HEALTH MINISTRY OF REPUBLIC OF MODOVA
STATE UNEVERSITY OF MEDICINE AND PHARMACY
"NICOLAE TESTEMITANU"

Department Social Medicine and Health
Management "Nicolae Testemitanu"

Elena Raevschi, Dumitru Tintiuc

# BIOSTATISTICS & RESEARCH METHODOLOGY

*Methodological recommendation*
*for medical students*

CHISINAU
2012

**HEALTH MINISTRY OF REPUBLIC OF MODOVA**
**STATE UNEVERSITY OF MEDICINE AND PHARMACY**
**"NICOLAE TESTEMITANU"**

**Department Social Medicine and Health Management "Nicolae Testemitanu"**

**Elena Raevschi, Dumitru Tintiuc**

# BIOSTATISTICS & RESEARCH METHODOLOGY

***Methodological recommendation for medical students***



**CHISINAU**
**Editorial-Polygraphic Center *Medicina***
**2012**

CZU 311.2:614(076.5)
R 19

Recommended to printing by Central Methodological Council of the State University of Medicine and pharmacy "Nicolae Testemitanu", protocol nr.1 from 2012 September, 14

**Authors:**

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

*Dumitru TINTIUC* – PhD, professor, Chairman of the Department of Social Medicine and Health Management, University of Medicine and Pharmacy "Nicolae Testemitanu".

**Reviewers:**

*Ion MEREUTA* – PhD, professor, University of Medicine and Pharmacy "Nicolae Testemitanu".

*Serghei POLIUDOV* – PhD, associate professor, Department of Social Medicine and Health Management, University of Medicine and Pharmacy "Nicolae Testemitanu".

This methodological recommendation is designed based on the international experience following the current didactic requirements. It corresponds to the analytical program (syllabus) of the Biostatistics and Research Methodology discipline for Medicine 2 faculty's students.

DESCRIEREA CIP A CAMEREI NAȚIONALE A CĂRȚII

Raevschi, Elena.

Biostatistics & Reasearch Methodology: Methodological recommendation for medical students / Elena Raevschi, Dumitru Tintiuc; Stat Univ. of Medicine and Pharmacy "Nicolae Testemițanu", Dep. Social Medicine and Health Management. – Ch.: Medicina, 2012. – 94 p.

250 ex.

ISBN 978-9975-113-56-4.

311.2:614(076.5)

R 19

ISBN 978-9975-113-56-4

## 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 a qualitative 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.

# 1. INTRODUCTION TO BASIC BIOSTATISTICS & MEDICAL RESEARCH. METHODOLOGY

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.

***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.*

On 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.

On practice 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 understandding its optimal application in getting a right scientific research results.
- 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.

On integration level

- To appreciate theoretic-applicable value of Medical Research Methodology in different disciplines in the health field.

Health research is an interdisciplinary field, mostly achieved through a large specter of knowledge. "Basic Biostatistics and Medical Research Methodology" 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 basic knowledge in the math and fundamental and applicative medicine fields. Knowledge of a cumputer is an indispensable requirement.

## 2. DESCRIPTIVE STATISTICS: DATA PRESENTATION

- **Definition of Variable**
- **Types of Data (variable)**
- **Scales of Measurement**
- **Tables**
- **Graphs**

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.

### 2.1 Definition of Variable

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

For example in a study population variable can be: age, data of birth, nationality, number of children, etc.

### 2.2 Types of Data (variable)

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

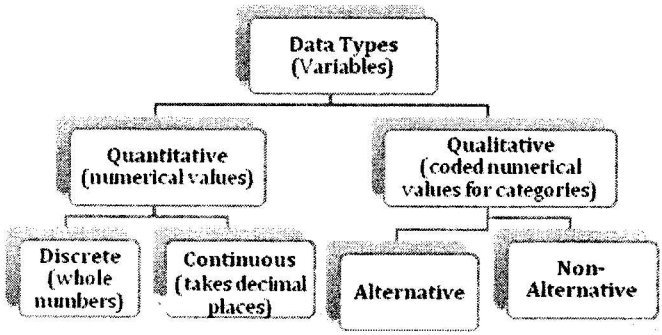



*Figure 2-1.* **Data Types**

**Discrete data** represent measurable quantities taking only specified values (integers) that differ by fixed amounts; no intermediate values are possible.

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

**Continuous data** 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.

**Alternative** *(Dichotomous or binary)* **data** represent measurable categories in that outcome can take only one of two values: yes or no

**Non-Alternative data** represent measurable categories in that outcome can take many values.

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

## 2.3 Scales of Measurement

Depending of the nature of the variable, we have different measurements scales.

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 (classificatory).
- Ordinal (ranking).
- Numerical (Interval and Ratio scales).

**Nominal scales** (dichotomous or binary or categorical scales):

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 (qualitative variables).

For example, the variable "*gender*" can have only in two values: *male and female – Alternative (Dichotomous variable).* "*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 variable.*

The sequence in which subgroups are listed makes no differrence, as there is no relationship among subgroups.

**Ordinal scales:**

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.*

*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.* 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.

**Numerical scales:**

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

- ***Interval scale:***

An interval scale has all the characteristics of an ordinal scale. This has no absolute zero value e.g. body temperature.

*For example*:

Celsius scale: 0*C to 100*C

Fahrenheit scale: 32*F to 212*F

- ***Ratio 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.*

The measurement of variables like serum cholesterol level, income, age, body height and weight are examples of this scale. A person who is 40 year old is *twice* as old as one who is 20 year old.

## 2.4 Tables

The tables are commonly used to display data observation:

***a) Simple frequency distribution Table (S.F.D.T.)***

*Table 2.1* **Distribution of students at the University of Medicine of Moldova according to their smoking habit, 2011**
*(Simple tabular presentation)*

| Faculty | Number of students | |
|---|---|---|
| | Absolute number of students | % |
| **1. Medicine** | | |
| **2. Public Health** | | |
| **3. Pharmacy** | | |
| **Total:** | | 100,0 |

***b) Grouping frequency distribution Table***

***Table 2.2*** **Distribution of students at the Universityaccording to their smoking habit, 2011** *(Grouping tabular presentation)*

| Faculty | Sex | | Age (when they smoked the first cigarette) | | | Total |
|---|---|---|---|---|---|---|
| | male | female | < 15 years | 15 – 18 years | > 18 years | |
| **1. Medicine** | | | | | | |
| **2. Public Health** | | | | | | |
| **3. Pharmacy** | | | | | | |
| **Total:** | | | | | | |

## *c) Complex frequency distribution Table*

*Table 2.3* **Distribution of students at the University according to their smoking habit, 2011** *(Complex tabular presentation)*

| Faculty | Number of cigarettes smoked a day | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | <10 | | | 11–20 | | | >20 | | | |
| | m | f | both | m | f | both | m | f | both | |
| 1.Medicine | | | | | | | | | | |
| 2. Public Health | | | | | | | | | | |
| 3.Pharmacy | | | | | | | | | | |
| Total: | | | | | | | | | | |

## 2.5 Graphs.

1. **Line Graphs** – commonly used to display trends over time

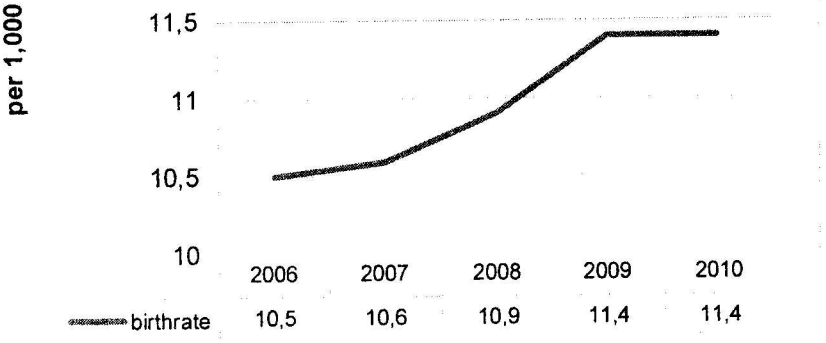



*Figure 2-2.* **Birthrate in Republic of Moldova, 2006–2010**

Line graphs illustrate the relationship between continuous quantities. Each value of the *x-axis* has a single corresponding values on the on the *y-axis*. Adjacent points are connected by straight lines.

2. **Bar Charts** – compare multiple values

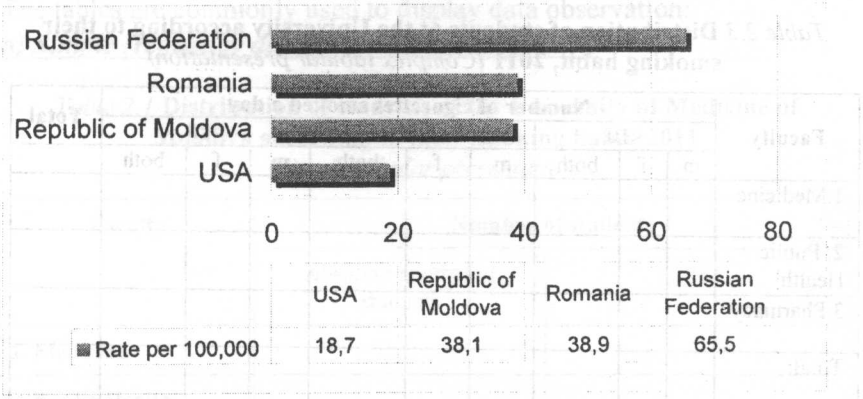



*Figure 2-3.* **Age-standardized prevalence estimates for smoking cigarette: Daily users – Males, Aged 15+, 2008**

Bar charts are popular type of graph used to display a frequency distribution for nominal or ordinal data. In bar charts the various categories into which the observations fall are presented along a horizontal axis. The bars have to be of equal distance being separated from one other so as not to imply continuity.

3. **Column Charts** ***(Vertical Bar Charts)*** – compare values across categories

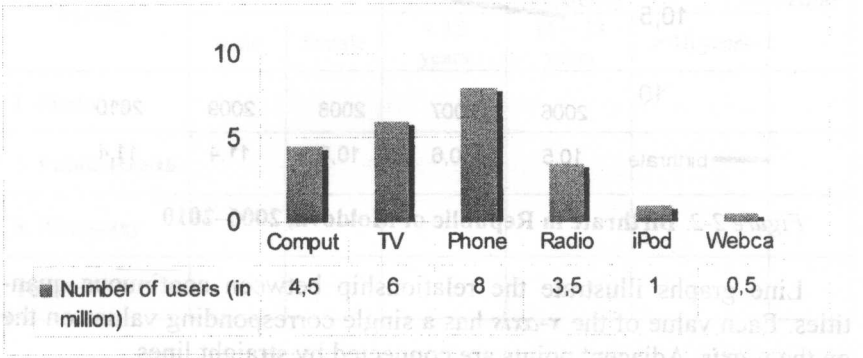



*Figure 2-4.* **Most Used Technology**

Column charts are popular type of graph used to display a frequency distribution for nominal or ordinal data. A vertical column is drown above each category such the height of the bar represents either the frequency or the relative frequency of observations within that class. The columns have to be of equal distance being separated from one other so as not to imply continuity.

4. **Pie Charts** – display contribution of each value to a total

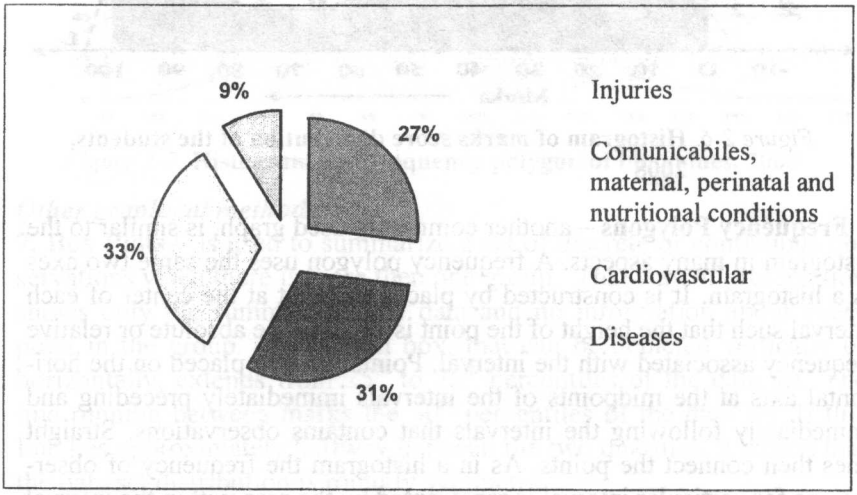



*Figure 2-5.* **Distribution of major causes of death including CVDs, WHO 2008**

5. **Histograms** – the most informative way to present relative frequentcies. A histogram looks a bit like a column charts, but:

- whereas a column chart is a pictorial representation of a frequency distribution of either nominal or ordinal data *(categorical data)*, a histogram display a frequency distribution for discrete or continuous data *(numerical data)*.
- histograms give an idea of the shape of the relative frequency distribution. Column 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.

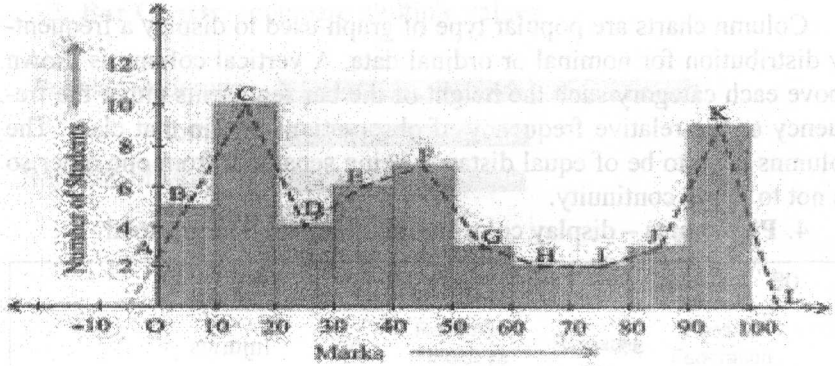
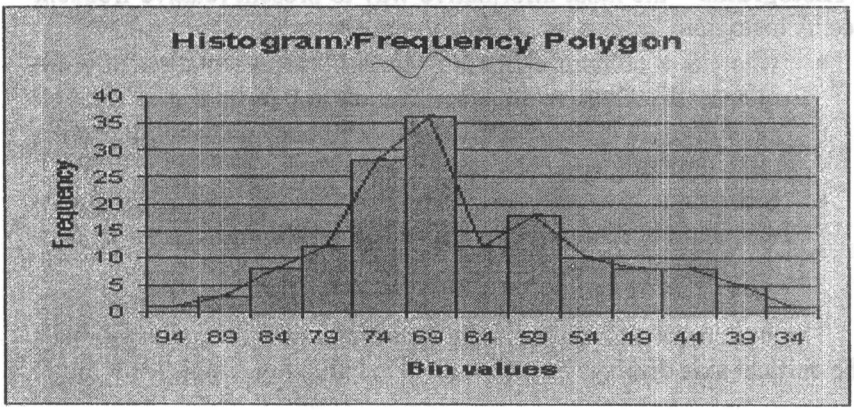



*Figure 2-6.* **Histogram of marks score distribution of the students, 2008**

6. **Frequency Polygons** – 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.



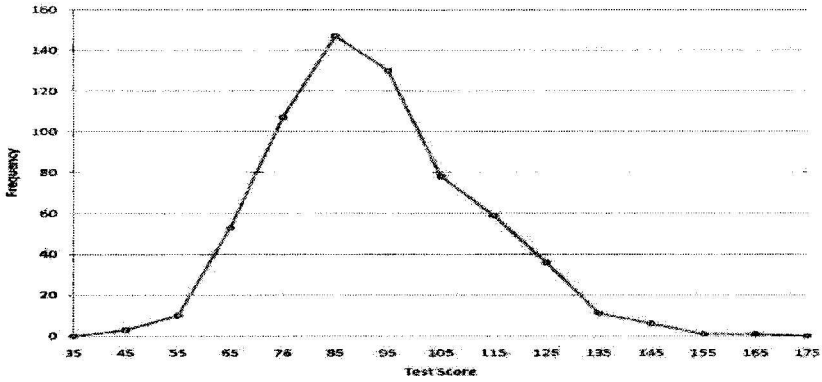



*Figure 2-7.* **Histogram and frequency polygon of bin values, 2008**

***Other graphical methods:***

**7. Box Plots** – is used to summarize a set of discrete or continuous observations when there is more then one group. That way of presentation shows only the summary of the data and no information about every point in the group. The central box, that can be depicted vertically or horizontally, extends from $25^{th}$ to $75^{th}$ percentiles of the data set. The line running between marks the $50^{th}$ 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 represent the adjacent values of the plot which are the most extreme observations, but not more than 1.5 times interquartile range. 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.

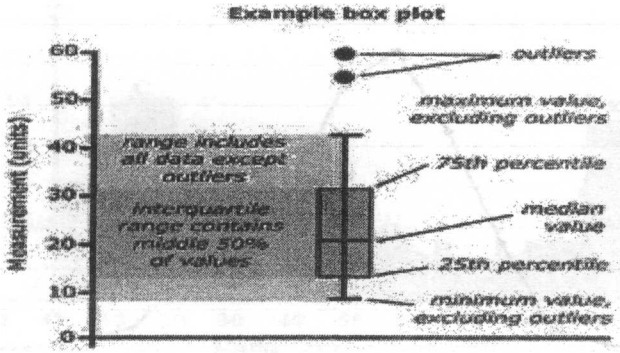



*Figure 2-8.* **Box plot example**

8. **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.

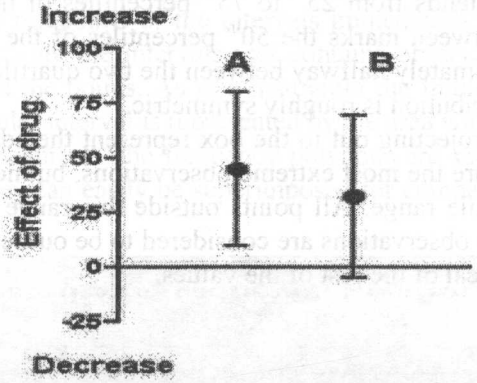



*Figure 2-9.* **Error bar charts of effect of drug for patients with and without a pulmonary embolism**

9. **Scatter Plots** – a two-way scatter plot is used to depict the relationship between two different continuous measurements. Each point on the graph represents a pair of values simultaneously.

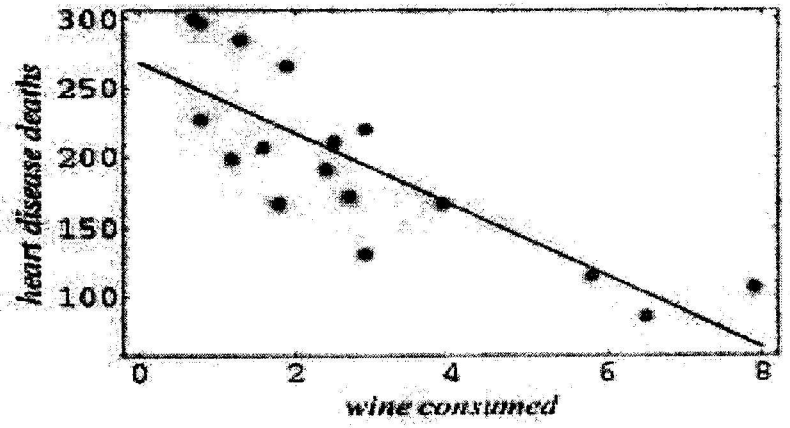



*Figure 2-10.* **Illustration of a scatter plot**

**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 birth 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 Municipal Hospital, 2007–2011**

| Years | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|
| Divergence, % | 11 | 9.8 | 8.0 | 9.2 | 8.2 |

4. Use the following data to display it bz all appropriate graf. State your decision.

**Acute viral hepatitis morbidity in the Republic of Moldova, 2010**

| Type | HAV | HBV | HCV | HDV | HEV |
|---|---|---|---|---|---|
| % | 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. How do ordinal data differ from nominal data? Give examples.
3. How do alternative variable differ from non-alternative one? Give examples.
4. Definition of measurements scale. Classification. Give an example for each type.
5. What kind of data presentation do you know? State the difference between them.
6. Tabular data presentation: contents and types. Give an example for each type.
7. Graph data presentation: contents and types. Give an example for each type.
8. Name appropriate graph data presentation for nominal variable. Give examples.
9. Name appropriate graph data presentation for ordinal variable. Give examples.
10. Name appropriate graph data presentation for numerical variable. Give examples.

## 3. DESCRIPTIVE STATISTICS: NUMERICAL DATA SUMMARIZING

- Measures of Central Tendency
- Measures of Variability.

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

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

***Descriptive Statistics measures are:***

A. Central tendency: Mean, Median, and Mode

B. Variability: Range, Interquartile range, Variance, Standard deviation, Coefficient of Variation

### 3.1 Measures of Central Tendency

Measures of the Central Tendency are the most useful summary numbers, which characterize the middle (the center) of the set data. The three measures used in medicine are mean, median, and mode. All three are used for numerical data, and the median is used for ordinal data as well.

**Mean** $(\bar{X})$

The mean measure the middle of the distribution of a numerical variable.

The mean (denoted by x-bar) is calculated dividing the sum ($\Sigma$) of the individual data ($x_i$) by the number of observations (n):

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

$$\bar{X} = \frac{\Sigma x_i}{n-1}$$ – when n<120

$$\bar{X} = \frac{\Sigma x_i}{n}$$ – when n>120

**b)** ***The Weight Mean*** – used for data set when some values occur more than one time.

$$\bar{X} = \frac{\Sigma x_i \times f}{n}$$, when "f "– is frequency of individual data

$$\bar{X} = \frac{\Sigma x_i \times f}{n-1}$$, – when n<120

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.

**Median ($M_d$)**

The median 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. Me-dians 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 (n+1)/2 observation.

If an even number of observation, the median $M_d$ will be the midpoint between the middle two observations. Ex.: Median of 14 observations is the midpoint between $7^{th}$ and $8^{th}$.

**Mode ($M_o$)**

Mode 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.

***For correct practical application of central tendency measurements two factors are important:***

1. The scale of measurements.
2. The shape of the distribution of data set.

If outlying observations occur in only one direction, the distribution is called a skewed distribution. There are two types of the ***skewed distribution:***

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

*A **symmetric distribution*** has the same shapes on both sides of the mean.

Relationship of central tendency measurements in this case is:

$M_o = M_d = \bar{X}$

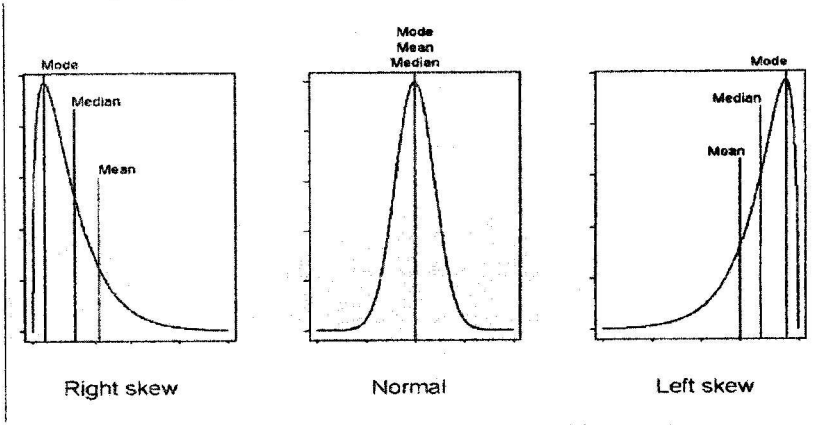



*Fig. 3-1.* **Illustration of data type distribution** *(skewness)*

Skewness ranges from -3 to 3. Acceptable range for normality is skewness lying between -1 to1. Normality should not be based on skewness alone; the kurtosis of the bell-curve (the peakness) is important too (see Fig.3–2). Acceptable range for normality is kurtosis lying between -1 to 1.

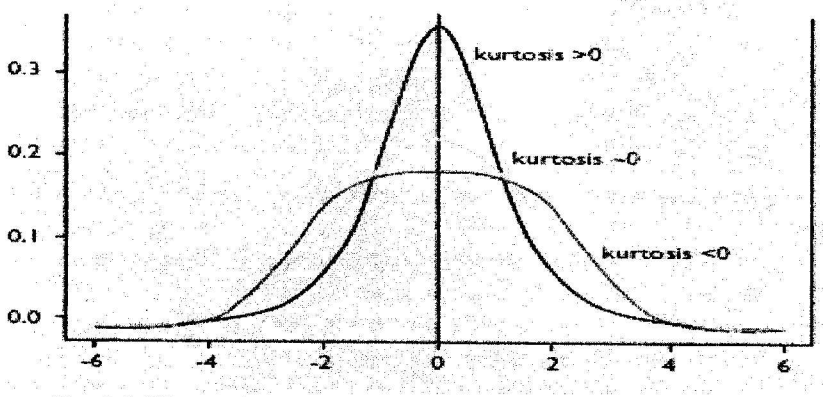



*Fig. 3-2.* **Illustration of data type distribution** *(kurtosis)*

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 3-1.* **Flowchart for normality checking**

| | |
|---|---|
| | **Small samples** (n<30); *always assume not normal* |
| | **Moderate samples** (30 – 100)<br>*If formal test is significant, accept non-normality otherwise double-check using graphs, skewness and kurtosis to confirm normality.* |
| | **Large samples** (n>100)<br>*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: The Pearson skewness coefficient ($C_{as}$), defined by:

$C_{as}$ =(mean-mode)/ standard deviation

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.

According these two factors (scale of measurements and shape of distribution) it is possible to make ***a right choice for central tendency measurements best application:***

- 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.

### 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. Certainly it is necessary to describe the dispersion (the spread) of the data set. This is the target of Measures of Variability: Range, Variance, Standard Deviation and Coefficient of Variation.

*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)

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.

**Range**

Range is heavily influenced by the most extreme values and ignores the rest of the distribution.

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

$$R=\text{Max}(x_i) - \text{Min}(x_i)$$

*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

The range is used with numerical data when the purpose is to emphasize extreme values.

**Interquartile Range** (IQR)

A second measures of variability that is used to limit the influence of extreme values.

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.

$IQR = Q_3 - Q_1$

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.

**The Variance ($S^2$) and Standard Deviation (S)**

The variance is the measure of how spread out a distribution is.

***The variance definition is the average squared deviation of each number from its mean.***

$$S^2 = \frac{\Sigma(Xi - \bar{X})^2}{n-1}$$

*where:*

*Xi – individual values of data set; $\bar{X}$ – mean; n – number of observations*

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

***The Standard Deviation (S)***

The standard deviation 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.***

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.

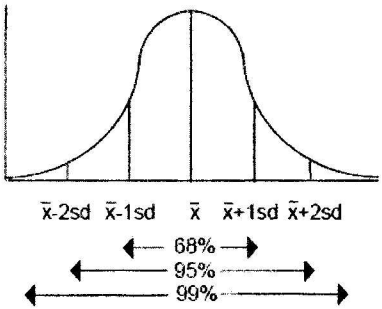



*Fig. 3-3.* **The Normal Distribution**

Knowing variance, is very easy to calculate standard deviation, which is square root taken by variance.

$$S = \sqrt{S^2} = \sqrt{\frac{\Sigma(X_i - \bar{X})^2 \times f}{n-1}}$$

*where:*

*$S^2$ – variance; Xi – individual values of data set; $\bar{X}$ - mean; n – number of observations*

The standard deviation is used when the mean is used (symmetric numerical data)

**The Coefficient of Variation** (CV)

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%.

$$CV = \frac{SD}{\bar{X}} \times 100\%$$

*where:*

*SD – Standard Deviation; $\bar{X}$ - mean*

| Coefficient of Variation, % | Level of Variability (Dispersion) |
|---|---|
| <10% | Low |
| 10-35% | Medium |
| >35% | High |

As it has no units, the coefficient of variation can be used to compare two tests 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 following scale:

For a 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.

*Example:*

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. Interpret them.

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:

| **Xi** | **Frequency, f** | Xi x f | $(Xi - \bar{X})$ | $(Xi - \bar{X})^2$ | $(Xi - \bar{X})^2 \times f$ |
|---|---|---|---|---|---|
| 24 | 2 | 48 | | | |
| 26 | 4 | 104 | | | |
| 30 | 3 | 90 | | | |
| 31 | 5 | 153 | | | |
| 33 | 2 | 66 | | | |
| 35 | 3 | 105 | | | |
| | n=19 | 566 | | | |

3. Calculate the Mean:

$\bar{X} = \frac{\sum Xi \times f}{n}$, when "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 $M_d$ will be the (n+1)/2 observation.

$M_d = (19+1)/2 = 10$;

The $10^{th}$ observation is the median (only in arranged set).

$M_d = 31$.

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

$$S^2 = \frac{\Sigma(X_i - \bar{X})^2 \times f}{n}$$

$X_{i1} = 24 - 29.8 = -5.8$
$X_{i2} = 26 - 29.8 = -3.8$
$X_{i3} = 30 - 29.8 = +0.2$
$X_{i4} = 31 - 29.8 = +1.2$
$X_{i5} = 33 - 29.8 = +3.2$
$X_{i6} = 35 - 29.8 = +5.2$

Complete the table with the calculate data:

| $X_i$ | Frequency, f | $X_i \times f$ | $(X_i - \bar{X})$ | $(X_i - \bar{X})^2$ | $(X_i - \bar{X})^2 \times f$ |
|---|---|---|---|---|---|
| 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 |

$S^2 = \frac{233.96}{19} = 12.3$;

then calculate the standard deviation: $S = \sqrt{S^2} = \sqrt{12.3} = 3.5$

7. Calculate the coefficient of variation:

$CV = \frac{SD}{\bar{X}} \times 100$; $\quad CV = \frac{3.5}{29.8} \times 100\% = 11.7\%$

*Interpretation:*

- The ages of the data set patients that you are studying is 29.8±3.5 years.
- Data dispersion level is medium (CV=11.7%) – appropriate for a scientific research.
- The distribution of data is almost symmetric (acceptable skewed to the left): Mode=31 Median=31 Mean=29.8.
- The mean is representative for this data set.

**Exercises:**

1. Using the following data set:
   - Find the measures of central tendency.
   - Make appropriate graph presentation.
   - Analyze type of data distribution.

   A study of 52 patients was conducted investigating the BMI data: 29; 22; 21; 24; 24; 24; 37; 23; 35; 35; 27; 28; 35; 35; 27; 28; 28; 34; 50; 28; 28; 29; 32; 37; 31; 29; 22; 43; 27; 28; 29; 28; 34; 50; 28; 28; 29; 32; 37; 31; 29; 22; 43; 27; 28; 29; 27; 21; 28; 33; 37; 39.

2. Using the following data set:
   - Find all measures of central tendency. Interpret them.
   - Make appropriate graph presentation.
   - Analyze type of data distribution.

   A study of 42 patients was conducted investigating the blood glucose level data: 3.3; 3.5; 3.0; 2.9; 5.5; 5.7; 6.6; 3.3; 7.8; 9.0; 12.0; 9.0; 6.0; 11.0; 11.8; 11.8; 11.0; 5.5; 3.3; 4.0; 4.2; 4.0; 5.7; 6.6; 3.3; 7.8; 9.0; 12.0; 2.9; 9.0; 6.0; 6.0; 11.0; 11.8; 11.8; 11.0; 11.8; 5.5; 11.8; 6.0; 3.3; 3.5 (moll/l).

3. Using the data set from p.1:
   - Find all measures of variability. Interpret them.
   - 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 Descriptive Statistics are more appropriate for?

4. Using the data set from p.2:
   - Find all measures of variability. Interpret them.
   - 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 Descriptive Statistics are more appropriate for?

**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. Standard deviation: meaning, characteristics, and preferred use conditions. Rule of calculation.
11. Coefficient of variation: meaning, characteristics, and preferred use conditions. Rule of calculation.

# 4. DESCRIPTIVE STATISTICS: NOMINAL AND ORDINAL DATA SUMMARIZING

- Relative values: The types and methods of calculation. Graph Prezentation.
- Vital statistics indicators.
- Adjusted Rate: Direct Method of Standardization.

## 4.1Relative values: The types and methods of calculation. Graph Presentation

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.

Relative values type:

- Rate.
- Proportion.
- Ratio.

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

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

$a$ – 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 (eg, 100 (%); 1,000 ($^0/_{00}$); 10,000 ($^0/_{000}$); 100,000 ($^0/_{0000}$ )

Rates are mostly used to established the morbidity, natality and mortality phenomenon 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, Column chart.

**Proportions** are extensive statistical indicators expressing the structure of a phenomenon. A proportion is defined as a part of a phenolmenon divided by whole phenomenon. Mostly this indicator is shown in %.

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

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

*Graph presentation:* Pie chart

**Ratios** are the rapport between 2 independent phenomenons. It is defined as a part divided by another part. So, the ratio is the number of observations in a group with a given characteristic divided by the number of observations without the given characteristic:

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

*Graph presentation:* Line graph, Bar chart, Column chart.

## 4.2 Vital Statistics Indicators

Vital statistics describe the life of the population: birth, deaths, marriages, divorces, and diseases occurrences.

Mostly vital statistical indicators are represented by rates. Somme of the most commonly used rates are briefly defined in the following items:

**Mortality Rate:** The number of deaths that occur during the specified time period, divided by the total number of population 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. Rates that are computed within relatively small, well-defined subgroups are called ***specific rates.***

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.*

**Morbidity Rate:** The number of person who develop a disease during the specified time period divided by the number of people who were at the risk for 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 number of people who were at the risk for the same period of time.

**Prevalence** is defined as the number of persons with a given disease at a given point in time divided by the 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 rate and specific rate as well.

*Example:*

Number of population is 75,000 in a specified year and locality. In that year were born 1908 and died 897 of individuals. In that locality were 40 doctors: 20 – physicians;

10 – surgeons; 10 – other.

Compute all main possible statistical indicators.

*1. Rates: (intensive indicators).*

Birth rate = $\frac{1.908}{75.000} \times 1{,}000 = 25.44\,^{0}/_{00}$.

Crude mortality rate = $\frac{897}{75.000}\ 1{,}000 = 11.96\ ^{0}/_{00}$.

*2. Proportion (extensive indicator).*

Percentage of physicians = $\frac{20}{40} \times 100\% = 50\%$.

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

*3.Ratio (rapport indicator)*

Medical supply = $\frac{40}{75.000} \times 10{,}000 = 5.3$.

5.3 doctors for every 10,000 populations.

### 4.3 Adjusted Rate: Direct Method of Standardization.

Crude rates can be used to make comparisons between two different populations only if the populations are homogenous in all characterristics. 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.

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.

*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 |

**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 (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 then 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 then 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 males population of 220, having the same sex-specific viral hepatitis morbidity rate as at factory 'A" is (2%):

$100 - 2$

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

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:

Female factory "A".

Males and Females factory "B".

Total expected number 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:

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

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

**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").

**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 indicators.
- 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 500 children were born, 110 individuals died and 400 individuals develop cardiovascular disease for the first time in their life.

- Compute all possible vital statistics indicators.
- Make appropriate graph presentation for them.

**4.**Consider the following data comparing acute abdomen lethality cases by start disease terms of hospitalization at the hospitals "A" and "B":

| **The term, 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 hospitals? Explain these results (interpretation and conclusions).

**5.** Consider the following data comparing the hospital mortality at the hospitals "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?

## 5. CORRELATION

- Pearson's Correlation Coefficient.
- Spearman's Rank Correlation Coefficient.

A lot of medical researches are related to relationship between two or more characteristics. For this kind of purpose is appropriate to use correlation that is able to examining the relationship between two variables.

### 5.1 Pearson's Correlation Coefficient

Pearson's correlation coefficient is one measure of the relationship between two *numerical* characteristics, symbolized by "X" and "Y". The correlation coefficient is denoted by "r", it is calculated using the formula:

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

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: $-1 \leq r \leq 1$.

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

- ***Positively correlation*** ( + ), when coefficient of correlation is $0 \leq r \leq 1$ : "Y" tends to increase in magnitude as "X" increases;
- ***Negatively correlation*** ( - ), when coefficient of correlation is $-1 \leq r \leq 0$ : "Y" decreases as "X" increases;

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.

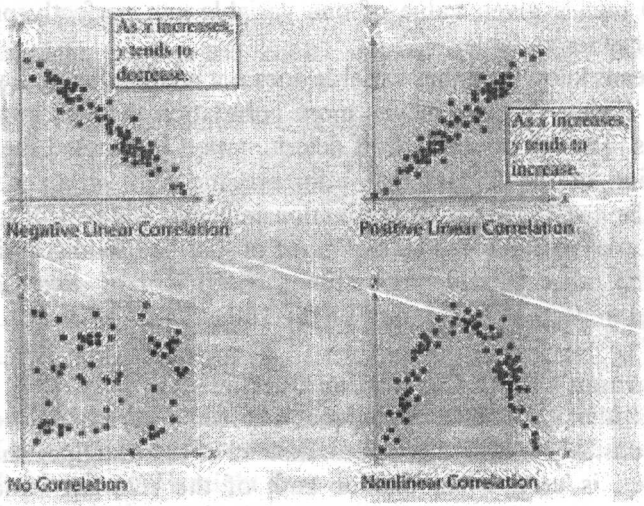



*Figure 5.1.* **Scatter plots showing possible relationship between "X" and "Y".**

There are following crude rule for ***interpreting the size of correlation:***

- The correlation coefficient equal $\pm$ 1 denote a perfect linear and ***strong relationship;***
- Correlations from 0 to 0.25 (or – 0.25) indicate ***little or no relationship***;
- Correlations from 0.25 to 0.50 (or – 0.25 to – 0.50) indicate ***a fair degree of relationship;***
- Correlations from 0.50 to 0.75 (or – 0.50 to – 0.75) indicate ***a good relationship***;
- Correlations greater than 0.75 (or – 0.75) indicate ***a very good to excellent relationship.***

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

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

## 5.2 Spearman's Rank 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.

**Exercises:**

**1. 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 10 newborns Apgar score at birth data:

| **Nr. child** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Pregnancy term, weeks** | 26 | 33 | 40 | 36 | 38 | 29 | 29 | 29 | 40 | 39 |
| **Apgar score** | 6 | 8 | 10 | 9 | 9 | 7 | 7 | 8 | 10 | 10 |

**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 9 newborns length (cm) and weight (kg) data:

| Nr. child | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Length | 56 | 50 | 49 | 45 | 48 | 49 | 43 | 50 | 47 |
| Weight | 4.0 | 3.5 | 3.0 | 2.1 | 2.4 | 2.7 | 2.2 | 3.4 | 2.4 |

**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?

# 6. PROBABILITY THEORY AND HYPOTHESES TESTING INTRODUCTION

- The Meaning of Probability Theory.
- Populations and Samples. The Sampling.
- Hypotheses General Concepts.
- Estimation and Hypotheses Testing.

## 6.1 The Meaning of Probability Theory

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):

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

Then, Probability of an event not to occur (q) is defend:

$$q = \frac{n - m}{n} = 1 - \frac{m}{n} = 1 - P; \quad \text{or} \quad q = 1 - P;$$

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.

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.

**Central limits theorem** states the properties of the sampling distribution:

1. The mean of the sampling distribution is identical to the population mean.
2. The standard deviation of the distribution of the sample means is equal to $\sigma/\sqrt{n}$. This quantity is known as the ***standard error of the mean*** ($SE_{\bar{x}}$; $\sigma_{\bar{x}}$);
3. Provide that "n" is large enough; the shape of the sampling distribution is approximately normal.

## 6.2 Populations and Samples. The sampling

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

***Population*** – statisticians term used to describe a large collection of items that have something in common. Almost having a meaning as a totality.

Ex.: population (as a demography term), patients, investigations, hospital discharges, etc.

***Sample*** – a segment of the population selected to represent the population as a whole.

***Sampling*** – the enquiry that utilizes special methods taking a sample.

***Study units (sampling units)*** – the individual elements in the population or sample of interest.

***Target population*** – the ideal population that you would like to describe.

***Study population*** – the group from which we can select a sample.

**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.

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 or biased. Therefore, in a 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.

***The types of sampling:***

**I. Probability sampling**

**1.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.

**2. 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.

**3.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.

**4.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.

**II. Nonprobability sampling**

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

***Convenience sample*** – the researcher selects the easiest population members from which obtain the information.

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

*Table 6-1.* **Population and Sample symbols**

| Statistics Concepts | Population Parameter | Sample Statistics |
|---|---|---|
| | TRUTH | ESTIMATE OF THE TRUTH |
| | Greek Symbols | Roman Symbols |
| Mean | $\mu$ | X |
| Standard Deviation | $\sigma$ | S |
| Variance | $\sigma^2$ | $S^2$ |
| Size (*number of observations*) | N | n |

## 6.3 Hypothesis General Concepts

### 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 clari-ty, specificity and objectivity.

**Hypotheses types:**

- ***Non-directional*** – states that relation or difference between variables exists
- ***Directional*** – states the expected direction of the relation or difference.
- ***Null*** ($H_0$) – states that there is no significance relation or difference between variables.
- ***Alternative*** ($H_0$) – is a second statement that contradicts the null hypothesis.

## 6.4 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. For make estimations the researcher uses the principles of hypotheses testing.

Statistical hypothesis testing involves stating null hypothesis and an alternative hypotheses 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.

Actually 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 and/or a confidence interval.

- Directional hypotheses are testing by One-Sided Test.
- Non-directional hypotheses are testing by Two-Sided Test.

**Confidence interval** (CI) 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; $SE_x$ – standard error

According to the rules of normal distribution, the probability 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).

***Steps in hypothesis testing:***

1. Formulation of hypotheses: $H_0$ and $H_1$.
2. Decide about the appropriate statistic test.
3. Select the level of significance for the statistic test ($\alpha$). There is a probability to reject the true $H_0$. 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 statistic test must attain for to be declared significant.
5. Perform the calculation.
6. State the conclusions.

**Hypotheses testing errors**

Two types of error may occur in hypothesis testing:

**Type I Error**

Is rejecting the null hypothesis when it is really true.

$\alpha$ is the probability to make type 1 error; concluding that the observed group difference was true effect when in fact it was du to chance or systematic error.

- As convicting the innocent man.
- False Positive rate.

**Type II Error**

Is do not reject the null hypothesis when it is false.

$\beta$ is the probability to make type 2 error; concluding that the observed group was du to a chance or random error, when in fact it was a true effect.

- As absolving the guilty man.
- False Negative rate.

**Power of the study** is the ability of a study to detect a true differrence: probability of rejecting the null hypothesis when it is really false or concluding that the alternative hypothesis is true when is really true.

Power is defined as (1- $\beta$) or (1 – a type II Error).

Power is very important in hypothesis testing. Traditional values used for Power are 0.80 and 0.95.

**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- $\alpha$) or (1 – a type I Error)

**Significance Level** ($\alpha$) 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 $H_0$ the significance level should be enough small (0.05; 0.01; 0.001).

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

**P-values**

There is a concept related to significance and to the $\alpha$ level. 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 calculated after the statistical has been performed; if the P-value is less than $\alpha$, the null hypothesis is rejected.

**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 following statistical concept: population, sample, sampling, unit of observation, target population, and study population.
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. Give the definition of hypothesis and state its types.
12. What are the confidence interval and limits?
13. Describe the two types of errors that can be made when you conduct a test of hypothesis.
14. 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.
15. What is a power of the study? What does the power mean in words?
16. What is the confidence level? What does the confidence level mean in words?
17. What is the significance level? What does the significance level mean in words?
18. What is a P-value? What does the P-value mean in words?

# 7. HYPOTHESIS TESTING PARAMETRIC AND NON-PARAMETRIC METHODS INTRODUCTION

There are parametric methods of hypothesis testing appropriate to parametric data (numerical measures) and nonparametric methods hypothesis testing appropriate to nonparametric data (ordinal or nominal measures).

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 research question in terms of statistical hypotheses.
- Data Distribution type (normal or skewed).
- Homogeneity of variance.

*Table 7.1* **Examples of test Statistics**

| Parametric | Non-Parametric |
|---|---|
| T-value:<br>one sample, paired and unpaired t-tests | Chi square test ($\chi^2$) |
| Z-value: z test | Fisher's Test |
| | U- value: Mann-Whitney test |
| | Wilcoxon rank sum |
| | McNemar test |
| **Advanced Statistical Hypothesis Tests** | |
| Repeated Measures ANOVA | |
| One way ANOVA | |
| Two way ANOVA | |
| Post Hoc Analysis | |
| ANCOVA | |
| MANOVA | |
| Simple and Multiple Linear Regression | |
| Logistic Regression | |
| Survival Analyses | |
| Nonlinear Regression | |

**Comparing Means in two groups with the t-Test**

Actually we can use t Test under the condition:

- The data fallow the normal distribution.
- Homogeneity of variance (the population variance are equal).
- Numerical measures of data.
- Having no more 2 groups.

If your data don't respect these conditions you should rather apply non-parametric tests:

*Table 7.2* **Parametric vs. non-Parametric**

| **Parametric** | **Non-Parametric** |
|---|---|
| One Sample T-test | Sign Test/Wilcoxon Signed Rank Test |
| Pared T-test | Sign Test/Wilcoxon Signed Rank Test |
| 2 Samples 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.

The t statistic for testing the mean difference in two independent groups has the difference between the means in the numerator and the standard error of the mean difference in the denominator; in symbols it is:

$$t_{observed} = \frac{D}{\sigma D} = \frac{|\bar{X}_1| - |\bar{X}_2|}{\sqrt{SE_1^2 + SE_2^2}}$$

t – observed ( for comparing to critical value - $t_{table}$ )

D – mean difference;

$\sigma_D$ – standard of the mean error difference;

$\bar{X}_1$ and $\bar{X}_2$ – compared means,

$SE_1$ and $SE_2$ standard errors of the compared means.

***Draw and state the conclusions:***

After performing the calculation of $t_{observed}$ we have to find the $t_{tabel}$ (critical value) from the special table (see tab.6.2) in order to compare it:

- If $t_{observed}$ is larger then critical value the null hypothesis is rejected such the test result is said the difference between the compared means to be statistically significant.
- If $t_{oserved}$ 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.

***Critical value*** *($t_{tabel}$)*

If the numbers of observations "n" >120 then the critical value is already known: t=1.96 (when $\alpha = 0.05$); t=2.58 (when $\alpha = 0.01$);); t=3.29 (when $\alpha = 0.001$).

If the numbers of observations "n" < 120 then the critical value is taken from the special table, according their degrees freedom which are defined as $(n_1+n_2 -2)$ (see tab.6.2).

*Table 7.3* **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 |

Source: Beth Dawson, Robert G.Trapp *Basic and Clinical Biostatistics* 2004: adapted from Table 12 in Pearson ES, Hartley HO (editors): Biometrika Tables for Satitisticians, 3rd ed, Vol 1. Cambridge University Press, 1966.

**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 representtative.
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.

$H_0$: the sample means difference are not statistically significance

$H_1$: the sample means difference are not statistically significance

$P > 0.05 \Rightarrow H_0$ accepted

$P < 0.05 \Rightarrow H_0$ rejected

**Variant 1**

Two samples (n=10 each) the cholesterol blood level results are:

| Observations number | | 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Blood Cholesterol, mg/dl | Gr.1 | 168 | 258 | 228 | 247 | 156 | 172 | 165 | 210 | 264 |
| | Gr.2 | 136 | 148 | 125 | 121 | 157 | 148 | 116 | 140 | 161 |

**Variant 2**

Two samples (n=10 each) the systolic blood pressure results are:

| Observations number | | 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| SBP, mm/Hg | Healthy | 130 | 130 | 120 | 110 | 90 | 120 | 125 | 115 | 135 |
| | Sick | 170 | 175 | 160 | 170 | 170 | 190 | 185 | 185 | 170 |

**Review questions:**

1. What are the parametric methods of hypotheses testing?
2. What are the non-parametric methods of hypotheses testing?
3. When would you prefer to apply hypotheses testing parametric methods?
4. When would you prefer to apply hypotheses testing non-parametric methods?
5. State the main considerations parametric methods vs. non-parametric methods.

## 8. THE RESEARCH METHODOLOGY INTRODUCTION

- Research Definition, Characteristics and Types.
- The steps of Research Process Contents.
- Formulating the Research Problem.
- Reviewing the Literature. Use of the Medical Database.
- The Formulation of the Aim and Objectives.

### 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 wi-thout introducing your own interest.

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:

- 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.

- 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.

(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.**

**Research** can be classified from three perspectives:

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

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

1. ***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.
2. ***Applied research*** is done to solve specific, practical questions for policy formulation and understanding of phenomena.

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

1. ***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.
2. ***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.

3. ***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.
4. ***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.
5. ***Exploratory research*** is undertaken to explore an area where little is known or investigate the possibilities of undertaking a particular research study (feasibility study/pilot study)

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

1. ***Quantitative research*** determines the extent of a problem, issue or phenomenon by quantifying the variation. The main question is: how many?
2. ***Qualitative research*** is more appropriate to explore the nature of a problem, issue or phenomenon without quantifying it. Main objecttive 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.

## 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 steps 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:***

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 recommenddations).
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 have 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:

**Step 1.** ***Identify*** a broad field of interest.

**Step 2.** ***Dissect*** the broad field into sub areas.

**Step 3.** ***Select*** what sub areas is of most interest.

**Step 4.** ***Raise*** research questions.

**Step 5.** ***Establish*** of the hypotheses.

**Step 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 are 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.

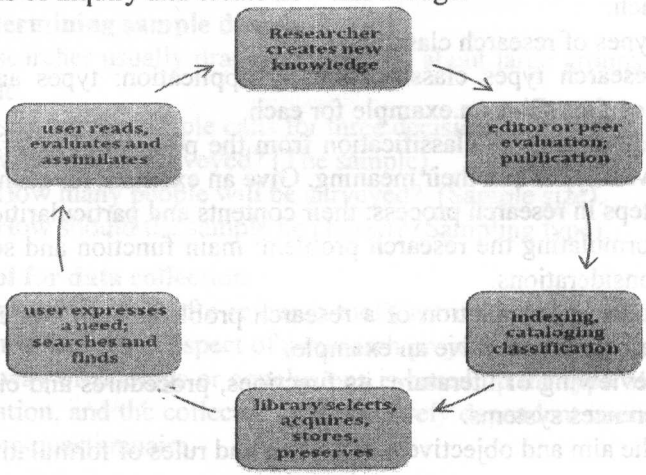



*Figure 8-2.* **Information Cycle**

Bibliographic documentation is a permanent process of professional update.

### 8.5 The 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.

**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.

# 9. PREPARING THE RESEARCH DESIGN

- Research Design Definition and steps.
- Determining Sample Design.
- Tool for Data Collection.
- Study Design Classification.
  - Case-Series Studies.
  - Cross-Sectional Studies.
  - Case-Control Studies.
  - Cohort Studies.
  - Clinical Trials Studies.

## 9.1 The research design definition and steps

| Research design is the conceptual structure within which research would be conducted. |
|---|

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 establish source of information – determining sample design.
- To elaborate tools for data collection.
- To adopt study design.

## 9.2 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).

## 9.3 Tool for data collection

The construction of a research tool (instrument) for data collection is the most important aspect of a research project. 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 objectives 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 developped 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** include 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 9-1.* **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 fallow table:

*Table 9-2.* **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 design of questionnaire:***

- Title.
- Instructions.
- General information about respondent.
- The questions.
- Thank you.

## 9.4 Study design classification

There are several different schemes of classifying study design. One of them is indicated in the followed table 9–3 and divides studies into those in which the subjects were observed, called ***observational studies***, and those in which some intervention was performed, called ***experimental studies***. 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.

*Table 9-3.* **Classification of study design**

<table>
<tr><td rowspan="5">Descriptive studies</td><td rowspan="4">I. Observational studies</td><td>1.Case series</td></tr>
<tr><td></td><td>2.Cross-sectional</td></tr>
<tr><td rowspan="3">Analytical studies</td><td>3.Case-control</td></tr>
<tr><td>4.Cohort</td></tr>
<tr><td rowspan="2">II. Experimental studies</td><td>1.Controlled trials</td></tr>
<tr><td>2.Studies with no controls</td></tr>
<tr><td colspan="3">III. Meta-analyses</td></tr>
</table>

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 e ethical considerations.

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. 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 then 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 subject at one time.

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.

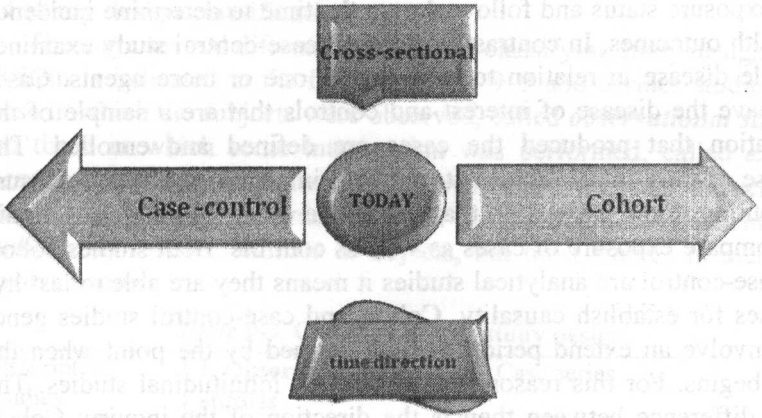



*Figure 9-1.* **Schematic diagram of the time relationship among different observational study designs.**

**9.5 Case series studies**

It is the simplest design in which the author describes some interesting or intriguing observations that occurred from a small number of patients. When certain characteristics of a group (or series) of patients (or case) 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.

**9.6 Cross-sectional studies** (Prevalence study, transversal study, survey).

This type of observational study analyzes data collected on a group of subjects at one time rather than over a period of time. 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 od time).

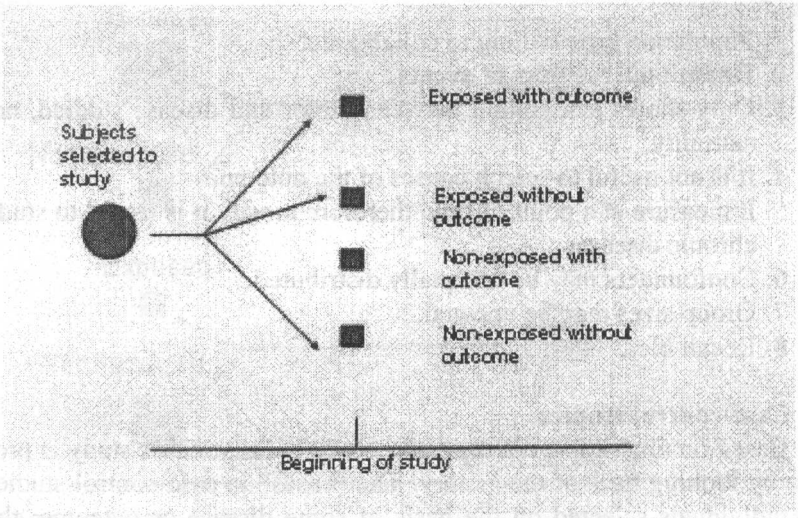



*Figure 9-2.* **Schematic diagram of cross-sectional study design**

Cross-sectional study is used for measure prevalence of a disease and look at potential risk factors or cause. Because cross-sectional studies examines 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 to chi-square and other non-parametric tests, t test; analysis of variance; 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 measure 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.

**9.7 Case-control studies**

Based on time it is a ***retrospective*** study because cases study is provide by looking back at the history. 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 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. 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, employment.
- Interview.
- Self administrated questionnaire.
- Direct measurement.

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

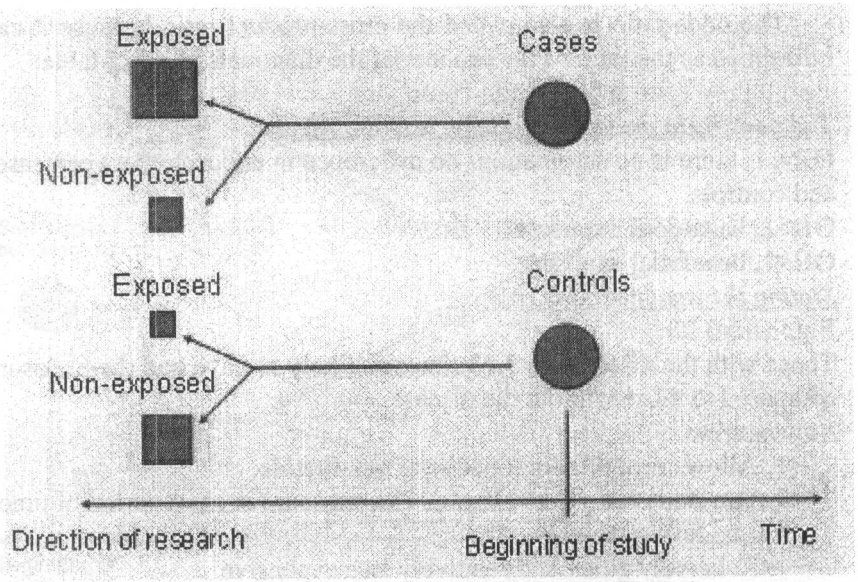



*Figure 9-3.* **Schematic diagram of case-control study design**

***Analysis*** of case-control study includes the calculation of the association measure called ***"odds ratio"***. An odds is defined as the probability that an event will occur divided by the probability that the same event will not occur.

Odds Ratio = (odds event 1)/(Odds event 2).

The odds ratio provides a way to look at risk in case-control studies. The odds ratio is easy to calculate when the observations are given in a 2x2 table:

*Table 9-4.* **Case –control study multiple 2x2 table**

| **Exposure factor** | **Outcome (disease)** | |
|---|---|---|
| | Yes (cases) | No (controls) |
| **Yes** | a | b |
| **No** | c | d |
| **Total** | a+c | b+d |

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

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.

*Stating the result of odds ratio:*

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.

**9.8 Cohort studies** (Incidence study, longitudinal study, 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. 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?".

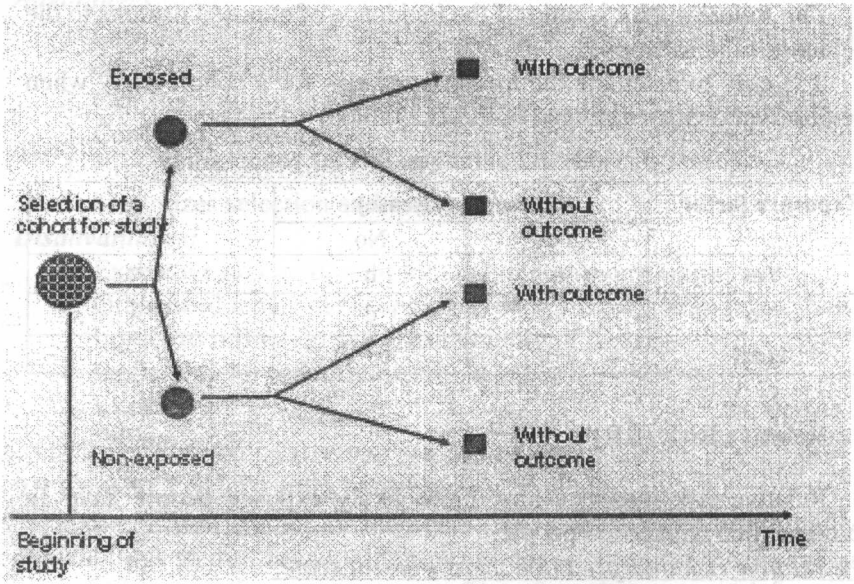



*Figure 9-4.* **Schematic diagram of cohort study design**

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 non-exposed.

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

*Table 9-5.* **Table 2x2 arrangements for cohort study**

| Exposure factor | Outcome (disease) | | total |
|---|---|---|---|
| | Yes | No | |
| **Yes** | a | b | a+b |
| **No** | c | d | c+d |
| **Total** | a+c | b+d | |

$$\textbf{Relative Risk (RR)} = \frac{a/a+b}{c/c+d}$$

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

*The relative risk interpretation measurement scale:*

**RR= 1,** null value (no association) – no difference in disease between exposed and unexposed groups.

**RR>1**, hazardous exposure

**RR<1**, beneficial exposure

*Stating the result of relative risk:*

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.

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

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.

*Stating the result of attributable risk:*

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.

**9.9 Clinical Trials Studies** (experimental studies).

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 into big groups:

**I. Controlled trials:**

1. Parallel or concurrent controls:
   *a) Randomized.*
   *b) Not randomized.*
2. Sequential controls:
   *a) Self-control.*
   *b) Crossover.*
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.

**1. 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):

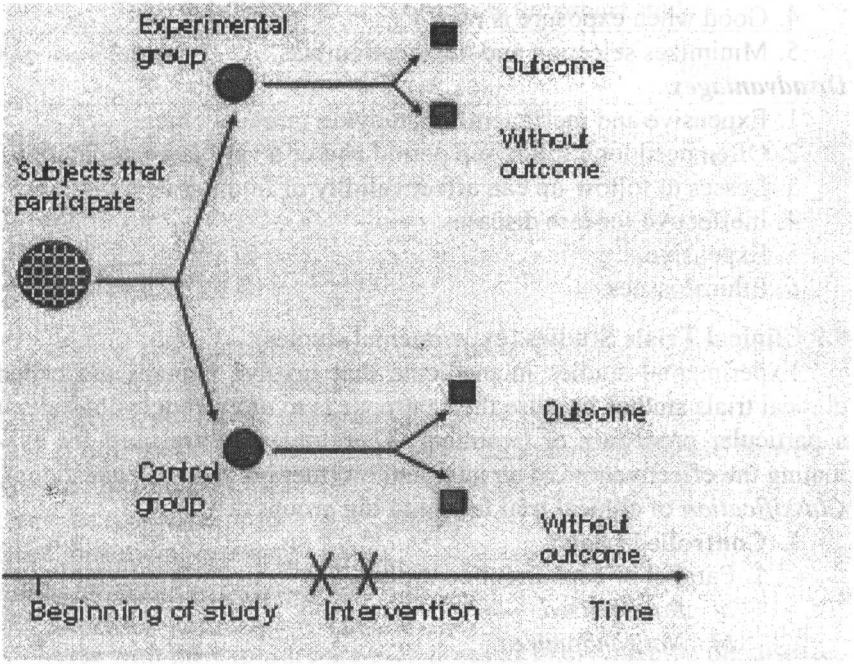



*Figure 9-5.* **Schematic diagram of randomized controlled trial design.**

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 planed for the same time period and the same study.

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 intervenetions only.

**a) 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 group, than this a way to reduce bias.

**b) 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.

**2. Sequential controlled trials**

**a) Self-Control Trials** 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.

**b) 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 (wash out period) without it. Then the intervention (1) is administrated to control group and intervenetion (2) is administrated to experimental group.

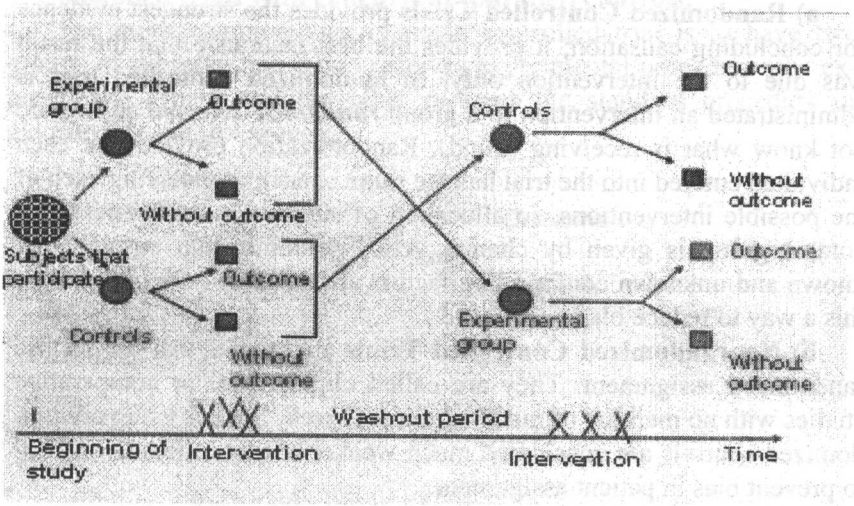



*Figure 9-6.* **Schematic diagram of crossover trial (sequential controlled) design**

**3. Trials with External Controls** are the studies when investigators compare the results of another researcher or with the results of a previous study. Also, they are called historical controls.

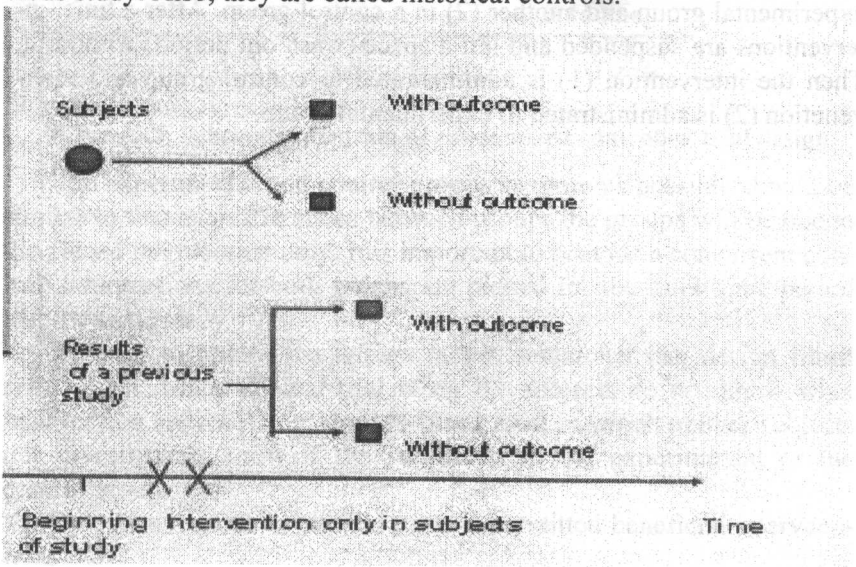



*Figure 9-7.* **Schematic diagram of trial with external control design**

***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 9-6.* **Table 2x2 arrangements for clinical trial study**

| **Outcome** | **Exposure factor (intervention)** | | **total** |
|---|---|---|---|
| | Experimental treatment | Placebo | |
| **Yes** | a | b | a+b |
| **No** | c | d | c+d |
| **Total** | a+c | b+d | a+b+c+d |

*1. **Experimental Event Rate (EER)*** – rate of event (risk of disease) in a group of subjects that received experimental treatment:

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

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

$$\mathbf{CER} = \frac{b}{b+d}$$

*3. **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.

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

*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.

*4. 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.

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

e.g.: In the aspirin study the EER for an 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 = 0.30 – 0.14 = 0.16.

The interpretation:

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.

*5. 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.

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

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.

*6. Relative Risk Reduction (RRR)* – defined as amount of risk reducetion relative to the baseline:

$$\mathbf{RRR} = \frac{CER - EER}{CER} = 1 - \mathrm{RR}$$

e.g.: In the previous aspirin study

$$\mathrm{RRR} = \frac{0.30 - 0.14}{0.30} = 0.53 \text{ or } 53\%.$$

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%.

*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.

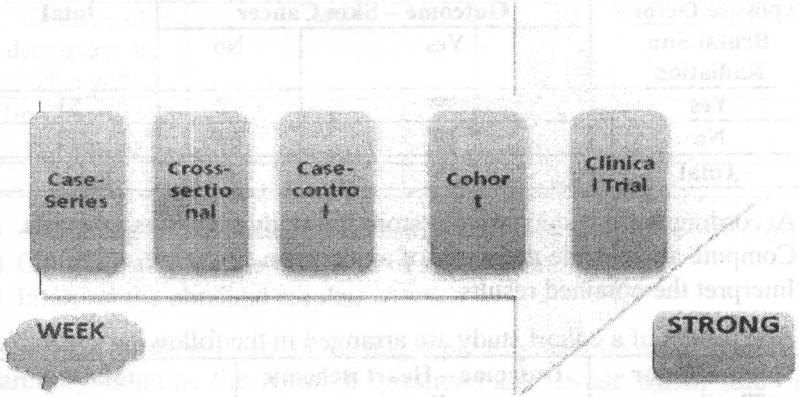



*Figure 9-8.* **Schematic diagram of study design evidence strength**

**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.

**3.** 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 |

a) According these results try to restore the study scenarios in words.

b) Compute all possible measures of association.

c) Interpret the obtained results.

**4.** 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 |

a) According these results try to restore the study scenarios in words.

b) Compute all possible measures of association.

c) Interpret the obtained results.

**5.** 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 |

a) According these results try to restore the study scenarios in words.

b) Compute all possible measures of association.

c) Interpret the obtained results.

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

| **Outcome** | **Exposure (intervention)** | | **total** |
|---|---|---|---|
| | Metistatin (Anticholesterolemiant) | Placebo | |
| Cholesterol decreasing by 20 mg/dl | 43 | 39 | 82 |
| No Cholesterol decreasing | 3 | 8 | 11 |
| **Total** | 46 | 47 | 93 |

a) According these results try to restore the study scenarios in words.
b) Compute all possible measures of association.
c) Interpret the obtained results.

**7.** 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.

Is that the study design used by the investigators?

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

**Review questions:**

1. Research design definition and steps.
2. Steps of questionnaire construction.
3. Basic types of questions structure. Their contents.
4. Types of survey methods. Their contents.
5. General design of questionnaire.
6. Study design classification.
7. Observational versus experimental study design: the meaning and particularity. Advantages and disadvantages. Give an example.
8. Which type of study design is best depending on research questions?
   a. Therapy question.
   b. Diagnosis/screening.
   c. Prognosis.
   d. Occurrence.
   e. Causation.

9. State the main difference between the following study design: observational descriptive and observational analytic. Give an example.
10. State the main difference between the following study design: case-control and cohort study. Give an example.
11. State the main difference between the following study design: case-series and case-control. Give an example.
12. The main association measures of cross-sectional study: definition and meaning interpretation.
13. The main association measures of case-control study: definition and meaning interpretation.
14. The main association measures of cohort study: definition and meaning interpretation.
15. Clinical trial definition and classification.
16. What does mean to be controlled and uncontrolled clinical trial?
17. What does mean to be randomized clinical trial?
18. Controlled clinical trials with concurrent controls: definition, types and schematic diagram.
19. Sequential clinical trials with concurrent controls: definition, types and schematic diagram.
20. Trials with historical control: definition and schematic diagram.
21. The main analyses measures of clinical trial study: definition an d meaning interpretation.
22. State the study design evidence strength.

## 10. REPORTING THE FINDINGS OF THE RESEARCH

- Writing research report
- Oral presentation of medical research

### 10.1 Writing research report

Writing the report is the last step of the research process. The reports informs 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:

*Title page:*

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

*Project body:*

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

### 10.2 Oral presentation of medical research

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.

*Figure and tables:*

Same principles as the writing report.

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

## Bibliography


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

Pagano M., Gauvreau K., *Principles of Biostatistics*, Second Edition, Belmont, CA, USA, 2000.

Berry G., Matthews JNS, Armitage P., Statistical Methods in Medical Research, 4th Edition, Blackwell Scientific, 2001.

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

Daniel W.W., Bistatistics: *A Foundation for Analysis in the Health Sciences*, 7th ed. Wiley, 1998.

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.

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.

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.

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.

Tintiuc D,Badan V, Raevschi Elena, Grossu Iu, Grejdeanu T, Vicol Corina, Margine L: Biostatistica si Metodologia Cercetarii Stiintifice, CEP Medicina, Chisinau, 2011.

Weinstein MC, Fineberg HV: Clinical Decision Analysis. WB Saunders, 1998.

## Appendices

1. Project presentation guidelines.
2. Potential Student Topics for Project.
3. Project scoring.

### *1. Project Presentation Guidelines*

The purposes of this project are to:

1. Learn something new about a specific chronic disease.
2. Synthesize information in order to present it in a concise, coherent way to others.
3. Gain public presentation skills.

The topic for your presentation have to be chosen from the followed list or another topic with the approval of the course instructor. The presentation topic must be approved by your course instructor by October 15. First come, first serve since only one presentation will be allowed per topic from groups working together at the class.

Time your presentation to make sure it takes no longer than 15 minutes. A 5-minute discussion period will follow so consider potential questions.

- **TERM presentation**

Each student must prepare a presentation on a subject of their choosing. Students must prepare their oral presentation (15 min) according approved topic by December 1st, as the presentations will be provided during the December classes, according the approved schedule by November 15. All changes in the schedule for December presentation are allowed by the end of November only with permission of the course instructor. The students having been missed their scheduled presentation will be qualified as a failing grade.

**Resources:**

**CDC statistics** http://www.cdc.gov/DataStatistics/
**WHO statistics** http://www.who.int/whosis/en/
**Healthy People 2010 Info Access Project** (pre-formulated searches): http://phpartners.org/hp/
**National Library of Medicine Selected Data Tools and Stats:** http://www.nlm.nih.gov/hsrinfo/datasites.html
http://www.healthierlongerlife.org/ etc.

PowerPoint slides – no more than 7 lines of text per slide.
Use slides as a reminder of what you want to talk about rather than reading from the slide word-for-word.
Use font size of 32+ in general but footnotes can be smaller.
Between 12–18 slides total (general rule is 1 slide per minute).
Make sure the colors you choose project well.
Use a footnote to reference data citations, tables, and figures, etc that have been taken from a copyrighted publication.
Recommended footnote formats:

1. Jones P, et al. *N Engl J Med* 2005; 352: 234–238 **or.**
2. Jones P, et al (2005).
3. American Heart Association (2006) www.amheart.org.

References & other resources should be listed on a final slide in the full bibliographic citation format:

1. Jones P, Smith X, Peters Q, et al. Comparison of observational stu-dies and randomized, controlled trials. *N. Engl J. Med.* 2005; 352: 234–238.
2. American Heart Association. *Heart Attack and Stroke Facts;* 2006. http://www.amheart.org. Accessed April 1, 2006.

Include (see project scoring sheet for more details on content):
**1 slide** – Title slide with your name & topic
**1–3 slides** – Define the disease – how is it defined?
– Types of disease or tumors (if applicable).
– Most important aspects of biology, pathology, clinical presentation, long-term complications, and diagnosis.
**1–5 slides** – Provide an overview of the epidemiology – this would include incidence, prevalence, and survival in:
– Population (your country/global).
– Pertinent subgroups by age, sex, race/ethnicity, region, etc (as applicable).
**2–5 slides** – Describe exposures or risk factors for:
– Population.
– Subpopulations if a group has significantly higher incidentce/prevalence of disease.
**1–3 slides** – Prevention/treatment/screening (as applicable).
**1–2 slides** – Summary /conclusions and future research directions.
**1 slide** – References.

**PowerPoint Tips** ***(from the American Heart Association)***
The graphics you project on the screen to support what you say should help clarify ideas, emphasize key points, show relationships, and provide the visual information your audience needs to understand your message.

**Here are a few suggestions:**

- Keep the **WORDING** clear and simple.

– Abbreviate your message. Don't use complete sentences. Outline the thought to provide a memory trigger.
– Try to follow the "7–7" rule. Create up to seven words across, seven lines down.
– Don't hesitate to continue the same title on the next slide with more information.
– The more you break this rule, the less the audience will see.
– Use only two font styles per slide. A typical design might be to use Times-Roman for your titles and Ariel for the text below.
– Upper and lower case lettering is more readable than all capital letters. Current styles indicate that using all capital letters means you are shouting.
– In bullet point lines, capitalize the first word and no other words unless they normally appear capped.

- Keep the **DESIGN** consistent and appealing to the eye.

– Use colors sparingly; two to three at most. Be consistent from slide to slide. For example, use one color for all of your titles, another for the text body, etc.
– Light backgrounds work. White can be too bright depending on competing light levels in the room. Light browns and blues work well. Lettering could be black, dark blue, dark purple or dark green.
– Dark backgrounds work, too. The most effective background colors are blue, turquoise, purple, magenta, teal, etc. Lettering could be white, yellow, cyan, pink and lighter versions of most other colors.
– Don't use red in any fonts or backgrounds. It is an emotionally overwhelming color that is difficult to see and read.
– Graduated backgrounds are more interesting than plain ones.
– Textured backgrounds can add style to your talk. Simple, light textured backgrounds work well. Complicated textures make the content hard to read.

- Graphics, illustrations, cartoons, artwork and photographs will bring another dimension to your presentation. Determine your comfort level and match the graphics with your message and your speaking style.
- A note about photographs: when importing pictures, be sure they are no larger than 2 megabytes and are in a **.jpg** format. Larger files can slow down your show.
- Experiment with animating the transition between slides as well as animating the content within the slide.

• Keep **GRAPHS, CHARTS AND DIAGRAMS** simple, if possible.

You have been staring at your data for months. The audience only gets a minute. Simple graphs, etc. are absorbed more quickly than complex, cluttered ones.

- Use bar graphs and pie charts instead of tables of data. The audience can then immediately pick up relationships.
- Place labels outside pie charts.
- Simplify scales on the X- and Y-axis.

• **GENERAL HINTS.**

- If using a laser pointer, don't move it too fast. For example, if circling a number on the slide, do it slowly.
- Don't point the laser at the audience.
- Look at the audience, not at the slides, whenever possible.
- Most slide programs are made to be user friendly so that everyone can make their own slides. However, if possible, don't hesitate to ask a professional slide production person to go over your show and offer hints.
- Run "spell check" on your show when finished.

*Your goal is to design a presentation that delivers your message clearly, efficiently and in an interesting manner. How you design your show will reflect your speaking style and your personality. You efforts will pay off in the long run and ensure that each presentation will communicate the importance of the content and the passion you have talking about it to your audience.*

## *2. Potential Student Topics for Project*

***Autoimmune diseases:***

Systemic lupus erythematosis.
Rheumatoid arthritis.
Ankylosing spondylitis.
Sarcoidosis.

***Cancer:***

Lung, Colorectal, Breast, Cervical, Prostate, Bladder, pharyngeal and oral cavity cancer, Melanoma, Limphoma, Leukemia, Hepatocellular carcinoma. Other skin cancers, Ovarian cancer, Pancreas cancer, Sarcomas (bone), Stomach cancer, Uterine/endometrial cancer.

***Cardiovascular diseases:***

Coronary Heart Disease.
Cerebrovascular Disease.
Heart Failure.
Peripherial Arterial Disease.
Atrial fibrillation.
Rheumatic heart disease.
Aortic stenosis.
Deep vein thrombosis/pulmonary embolism.
Congenital heart disease.

***Chronic Respiratory Diseases:***

Asthma.
Cystic fibrosis.
Sleep apnea.
Primary pulmonary hypertension.
Workers' Pneumoconiosis.
Allergic alveolitis.

***Eye:***

Cataracts.
Glaucoma.
Macular degeneration.

***Gastrointestinal:***

Peptic ulcer disease.
Chronic pancreatitis.

Gallbladder disease.
Celiac disease.
Ulcerative colitis.
See Liver below.

***Hematologic disorders:***
Sickle cell anemia.
Hemachromatosis.

***Hormone disorders:***
Hypothryroidism.

***Infectious:***
HIV/AIDS.

***Kidney disease:***

***Liver Diseases:***
Alcoholic liver disease.
Cirrhosis.
Hemachromatosis.
Hepatitis A.
Hepatitis B.
Hepatitis C.
Nonalcoholic fatty liver disease.
Other chronic hepatitis.

***Mental Disorders:***
Schizophrenia.
Mental retardation.
Depressive disorders.

***Musculoskeletal Diseases and arthritis :***
Osteoarthritis.
Rheumatoid arthritis.
Fibromyalgia.
Gout.
Osteoporisis.

***Neurologic disorders:***
Epilepsy.
Parkinson's disease.
Alzheimer's disease.
Autism spectrum disorders.
Traumatic brain injury.

Spinal cord injury.
Fibromyalgia.
Migraine.
Myasthenia gravis.

***Psychiatric:***

Anorexia nervosa.
Bulimia.
Domestic abuse (adult).
Child abuse.

***Skin diseases:***

Psoriasis.

## *3. Project Scoring*

B/RM Fall 2012

Student Name: _____ gr.nr _____ Date: ___ / ___ / ___

Project Scoring (100 points)

**1. Define the disease (10 points)** ----------

- Describing the types of disease or tumors.
- Describe the most important aspects of biology, pathology, clinical presentation, and diagnosis.
- Describe long-term complications.

**2. Incidence, prevalence, other statistical indicators (40 points)** ____

Provide descriptive data on:
10-Population as a whole.
Provide numeric information when possible:

- rude rate Cases per unit of population as appropriate (e.g. 2/100,000).
- roportion of population affected, if applicable (e.g., a relatively common disease).

10-Populations at risk [Age, sex, race/ethnicity, region (country and world if available), etc.) Specific rate.
10-Relative importance compared to other chronic diseases.
10-Explanation of indicators.

**3. Describe exposures or risk factors (30 points)** _____

10 - Relative and attributable (if available) risks associated with each risk factor.
10 - Relative importance for different subpopulations, if applicable (e.g., if disease rates in an age group/sex/racial ethnic group/region are markedly different, discuss whether risk factor prevalence differs between/among groups as a potential explanation).
10 - Explanation of Relative and attributable risks meaning in the topic context

**4.Provide an overview of what can be done: (6 points)** _____

2-Prevention – is it possible?
    Primary
    Secondary
2-Treatments
2-Screening - is it appropriate?
If so, who should be screened – population, high risk groups, relatives, etc).

**5. Summary (4 points)** _____

- What is known of the disease to date.
- Future areas for investigation.

**6. Formatting and presentation (10 points)** _____

__ of 2 points for organization
__ of 2 points for spell checking
__ of 2 points for well formatted slides (subtract points if too small to read, too much info)
__ of 2 points for referencing material used on slides
__ of 2 points for staying within time limit

**NOTES:**
**Total** _____