MINISTRY OF HEALTH OF THE REPUBLIC OF MODOVA
NICOLAE TESTEMITANU STATE UNVERSITY OF MEDICINE AND PHARMACY
OF THE REPUBLIC OF MOLDOVA

Department
Nicolae Testemitanu Social Medicine and Management

Elena Raevschi, Olga Penina

BASIC BIOSTATISTICS
& RESEARCH METHODOLOGY

(Second Edition)

CHISINAU
Centrul Editorial-Poligrafic *Medicina*
2023

CZU 311.2:614.2(075)
R 19

Recommended to printing by Quality Management Board at the
Nicolae Testemitanu State University of Medicine and Pharmacy,
minutes no. 02 of 22 November, 2022

Authors:

Elena RAEVSCHI – PhD habilitate, associate professor, Nicolae Testemitanu Department of Social Medicine and Health Management, Nicolae Testemitanu University of Medicine and Pharmacy

Olga PENINA – PhD, associate professor, Nicolae Testemitanu Department of Social Medicine and Health Management, Nicolae Testemitanu University of Medicine and Pharmacy

Reviewers:

Larisa SPINEI – PhD habilitate, professor, Nicolae Testemitanu Department of Social Medicine and Health Management, Nicolae Testemitanu University of Medicine and Pharmacy

Alexandru CORLATEANU – PhD habilitate, professor, Nicolae Testemitanu Department of Internal Medicine, Nicolae Testemitanu University of Medicine and Pharmacy

The material is published in the author's edition

The Second Edition of Methodic recommendation BASIC BIOSTATISTICS & RESEARCH METHODOLOGY is performed based on the international experience and the updated didactic standards, as well. It corresponds to the requirements of the syllabus of the discipline Biostatistics and Scientific Research Methodology for medical students of the Nicolae Testemitanu State University of Medicine and Pharmacy.

DESCRIEREA CIP A CAMEREI NAŢIONALE A CĂRŢII DIN REPUBLICA MOLDOVA
Raevschi, Elena.

Basic Biostatistics & Research Methodology / Elena Raevschi, Olga Penina; Ministry of Health of the Republic of Moldova, *Nicolae Testemitanu* State University of Medicine and Pharmacy of the Republic of Moldova, Department *Nicolae Testemitanu* Social Medicine and Management. – 2nd ed. – Chişinău: CEP *Medicina*, 2023. – 141 p.: fig., tab.

Bibliogr.: p. 137-138. – În red. aut. – [200] ex.

ISBN 978-9975-82-311-1.

311.2:614.2(075)
R 19

Contents

Preface

Health research is an interdisciplinary field often being achieved by using a large spectrum of knowledge contributions. Biostatistics and Research Methodology as a discipline is required to conduct research according to the worldwide standards. This paper is prepared in accordance with the discipline syllabus and contains comprehensive and well-structured information about all steps necessary for an appropriate scientific study. This paper was written for students of the health sciences as an introduction to the Biostatistics and Research Methodology destined to help conducting the undergraduate research theses.

New to the Second Edition

The second edition includes revised structure, contents and expanded discussion on many topics throughout the book, and additional figures and tables to help clarify concepts for descriptive statistics, and, especially, for inferential statistics which has been completed with more detailed concepts of hypothesis testing using parametric and nonparametric tests, as well as correlation and regression analyses. Tables containing critical value for t-distribution corresponding to commonly used areas under the curve and chi-square distribution have been added to Appendix A and Appendix B. We have also added new chapters, as Research Ethics Introduction and Plagiarism Preventing, as well, exercises, including questions reviewing the basic concepts covered each chapter.

Acknowledgements

We thank the professors who reviewed the manuscript: Larisa Spinei, PhD habilitate, professor, Nicolae Testemitanu Department of Social Medicine and Health Management, Nicolae Testemitanu

University of Medicine and Pharmacy, and Alexandru Corlateanu, PhD habilitate, professor, Nicolae Testemitanu Department of Internal Medicine, Nicolae Testemitanu University of Medicine and Pharmacy.

Our thanks to all professors and teaching assistance who have helped us teach the course and who have made many valuable suggestions.

Authors

CHAPTER 1. INTRODUCTION TO BASIC BIOSTATISTICS AND MEDICAL RESEARCH METHODOLOGY FIELD

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

The course Biostatistics and scientific research methodology covers the theoretical and practical aspects related to the realization of a scientific research and the statistical data analysis. The course has the content similar to other European universities with up-to-date information, and it equips students with necessary baggage of knowledge in order 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:

To help the students to understand the basic concept of Biostatistics in such a way that they can use it to plan and to analyze data in simple biomedical researches.

At the knowing and understanding level:

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

- To know the principles, technology, methods and technics 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, also limits of modern methods used in Scientific Research.

At the application level:

- To analyze definitions, theoretical and practical methods of Methodology of Scientific Research.
- To use statistical methods and techniques in the scientific process.
- To demonstrate capability of analysis, interpretation and presentation of scientific research results.
- To use base knowledge of biostatistics necessary for understanding its optimal application in getting a 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 specialty 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 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 to use information critically from scientific publications in own researches using the new information and communication technologies.

Study outcomes. The student at the end of the course will be able:

- To explain the basic concepts with regard to the organization of a 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, 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, mostly achieved through a large specter of knowledge. Biostatistics and Scientific Research Methodology in medicine field is a discipline, which allows integrating and analyzing obtained knowledge through studies of fundamental and applicable disciplines.

This discipline is necessary for evaluation of research activities with modern research standards compliance. Being a discipline of integration, it correlates with other disciplines that use Statistics.

For a better understanding of the discipline, it is necessary to possess the basics of high school mathematics, as well the basics in biomedicine.

CHAPTER 2. DESCRIPTIVE STATISTICS: DATA PRESENTATION

Descriptive Statistics organizes and summarizes a large set of data observations by a few meaningful numbers. These allow to digest and to figure out large quantities of data, and to effectively communicate to others important aspects of research. Descriptive statistics does not explain or interpret the results: only *suggest hypotheses.*

2.1 Basic Concepts

2.1.1 Variable

Definition of Variable

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

Examples: age, data of birth, nationality, number of children, blood pressure*, etc*.

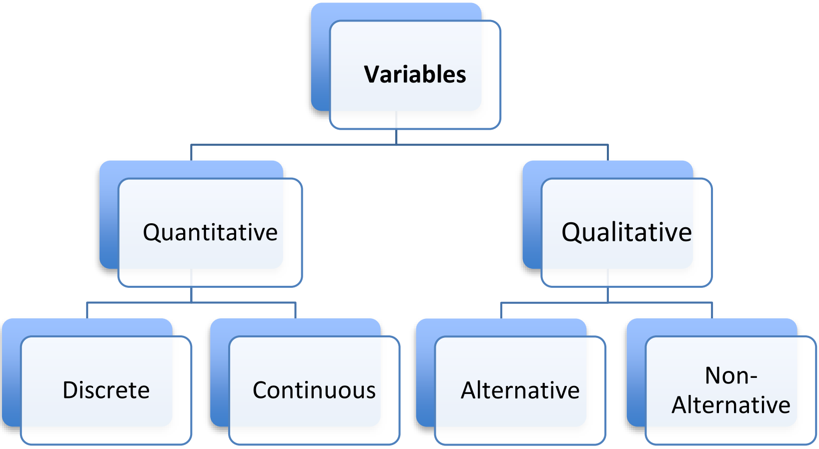


Figure 2.1 Types of Variables

Classification of Variables

To be able to correctly present descriptive (and inferential) statistics, it necessary to understand the data types that usually encountered in any research study(*Figure 2.1*).

Quantitative Variables (Numerical) represent the variables which can be quantified. They are classified as follows: (1) Discrete and (2) Continuous variables.

(1) *Discrete Variables* represent measurable quantities taking only specified values that differ by fixed amounts (whole numbers); no intermediate values are possible, and has values equal to integers.

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

(2) *Continuous Variables* represent measurable quantities but are not restricted to taking on certain specified values (such as integers) having values on a continuum.

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

Qualitative Variables (Categorical) represent the variables which cannot be quantified. They are classified as follows: (1) Alternative and (2) Non-Alternative variables.

(1) *Alternative Variables (Dichotomous or binary)* represent measurable categories in that outcome can take only one of two values: “Yes” or “No”, *etc*.

(2) *Non-Alternative Variables* represent measurable categories in that outcome can take many categories values.

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

In terms of analyses performing the categories of qualitative variables are coded by numbers.

2.1.2 Population

Statisticians term used to describe a large totality of items that have something in common being the subject of a study:

a) *Target population* – known *as theoretical statistic population:* is the group to whom we wish to generalize our findings.

b) *Study population* – known *as accessible population:* is the actual frame of the study, from which we drew our sample.

2.1.3 Sample

The sample is a subset of the study population selected to represent the entire target population of research interest.

Unit of observation, sometimes also called statistical unit – the entity on which information is collected in the population or sample of interest. An observation is the value, at a particular period, of a particular variable having simple or complex entity.

Common *examples* include individual, household, community, school, etc. Clearly identifying the unit of observation is important for a logical survey design, organized data collection, and an objective analysis.

2.1.4 Parameters and Statistics

Table 2.1

Population and Sample Symbols

	Parameters (from Population Data)	**Statistics** (from 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

Statistic is a characteristic, or value, derived from sample data.

Parameter is a characteristic, or value, derived from population data.

2.2 Scales of Measurement

There are different scales of measurement depending on the nature of the variables.

The scale of measurement has implications for the way information is displayed and summarized and determines the statistical methods for analyzing the data. There are 3 scales of measurements in statistics:

- Nominal
- Ordinal
- Numerical (Interval and Ratio scales)

2.2.1 Nominal Scale

(= classificatory scale):

These only have categorical nature.

A variable measured on a nominal scale may have one, two or more subcategories depending upon the extent of variation of qualitative variables.

For example, the variable "*gender*" can have only two values: *male and female.* Nominal data that take one of two distinct values – such as *male* and *female* – are considered *Alternative categorical variable or Dichotomous variable*.

However, not all nominal data need be dichotomous. Often there are three or more possible categories into which the observations can fall. For example," Anemias" can be classified in many sub-categories as microcytic (including iron deficiency), macrocytic or megaloblastic (including vitamin B_{12} deficiency), and normocytic (often associated with chronic disease) – *Non-Alternative categorical variable.* The sequence in which subgroups are listed makes no difference, as there is no relationship among subgroups.

2.2.2 Ordinal Scale

(= ranking scale):

Besides categorizing individuals, objects, responses or a property into subgroups on the basis of common characteristic, it ranks the subgroups in a certain order.

They are arranged either in ascending or descending order according to the extent a subcategory reflect the magnitude of variation in the variable. When the order among categories becomes important, the data are referred to as *ordinal* scale of measurement.

For example, "income" can be measured either using quantitative variables such as lei, dollars or qualitative variables such as "above average", "average" and "below average". *The "distance" between these subcategories is not equal as there is no quantitative unit of measurement.* For example, to illustrate Apgar scores, which describe the maturity of newborn infants, range from 0 to 10, with lower scores indicating depression of cardiorespiratory and neurologic functioning and higher scores indicating good functioning. The difference between the scores of 8 and 9 doesn't have the same clinical implications as the difference between scores of 0 and 1.

So, in ordinal scales the interval between 2 points of measurements in sequence is not the same.

2.2.3 Numerical Scale

The interval between 2 points of measurements in sequence is always the same. There are 2 types of numerical scales:

⇒ *Interval numerical scale*

An interval scale has all the characteristics of a nominal and ordinal scale and, additionally, the data can be arranged in a hierarchical order with the same distance between them. Interval scale has no absolute zero value: can takes meaningful values below and above the zero.

For example,

Celsius scale:	0*C to 100*C
Fahrenheit scale:	32*F to 212*F

$\Rightarrow$ *Ratio numerical scale*

A ratio scale has all the properties of nominal, ordinal and interval scales plus its own property: the zero point of a ratio scale is fixed, which means it has absolute zero value (there are no values below zero).

The measurement of variables like serum cholesterol, income, age, body height and weight are examples of this scale.

2.3 Tables

It is very important to identify the statistical technics that are most appropriate for describing each kind of variable in order do not lose a certain amount of information when data are summarized.

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

One type of table that is commonly used to display data is known as *frequency distribution*.

For nominal and ordinal data, a frequency distribution consists of a set of categories along with the numerical counts that correspond to each one. As a simple illustration of this format, *Table 2.2* displays the numbers of students (numerical counts) who did and did not smoke (classes or categories) in a university in 2021. A more complex example is given in *Table 2.3*, which specifies the numbers of cigarettes smoked per adult in the *country X* in various years.

To display discrete or continuous data often need grouped frequency distribution approaches. This implies to break down the range of values of the observations into a series of distinct, nonoverlapping intervals. Once the upper and lower limits for each interval have been selected, the number of observations whose values fall within each pair of limits is counted, and the results of grouped frequencies distribution are arranged as a table; the grouped frequencies of shock-index in 931 patients corresponding to each interval are presented in *Table 2.4.*

Table 2.2
Distribution of students in a university according to their smoking habit, 2021

Smoking Habit	Number of Students
Yes	950
No	5050

Table 2.3
Absolute frequencies of cigarette consumption per adult, Country *X*, 1930-2020

Year	Number of Cigarettes
1930	54
1940	151
1950	665
1960	1485
1970	1976
1980	3522
1990	4171
2000	3985
2010	3581
2020	2528

Table 2.4
Grouped absolute frequencies of shock-index score in 931 patients

Shock Index Score	Number of Patients
0.30-0.39	38
0.40-0.49	104
0.50-0.59	198
0.60-0.69	199
0.70-0.79	155
0.80-0.89	102
0.90-0.99	60
1.00-1.09	37
1.10-1.19	19
1.20-1.29	19

Tables are most informative when they are not overly complex. As a general rule, tables and columns within them should always be clearly labeled. If units of measurements are involved, they should be specified.

2.3.2 Relative Frequency

It is sometimes useful to know the proportion of values that fall into 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 appears in that interval. The relative frequency is computed by dividing the numbers of values within an interval by the total number of values in the table expressed in %.

Relative frequencies are useful for comparing sets of data contain unequal numbers of observations. *Table 2.5* displays the absolute and relative frequencies of shock-index score in 931 patients depicted in *Table 2.4.*

Table 2.5
Absolute and relative frequencies of shock-index score in 931 patients

Shock Index Score	**Frequency, (Number of Patients)**	**Relative Frequency (%)**
0.30-0.39	38	4.1
0.40-0.49	104	11.2
0.50-0.59	198	21.3
0.60-0.69	199	21.4
0.70-0.79	155	16.6
0.80-0.89	102	11.0
0.90-0.99	60	6.4
1.00-1.09	37	4.0
1.10-1.19	19	2.0
1.20-1.29	19	2.0
Total	**931**	**100.0**

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* lists the cumulative relative frequencies for the shock-index score for 931 patients depicted in *Table 2.5.*

Table 2.6
Absolute, relative and cumulative relative frequencies of shock-index score for 931 patients, Emergency Care Unit, 2020

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
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

A second way to summarize and display data is using graphs, which are easier to read than tables but they often supply a lesser degree of details. The most informative graphs are relatively simple and self-explanatory. As tables, they should be clearly labeled and units of measurements should be indicated.

2.4.1 Line Graphs

Line Graphs are commonly used to display trends over the time of nominal data. Each value of the *x-axis* has a single corresponding value on the on the *y-axis*. Adjacent points are connected by straight lines.

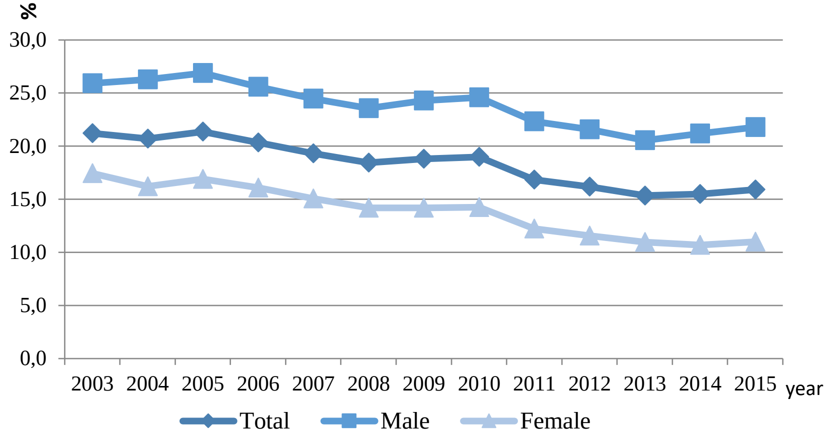
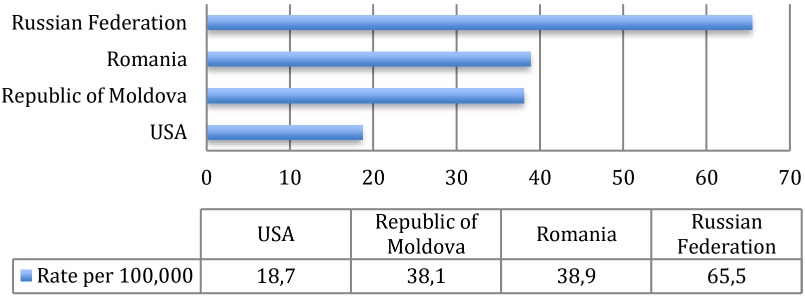


Figure 2.2 Unconditional probability of dying between ages 30 and 70 years from major cardiovascular diseases, Republic of Moldova, 2003-2015 (%)

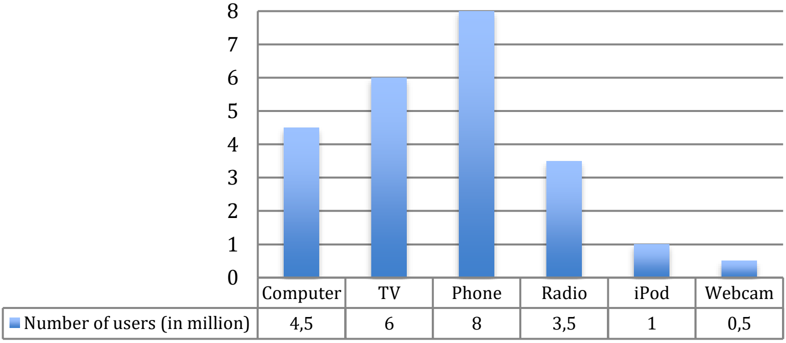
2.4.2 Bar Charts

Bar charts are popular type of graph used to display a frequency distribution for nominal or ordinal data. The bars have to be of equal distance being separated from one other so as not to imply continuity. The horizontal bar charts compare multiple values and the various categories into which the observations fall are presented along a horizontal axis, illustrated in *Figure 2.3.* A vertical bar charts compares values across categories, and is drawn above each category such that the height of the bar represents either the frequency or relative frequency of observations within that class, presented in *Figure 2.4.*



	USA	Republic of Moldova	Romania	Russian Federation
■ Rate per 100,000	18,7	38,1	38,9	65,5

Figure 2.3 Age-standardized prevalence estimates for smoking cigarette per 100,000 population: Daily users – Males, Aged 15+, 2018.



	Computer	TV	Phone	Radio	iPod	Webcam
■ Number of users (in million)	4,5	6	8	3,5	1	0,5

Figure 2.4 Most Used Technology

2.4.3 Pie Charts

Pie charts display contribution of each value to a total and it used to display a frequency distribution for nominal data.

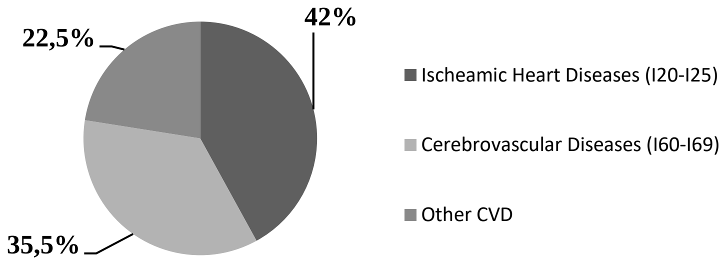


Figure 2.5 Global distribution of cardiovascular deaths due to heart attacks, strokes and other types of cardiovascular diseases, 2020

2.4.4 Histograms and Frequency Polygons

The histograms are the most informative way to present absolute and relative frequencies *(Figure 2.6)*. A histogram looks a bit like a bar charts, but:

- whereas a bar chart is a pictorial representation of a frequency distribution of either nominal or ordinal data *(categorical data)*, a histogram displays a frequency distribution for discrete or continuous data *(numerical data).*
- histograms give an idea of the shape of the relative frequency distribution. Bar charts are just tallies and can’t tell about distribution shapes.

The horizontal axis displays the true limits of the various intervals. The vertical axis displays the absolute or relative frequencies values.

The frequency polygon another commonly used graph, is similar to the histogram in many aspects. A frequency polygon uses the same two axes as a histogram. It is constructed by placing a point at the center of each interval such that the height of the point is equal to the absolute or relative frequency associated with the interval. Points are also placed on the horizontal axis at the midpoints of the intervals immediately preceding and immediately following the intervals that contains observations. Straight lines then connect the points. As in a histogram the frequency of observations for particular interval is represented by the area within the interval beneath the line segment. The frequency polygons are superior to histograms because they can easily be superimposed for compare two or more sets of data displayed in *Figure 2.7.*

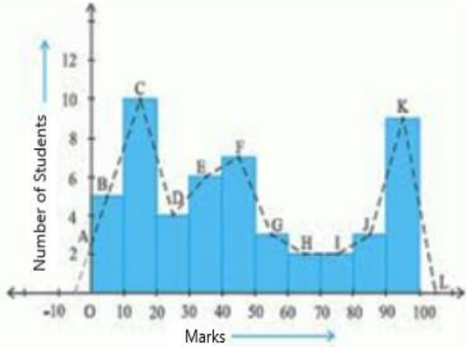


Figure 2.6 Histogram of marks score distribution of the students

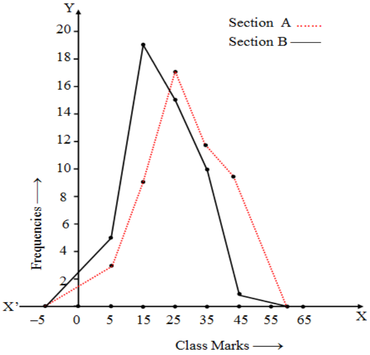


Figure 2.7 Frequency polygons of marks score distribution of the students in section A and B

2.4.5 Box Plots

Box plots are used to summarize a set of discrete or continuous data when there is more than one group. The box plot, sometimes is called a box-and-whisker plot. That way of presentation shows only the summary of the data and no information about every point in the group.

An example of box plot is given in *Figure 2.8* which specifies its elements. The central box, that can be depicted vertically or horizontally, extends from 25th to 75th percentiles of the data set. The line running between marks the 50th percentiles of the data set. If this line lies approximately halfway between the two quartiles, it means that the data set distribution is roughly symmetric.

The lines projecting out to the box (whiskers) represent the adjacent values of the plot which extend 1.5 times the interquartile range above and below the 75th and 25th percentiles. All points outside this range are represented by circle. These observations are considered to be outliers or data points that are not typical of the rest of the values.

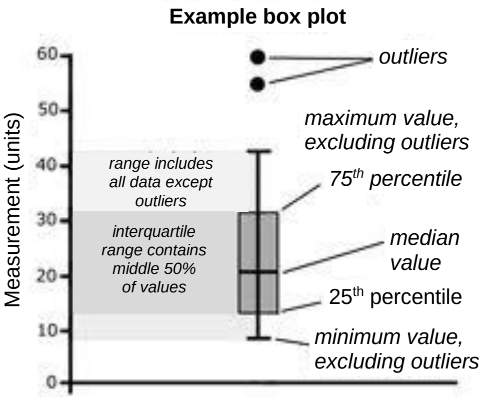

Example box plot
Measurement (units)
60
50
40
30
20
10
0
outliers
maximum value, excluding outliers
75th percentile
median value
25th percentile
minimum value, excluding outliers
range includes all data except outliers
interquartile range contains middle 50% of values

Figure 2.8 Box plot structure components

2.4.6 Error Bar Plots

Error bar plot is often use in the medical literature comparing two or more groups. The circle designates the mean, and the bars illustrate the standard deviation, although some authors use mean and standard error. The error bars indicate the similarity of the distribution, just as box plots do.

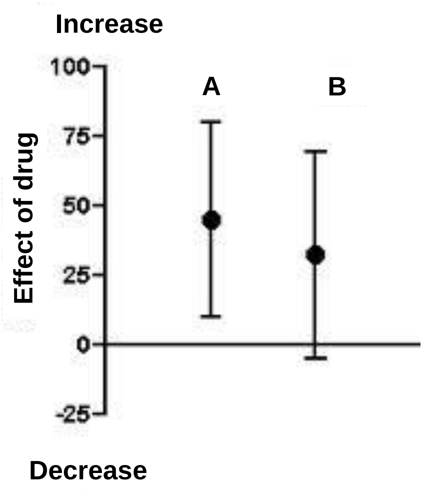


Figure 2.9 Error bar charts of effect of drug for patients with (A) and without (B) a pulmonary embolism

2.4.7 Two-Way Scatter Plots

A two-way scatter plot is used to depict the relationship between two different continuous variables. Each point on the graph represents a pair of values simultaneously.

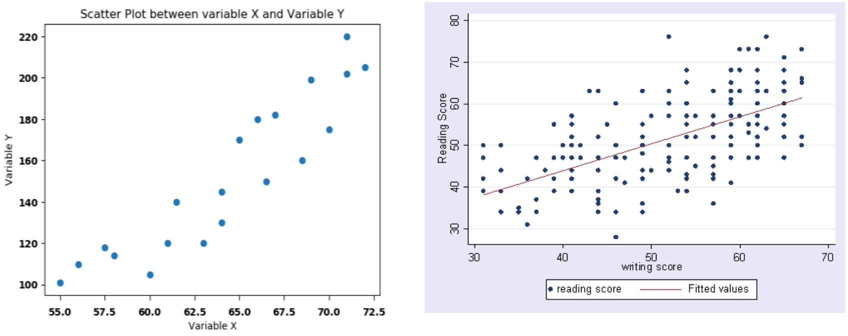


Figure 2.10 Two-way scatter plot graph presentation

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 examination in a given year;
   c. Serum cholesterol level of healthy individuals;
   d. Presence of diarrhea in group of infants.
2. State the type of the variable and appropriate measurement scale for the following sets of data:
   a. The age onset of breast cancer in females;
   b. Body temperature of the patients;
   c. Discharged patient outcome;
   d. Number of births in a given year.
3. Use the following data to display it by all appropriate graph. State your decision.

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

4. Use the following data to display it by all appropriate graph. State your decision.

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

5. Propose a set of data that can be displayed by line chart. State your decision.
6. Propose a set of data that can be displayed by bar chart. State your decision.

Review Questions

1. What are descriptive statistics?
2. Classification of variables. Give examples.
3. How does alternative variable differ from non-alternative one? Give examples.
4. Definition of measurements scale. Classification. Give an example for each type.
5. How do ordinal data differ from nominal data? Give examples.
6. What kind of data presentation do you know? State the difference between them.
7. When you construct a table, when might it be beneficial to use relative rather than absolute frequencies distribution?
8. Graph data presentation: contents and types. Give an example for each type.
9. Name appropriate graph data presentation for nominal variable. Give examples.
10. Name appropriate graph data presentation for ordinal variable. Give examples.
11. Name appropriate graph data presentation for numerical variable. Give examples.

CHAPTER 3. DESCRIPTIVE STATISTICS: SUMMARIZING NUMERICAL DATA

As you have already known Descriptive Statistics is used to organize and describe the characteristics of a set of data.

Descriptive Statistics has no hypothesis and doesn't analyze data.

Descriptive statistics of numerical data allows us to make concise, quantitative statement that characterize the distribution of values as a whole. Together, the various types of descriptive statistics can provide a great deal of information about a set of observations.

Descriptive Statistics numerical summary measures are:

⇒ 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 the most useful numbers, which characterize the middle (the center) of the set of data where observations tend to cluster. The three measures commonly used in medicine are: mean, median, and mode. All three are used for numerical data summarizing, 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 dividing the sum (Σ) of the individual values (x_i) by the number of observations (n):

a) *The Simple mean* – used for data set when all values occur one time only (f=1).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

b) *The Weighted Mean* – used for data set when some values occur more than one time (f>1).

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\, f,$ *where* "f "– is 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 data set (1;2;2;3) is 8/4 or 4. 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. So, the mean 3 is more appropriate for the set data, then the mean 6.

3.1.2 Median

The median (M_d) divides the ordered array into two equal parts. The median is the middle point in the observation data set, then a half of observations are smaller and half are larger.

The median is less sensitive to extreme values than the mean is. Medians frequently are used to measure the middle of the distribution of an ordinal or numerical characteristic that is skewed. When the data are not symmetric the median is the best measures of central tendency.

Before to calculate median, you have to arrange the observations from smallest to largest:

⇒ If, an *odd number of observations*, the median M_d will be the middle value in arranged data or the $[(n+1)/2]^{th}$ of observation.

For example, the median of data set with n=5 (1; 2; $\underline{4}$; 5; 6) is $M_d = 4$.

⇒ If, an *even number of observations*, the median M_d will be the midpoint between the middle two observations. Ex.: Median of 14 observations is the midpoint between 7th and 8th or the

$(n/2)^{th}$ and $[(n/2)+1]^{th}$ positions of observations in the ordered data set. The median is the mean of the two middle most values.

For example, the median of data set with n=4 (1; <u>2; 4</u>; 5) is $M_d = 3$.

3.1.3 Mode

The Mode (M_o) is a value that occurs most frequently in data set.

Ex.: 3; 4; 5; 6; 6; 6; 7; 8; 9, Mo=6.

There is no mode, if all values are different. May be more than one mode: 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 practical application for a given set of date depends on the shape of the distribution of data or the way in which the values are distributed *(Figure 3.1):*

⇒ Normal Distribution

Normal distribution of data set values is symmetric around its center and has the same shapes on both sides of the mean, and is called bell-shaped curve, as well. Then the mean, the median, and the mode should all be roughly the same. When the data are symmetric the mean is the best measures of central tendency in terms of date set representativeness assuring.

Relationship of central tendency measurements in this case is:

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

⇒ Skewed Distributions

If outlying observations occur in only one direction, the distribution is called a skewed distribution. When the data are not symmetric the median is the best measures of central tendency in terms of data set representativeness assuring. Because the mean is sensitive to extreme observations, it is pulled in the direction of the outlying data values and as a result might end up either excessively inflated or excessively deflated.

There are two types of the *skewed distributions:*

1. Negatively (skewed to the left) – outlying values are small.

Relationship of central tendency measurements in this case is:

$\bar{x} < M_d < M_o$

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

Relationship of central tendency measurements in this case is:

$M_o < M_d < \bar{x}$

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

The following guidelines are useful in deciding which measure of central tendency is most appropriate for a given set of data:

- The mean is used for numerical data and for symmetric (not skewed) distribution.
- The median is used for ordinal data or for numerical data if the distribution is skewed.
- The mode almost is used for bimodal distribution.

Remember that when we summarize a set of data, information is always lost. Thus, although it is helpful to know where the center of data set lies, this information is usually not sufficient to characterize an entire distribution of measurements.

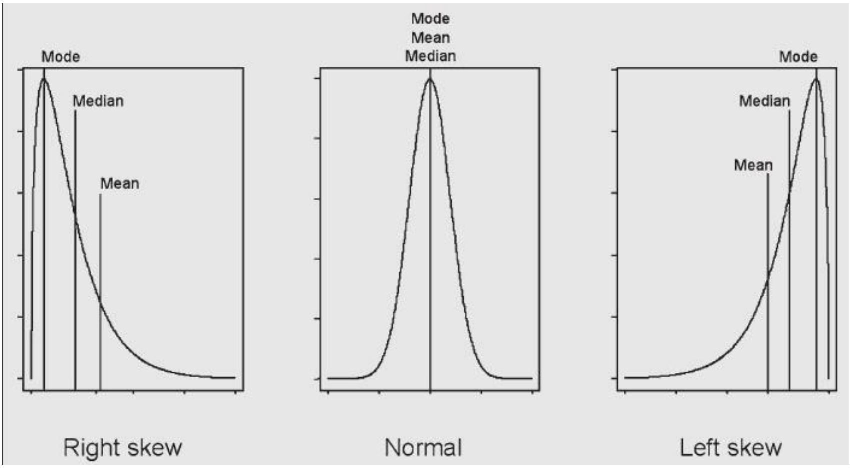


Figure 3.1 Possible distribution of data values

3.2 Measures of Variability

The measurements of central tendency describe only the middle of the set data, being used alone they are not able to describe adequately it.

To know how good our measure of central tendency actually is, we need to have some idea about the data variation. Do all the observations tend to be quite similar and therefore lie close to the center, or are they spread out across a broad range of values? As in following example, in each of the two very different distributions of data values, 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)

As we see the measurements of the central tendency are the same even the set data are different. It's why for appropriate descriptive statistics of set data measurements of central tendency must be used with variability measurements. Certainly, it is necessary to describe the dispersion of the data set which is the target of the measures of variability (dispersion): Range, Interquartile Range, Variance, Standard Deviation and Coefficient of Variation.

3.2.1 Range

Range is defined as the difference between the largest and smallest values in the data set. Although the range is easy to compute:
Range (R) =Max (x_i) – Min (x_i)

Example:
Set 1: -200; -20; -10; 7; 10; 20; 200 (n=7; Mean $\bar{x}$= 1)
Set 2: -20; -5; -2; 7; 2; 5; 20 (n=7; Mean $\bar{x}$=1)
R1 = 400 and R2 = 40

Range is heavily influenced by the most extreme values and ignores the rest of the distribution. Therefore, like mean, it is highly sensitive to exceptionally large or small values. The range is used with numerical data when the purpose is to emphasize extreme values.

3.2.2 Interquartile Range

A second measures of variability that is used to limit the influence of extreme values is called *interquartile range* (IQR).

Quartiles are the values that divide ordered data set into quarters. Interquartile range is defined as the difference between the 25th and 75th percentiles (the percentage of the distribution), also called the first and third quartiles (Q_1 and Q_3). The 50th percentile ($Q_{2)}$ is the middle of data set being the median, as well. *Figure 3.2* displays the quartiles and interquartile range by box plot graph presentation.

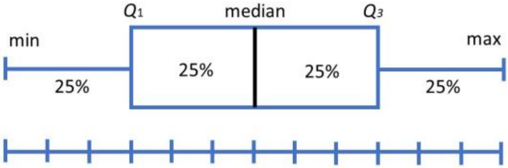


Figure 3.2 Graph presentation of quartiles and interquartile range

The *interquartile range* is calculated by subtracting the 25th percentile from the 75th percentile; consequently, it encompasses the middle 50% observations, as follows:

$IQR = Q_3 – Q_1$

For compute the quartiles rank the orders from lowest to highest and use:

$Q_1= (n+1)/4$ ranked values
$Q_2= (n+1)/2$ ranked values
$Q_3= 3(n+1)/4$ ranked values

- Interquartile Range (IQR) is used in two situations:
  1. When the median is used (ordinal data or skewed numerical data)
  2. When the mean is used but the target is to compare individual observations with a set of norms.
- Interquartile range is appropriate to be used for describe the central 50% of the distribution, regardless of its shape.

3.2.3 Variance and Standard Deviation

The *variance* (S^2) is the measure of how spread out a distribution is quantifying the amount of variability, or spread, around the mean of measurements.

The *variance* is calculated as the average distance of the individual's values from the mean, or defined as the average squared deviation of each number from its mean. More explicitly, the variance is calculated by subtracting the mean of a set of values from each of the observations, squaring these deviations, adding them up, and dividing by 1 less than the number of observations in the data set. Representing the variances by S^2,

$$\mathrm{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \mathrm{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i$$

where:
x_i – individual values of data set; $\bar{x}$ – mean; *n* – number of observations, f_i – frequency associated with the *i^th^* interval.

The variance has rather theoretical significance than practical being a first step of Standard Deviation calculation.

The *standard deviation (S)* is the most commonly used method to describe the variability of a data set, being a measure of the spread around the mean.

The standard deviation is an estimate of the average distance of the values from their mean and is calculated as the positive square root of the variance.

Knowing variance, is very easy to calculate standard deviation (S), which is square root taken by variance:

$$\mathrm{S} = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_{i-}\bar{x})^2}{n-1}} \qquad \mathrm{S} = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_{i-}\bar{x})^2 f_i}{n-1}}$$

where:
S^2 – variance; x_i – individual values of data set; $\bar{x}$ – mean; *n* – number of observations, f_i – frequency associated with the i^{th} interval.

The standard deviation is used when the mean is used (symmetric numerical data). Together, the mean and standard deviation of a set of data can be used to summarize the characteristics of the entire distribution of values. The standard deviation has the same units of measurements as the mean. Therefore, it is meaning less to compare standard deviations for two unrelated quantitates.

3.2.4 Coefficient of Variation

Coefficient of variation is a relative variation rather than absolute variation as standard deviation is. The coefficient of variation is used when the intent is to compare distributions measured in different scales.

The coefficient of variation 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; $\bar{x}$ – mean.

As it has no units, the coefficient of variation can be used to compare two sets that are measured on different scales.

Even not making comparison, the coefficient of variation can be used to appreciate the variability level in a single set data, according the scale displayed in *Table 3.1.*

Table 3.1

Scale of Coefficient of Variation

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

For scientific research a set data must have low or medium level of variability, because if the level of dispersion of data is high the mean is not representative for that data set.

The following guidelines are useful in deciding which measure of variability is most appropriate for a given set of data:

1. The standard deviation is used when the mean is used (i.e., with symmetric numerical data).
2. Interquartile range is used:
    a. When the median is used (i.e., with ordinal data or with skewed numerical data).
    b. When the mean is used but objectives is to compare individual observations with a set of norms.
    c. To describe the central 50% of a distribution, regardless of its shape.
3. The range is used with numerical data when the purpose is to emphasize extreme values.
4. The coefficient of variation is used when the intent is to compare distributions measured on different scales.

3.3 Normal Distribution and its Properties

Normal distributions have key characteristics that are easy to spot in graph as is presented in *Figure 3.3:*

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

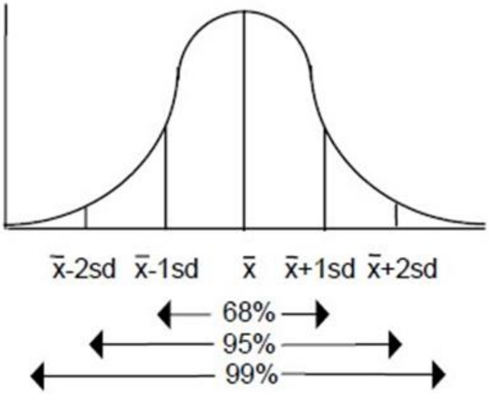


Figure 3.3 The Normal Distribution Properties

If the data is normally distributed (bell-shaped curve), approximately 68% of data will lie within 1 standard deviation, approximately 95% within 2 standard deviations, and approximately 99% of data will lie within 3 standard deviations.

The standard deviation, along with the mean, can be helpful in determining skewness when only summary statistics are given: if the mean minus 2 standard deviations contains zero (i.e., the mean is smaller than 2 standard deviations), the observations are probably skewed.

3.4 Skewness and Kurtosis in terms of Normality Checking

As you know, the skewness is the degree of distortion from the normal distribution (bell shaped). Acceptable range for normality is skewness lying between -1 to +1. Normality checking should not be based on skewness alone.

Therefore, the kurtosis of the bell-curve is important to take into consideration, as well. The shape of the vertex (the peakness) of the curve is known as kurtosis (see *Figure 3.4*). Kurtosis is all about the tails of the distribution. It is actually the *measure of outliers* present in the distribution. Acceptable range for normality is kurtosis lying between -1 to +1.

There are three types of kurtoses:

⇒ *Platy Kurtic* – If the peak of the curve is flat compared to normal and the tails are long

⇒ *Meso Kurtic* – If the peak of the curve is normal and tails on both side of mean are also normal

⇒ *Lepto Kurtic* – if the peak of the curve is narrow and tails on both sides of the mean are small

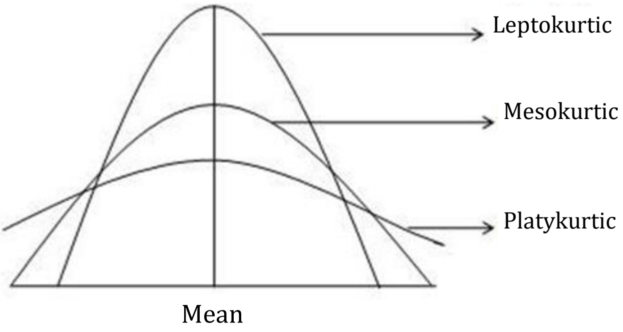


Fig. 3.4 Types of Kurtosis

3.5 Pearson’s Coefficient of Skewness

There are many formal statistical tests for normality checking using the computer. One caution to using the formal test is that these tests are very sensible to the sample sizes of the data:

Table 3.2

Flowchart for normality checking

1.	*Small samples* (n<30) – Always assume not normal
2.	*Moderate samples* (30 – 100) – If formal test is significant, accept non-normality otherwise double-check using graphs, skewness and kurtosis to confirm normality.
3.	*Large samples* (n>100) – If formal test is not significant, accept normality otherwise double-check using graphs, skewness and kurtosis to confirm non-normality.

Karl Pearson suggested simple calculation as a measure of asymmetry or skewed distribution.

The Pearson skewness coefficient (C_{as}), defined by:

$$C_{as} = \frac{Mean - Mode}{Standard\ Deviation} = \frac{\bar{x} - M_O}{S}$$

Skewness coefficient is an abstract value that determine the level and the type of skewness and its value fall in the interval (-1, +1):

$C_{as} = 0$, the distribution is symmetric;

$C_{as} \Rightarrow 0$, the distribution is easy skewed;

$C_{as} \Rightarrow (+/- 1)$, the distribution is heavily skewed;

C_{as} (interval 0; +1), positively skewed;

C_{as} (interval -1; 0), negatively skewed.

Review Exercises

Example solving:

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			
26	4	104			
30	3	90			
31	5	153			
33	2	66			
35	3	105			
Total	n=19	566			

3. Calculate the Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\, f\ ,$$

where "f "– is frequency of individual data

$$\bar{x}= \frac{24\times2+26\times4+30\times3+31\times5+33\times2+35\times3}{19} = \frac{566}{19} = 29.8$$

4. The Mode = 31 (frequency is five)
5. Calculate the Median

'19' is an odd number of observations, the median position will be the (n+1)/2 observation.

Median position is (19+1)/2=10

The 10th observation is the median (only in arranged set of data)

M_d = 31

6. For calculate the standard deviation we should find out about the variance at first:

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

X_{i1} = 24 – 29.8= - 5.8

X_{i2} = 26 – 29.8 = - 3.8

X_{i3} = 30 – 29.8 = +0.2

X_{i4} = 31 – 29.8 = +1.2

X_{i5} = 33 – 29.8 = +3.2

X_{i6} = 35 – 29.8 = +5.2

Continue to complete the frequency table with the calculate data:

$\boldsymbol{x_i}$	***Frequency, f***	$\boldsymbol{x_i f_i}$	$\boldsymbol{x_i - \bar{x}}$	$(\boldsymbol{x_i - \bar{x}})^2$	$(\boldsymbol{x_i - \bar{x}})^2 f_i$
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

$S^2 = \frac{233.96}{18} = 13$;

then calculate the standard deviation:

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

7. Calculate the coefficient of variation:

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

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

Conclusions:

- The ages of the data set patients that you are studying is 29.8±3.6 years.
- Data variability level is medium (CV=12.1%) – appropriate for scientific research.
- The distribution of data is almost symmetric (acceptable skewed to the left): Mode=31, Median=31, and Mean=29.8.
- The mean is representative for this data set.

Review Exercises:

1. Using the following data set:
   - Find the measures of central tendency;
   - Analyze relationship between the measures of central tendency and identify the shape of the distribution. State your decision.

A study of 10 patients was conducted investigating the BMI data: 29; 22; 24; 37; 23; 35; 35; 27; 50; 37.

2. Using the following data set:
   - Find all measures of central tendency;
   - Analyze relationship between the measures of central tendency and identify the shape of distribution. State your decision.

A study of 11 patients was conducted investigating the blood glucose level data: 3.3; 12.0; 9.0; 6.0; 11.0; 11.8; 11.8; 11.0; 5.5; 3.3 (moll/l).

3. Using the data set from p.1:
   - Find all measures of variability: State your decision;

- Make your decision how representative is the mean for this data set?
- According to scale of measurements and shape distribution of this data set which summary measures of central tendency are more appropriate for?
- Construct a box plot.

4. Using the data set from p.2:
   - Find all measures of variability: State your decision;
   - Make your decision how representative is the mean for this data set?
   - According to scale of measurements and shape distribution of this data set which summary measures of central tendency are more appropriate for?
   - Construct a box plot.

Review Questions

1. The mean: definition, types, and its rules of calculation. Give an example.
2. The median: definition and its rules of calculation. Give an example.
3. The mode: definition and its rules of calculation. Give an example.
4. Compare the mean, median and mode as measures of central tendency.
5. Under what conditions is use of the mean preferred?
6. Under what conditions is use of the median preferred?
7. Under what conditions is use of the mode preferred?
8. Variability measures: reasons for application.

9. Range: meaning, characteristics, and preferred use conditions. Rule of calculation.
10. Quartiles: meaning, characteristics, and rule of calculation.
11. Interquartile range: meaning, characteristics, and preferred use conditions. Rule of calculation.
12. Standard deviation: meaning, characteristics, and preferred use conditions. Rule of calculation.
13. Coefficient of variation: meaning, 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 WITH NUMBERS

4.1 Methods to Describe Nominal Data

As usually obtained statistic information in a research is presented by *absolute values.* These values are difficult to interpret because they are not able to make a comparison, synthesis or correlation among different characteristics.

To make comparisons among groups more meaningful relative values may be used instead of absolute numbers.

Nominal data can be measured using several methods:

- Proportions and percentages
- Ratios
- Rates

4.1.1 Proportions and Percentages

Proportions are extensive statistical indicators expressing the structure of a phenomenon, and is defined as a part of phenomenon divided by whole. The proportion is calculated dividing number, a, of observations with a characteristic of interest by the total number of observations $a+b$. A percentage is simply the proportion multiplied by 100%. That is,

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

The proportion is a static indicator, which never make the association between environment and phenomenon and never allow the evaluation of its dynamic changes: it makes a statistic balance in a specified moment of time only.

The proportion is 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 a part divided by another part, which are 2 independent phenomena. So, the ratio is the number of observations in a group with a given characteristic (a) divided by the number of observations without the given characteristic (b):

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

Graph presentation: Line graph, Bar chart.

4.1.3 Rates

Rates are intensive statistical indicators expressing the frequency (the level) of a phenomenon computed over a specified period of time.

Rate= $\frac{a}{a+b} \times base$

a – number of observations with a given characteristic (such as those who died in a specified year and place)

$a+b$ – total number of observation (such as number of population)

base – is a multiplier (e.g, 100; 1,000; 10,000; 100,000)

Rates are commonly used in the calculation of vital statistics and allow:

- To determine the frequency of a phenomenon spreading in a specified research group;
- To make a comparison of a different groups by their frequency level of a homogenous phenomenon;

- To detect the dynamic changes in the phenomenon frequency spreading on the base of a specified research group.

Graph presentation: Line graph, Bar chart.

4.2 Health Status Indicators in Terms of Descriptive Statistics

Health status indicators measure the health of the population. There are three main types of the health status indicators:

⇒ Mortality indicators;

⇒ Morbidity indicators;

⇒ Disability indicators.

Mostly all health status indicators are represented by rates, and by proportion and ratio when needed. Somme of the most commonly used rates are briefly defined in the following items:

4.2.1 Mortality Rates

Mortality rate is defined as the number of deaths that occur during the specified period of time, divided by the total number of people who were at the risk of dying for the same period of time.

⇒ *A crude rate* is a rate computed over all individuals in a given entire population, regardless differences caused by age, gender, race, *etc*.

⇒ *A specific* rate is a rate computed within relatively small, well-defined subgroups**:**

Mortality rate calculated for individuals age groups are known as *age-specific death rates;* for individual sex group – *sex-specific mortality rate*; for cause group – *cause-specific mortality rate.*

4.2.2 Morbidity Rates

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

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

Incidence is defined as the number of new cases that have occurred during the specified time period divided by the total number of people who were at the risk for the same period of time.

Prevalence is defined as 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.

Prevalence and incidence rates are used to evaluate disease patterns and make future projections.

Morbidity rate provide a standard way to evaluate crude rates and specific rates, as well.

Example solving:

In a specified year and locality number of population is 75,000. In that year died 897 of individuals. In that locality were 40 doctors: 20 – physicians; 10 – surgeons; 10 – other.

Compute all possible measures of nominal data.

1. Proportion and Percentage

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

Percentage of surgeons – the same way that is considered up.

2. Ratio

Doctors supply = $\frac{40}{75{,}000} \times 10{,}000 = 5.3$ *per 10,000 population*

3. *Rate*

Crude mortality rate = $\frac{897}{75{,}000} \times 1{,}000 = 11.9$ *per 1,000 population*

4.3 Adjusted Rates: Direct Method of Standardization

Crude rates can be used to make comparisons between two different populations only if the populations are homogenous in all characteristics. Therefore, if the populations are different by factors such as gender, age, *etc*. instead of crude rate must be used adjusted crude rate by gender, age, *etc*. for comparison making; otherwise comparison will not be valid.

The direct method of standardization focuses on computing the conventional rates that would result if instead of having different characteristics distribution, all groups being compared were to have the same standard composition. So, adjusted rates are conventional values (not real) that make sense only for comparison process and cannot to be used separately.

Example:

Viral Hepatitis Morbidity in factory "A" and "B" for a given year

Sex	Factory "A"		Factory "B"		**Step 1**		**Step 2**	**Step 3**	
	Total nr of Workers (2)	Nr. of develop disease workers	Total nr of Workers (4)	Nr. of develop disease workers	Specific Rates, %		Standard selection (2) +(4)	Expected number	
					A	B		A	B
Males	50	1	170	4	2.0	2.3	220	4.4	5.06
Females	200	10	30	3	5.0	10.0	230	11.5	23.0
Total	250	11	200	7	4.4	3.3	450	15.9	28.06
					Step 4 (Adjusted rates)		100	3.5	6.2

Determining an adjusted rate is a relatively simple process having the following steps:

1. To compute the rates for each comparison group
2. To select the standard distribution
3. To compute the expected number for each group
4. To calculate adjusted rate for each group

Step1

To compute the sex specific viral hepatitis morbidity rates (VHMR) for each comparison group:

Factory A (males) = $\frac{1}{50} \times 100 = 2\%$

Factory A (females) = $\frac{10}{200} \times 100 = 5\%$

Factory A (total VHMR) = $\frac{11}{250} \times 100 = 4.4\%$

The same calculation is provided for factory "B".

Making a comparison we have noted a paradox: the both sex specific rates at the factory "A" (f-5%; m-2%) are lower than sex specific rates at the factory "B"(f-10%; m-2.3%), but viral hepatitis morbidity rate for total number of workers at the factory "A" (4.4%) is higher than at the factory "B" (3.3%). That means the morbidity rates were influenced by gender characteristics.

Step 2

To select the standard distribution

We then calculate the numbers of standard population distribution while retaining its own individual sex-specific morbidity rates. In our example the standard population (called the reference population) is the sum of column (2) and (4) according the gender and total workers data. So, we have found out the standard for males 220 (50+170), for female 230 (200+30), and the total standard population 450 (220+230).

Actually, which population is chosen, as the standard does not matter; in fact, a set of frequencies corresponding to a totally separate references population can be used. The point is that the same set of numbers must be applied to both populations.

Step 3

To compute the expected number for each group

The point is to find out how many sick individuals will be expected in the standard male's population of 220, having the same sex-specific viral hepatitis morbidity rate as at factory 'A" is (2%). That is,

100 – 2

220 – x

x = $\frac{220 \times 2}{100} = 4.4$

So, expected number of sick males' workers for standard population of 220 is 4.4.

Fallow the same way for the next calculation of expected number for groups: female factory "A" and males and female's factory "B".

Total expected number for group of factory "A" and "B": sum of expected number for males and females according the factory:

Total expected number factory "A" = 4.4+11.5 = 15.9

Total expected number factory "B" = 5.06+23.0 = 28.06

Step 4

To calculate adjusted rate for each group

The sex-adjusted viral hepatitis morbidity rate for each factory is then calculated by dividing its total expected number of sick individuals by the total number of standard population:

Factory "A" = $\frac{15.9}{450} \times 100\% = 3.5\%$

Factory "B" = $\frac{28.06}{450} \times 100\% = 6.2\%$

Summarizing the results and its interpretation:

	Factory "A"	Factory "B"	Comparison Results
Specific Rates	4.4	3.3	A > B; false
Adjusted rates	3.5	6.2	A < B; true

Conclusions:

Making comparison of sex-adjusted viral hepatitis morbidity rate: This is the opposite of what we observed when we looked at the specific rates, implying that this specific morbidity rates were indeed influenced by the gender structure of the underlying groups (Factory "A" and "B").

Review Exercises

1. In a locality A in a specified year were registered 2,500 illness: 800 of them – cardiovascular diseases; 500 pulmonary diseases; 450 – injuries, and other – 750. The number of population is 900,000.
   - Compute all possible vital statistics.
   - Make appropriate graph presentation for them.
2. In a locality B in a specified year the number of population is 78,000. In this year 110 individuals died and 400 individuals develop cardiovascular disease for the first time in their life.
   - Compute all possible vital statistics.
   - Make appropriate graph presentation for them.
3. Consider the following data comparing acute abdomen lethality at the hospital "A" and "B":

The term of hospitalization, hours	Hospital "A"		Hospital "B"	
	Nr. of patients	Nr. of lethality cases	Nr. of patients	Nr. 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 does the adjusted rates differ from the crude rates in each of these 2 hospitals? Explain these results (interpretation and conclusions).

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

Disease	Hospital "A"		Hospital "B"	
	Nr. of patients	Nr. of deaths	Nr. of patients	Nr. of deaths
Gastrointestinal	1,200	24	1,700	40
Malign tumor	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 crud 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 use of the relative values is preferred. Give an example.
2. Types of relative values: what is the difference and similarity among them. Give an example for each type.
3. Rates: particularity, rules of calculation, conditions of application and appropriate graph presentation. Give an example.
4. Proportions: particularity, rules of calculation, conditions of application and appropriate graph presentation. Give an example.
5. Ratios: particularity, rules of calculation, conditions of application and appropriate graph presentation. Give an example.

6. The Vital Statistics: definition and the main used rates.
7. What is the difference between crude and specific rates?
8. What is the difference between mortality and morbidity rates?
9. What is the difference and similarity between prevalence and incidence?
10. Under what conditions is use of the adjusted rate preferred?
11. Direct method of standardization: definition and its process steps.
12. Under what circumstance should crude, specific, and adjusted rates each be used?

CHAPTER 5. CORRELATION AND REGRESSION

Biomedical research is often related to relationship between two or more variables. For this kind of purpose is appropriate to use correlation that is able to examining the relationship between two variables. There are two basic kinds of correlational techniques:

1. *Correlation*, which is used to establish and quantify the *strength and direction* of the relationship between two variables.
2. *Regression*, which is used to express the *functional relationship* between two variables, so that the value of one variable can be *predicted* from knowledge of the other.

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 numerical variables normally distributed, symbolized by "X“ (independent or explanatory variable) and "Y" (dependent or outcome variable).

The correlation coefficient is denoted by ”r”, it is calculated using the formula:

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

The correlation coefficient is a dimensionless number (no units of measurement).

The maximum values that "r" can achieve is 1, and its minimum values is -1.

Therefore, for any given set data the coefficient of correlation is: $-1 \leq r \leq 1$

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

Coefficient of correlation direction is based on the relation noted above:

⇒ *Positive correlation* (+), when coefficient of correlation is $0 < r \leq 1$: “Y” tends to increase in magnitude as “X” increases; high values on one variable (salt intake) are associated with high values of the other variable (blood pressure).

⇒ *Negative correlation* (-), when coefficient of correlation is $-1 \leq r < 0$: “Y” decreases as “X” increases; high values on one variable (cigarette consumption) are associated with low values of the other variable (life expectancy).

The values r = 1 and r = -1 occur when there is exact linear relationship between X and Y, if r = 0 means no linear relationship or no correlation exists between two variables.

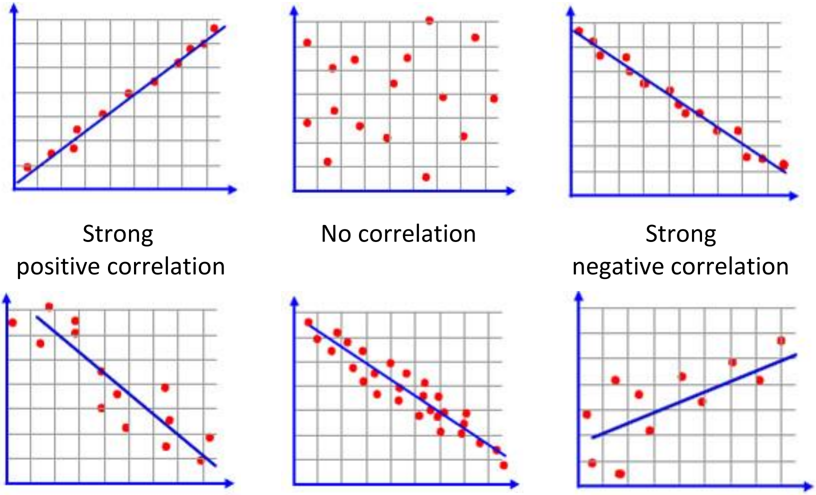
There are following scale for interpreting *the strength* of the correlation based on size of the coefficient:

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

⇒ from 0.25 to 0.50 (±): moderate correlation*;*

⇒ from 0.50 to 0.75 (±): strong correlation;

⇒ from 0.75 to 1 (±): very strong.



Strong positive correlation

No correlation

Strong negative correlation

Weak negative correlation

Moderate negative correlation

Weak positive correlation

Figure 5.1 Two-way Scatterplots and Correlation

Note:

⇒ Correlation *does not imply causation*; there is only a measure of straight-line relationship.

⇒ Inferring a causal relationship between two variables on the basis of a correlation is a common and fundamental error.

Spearman's Rank-order Correlation Coefficient

Like other parametric techniques Pearson's correlation coefficient is very sensitive to outlying values. Instead of that the Spearman's rank correlation is used when one or both of the relevant variables are ordinal (or one ordinal and one numerical) characteristics and when the numerical observations are skewed with extreme values.

The calculation of the Spearman rank correlation, symbolized as r_s involves rank ordering the values on each of the characteristics from lower to highest; the ranks are then treated, as they were the actual values themselves.

As the Pearson's correlation coefficient, the Spearman rank correlation coefficient ranges in value from -1 to 1. Values of r_s close to the extremes indicate a high degree of correlation between X and Y; values near 0 imply a lack of linear association between the two variables.

5.1.2 Coefficient of Determination

Sometimes the correlation is squared (r^2) to form a useful statistic called the *coefficient of determination.*

It is a statistical term that tells us how good one variable is at predicting another.

r^2 = 1.0 means given value of one variable can perfectly predict the value for another variable

r^2 = 0 means knowing either variable does not predict the other variable

The higher r^2 value means more correlation there is between two variables. Though coefficient of determination is denoted the common association of the factors that influence the two variables. In other words, the coefficient of determination indicates the part of total value dispersion of variable can be explained or justified by dispersion of the values the other variable. Sometimes coefficient of determination is presented in percent, being multiply by 100. In this case, the coefficient of determination r^2 expresses the proportion of the variance in one variable that is accounted for, or "explained," by the variance in the other variable.

5.1.3 Test of Significance

The correlation coefficient can be tested for significance with the help of student's test. Sometimes the correlation coefficient value may be high but it may not be found significant due to small number of observations.

⇒ Null hypothesis, $H_0 : \rho = 0$,

where, "ρ" is called the population correlation coefficient, expressing the degree of association between two continuously measured variables for a complete population of interest.

⇒ Alternative hypothesis, $H_1 : \rho \neq 0$,

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

where,

r – the sample coefficient of correlation;
n – sample size (number of observations).

⇒ If t (calculated) value > t (tabulated) with (n-2) degrees of freedom at chosen level of significance, the null hypothesis is rejected and alternative hypothesis is accepted. That means the correlation between two variables in question is significant.

The correlation coefficient can be tested for significance with confidence interval calculation, as well, which is a range of values that is likely to contain a population correlation coefficient with a certain level of confidence.

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 makes no such distinction, as two variables involved are treated symmetrically. Like correlation analyses, simple linear regression is a technique that used to explore the nature of 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 actually the same "line of best fit" to the scatter-gram as that used in calculating the correlation coefficient.

The simple linear equation is:

$$Y_x = a + bX$$

where,

Y_x – expected value of Y (dependent variable);

a – is a constant, known as the "intercept constant" because it is the point where the *Y* axis is intercepted by the regression line (in other words, the value of *Y* when *X* is zero);

b – regression coefficient, it shows the change in *Y* when *X* increases by 1 unit;

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

The regression coefficient (b) approaches are:

$b = 0$; the variable Y does not dependent on the variable X;

$b \neq 0$; the variable Y dependent on the variable X as follows:

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

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

Once the values of *a* and *b* have been established, 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,* in mL/min/kg) can be predicted from the hepatic clearance rate of indocyanine green dye (*X,* in mL/min/kg), according to the equation $Y_x = 0.30 + 1.07X$, thus enabling anesthesiologists to reduce the risk of lidocaine overdosage by testing clearance of the dye (Pagano M., Gauvreau K., 2000).

5.2.2 Multiple Linear Regression

Multiple linear regression techniques are applied when more than one continuous variable is used to predict the expected value of Y, thus increasing the overall percentage of variance in Y that can be accounted for; a multiple regression equation is therefore:

$$Y_x = a + b_1 X_1 + b_2 X_2 + \ldots b_n X_n$$

5.2.3 Logistic Regression

When studying linear regression, the response variable Y was continuous, and was assumed to fallow a normal distribution. We were concerned with predicting the *mean* value of the response variable Y corresponding to a given value for explanatory variable X (simple linear regression) or a given set of values for explanatory variables (multiple linear regression).

There are many situations in which the response of interest is dichotomous rather that continuous. Just as we estimated the mean value of the response when Y was continuous, we would like to be able to estimate the probability associated with dichotomous response (which, of course, is also its mean) for various values of an explanatory variable. To do this, we use a technique known as logistic regression.

Review Exercises

1. *Using the following data set:*

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

A group of 16 newborns Apgar score at birth data:

Nr. of observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Pregnancy term, weeks	26	33	40	36	38	29	29	29	40	39	27	27	27	33	29	30
Apgar score	6	8	10	9	9	7	7	8	10	10	8	7	7	8	7	7

2. *Using the following data set:*

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

A group of 16 newborns length and weight data:

Nr. of observations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Length, cm	56	50	49	45	48	49	43	50	47	43	51	54	47	43	49	50
Weight, kg	4.0	3.5	3.0	2.1	2.4	2.7	2.2	3.4	2.4	2.1	3.6	3.7	2.3	2.1	2.8	3.3

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 use of the correlation preferred?
2. What 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 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 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 use of the logistic regression preferred?

CHAPTER 6. INFERENTIAL STATISTICS: PROBABILITY THEORY AND HYPOTHESES TESTING INTRODUCTION

Inferential statistics are statistics, derived from sample data, that are used to make conclusions about the population parameters from which that sample was drawn. We use inferential statistics to determine whether some phenomenon observed in a sample represents an actual phenomenon in the population from which the sample was drawn.

6.1 Probability Theory

6.1.1 General Concepts

Probability plays a key role in inferential statistics. When it comes to deciding whether a result in a study is *statistically significant* (i.e., our result is not due to chance), we must rely on probability to make the determination.

Probability theory is essential to many human activities using large sets of data that need to be analyzed. The point is an experiment can be repeated many times called a trial and one or more outcomes can result from each trial.

Therefore, classical definition of probability states that: the probability of an event to occur (p) is number of cases favorable to the event (m) over the number of total outcomes possible (n):

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

Then, probability of an event not to occur (q) is defend, as follows:

$\text{q} = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - \text{p};$ or q = 1 - p; then p + q = 1

So, the sum of event probability to occur and not to occur is equal to 1, in percent – 100%. Therefore, the values of "P" lies between "zero" and 1 or 0 – 100%. In this way, the event probability to occur rise having "P" closer to 1 or to 100%, and conversely the event probability not to occur rise having "P" closer to "zero".

Two of the major representative mathematical results describing such patterns are the Law of Large Number and Central Limits Theorem.

6.1.2 Law of Large Numbers

In probability theory the *Law of large number* is a theorem that describe the results of performing the same experiment a large number of times. According to the law of large number:

- The average of the experiment results will tend to become closer to the expected values as more trials are performed.
- By approach to a specific number of trials the average of the experiment results became as closer as possible to expected values.

Only sufficiently large number of trials is able to really reproduce the regularity of the studying phenomena, to generalize from a sample to a larger population.

6.1.3 The Central Limit Theorem

The central limits theorem states the properties of the random sampling distribution, as follows:

1. The mean of the sampling distribution is identical to the population mean μ .

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

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

where,

$SE_{\bar{x}}$ – standard error of the mean; σ – population standard deviation;
n – sample size.

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

6.1.4 Using the Standard Error

As the formula shows, the standard error is dependent on the size of the sample: the larger “n” becomes, the more closely will the sample means represent the true population mean.

We use standard error to indicate the uncertainty around the estimate of the mean measurement. It tells us how well our sample data represents the whole population how much the sample mean would vary if you were to repeat a study using new samples from within a single population. Small samples with large standard deviation produce large standard errors.

The standard error is useful when we want to calculate a confidence interval which is the interval estimation based on sample statistics in order to make conclusions about population parameters.

6.2 Populations and Samples: Sampling

6.2.1 Sampling Definition

According to the probability theory we can make inference about a specified population characteristic using the information contained in a sample of the subjects.

Sampling is the enquiry that utilizes special methods taking a sample or the process selecting the group that you will actually collect data from in your research.

Reasons for sampling:

1. Save time - faster study.
2. Save money.
3. Often more accurate results.
4. Give possibility to reduce heterogeneity.
5. Give possibility to estimate the error.

6.2.2 Sampling Methods

Several methods of sampling are used in medical research. A key issue is that any method should be random to use inferential statistics (probability samples).

The sample will lead a real and valid inferences only being representative by qualitative and quantitative criteria. If it does not the conclusions made about population may be distorted. Therefore, in research it is better to use *probability samples*, in which the probability of being included in the sample is known for each subject in the population.

Inferential statistics can be used only for probability samples. For non-probability samples only, descriptive statistics can be used.

The *methods of sampling* are defended, as follows,

A. Probability sampling:

⇒ Simple Random Sampling

Every member of the population has an equal chance to be selected for the study. One way to select a simple random sample is to list and number each study unit, mix them up thoroughly, and then select units from this sampling frame until the required sample size is achieved.

Another way is to use a computer or a table of random numbers to identify the units to be included.

⇒ Systematic Sampling

Systematic Sampling can be used if a complete list of the population is available, where one of which every k^{th} item is selected; k is determined by dividing the numbers of items in the sampling frame by the desired sample size taking a Systematic Random Sample.

⇒ Stratified Sampling

Population is divided into mutually exclusive groups (stratum) such as age groups and random samples are drawn from each group (stratum) taking a Stratified Random Sample.

⇒ Cluster Sampling

Population is divided into mutually exclusive groups (blocks) such as age groups and random samples are drawn from each group (blocks) taking a Cluster random Sample. Clusters are commonly based on geographic areas.

B. Nonprobability sampling:

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

⇒ Convenience Sampling – the researcher selects the easiest population members from which obtain the information.

⇒ Quota Sampling – the researcher finds and interviews a prescribed number of people in each of several categories.

6.3 Estimation and Hypothesis Testing

Estimation and hypothesis testing are the big point of Inferential Statistics that enable the researcher to draw conclusions about data and the relationships between variables.

Estimation is the sample statistics applying process in order to make conclusions about population parameters using the information contained in a sample of observations. For make estimations the researcher uses the principles of hypotheses testing.

Two methods of estimation are commonly used. The first is called *point estimation* which involves using the sample data to calculate a single number to estimate the parameter of interest. A point estimate does not provide any information about the variability of the estimator: it is not known how close the sample mean ($\bar{x}$) is to population mean (μ) in any given situation. Consequently, a second method of estimation, known as *interval estimation*, is often preferred. The main difference between point and interval estimation is the values that are used. Point estimation uses a single value, while interval estimation uses a range of reasonable values that are intended to contain the parameter of interest – the population mean (μ), in this case, with a certain degree of confidence in order to infer information about the population. This range of values is called a *confidence interval* (CI).

Statistical hypothesis testing involves stating null hypothesis and an alternative hypothesis and then doing a statistical test to see which hypotheses should be concluded. The main meaning of hypothesis testing is to disprove the null hypothesis and accept the alternative.

Hypothesis testing is a mathematical formula that generates a test statistic from data that is then compared to a table or performed by computer software to generate a p-value (*point estimation*) and/or a confidence interval (*interval estimation*).

6.4 Confidence Intervals

Confidence interval is an interval estimate of a population parameter. Confidence interval can be computed for any population parameter being represented by: mean, proportion, relative risk, odds ratio, and correlation, as well as for difference between two means, two proportions, *etc.*

The ends of the confidence interval are called *confidence limits.*

Confidence intervals (CI) are defined as the values interval determined by sample mean and standard error where it is expected to find population mean,

CI= $\bar{x} \pm SE_{\bar{x}}$; *where:* $\bar{x}$ – sample mean; $SE_{\bar{x}}$ – standard error

According to the rules of normal distribution, the probability or degree of confidence to find the population mean in that interval is:

68.26% → ($\bar{x}$- SE) > CI < ($\bar{x}$+ SE)

94.95% → ($\bar{x}$ -2SE) > CI < ($\bar{x}$+ 2SE)

99.73% → ($\bar{x}$- 3SE) > CI < ($\bar{x}$+ 3SE)

Table 6.1

Confidence interval interpretation

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

6.5 Hypothesis Testing: Basic Theoretical Concepts

6.5.1 Hypothesis Definition

Hypothesis is based on observation an educated guess, assumption or idea about a phenomenon interesting to be study.

The main function of hypotheses is to focus the research bringing clarity, specificity and objectivity.

6.5.2 Hypotheses Types

⇒ *Null hypothesis (H_0)* – states that there is no significance relation or difference between variables in question (e.g., two population means will not differ),

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

⇒ *Alternative hypothesis (H_1)* – is a second statement that contradicts the null hypothesis and states that there are differences between the groups (two population means will differ).

- *Non-directional or Two-tailed* alternative– states that difference between variables exists without to specify which value will be larger: it is tested by Two-Sided Test / Two-Sided CI.

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

- *Directional* or *One-tailed* alternative – states the expected direction of the difference specifying which value will be larger: it is tested by One-Sided Test / One-Sided CI.

  H_1: $\bar{x}_1 > \bar{x}_2$ *or* H_1: $\bar{x}_1 < \bar{x}_2$

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

6.5.3 Types of Error

Two types of error may occur in hypothesis testing:

⇒ *Type I Error*

Is rejecting the null hypothesis when it is really true.

The probability to make type I error (α) it is also known as a *rejection error*: concluding that the observed group difference was true effect, when in fact it was due to chance or systematic error. Alpha error (α) Accepts the false Alternative Hypothesis (the rule of "AAA")

- As convicting the innocent man
- "False positive" error: a positive conclusion has been reached about a hypothesis that is actually false.

$\Rightarrow$ *Type II Error*

Is do not reject the null hypothesis when it is false

The probability to make type 2 error (β) it is also known as an *acceptance erro*r: concluding that the observed group difference was due to a chance or random error, when in fact it was a true effect. Beta Error (β) Accepts the false Null Hypothesis (the rule of "BEAN")

- As absolving the guilty man
- "False negative" error: a negative conclusion has been drawn about a hypothesis that is actually true.

6.5.4 Power of the Study

Power of the study is the ability of a study to detect a true difference: probability of rejecting the null hypothesis when it is really false or concluding that the alternative hypothesis is true when it is really true. In other words, it is probability to avoid a type II error.

Power is defined as: (1- β) or (1 – a type II Error).

A study is required to have a power of 0.8 (or a β of 0.2) to be acceptable – in other words, a study that has a less than 80% chance of detecting a false null hypothesis is generally considered to be unacceptable. Increasing the sample size is the most practical and important way of increasing the power of a statistical test.

6.5.5 Confidence Level

Confidence level is the ability of a study does not detect a false difference: probability to accept the null hypothesis when it is really true.

Confidence level is defined as (1- α) or (1 – a type I Error).

6.5.6 Significance Level

Significance level (α) of a test is the probability that the test statistic will reject the null hypothesis when it is really true (concluding there is a difference when there is not). In order do not reject by mistake null hypothesis (H_0) the significance level should be enough small (0.05; 0.01; 0.001).

The significance level (α) is a probability to make type I error you set before calculating the statistical test.

Note that statistical significance does not imply clinical or scientific significance; the test result could actually have little practical consequence.

6.5.7 p-value

The *p-value* is a concept related to significance level (α). The p-value is a probability of obtaining the results that the null hypothesis is true or to occur results by chance. If this probability is enough small then null hypothesis is rejected. The p-value is a probability to make type I error you find after calculating the statistical test; if the p-value is less than significance level (α), the null hypothesis is rejected and you can conclude that the difference between two means is statistically significant and not due to chance.

6.6 Hypothesis Testing Process General Approaches

We again concentrate on drawing some conclusion about a population parameter, using the information contained in a sample of observations. As we saw in the preceding points, one approach is to construct a confidence interval for population mean (μ); another is to conduct a *statistical test.*

To perform a statistical test, we must take into account the general approaches of the hypothesis testing process, as follows:

1. Formulate the hypotheses: null hypothesis (H_0) and alternative hypothesis (H_1)
2. Decide about the appropriate statistical test.
3. Select the level of significance (α) for the statistical test. There is a probability to reject the true null hypothesis (H_0). In order do not reject by mistake H_0 the significance level should be enough small (0.05; 0.01; 0.001).
4. Determine the value which the statistical test must attain for to be declared significant.
5. Perform the calculation.
6. State the conclusions.

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 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 a skewed to the left distribution. The sampling distribution of the serum cholesterol mean 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. Mean systolic blood pressure in a group of 500 individuals selected from a total of 100,000 individuals of the locality A is 130 mmHg and standard deviation is 15 mmHg. Calculate: standard error, confidence interval and confidence limits (for 95%) 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 is 25 mm Hg.
   1. Calculate a 95% confidence interval for the true mean systolic blood pressure among the population of 60-year-old men with diabetes mellitus.
   2. 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 statistical inference?
2. What is the meaning of the probability theory?
3. Explain the statements of the law of large number and their applications in research.
4. Explain the statements of the central limit theorem and their applications in research.
5. Give the definition for the sampling.
6. What is important that a sample drawn from a population be random?

7. When might you prefer to use systematic sampling rather than simple random sampling?
8. When would you prefer stratified sampling?
9. When would you prefer cluster sampling?
10. What is purpose of a test of hypothesis?
11. Explain the difference between point and interval estimation.
12. Give the definition of hypothesis and state its types.
13. What are the confidence interval and limits?
14. What are the factors that affect the length of a confidence interval for a mean?
15. Describe the two types of errors that can be made when you conduct a test of hypothesis.
16. Explain the analogy between type I and type II errors in a test of hypothesis and the false positive and false negative results that occur in diagnostic testing.
17. What is a power of the study? What does the power mean in words?
18. What is the confidence level? What does the confidence level mean in words?
19. What is the significance level? What does the significance level mean in words?
20. What is a p-value? What does the P-value mean in words?

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

7.1 Parametric and Nonparametric Test Assumptions

There are parametric methods of hypothesis testing appropriate to parametric data (interval and ratio scale) and nonparametric methods hypothesis testing appropriate to nonparametric data (ordinal or nominal scale).

Making an appropriate choice for statistical methods depends mainly on:

⇒ Data measures type (numerical, nominal or ordinal);
⇒ Independent or related (pared) samples;
⇒ Sample size (n>30 or n<30);
⇒ Number of groups (one, two or more);
⇒ Directional or non-directional alternative hypotheses;
⇒ Data distribution type (normal or skewed);
⇒ Homogeneity of variance.

If your data don't respect these conditions, you should rather apply nonparametric tests as indicated in *Table 7.1*

Table 7.1

Parametric *vs.* Nonparametric test

STATISTICAL TEST	
Parametric	**Nonparametric**
One sample t-test	Sign Test/Wilcoxon Signed Rank Test
Pared t-test	Sign Test/Wilcoxon Signed Rank Test
Two-sample t-test	Mann Whitney U-test/ Wilcoxon Signed Rank Test
ANOVA	Kruskal Wallis Test

Source: *Chan Y.N. Biostatistics 102: Quantitative Data – Parametric and Non-Parametric Tests*, Singapore Med J 2003 Vol 44(8) : 391-396.

7.2 Parametric Test: Comparison of Two Means

7.2.1 Conditions to use t-test

In order to compare two means, we can use t-test under the following conditions:

1. The data fallow the normal distribution;
2. Homogeneity of variance (the population variance is equal);
3. Numerical measures of data (interval or ratio scale);
4. Having no more than 2 groups taking into account:
   - ⇒ Type t tests for one sample: the comparation between a sample mean and a population mean
   - ⇒ Type t tests for two independent samples: the comparation between means of two independent samples
   - ⇒ Type t tests for two pared samples: the comparation between two repeated measurements of the same group

7.2.2 Steps hypothesis testing about the mean

To test a hypothesis about the mean, the steps are the follows:

1. State the null and alternative hypotheses, H_0 and H_1.
2. Select the decision criterion α (or “level of significance”).
3. Establish the critical values.
4. Draw a random sample from the population, and calculate the mean of that sample.
5. Calculate the standard deviation (S) and estimated standard error of the sample ($\mathrm{SE}_{\bar{x}}$).
6. Calculate the value of the test statistic t that corresponds to the mean of the sample (t_{calc}).
7. Compare the calculated value of t with the critical values of t, and then accept or reject the null hypothesis.

7.2.3 The critical value establishment

The critical value ($t_{critical}$) is computed based on the given significance α and the type of probability distribution of the idealized model. The critical values are established by referring to a table of t scores (see Appendix A).

- ⇒ If the numbers of observations “n” >120 then the critical value is already known: t=1.96 ($\alpha = 0.05$); t=2.58 ($\alpha = 0.01$); t=3.29 ($\alpha = 0.001$).
- ⇒ If the numbers of observations “n” < 120 then the critical value is taken from the special table (see Appendix A), according their significance level (α) in area 1 or 2 tails, and degrees of freedom (df), which are defined as:

$$df= (n_1+n_2) -2$$

where:

n_1 – number of observations in the sample 1
n_2 – number of observations in the sample 2

7.2.4 The test statistics value calculation: Two-sample t-test

The t statistic (t_{calc}) calculation for comparison of means in two independent groups is computed according the following equation,

$$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{{SE_1}^2 + {SE_2}^2}}$$

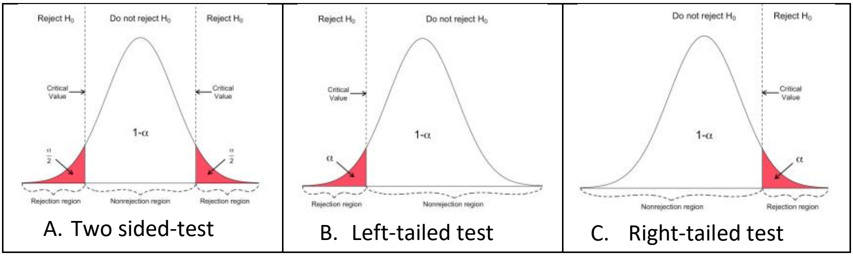
where:

$\bar{x}_1$: the mean for sample 1
$\bar{x}_2$: the mean for sample 2
SE_1 : standard error of the mean for sample 1
SE_2 : standard error of the mean for sample 2

7.2.5 The critical value approaches

The critical value divides the area under probability distribution curve in rejection(s) and nonrejection region. The following three diagrams in the *Figure 7.1* show a right tailed test, left tailed tests, and a two-sided test. The idealized model in the figures, and thus H_0, is described by a bell-shaped normal probability curve.

If the sample mean falls in the area of rejection, the null hypothesis is rejected, and the alternative hypothesis is accepted.



A. Two sided-test B. Left-tailed test C. Right-tailed test

Figure 7.1 The critical value approaches to hypothesis testing

In a two-sided test the null hypothesis is rejected if the test statistics (t_{calc}) is either too small or too large. Thus, the rejection region for such a test consists of two parts: one on the left and one on the right, as displayed in the *Figure 7.1 (A).*

For a left-tailed test, the null hypothesis is rejected if the test statistic is too small. Thus, the rejection region for such a test consists of one part, which is left from the center, as displayed in the *Figure 7.1 (B).*

For a right-tailed test, the null hypothesis is rejected if the test statistic is too large. Thus, the rejection region for such a test consists of one part, which is right from the center, as displayed in the *Figure 7.1 (C).*

By applying the critical value approaches, it is determined, whether or not the observed test statistic (t_{calc}) is more extreme than a defined critical value. Therefore, the observed test statistic (calculated on the basis of sample data) is compared to the critical value, which is some kind of cutoff value. If the test statistic (t_{calc}) 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,

- ⇒ If t_{calc} is larger than critical value the null hypothesis is rejected such the test result is said the difference between the compared means to be statistically significant.
- ⇒ If t_{calc} is less then critical value the null hypothesis is not rejected such the test result is said the difference between the compared means to be not statistically significant.

7.3 Nonparametric Tests

The previous sections stated the main conditions for testing hypotheses about the means, using t-tests. There are other statistical tests that do not share the requirements for parametric tests. They are known as nonparametric tests and do not assume that the population is normally distributed, so they are called distributions free tests. Nonparametric tests are used to test nominal, ordinal or skewed numerical data.

Such tests, however, have the disadvantages that they are generally less powerful than parametric tests.

7.3.1 Chi-Square

The most important nonparametric test is the chi-square (χ^2) test, which is used for testing hypotheses about nominal scale data. Chi-square is basically a test of proportions, telling us whether the proportions of observations falling in different categories differ significantly from those that would be expected by chance.

As with other tests, chi-square involves calculating the test statistic (χ^2_{calc}) according to a standard formula and comparing it with the critical value according their degrees of freedom (df) and appropriate for the selected level of significance (α) from chi-square tables distribution (see Appendix B).

The degrees of freedom for the chi-square are calculated using the following formula:

df = (r-1) (c-1)

where:

r – number of rows;
c – number of columns.

If the observed chi-square test statistic (χ^2_{calc}) is greater than the critical value, the null hypothesis can be rejected.

Review Exercises

Using the fallowing data sets:

1. Compute the mean for the 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 t-test and state the conclusions about means difference statistical significance.

H_o: the sample means difference is not statistically significant
H_1: the sample means difference is statistically significant

$p > 0.05$ ⇛ H_0 accepted
$p < 0.05$ ⇛ H_0 rejected

Data set 1

Two samples (n=15 each) the cholesterol blood level results are:

Observations number		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Blood Cholesterol, mg/dl	Gr.1	168	258	228	247	156	172	165	210	264	220	258	200	195	245	189
	Gr.2	136	148	125	121	157	148	116	140	161	122	128	122	137	139	128

Data set 2

Two samples (n=12 each) the systolic blood pressure results are:

Observations number		**1**	**2**	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
SBP, mm/Hg	Gr.1	130	130	120	110	90	120	125	115	135	140	120
	Gr.2	170	175	160	170	170	190	185	185	170	160	190

Review Questions

1. What are the parametric methods of hypotheses testing?
2. What are the nonparametric methods of hypotheses testing?
3. When would you prefer to apply hypotheses testing parametric methods?
4. When would you prefer to apply hypotheses testing nonparametric methods?
5. State the main considerations parametric methods vs. nonparametric methods.
6. When should you use the two-sample t-test?

CHAPTER 8. THE RESEARCH METHODOLOGY INTRODUCTION

8.1 Research Definition, Characteristics and Types

> *Research* is a structured activity utilizing appropriate scientific methodology to solve problems and create new knowledge that is generally applicable.

Research is a process of collecting, analyzing and interpreting information to answer question having the following *characteristics*:

1. *Validity:* this concept means that correct procedures have been applied to find answer to a question.
2. *Unbiased and objective:* each step of the research process and each conclusion have been taken to the best of your ability and without introducing your own interest.

   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 analyzed.

   Generally, three types of bias are distinguished: confounding, selection bias and information bias. Confounding is distinguished from selection and information bias in that when it appears: collecting, analyses or interpretation data. Thus, we need to be extra careful at the design and execution every stage of scientific study.

   Confounding is a bias that results when the risk factors being studied is so mixed up with other possible risk factors that is single effect is very difficult to distinguish.

   Selection bias is a distortion that results from how the study is designed at every step from formulating of research problem till generalization data plane:

(For example: Nonresponse bias, Exclusion bias, sample volume bias, etc.)

Information bias is a distortion almost from the period of information registration:

- ⇒ Random error – this error is part of human being occurring because researcher is not being sufficiently careful. This kind of error doesn't affect so heavily the final results of the study as systematic error does.
- ⇒ Systematic error – that is due to systematic measurement error or misclassification of subjects by interviewers (not sufficiently instructed). This kind of error heavily affects the final results of the study.

(For example: interviewing bias, recall bias, reporting bias, *etc*.).

3. *Repeatability (Fidelity):* this concept implies that whatever you conclude on basis of your findings can be reproduced again (verified) by you and others.
4. *Comparability:* the process of investigation must be foolproof and free from drawbacks. The research conclusions and results must be able to withstand critical and comparison scrutiny.
5. *Systematic:* all undertaken investigation procedure must follow a certain logical sequence. The different steps cannot be taken in a hazard way. Some procedures must follow others.
6. *Relevance:* There are two key types of relevance,
   - ⇒ *scientific* relevance, where a study increases our understanding of a disease or a process;
   - ⇒ *societal* relevance, where society directly benefits as a result of this increased understanding.

As shown in *Table 8.1,* research can be classified from three perspectives:

- Application of research study
- Inquiry mode employed
- Objectives in undertaking the research

Table 8.1
The Types of the Research by perspectives

TYPES OF RESEARCH		
Application of research study	***Inquiry mode employed***	***Objectives in undertaking the research***
1. pure	1. quantitative	1. historical
2. applied	2. qualitative	2. descriptive
		3. correlational
		4. experimental
		5. exploratory

From the point of view of application there are two broad categories of research:

⇒ *Pure research* involves developing and testing theories and hypotheses that are intellectually challenging to the researcher but may or not have practical application at the present time.

⇒ *Applied research* is done to solve specific, practical questions for policy formulation and understanding of phenomena.

From the process adopted to find answer to research questions the two approaches are:

⇒ *Quantitative research* determines the extent of a problem, issue or phenomenon by quantifying the variation. The main question is: how many?

⇒ *Qualitative research* is more appropriate to explore the nature of a problem, issue or phenomenon without quantifying it.

Main objective is to describe the variation in a phenomenon having the question: how is?

Both approaches have their place in research. Both have their strengths and weaknesses.

From the point of view of objectives undertaking the research can be classified:

⇒ *Historical research* has the purpose to arrive at conclusions concerning trends, causes or effects of past occurrences. This may help in examining present events and anticipating future events. The data are not gathered by administrating instruments to individuals. They are collected from original documents or by interviewing the eyewitness. The data thus collected are subjected to scientific analyses to assess its authenticity and accuracy.

⇒ *Descriptive research* describes systematically a situation, problem, phenomenon, etc. Descriptive research deals with collecting data and answering questions concerning the current status of the subject of study. It concerns with determining the current practices status or features of situations. Another aspect of descriptive research is that data collection is either done through asking questions from individuals in the situation (through questionnaires or interviews) or by observation.

⇒ *Correlational research* attempts to discover or establish the existence of a relationship between two or more aspects of a situation. Descriptive and historical researches provide a picture of events that are currently happening or have occurred in the past. Researchers often want to go beyond mere description and begin to figure out the relationship *by observation* the phenomena and to determine the degree of that relationship as well. The relationship thus determined could be used for making predictions and hypotheses testing. Correlational researches are studies that are often conducted to test the reliability and predictive validity of instruments used for division making concerning selection of individuals for the likely success in a course of the study.

⇒ *Experimental research* help establish the presence of a relationship and causality not by observation only (as correlational studies) but – in base of experiment providing: the investigator's actions are involved in phenomena process.

⇒ *Exploratory research* is undertaken to explore an area where little is known in order to investigate the possibilities of undertaking a particular research study (feasibility study/pilot study)

8.2 The Steps of Research Process Contents

For a research process there are two important decisions to make:

1. What you want to find out about.
2. How to go about finding their answers.

There are practical steps through which you must pass in your research process in order to find answers to your research questions. The way to finding answers to your research questions constitutes research methodology. At each step in the research process, you are required to choose from multiplicity of methods and techniques of research methodology that will help to best achievement of research objectives. To provide systematic principle of research it is necessary to follow carefully the study process steps, as follows:

1. Formulating the research problem;
2. Literature review;
3. Developing the aim and objectives of the study;
4. Preparing the research design;
5. Collecting the data;

6. Data analyses;
7. Generalization and interpretation (draw conclusions and re-commendations);
8. Data presentation (writing a report and oral public presentation).

8.3 Formulating the Research Problem

It is the first and more important step in the research process. There are many considerations in a research problem selection:

⇒ *Interest* – one should make a selection of a topic of great interest to sustain the required motivation

⇒ *Magnitude* – it is extremely important to select a topic that can be managed within the time and disposal resource.

⇒ *Relevance* – the future study has to add to the existing body of knowledge.

⇒ *Availability of data*

⇒ *Ethical issues*

The process of formulating research problem has a few steps. Working through these steps requires an appropriate level of knowledge in the broad subject area within the study is to be undertaken. Without such knowledge it is difficult to clearly and adequately formulate the research problem.

The way and quality of research problem formulating is determined by every step that follows:

1. *Identify* a broad field of interest;
2. *Dissect* the broad field into sub areas;
3. *Select* what sub areas is of most interest;
4. *Raise* research questions;
5. *Establish* of the hypotheses;
6. Double *check*.

8.4 Reviewing the Literature

This step is essential preliminary task in order to find out about available body of knowledge in your interest area. In addition, literature review is integral part of entire research process and makes valuable contribution to every operational step. Reviewing literature can be time consuming, daunting and frustrating, but is also rewarding.

Its functions are:

1. Bring clarity and focus to your research problem
2. Improve the study methodology
3. Broaden investigator's knowledge
4. Contextualize investigator's findings

Procedures for reviewing literature are:

⇒ Search for existing literature in your area of the study

⇒ Review the selected literature

⇒ Develop a theoretical framework

⇒ Develop a practical framework

⇒ *Search for existing literature*

To effectively search for literature in your field of interest it is imperative to have in mind at least some idea of broad subject field and of the investigate problem, in order to set parameters for your search. Actually, there are many on-line medical databases:

MedLine (OVID) – http://gateway.ovid.com
Mdconsult – http:// home.mdconsult.com
HINARI – http:// www.who.int/hinari/usinghinari/en/index.html
PubMed – http:// www.ncbi.nlm.gov./pubmed
And many others…

For an academic paper, you should use books and articles as well as Web sites that collect important information to your topic.

During your literature review investigator can note the detailed description of methodologies and instruments used in previous research. In addition, this detailed description is the best way to determine possible approaches appropriate to their needs, such as sampling techniques, interviewing, data collection methods, and interventions.

Cite what you find using a standard format. The bibliography of literature reviewing should give a clear, complete description of the sources that were used while preparing the report. It is an alphabetical list as per the author's surname.

There are three *systems to citing references* used more frequently:

⇒ The *Harvard System* (the oldest) – "author-date" system

⇒ The *Vancouver System* (launched in 1978 Canada) – a variant of sequential numerical system

⇒ *Letter-number systems* = a hybrid system

As Sir Isaac Newton commented: "If I see further, it is because I stand on the shoulders of giants". In other words, every researcher relies on information created by others who came before. In academic writing, researchers document the information they rely on, carefully giving credit to the authors of original information that supports their own writing.

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

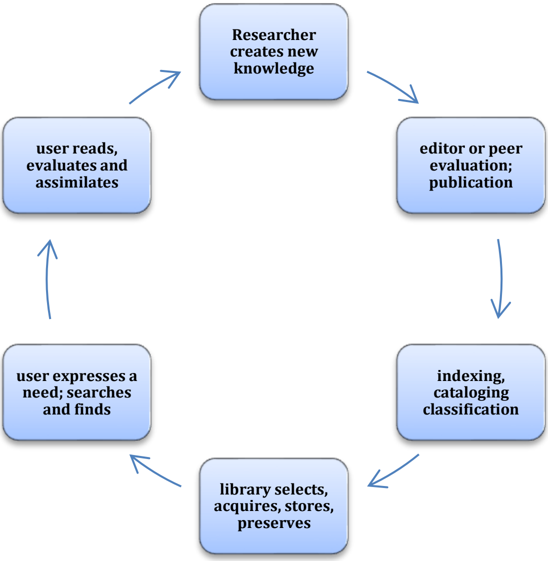

Researcher creates new knowledge
editor or peer evaluation; publication
indexing, cataloging classification
library selects, acquires, stores, preserves
user expresses a need; searches and finds
user reads, evaluates and assimilates

Figure 8.1 Information Cycle

Bibliographic documentation is a permanent process of professional update.

8.5 Formulation of the Aim and Objectives

Aim is the goal that is set out to attain in the study being an overall statement of the study.

Objectives are the required tasks to be fulfilled for aim's accomplishment.

It is extremely important to formulate them clearly and specifically.

The objectives should be numerically listed. Each objective should contain only one aspect of the study. When formulating objectives, it is preferably to use 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 Research Design

8.6.1 Research Design Definition

> *Research design* is the conceptual structure within which research would be conducted.

8.6.2 Research Design Steps

The function of research design is to provide for the collection of relevant information with minimal expenditure of effort, time and money.

The preparation of research design appropriate for a particular research problem, involves the consideration of the following steps:

⇒ To determine *sample design*

⇒ To elaborate *tools* for data collection

⇒ To adopt *study design*

8.6.3 Determining Sample Design

Researcher usually draws conclusions about large groups by taking a sample.

Designing the sample calls for three 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

The construction of a research tool for data collection is one of the most important aspect of a research protocol. Because anything you say by way of findings or conclusions is based up the type of collected information, and the collected data is entirely dependent upon the question from questionnaire.

Guidelines to construct a questionnaire:

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

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

Step III: Take each research question listed in step I and each objective listed in step III and list the information required to answer it.

Step IV: Formulate questions to obtain this information.

A questionnaire consists of a set of questions presented for answers. Many ways to ask questions exist, it is important to have questions clear and obtained desired information. The questionnaire should be developed and tested carefully before being used on a large scale.

There are three basic types of question structure:

⇒ Closed

⇒ Open-ended

⇒ Combination of both

Closed question structure includes all possible answers-prewritten response categories, and respondents are asked to choose among them. (e.g. multiple-choice questions, scale questions). That type of questions used to generate statistics in quantitative research and their payoff is the ease with which the answers can be analyzed and reported.

Open-ended questions are ones that permit the subject to respond in his or her own words. A primary advantage of open-ended questions is the capability to report some of the prototypic answers using a subject's own words. Questionnaire doesn't contain boxes to tick but instead leaves a blank section for the respondent's answer. As there are not standard answers to these questions data analysis is more complex.

Combination of both questionnaire structure begins with a series of closed questions and finish with a section of open-ended questions or more detailed response.

Table 8.2

Open versus closed questions

	Use open	Use closed
Purpose	Actual words or quotes	Most common answers
Respondents	Capable to provide answers	Willing to answer only if easy and quick
Asking the questions	Choices are unknown	Choices can be anticipated
Analyzing results	Content analyses; time consuming	Counting or scoring
Reporting results	Individual or grouped responses	Statistical data

Source: Beth Dawson, Robert G. Trapp *Basic and Clinical Biostatistics* 2004:

Most surveys use self-administrated questionnaires – in person or via mail, or email, or interviews – again in person or over the phone. Advantages and dis advantages exist for each method, some of which are illustrated in the *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: Beth Dawson, Robert G. Trapp *Basic and Clinical Biostatistics* 2004:

General structure of questionnaire:

⇒ Title

⇒ Instructions

⇒ General information about respondent

⇒ The questions

⇒ Thank you

8.6.5 Study Design Classification

There are several different schemes of classifying methods of study design. One of them adjusted is indicated in the followed *Table 8.4.* This classification divides studies design into those in which the subjects were observed, called *observational studies*, those in which some intervention was performed, called *experimental studies*, and those in which *primary studies* are pre-appraised, *or* filtered (*secondary studies*).

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	– Narrative reviews – Systematic reviews – Meta-analyses

Each type of design study simply represents a different way of harvesting information. The selection of one design over another depends on the particular research questions, concerns about validity, and practical an ethical consideration.

⇒ *Observational* studies provide information on exposures that occur in natural settings, and they are not limited to preventions and treatments. Furthermore, they do not suffer from the ethical issues of experimental studies.

The two principal types of observational studies are cohort and case-control studies. A classical cohort study examines one or more health effects of exposure to a single agent. Subjects are defined according to their exposure status and followed over the time to determine incidence of health outcomes.

In contrast, a classical case-control study examines a single disease in relation to exposure to one or more agents. Cases who have the disease of interest and controls that are a sample of the population that produced the cases are defined and enrolled. The purpose of the control group is to provide information on the exposure distribution in the population that rise to the cases. Investigators obtain and compare exposure of cases as well as controls.

Both studies cohort and case-control are analytical studies it means they are able to test hypotheses for establish causality. Cohort and case-control studies generally involve an extend period of time defined by the point when the study begins. For this reason, both are called longitudinal studies. The major difference between them is the direction of the inquiry: Cohort study is forward looking, from a risk factor to outcomes, whereas case-control study is backward looking, from an outcome to risk factors. In this context there is a cross-sectional study analyses data collected on a group of subjects at one time. The time relationship among different observational study designs are displayed in *Figure 8.2*.

⇒ *Experimental studies* involve an intervention, such as a drug, a procedure, or a treatment. Both observational and experimental studies may involve animals or objects, but most studies in medicine involve people. Experimental studies, which provide interventions are often infeasible because of difficulties enrolling participants, high costs, and big ethical issues, most research, are conducted using observational studies.

⇒ *Systematic Review* is a secondary study consisting of a summary of the clinical literature according the special methodology. A systematic review is a critical assessment and evaluation of all primary research studies that address a particular clinical issue. The researchers use an organized method of locating, assembling, and evaluating a body of literature on a particular topic using a set of specific criteria. A systematic review typically includes a description of the findings of the collection of research studies. The main objective of a systematic review is to formulate a well-defined research question and use qualitative and quantitative methods to analyze all the available evidence attempting to answer the question. In contrast, *narrative reviews* can address one or more questions with a much broader scope. Narrative reviews often do not meet important criteria to help mitigate bias – frequently they lack explicit criteria for article selection and frequently there is no evaluation of selected articles for validity, as examples.

⇒ *Meta-Analysis* is a secondary study consisting of studies using a quantitative method of combining the results of independent primary studies (usually drawn from the published literature) and synthesizing summaries and conclusions which may be used to evaluate therapeutic effectiveness, plan new studies, *etc*. It is often an overview of clinical trials.

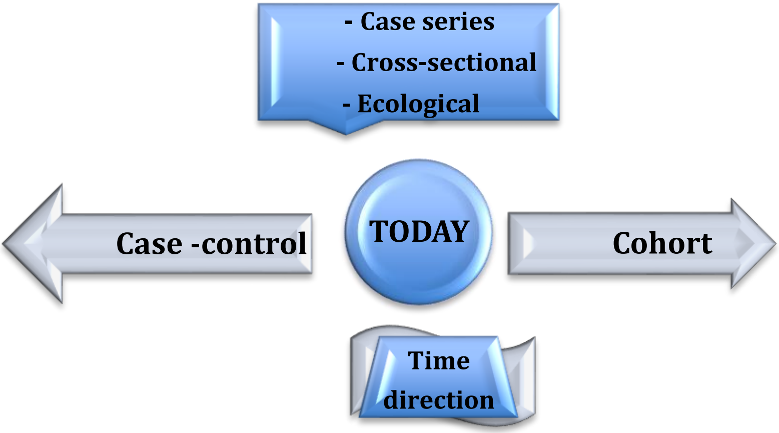


Figure 8.2 Time Relationship among different Observational Study Design

Additional observational study designs include cross-sectional studies that examine the relationship between a disease and an exposure among individuals in a defined population at a point in time.

Thus, it takes a snapshot of a population and measures the exposure prevalence in relation to the disease prevalence. Cross-sectional study is a descriptive study it means it is able only to suggest hypotheses for the future cohort or case-series study. Because cross-sectional study is not able to test hypotheses for establishing causality as case-control and cohort studies do. The goal of all study designs is to determine the relationship between an exposure and a disease with validity and precision using minimum of resources.

8.7 Study Design Evidence Strength

The evidence pyramid shows designs from the strongest to the weakest (top to bottom) and the frequencies of such studies (with). As you can see from *Figure 8.3*, the studies with strong evidence are in short supply.

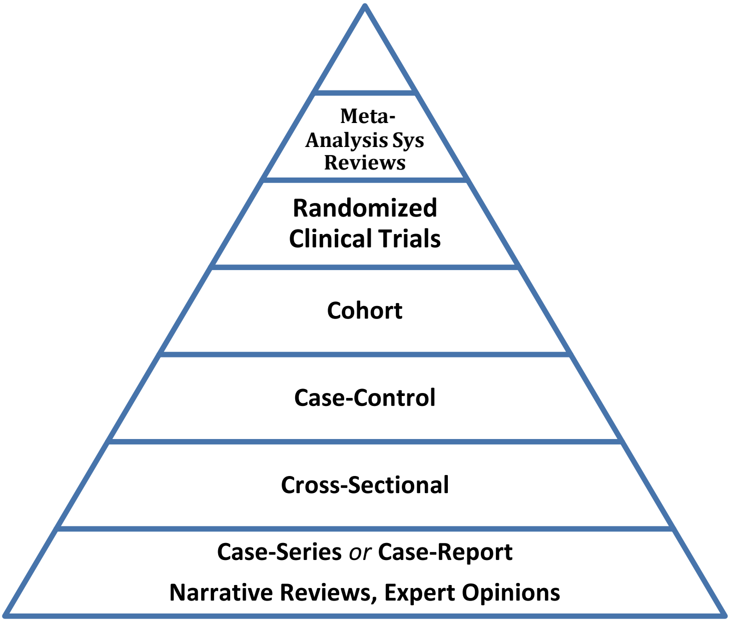

Meta-Analysis Sys Reviews
Randomized Clinical Trials
Cohort
Case-Control
Cross-Sectional
Case-Series *or* Case-Report
Narrative Reviews, Expert Opinions

Figure 8.3 Research Study Design Evidence Strength Pyramid

Descriptive observational research, narrative review and expert opinion reside at the bottom of the pyramid while secondary studies (Systematic Review and Mete-Analysis) with the highest internal validity, characterized by a high degree of quantitative analysis, review, analysis, and stringent scientific methodology, are at the top of the pyramid.

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 course, assuming that some number of hours is required each year. In addition, you want to learn topics they would like to have included in the future programs. How would you select the sample of physicians to include in your study survey? Explain your choice.
   a) All physicians who attend last year’s programs
   b) All physicians who attend the two upcoming programs
   c) A random sample of physicians who attend last year’s programs
   d) A random sample of physicians obtained from a list maintained by the state medical society
   e) A random sample of physicians in each county 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 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. Give an example for each.
6. Research types classification from the point of view of objectives: types and their meaning. Give an example for each.
7. Steps in research process: their contents and particularities.
8. Formulating the research problem: main function and selection considerations.
9. Steps in formulation of a research problem: their contents and particularities. Give an example.
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. General design of questionnaire.
17. Study design classification.
18. Observational versus experimental study design: the meaning and particularity. Advantages and disadvantages. Give an example.

19. Which type of study design is best depending on research questions?
    a) Therapy question
    b) Diagnosis/screening
    c) Prognosis
    d) Occurrence
    e) Causation
20. State the main difference between the following study design: observational descriptive and observational analytic. Give an example.
21. State the main difference between the following study design: case-control and cohort study. Give an example.
22. State the main difference between the following study design: case-series and case-control. Give an example.
23. Classify study designs according with strong evidence. Give an example.

CHAPTER 9. OBSERVATIONAL DESCRIPTIVE STUDIES

9.1 Case-series / Case-report study

It is the simplest design in which the author describes some interesting or intriguing observations that occurred from a small number of patients (case-series study) or even one patient with unusual situation (case-report study). When certain characteristics of a group (or series) of patients (or cases) are described in a published report, the result is called *case-series studies*. This type of study design lead to the generation of hypotheses that are subsequently investigated in a cross-sectional, case-control or cohort study.

Used for:

⇒ Recognition of new disease/outcome.

⇒ Formulation of hypotheses.

Advantages:

1. They are easy to write.
2. The observations can be extremely useful to other investigators.

Disadvantages:

1. They are susceptible to many biases.
2. They are not able for conclusive decisions.

9.2 Cross-sectional study

This type of observational study analyzes data collected on a group of subjects at one time rather than over a period of time, they are called also *Transversal studies*. Cross- sectional study is designed to determine "What is happening?" right now. Subject are selected and information is obtained in a short period of time (point of time) as is displayed in *Figure 9.1.*

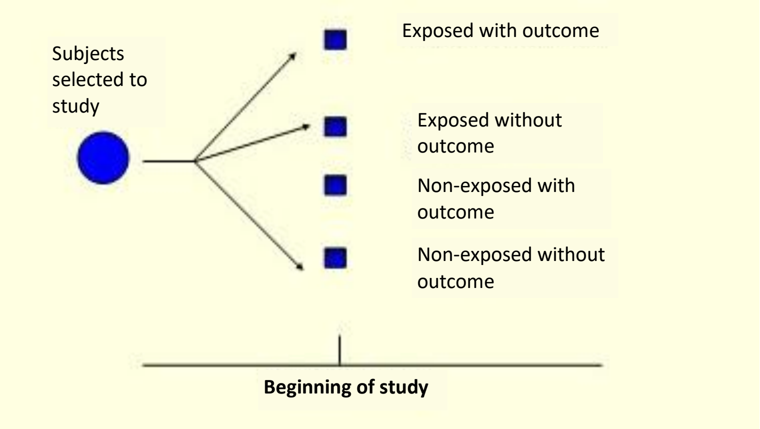


Figure 9.1 Flowchart of cross-sectional study design

Cross-sectional study is used for measure prevalence of a disease and look at potential risk factors or cause. Because cross-sectional studies examine relationship between exposure and diseases 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.

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

As a *statistical procedure for cross-sectional study data analysis* is used: calculation of all possible proportion, rates (adjusted, as well) and ratios. In the same time, is appropriate to compute confidence limits for proportion or means, and correlations. In addition, could be applied parametric t test, analysis of variance, chi-square and other non-parametric tests, and regression, including logistic regression.

Advantages:

1. They are useful to know the burden of a disease in a group – prevalence rate can be obtained.

2. Cheap and fast.
3. Useful to evaluate diagnostic procedure.
4. To study common risk factors.
5. To study common outcomes.

Disadvantages:

1. Population little willing to collaborate.
2. Doesn’t tell the flow of events.
3. Only shows association between factor and disease studied, not causality.
4. It is not useful to search causes of the outcome.
5. It measures at a point of time therefore mostly it is useful to study chronic diseases.
6. Confounders may be unequally distributed.
7. Group sizes may be unequal.
8. Recall bias.

Review Questions

1. Case-series study definition and contents.
2. Case-series study advantages.
3. Case-series study disadvantages.
4. Which statistical analyses is appropriate in case of case-series study design?
5. Cross-sectional study definition and synonyms.
6. Flowchart content of cross-sectional study.
7. Cross-sectional study advantages.
8. Cross-sectional study disadvantages.
9. Which statistical analyses is appropriate in case of cross-sectional study design?

CHAPTER 10. OBSERVATIONAL ANALYTICAL STUDIES

10.1 Case-control study

Based on time, case-control study design is a *retrospective* study because cases study is provided by looking back at the history. Although a case-control study provides the information about causality, recall bias is a common problem.

The “cases” in case-control studies are individuals selected on the basis of some disease or outcome; the “controls” are individuals without the disease or outcome. The investigators must use matching to associate controls with cases on characteristics such as age and sex. So, both cases and controls should be match able except for exposure to the factor under study; reasoning is to take account for any potentially confounding variables: matching in a case-control study reduces the influence of confounding.

Groups chosen study is based on disease status (dependent variable) and both groups will be asked of their exposure to a factor(s) (independent variable).

Case-control studies are used for looking at potential causes of diseases (causation research question).

Methods of data collection is based on:

⇒ Available records from hospital, vital statistics, and other registers;

⇒ Interview;

⇒ Self-administrated questionnaire;

⇒ Direct measurement.

The exposure histories (rate) of the cases and controls will then be compared. The case-control studies ask the question; “What happened?”.

In a case-control study, we know the outcome and look for the exposure in the past or retrospectively, as is displayed in *Figure 10.1.*

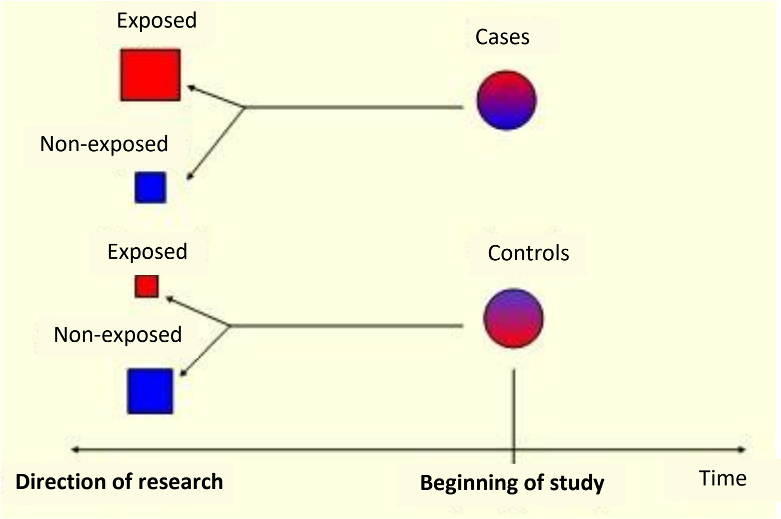


Figure 10.1 Flowchart of case-control study design

Analysis of case-control study includes the calculation of the association measure called *"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.

$$\textit{Odds Ratio} = \frac{\text{Odds Event 1}}{\text{Odds Event 2}}$$

The odds ratio provides a way to look at risk in case-control studies. Odds Ratio shows how many times a case is more likely to have been exposed to a risk factor as compared to a control.

The odds ratio is easy to calculate when the observations are given in a 2x2 table *or* contingency table which. A contingency table is a special type of frequency distribution table, where two variables are shown simultaneously:

Table 10.1
Case-control study 2x2 contingency table

Exposure factor	**Outcome *(disease)***		**Total**
	Cases	Controls	
Exposed	a	b	a + b
Nonexposed	c	d	c + d
Total	a + c	b + d	a + b + c + d

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

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

The odds ratio measurement interpretation scale:

- ⇒ OR= 1, there is no association: no difference in exposure between cases and controls;
- ⇒ OR>1, hazardous exposure;
- ⇒ OR<1, beneficial exposure.

Test of significance

The odds ratio (OR) measurement can be tested for significance with confidence interval calculation, which is a range of values that is likely to contain a population OR with a certain level of confidence. As we know, if an odds ratio (OR) is 1, it means there is no association between the exposure and outcome. So, if the 95% confidence interval for an OR includes 1, it means the results are not statistically significant.

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

> E.g.: OR=3.23
> Those with the disease are 3.23 times as likely to have had the exposure compared to those without the disease.

Advantages:

1. Allows examination of several risk factors.
2. Can study long-term effects of an exposure in short period of time.
3. Use fewer subjects.
4. Relatively quick and relatively less expensive.
5. Suitable for rare diseases.

Disadvantages:

1. Because of their retrospective nature of data collection, there is a greater chance of bias and confounders.
2. Selection of an appropriate control group can be difficult.
3. Recall bias: retrospective nature.
4. Cannot tell about incidence or prevalence.
5. Difficult to establish time relationship between exposure and outcome.

10.2 Cohort study

A cohort study is also called: *Incidence study, Longitudinal study or Prospective study*.

A cohort is a group of people who have something in common and who remain part of a group over an extended time. In terms of statistical meaning, *a cohort* is a group of people that does not have the disease of interest is selected and then observed for an extended period.

In a cohort study, the researcher will collect the information about new cases of a disease (incidence) during regular examinations in the future (prospective).

The cohort studies include the group or groups of individuals (cohort) that are studied over the time as to the onset of new cases of disease and factors associated with the onset of the disease. Cohort studies ask the question: "What will happen?". In a cohort study, we know the exposure, and we follow up the subjects over a period of time looking for the outcome in question (in the future or prospectively), as is displayed in *Figure 10.2.*

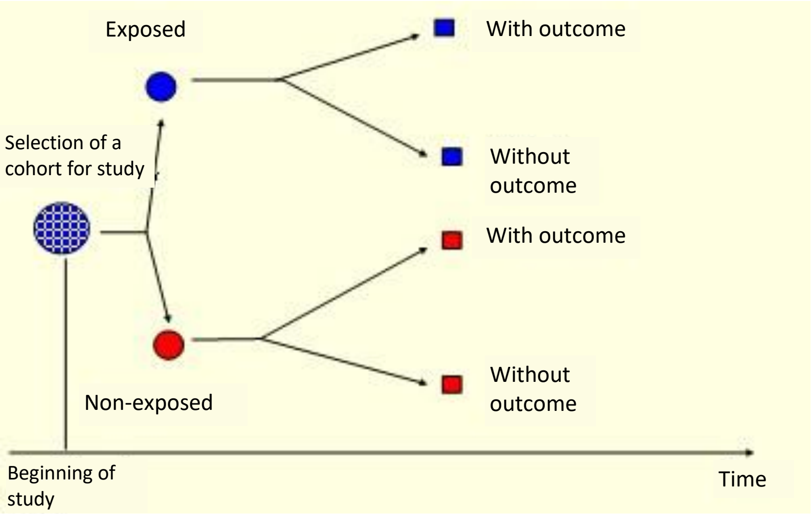


Figure 10.2 Flowchart of cohort study design

The typical cohort studies are usually *prospective* because risk factor exposure and subsequent health outcomes are observed after the beginning of the study. Cohort studies use groups that are similar in all respects except exposure. Select a group of people free of disease and classify them in the analysis as to the level exposure.

Cohort studies are used for:

⇒ Measure the incidence of disease.

⇒ Looking at the causes of diseases.

⇒ Determining prognosis.

⇒ Establishing timing and directionality of events.

Obtaining data in cohort study is possible by personal interviews, medical examination or special test and environmental survey.

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

Analysis of cohort study for estimate the relationship between a risk factor and occurrence of a given outcome use the calculation of association measures:

⇒ Relative Risk (RR)

⇒ Attributable Risk (AR)

The Relative Risk is ratio of the incidence of exposed persons to the incidence of nonexposed.

It is easy to calculate the measures of risk for a cohort study when the observations are arranged in the 2x2 table:

Table 10.2

Table 2x2 arrangements for cohort study

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

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

Relative risk indicates how more likely exposed people have an outcome then unexposed people.

The relative risk interpretation measurement scale:

⇒ RR=1, there is no association – no difference in disease between exposed and unexposed groups;

⇒ RR>1, hazardous exposure;

⇒ RR<1, beneficial exposure.

Test of significance

The relative risk (RR) measurement can be tested for significance with confidence interval calculation, which is a range of values that is likely to contain a population RR with a certain level of confidence. As we know, if a relative risk (RR) is 1, it means there is no association between the exposure and outcome. So, if the 95% confidence interval for an RR includes 1, it means the results are not statistically significant.

The way we would interpret the result of statistically significant Relative Risk is as follows:

> E.g.: RR=3.23
>
> Those with the exposure are 3.23 times as likely develop the disease compared to those without the exposure.

The Attributable Risk is the ratio of the difference between the incidence of exposed and unexposed persons to the incidence of exposed persons represented in percent.

$$\text{Attributable Risk (AR)} = \frac{\text{Incidence of exposed} - \text{Incidence of nonexposed}}{\text{Incidence of exposed}}\, x\ 100\% = \frac{\left(\frac{a}{a+b}\right) - \left(\frac{c}{c+d}\right)}{\frac{a}{a+b}} \times\ 100\%$$

Attributable risk indicates the part of the persons with outcome among exposed due to their exposure that suggest the idea that having been not exposed they could avoid the outcome occurrence.

The way we would interpret the result of Attributable Risk is as follows:

> E.g.: AR=80%
>
> 80% of the people with outcome among exposed are attributed to their exposure.

Advantages:

1. Opportunity to measure risk factors before disease occurs: evidence of causality.
2. Can study multiple diseases outcomes.
3. Can yield incidence rate as well as relative risk estimates.
4. Good when exposure is rare.
5. Minimizes selection and information bias.

Disadvantages:

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

Review Exercises

1. The results of a cohort study are arranged in the following 2x2 table:

Exposure factor – **Brutal Sun Radiation**	Outcome – **Skin Cancer**		Total
	Yes	No	
Yes	39	12	51
No	10	64	74
Total	49	76	125

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

2. The results of a cohort study are arranged in the following 2x2 table:

Exposure factor – **The Sport Practice**	Outcome – **Heart Ischemic Disease**		Total
	Yes	No	
Yes	1024	2376	3400
No	1205	604	1809
Total	2229	2980	5209

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

3. The results of a case-control study are arranged in the following 2x2 table:

Exposure factor – **Diabetes**	Outcome – **Myocardial Infarction**		Total
	Yes	No	
Yes	60	40	100
No	340	360	700
Total	400	400	800

1. According 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. Case-control study design definition and synonyms.
2. The main association measures of case-control study: definition and meaning of interpretation.
3. Case-control study advantages.
4. Case-control disadvantages.
5. Cohort study design definition and synonyms.
6. The main association measures of cohort study: definition and meaning of interpretation.
7. Cohort study advantages.
8. Cohort study disadvantages.
9. State the main difference between the following study designs: case-control and cohort study. Give an example.

CHAPTER 11. EXPERIMENTAL STUDIES

11.1 Clinical Trials Studies Classification

Experimental studies in medicine that involve humans are called clinical trials studies because their purpose is to draw conclusions about a particular procedure or treatment. Therefore, they are used for evaluating the effectiveness of an intervention (therapy research questions).

Classification of clinical trial falls in two big groups:

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. Studies with no controls

Controlled trials are studies in which the experimental drug or procedure is compared with another drug or procedure as usually previously accepted or placebo treatment.

Uncontrolled trials are studies in which the experimental drug or procedure is described being not compared with another treatment.

Because the purpose of an experiment is to determine if the intervention makes a difference, studies with controls are have greater validity in medicine than uncontrolled studies.

11.2 Controlled Trials with Concurrent (*parallel*) Controls

The more common way to make controlled trial is to have two groups of subjects: one group receives experimental procedure (the experimental group) and the other receives the standard procedure or placebo (the control group), as it is displayed in *Figure 11.1.*

The experimental and control groups as more as possible should be similar so that any differences between groups the groups will be due to the planed intervention only. It is important to provide a concurrent control: interventions for both groups are planned for the same time period and the same study.

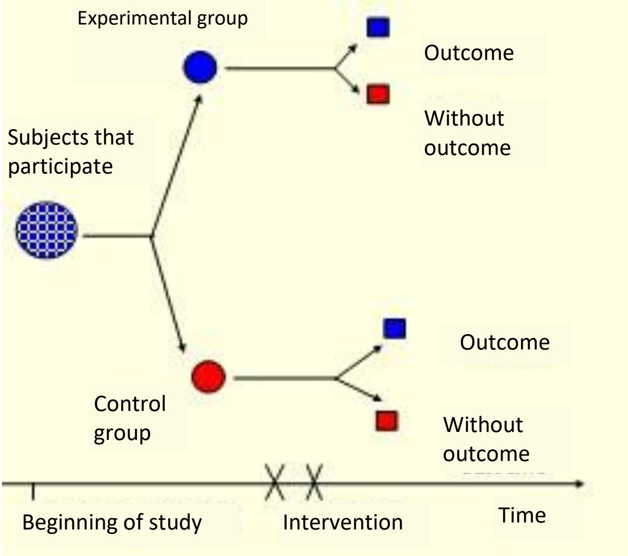


Figure 11.1 Flowchart of randomized controlled trial with parallel controls

In order to reduce the human factor, researcher can design *blind trial* which means clinical trial when the subjects do not know what intervention are receiving, *double-blind trial* in which neither subjects nor investigators know if the subject is in the experimental or the control group.

By ethics reasons in clinical trials is permitted beneficial interventions only.

⇒ *Randomized Controlled Trials* provides the strongest evidence for concluding causation: it provides the best insurance that the result was due to the intervention only.

In randomized controlled trial is administrated an intervention to a group *randomize selected* and we do not know what is receiving (blind). Randomization provide that each individual entered into the trial has the same chance of receiving each of the possible interventions, so allocation of subjects in experimental or control group is given by chance. Also, randomization ensures that known and unknown confounding factors are equal in both groups, then this a way to reduce bias.

⇒ *Nonrandomized Controlled Trials* are studies that do not use randomized assignment.

They are called clinical trial or comparative studies with no mention of randomization as well. Studies using nonrandomized controls are considered much weaker because they do nothing to prevent bias in patient assignment.

11.3 Sequential Controlled Trials

⇒ *Self-Control Trial*s are studies in which the same group works as control group.

A moderate level of control can be obtained by using the same group of subjects for both experimental and control options.

⇒ *Crossover Trial* is when it is administrated an intervention (1) to experimental group and another (2) in a control group.

After a time, interventions are suspended and left a space without it (*washout period*). Then the intervention (1) is administrated to control group and intervention (2) is administrated to experimental group, as it is displayed in *Figure 11.2.*

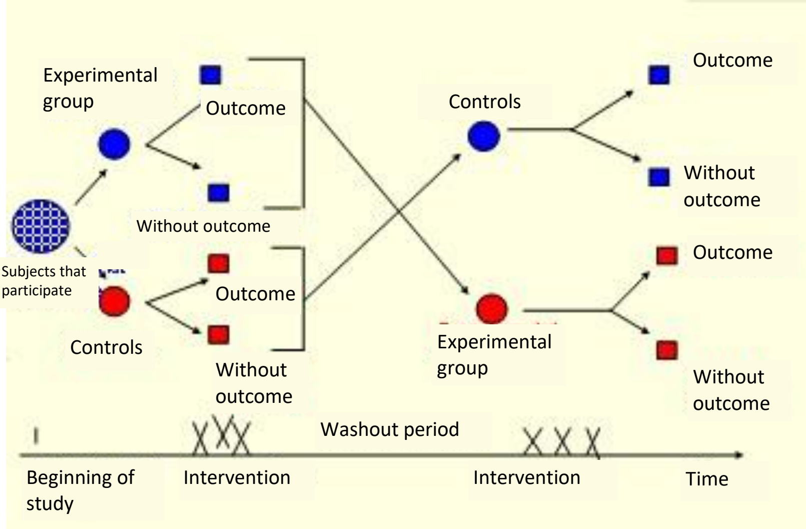


Figure 11.2 Flowchart of clinical trial with crossover sequential controls

11.4 Trials with External Controls

Controlled trials with external controls are the studies when investigators compare the results of another researcher or with the results of a previous study, as it is displayed in *Figure 11.3.* Also, they are called historical controls.

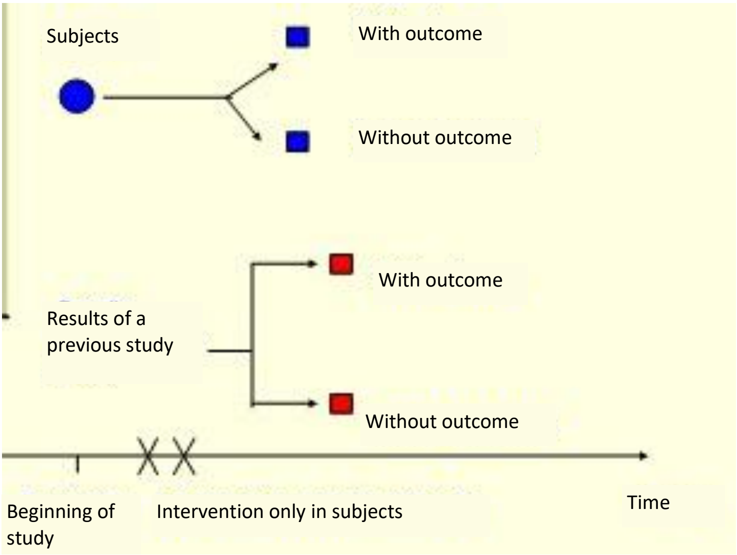


Figure 11.3 Flowchart of clinical trial with external controls

Statistical analysis of clinical trial study includes the calculation of the association measure as:

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

It is easy to calculate the association measures for a clinical trial study when the results are arranged in the 2x2 table:

Table 11.1

Table 2x2 arrangements for clinical trial study

Exposure factor (intervention)	**Outcome *(disease)***		**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*) **–** rate of event (risk of disease) in a group of subjects that received experimental treatment:

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

⇒ *Control Event Rate (CER)* – rate of event (risk of disease) in a group of subjects that received traditional treatment or placebo:

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

⇒ *Relative Risk (RR)* – is ratio of the risk of disease in an exposed subjects' group (received experiment treatment) to the rate of risk disease in a non-exposed subjects' group (didn't receive experiment treatment). Relative risk indicates how more likely exposed people have an outcome then unexposed people.

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

The relative risk interpretation measurement scale:

RR= 1, the intervention is an indifferent factor.
RR>1, intervention is a risk factor
RR<1, intervention is a beneficial/protective factor

⇒ *Absolute Risk Reduction (ARR)* provides a way to assess the reduction in risk compared with the baseline risk and indicates how many subjects avoid the event occurrence for every 100 subjects.

$$\text{ARR} = |EER - CER|$$

e.g.: In the aspirin study the EER for a cardiovascular disease (CVD) was 0.14 in the aspirin group (experimental group), and the control group event rate (CER) was 0.30. in this case ARR = $|0.14 - 0.30|$ = 0.16

The way we would interpret the result of Absolute Risk Reduction is as follows:

> The risk of CVD is 14 subjects per 100 in the group taking aspirin, and 30 in the group-taking placebo. Taking aspirin every 16 subjects per 100 people avoid the occurrence of CVD or 160 per 1000 people, *etc*.

⇒ *Relative Risk Reduction (RRR)* – defined as amount of risk reduction relative to the baseline

$$\text{RRR} = \frac{|EER-CER|}{CER} = \frac{AAR}{CER}$$

e.g.: In the previous aspirin study RRR= $\frac{|0.14-0.30|}{0.30} = 0.53$ or 53%

The relative risk reduction tells us that, relative to the baseline risk of 30 CVD per 100 people, giving aspirin reduces the risk by 53%.

⇒ *Number needed to treat* is an added advantage of interpreting risk data in terms of absolute risk reduction. The main meaning is to find out the number needed to treat in order to prevent one event.

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

e.g: in the previous example of aspirin study NNT =1/0.16= 6.25, then to avoid one CVD is needed to be aspirin treated 6.25 patient. This type of information helps clinicians evaluate the relative risks and benefits of a particular treatment.

Advantages:

1. Give a strong causality evidence
2. Less bias
3. Historic controls can be used in preliminary study

Disadvantages:

1. Expensive
2. Ethical issues
3. They need time
4. Participant compliance

Review Exercises

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

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

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

1.2. Compute all possible measures of association.

1.3. Interpret the obtained results.

2. Select a study with an interesting topic to you from current journals. Carefully examine the research questions and decide which study design would be optimal to answer the question.

2.1. Is that the study design used by the investigators?

2.2. If not, what are the reasons for the study design used? Do they make sense?

Review Questions

1. Clinical trial definition and classification.
2. What does mean to be controlled and uncontrolled clinical trial?
3. What does mean to be randomized clinical trial?
4. Controlled clinical trials with concurrent controls: definition, types and flowchart of appropriate design.
5. Sequential clinical trials with concurrent controls: definition, types and flowchart of appropriate design.
6. Trials with historical control: definition and flowchart of appropriate design.
7. The main analyses measures of clinical trial study: definition and meaning of interpretation of each of them.

CHAPTER 12. REPORTING THE RESEARCH FINDINGS: GENERAL APPROACHES

12.1 Writing Research Report

Writing the report is the last step of the research process. The reports inform what you have done, what you have discovered and what conclusions you have drawn from your findings. The report should be written in academic style using a formal and not journalistic language.

Traditional written reports have following format:

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

B. *Project body*
   - Introduction
   - Review of related literature
   - Material and methods (Research design)
   - Results (Data analysis and interpretation)
   - Discussions (Summary)
   - Conclusions
   - Recommendations
   - References / Bibliography
   - Appendices

12.2 Public Presentation of Medical Research Results

The purpose of oral presentation of medical research is to submit a scientific work.

Oral presentation general design:

- Title, authors (the first slide)
- Introduction (1-2 slides)
- Aim and objectives (1-2 slides)
- Material and Methods (1-2 slides)
- Results – only most important (2-3 slides)
- Discussions (1 slide)
- Conclusions (1 slide)
- Closing (1 slide)

A presentation of about 10 minutes should be no more than 8-10 slides. As usually one slide takes one minute. Choice of graphic presentation (figure and tables) depends of objectives but the figures are preferably.

There are some recommendations for practical realization:

⇒ *Title*

Single line
Bold or different color

⇒ *Text*

Easily readable from the rear rows of the room. Should not exceed 5-7 lines in one slide.

> Broadly used **7x7 rule** says that, for each slide in your presentation, you should use no more than: 7 lines (or bullets) per slide. 7 (or fewer) words per line.

⇒ *Figure and tables*

Same principles as the writing report

The tables should not exceed 3-4 columns and 5-7 rows

12.3 Structure and Principles of graduation paper Development at Nicolae Testemitanu State University of Medicine and Pharmacy

The graduation paper must demonstrate the student's ability to work with literature relevant to the subject. The graduation paper should be developed correctly from the point of view of methodology, data analysis and interpretation, have a logical structure, be written in scientific language. The aspect must be in accordance with academic standards and comply with the scientific writing recommendations accordingly to Nicolae Testemitanu State University of Medicine and Pharmacy Regulation on the development and defense of the graduation paper 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' behavior.

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

Objectives of Research ethics are:

1. To protect human participants, their dignity, rights, and welfare.
2. To make sure that research is directed in a manner that assists welfares of persons, groups and/or civilization as a whole.
3. To inspect particular research events for their ethical reliability.

13.2 Research Ethics Principles

Research ethics are based on three fundamental approaches:

⇒ Respect for persons;

⇒ Beneficence;

⇒ Justice.

There are mainly five principles of research ethics based on its fundamental approaches:

1. Minimizing the risk of harm;
2. Obtaining informed consent;
3. Protecting anonymity and confidentiality;
4. Avoiding deceptive practices;
5. Providing the right to withdraw.

Main tips to ensure ethics at different steps of research:

⇒ Collect the facts and talk over intellectual belongings openly;

⇒ Outline the ethical matters;

⇒ Detect the affected parties (stakeholders);

⇒ Ascertain the forfeits;

⇒ Recognize the responsibilities (principles, rights, justice);

⇒ Contemplate your personality and truthfulness;

⇒ Deliberate innovatively about possible actions;

⇒ Respect privacy and confidentiality;

⇒ Resolve on the appropriate ethical action and be willing to deal with divergent point of view.

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 list, acknowledgements, etc.

Types of plagiarism are:

⇒ 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 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 in order to elaborate your own.

⇒ Self-plagiarism – reusing work that you have already published or submitted for a class. 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 it was not intentional, *it is still plagiarism and not acceptable.*

14.2 Preventing plagiarism techniques

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

⇒ Be confident in understanding what Plagiarism is;

⇒ Provide your own ideas by find 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 in 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, even so you still have to give appropriate academic credit;

⇒ Cite yourself as well;

⇒ Use plagiarism detection software.

Applying these techniques will help you ensure your work is your own.

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.

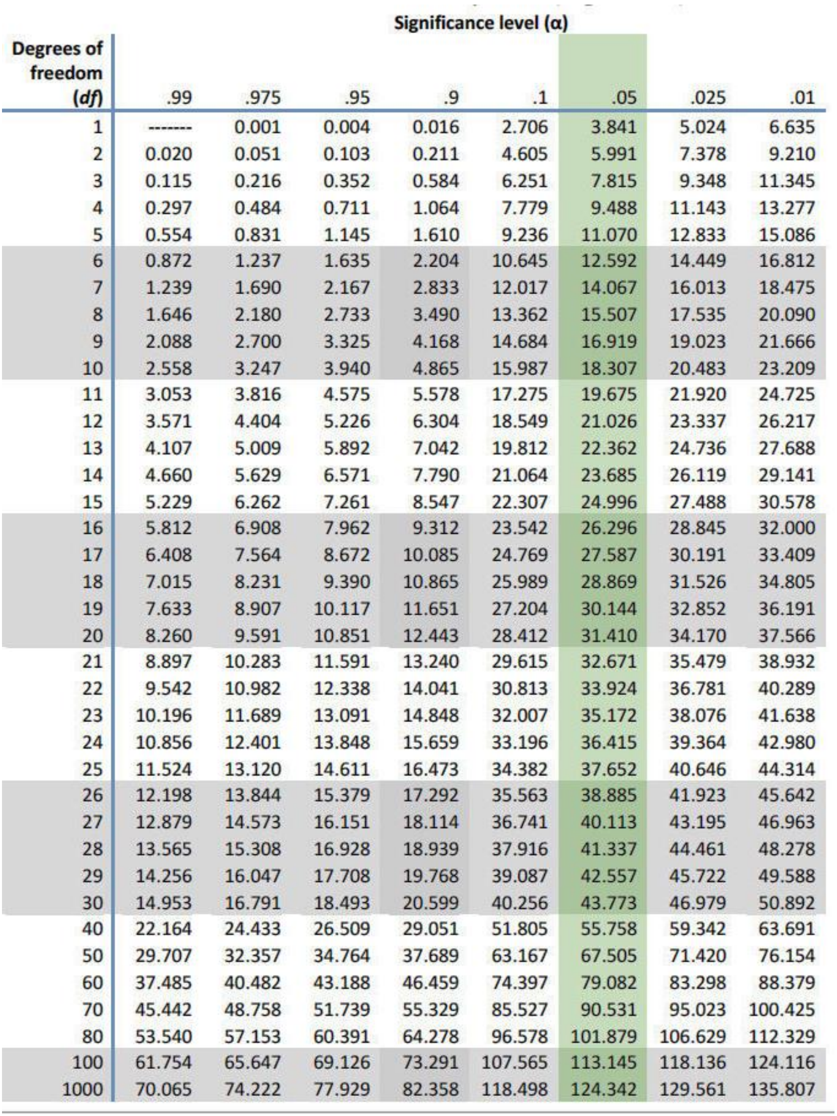
WEINSTEIN MC, FINEBERG HV: *Clinical Decision Analysis*, WB Saunders, 1998.

APPENDIX A: **Critical values for the "t" distribution corresponding to commonly used areas under the curve**

Degrees of freedom	Area in 1 Tail				
	0.05	0.025	0.01	0.005	0.0005
	Area in 2 Tails				
	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: **Chi-square distribution table**



	Significance level (α)							
Degrees of freedom (*df*)	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807