of information quickly. For example, instead of looking at the test scores of 100 students one by one, you can look at the average score to get a general idea of how the class performed.

These key numbers can help you see patterns and trends in the data, making it easier to understand and share important information from your research. However, it's important to remember that descriptive statistics do not explain why these patterns exist or what they mean. Instead, they give you clues that can lead to forming questions or ideas, which are known as hypotheses. A hypothesis is like a possible explanation that you can test with further research.

2.1 Basic Concepts

2.1.1 Variable

Definition of Variable: A variable is a characteristic of interest that has different values for different subjects or objects included in a study.

Examples: Age, date of birth, nationality, number of children, blood pressure.

Classification of Variables: To correctly present descriptive statistics, it is necessary to understand the data types commonly encountered in research studies (*Figure 2.1*).

Quantitative (Numerical) Variables. These variables can be quantified and are classified as follows:

1. *Discrete Variables:* These represent quantities that can only take on specific, distinct values, typically whole numbers, with no possible intermediate values.

Examples: number of patients, number of new cases of cardiac diseases, number of newborns in a specified year, *etc*.

2. *Continuous Variables*: These represent quantities that can take on any value within a range, not limited to specific discrete values.

Examples: age, blood glucose level, blood pressure, *etc*.

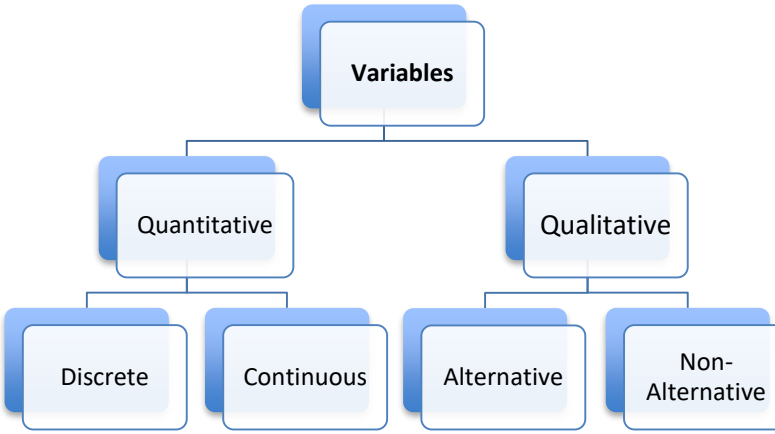

Figure 2.1 Types of Variables

Qualitative (Categorical) Variables. These variables cannot be quantified and are classified as follows:

1. *Alternative Variables (Dichotomous or Binary)*: Represent categories where outcomes can assume only one of two possible values.

Examples: "Yes" or "No"; "Male" or "Female", etc.

2. *Non-alternative Variables*: Represent categories where outcomes can assume multiple values.

Examples: blood type, severity disease level, *etc*.

In analyses, qualitative variables are frequently encoded using numerical values.

2.1.2 Population

A population refers to the entire group of individuals or items that share a common characteristic and are the focus of a study. A population includes every member that meets the criteria defined by the study. For example, a population could be all the students in a particular school, all the patients in a hospital, or all the trees in a forest.

- *Target Population:* This is the entire group to which the findings of the study are intended to be generalized. It is also known as the theoretical statistic population.
- *Study population:* This is the specific portion of the target population that is accessible and from which data is collected for the study. It is also known *as* the *accessible* population.

2.1.3 Sample

A sample is a subset of the study population. *The unit of observation*, or statistical unit, is the entity on which information is collected, such as an individual, household, community, school, etc. Clearly identifying the unit of observation is important for logical survey design, organized data collection, and objective analysis.

Target vs Study Populations: Cardiovascular Disease Research.

- **Target Population**: all adults aged 40 and above who have been diagnosed with cardiovascular disease worldwide.
- **Study Population:** Adults aged 40 and above with cardiovascular disease who are patients at major hospitals in the Republic of Moldova.

The study population serves as a representative subset of the broader target population.

2.1.4 Parameters and Statistics

Statistic is a characteristic or value derived from sample data.
Parameter is a characteristic or value derived from population data.

Table 2.1 Population and Sample Symbols

	Parameters (population data)	**Statistics (sample data)**
	Greek Symbols	*Roman Symbols*
Mean	μ	$\bar{x}$
Standard Deviation	σ	S
Variance	σ^2	S^2
Size (number of observations)	N	n

2.2 Scales of Measurement

Different scales of measurement depend on the nature of the variables. They determine how the data is displayed and summarized and the statistical methods for analysing the data. There are 3 scales of measurement in statistics:

- Nominal (Classificatory) scale;
- Ordinal (Ranking) scale;
- Numerical scale (Interval and Ratio scales).

2.2.1 Nominal Scale

Nominal scale data consists of categories that do not have any intrinsic order. A variable measured on a nominal scale can have two or more subcategories, depending upon the extent of variation in a qualitative variable.

Example 1. The variable *"gender"* typically has two categories: male and female. Nominal data that take on one of two distinct categories, such as male or female, are considered *alternative categorical variables* or *dichotomous variables*.

Example 2. However, not all nominal data are dichotomous. Many nominal variables have three or more categories. For example, "anaemia" can be classified into multiple subcategories: microcytic

anaemia, macrocytic or megaloblastic anaemia and normocytic anaemia. This variable is considered a non-alternative categorical variable. The order in which these categories are listed does not matter because there is no inherent ranking or relationship among categories.

Nominal scale data is used to label variables without providing any quantitative value. Each category is unique, and the sequence of categories is irrelevant.

2.2.2 Ordinal Scale

The ordinal scale, also known as the ranking scale, categorizes individuals, objects, responses, or properties into subgroups based on a common characteristic, and then ranks these subgroups in a specific order.

Subgroups are arranged either in ascending or descending order based on the extent to which a subcategory reflects the magnitude of variation in the variable. When the order among categories becomes meaningful, the data are referred to as being on an ordinal scale of measurement.

Example 1. "Income" of a patient can be measured using a qualitative variable with categories "above average", "average", and "below average". The "distance" between these subcategories is not equal, as there is no quantitative unit of measurement.

Example 2. Apgar score, which describes the maturity of newborn infants, ranges from 0 to 10. Lower scores indicate depression of cardiorespiratory and neurologic functioning, while higher scores indicate good functioning. The difference between scores of 8 and 9 does not have the same clinical implications as the difference between scores of 0 and 1. This illustrates that in ordinal scales, the intervals between points of measurement are not equal.

Ordinal scale data not only categorizes variables but also ranks them, though the intervals between ranks are not necessarily uniform or measurable.

2.2.3 Numerical Scale

The numerical scale is characterized by equal intervals between sequential points of measurement. There are two types of numerical scales:

⇒ *Interval Scale*

An interval scale encompasses all the characteristics of nominal and ordinal scales. Additionally, it allows the data to be arranged in a hierarchical order with equal distances between points. Interval scale data does not have an absolute zero (i.e., a point that signifies the complete absence of the variable), meaning it can take meaningful values both below and above zero.

Example.

Celsius scale: 0°C to 100°C

Fahrenheit scale: 32°F to 212°F

The Celsius and Fahrenheit temperature scales measure temperature with equal intervals between degrees. Both scales include values below zero (for Celsius) or below the freezing point of water (32°F for Fahrenheit), indicating temperatures below the freezing point of water.

⇒ *Ratio Scale*

A ratio scale has all the properties of nominal, ordinal and interval scales, with an additional characteristic, a fixed zero point. This means it has an absolute zero value, indicating the complete absence of the variable (there are no values below zero).

Examples of variables measured on a ratio scale include serum cholesterol level, income, age, body height and weight.

The table below highlights the essential differences between the four scales of measurement (*Table 2.2*).

Table 2.2 Difference between the four scales of measurement

Scale	Indicates difference	Indicates direction of difference	Indicates amount of difference	Absolute zero
Nominal	+			
Ordinal	+	+		
Interval	+	+	+	
Ratio	+	+	+	+

2.3 Tables

A table is the simplest way of summarizing a set of data and can be used for all types of variables.

2.3.1 Frequency Distributions

A commonly employed method for summarizing data is *the frequency distribution table*. This table categorizes data and displays the corresponding numerical counts for various types of data. For nominal and ordinal data, the frequency distribution organizes categories along with their counts. For instance, Table 2.3 presents the distribution of newborns by sex in the Republic of Moldova for the year 2022, utilizing nominal data.

Table 2.3 Frequency Distribution for Nominal Data: Distribution of Births by Sex in the Republic of Moldova in 2022

Newborn's Sex	Number of Births
Male	13950
Female	13068

To represent numerical data, both ungrouped and grouped frequency distributions are employed. For example, Table 2.4 illustrates an ungrouped frequency distribution by showing the annual number of births in the Republic of Moldova from 2000 to 2020, using interval data.

Table 2.4 Frequency Distribution for Interval Data: Annual Number of Births in the Republic of Moldova, in 2000-2022

Year	Number of Births
2000	36939
2010	40474
2020	30834
2022	27018

When numerical data is highly detailed, *grouped frequency distributions* are particularly useful. This approach involves dividing the range of values into distinct, non-overlapping intervals. After establishing these intervals, the number of observations within each interval is counted. *Table 2.5* exemplifies this method by detailing the grouped frequency distribution of births by mother's age in the Republic of Moldova for the year 2022, utilizing ratio data.

Table 2.5 Grouped Frequency Distribution for Ratio Data: Number of Births by Mother's Age in the Republic of Moldova, in 2022

Mother's Age Group	Number of Births
Less than 24 years old	7117
25-34 years old	15032
35-44 years old	4843
45 years old and over	26

Tables are most effective when they are clear and well-organized. Thus, tables and their columns should always be clearly labelled, and units of measurement should be specified as needed.

2.3.2 Relative Frequency

It is sometimes useful to know the proportion of values that fall into a given interval in a frequency distribution rather than the absolute number. The *relative frequency* for an interval is the proportion of the total number of observations that appear in that interval. The relative frequency is computed by dividing the number of values within an interval by the total number of values in the table, expressed as a percentage. Relative frequencies are useful for comparing sets of data containing unequal numbers of observations.

The *cumulative relative frequency* for an interval is the percentage of the total number of observations that have a value less than or equal to the upper limit of the interval. The cumulative relative frequency is calculated by summing the relative frequencies for the specified interval and all previous ones.

Table 2.6 displays the absolute, relative and cumulative relative frequencies of the shock-index score in 931 patients.

Table 2.6 Absolute, relative and cumulative relative frequencies of a shock-index score for 931 patients

Shock Index Score	Frequency (Number of Patients)	Relative Frequency (%)	Cumulative Relative Frequency (%)
0.30-0.39	38	4.1	4.1
0.40-0.49	104	11.2	15.3
0.50-0.59	198	21.3	36.6
0.60-0.69	199	21.4	58.0
0.70-0.79	155	16.6	74.6
0.80-0.89	102	11.0	85.6
0.90-0.99	60	6.4	92.0

Shock Index Score	Frequency (Number of Patients)	Relative Frequency (%)	Cumulative Relative Frequency (%)
1.00-1.09	37	4.0	96.0
1.10-1.19	19	2.0	98.0
1.20-1.29	19	2.0	100.0
Total	931	100.0	

2.4 Graphs

Graphs (charts) are a second way to summarize and display data. They are often easier to read than tables but may provide less detailed information. The most informative graphs are relatively simple and self-explanatory. As with tables, graphs should be clearly labelled and units of measurement should be indicated.

2.4.1 Line charts

Line graphs are commonly used to display trends over time for numerical data. Each value on the *x-axis* corresponds to a single value on the *y-axis*, and adjacent points are connected by straight lines (*Figure 2.2*).

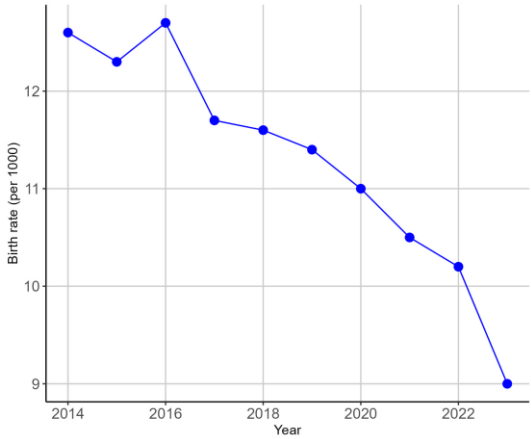

Figure 2.2 Birth Rate in the Municipality of Chisinau in 2014-2023, per 1000 population

2.4.2 Bar Charts

Bar charts, whether vertical or horizontal, are a widely used type of graph for illustrating frequency distributions of nominal or ordinal data. In these charts, bars must be of uniform width and spaced apart to prevent the misleading impression of continuity. Bar charts are effective for comparing multiple values, with categories displayed along either the vertical or horizontal axis, as shown in *Figure 2.3.*

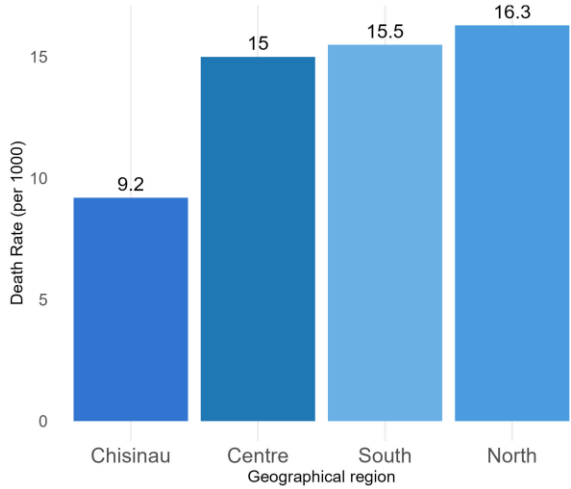

Figure 2.3 Crude Mortality Rate by Geographical Regions in the Republic of Moldova in 2023, per 1000 population

2.4.3 Pie Charts

Pie charts illustrate the proportion of each value relative to the whole and are used to show frequency distributions of nominal data.

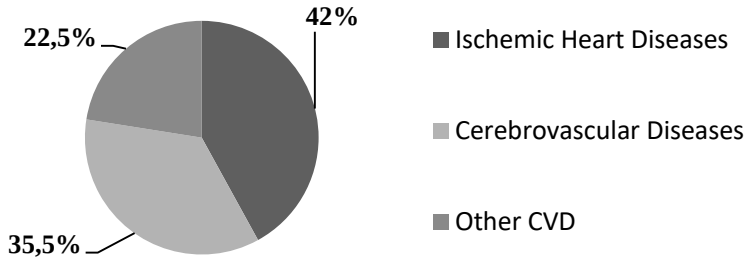

Figure 2.4 Global distribution of cardiovascular deaths due to heart attacks, strokes and other types of cardiovascular diseases (CVD), 2020

2.4.4 Histograms and Frequency Polygons

Histograms are one of the most informative methods for presenting both absolute and relative frequencies of numerical data. Although they resemble bar charts, there are key differences:

- Bar charts depict frequency distributions of categorical data (nominal or ordinal), whereas histograms illustrate frequency distributions for numerical data (discrete or continuous).
- Histograms reveal the shape of the frequency distribution, whereas bar charts only tally values without showing the distribution's shape.

In a histogram, the horizontal axis represents the limits of various intervals, while the vertical axis shows the absolute or relative frequency.

A frequency polygon, similar to a histogram but using points connected by straight lines rather than bars, provides a graphical representation of the dataset's distribution. To create a frequency polygon, plot the midpoints of each class interval on the horizontal axis and their corresponding frequencies on the vertical axis, then connect these points with straight lines. Frequency polygons and histograms can be superimposed for comparative analysis, as demonstrated in *Figure 2.5* with the dataset from *Table 2.6*.

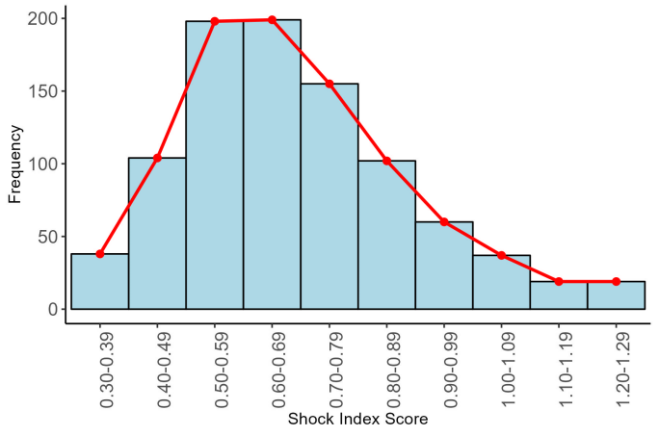

Figure 2.5 Histogram and Frequency Polygon of Shock Index Scores

2.4.5 Box Plots

Box plots, also known as box-and-whisker plots, are a valuable tool for summarizing the distribution of one or more sets of numerical data. Unlike other types of plots, box plots provide a succinct overview of the data distribution without showing every individual data point.

Figure 2.6 illustrates a box plot and its components. The central box, which can be oriented either vertically or horizontally, spans from the 25th percentile to the 75th percentile of the dataset. A line within the box represents the median (50th percentile), reflecting the central tendency of the data. When the median is positioned near the centre between the quartiles, it suggests a symmetrical distribution of the data.

The lines extending from the box, known as whiskers, represent the range of typical values in the dataset. These whiskers extend to 1.5 times the interquartile range (the difference between the 75th and 25th percentiles) above and below the box. Data points outside these whiskers are marked as circles, indicating outliers or values that fall outside the typical range.

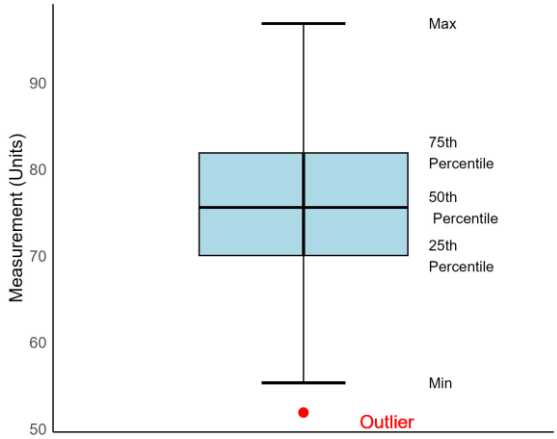

Figure 2.6 Box Plot and its Components

2.4.6 Error Bar Plots

Error bar plots are common in medical research for comparing groups. They depict the mean with a circle and illustrate variability using either standard deviation or standard error bars (*Figure 2.7*). Error bar plots provide insights into distribution similarities between groups.

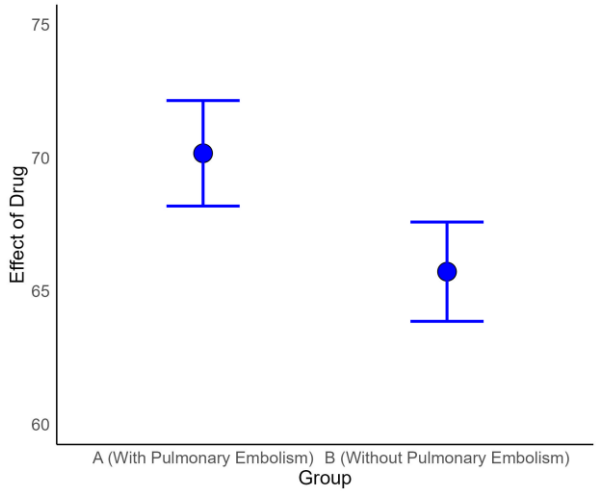

Figure 2.7 Error bar charts of the effect of the drug for patients with (A) and without (B) a pulmonary embolism

2.4.7 Two-Way Scatter Plots

A two-way scatter plot is used to depict the relationship between two different numerical variables. Each point on the graph represents a pair of values simultaneously (*Figure 2.8*).

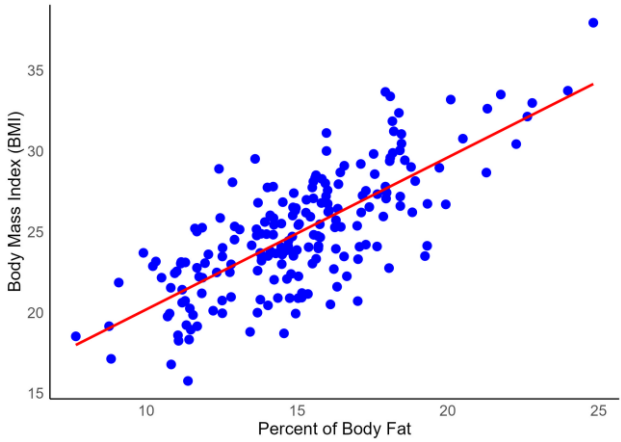

Figure 2.8 Scatter Plot Illustrating the Relationship between BMI and Percentage of Body Fat

Review Exercises

1. State the type of the variable and appropriate measurement scale for the following sets of data:
   a) Salaries of 125 physicians in a clinic
   b) The test scores of all medical students taking winter examinations in a given year
   c) Serum cholesterol level of healthy individuals
   d) Presence of diarrhoea in a group of infants
   e) The age onset of breast cancer in females
   f) Body temperature of the patients
   g) Discharged patient outcome
   h) Number of births in a given year

2. Use the following data to display it by all appropriate graphs. State your decision.

a) Clinical and Pathological Diagnoses Divergence in the Hospital, 2017-2021

Years	2017	2018	2019	2020	2021
Divergence, %	11	9.8	8.0	9.2	8.2

b) Acute viral hepatitis morbidity in the Republic of Moldova, 2021

Type	A	B	C	D	E
%	34.4	41.4	17.6	3.8	2.8

3. Propose a set of data that can be displayed by line chart. Explain your decision.
4. Propose a set of data that can be displayed by bar chart. Explain your decision.

Review Questions

1. What are descriptive statistics?
2. What are the classifications of variables? Provide examples.
3. How does an alternative variable differ from a non-alternative one? Give examples.
4. Define measurement scales and classify them. Provide an example for each type.
5. How do ordinal data differ from nominal data? Give examples.
6. What types of data presentation methods do you know? Explain the differences between them.
7. When constructing a table, when it might be beneficial to use relative frequencies rather than absolute frequencies?
8. Describe graph data presentation: content and types. Give an example for each type.

9. What is an appropriate graph for presenting nominal variables? Give examples.
10. What is an appropriate graph for presenting ordinal variables? Give examples.
11. What is an appropriate graph for presenting numerical variables? Give examples.

CHAPTER 3. DESCRIPTIVE STATISTICS: SUMMARIZING NUMERICAL DATA

Key Concepts

- *Descriptive statistics* merely describe, organize or summarize data.
- *Mode, Median* and *Mean* are used for numerical data (interval and ratio). The Median is also used for ordinal data.
- The mode and median are *insensitive* to outliers in a distribution. The mean is *sensitive* to outliers.
- If the original scores of a distribution are not available, the *weighted mean* can be estimated from a frequency table.
- In a *normal* (bell-shaped) distribution, all three measures of central tendency are identical.
- In a *positively* (right) skewed distribution, the mode is less than the median.
- In a *negatively* (left) skewed distribution, the median is less than the mode.
- The *mean* is used for numerical data and normal distributions.
- The *median* is used for ordinal data or numerical data if the distribution is skewed.
- The *mode* is used primarily for bimodal distribution.
- *An unimodal* distribution has one mode, *a bimodal* – two modes and *a uniform* distribution has no mode.
- The *range* is sensitive to extreme scores in a distribution.
- The *interquartile range* (IQR) is used to describe the central 50% of a distribution regardless of its shape.
- The IQR is the difference between the 75^{th} and 25^{th} percentiles.
- *Variance* (s^2) represents the amount of variability around the mean of a set of data.
- *Standard deviation* (s) is the average distance between the individual values in a distribution and the mean calculated for this distribution.

- The *coefficient of variation* (CV) is a measure of relative spread that permits the comparison of observations measured on different scales.
- The *standard deviation* is used when the mean is used (symmetric numerical data).
- The *IQR* is used when the median is used (ordinal data or skewed numerical data).
- *An empirical rule* (68-95-99.7 rule) can be applied only if the data have a normal distribution.
- *Pearson's skewness coefficient* is used to find the skewness in a sample.

Descriptive statistics are utilized to organize and describe the characteristics of a data set. Unlike inferential statistics, descriptive statistics do not involve hypothesis testing or data analysis. They enable us to succinctly characterize the distribution of values as a whole.

Descriptive Statistics for Summarizing Numerical Data:

⇒ *Measures of Central Tendency*: Mean, Median and Mode

⇒ *Measures of Variability (Dispersion)*: Range, Interquartile Range, Variance, Standard Deviation, Coefficient of Variation.

3.1 Measures of Central Tendency

Measures of central tendency are useful numbers that characterize the middle (centre) of the data set where observations tend to cluster. The three measures commonly used in medicine are mean, median, and mode. All three are used for summarizing numerical data, and the median is used for ordinal data as well.

3.1.1 Mean

The most frequently used measure of central tendency is the arithmetic mean. The mean is denoted by X-bar ($\bar{x}$) and is calculated by dividing the sum (Σ) of the individual values (X_i) by the number of observations (n):

a) *Simple mean* – used for ungrouped frequency distribution.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

b) *The Weighted Mean* – used for grouped frequency distribution.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\, f$$

Where "f "– is the frequency of individual values

The mean is commonly used to describe numerical data that is normally distributed. It is very sensitive to extreme values in the data set, also known as outliers. For example, the mean of the data set (1;2;2;3) is 8/4 or 2. If the number 19 is substituted for the 3, the data set becomes (1;2;2;19), and the mean is 24/4 or 6. Thus, the mean 2 is more appropriate for the set data, than the mean 6.

3.1.2 Median

The median (M_d) divides the ordered data set into two equal parts. The median is the middle point in the observation data set, with half of the

observations being smaller and half larger. The median is less sensitive to extreme values than the mean. It is frequently used to measure the middle of the distribution of an ordinal or skewed numerical characteristic. When the data are not symmetric, the median is the best measure of central tendency.

To calculate the median, arrange the observations from smallest to largest:

- ⇒ If the number of observations is *odd*, the median M_d will be the middle value in the arranged data or the [(n+1)/2]th observation. For example, the median of the data set with n=5 (1; 2; 4; 5; 6) is $M_d = 4$.
- ⇒ If the number of observations is *even*, the median M_d will be the midpoint between the middle two observations. For example, the median of the data set with n=4 (1; 2; 4; 5) is $M_d = 3$.

3.1.3 Mode

The Mode (M_o) is a value that occurs most frequently in the data set. For example, in the data set (3; 4; 5; 6; 6; 6; 7; 8; 9), Mo=6. There is no mode if all values are different (uniform distribution). There can be more than one mode in a distribution called bimodal or multimodal. Mode is not used frequently in practice.

3.1.4 Empirical Relationship Between the Measures of Central Tendency

The best measure of central tendency for a given set of data depends on the shape of the data distribution:

- ⇒ *Normal Distribution*: In a normal distribution, data values are symmetric around the centre and form a bell-shaped curve. The mean, median, and mode should all be roughly the same. In this case, the mean is the best measure of central tendency.

$$\bar{X} = M_d = M_o$$

⇒ *Skewed Distributions:* In a skewed distribution, outlying observations occur in one direction. The median is the best measure of central tendency in skewed distributions. The mean is sensitive to extreme values and may be excessively inflated or deflated.

There are two types of *skewed distributions:*

- Negatively Skewed (skewed to the left): Outlying values are small.

$$\bar{X} < M_d < M_o$$

- Positively Skewed (skewed to the right): outlying values are large.

$$M_o < M_d < \bar{X}$$

When data are skewed to the right, the mean lies to the right of the median, and when skewed to the left, the mean lies to the left of the median *(Figure 3.1).*

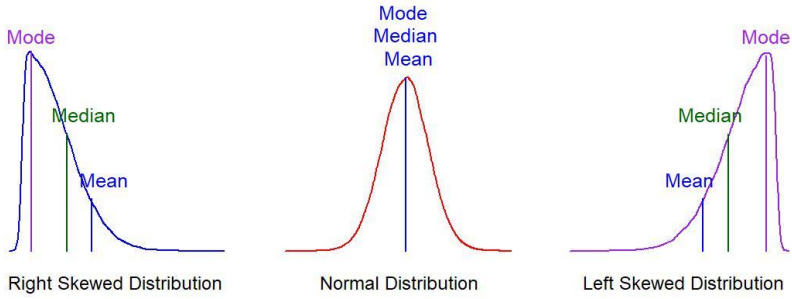

Figure 3.1 Possible data distributions

Guidelines for Choosing the Appropriate Measure of Central Tendency:

The correct practical application of measures of central tendency depends on the scale of measurement and the shape of a distribution.

1. The mean is used for numerical data with a symmetric (normal) distribution.
2. The median is used for ordinal data or numerical data with a skewed distribution.
3. The mode is almost always used for bimodal distributions.

3.2 Measures of Variability

Measurements of central tendency describe only the middle of the data set. To know how good a measure of central tendency is, we need to understand data variability. This includes knowing whether observations are similar and close to the centre or spread out across a broad range of values. In the following example, we see two very different distributions of data values where the mean, median, and the mode are equal:

Data Set 1: -200; -20; -10; 7; 10; 20; 200 (n=7; Mean $\bar{x}$=1; M_d=7)
Data Set 2: -20; -5; -2; 7; 2; 5; 20 (n=7; Mean $\bar{x}$=1; M_d=7)

Despite the significant differences in the data sets, the measures of central tendency are the same. This example illustrates the importance of using measures of central tendency in conjunction with measures of variability to appropriately describe a data set. To comprehensively describe the variability of the data, we need to use measures such as the Range, Interquartile Range, Variance, Standard Deviation and Coefficient of Variation.

3.2.1 Range

Range is the difference between the largest and smallest values in the data set.

Range (R) =Max (x_i) – Min (x_i)

Example:

Data Set 1: (-200; -20; -10; 7; 10; 20; 200) with n=7; Mean $\bar{X}$= 1

Data Set 2: (-20; -5; -2; 7; 2; 5; 20) with n=7; Mean $\bar{X}$=1

R_1 = 400

R_2 = 40

The range is heavily influenced by extreme values and ignores the rest of the distribution. It is highly sensitive to exceptionally large or small values and is used with numerical data to emphasize extreme values.

3.2.2 Interquartile Range

The *interquartile range* (IQR) minimizes the influence of extreme values in the data set. Quartiles are the values that divide an ordered data set into four equal parts. The IQR is calculated as the difference between the 75th percentile (Q_3) and the 25th percentile (Q_1), representing the middle 50% of observations.

$$IQR = Q3 - Q1$$

The 50th percentile (Q_2) is the median of the data set, marking its midpoint. *Figure 3.2* illustrates the quartiles and the interquartile range using a box plot.

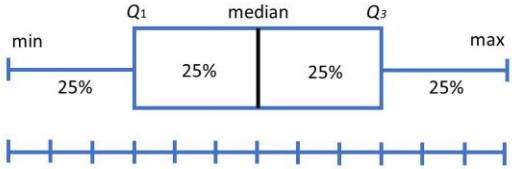

Figure 3.2 Graph presentation of quartiles and interquartile range

To compute the quartiles, rank the data from lowest to highest:

$Q_1 = (n+1) / 4$

$Q_2 = (n+1) / 2$

$Q_3 = 3(n+1) / 4$

The first quartile is the middle value between the smallest value and the median in a distribution. It indicates the 25th percentile. The second quartile is the median in a distribution and indicates the 50th percentile. The third quartile is the middle value between the median and the largest value in a distribution. It indicates the 75th percentile.

The interquartile range is used when the median is used (ordinal data or skewed numerical data). It is appropriate for describing the central 50% of the distribution regardless of its shape.

3.2.3 Variance and Standard Deviation

Variance (s^2) quantifies the amount of variability or spread around the mean. It is calculated as the average squared deviation of each number from the mean. For ungrouped data, variance is calculated using the formula:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

For grouped data, variance is calculated considering the frequency:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i$$

Where:
X_i – individual values of the data set
$\bar{X}$ - mean
n – number of observations
f_i – frequency

The *standard deviation (S)* is the square root of the variance and measures the spread around the mean. For ungrouped data:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

For grouped data:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}{n-1}}$$

Where:
S^2 – variance

The standard deviation is used when the mean is used (symmetric numerical data) and, along with the mean, summarizes the characteristics of the entire distribution of values. The standard deviation has the same units of measurement as the mean.

3.2.4 Coefficient of Variation

The *coefficient of variation* (CV) is a relative measure of variability. It is used to compare distributions measured on different scales. The CV is

defined as the standard deviation divided by the mean and multiplying by 100%.

$$CV = \frac{S}{\bar{x}} \times 100\%$$

Where:

S – Standard deviation

$\overline{x}$ – Mean

The CV is useful for comparing variability levels in different data sets and is represented as a percentage. The CV can be used to estimate the variability level in a data set according to the scale presented in *Table 3.1.*

Table 3.1 The scale of Coefficient of Variation

Coefficient of Variation %	**Level of Variability**
<10	Low
10-35	Medium
>35	High

For scientific research, a set of data should have low or medium variability to ensure that the mean is representative of the data set.

Guidelines for Choosing the Appropriate Measure of Variability:

1. The standard deviation is used when the mean is used (symmetric numerical data).
2. The interquartile range is used:
   a) When the median is used (ordinal data or skewed numerical data).
   b) When comparing individual observations with a set of norms.
   c) When describing the central 50% of a distribution regardless of its shape.

3. The range is used with numerical data to emphasize extreme values.
4. The coefficient of variation is used to compare the variability of different data sets.

3.3 Normal Distribution and its Properties

Normal distributions have key characteristics that are easy to identify in a graph, as presented in *Figure 3.3:*

1. The mean, median and mode are equal.
2. The distribution is symmetric about the mean: half of the values fall below the mean and half above it.
3. The distribution can be described by two values: the mean and the standard deviation.
4. The total area under the curve is 1.

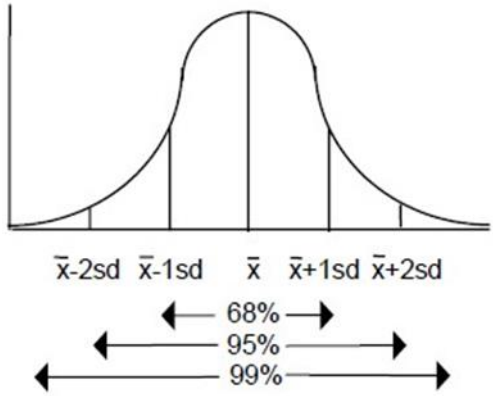

Figure 3.3 The Normal Distribution Properties

If the data is normally distributed (forming a bell-shaped curve), approximately 68% of data will lie within 1 standard deviation, approximately 95% within 2 standard deviations, and approximately 99.7% within 3 standard deviations.

3.4 Skewness and Kurtosis

Skewness represents the degree of deviation in a data set relative to the mean, indicating the direction and extent to which the distribution deviates from a symmetric normal distribution (bell curve). There are different formulas to measure skewness. The following two Pearson's coefficient of skewness (S_k) are used the most often for numerical data (interval or ratio).

Pearson's first coefficient of skewness (mode skewness):

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{\overline{\text{X}} - \text{M}_\text{o}}{\text{S}}$$

When using Pearson's first coefficient of skewness, values lie between -1 and +1. A value of 0 indicates a perfectly symmetrical distribution. Values closer to -1 or +1 indicate increasingly negative or positive skewed distributions.

Pearson's second coefficient of skewness (median skewness):

$$S_k = \frac{3 \times (\text{Mean} - \text{Median})}{\text{Standard Deviation}} = \frac{3 \times (\overline{\text{X}} - \text{M}_\text{d})}{\text{S}}$$

When using Pearson's second coefficient of skewness, values lie between -3 and +3. A value of 0 indicates a perfectly symmetrical distribution. Values closer to -3 or +3 indicate increasingly negative or positive skewed distributions.

Interpretation of the Pearson's second coefficient of skewness:

$S_k = 0$: The distribution is perfectly symmetric.

S_k is between -0.5 and 0.5: The distribution is almost symmetrical.

S_k is between -1 and -0.5: Moderate negative skewness.

S_k is between 0.5 and 1: Moderate positive skewness.

S_k is lower than -1 or greater than 1, the data is highly skewed.

Kurtosis is another measure of the shape of a distribution and refers to the shape of the curve's peak (see *Figure 3.4*). Whereas skewness evaluates the degree of asymmetry, kurtosis signifies the sharpness of the peak in a frequency distribution. Kurtosis is used to find the presence of outliers in the data.

There are three types of kurtosis:

⇒ *Platykurtic* – the peak of the curve is flat compared to normal, and the tails are long (negative kurtosis).

⇒ *Mesokurtic* – the peak of the curve is normal, and the tails on both sides of the mean are also normal.

⇒ *Leptokurtic* – the peak of the curve is narrow, and the tails on both sides of the mean are short (positive kurtosis).

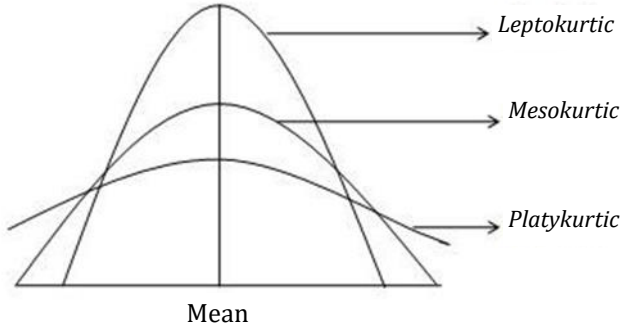

Figure 3.4 Types of Kurtosis

3.5. Example Calculation

Suppose the ages of the 19 patients that you are studying are:
31; 24; 26; 30; 24; 35; 35; 31; 35; 33; 26; 33; 26; 31; 26; 30; 31; 31; 30.
Calculate central tendency and variability measures. State your decision.

1. **Order the Data:**

24; 24; 26; 26; 26; 26; 30; 30; 30; 31; 31; 31; 31; 31; 33; 33; 35; 35; 35.

2. **Arrange Data in the Frequency Table:**

X_i	*Frequency*, f	$x_i f_i$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^2 f_i$
24	2	48	-5.8	33.64	67.28
26	4	104	-3.8	14.44	57.76
30	3	90	+0.2	0.04	0.12
31	5	153	+1.2	1.44	7.2
33	2	66	+3.2	10.24	20.48
35	3	105	+5.2	27.04	81.12
Total	n=19	566			233.96

3. **Calculate the Mean:**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i f_i$$

$$\bar{X} = \frac{24 \times 2 + 26 \times 4 + 30 \times 3 + 31 \times 5 + 33 \times 2 + 35 \times 3}{19} = \frac{566}{19} = 29.8$$

4. **Determine the Mode:**

$Mode\ (M_o) = 31$

5. **Calculate the Median**

Since n = 19 odd, the median position is $\frac{n+1}{2} = \frac{19+1}{2} = 10$.

The 10th observation in the ordered data is 31.

$Median\ (M_d) = 31$

6. Calculate the Variance and Standard Deviation:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i = \frac{233.96}{18} = 13$$

$$s = \sqrt{s^2} = \sqrt{13} \approx 3.6$$

7. Calculate the Coefficient of Variation:

$$CV = \frac{S}{\bar{x}} \times\ 100\%$$

$$CV = \frac{3.6}{29.8} \times\ 100\% = 12.1\%$$

8. Conclusions:

- The mean age of the patients is 29.8±3.6 years.
- The data variability level is medium (CV=12.1%), which is appropriate for scientific research.
- The distribution of data is almost symmetric, with Mode, Median and Mean values close to each other (slightly skewed to the left).
- The mean is representative of this data set.

Review Exercises:

1. Using the following BMI data set of 10 patients:
   Data: 29, 22, 24, 37, 23, 35, 35, 27, 50, 37.
   - Find the measures of central tendency.
   - Analyze the relationship between the measures of central tendency and identify the shape of the distribution. State your conclusion.

2. Using the following blood glucose level data set of 11 patients:
   Data: 3.3, 12.0, 9.0, 6.0, 11.0, 11.8, 11.8, 11.0, 5.5, 3.3 (mmol/l).
   - Find all measures of central tendency.
   - Analyze the relationship between the measures of central tendency and identify the shape of the distribution. State your conclusion.

3. Using the BMI data set from Exercise 1:
   - Find all measures of variability. State your conclusion.
   - Determine how representative the mean is for this data set.
   - Based on the scale of measurements and the shape of the distribution, decide which summary measures of central tendency are more appropriate.
   - Construct a box plot.

4. Using the blood glucose level data set from Exercise 2:
   - Find all measures of variability. State your conclusion.
   - Determine how representative the mean is for this data set.
   - Based on the scale of measurements and the shape of the distribution, decide which summary measures of central tendency are more appropriate.
   - Construct a box plot.

Review Questions

1. The mean: definition, types, and rules of calculation. Give an example.
2. The median: definition and rules of calculation. Give an example.
3. The mode: definition and rules of calculation. Give an example.
4. Compare the mean, median, and mode as measures of central tendency.
5. Under what conditions is the use of the mean preferred?
6. Under what conditions is the use of the median preferred?
7. Under what conditions is the use of the mode preferred?
8. Variability measures: reasons for application.
9. Range: definition, characteristics, and preferred use conditions. Rule of calculation.
10. Quartiles: definition, characteristics, and rule of calculation.
11. Interquartile range: definition, characteristics, and preferred use conditions. Rule of calculation.
12. Standard deviation: meaning, characteristics, and preferred use conditions. Rule of calculation.
13. Coefficient of variation: definition, characteristics, and preferred use conditions. Rule of calculation.
14. Define the normal distribution and its properties.

CHAPTER 4. DESCRIPTIVE STATISTICS: SUMMARIZING NOMINAL AND ORDINAL DATA

Key Concepts

- *Qualitative data* can be measured using several methods: ratios, proportions and rates.
- A *ratio* is a number of observations with a given characteristic "a" divided by a number of observations without that characteristic "b".
- A *proportion* is a number of observations with a given characteristic "a" divided by the total number of observations "a+b".
- A *rate* is similar to proportion but uses a multiplier (e.g., 1,000, 10,000, or 100,000) and is computed over a specified period.
- Bar/column charts and line charts are used to represent ratios and rates graphically.
- A pie chart is used to represent graphically proportion.
- A *proportion* expresses the structure of a phenomenon and is a static indicator.
- A *rate* expresses the frequency (level or intensity) of a phenomenon over time and is a dynamic indicator.
- *Rates* are crucial in epidemiology and evidence-based medicine; they form the basis for calculating vital statistics, which describe the health status of populations.
- *Rates* are usually computed for a year (annual rates).
- *Rates* can be crude, specific and standardized.
- *Rates* must be standardized when comparing populations with significant differences in their structure (e.g., by age and sex).
- *Prevalence* and *incidence* are two important measures of morbidity.

4.1 Methods to Describe Categorical Data

In research, statistical information is usually presented as absolute values. These values can be difficult to interpret as they do not allow for meaningful comparisons, synthesis or correlation among different characteristics. To make comparisons among groups more meaningful, relative values may be used instead of absolute numbers. Categorical (qualitative) data can be measured using three methods:

- Proportion and percentages
- Ratios
- Rates

4.1.1 Proportions and Percentages

Proportions are statistical indicators that express the structure of a phenomenon. They are defined as the part of a phenomenon divided by the whole. The proportion is calculated by dividing the number of observations with a characteristic of interest (*a*) by the total number of observations ($a+b$). A percentage is simply the proportion multiplied by 100%.

$$Proportion\ (Percentage) = \left(\frac{a}{a+b}\right) \times 100\%$$

Where:

- a is the number of observations with the characteristic of interest
- b is the number of observations without the characteristic.

Example of proportion:

$$Proportionate\ mortality = \frac{Number\ of\ deaths\ from\ a\ cause\ X\ in\ a\ year}{Total\ number\ of\ deaths\ in\ a\ year} \times 100$$

Number of deaths from deaths from cancer in the Republic of Moldova in 2023 = 5,959
Number of deaths from all causes in the Republic of Moldova in 2023 = 33,733

$$Proportionate\ mortality\ from\ cancer\ in\ 2023 = \frac{5{,}959}{33{,}733} \times 100 = 18\%$$

Proportions are called *extensive* indicators as they show the structure of a phenomenon. They are *static* indicators that provide a snapshot of a phenomenon at a specified moment and do not account for dynamic changes. They are useful for ordinal and numerical data as well as nominal data, especially when the observations have been placed in a frequency table.
Graph presentation: Pie chart.

4.1.2 Ratios

A ratio is defined as one part divided by another part representing two independent phenomena. The ratio is calculated by dividing the number of observations in a group with a given characteristic (a) by the number of observations without the given characteristic (b):

$$Ratio = \frac{a}{b}$$

Examples of ratios:

$$Sex\ ratio = \frac{Number\ of\ males}{Number\ of\ females}$$

Number of male newborns in the Republic of Moldova in 2023 = 12,239
Number of female newborns in the Republic of Moldova in 2023 = 11,794

$$Sex\ ratio\ at\ birth\ in\ 2023 = \frac{12239}{11794} = 1.04$$

In this example, the ratio has no units. However, the ratio can be rescaled by multiplying it by a base, such as 100 or 1,000. The sex ratio at birth in the

Republic of Moldova in 2023 was 1.04 or 104% (1.04 x 100%), i.e. 104 male newborns for every 100 female newborns.

$$Medical\ supply = \frac{Number\ of\ doctors}{Total\ population} \times base$$

Number of doctors in the Republic of Moldova in 2022 = 12,600
Number of the total population in the Republic of Moldova in 2022 = 2,565,030

$$Medical\ supply\ in\ 2023 = \frac{12600}{2565030} \times 1{,}000$$
$$= 5\ doctors\ per\ 1000\ population$$

Graph presentation: Line graph, bar/column chart.

4.1.3 Rates

A rate is similar to a proportion but includes a base and a time dimension. A rate is always computed for a specified period, usually a year (annual rate). *Rates* are *intensive* statistical indicators expressing the frequency or level of a phenomenon over a specified period. They are calculated as follows:

$$Rate = \left(\frac{a}{a+b}\right) \times base$$

Where:

- *a* is the number of observations with a given characteristic (e.g., number of deaths or births in a given year and place);
- *a+b* is the total number of observations (e.g., total population in a given year);
- *base* is a multiplier (e.g., 100; 1000; 10000; 100000).

Example of rate:

For example, a study lasted 1 year and the proportion of patients who developed a disease was 0.02. The rate per 1,000 patients would be (0.02) x (1,000) or 20 diseases per 1,00 patients in a year.

Graph presentation: Line graph, bar/column chart.

4.2 Health Status Indicators in Descriptive Statistics

Health status indicators assess the health of a population using vital statistics. They are categorized into three main types:

⇒ Mortality indicators;

⇒ Morbidity indicators;

⇒ Disability indicators.

Most health status indicators are represented by rates and sometimes by proportions and ratios. Mortality rates and morbidity rates are the most commonly used rates.

4.2.1 Mortality Rates

Mortality rate or death rate is defined as the number of deaths that occur during a specified period divided by the total number of people at risk of dying during the same period.

⇒ *Crude rate:* Computed over all individuals in a given population regardless of differences caused by age, gender, race, etc.

⇒ *Specific rate*: Computed within relatively small, well-defined population subgroups**.** For example:

- Age-specific death rates;
- Sex-specific death rates;
- Cause-specific death rates.

4.2.2 Morbidity Rates

Morbidity rate is defined as the number of individuals who developed a disease during a specified period divided by the total number of people at risk during the same period.

Incidence and *Prevalence* are the main measures of morbidity and are commonly used to evaluate the population's health status in many medical and epidemiological researches.

Incidence: The number of new cases that occur during a specified period divided by the total number of people at risk during that period.

Prevalence: The number of individuals with a given disease at a given point in time divided by the total population at risk for that disease at that time.

Morbidity rates provide a standard way to evaluate both crude and specific rates.

Example Calculation:

In a specified year and locality, the population is 75,000. In that year, 897 individuals died. In that locality, there were 40 doctors: 20 physicians, 10 surgeons and 10 others. Compute statistical indicators summarizing nominal data.

1. *Proportion and Percentage*

$$Proportion\ (Percentage) of\ physicians = \frac{20}{40} \times 100\% = 50\%$$

2. *Ratio*

$$Medical\ supply = \frac{40}{75000} \times 10{,}000 = 5.3\ doctors\ per\ 10000\ population$$

3. *Rate*

$$Crude\ death\ rate = \frac{897}{75000} \times\ 1000 = 11.9\ deaths\ per\ 1000\ population$$

4.3 Standardized Rates: Direct Method of Standardization

Crude rates can be used to compare two different populations only if the populations are homogeneous in all characteristics. However, if the populations differ by characteristics such as gender or age, using crude rates can lead to incorrect results. This is because crude rates are strongly influenced by the age structure of a population. For example, one population may have a higher proportion of elderly individuals than

the other. In such cases, crude rates must be adjusted or standardized to allow valid comparisons.

Standardization adjusts crude rates to eliminate the effects of differences in population composition, primarily by age and sex. Two principal methods of standardization exist: direct and indirect standardization. The direct method of standardization calculates the rates that would result if all groups being compared had the same standard composition. Adjusted rates are thus conventional values designed solely for comparison purposes and cannot be used independently.

The direct method of standardization of rates includes four steps:

1. Compute specific rates for each group;
2. Select the standard population;
3. Compute the expected number of events for each group;
4. Calculate adjusted rates.

Below is a step-by-step example of the calculation of standardized death rates for the two regions.

Example Calculation: Direct Method of Standardization of Death Rates.

Step 1: Compute Specific Rates for Each Group

Sex	**Region A**		**Region B**		**Sex-specific death rate (per 1,000)**	
	Persons	Deaths	Persons	Deaths	Region A	Region B
Males	50	1	170	4	**20**	**24**
Females	200	10	30	3	**50**	**100**
Total	250	11	200	7	**44**	**35**

For example, the sex-specific death rate for males in Region A = $\frac{1}{50} \times 1{,}000 = 20$ deaths per 1,000 males.

Step 2. Select the Standard Population

Sex	**Region A**		**Region B**		**Standard population**
	Persons	Deaths	Persons	Deaths	Persons from A + persons from B
Males	50	1	170	4	**220**
Females	200	10	30	3	**230**
Total	250	11	200	7	**450**

In this example, the standard population is the sum of populations from both regions. In other words, we assume that the population structure by sex is the same in both regions.

Step 3. Compute the Expected Number of Events for Each Group

Sex	Specific rate (per 1,000)		Standard population	Expected events (deaths)	
	Region A	Region B	Persons from A + persons from B	Region A	Region B
Males	20	24	220	4	5
Females	50	100	230	12	23
Total				**16**	**28**

$$\text{Number of expected events} = \frac{Specific\ rate \times standard\ population}{1{,}000}$$

For example, the number of expected deaths for region A in males = $\frac{20\times220}{1{,}000} = 4$ deaths. In other words, if the number of the male population in region A was the standard population (220 persons), we would expect 4 deaths here.

Step 4. Calculate the standardized (adjusted) rate for each group (region)

Sex	Standard population	Expected events (deaths)		Standardized rate (per 1,000)	
		Region A	Region B	Region A	Region B
Males	220	4	5		
Females	230	12	23		
Total	450	16	28	**36**	**62**

$$Standardized\ rate = \frac{Total\ number\ of\ expected\ events}{Total\ standard\ population} \times 1{,}000$$

For example, the standardized death rate for Region A = $\frac{16}{450} \times 1{,}000 = 36$ deaths per 1,000 standard population.

Conclusion: Comparison of crude and standardized rates

Crude rate (per 1,000)		**Standardized rate (per 1,000)**	
Region A	Region B	Region A	Region B
44	**35**	**35**	**63**

The intensity (level) of mortality is higher in region B. The level of mortality in Region B is 1.8 times higher than in Region A (63 ÷ 35 = 1.8).

Review Exercises

1. In locality A in a specified year, 2,500 illnesses were registered: 800 of them were cardiovascular diseases; 500 were pulmonary diseases; 450 were injuries; and 750 were other diseases. The population number is 900,000.
   - Compute all possible vital statistics.
   - Make an appropriate graph presentation for them.

2. In locality B in a specified year, the population number is 78,000. This year, 110 individuals died and 400 individuals developed cardiovascular disease for the first time in their lives.
   - Compute all possible vital statistics.
   - Make an appropriate graph presentation for them.

3. Consider the following data comparing acute abdomen lethality at hospitals "A" and "B":

Term of hospitalization, hours	**Hospital "A"**		**Hospital "B"**	
	Number of patients	Number of lethality cases	Number of patients	Number of lethality cases
< 6	650	72	490	34
6-12	450	83	380	66
>24	131	23	736	206
Total:	1,231	178	1,606	306

   - Compute the crude rates and compare these rates.
   - How do the adjusted rates differ from the crude rates in each of these hospitals? Explain these results (interpretation and conclusions).

4. Consider the following data comparing hospital mortality at hospitals "A" and "B":

Disease	Hospital "A"		Hospital "B"	
	Number of patients	Number of deaths	Number of patients	Number of deaths
Gastrointestinal	1,200	24	1,700	40
Malign tumour	190	55	100	30
Cardiovascular	160	100	1100	72
Total:	1,650	179	2,900	142

- Compute the crude rates and compare these rates.
- How does the adjusted mortality rate differ from the crude mortality rate in each of these hospitals? Explain these results (interpretation and conclusions).

Review Questions

1. Absolute and relative values: their meaning and application in Biostatistics. Under what conditions is the use of the relative values preferred? Give an example.
2. Types of relative values: what are the differences and similarities among them? Give an example for each type.
3. Proportions, ratios and rates: particularity, rules of calculation, conditions of application and appropriate graph presentation. Give an example.
4. The Vital Statistics: definition and the main used rates.
5. What is the difference between crude and specific rates?
6. What is the difference between mortality and morbidity rates?
7. What are the differences and similarities between prevalence and incidence?
8. Under what conditions is the use of the adjusted rate preferred?
9. Direct method of standardization: definition and its process steps.
10. Under what circumstances should crude, specific and adjusted rates each be used?

CHAPTER 5. CORRELATION AND REGRESSION

Key Concepts

- *Correlation coefficient* (*r*): Measures the strength and direction of the linear relationship between two variables.
- *Variable X*: Often referred to as the independent or explanatory variable.
- *Variable Y*: Known as the dependent or outcome variable.
- The value of the correlation coefficient varies from -1 to +1.
- *Direction of Correlation*: Indicated by the sign (+ or -) of the coefficient.
- *Strength of Correlation*: Indicated by the magnitude of the coefficient.
- *Positive Correlation* (+): High values on one variable are associated with high values on the other variable.
- *Negative Correlation* (-): High values on one variable are associated with low values on the other variable.
- *Pearson's Correlation Coefficient* (r): Evaluates the linear relationship between two numerical variables.
- *Spearman's Rank Correlation Coefficient* ($\boldsymbol{r_s}$): Evaluates the relationship between two variables using rank orders.
- *Regression:* Predicts the value of a dependent variable (Y) based on one or more independent variables (X).
- *Linear regression* can be simple and multiple.

Biomedical research frequently investigates the relationship between two or more variables. For example, is there a relationship between salt consumption and blood pressure or between cigarette smoking and life expectancy? Understanding these relationships is crucial for uncovering associations and making predictions. To examine these relationships, two fundamental statistical techniques are commonly used: correlation and regression.

1. *Correlation:* The technique is used to establish and quantify the strength and direction of the relationship between two variables.
2. *Regression*: This method is used to express the functional relationship between two or more variables. Regression analysis allows the researcher to predict the value of one variable based on the value of another variable.

5.1 Correlation

5.1.1 Types of Correlation Coefficient

Pearson's Correlation Coefficient

Pearson's correlation coefficient is a parametric measure of the relationship between two normally distributed numerical variables. The Independent or explanatory variable is "X" and the dependent or outcome variable is "Y".

The correlation coefficient is symbolized by "r" and is calculated using the formula:

$$r = \frac{\Sigma\ (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

Where:

X_i: the individual score of the independent variable

Y_i: the individual score of the dependent variable

$\bar{X}$: the mean of the independent variable

$\bar{Y}$: the mean of the dependent variable

The correlation coefficient is a dimensionless number, meaning that it has no units of measurement. Tits value ranges from -1 to 1. For any given set of data, the correlation coefficient (r) satisfies the inequality:

$$-1 \leq r \leq 1$$

The direction of the relationship is indicated by the sign of the coefficient, while the strength of the relationship is indicated by the magnitude of the coefficient.

Direction of Correlation:

⇒ *Positive correlation* $(0 < r \leq 1)$ indicates that as one variable increases, the other variable also tends to increase. For example, higher salt intake is associated with higher blood pressure.

⇒ *Negative correlation* $(-1 \leq r < 0)$ indicates that as one variable increases, the other variable tends to decrease. For instance, higher cigarette consumption is associated with lower life expectancy.

Specific values:

$r = 1$: Indicates a perfect positive linear relationship.

$r = -1$: Indicates a perfect negative linear relationship.

$r = 0$: Indicates no linear relationship (no correlation) between the variables.

Interpreting the Strength of Correlation:

⇒ 0 to 0.25 (±): No or weak correlation;

⇒ 0.25 to 0.50 (±): Moderate correlation*;*

⇒ 0.50 to 0.75 (±): Strong correlation;

⇒ 0.75 to 1 (±): Very strong;

⇒ ± 1: Perfect correlation.

This scale helps in understanding how closely the two variables are related. Higher absolute values of r denote a stronger relationship, while values closer to zero indicate a weaker relationship.

Scatterplots

Scatterplots are a useful visual tool for displaying the relationship between two numerical variables. They help in assessing the linearity of the relationship and identifying the outliers.

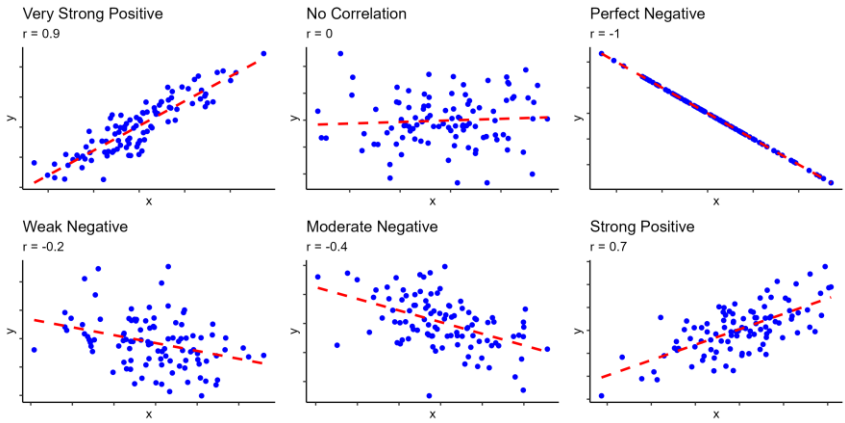

Figure 5.1 Scatter Plots Illustrating Various Strengths of Correlation

Example of calculating Pearson's Correlation Coefficient

Consider the relationship between the Length (independent variable) and the Weight (dependent variable) of nine newborns (*Table 5.1*). To calculate Pearson's correlation coefficient:

1. Calculate the mean of X and Y variables.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{437}{9} = 48.6 \qquad \bar{Y} = \frac{\sum Y_i}{n} = \frac{25.7}{9} = 2.9$$

2. Calculate the deviation scores of each X and Y variables.

$$(X_i - \bar{X}) \qquad \text{and} \qquad (Y_i - \bar{Y})$$

3. Square deviation scores for variables X and Y:

$$(X_i - \bar{X})^2 \qquad \textit{and} \qquad (Y_i - \bar{Y})^2$$

4. Calculate the coefficient using the formula:

$$r = \frac{\Sigma\,(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}} = \frac{17.5}{\sqrt{106.2 \times 3.5}} = \frac{17.5}{19.2} = 0.91$$

Table 5.1. Data on length (cm) (X_i) and weight (kg) (Y_i) gathered for nine new-borns

Number of a child	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X}) \times (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	43.0	2.2	30.9	0.4	3.6	30.9	0.4
2	45.0	2.1	12.6	0.6	2.7	12.6	0.6
3	47.0	2.4	2.4	0.2	0.7	2.4	0.2
4	48.0	2.4	0.3	0.2	0.3	0.3	0.2
5	49.0	3.0	0.2	0.0	0.1	0.2	0.0
6	49.0	2.7	0.2	0.0	-0.1	0.2	0.0
7	50.0	3.5	2.1	0.4	0.9	2.1	0.4
8	50.0	3.4	2.1	0.3	0.8	2.1	0.3
9	56.0	4.0	55.4	1.3	8.5	55.4	1.3
Total	**437**	**25.7**	**106.2**	**3.5**	**17.5**	**106.2**	**3.5**

Conclusion: The correlation coefficient r is 0.91. The relationship between the length and weight of newborns is positive and very strong.

Spearman's Rank Correlation Coefficient

Spearman's rank correlation is a non-parametric measure used for the following types of data:

- For two ordinal variables;
- For one ordinal and one numerical variable;
- For two numerical variables if either of the two variables is not normally distributed.

This coefficient is symbolized as r_s and is calculated by ranking the values of each variable and then applying Pearson's formula to the ranks. The formula for Spearman's rank correlation is as follows:

$$r_s = \frac{\Sigma\,(R_x - \bar{R}_x)\left(R_y - \bar{R}_y\right)}{\sqrt{\Sigma(R_x - \bar{R}_x)^2 \Sigma\left(R_y - \bar{R}_y\right)^2}}$$

Where:

R_x and R_y are the ranks of the X and Y variables

$\bar{R}_x$ and $\bar{R}_y$ are the mean ranks for the X and Y variables

As the Pearson's correlation coefficient, the Spearman rank correlation coefficient ranges in value from -1 to 1. Values of r_s closer to the extremes indicate a high degree of correlation between X and Y; values closer to 0 imply a weaker relationship.

Other types of Correlations

- Point-Biserial Correlation Coefficient: Used when one variable is numerical and the other is dichotomous.
- Phi Coefficient: used for two dichotomous variables.

Important Considerations

⇒ *Correlation does not imply causation*. A significant correlation between two variables does not mean that changes in one variable cause changes in the other variable. The correlation coefficient is only a measure of the relationship between two variables. Inferring a causal relationship between two variables based on a correlation is a common and fundamental error.

⇒ The fact that a correlation is present between two variables in a sample does not mean that the correlation exists in the population. Statistical tests, such as t-tests, are required to determine the significance of the correlation (see Chapter 6).

5.1.2 Coefficient of Determination

The coefficient of determination, denoted as R^2, is the square of the correlation coefficient. It represents the proportion of the variance in one variable that is explained by the variance in the other variable. When the two variables are correlated, there is a certain amount of the shared variance between them. The stronger the correlation, the greater the amount of the shared variance and the higher the coefficient of determination. The value of R^2 ranges from 0 to 100% (*Figure 5.2*).

Example of Calculating Coefficient of Determination

Consider a study investigating the relationship between hours of study per week (variable X) and exam scores (variable Y) among medical students. Suppose the study finds a correlation *r* of 0.70. To calculate the coefficient of determination:

$R^2 = (0.70)^2 = 0.49$ or 49%

This result indicates that 49% of the variance in exam scores is explained by the variance in the number of study hours. Therefore, nearly half of the difference in exam scores among students can be attributed to differences in their study habits.

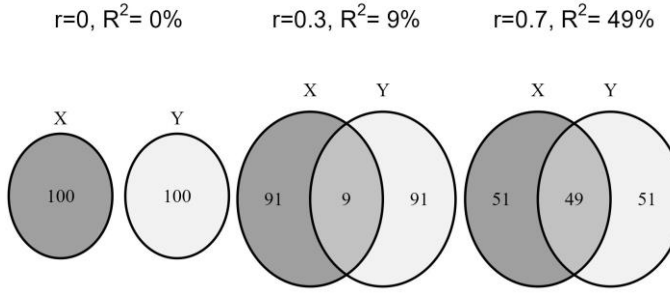

Figure 5.2 Graphical explanation of the coefficient of determination (R^2)

5.2 Regression: General Approaches

5.2.1 Simple Linear Regression

If two variables are highly correlated, it is possible to predict the value of the dependent variable from the value of the independent variable by using regression methods. Correlation analyses make no such distinction, as the two variables involved are treated symmetrically. Like correlation analyses, simple linear regression is a technique that is used to explore the nature of the relationship between two continuous variables. The primary difference between these two statistical

analytical methods is that regression enables us to investigate the change in one variable Y (dependent or outcome variable), which corresponds to a given change in the other variable X, known as the independent or explanatory variable by *regression equation*, which quantifies the straight-line relationship between the two variables. This straight line, or *regression line*, is the same "line of best fit" to the scatterplot as that used in calculating the correlation coefficient.

In the case of the simple linear equation, the value of one variable (X) is used to predict the value of the other variable (Y) using the regression equation. The equation for the simple linear regression is as follows:

$$Predicted\ value\ Y = \alpha + \beta X$$

Where:

Y – the expected value of Y (dependent variable)

α – a constant known as the "intercept constant"

β – the "slope constant" of the regression line

X – the value of the variable X (independent variable)

The equation of the simple linear regression is known as the slope-intercept form. The slope (β) is the number being multiplied by the value of X and the intercept (α) is the number being added or subtracted. The slope constant shows the change in Y when X increases by 1 unit. The intercept constant is a value of Y when X is 0 (it is the point where the Y axis is intercepted by the regression line).

The regression coefficient (β) approaches are as follows:

If $\beta = 0$, the variable *Y* is not dependent on the variable *X*

If $\beta \neq 0$, the variable *Y* is dependent on the variable *X* as follows:

$\beta > 0$, the positive direction of the relationship between Y and X

$\beta < 0$, the negative direction of the relationship between Y and X

Once the values of *a* and *b* are determined, the expected value of *Y* can be predicted for any given value of *X.* For example, it has been

shown that the hepatic clearance rate of lidocaine (*Y* variable, mL/min/kg) can be predicted based on the hepatic clearance rate of indocyanine green dye (*X* variable, mL/min/kg), using the equation $Y = 0.30 + 1.07X$. This approach allows anaesthesiologists to mitigate the risk of lidocaine overdosage by evaluating the clearance of the dye (Pagano M., Gauvreau K., 2000).

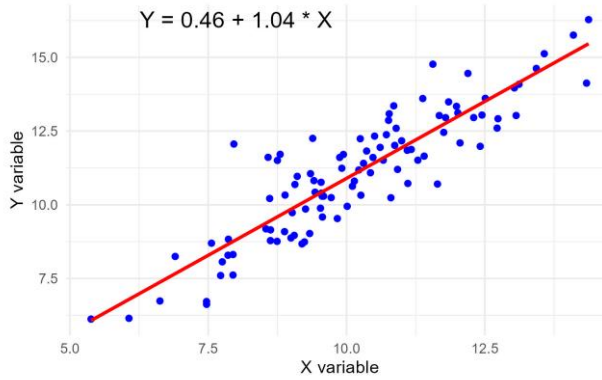

Figure 5.3 Scatterplot with regression line for a set of hypothetical data

5.2.2 Multiple Linear Regression

Multiple linear regression techniques are applied when more than one continuous variable X is used to predict the expected value of *Y*. The multiple regression equation is formulated as follows:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

5.2.3 Logistic Regression

In the case of linear regression, the response variable Y is numerical and assumed to follow a normal distribution. *Logistic regression* addresses situations where predictor variables (independent X) are numerical and

predicted variables (dependent Y) are nominal (e.g., survival vs death, cured vs uncured). Instead of predicting a mean value, logistic regression estimates the probability associated with the dichotomous response for various values of an explanatory variable.

Review Exercises

1. A group of 10 students were observed for the number of hours they studied and their corresponding exam scores.

Data Set 1. Hours Studies vs. Exam Scores

Nr. of observations	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Hours Studied	2	3	4	5	6	7	8	9	10	11
Exam Scores	56	61	68	70	72	75	78	80	85	88

- Calculate the appropriate coefficient of correlation. State your choice.
- Interpret the computed coefficient of correlation.
- Create a scatter plot for these data.

2. A group of 10 individuals recorded their daily exercise time (in minutes) and their corresponding weight loss (in kg) over a month.

Data Set 2. Daily Exercise Time vs. Weight Loss

Nr. of observations	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Daily Exercise Time (minutes)	15	20	25	30	35	40	45	50	55	60
Weight Loss (kg)	0.5	0.8	1.0	1.3	1.5	1.8	2.0	2.3	2.5	2.8

- Calculate the appropriate coefficient of correlation. State your choice.
- Interpret the computed coefficient of correlation.
- Create a scatter plot for these data.

3. If the relationship between two measures is linear and the coefficient of correlation has a value near 1, a scatterplot of observations:

a) Is a horizontal straight line
b) Is a vertical straight line
c) Is a straight line that is neither horizontal nor vertical

d) Has a negative slope
e) Has a positive slope

4. If the relationship between two measures is linear and the coefficient of correlation has a value near -1, a scatterplot of observations:

a) Is a horizontal straight line
b) Is a vertical straight line
c) Is a straight line that is neither horizontal nor vertical
d) Has a negative slope
e) Has a positive slope

Review Questions

1. Under what conditions is the use of the correlation preferred?
2. What are the strengths and limitations of Pearson's correlation coefficient?
3. How does Spearman's rank correlation differ from the Pearson correlation?
4. When you are investigating the relationship between two continuous variables, why is it important to create a scatter plot data?
5. If a test of hypothesis indicates that the correlation between two variables is not significantly different from 0, does this necessarily imply that the variables are independent? Explain.
6. What is the main distinction between correlation analyses and regression analyses?
7. Under what conditions is the use of the simple linear regression is preferred?
8. Under what conditions is use of the multiple linear regression preferred?
9. Under what conditions is the use of logistic regression preferred?

CHAPTER 6. INFERENTIAL STATISTICS: PROBABILITY THEORY AND HYPOTHESES TESTING

Key Concepts

- *Inferential statistics* are derived from sample data and are used to make inferences (conclusions) about the population parameters.
- Inferential statistics can be used only for *probability samples*. For *non-probability samples*, only descriptive statistics can be used.
- *Probability* plays a key role in inferential statistics. When it comes to deciding whether a result in a study is statistically significant, we must rely on probability to make the determination.
- The sampling distribution of means will always tend to be normal, irrespective of the shape of the population distribution from which the samples were drawn (*Central Limit Theorem*).
- *Standard error* is the average difference between the population mean and an individual sample mean.
- *Standard error* is inversely related to the square root of the sample size (n). Small samples produce large standard errors.
- The *null hypothesis* (H_0) always suggests that there is an absence of effect in a population (e.g., two population means will not differ).
- The *alternative hypothesis* (H_1) predicts that there are differences between the groups (e.g., two population means will differ).
- A *two-tailed* or *non-directional alternative hypothesis* gives no speculation about which value will be larger.
- A *one-tailed* or *directional alternative hypothesis* specifies which value will be larger.
- The rule of "AAA" for a Type I error: Alpha error Accepts the false Alternative Hypothesis.

- The rule of "BEAN" for a Type II error: Beta error Accepts the false Null Hypothesis.
- *Confidence level* is the probability of accepting the Null when it is true.
- *Power of the study* is the probability of rejecting the Null when it is false.
- *Significance level (α-level)* is the probability of making a type I error, set before calculating the statistical test.
- The α-level is typically set at 0.05 or less.
- *The p-value* is the probability of making a type I error, found after calculating the statistical test.
- If the p-value < α level (p-value < 0.05), you can reject the Null. You can conclude that the difference between the two means is statistically significant.
- *A confidence interval* is an interval calculated using sample statistics that contains the population parameter (e.g., mean) within a certain degree of confidence (e.g., 95% or 99% confidence).

Inferential statistics use sample data to make conclusions about population parameters. These statistics help us determine whether an observed phenomenon in a sample truly exists in the broader population from which the sample was drawn.

6.1 Probability Theory

6.1.1 General Concepts

Probability is crucial in inferential statistics, especially when deciding if a study result is *statistically significant*, i.e., our result is not due to chance. Probability theory is fundamental to analysing large data sets and repeated experiments, called trials, which yield various outcomes.

The classical definition of probability (*p*) states that the probability of an event occurring is the number of favourable outcomes (m) divided by the total number of possible outcomes (n):

$$p = \frac{m}{n}$$

The probability of an event not occurring (q) is:

$$q = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p$$

Thus, the sum of the probabilities of an event occurring and not occurring equals 1 (or 100%). The value of *p* ranges from 0 to 1 (0% to 100%), with a higher p indicating a higher likelihood of the event occurring.

Two significant mathematical concepts in probability theory are the Law of Large Numbers and the Central Limit Theorem.

6.1.2 Law of Large Numbers

The *Law of Large Numbers* states that as the number of trials in an experiment increases, the average of the results becomes closer to the expected value. By approaching a specific number of trials, the average of the experiment results becomes as close as possible to the expected value.

6.1.3 Central Limit Theorem

The *Central Limit Theorem* describes the properties of the sampling distribution of the sample means:

1. The mean of sampling distribution is equal to the population mean (μ).

2. The standard deviation of the sampling distribution of the means is known as the *standard error of the mean* ($SE_{\bar{x}}$;). The standard error of the mean is calculated by dividing the population standard deviation (ϭ) by the square root of the sample size (*n*):

$$SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Where:

$SE_{\bar{x}}$: standard error of the mean

σ: standard deviation of the population

n – sample size.

In other words, the standard error represents the average difference between the population mean and the mean of an individual sample.

3. Provide that *n* is large enough, the shape of the sampling distribution is approximately normal, regardless of the shape of the population distribution from which the samples were drawn.

6.1.4 Using the Standard Error

The standard error depends on the sample size; the larger the sample size, the more closely the sample mean ($\bar{X}$) represents the population mean (μ). The standard error indicates the uncertainty around the mean estimate, reflecting how much the sample mean would vary with repeated sampling. Large standard errors result from small samples with large standard deviations. The standard error is used to calculate confidence intervals, which help infer population parameters.

6.2 Sampling

6.2.1 Sampling Definition

Probability theory allows us to make inferences about population characteristics using sample data. *Sampling* involves selecting a group from the population to collect data for research. Reasons for sampling include saving time and money, obtaining more accurate results, reducing heterogeneity and estimating errors.

6.2.2 Sampling Methods

Several sampling methods are used in medical research, each requiring randomness to use inferential statistics effectively. *Probability samples* ensure each subject's inclusion probability is known, leading to valid inferences. Non-probability samples only allow for descriptive statistics.

A. Probability Sampling:

- ⇒ *Simple Random Sampling*: Each population member has an equal chance of selection. Units are selected randomly until the sample size is reached.
- ⇒ *Systematic Sampling*: Every k^{th} item is selected from a population list, with *k* determined by dividing the population size (N) by the desired sample size (n).
- ⇒ *Stratified Sampling*: The population is divided into mutually exclusive groups (strata), and random samples are drawn from each stratum.
- ⇒ *Cluster Sampling*: At first, a set of groups or "clusters" is selected randomly from a population, and then cases are selected at random from clusters. Clusters are commonly based on geographic areas.

B. Non-Probability Sampling:

Non-probability samples are those in which the probability that a unit is selected is unknown.

⇒ Convenience Sampling: The researcher selects the easiest population members to obtain information.

⇒ Quota Sampling: The researcher interviews a prescribed number of people in each of several categories.

6.3 Estimation and Hypothesis Testing

Estimation and hypothesis testing are key components of inferential statistics, allowing researchers to draw conclusions about data and relationships between variables.

Estimation:

- *Point Estimation:* Uses sample data to calculate a single number to estimate the parameter of interest, such as a population mean, without providing information about the variability of the estimate. It is not known how close the sample mean ($\bar{X}$) is to the population mean (μ).
- *Interval estimation:* Uses a range of values to estimate the parameter, providing a confidence interval (CI) that contains the population mean (μ) with a certain degree of confidence.

Hypothesis Testing: Involves stating a null hypothesis (H_0) and an alternative hypothesis (H_1) and performing a statistical test to determine which hypothesis to accept. The goal is to disprove the null hypothesis and accept the alternative.

6.4 Confidence Intervals

A confidence interval is an interval estimate of a population parameter, represented by mean, proportion, correlation coefficient or differences between two means or proportions. The ends of the interval are called *confidence limits.* Confidence intervals (CI) are calculated using the sample mean and standard error:

$$CI = \bar{X} \pm z \times SE_{\bar{x}}$$

Where:

$\bar{X}$: Sample mean

$SE_{\bar{x}}$: Standard error

z: the z-value equal to 1.96 for CI_{95} and 2.56 for CI_{99}.

Example of Calculating Confidence Interval

This example illustrates how to find the Confidence Interval with a 95% degree of confidence (CI_{95}) for the following data set.

Table 6.1 Data on the Level of Cholesterol Collected for 10 Patients

Observation unit	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Blood cholesterol level (mg/dl), x_i	168	258	228	230	247	156	172	165	210	264

1. Find the sample mean.

$\bar{X} = \frac{\sum x_i}{n}$ $\bar{X} = 210\ mg/dl$

2. Find the standard deviation

$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$ $s = 41.1\ mg/dl$

3. Find the standard error (SE)

$SE = \frac{s}{\sqrt{n}}$ $SE = \frac{41.1}{3{,}16} = 13.1\ mg/dl$

4. Find the lower and upper limits of the confidence interval CI_{95}

$\mathrm{CI}_{95} = \overline{X} \pm z \times SE$

$CI_{95} = 210 \pm 1.96 \times 13.1$

CI_{95}: from 183.5 (lower limit) to 236 (upper limit)

Interpretation of the results: If you repeat the experiment 100 times, only 5 times out of 100, the population mean (μ) will be lower or higher than the limits of the confidence interval. In 95 times out of 100, the population mean will be between the lower and upper limits of the confidence interval. For CI_{95} the probability to find the population mean within its limits is 0.95. The probability of finding the population mean out of these limits is 0.05.

Comparing Two Means Using Confidence Intervals

In medical research, comparing the means of different groups is a common task to determine if a treatment or condition has a significant effect. By comparing the confidence intervals (CIs) of different groups, we can conclude if the difference between the two means is meaningful, i.e., statistically significant, or not.

Figure 6.1 compares the pulse rates of three groups of patients: Group B and Group C are compared to the reference Group A. The pulse rates are represented with their means and 95% confidence intervals (error bars). The pulse rate for Group B is significantly higher than that of Group A since the 95% CI for Group B does not overlap with the 95% CI for Group A (p-value for t-test < 0.05). The difference in pulse rates between Group A and Group C is not statistically significant; the overlap of the CIs suggests that the observed difference could be due to random variation (p-value for t-test > 0.05).

Table 6.2 presents a framework for interpreting the overlap of confidence intervals and its corresponding implications for the statistical significance of the differences between groups.

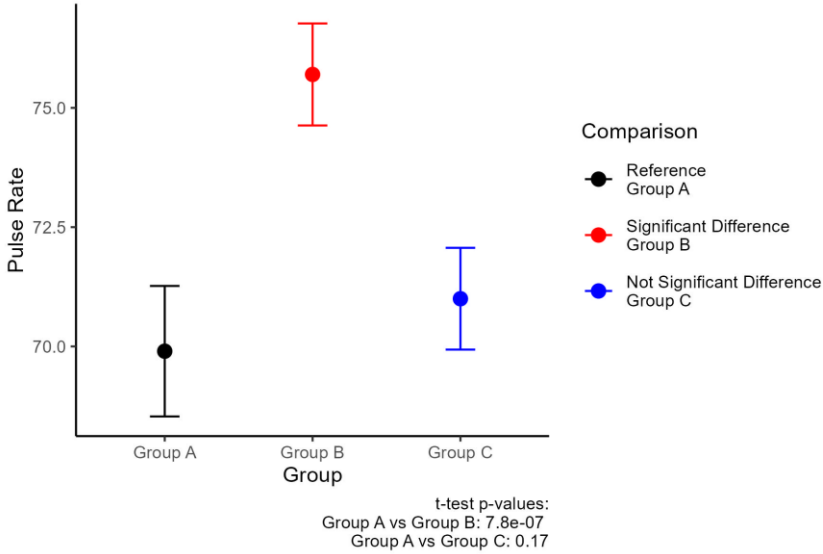

Figure 6.1 Comparison of Pulse Rates across Three Groups Using 95% Confidence Intervals

Note: Compared to Group A (reference group), the difference in the mean pulse rate is statistically significant only for Group B (no overlap between CIs), but not for Group C (overlap between CIs).

Table 6.2 Compare Two Means Using Confidence Intervals (CIs)

Overlap of CIs	Statistical significance between comparing groups
None	Highly significant difference
Slight	Possible significant but not highly significant
Large	Definitely not significant

6.5 Hypothesis Testing: Basic Theoretical Concepts

6.5.1 Hypothesis Definition

A hypothesis is an educated guess or assumption about a phenomenon. It focuses on research, bringing clarity, specificity and objectivity.

6.5.2 Hypotheses Types

⇒ *Null hypothesis (H_0)*: States that there is no significant difference or relationship between variables (e.g., two population means are equal).

$$H_0: \mu_1 = \mu_2$$

⇒ *Alternative hypothesis (H1):* Contradicts the null hypothesis, stating that there are differences between groups (e.g., two population means differ).

- *Non-directional or Two-tailed Alternative Hypothesis*: States a difference exists without specifying which value is larger, tested by a two-sided test.

$$H_1: \mu_1 \neq \mu_2$$

- *Directional or One-tailed Alternative Hypothesis*: States the expected direction of the difference, specifying which value is larger, tested by a one-sided test.

$$H_1: \mu_1 > \mu_2 \qquad \text{or} \qquad H_1: \mu_1 > \mu_2$$

Together, the null and alternative hypotheses cover all possible values of the population mean (μ): consequently, one of the two statements must be true.

6.5.3 Type I and Type II Errors

Two types of error may occur in hypothesis testing:

- ⇒ *Type I Error*: Rejecting the null hypothesis when it is true (false positive). The probability of making a Type I error (α) is also known as a *rejection error.*
- ⇒ *Type II Error:* Not rejecting the null hypothesis when it is false (false negative). The probability of making a Type 2 error (β) is also known as an *acceptance erro*r.

6.5.4 Power of the Study

The power of a study refers to its ability to detect a true difference when it exists. Specifically, it is the probability of rejecting the null hypothesis when it is false, thereby concluding that the alternative hypothesis is true when it is really true. This means the power of a study is the probability of avoiding a Type II error (β). Mathematically, it is expressed as:

$$Power\ of\ the\ study = 1 - \beta$$

Where: β is the probability of making a Type II error.

For a study to be considered acceptable, it is generally required to have a power of at least 0.8 (or a β of 0.2). In other words, a study should have at least an 80% chance of detecting a true difference if it exists. A study with less than 80% power is typically unacceptable.

One of the most practical and important ways to increase the power of a study is by increasing the sample size. Larger sample sizes reduce the standard error, making it easier to detect true differences or effects.

6.5.5 Confidence Level

The confidence level refers to the ability of a study not to detect a false difference. It is the probability to accept the null hypothesis when it is really true. The confidence level is defined as:

$$Confidence\ level = 1 - \alpha$$

Where: α is the probability of making a Type I error.

Table 6.3 summarizes the four possible outcomes of hypothesis testing. It is important to remember that we always test the null hypothesis, which in an unknown reality can be either true or false.

Table 6.3 Four Possible Outcomes of Testing the Null Hypothesis (H_0)

	Null is TRUE	**Null is FALSE**
ACCEPT	Correct Decision ***Confidence level*** 1-α	Type II Error β
REJECT	Type I Error α	Correct Decision ***Power of the study*** 1-β

6.5.6 Significance Level

Before we can reject the Null hypothesis, we need to be reasonably certain that any observed difference between the sample statistic ($\bar{X}$) and the population parameter (μ) is not due to chance. How different must a sample mean be from a population mean before considering the difference meaningful or statistically significant? This criterion is known as the *significance level* or α-level, which represents the probability of making a Type I error, i.e. the probability of rejecting the null hypothesis when it is actually true.

The significance level (α-level) is the probability of making a Type I error, set *before* calculating the statistical test. To minimize the chance of incorrectly rejecting the null hypothesis, the significance level should be sufficiently small. It is commonly set at 0.05 or lower (0.01 0.001). The smaller the α-level, the less likely it is to commit a Type I error.

6.5.7 p-value

The *p-value* is a concept closely related to the significance level (α-level). It represents the probability of obtaining the observed results if the null hypothesis is true. It is the probability of making a Type I error *after* calculating the statistical test.

After performing a statistical test, you compare the p-value to the α-level. If the p-value is less than the α-level (e.g., p-value < 0.05), you can reject the null hypothesis and accept the alternative hypothesis. This indicates that the difference between the two means is statistically significant and unlikely to be due to chance.

Conversely, if the p-value is greater than the α-level (e.g., p-value > 0.05), you must accept the null hypothesis and conclude that the difference between the two means is not statistically significant and is likely due to chance.

6.6 Key Steps in Hypothesis Testing

In hypothesis testing, we aim to draw conclusions about a population parameter using sample data. One approach is to construct a confidence interval for a population mean (μ); another is to conduct a statistical test. The general steps in the hypothesis testing process are:

1. Formulate the hypotheses: null hypothesis (H_0) and alternative hypothesis (H_1);
2. Select the appropriate statistical test;
3. Select the level of significance (α-level).
4. Determine the critical value that the statistical test must reach to be declared significant.
5. Perform the calculations.
6. State the conclusions.

The steps of hypothesis testing with examples will be discussed in detail in Chapter 7.

Review Exercises

1. What is the purpose of a test of hypothesis?
2. Briefly explain the relationship between confidence intervals and hypothesis testing.
3. Under what circumstances might you use a one-sided test of a hypothesis rather than a two-sided test?
4. Describe the two types of errors that can be made when you conduct a test of hypothesis.
5. The level of serum cholesterol of a sample of 400 adult men is skewed to the left distribution. The sampling distribution of the serum cholesterol means is:
    a) Skewed to the left
    b) Skewed to the right
    c) Normal
    d) Not possible to determine
6. The standard error of a statistic is:
    a) The mean of the sampling distribution
    b) The standard deviation of the sampling distribution
    c) The mean divided by the square root of the sample size (n)
7. The mean systolic blood pressure in a group of 500 individuals selected from a total of 100,000 individuals of locality A is 130 mmHg and the standard deviation is 15 mmHg. Calculate: the standard error and 95% confidence interval for the population mean. Interpret your results.
8. A study is conducted concerning the systolic blood pressure of 60-year-old men with diabetes mellitus. In this study, a random sample of 300 60-year-old men with diabetes mellitus was

selected and the mean systolic blood pressure was 160 mm Hg and the sample standard deviation was 25 mm Hg.

a) Calculate a 95% confidence interval for the mean systolic blood pressure among the population of 60-year-old men with diabetes mellitus.
b) Suppose that the sample size was 150 instead of 300, but the sample mean and the sample standard deviations are the same. Does the confidence interval get wider or narrower? Why?

Review Questions

1. What is the concept of inferential statistics?
2. What is the role of probability theory in statistics?
3. Describe the Law of Large Numbers and its relevance to research applications.
4. Explain the Central Limit Theorem.
5. What is sampling, and why is it essential in statistical analysis?
6. Why is random sampling crucial for ensuring the validity of research results?
7. In what scenarios might systematic sampling be preferred over simple random sampling?
8. When is stratified sampling the most appropriate method?
9. Under what circumstances is cluster sampling the most suitable approach?
10. What is the primary purpose of conducting hypothesis testing?
11. How do point estimation and interval estimation differ?
12. Define a hypothesis and explain the various types of hypotheses.

13. What are confidence intervals and confidence limits, and how are they used in statistical analysis?
14. What factors influence the width of a confidence interval for a mean?
15. Describe the two main types of errors that can occur in hypothesis testing.
16. Clarify the distinction between Type I and Type II errors in hypothesis testing.
17. What does the power of a study refer to, and how can it be described?
18. What is the confidence level, and how is it interpreted?
19. What does the significance level represent, and how can it be explained?
20. What is the p-value, and what does it signify in the context of statistical testing?

CHAPTER 7. HYPOTHESIS TESTING: PARAMETRIC AND NON-PARAMETRIC METHODS

Key Concepts

- *Parametric* statistical tests are used for numerical (interval or ratio) scale data, assuming that the data in the population are normally distributed.
- *t-tests* are examples of parametric statistical tests.
- *Non-parametric* statistical tests are used for categorical data (nominal or ordinal).
- The *chi-square* (χ^2) test is an example of a non-parametric statistical test. The chi-square test is used to test for differences in proportions and associations between two variables.
- Non-parametric statistical tests are less powerful than parametric tests.
- The type of hypothesis test—two-tailed, left-tailed or right-tailed—depends on the alternative hypothesis.
- If the absolute value of the test statistic is greater than the critical value, reject the null hypothesis. It means that the p-value is less than α-level.
- If the absolute value of the test statistic is less than the critical value, do not reject the null hypothesis. It means that the p-value is greater than the α-level.

7.1 Parametric and Non-Parametric Tests

Hypothesis testing methods are divided into parametric and non-parametric categories based on the type of data being analysed. Parametric methods are suitable for interval and ratio scale data, assuming that the data follow a normal distribution in a population. In contrast, non-parametric methods are used for ordinal or nominal scale data and do not require the assumption of a normal distribution.

Making an appropriate choice for statistical methods depends mainly on several key factors:

⇒ Type of data (numerical, nominal or ordinal);

⇒ Nature of samples (independent or related/pared);

⇒ Sample size (n>30 or n<30);

⇒ Number of groups (one, two or more);

⇒ Type of alternative hypothesis (directional or non-directional);

⇒ Data distribution type (normal or skewed);

⇒ Homogeneity of variance.

Some parametric and non-parametric tests are shown in *Figure 7.1*.

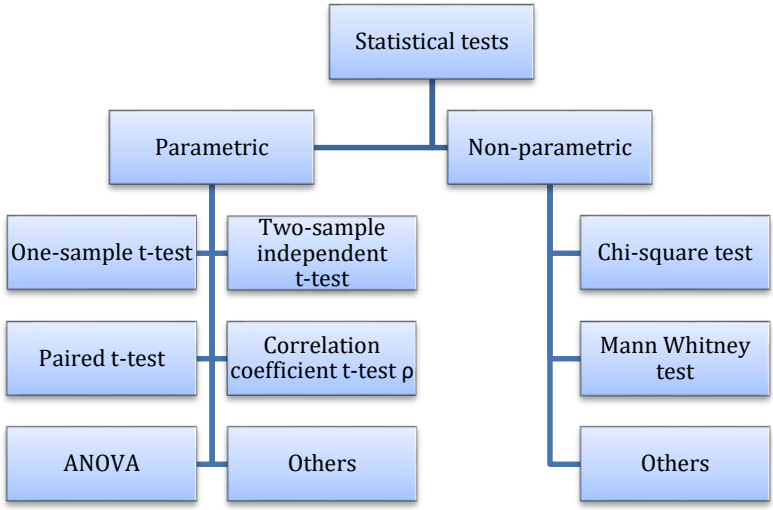

Figure 7.1 Parametric *vs* Non-Parametric Statistical Tests

7.2 General Approach to Hypothesis Testing

7.2.1 Steps of Hypothesis Testing

Although a large number of statistical tests exist, the steps of hypothesis testing remain the same.

Step 1. Formulate the Hypotheses

- Null Hypothesis (H_0): This is the hypothesis that there is no effect or no difference. It is the hypothesis that the researcher aims to reject.
- Alternative Hypothesis (H_1): This is the hypothesis that there is an effect or a difference. It is the hypothesis that the researcher aims to prove.

Step 2. Select the Appropriate Statistical Test

- The choice of the statistical test depends on the type of data you have and the design of the study. Common tests include t-tests, chi-square test, ANOVA, etc.
    - One-sample t-test: Used to compare the mean of a single sample to a known mean.
    - Two-sample Independent t-test: Used to compare the means of two independent groups.
    - Paired t-test: Used to compare means from the same group at different times.
    - Correlation Coefficient t-test ρ (rho): Used to assess the strength and direction of the relationship between two continuous variables.
    - Chis-square test: Used for categorical data to assess how likely it is that an observed distribution is due to chance.
    - ANOVA: used to compare means among three or more groups.

Step 3. Select the Level of Significance (α-level)

- The α-level represents the probability of rejecting the null hypothesis when it is true (Type I error).
- Commonly used α levels are 0.05, 0.01 and 0.001.

Step 4. Determine Degrees of Freedom and the Critical Value

- Degrees of Freedom (df): This depends on the statistical test and the sample size (n).
  - For one-sample t-test, $df = n - 1$.
  - For a two-sample independent t-test, $df = n_1 + n_2 - 2$.
  - For the Correlation Coefficient t-test ρ, $df = n - 2$.
  - For a chi-square test, *df* is typically calculated based on the number of categories.
- *Critical Value* (t-critical): This value is obtained from statistical tables (t-distribution, chi-square distribution, etc.) based on the α level and the degrees of freedom. The statistical tables can be found in Annex A (t-distribution table) and Annex B (chi-square table).

Step 5. Perform the Calculations

- Compute the test statistic using the appropriate formula for the chosen test.

Step 6. State the Conclusions

- Compare the calculated test statistic to the critical value.
  - If the absolute value of the test statistic (|t-calculated|) is greater than the critical value, reject the null hypothesis. It means that the p-value is less than α-level, indicating statistical significance.
  - If the absolute value of the test statistic (|t-calculated|) is less than the critical value, do not reject the null

hypothesis. It means that the p-value is greater than α-level, indicating the lack of statistical significance.

7.2.2 Hypothesis Testing: Two-tailed, Left-Tailed and Right-Tailed Tests

Hypothesis testing helps us determine whether a claim about a population parameter is true. Hypothesis testing involves stating a null hypothesis (H_0) and an alternative hypothesis (H_1). There are three types of hypothesis tests based on the nature of the alternative hypothesis:

⇒ Two-tailed test: The alternative hypothesis contains the "≠" sign.

⇒ Left-tailed test: The alternative hypothesis contains the "<" sign.

⇒ Right-tailed test: The alternative hypothesis contains the ">" sign.

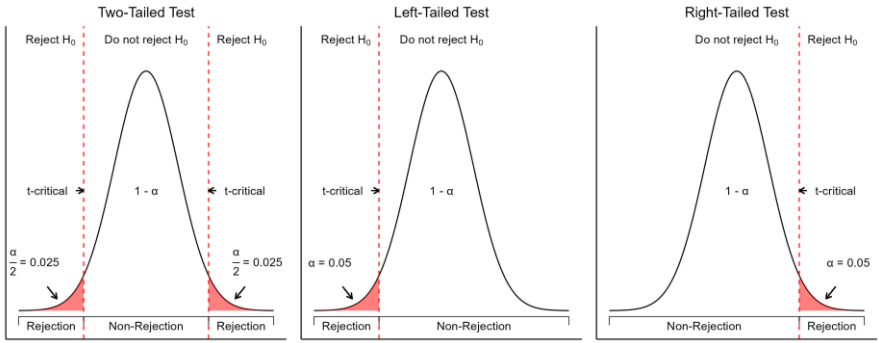

Figure 7.2 Illustration of Critical Values, Rejection and Non-Rejection Regions in Hypothesis Testing

Figure 7.2 illustrates two-tailed, left-tailed and right-tailed tests using a probability distribution plot represented by a bell-shaped normal curve. The area under the curve is equal to 1 or 100%. A critical t-value

(a dashed line in Figure 7.2) divides the area under the curve in rejection and non-rejection regions of the null hypothesis.

For a *two-tailed test*, the null hypothesis is rejected if the test statistic (t-calculated) is either too small or too large, creating two rejection regions: one on the left and one on the right. The null hypothesis is rejected if the test statistic is either lower or higher than the critical values.

For a *left-tailed test*, the null hypothesis is rejected if the test statistic (t-calculated) is too small, placing the rejection region on the left side of the distribution. The null hypothesis is rejected if the test statistic falls below the critical value on the left.

For a *right-tailed test*, the null hypothesis is rejected if the test statistic (t-calculated) is too large, placing the rejection region on the right side of the distribution. The null hypothesis is rejected if the test statistic exceeds the critical value on the right.

Using the critical value approach, we determine whether the calculated test statistic is more extreme than the critical value. The calculated test statistic is compared to the critical value, which serves as a cutoff point. If the test statistic is more extreme than the critical value, the null hypothesis is rejected. If the test statistic is not as extreme as the critical value, the null hypothesis is not rejected.

7.3 Parametric Tests

7.3.1. One-Sample t-Test

Definition: A one-sample t-test is used to determine whether the mean of a single sample is significantly different from a known or hypothesized population mean. This parametric test is useful when comparing the sample mean to a standard or expected value.

Conditions to Use One-Sample t-Test

- The data should be continuous (interval/ratio scale).
- The sample should be randomly selected.
- The sample data should be approximately normally distributed.
- The population standard deviation is unknown.

Example

Suppose, a sample of 15 medical students reports their hours of sleep during the final exam week. The data collected is as follows:

7, 6, 5, 6, 7, 8, 6, 7, 5, 6, 7, 8, 5, 6, 7

You want to test if the mean number of hours of sleep during the final exam week ($\bar{X}$) is significantly different from the recommended 8 hours (hypothesized population mean, μ).

Solution

1. **Formulate the Hypotheses:**

- Null hypothesis (H_0): $\mu = 8$ (The mean sleep duration is 8 hours)
- Alternative hypothesis (H_1): $\mu \neq 8$ (The mean sleep duration is not 8 hours). In this example, the alternative hypothesis is two-tailed. It means that the statistical test is two-tailed.

2. **Select the Appropriate Statistical Test:**

- Use a one-sample t-test since we are comparing the sample mean to a known population mean.

3. Select the Level of Significance (α-level):

- Choose $\propto = 0.05$.

4. Determine Degrees of Freedom (*df*) and the Critical Value:

- $df = 15 - 1 = 14$
- The critical t-value for a two-tailed test with df=14 and α=0.05 can be found using the t-distribution table (see Appendix A): t-critical = ±2.145.

5. Perform the Calculations:

- Calculate the sample mean ($\bar{X}$):

$$\bar{X} = \frac{\sum X_i}{n} = \frac{96}{15} = 6.4$$

- Calculate the sample standard deviation (s):

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{\sum(7-6.4)^2 + (6-6.4)^2 \ldots + (7-6.4)^2}{15-1}} \approx 1.12$$

- Calculate the t-test statistic using one-sample t-test formula:

$$t = \frac{\bar{X} - \mu}{SE_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{6.4 - 8}{1.12/\sqrt{15}} = \frac{-1.6}{0.289} \approx -5.54$$

6. State the Conclusions:

- Compare the calculated t-value to the critical t-value:
  - The calculated t-value is -5.54
  - The critical t-value is ±2.145.
- Conclusion: Based on our calculations, the test statistic was found to be -5.54. The absolute value of the test statistic (5.54) is greater than the absolute value of the critical value (2.145). Therefore, we reject the null hypothesis and conclude that the mean number of hours medical students sleep during their final exam week is significantly different from the recommended 8 hours (p-value < 0.05).

Illustration of Critical Values and Rejection Regions

In our example, the one-sample t-test is two-tailed, as indicated by the two critical values at ±2.145, corresponding to a significance level (α) of 0.05 and *df* of 14 (*Figure 7.3*). The shaded areas represent the rejection regions, where the null hypothesis would be rejected. The central unshaded region indicates where the null hypothesis would not be rejected. The calculated t-value (-5.54) lies within the rejection region (beyond the critical value of -2.145), leading to the rejection of the null hypothesis.

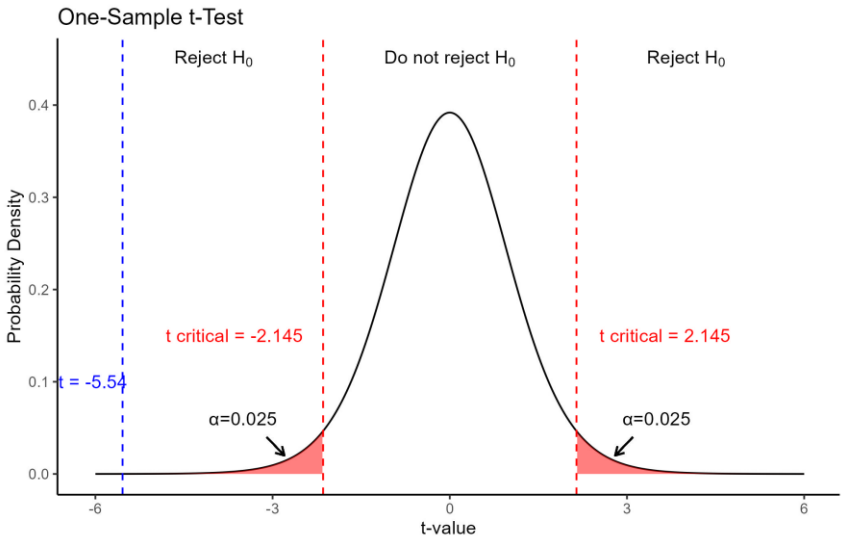

Figure 7.3 One-Sample t-Test: Visualization of Critical Values and Rejection Regions for a Two-Tailed Test

7.3.4 Two-Sample Independent t-Test

Definition: A two-sample independent t-test compares the means of two independent groups to determine whether there is statistical evidence that the associated population means are significantly different.

Conditions to Use Two-Sample Independent t-Test

- The two samples should be randomly selected
- The two samples must be independent of each other.
- The dependent variable should be measured on an interval or ratio scale.
- The data in each group should be approximately normally distributed.
- The test can be applied with either equal or unequal variances in the two groups. Different formulas are used depending on the equality of variances.

Example

A researcher wants to determine if there is a significant difference in the average recovery time (in days) between patients treated with Drug A and those treated with Drug B for a particular disease. The researcher collects a sample of recovery time from two independent groups of patients: one group treated with Drug A and the other group treated with Drug B.

Data

- Group A (Drug A):
  - Sample mean ($\bar{X}_A$): 8 days
  - Sample standard deviation (s_A): 2 days
  - Sample size (n_A): 31

- Group B (Drug B):
  - Sample mean ($\bar{X}_B$): 6 days
  - Sample standard deviation (s_B): 1.5 days
  - Sample size (n_B): 31

Solution

We will apply a right-tailed test, assuming unequal variances.

1. **Formulate the Hypotheses:**

- Null hypothesis (H_0): $\mu_A = \mu_B$. There is no difference in the average recovery time between patients treated with Drug A and Drug B. You can also write the H_0 as $\mu_A - \mu_B = 0$.
- Alternative hypothesis (H_1): $\mu_A > \mu_B$. The average recovery time for patients treated with Drug A is greater than the average recovery time for patients treated with Drug B. You can also write the H_1 as $\mu_A - \mu_B > 0$.

2. **Select the Appropriate Statistical Test:**

- We use a two-sample independent t-test with unequal variances (Welch's test).

3. **Select the Level of Significance (α-level):**

- Choose $\propto = 0.05$.

4. **Determine Degrees of Freedom (*df*) and the Critical Value:**

- $df = 31 + 31 - 2 = 60$
- The critical t-value for a one-tailed test with *df*=60 and α=0.05 can be found using the t-distribution table (see Appendix A). t-critical = 1.671.

5. **Perform the Calculations**

- Calculate the t-test statistic using a two-sample independent t-test formula (with unequal variances):

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{{s_A}^2}{n_A} + \frac{{s_B}^2}{n_B}}} = \frac{8 - 6}{\sqrt{\frac{2^2}{31} + \frac{1.5^5}{31}}} = \frac{2}{\sqrt{0.129 + 0.072}} \approx 4.46$$

6. State the Conclusions

- Compare the calculated t-value to the critical t-value:
  - The calculated t-value is 4.46
  - The critical t-value is 1.671.
- Conclusion: Based on our calculations, the test statistic was found to be 4.46. The absolute value of the test statistic is greater than the critical value of 1.671. Therefore, we reject the null hypothesis and conclude that there is significant evidence to suggest that the average recovery time for patients treated with Drug A is greater than with Drug B (p-value < 0.05).

Illustration of Critical Values and Rejection Regions

In our example, the calculated t-value (4.46) lies within the rejection region (it is more than the critical value of 1.671). This leads to the rejection of the null hypothesis. This two-sample independent t-test is a right-tailed test. The shaded area on the right represents the rejection region, where the null hypothesis would be rejected. The rest unshaded region indicates where the null hypothesis would not be rejected (*Figure 7.4*).

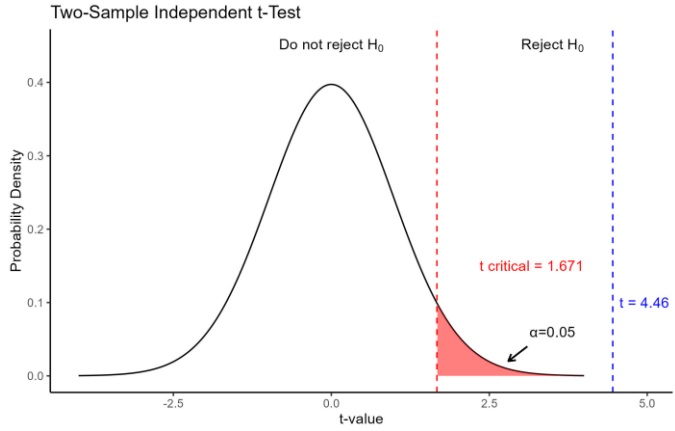

Figure 7.4 Two-Sample Independent t-Test: Visualization of Critical Value and Rejection Region for a Right-Tailed Test

7.3.5 Correlation Coefficient t-test

Definition: A t-test for the correlation coefficient ρ (rho) determines whether there is a statistically significant linear relationship between two continuous variables.

Conditions to Use a Correlation Coefficient Test

- The two variables should be continuous.
- The data should follow a bivariate distribution.
- The sample should be randomly selected.
- The relationship between the variables should be linear.

Example

A researcher is investigating the relationship between the number of weekly exercise hours undertaken by patients and their cholesterol levels. Data was gathered from 30 patients. The objective is to determine whether there is a statistically significant negative correlation between the duration of physical exercise and cholesterol

levels. Both exercise hours (per week) and cholesterol levels (mg/dL) were documented. The calculated correlation coefficient (r) is -0.15.

Solution

1. **Formulate the Hypotheses:**

- Null hypothesis (H_0): $\rho = 0$. There is no correlation between exercise hours and cholesterol levels.
- Alternative hypothesis (H_1): $\rho < 0$. There is a negative correlation between exercise hours and cholesterol levels.

2. **Select the Appropriate Statistical Test:**

- We use the Pearson correlation test.

3. **Select the Level of Significance (α-level):**

- Choose $\propto = 0.05$.

4. **Determine Degrees of Freedom (*df*) and the Critical Value:**

- $df = 30 - 2 = 28$
- The critical t-value for a one-tailed test with *df*=28 and α=0.05 can be found using the t-distribution table (see Appendix A). It is 1.701.

5. **Perform the Calculations**

- Given r = -0.15. Calculate the t-test statistic using the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t = \frac{-0.15\sqrt{30-2}}{\sqrt{1-(-0.15)^2}} = \frac{-0.15\sqrt{28}}{\sqrt{1-0.0225}} = \frac{-0.15 \times 5.2915}{0.9887} = \frac{-0.7937}{0.9887} = -0.803$$

6. **State the Conclusions**

- Compare the calculated t-value to the critical t-value:
  - The calculated t-value is -0.803
  - The critical t-value is 1.701.

- Conclusion: Based on our calculations, the test statistic was found to be -0.803. The absolute value of the test statistic (0.803) is less than the critical value of 1.701. Therefore, we do not reject the null hypothesis and conclude that there is no statistically significant negative correlation between exercise hours and cholesterol levels.

Illustration of Critical Values and Rejection Regions

In *Figure 7.5*, the left-tailed test for the correlation coefficient is illustrated. The critical t-value is marked with a dashed line at -1.701. If the calculated t-value falls within the shaded area, the null hypothesis is rejected. In this example, the calculated t-value of -0.803 lies in the non-rejection region, leading to the conclusion that there is no statistically significant negative correlation between exercise hours and cholesterol levels.

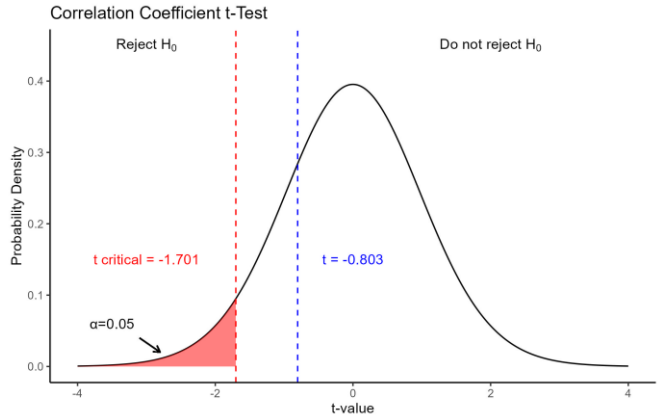

Figure 7.5 Correlation Coefficient t-Test: Visualization of Critical Values and Rejection Regions for a Left-Tailed Test

7.4 Non-parametric Tests

The non-parametric tests do not assume that the population is normally distributed, so they are called distribution-free tests. Non-parametric tests are used to test nominal, ordinal or skewed numerical data.

7.4.1 Chi-Square Test

Definition: The Chi-Square test is a statistical method used to determine if there is a significant association between two categorical variables.

Conditions to Use a Chi-Square Test

- The data must be presented in the form of counts or frequencies in a contingency 2x2 table.
- The categories should be mutually exclusive.
- The sample size should be sufficiently large (expected frequencies in each cell should be at least 5).

Example

A researcher is studying the relationship between smoking status (smoker vs. non-smoker) and the incidence of lung cancer (present vs. not present). They collect data from 100 patients, as presented in *Table 7.1*. It is necessary to determine whether there is a statistically significant association between smoking status and lung cancer.

Table 7.1 Smoking status and lung cancer frequencies in a 2x2 table (recorded frequencies)

	Lung Cancer Present	Lung Cancer Not Present	Total
Smoker	30	20	50
Non-smoker	10	40	50
Total	40	60	100

Question: Is there a statistically significant association between smoking status and lung cancer?

Solution

1. **Formulate the Hypotheses:**

- Null hypothesis (H_0): There is no association between smoking status and lung cancer.
- Alternative hypothesis (H_1): There is an association between smoking status and lung cancer.

2. **Select the Appropriate Statistical Test:**

- We use the Chis-Square test.

3. **Select the Level of Significance (α-level):**

- Choose $\propto = 0.05$.

4. **Determine Degrees of Freedom (*df*) and the Critical Value:**

- $df = (number\ of\ rows - 1) \times (number\ of\ columns - 1)$
- $df = (2-1) \times (2-1) = (2-1) \times (2-1) = 1$
- The critical value for *df*=1 and α=0.05 can be found using a Chis-Square distribution table (see Appendix B). It is 3.841.

5. **Perform the Calculations:**

- Calculate the expected frequencies for each cell:

$$E_{ij} = \frac{(Row\ Total) \times (Column\ Total)}{Grand\ Total}$$

For example, for the cell "Smoker & Lung Cancer Present":

$$E_{1,1} = \frac{50 \times 40}{100} = 20$$

In the same way, calculate the expected values in other cells as presented in Table 7.2.

Table 7.2 Smoking status and lung cancer frequencies in a 2x2 table (expected frequencies)

	Lung Cancer Present	Lung Cancer Not Present	Total
Smoker	20	30	50
Non-smoker	20	30	50
Total	40	60	100

- Calculate the Chi-Square test (χ^2) using the observed (O_{ij}) and expected frequencies (E_{ij}):

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi^2 = \frac{(30-20)^2}{20} + \frac{(20-30)^2}{30} + \frac{(10-20)^2}{20} + \frac{(40-30)^2}{30} = 16.66$$

6. State the Conclusions:

- Compare the calculated Chi-square value to the critical t-value:
  - The calculated Chi-square is 16.66.
  - The critical value is 3.841.
- Conclusion: The calculated Chi-square test (16.66) is greater than the critical value of 3.841. Therefore, we reject the null hypothesis and conclude that there is a statistically significant association between smoking status and lung cancer (p-value < 0.05).

Review Exercises

Using the following data sets:

1. Compute the mean for both groups.
2. Calculate variation measures and find out if the means are representative.
3. Compute the confidence interval for $\alpha = 0.05$.
4. Compare the means using a t-test and state conclusions about the statistical significance of the means difference.

Hypotheses:

H_0: The difference between the sample means is not statistically significant.

H_1: The difference between the sample means is statistically significant.

$p > 0.05 \Rightarrow H_0$ accepted

$p < 0.05 \Rightarrow H_0$ rejected

Data Set 1: Blood Cholesterol Levels Results for Two Independent Samples (n=15 each), mg/dl:

Nr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Group 1	168	258	228	247	156	172	165	210	264	220	258	200	195	245	189
Group 2	136	148	125	121	157	148	116	140	161	122	128	122	137	139	128

Data Set 2: Systolic Blood Pressure Results for Two Independent Samples (n=12 each), mmHg:

Nr.	**1**	**2**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Group 1	130	130	120	110	90	120	125	115	135	140	120
Group 2	170	175	160	170	170	190	185	185	170	160	190

Review Questions

1. What are the parametric methods of hypothesis testing?
2. What are the non-parametric methods of hypothesis testing?
3. When would you prefer to apply parametric methods of hypothesis testing?
4. When would you prefer to apply non-parametric methods of hypothesis testing?
5. State the main considerations of parametric methods vs. non-parametric methods.
6. When should you use the one-sample t-test?
7. When should you use the two-sample independent t-test?
8. When should you use the correlation coefficient t-test?
9. When should you use the Chi-Square test?

CHAPTER 8. INTRODUCTION TO THE RESEARCH METHODOLOGY

Key Concepts

- From the *point of view of application*, research can be classified as pure and applied.
- From the *point of view of the research methodology* used to answer a research question, research can be quantitative and qualitative.
- Study designs in medicine fall into two categories: studies in which subjects are observed (*observational*) and studies in which the effect of an intervention is observed (*experimental*).
- *Descriptive* study designs can only formulate a hypothesis about a possible relationship between a risk factor and an outcome.
- *Analytic* study designs can evaluate/analyse the relationship between a risk factor and a disease. They can test a hypothesis about causality.
- *Systematic Review* is a secondary study consisting of a summary of the clinical literature.
- *Bias* occurs when the way a study is designed or carried out causes an error in the results and conclusions. Bias can be due to the manner in which subjects are selected or data are collected and analysed.
- *Recall bias* is a systematic error and occurs when participants in a study are systematically more or less likely to recall information on exposure depending on their outcome status.

8.1 Research Definition, Characteristics and Types

8.1.1 Definition and Characteristics of Research

> *Research* is a systematic activity that employs appropriate scientific methodologies to address problems and generate new, broadly applicable knowledge.

Research involves the process of collecting, analysing and interpreting information to answer questions. It has the following *characteristics*:

1. *Objectivity:* Research must be free from personal and other biases. *Bias* can occur if a study is designed or conducted in a way that leads to errors in the results and conclusions. Bias can be caused by improper selection of subjects as well as by unsuitable methods of data collection and analysis.
2. *Validity:* Validity is synonymous with accuracy of research (accuracy of procedures, research instruments, tests, etc.). Two types of validity exist:
    - *Internal validity*: The results are valid for the sample of subjects who were actually studied.
    - *External validity*: The results are valid for the population from which the sample was drawn. If the results can be applied to other populations, the study is considered *generalizable*.
3. *Reliability:* Reliability, synonymous with *repeatability* and *reproducibility*, implies that findings can be reproduced and verified by the original researcher and others.
4. *Comparability:* The investigation process must be thorough and free from flaws, allowing research conclusions and results to withstand critical scrutiny and comparison.
5. *Systematic Approach:* All investigation procedures must follow a logical sequence. Each step should be conducted in an orderly

manner, ensuring coherence and consistency throughout the research process.

6. *Relevance:* There are two key types of relevance:

⇒ *Scientific Relevance*: The study advances our understanding of a specific scientific concept, process or disease, contributing to the broader body of scientific knowledge;

⇒ *Societal Relevance*: The study provides direct benefits to society, such as improving public health, informing policy decisions or enhancing quality of life through practical applications of the research findings.

8.1.2 Random Errors and Systematic Biases in Research

In research, errors and biases can significantly impact the validity and reliability of findings. Understanding the difference between random errors and systematic bias is essential for designing studies and interpreting results accurately.

⇒ **Random errors**

Random errors are unpredictable variations that occur during the measurement process. These errors arise from fluctuations in measurement instruments or other unpredictable factors. While random errors can affect the precision of results, they do not typically lead to consistent biases in one direction. Random errors do not affect heavily study results.

Example: Variability in blood pressure readings can occur due to slight differences in the position of the cuff or the subject's posture during measurements. Each reading might slightly differ, but on average, they should centre around the true value.

⇒ **Systematic Biases**

Systematic biases, also known as systematic errors, are consistent and repeatable inaccuracies that occur due to flaws in the study design,

data collection or analysis methods. These biases can lead to incorrect conclusions by consistently skewing results in a particular direction. Systematic biases affect heavily study results.

Generally, three types of systematic bias are distinguished:

1. **Confounding**: The bias occurs when the main risk factor being studied is mixed up with another variable, making it hard to isolate the true effect of the main risk factor.

Example: In a study investigating the relationship between coffee consumption and heart disease, if coffee drinkers also tend to smoke more, it becomes challenging to isolate the impact of coffee consumption alone on heart disease risk.

2. **Selection bias**: This is a systematic error that occurs when the participants or subjects included in a study are not representative of the target population, leading to skewed or biased results.

Examples:

- ✓ *Non-Response Bias*: if individuals who do not respond to a survey differ significantly from those who do, the results may not be representative of the entire population.
- ✓ *Exclusion Bias*: If certain groups are systematically excluded from the study, the findings may not be generalizable. For example, excluding elderly patients from a clinical trial may lead to results that are not applicable to older adults.
- ✓ *Sample Volume Bias*: If the sample size is too small or not adequately randomized, the results may not accurately reflect the broader population.

3. **Information bias:** This distortion occurs during data collection, often due to systematic measurement errors or misclassification of subjects.

Examples:

- ✓ *Interviewing Bias*: if interviewers are not sufficiently trained or if they unintentionally influence responses, the data collected may be biased.
- ✓ *Recall Bias*: In retrospective studies, participants may not accurately remember past events or exposures, leading to biased data. For example, patients with a disease may remember their exposure to risk factors differently than healthy individuals.
- ✓ *Reporting Bias*: If participants selectively reveal information, such as underreporting socially undesirable behaviours (e.g., smoking or alcohol consumption), the study's findings may be skewed.

8.1.3 Types of Research

Research can be classified into three primary perspectives, as outlined in *Table 8.1*:

- Application of the research
- Methodology of the research
- Objectives of the research

Table 8.1 Types of Research by Perspectives

Application of the research	***Methodology of the research***	***Objectives of the research***
1. pure	1. quantitative	1. historical
2. applied	2. qualitative	2. descriptive
		3. correlational
		4. experimental
		5. exploratory

Application of the Research

⇒ *Pure Research*: This type of research focuses on developing and testing theories and hypotheses that are intellectually challenging. While it may not have immediate practical applications, it advances theoretical knowledge.

⇒ *Applied Research*: This research is designed to address specific, practical questions. It aims to solve real-world problems.

Methodology of the Research

⇒ *Quantitative Research*: This approach quantifies the extent of a problem, issue, or phenomenon by measuring variables and statistical analysis. The central question is: "How many?"

⇒ *Qualitative Research*: This approach explores the nature of a problem, issue, or phenomenon without quantifying it. It focuses on understanding the "how" and "why" of a situation through descriptive data.

Both quantitative and qualitative approaches have unique strengths and are often complementary.

Objectives of the Research:

⇒ *Historical Research*: Aims to draw conclusions about past events, trends, causes, or effects. It often involves analysing primary sources or interviewing eyewitnesses. This type of research helps in understanding current events and predicting future trends.

⇒ *Descriptive Research*: Systematically describes a situation, problem, or phenomenon. It focuses on collecting data to answer questions about the current status of the subject. Methods include questionnaires, interviews, or observations to gather information about present conditions.

⇒ *Correlational Research*: Seeks to identify and measure relationships between two or more variables. It goes beyond description to

explore how variables are related and can be used to make predictions or test hypotheses. This type of research is often used to validate predictive tools and instruments.

⇒ *Experimental Research*: Establishes causality by actively manipulating variables and observing the effects. Unlike correlational research, experimental research involves intervention by the researcher to determine cause-and-effect relationships.

⇒ *Exploratory Research*: Conducted when little is known about a topic, it aims to investigate the feasibility of a study or identify areas for further research. Often used as a preliminary study or pilot study to explore new ideas and gather preliminary data.

8.2 The Steps of Research Process

In conducting research, two crucial decisions must be made:

1. What you want to discover.
2. How to go about discovering it.

To find answers to your research questions, you need to follow a series of practical steps. The approach you take to answer these questions is known as research methodology. At each step of the research process, you will select from a range of methods and techniques designed to best achieve your research objectives. For a systematic and effective research process, follow these steps:

1. **Define the Research Problem**: Clearly articulate the issue or question you want to explore. This involves identifying the research gap and specifying what you aim to investigate.
2. **Conduct a Literature Review**: Examine existing research and literature related to your topic. This helps to understand the current state of knowledge, refine your research problem and pinpoint gaps that your study may address.

3. **Formulate the Aim and Objectives**: Clearly define the primary aim of your research and break it down into specific, measurable objectives. These will guide the development of your research design and methodology.
4. **Develop the Research Design**: Create a detailed plan outlining the research methodology, including study design, sampling methods, data collection procedures, and analysis techniques.
5. **Collect Data**: Gather the necessary data using the chosen methods. This could involve surveys, interviews, experiments, or secondary data analysis, depending on your research design.
6. **Analyse Data**: Process and analyse the collected data. Use appropriate statistical techniques based on your research objectives and data type.
7. **Generalize and Interpret Results**: Draw conclusions based on your data analysis. Interpret the findings in relation to your research questions and provide recommendations based on these results.
8. **Present Findings**: Communicate your research findings through a well-structured report and, if applicable, an oral presentation.

8.3 Formulating the Research Problem

Formulating the research problem is the first and most critical step in the research process. A well-defined research problem provides direction and focus for the entire study. Consider the following factors when selecting a research problem:

⇒ *Interest*: Choose a topic that genuinely interests you. Passion for the topic will help maintain motivation throughout the research process.

⇒ *Magnitude*: Ensure the topic is manageable within the available time and resources.

⇒ *Relevance*: The research should contribute to the existing body of knowledge.

⇒ *Data Availability*: Verify that the necessary data can be accessed or collected.

⇒ *Ethical Considerations*: Ensure that the research adheres to ethical standards.

The process of formulating a research problem involves several key steps. Adequate knowledge of the broad subject area is essential to clearly and effectively define the research problem. Follow these steps to formulate a robust research problem:

1. **Identify a Broad Field of Interest**: Start by selecting a general area that aligns with your interests and expertise.
2. **Dissect the Broad Field into Sub-Areas**: Break down the broad field into more specific sub-areas to narrow down the focus.
3. **Select Sub-Areas of Interest**: Choose the sub-areas that are most intriguing and relevant to your research goals.
4. **Develop Research Questions**: Formulate specific questions that your research aims to answer. These questions should guide the direction of your study.
5. **Establish Hypotheses**: Based on your research questions, develop hypotheses that propose potential answers or explanations.
6. **Double check**: Reevaluate the research problem and its formulation. Ensure that the problem is clearly defined, feasible, and aligned with your research objectives.

8.4 Literature Review

The literature review is a crucial initial step in the research process, essential for understanding the existing body of knowledge in your area of interest. This step not only provides context for your research but also contributes significantly to refining and guiding every phase of your study.

Functions of a Literature Review:

1. Clarifies and Focuses the Research Problem
2. Improve Study Methodology
3. Broadens Knowledge
4. Contextualize Findings

Procedures for Conducting a Literature Review:

⇒ **Search for Existing Literature**: Begin by identifying and gathering relevant literature related to your research area. Use academic databases and libraries to locate sources.

⇒ **Review the Selected Literature**: Critically analyse and synthesize the literature to understand the current state of research on your topic.

⇒ **Develop a Theoretical Framework**: Construct a framework that outlines the theories and concepts that will guide your research.

⇒ **Develop a Practical Framework**: Design a practical framework that incorporates methodologies, techniques, and approaches relevant to your study.

⇒ **Organize and Document Sources**: Systematically catalog the literature you review.

Tools and Resources:

To effectively search for literature, start with a clear understanding of your research topic and set parameters for your search. Utilize the following online databases and resources:

- MedLine https://www.nlm.nih.gov/medline/medline_overview.html
- Research4Life https://portal.research4life.org/
- HINARI https://www.emro.who.int/information-resources/hinari/hinari.html
- PubMed https://pubmed.ncbi.nlm.nih.gov/
- And many others…

For an academic paper, you should use books and articles as well as Web sites that collect important information on your topic. In academic writing, it is important to use a variety of sources, including books, scholarly articles, and credible websites. During your review, take detailed notes on methodologies and instruments used in previous research. This will help you select appropriate approaches for your own study, such as sampling techniques, data collection methods, and analytical procedures. Ensure that you cite your sources using a standardized format. The bibliography of your literature review should be a complete and clear list of all the sources referenced, organized alphabetically by the author's surnames.

The three most commonly used citation styles are:

⇒ ***Harvard System*****:** An “author-date” citation style.

- In-Text Citation: (Smith, 2020)
- Reference List: Smith, J. (2020). *Understanding Modern Research*. Oxford University Press.

⇒ ***Vancouver System*:** A numerical citation style was introduced in Canada in 1978.

- In-Text Citation: (1)
- Reference List: Smith, J. *Understanding Modern Research*. Oxford University Press; 2020.

⇒ ***Letter-Number Systems*:** A hybrid citation system combining elements of both numerical and alphabetical systems.

- In-Text Citation: (Smith 2020a)
- Reference List: Smith J. *Understanding Modern Research*. Oxford University Press; 2020a.

The diagram below shows how information moves through different phases to become a part of the published body of knowledge. This knowledge is then available to researchers to build upon, inspire new areas of inquiry and create new knowledge.

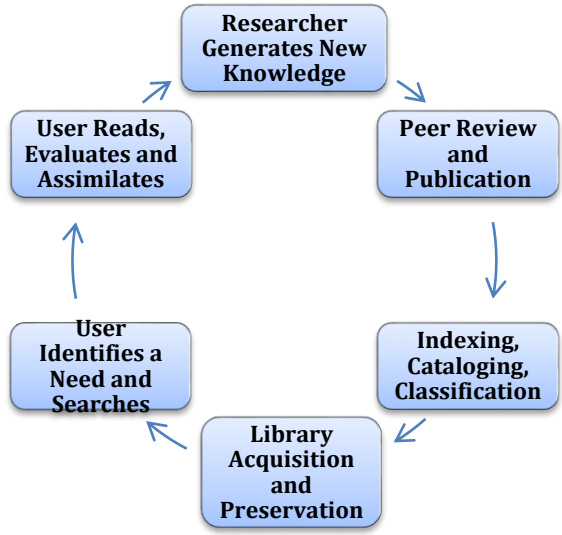

Figure 8.1 Information Cycle

8.5 Formulation of the Aim and Objectives of the Study

The *aim* is the goal that the study sets out to achieve, serving as an overall statement of the study's purpose. *Objectives* are the specific tasks required to accomplish this aim.

It is crucial to formulate these objectives clearly and specifically. Objectives should be numerically listed, with each objective addressing only one aspect of the study. When formulating objectives, it is preferable to use action-oriented words or verbs. Therefore, the objectives should start with words such as: to determine, to find out, to ascertain, to measure, to explore, etc.

8.6 Preparing the Research Design and Collecting the Data

8.6.1 Research Design Definition and Steps

> *Research design* is the conceptual framework within which research is conducted.

Its function is to ensure the collection of relevant information with minimal expenditure of effort, time, and money. Preparing an appropriate research design for a particular research problem involves the following steps:

- ⇒ To determine the *sample design*
- ⇒ To elaborate *tools* for data collection
- ⇒ To adopt a *study design*

8.6.3 Determining Sample Design

Researchers typically draw conclusions about a population by studying a sample. Designing the sample involves three key decisions:

1. Who will be surveyed? (*The sample*)
2. How many people will be surveyed? (*Sample size*)
3. How should the sample be chosen? (*Sampling type*)

8.6.4 Tool for Data Collection

Constructing a research tool for data collection is a critical aspect of a research protocol. Findings and conclusions are based on the collected data, which in turn depends on the questions posed in the questionnaire.

Guidelines for constructing a questionnaire:

Step I: Clearly define and list all specific objectives or research questions for the study.

Step II: For each objective or research question, list all associated questions that you want to answer through the study.

Step III: For each research question listed in Step I and each objective listed in Step II, determine the information required to answer them.

Step IV: Formulate questions to obtain this information.

A questionnaire consists of a set of questions presented for respondents to answer. There are many ways to ask questions, but it is important to ensure clarity and to obtain the desired information. The questionnaire should be developed and tested carefully before being used on a large scale.

Types of Question Structure:

- ⇒ Closed
- ⇒ Open-ended
- ⇒ Combination of both

Closed Questions: These include all possible answers in prewritten response categories, and respondents choose from among them (e.g., multiple-choice questions, scale questions). Closed questions are used to

generate statistics in quantitative research. Their main advantage is the ease with which answers can be analysed and reported.

Open-ended Questions: These permit respondents to answer in their own words. The primary advantage is the ability to capture respondents' thoughts using their own words. However, data analysis can be more complex.

Combination Questions: This structure starts with a series of closed questions and ends with open-ended questions for more detailed responses.

Table 8.2 Open versus Closed Questions

Criteria	Open Questions	Closed Questions
Purpose	Capture actual words or quotes	Most common answers
Respondents	Provide in-depth, detailed answers	Prefer quick and easy responses
Question Context	Choices are unknown	Choices can be anticipated
Analysis	Content analyses; time-consuming	Counting or scoring
Reporting	Individual or grouped responses	Statistical data

Source: Based on (Dawson and Trapp, 2004)

Most surveys use self-administrated questionnaires – in person or via mail, email or interviews – again in person or over the phone. Advantages and disadvantages exist for each method, some of which are illustrated in *Table 8.3.*

Table 8.3 Advantages and disadvantages of different survey methods

	Self-administered Mail/email	Self-administered in person	Interview by phone	Interview in person
Cost	++	+	-	-
Time	++	+	-	-
Standardization	+	+	+/-	+/-
Depth/detail	-	-	+	++
Response rate	-	++	+	++
Missing responses	-	+	++	++
+Advantages; - Disadvantages; +/- Neutral				

Source: (Dawson and Trapp, 2004)

General Structure of the Questionnaire:

⇒ Title

⇒ Instructions

⇒ General information about the respondent

⇒ The questions

⇒ Thank-you note

8.6.5 Study Design Classification

There are several schemes for classifying methods of study design. One such classification is shown in *Table 8.4*. This classification divides study designs into three categories: observational studies (where subjects are observed without intervention), experimental studies (where some intervention is performed) and secondary studies (where primary studies are pre-appraised or filtered).

Table 8.4 Classification of study design

DESIGN	METHOD	TYPES
I. Primary studies:		
– Observational studies	Descriptive	– Case series/report – Cross-sectional – Ecological *(population level)*
	Analytical	– Case-Control – Cohort
– Experimental studies	Analytical	– Clinical Trial – Community Trial *(population level)*
II. Secondary studies:		
– **Pre-appraised *or* filtered studies**	Quantitative Qualitative	– Systematic reviews – Narrative reviews – Meta-analyses

Study Design Overview

Study designs are methods for gathering information and assessing the relationship between risk factors (exposures or causes) and diseases (outcomes or effects). The primary goal is to establish causality, which is the relationship between exposure and outcome or cause and effect. The choice of study design is influenced by the research questions, concerns about validity, and practical or ethical considerations.

⇒ ***Observational studies***: These studies provide insights into exposures in natural settings without the ethical issues associated with experimental studies. They include:

- **Case-Series Studies**: Document a series of cases with a similar condition but do not include a comparison group.
- **Cross-Sectional Studies**: Assess data at a single point in time to provide a snapshot of the relationship between risk factors and outcomes.
- **Case-control Studies**: Compare individuals with a disease (cases) to those without it (controls) to identify differences in exposure.
- **Cohort Studies***:* Follow a group exposed to a risk factor over time to observe health outcomes.

Case-series and cross-sectional studies are descriptive; they hypothesize about relationships between risk factors and outcomes but do not establish causality. In contrast, case-control and cohort studies are analytical; they test hypotheses to establish causality. Cohort studies are longitudinal, collecting data on risk factors and outcomes at multiple points in time, whereas cross-sectional studies collect data at a single point. Cohort studies are prospective, tracking from risk factors to outcomes, while case-control studies are retrospective, tracing from

outcomes to risk factors. Cross-sectional studies do not have a time direction, as data is collected once (see *Figure 8.2)*.

⇒ ***Experimental Studies (Clinical Trials)****:* These involve interventions, such as drugs or treatments, and are designed to assess their effects.

⇒ ***Systematic Review****:* This secondary study critically assesses and summarizes the clinical literature on a specific topic using a structured methodology. Researchers systematically locate, assemble, and evaluate relevant primary studies based on predefined criteria. The review aims to answer a well-defined research question using both qualitative and quantitative methods.

⇒ ***Narrative Reviews****:* These reviews have a broader scope and may not follow rigorous criteria for article selection or evaluation. They often lack explicit criteria for inclusion and may not assess the validity of the studies reviewed, which can introduce bias.

⇒ ***Meta-analysis****:* A type of secondary study that quantitatively combines results from multiple primary studies to provide a comprehensive summary and conclusions. Meta-analyses often focus on evaluating therapeutic effectiveness or planning further research.

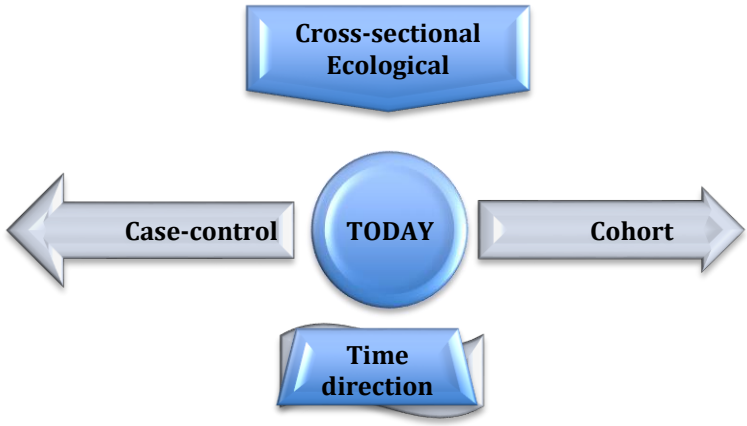

Figure 8.2 Time Relationship among Different Observational Study Designs

8.7 Study Design Evidence Strength

The evidence pyramid ranks study designs from the strongest to the weakest, with the highest-quality studies at the top and those providing weaker evidence at the bottom. As illustrated in *Figure 8.3*, studies that offer strong evidence are relatively scarce.

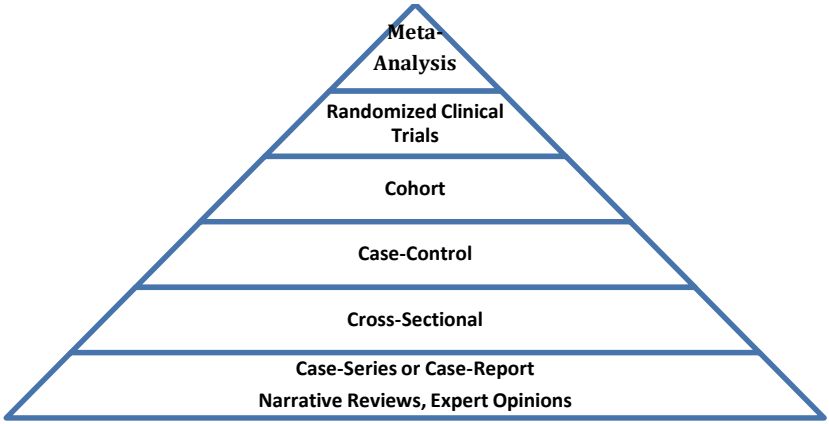

Figure 8.3 Research Study Design Evidence Strength Pyramid

Secondary studies, such as *Systematic Reviews* and *Meta-Analyses*, provide the highest internal validity. These studies synthesize findings from multiple primary studies to offer robust evidence of causality.

Experimental studies, such as randomized controlled trials (RCTs), are strong in establishing causal relationships through direct interventions. They are more controlled than observational studies but are less frequent due to practical and ethical challenges.

Cohort studies track groups exposed to risk factors over time to observe health outcomes, providing strong evidence of causality. *Case-control studies* compare individuals with a disease to those without it to identify exposure differences. While valuable, case-control studies are considered weaker than cohort studies because they are retrospective in nature.

Cross-sectional studies analyse data from a group of subjects at a single point in time, offering a snapshot of associations without determining causality. *Case-series studies, narrative reviews* and *expert opinions* are positioned at the bottom of the pyramid. These studies are less controlled, providing weaker evidence of causality.

Review Exercises

1. A clinic manager wants to survey a random sample of patients to learn how they view some recent changes made in the clinic operation. The manager has drafted a questionnaire and wants you to review it. One of the questions asks: "Do you agree that the new clinic hours are an improvement over the old ones?"
   - What advice will you give the manager about the wording of this question? Explain your choice.
2. Suppose you would like to know how far physicians are willing to travel to attend continuing education courses, assuming that some number of hours is required each year. In addition, you want to learn topics they would like to have included in future programs. How would you select the sample of physicians to include in your study survey? Explain your choice.
   a) All physicians who attended last year's programs
   b) All physicians who attend the two upcoming programs
   c) A random sample of physicians who attended last year's programs
   d) A random sample of physicians was obtained from a list maintained by the state medical society
   e) A random sample of physicians in each county was obtained from a list maintained by the county medical societies

Review Questions

1. The definition and characteristics of research.
2. The role of validity in the research process.
3. Characteristics of research: their meaning. Give an example for each.
4. Types of research classification.
5. Research types classification by application: types and their meaning.
6. Research types classification from the point of view of objectives: types and their meaning.
7. Steps in the research process: their contents and particularities.
8. Formulating the research problem: main function and selection considerations.
9. Steps in the formulation of a research problem: their contents and particularities.
10. Reviewing of literature: its functions, procedures and citing references systems.
11. The aim and objectives: definition and rules of formulating.
12. Research design definition and steps.
13. Steps of questionnaire construction.
14. Basic types of questions structure. Their contents.
15. Types of survey methods. Their contents.
16. Study design classification.
17. Observational versus experimental study design: the meaning and particularity. Advantages and disadvantages.
18. Which type of study design is best depending on research questions?
    a) Therapy question
    b) Diagnosis/screening

c) Prognosis
d) Occurrence
e) Causation

19. State the main difference between the following study design: observational descriptive and observational analytic.
20. State the main difference between the following study design: case-control and cohort study.
21. State the main difference between the following study design: case-series and case-control.
22. Classify study designs according to strong evidence.

CHAPTER 9. OBSERVATIONAL DESCRIPTIVE STUDIES

Key Concepts

❖ Case-series and cross-sectional studies are observational and descriptive studies.
❖ *Cross-sectional study* is called transversal, prevalence study and community survey.
❖ *A transversal* study collects information about exposure and outcome only once and at the same point in time.
❖ Each study design has specific *advantages* and *disadvantages*.

9.1 Case-series / Case-report study

A case-series or case-report study is the simplest study design in which the author describes some interesting or unusual observations that occurred in a small number of patients (case-series study) or even in a single patient with an unusual condition (case-report study). When certain characteristics of a group (or series) of patients (or cases) are described in a published report, it is called a *case-series study*. This type of study design often generates hypotheses that can further be investigated in case-control or cohort studies.

Used for:

⇒ Recognition of new diseases or outcomes.
⇒ Formulation of hypotheses.

Advantages:

1. Easy to write.
2. Observations can be extremely useful to other investigators.

Disadvantages:

1. Susceptible to many biases.
2. Not suitable for making conclusive decisions.

9.2 Cross-sectional study

A cross-sectional study, also known as a *transversal study*, analyses data collected from a group of subjects at one point in time rather than over a period. It is designed to determine "What is happening at present?". Subjects are selected, and data about exposure to risk factors and outcomes (disease) are gathered within a short period, as depicted in *Figure 9.1.*

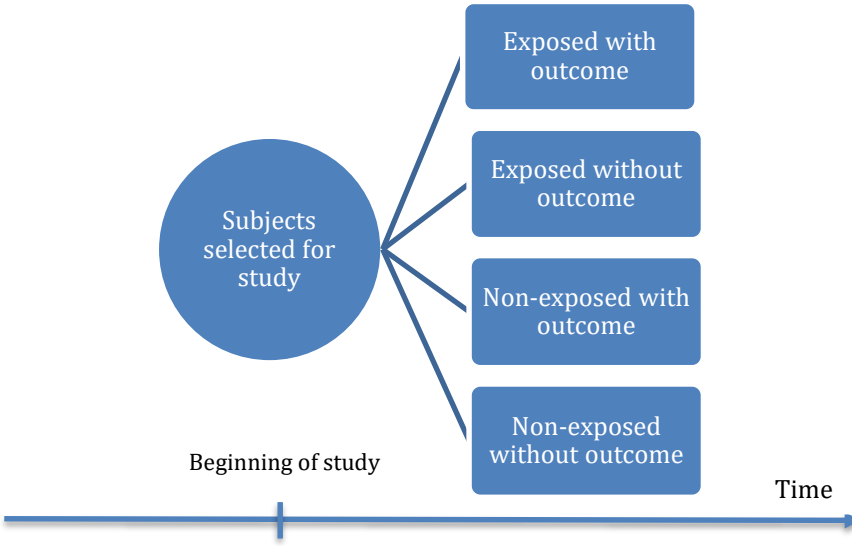

Figure 9.1 Flowchart of cross-sectional study design

A cross-sectional study is used to measure the prevalence of a disease and look at potential risk factors or causes. Because cross-sectional studies examine the relationship between exposure and disease prevalence in a defined population at a single point in time, they are called also *prevalence studies*. *Surveys* are generally cross-sectional

studies, although surveys can be a part of a cohort or case-control studies.

The cross-sectional study design is best to be used for diagnosis/screening, occurrence, surveys or establishing norm research questions.

Statistical Procedures for Cross-Sectional Study Data Analysis:

- Calculation of proportions, rates (adjusted as well) and ratios.
- Computation of confidence intervals for proportions or means.
- Correlation and regression analyses, including logistic regression.
- Application of parametric tests such as t-test, analysis of variance, chi-square and other non-parametric tests.

Advantages:

1. Useful for understanding the burden of a disease in a group – prevalence rate can be obtained.
2. Cheap and fast.
3. Useful for evaluating diagnostic procedures.
4. Helpful in studying common risk factors.
5. Useful in studying common outcomes.

Disadvantages:

1. Participants may be less willing to collaborate.
2. Does not indicate the sequence of events.
3. Shows an association between risk factors and the disease studied, not causality.
4. Not useful for identifying the causes of outcomes.
5. Mostly useful for studying chronic diseases.

6. Confounders may be unequally distributed.
7. Group sizes may be unequal.
8. Recall bias.

Review Questions

1. Define a case-series study and its contents.
2. What are the advantages of a case-series study?
3. What are the disadvantages of a case-series study?
4. Which statistical analysis is appropriate for a case-series study design?
5. Define a cross-sectional study and provide its synonyms.
6. Describe the content of the flowchart for a cross-sectional study.
7. What are the advantages of a cross-sectional study?
8. What are the disadvantages of a cross-sectional study?
9. Which statistical analysis is appropriate for a cross-sectional study design?

CHAPTER 10. OBSERVATIONAL ANALYTICAL STUDIES

Key Concepts

- A *case-control study* design includes a group of cases (those with an outcome) and a group of controls (those without an outcome).
- In a case-control study, the information about the exposure is collected *retrospectively*.
- Although a case-control study provides information about causality, *recall bias* is a common problem.
- *Matching* in a case-control study reduces the influence of confounding.
- The *odds ratio* shows how many times a case is more likely to have been exposed to a risk factor as compared to a control.
- A *cohort* is a group of people who do not have the disease of interest that is selected and then observed for an extended period.
- In a cohort study, the researcher will collect information about new cases of a disease (incidence) during regular examinations in the future (*prospective*).
- In a *longitudinal* study, two or more sets of observations are collected for every subject in the study over some time.
- A cohort study is called a *prospective*, *longitudinal*, *follow-up*, *incidence* study.
- *Relative Risk* shows how many times an exposed person is more likely to contract the disease as compared to an unexposed person.
- *Attributable risk* indicates the proportion of disease incidence that can be attributed to (or explained by) a specific exposure (among those who were exposed).
- In a case-control study, we know the outcome and look for the exposure in the past or retrospectively.

- ❖ In a cohort study, we know the exposure, and we follow up the subjects over some time looking for the outcome in question (in the future or prospectively).

10.1 Case-control study

A case-control study is a *retrospective* study design where information about risk factors is obtained by looking back at the history of participants. Although a case-control study can provide insights into causality, recall bias is a common issue.

Study Design

Cases: Individuals with a specific disease or outcome.

Controls: Individuals without the disease or outcome.

The investigators must use *matching* to associate controls with cases on characteristics such as age and sex. Both cases and controls should be similar except for their exposure to the risk factor under study, reducing the influence of potential confounding variables.

Participants are selected based on disease status (dependent variable), and both groups are asked about their exposure to potential risk factors (independent variable). Case-control studies are used for investigating the potential causes of diseases (causation research question).

Methods of Data Collection

- ⇒ Available records from the hospitals, vital statistics and other registers.
- ⇒ Interviews.
- ⇒ Self-administrated questionnaires.
- ⇒ Direct measurement.

The exposure histories of the cases and controls are then compared. Case-control studies seek to answer the question, “What happened?”. By

knowing the outcome, researchers look for the exposures in the past or retrospectively, as illustrated in *Figure 10.1.*

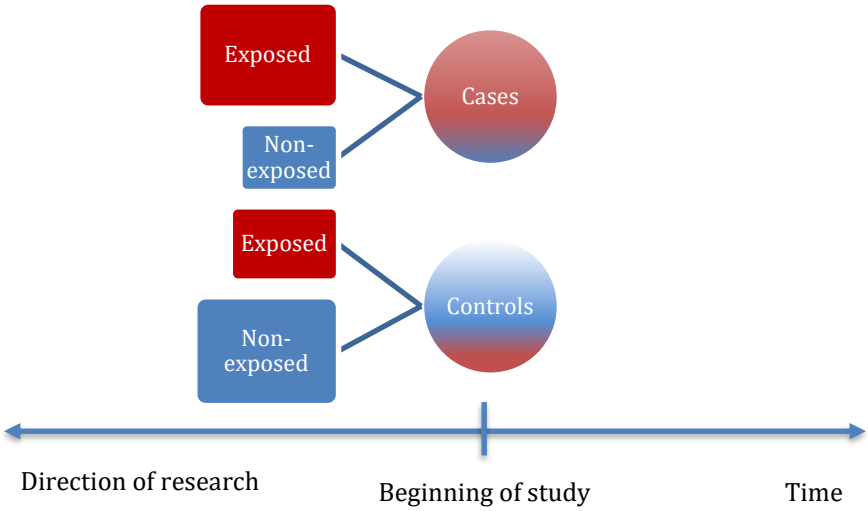

Figure 10.1 Flowchart of case-control study design

Data Analysis

The analysis of a case-control study involves calculating a measure of association known as the *Odds Ratio*. The odds are defined as the probability that an event will occur divided by the probability that the same event will not occur: p / (1-p). The odds ratio is calculated as follows:

$$Odds\ Ratio = \frac{\text{Odds that a case had been exposed to the risk factor}}{\text{Odds that a control had been exposed to the risk factor}}$$

The odds ratio provides a way to assess risk in case-control studies, showing *how many times a case is more likely to have been exposed to a risk factor as compared to a control*. It is easy to calculate when data

is presented in a 2x2 contingency table. A contingency table is a special type of frequency distribution table, where two variables are shown simultaneously (*Table 10.1*).

Table 10.1 2x2 Contingency Table for Case-Control Study

Exposure Factor	**Outcome (disease)**		**Total**
	Cases	Controls	
Exposed	a	b	a + b
Non-exposed	c	d	c + d
Total	a + c	b + d	a + b + c + d

$$\text{Odds Ratio (OR)} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

The odds ratio is also known as the cross-product ratio because it can be defined as the ratio of the product of the diagonals in a 2x2 table.

Interpretation of Odds Ratio

⇒ OR = 1: No association (no difference in exposure between cases and controls).

⇒ OR > 1: Hazardous exposure.

⇒ OR < 1: Beneficial exposure.

Test of significance

The odds ratio can be tested for significance using a confidence interval calculation, which is a range of values that is likely to contain a population OR with a certain level of confidence. If the 95% confidence interval for the OR includes 1, the results are not statistically significant.

Example Interpretation of OR

The way we would interpret the result of a statistically significant Odds Ratio is as follows:

> If OR=3.23, those with the disease are 3.23 times more likely to have had the exposure compared to those without the disease.

Advantages:

1. Allows examination of several risk factors.
2. Can study the long-term effects of exposure in a short period.
3. Requires fewer subjects.
4. Relatively quick and relatively cheap.
5. Suitable for rare diseases.

Disadvantages:

1. Higher chance of bias and confounders due to retrospective data collection.
2. Difficult to select an appropriate control group.
3. Recall bias can exist due to a retrospective nature.
4. Cannot determine incidence or prevalence.
5. Difficult to establish a temporal relationship between exposure and outcome.

10.2 Cohort study

A cohort study, also known as an i*ncidence study, longitudinal study or prospective study*, follows a group of individuals over an extended period to observe the development of new cases of a disease and associated risk factors.

Study Design

A *cohort* is a group of people who share a common characteristic but do not have the disease of interest at the study's start. They are observed over time to monitor the incidence of new cases. Researchers collect data on exposures and follow the cohort forward in time to observe new cases of a disease (incidence) during regular examinations in the future (*prospective* nature).

In a cohort study, the question asked is "What will happen?". Researchers know the exposure status at the beginning and follow the subjects over time to observe the outcomes, as depicted in *Figure 10.2.*

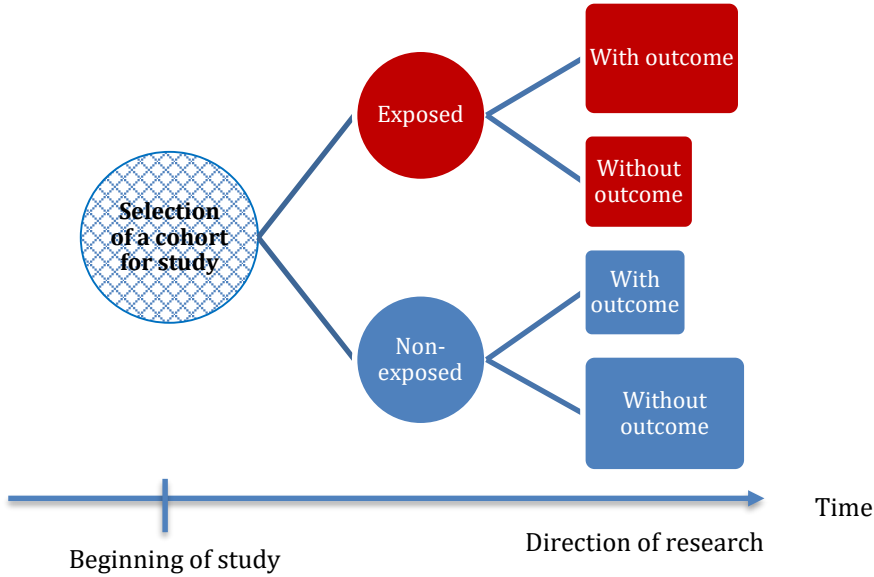

Figure 10.2 Flowchart of cohort study design

The typical cohort studies are usually *prospective* because health outcomes are observed after the beginning of the study. Cohort studies use groups that are similar in all respects except exposure.

Cohort studies are used for:

⇒ Measuring the incidence of disease.
⇒ Investigating the causes of diseases.
⇒ Determining prognosis.
⇒ Establishing the timing and directionality of events.

Methods of Data Collection

Obtaining data in a cohort study is possible through personal interviews, medical examinations and environmental surveys.

Data Analysis

The primary objective of the *analysis of cohort study* data is to compare the occurrence of outcomes in exposed and unexposed groups.

The following association measures are used to estimate the relationship between a risk factor and the occurrence of a given outcome:

$\Rightarrow$ Relative Risk (RR)

$\Rightarrow$ Attributable Risk (AR)

The *Relative Risk (RR)* is a ratio of the incidence rate in the exposed group to the incidence rate in the unexposed group. It is easy to calculate the RR for a cohort study when the data are arranged in the 2x2 table:

Table 10.2 2x2 Contingency Table for Cohort Study

Exposure factor	**Outcome (disease)**		**Total**
	Yes	No	
Exposed	a	b	a + b
Non-exposed	c	d	c + d
Total	a + c	b + d	a + b + c + d

$$Relative\ Risk\ (RR) = \frac{\text{Incidence of Exposed}}{\text{Incidence of Non} - \text{exposed}} = \frac{a\ /\ (a+b)}{c\ /\ (c+d)}$$

Relative risk indicates *how many times an exposed person is more likely to have an outcome as compared to an unexposed person*.

Interpretation of Relative Risk

⇒ RR = 1: No association (no difference in disease incidence between exposed and unexposed groups).

⇒ RR > 1: Hazardous exposure.

⇒ RR<1: Beneficial exposure.

Test of significance

Relative risk (RR) can be tested for significance using confidence intervals. If the 95% confidence interval for RR includes 1, the results are not statistically significant.

Example Interpretation of RR

> If RR = 3.23, individuals with the exposure are 3.23 times more likely to develop the disease compared to those without the exposure.

The *Attributable Risk (AR)* measures the proportion of disease incidence among the exposed that is due to the exposure. It is the ratio of the difference between the incidence of exposed and unexposed persons to the incidence of exposed persons represented in per cent.

$$Attributable\ Risk\ (AR) = \frac{\text{Incidence of Exposed} - \text{Incidence of Nonexposed}}{\text{Incidence of Exposed}}\ x\ 100\%$$

$$= \frac{(a\ /\ (a+b)) - \ (c\ /\ (c+d))}{a\ /\ (a+b)} \times\ 100\%$$

Example Interpretation of AR

> If AR = 80%, then 80% of the disease incidence among exposed can be attributed to the exposure.

Advantages

1. Ability to measure risk factors before disease occurs, providing evidence of causality.
2. Ability to study multiple disease outcomes.

3. Provides incidence rates and relative risk estimates.
4. Suitable for studying rare exposure.
5. Minimizes selection and information bias.

Disadvantages

1. Expensive
2. Inefficient for studying rare outcomes.
3. Requires a long follow-up period and/or a large population.
4. Losses to follow-up can affect the validity of findings.
5. Ineffective for rare diseases.
6. Ethical issues.

Review Exercises

Scenario 1: Cohort Study on Sun Radiation and Skin Cancer 2x2 Table

Exposure to Brutal Sun Radiation	Outcome (Skin Cancer)		Total
	Yes	No	
Yes	39	12	51
No	10	64	74
Total	49	76	125

1. According to these results try to restore the study scenarios in words.
2. Compute all possible measures of association.
3. Interpret the obtained results.

Scenario 2: Cohort Study on Sport Practice and Heart Ischemic Disease 2x2 Table

Exposure to Sports Practice	Outcome (Heart Ischemic Disease)		Total
	Yes	No	
Yes	1024	2376	3400
No	1205	604	1809
Total	2229	2980	5209

1. According to these results try to restore the study scenarios in words.
2. Compute all possible measures of association.
3. Interpret the obtained results.

Scenario 3: Case-Control Study on Diabetes and Myocardial Infarction 2x2 Table

Exposure to Diabetes	Outcome (Myocardial Infarction)		Total
	Yes	No	
Yes	60	40	100
No	340	360	700
Total	400	400	800

1. According to these results try to restore the study scenarios in words.
2. Compute all possible measures of association.
3. Interpret the obtained results.

Review Questions

1. Define a case-control study design and give its synonyms.
2. What are the main association measures used in a case-control study? Define them and explain their interpretation.
3. What are the advantages of a case-control study?
4. What are the disadvantages of a case-control study?
5. Define a cohort study design and give its synonyms.
6. What are the main association measures used in a cohort study? Define them and explain their interpretation.
7. What are the advantages of a cohort study?
8. What are the disadvantages of a cohort study?
9. What is the main difference between a case-control study and a cohort study? Provide an example.

CHAPTER 11. EXPERIMENTAL STUDIES

Key Concepts

- *Clinical trials* or experimental studies are divided into two main categories: controlled trials and uncontrolled trials.
- *Controlled Trials* compare an experimental treatment against a control (standard treatment, previous treatment, or placebo). They are considered more valid due to their ability to minimize bias and clearly determine the intervention's efficacy.
- *Uncontrolled Trials* lack a control group, making them less robust in terms of validity as they do not provide a comparative baseline.
- *Concurrent (Parallel) Controlled Trials* involve simultaneous testing of the experimental and control groups within the same study to ensure any differences are attributed to the intervention. *Blind trials* (single or double) are used to minimize bias.
- *Randomized Controlled Trials* (RCTs) are the gold standard for clinical trials, where participants are randomly assigned to groups, reducing selection bias and enhancing the reliability of the results.
- *Nonrandomized Controlled Trials* lack random assignment, which makes it difficult to ensure that group differences are solely due to the intervention, leading to weaker evidence.
- In *Crossover Trials*, patients receive both treatments (experimental and control) at different times, serving as their own controls, which minimizes variability.
- *Trials with External Controls* compare experimental results with historical data from other studies. While insightful, they are less reliable due to differences in study conditions and populations.

- *Statistical measures in Clinical Trials* include Experimental Event Rate (EER), Control Event Rate (CER), Relative Risk (RR), Absolute Risk Reduction (ARR), Relative Risk Reduction (RRR), and Number Needed to Treat (NNT).

11.1 Classification of Clinical Trials

Clinical trials are experimental studies involving humans, aimed at evaluating the effectiveness of medical procedures or treatments. They are essential for answering therapy-related research questions and can be broadly classified into two main categories: controlled trials and uncontrolled trials.

Classification of clinical trials

I. Controlled Trials
   - 1.1 Parallel or concurrent controls
     - *a) Randomized*
     - *b) Not randomized*
   - 1.2 Sequential Controls
     - *a) Self-Control*
     - *b) Crossover*
   - 1.3 External Controls

II. Uncontrolled Trials

Controlled trials compare the experimental drug or procedure against another treatment, which could be a standard treatment, a previously accepted treatment or a placebo.

Uncontrolled trials describe the outcomes of an experimental drug or procedure without comparing it to another treatment. These studies lack a control group and are generally considered less robust in terms of validity.

Controlled trials hold greater validity in medical research because they are designed to isolate the effects of the intervention, minimizing bias and allowing for a clearer determination of the intervention's true efficacy. Uncontrolled studies, while useful in some contexts, do not provide the same level of evidence due to the absence of a comparative baseline.

11.2 Controlled Trials with Concurrent (Parallel) Controls

A common method for conducting a controlled trial involves creating two groups of subjects: the experimental group, which receives the experimental procedure, and the control group, which receives the standard procedure or a placebo, as illustrated in *Figure 11.1.*

To ensure the validity of the results, it is crucial that the experimental and control groups are as similar as possible, so any observed differences can be attributed solely to the intervention. Providing concurrent control means that interventions for both groups are conducted simultaneously within the same study.

To minimize bias, researchers can design blind trials:

- *Single-Blind Trials:* Subjects do not know which intervention they are receiving.
- *Double-Blind Trials:* Neither the subjects nor the investigators know who is in the experimental or control group.

For ethical reasons, clinical trials are permitted only when the interventions are expected to be beneficial.

⇒ ***Randomized Controlled Trials*** **(RCTs)**

RCTs are the gold standard for determining causation, as they provide the strongest assurance that the observed results are due to the intervention alone.

In a randomized controlled trial (RCT), participants are randomly assigned to either the experimental group or the control group, ensuring that each individual has an equal chance of receiving any of the possible interventions. This randomization process helps eliminate selection bias, balancing both known and unknown confounding factors between the groups. Additionally, in blinded RCTs, neither the participants nor the researchers know which intervention is being administered, further reducing bias and enhancing the reliability of the results.

⇒ ***Nonrandomized Controlled Trials***

Nonrandomized controlled trials are studies that do not use random assignment to allocate subjects to the experimental or control groups. These trials, also known as clinical trials or comparative studies without randomization, are considered much weaker because they do not mitigate bias in patient assignment, making it difficult to ensure that differences between groups are solely due to the intervention.

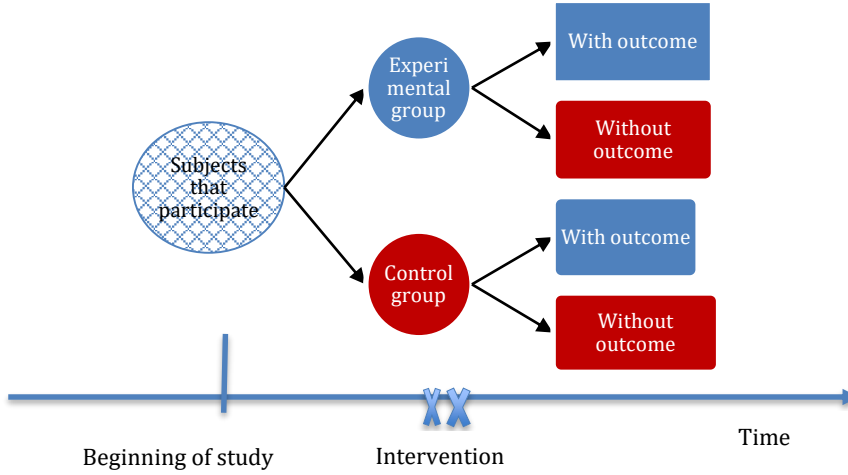

Figure 11.1 Flowchart of clinical trial with parallel controls

11.3 Sequential Controlled Trials

⇒ ***Self-Control Trials***

Self-control trials are studies where the same group of subjects serves as both the experimental and control group. This design allows for a moderate level of control by comparing the outcomes within the same group of individuals under different conditions.

⇒ ***Crossover Trials***

Crossover trial involves two groups of patients: one group initially receives the experimental treatment, while the other group receives the placebo or control treatment. After a specified period, both treatments are withdrawn for a *"washout"* period, during which no treatments are administered to ensure that the effects of the initial treatments have subsided. Following the washout period, the groups switch treatments: the first group now receives the placebo, and the second group receives the experimental treatment (*Figure 11.2*).

The crossover trial, known as a *within-subjects design*, is particularly powerful when used appropriately. In this study design, each patient receives both the active treatment and the placebo at different times, allowing for direct comparisons within the same individual. Consequently, each patient serves as their own control, enhancing the reliability and validity of the results by minimizing the variability between different subjects.

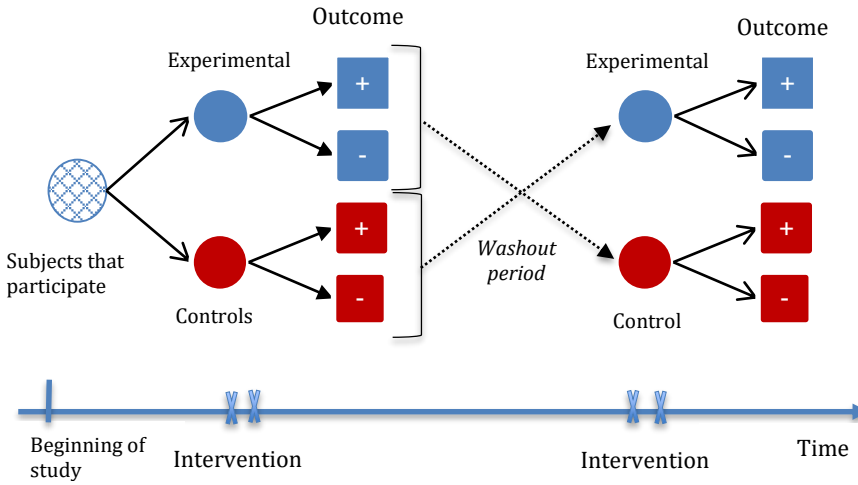

Figure 11.2 Flowchart of clinical trial with crossover sequential controls

11.4 Trials with External Controls

Controlled trials with *external controls* involve comparing the results of an experimental treatment with data from other studies or previously collected data, as shown in *Figure 11.3*. These comparisons, also known as historical controls, allow researchers to evaluate the effectiveness of an intervention by using existing data as a benchmark. While this method can provide valuable insights, it is generally considered less reliable than trials with concurrent controls due to potential differences in study conditions and populations.

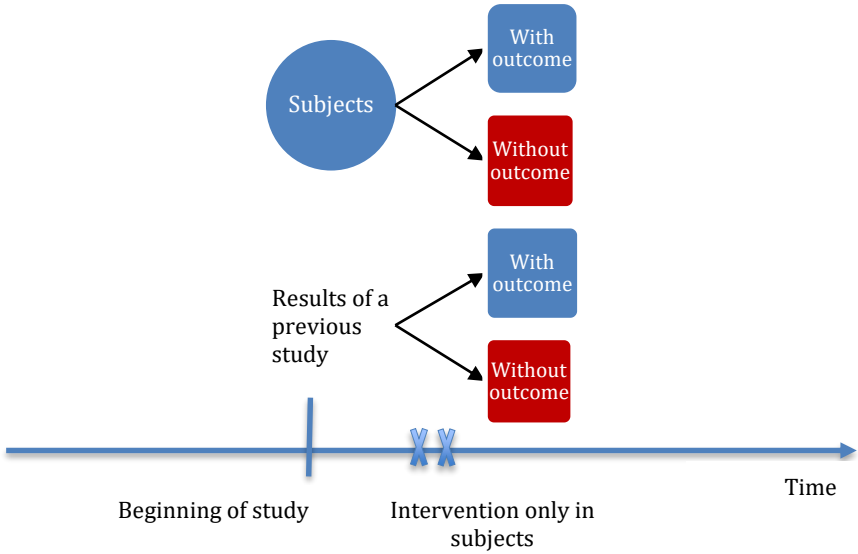

Figure 11.3 Flowchart of clinical trial with external controls

11.5 Statistical Analysis of Clinical Trials

Statistical analysis in clinical trials involves calculating various measures to assess the effectiveness of interventions. These measures include:

- Experimental Event Rate (EER)
- Control Event Rate (CER)
- Relative Risk (RR)
- Absolute Risk Reduction (ARR)
- Relative Risk Reduction (RRR)
- Number Needed to Treat (NNT)

To calculate these association measures, the clinical trial results are typically arranged in a 2x2 table:

Table 11.1 2x2 Table for Clinical Trial Study

Exposure factor (intervention)	**Outcome *(bad outcome)***		**Total**
	Yes	No	
Exposed: Experimental treatment	a	b	a + b
Nonexposed: Placebo	c	d	c + d
Total	a + c	b + d	a + b + c + d

⇒ ***Experimental Event Rate (EER)***: The rate of the event (risk of a bad outcome, such as a disease or death) in the group receiving the experimental treatment:

$$EER = \frac{a}{a+b}$$

⇒ ***Control Event Rate (CER)***: The rate of the event in the group receiving the placebo or control treatment:

$$CER = \frac{c}{c+d}$$

⇒ ***Relative Risk (RR)***: The ratio of the risk of disease in the experimental group to the risk of disease in the control group. It indicates how much more likely the exposed group is to experience the outcome compared to the unexposed group.

$$RR = \frac{EER}{CER}$$

Interpretation of RR:

> RR = 1: The intervention has no effect.
> RR > 1: The intervention is a risk factor.
> RR < 1: The intervention is a beneficial/protective factor.

⇒ ***Absolute Risk Reduction (ARR)***: Measures the reduction in the risk due to the intervention compared to the baseline risk and indicates how many subjects avoid the event for every 100 subjects treated. ARR is calculated as the difference between the Control Event Rate (CER) and the Experimental Event rate (EER):

$$ARR = CER - EER$$

Example: In a study, if the EER for cardiovascular disease (CVD) in the aspirin group is 0.14 and the CER in the control group is 0.30, then:

$$ARR = 0.30 - 0.14 = 0.16$$

Interpretation of ARR:

> The risk of CVD is 14 subjects per 100 in the group taking aspirin and 30 per 100 in the group taking placebo. Thus, 16 persons per 100 people avoid CVD by taking aspirin.

⇒ ***Relative Risk Reduction (RRR)***: Indicates the proportion of risk reduction relative to the baseline risk:

$$RRR = \frac{CER - EER}{CER} = \frac{ARR}{CER}$$

Example: Using the previous aspirin study,

$$RRR = \frac{0.16}{0.30} = 0.53 \text{ or } 53\%$$

Interpretation of RRR:

> Relative risk reduction is 53%. This means that relative to the baseline risk of 30 CVD cases per 100 people, aspirin reduces the risk by 53%.

⇒ ***Number Needed to Treat (NNT)***: Indicates the number of patients who need to be treated to prevent one adverse event.

$$NNT = \frac{1}{ARR}$$

Example: For the aspirin study, NNT is calculated as follows:

$$NNT = \frac{1}{0.16} = 6.25$$

Interpretation of NNT:

> Thus, approximately 6.25 patients need to be treated with aspirin to prevent one case of CVD. This information helps clinicians weigh the benefits and risks of a treatment.

Advantages:

1. Provide strong evidence for causality.
2. Reduce bias.
3. Allow to use historical controls for preliminary studies.

Disadvantages:

1. Expensive.
2. Ethical issues.
3. Requires time.
4. Relies on participant compliance.

Review Exercises

1. The results of a randomized controlled trial are arranged in the following 2x2 table:

Exposure Factor (Intervention)	**Outcome**		**Total**
	No Cholesterol decreasing	Cholesterol decreasing by 20 mg/dl	
Exposed: Experimental treatment: A drug-inhibitor of the enzyme 3-hydroxy-3-methylglutaryl coenzyme A reductase	3	43	46
Nonexposed: Placebo	8	39	47
Total	11	82	93

1.1. According to these results, describe the study scenarios in words.

1.2. Compute all possible measures of association.

1.3. Interpret the obtained results.

2. Select a recent study from a peer-reviewed journal that you find interesting.

2.1. Identify the main research question it addresses.

2.2. Determine which study design (e.g., randomized controlled trial, cohort study) would be best for answering the research question.

2.3. Verify if the study used the design you considered optimal. If not, look into why the authors chose their design.

Review Questions

1. What is the definition and classification of clinical trials?
2. What distinguishes a controlled clinical trial from an uncontrolled clinical trial?
3. What defines a randomized clinical trial?
4. What are controlled clinical trials with concurrent controls? Provide definitions, types, and a flowchart of the appropriate design.
5. What are controlled clinical trials with sequential controls? Provide definitions, types, and a flowchart of the appropriate design.
6. What are clinical trials with historical controls? Provide a definition and a flowchart of the appropriate design.
7. What are the main measures of association in clinical trial analysis? Define each measure and explain its interpretation.

CHAPTER 12. REPORTING THE RESEARCH FINDINGS: GENERAL APPROACHES

12.1 Writing a Research Report

Writing the research report represents the final step in the research process. This report communicates the objectives of your study, the methods used, the findings obtained, and the conclusions drawn. It should be written in a formal academic style, avoiding colloquial or journalistic language.

A traditional research report typically follows this format:

A. *Title page*
   - Title of the research project
   - Name of the researcher
   - Name of the institution
   - Date of publication

B. *Project Body*
   - **Introduction**: Provides background information and the research question or hypothesis.
   - **Literature Review**: Summarizes existing research relevant to the study.
   - **Material and Methods**: Describes the research design, procedures, and materials used.
   - **Results**: Presents the data analysis and interpretation of findings.
   - **Discussions**: Summarize the findings, explain their implications, and relate them to the literature.
   - **Conclusions**: Draw final conclusions based on the results and discussion.

- **Recommendations**: Suggests potential applications or further research based on the findings.
- **References / Bibliography**: Lists all sources cited in the report.
- **Appendices**: Includes supplementary material such as raw data, questionnaires, or detailed calculations.

12.2 Public Presentation of Medical Research Results

The purpose of an oral presentation is to effectively communicate the scientific findings of your medical research to an audience.

General Design of an Oral Presentation:

- **Title Slide**: Title of the presentation and names of the authors.
- **Introduction**: 1-2 slides providing background and context.
- **Aim and Objectives**: 1-2 slides outlining the research goals.
- **Material and Methods**: 1-2 slides describing the study design and methodology.
- **Results**: 2-3 slides highlighting the most important findings.
- **Discussion**: 1 slide summarizing key interpretations and implications.
- **Conclusions**: 1 slide with the main conclusions drawn from the study.
- **Closing**: 1 slide for final thoughts and acknowledgements.

Presentation Tips:

- **Duration:** Aim for a 10-minute presentation with 8-10 slides. Generally, allocate about one minute per slide.
- **Graphic Presentation:** Use figures and tables to illustrate your data, as they are often more effective than text. Ensure that all visuals are clear and relevant to your message.

Practical Recommendations:

- **Title Slide:** Use a single line with bold or contrasting colours for visibility.
- **Text:** Ensure text is legible from the back of the room. Limit text to 5-7 lines per slide, following the 7x7 rule: no more than 7 lines and 7 words per line.
- **Figures and Tables:** Maintain clarity and simplicity. Tables should have no more than 3-4 columns and 5-7 rows to avoid clutter.

12.3 Structure and Principles of Graduation Paper Development at Nicolae Testemitanu SUMPh

The graduation paper must demonstrate the student's ability to effectively engage with literature pertinent to their subject. It should be methodologically sound, with thorough data analysis and interpretation, and should follow a logical structure. Additionally, the paper must be written in scientific language, adhering to academic standards and the scientific writing guidelines set by Nicolae Testemitanu State University of Medicine and Pharmacy. Compliance with these standards ensures the paper meets the university's regulations for the development and defence of graduation papers. For detailed guidelines, refer to the university's official documentation at www.usmf.md.

CHAPTER 13. RESEARCH ETHICS INTRODUCTION

13.1 Research Ethics Definition and Objectives

Ethics are the set of rules that manage our expectations of our own and others' behaviour.

> *Research ethics* are the set of ethical guidelines that direct how scientific research should be conducted and disseminated.

Objectives of Research Ethics:

1. To protect participants, ensuring their dignity, rights and welfare.
2. To ensure that research is conducted in a manner that assists the welfare of persons, groups and society as a whole.
3. To evaluate specific research activities for their ethical integrity.

13.2 Research Ethics Principles

Research ethics are grounded in three fundamental approaches:

⇒ **Respect for persons**: Acknowledging the autonomy and rights of individuals.

⇒ **Beneficence**: Maximizing benefits and minimizing harm to participants.

⇒ **Justice**: Ensuring a fair distribution of the benefits and burdens of research.

There are five main principles of research ethics based on the fundamental approaches:

1. **Minimizing the risk of harm**: Taking steps to prevent harm to participants.
2. **Obtaining informed consent**: Ensuring participants are fully informed about the research and voluntarily agree to participate.

3. **Protecting anonymity and confidentiality**: Safeguarding participants' private information.
4. **Avoiding deceptive practices**: Being honest and transparent with participants.
5. **Providing the right to withdraw**: Allowing participants to leave the study at any time without penalty.

Key Tips for Ensuring Ethical Research:

- Collect facts and discuss intellectual property openly.
- Identify and address ethical issues.
- Recognize stakeholders and consider their interests.
- Determine potential sacrifices and risks.
- Acknowledge responsibilities (principles, rights, justice).
- Reflect on personal integrity and honesty.
- Think creatively about possible actions.
- Respect privacy and confidentiality.
- Decide on the appropriate ethical action and be prepared to handle differing viewpoints.

CHAPTER 14. PREVENTING PLAGIARISM: KEY PRINCIPLES

14.1 Plagiarism Meaning and Types

Plagiarism means using someone else's work or ideas without giving them full academic credit by citations, including in reference lists, acknowledgements, etc.

Types of Plagiarism:

⇒ **Direct Plagiarism**: Using word-for-word transcription of someone else's work without citation and quotation marks. There are many types of direct plagiarism:

- Global plagiarism: Using the entire text as your own.
- Paraphrasing Plagiarism: Reformulation of someone else's idea to present it as own.
- Patchwork Plagiarism: Stitching together parts of different works to elaborate your own.

⇒ **Self-plagiarism**: Reusing a work already published or submitted in an academic context. If you want to include any text, ideas, or data that you already submitted, be sure to inform by citing yourself.

⇒ **Accidental plagiarism**: Unintentional plagiarism is the accidental appropriation of the ideas and materials of others due to a lack of understanding of the conventions of citation and documentation. Even if unintentional, *it is still plagiarism and not acceptable.*

14.2 Preventing Plagiarism Techniques

There are a few simple approaches to consider to avoid plagiarism in your scientific writing:

⇒ Be confident in understanding what Plagiarism is;

⇒ Provide your own ideas by finding something new to say;

⇒ Use quotes to underline that you are using others' ideas;

- ⇒ Give full academic credit to all sources you use by correct citations and including them in the reference list;
- ⇒ Be careful with paraphrasing: When paraphrasing you have to write it in your own words and cannot just take out one word and replace it, even so, you still have to give appropriate academic credit;
- ⇒ Cite yourself as well;
- ⇒ Use plagiarism detection software.

BIBLIOGRAPHY

BERRY G., MATTHEWS JNS, ARMITAGE P. Statistical Methods in Medical Research, 4th Edition, Blackwell Scientific, 2001.

COLTON T. *Statistics in Medicine*, Little, Brown, 1974.

COMSTOCK G. *Research Ethics: A Philosophical Guide to the Responsible Conduct of Research,* 1st Edition. Cambridge University Press, 2013.

DANIEL W.W. Biostatistics: *A Foundation for Analysis in the Heal.th Sciences*, 7th ed. Wiley, 1998

DAWSON B., TRAPP G. R. *Basic and Clinical Biostatistics,* Fourth Edition, McGraw-Hill Companies, Inc., USA, 2004.

FEINSTEIN A.R. *Clinical Epidemiology: The Architecture of Research*, WB Saunders, 1985.

FISHER LD, VAN BELLE G. *Biostatistics: A Methodology for Health Sciences,* Wiley,1996.

FLEISS JL. *Design and Analysis of Clinical Experiments*, Wiley, 1999.

FLEISS JL. *Statistical Methods for Rates and Proportion*, 2nd Edition, Wiley, 1981.

GLANTZ, STANTON A. *Primer of Biostatistics,* University of California. 4th Edition, McGraw-Hill, Inc, 1994: *перевод на русский язык, Издательский дом »Практика», 1999.*

GLASER, ANTONY N. *High-Yield Biostatistics,* Medical University of South Carolina. 4th Edition, Lippincott Williams & Wilkins, a Wolters Kluwer, Philadelphia, 2014.

GREENBERG RS. *Prospective studies.* In Kotz S, Johnson NL (editors): *Encyclopedia of Statistics Sciences*, Vol. 7, pp.315-319. Wiley, 1986.

GREENBERG RS. *Retrospective studies.* In Kotz S, Johnson NL (editors): *Encyclopedia of Statistics Sciences*, Vol. 8, pp.120-124. Wiley, 1988.

HENNESEY DESENA L. *Preventing plagiarism. Tips and Techniques,* National Council of Teachers of English, 2007

HULLEY SB (ED), CUMMINGS SR, BROWNER WS ET AL. *Designing Clinical Research*, 2nd Edition Lippincott Williams and Wilkins, 2001.

INGELFINGER JA, WARE JH, THIBODEAU LA. *Biostatistics in Clinical Medicine*, 3rd Edition, Macmillian, 1994.

KANE RL. *Understanding Health Care Outcomes Research*, Aspen Publishers, 1997.

KRUGER RA, CASEY MA. *Focus Groups: A Practical Guide for Applied Research*. Sage, 2000.

LANDRIVON G., DELAHAYE F. *La Recherche Clinique. De l'idee a la publication*. RECIF. Masson, Paris, 1995: traducere limba română Edit DAN, 2002.

NAGESVARO RAO G. *Biostatistics and Research Methodology,* PharmaMed Press, 2018.

PAGANO M., GAUVREAU K., *Principles of Biostatistics*, Second Edition, Belmont, CA, USA, 2000.

RAEVSCHI E., TINTIUC D., *Biostatistics & Research Methodology*, Nicolae Testemitanu SUMPh, CEP Medicina, Chisinau, 2012.

REA LM, PARKER RA: *Designing and Conducting Survey Research: A comprehensive Guide*, 2nd Edition Jossey-Bass, 1997

SCHLESSELMAN JJ: *Case-Control Studies: Design, Conduct, Analysis*. Oxford, 1982.

TAYLOR B. R. Medical Writing: A Guide for Clinicians, Educators, and Researchers, 3rd Edition, Springer, 2018.

TINTIUC D., BADAN V., RAEVSCHI E., GROSSU IU., GREJDEANU T., ET AL. *Biostatistica si Metodologia Cercetarii Stiintifice*, USMF „Nicolae Testemitanu", CEP Medicina, Chisinau, 2011.

WEINSTEIN MC, FINEBERG HV: *Clinical Decision Analysis*, WB Saunders, 1998.

APPENDIX A: Critical values for the "t" distribution

Degrees of freedom	One-Tailed Test				
	0.05	0.025	0.01	0.005	0.0005
	Two-Tailed Test				
	0.10	0.05	0.02	0.01	0.001
1	6.314	12.706	31.821	63.657	636.62
2	2.920	4.303	6.965	9.925	31.598
3	2.353	3.182	4.541	5.841	12.924
4	2.132	2.776	3.747	4.604	8.610
5	2.015	2.571	3.365	4.032	6.869
6	1.943	2.447	3.143	3.707	5.959
7	1.895	2.365	2.998	3.499	5.408
8	1.860	2.306	2.896	3.355	5.041
9	1.833	2.262	2.821	3.250	4.781
10	1.812	2.228	2.764	3.169	4.587
11	1.796	2.201	2.718	3.106	4.437
12	1.782	2.179	2.681	3.055	4.318
13	1.771	2.160	2.650	3.012	4.221
14	1.761	2.145	2.624	2.977	4.140
15	1.753	2.131	2.602	2.947	4.073
16	1.746	2.120	2.583	2.921	4.015
17	1.740	2.110	2.567	2.898	3.965
18	1.734	2.101	2.552	2.878	3.922
19	1.729	2.903	2.539	2.861	3.883
20	1.725	2.086	2.528	2.865	3.850
21	1.721	2.080	2.518	2.831	3.819
22	1.717	2.074	2.508	2.819	3.792
23	1.714	2.069	2.500	2.807	3.767
24	1.711	2.064	2.492	2.797	3.745
25	1.708	2.060	2.485	2.787	3.725
26	1.706	2.056	2.479	2.779	3.707
27	1.703	2.052	2.473	2.771	3.690
28	1.701	2.048	2.467	2.763	3.674
29	1.699	2.045	2.462	2.756	3.659
30	1.697	2.042	2.457	2.750	3.646
40	1.684	2.021	2.423	2.704	3.551
60	1.671	2.000	2.390	2.660	3.460
120	1.658	1.980	2.358	2.617	3.373
⋈	1.645	1.960	2.326	2.576	3.291

APPENDIX B: **Critical values for the Chi-square distribution**

Degrees of freedom	**Significance level (α)**								
	0.995	**0.00**	**0.975**	**0.95**	**0.9**	**0.1**	**0.05**	**0.025**	**0.01**
1	0	0	0	0	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.1	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.3	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.2	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.7	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.6	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.27
12	3.07	3.57	4.4	5.23	6.3	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.6	5.23	6.26	7.26	8.55	22.31	25	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.3	28.85	32
17	5.7	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.2	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.4	13.85	15.66	33.2	36.42	39.36	42.98
26	11.16	12.2	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	16.49	20.6	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.5	17.79	19.81	21.66	23.95	44.9	48.6	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.9	61.16
42	22.14	23.65	26	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.2
50	27.99	29.71	32.36	34.76	37.69	63.17	67.5	71.42	76.15
55	31.73	33.57	36.4	38.96	42.06	68.8	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.4	79.08	83.3	88.38
65	39.38	41.44	44.6	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.2	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.9	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.09	74.22	77.93	82.36	118.5	124.34	129.56	135.81

USMF „Nicolae Testemițanu”
Centrul Editorial-Poligrafic *Medicina*
Formatul hârtiei 60x84 $^{1}/_{16}$ Tiraj: 200 ex.
Coli de autor 4,7 Comanda nr. 32
Chișinău, bd. Ștefan cel Mare și Sfânt, 165