MINISTRY OF HEALTHCARE OF THE REPUBLIC OF KAZAKHSTAN

NCJSC “Astana Medical University”

Sultanova Zh.D., Zhunissova U.M.

STATISTICAL ANALYSIS OF BIOMEDICAL DATA USING OF MS EXCEL AND SPSS STATISTICS PROGRAMS

Tutorial (part 1)

Nur-Sultan, 2021

UDC 61:519.25:004(075.8)
LBC 51.1я73
S 92

REVIEWERS:

- **Aimakhanova A.Sh** - KazNMU named after S. D. Asfendiyarov, can.ph-math.sc., associate professor, head of the chair of Biostatistics and Basics of Scientific Research.
- **Y.D. Dalenov** - NJSC «AMU», doctor of medical science, professor, head of the chair of preventive medicine nutrition.

Authors: Sultanova Zh.D., Zhunissova U.M.

S 92. Statistical analysis of biomedical data using of ms excel and SPSS Statistics programs. Tutorial / Zh.D. Sultanova, U.M. Zhunissova; NJSC Astana medical university, Nur-Sultan, 2021. - 80 p.

The tutorial corresponds to the standard educational program on the discipline of Biostatistics for medical specialties of higher educational institutions. It is intended for medical students studying in English to learn the basics of biostatistics course. And includes basic concepts of biostatistics, descriptive statistics, sample method, normal distribution law. As well as it contains practical works using the statistical software package MS Excel and SPSS Statistics.

The tutorial can be used by students in preparation for practical classes, tests and exams.

UDC 61:519.25:004(075.8)
LBC 51.1я73

Approved and recommended for publication by the Quality Assurance Committee of educational programs of NJSC “Astana Medical University” as additional educational literature.

Protocol number ____ from _____ «____», 20 y.

CONTENTS

List of notations

1. $X_1, X_2, \ldots X_i$ - values of random variable X
2. $\overline{X}$ - Mean or arithmetic mean or sample mean (read as X bar)
3. ν - statistical probability
4. P(A) - probability of the event A
5. $m_1, m_2, m_3, m_4, m_5 \ldots m_к$ - frequencies
6. n – total number of observations (sample size)
7. S_i - cumulative frequency
8. *M (X)* – expected value
9. μ - population mean
10. M_e - median
11. M_o - mode
12. D (x) –variance
13. σ – standard deviation
14. α - significance level
15. t - Student t-test
16. *df* – degree of freedom

PREFACE

Statistics is the science where by inferences are made about specific random phenomena on the basis of relatively limited sample material. The field of statistics has two main areas: mathematical statistics and applied statistics. Mathematical statistics concerns the development of new methods of statistical inference and requires detailed knowledge of abstract mathematics for its implementation. Applied statistics involves applying the methods of mathematical statistics to specific subject areas, such as economics, psychology, and public health. Biostatistics is the branch of applied statistics that applies statistical methods to medical and biological problems.

This tutorial introduces students to statistical methods of data presentation and processing. It is our aim to introduce basic concepts of statistics with sound theoretical explanations. Because biostatistics is basically an interdisciplinary applied subject, we offer many applied examples. We present examples of solving statistical problems using MS Excel and SPSS.

This introductory-level biostatistics is designed for students interested in medicine or other health-related areas.

1. INTRODUCTION TO BIOSTATISTICS

The modern study of the life sciences includes experimentation, data gathering, and interpretation. This text offers an introduction to the methods used to perform these fundamental activities.

The design and evaluation of experiments, known as the scientific method, is utilized in all scientific fields and is often implied rather than explicitly outlined in many investigations. The components of the scientific method include observation, formulation of a potential question or problem, construction of a hypothesis, followed by a prediction, and the design of an experiment to test the prediction.

1.1. DEFINITION OF BIOSTATISTICS

Statistics is the science of collecting, organizing, and interpreting numerical facts which we call data. The goal of statistics is to obtain information from data.

Biostatistics is a system of the knowledge about planning rules and analysis of the results of the experimental researches of biological objects. Studying biostatistics prepares you for designing biological experiments and clinical trials.

Biostatistics combines two fields of study:

Bio stands for biology, the study of living thinks, and

Statistics is concern with the collection, tracking, organization, summary, classification, analysis and evaluation of facts to reach some inference.

Biostatistics is the application of statistical methods to a wide variety of fields of biology or life sciences including human biology, agriculture, veterinary, microbiology and genetics.

Biostatistics also called biometry, literally meaning biological measurement. Biometry has Greek origin (bios = *life* + *metron* = measure). Although, two terms are used interchangeably, biometry is largely used for biological and agricultural applications and biostatistics for medical usages. It involves development and application of statistical techniques to scientific researches in health related fields including medicine, epidemiology, public health and areas of environmental study

Biostatistics is the science which deals with development and application of the most appropriate methods for the:

- ❖ Collection of data.
- ❖ Presentation of the collected data.
- ❖ Analysis and interpretation of the results.
- ❖ Making decisions on the basis of such analysis

Method of analysis is a collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

1.2. HISTORY OF BIOSTATISTICS

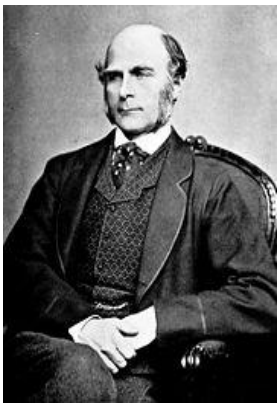
Statistics has ancient roots and is characterized initially by the first censuses of the population, its property, lands in the ancient and Middle Ages. These studies were of great practical importance, usually military or tax.

Biological statistics, as a scientific method of Biomedicine, in its historical development has passed a long and difficult way: from the verbal description of biological objects - to their measurement, from simple statistical data and tables – to the systematic statistical analysis of mass phenomena in biology and medicine. At the initial stages of its formation biostatistics was not yet a separate science, it developed for a long time in the structure of General statistics and only in the XIX century received the right to independent existence.

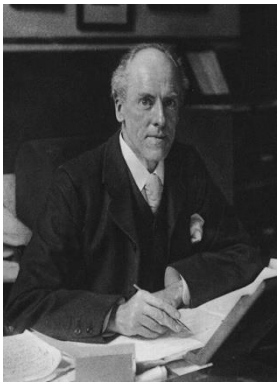
Biostatistics reasoning and modeling were crucial for the fundamental theories of modern biology. Leading figures in biostatistics relied on statistics and developed its use in biology.

Adolphe Quetelet (1796-1874), the Belgian astronomer and mathematician used statistical methods for the first time to the problems of biology, medicine and sociology and developed statistical theory as a general method of research, applicable to any observational science.

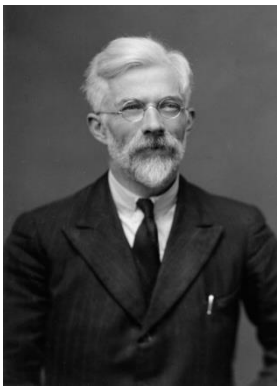
He was the first person to apply the statistical normal distribution to characteristics of human populations and he introduced the height-weight measure we know today as the body mass index.



Francis Galton (1822 –1911), cousin of Charles Darwin, was an English Victorian polymath, inventor, proto-geneticist, psychometrician, and statistician. He is called the **«Father of biostatistics».** With an IQ of two hundred, Galton was a best-selling author, produced over 340 papers and books. He also created the statistical concept of correlation and widely promoted regression toward the mean. He was the first to apply statistical methods to the study of human differences and inheritance of intelligence.



Karl Pearson (1857 - 1936), was important in the founding of the school of biometrics, which was a competing theory to describe evolution and population inheritance at the turn of the 20th century. In fact, Pearson devoted much time during 1893 to 1904 to developing statistical techniques for biometry. These techniques, which are widely used today for statistical analysis, include the chi-squared test, standard deviation, and correlation and



regression coefficients.

Ronald Fisher (1890 –1962) was a British statistician and geneticist. For his work in statistics, he has been described as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century statistics". From 1919 onward, he worked at the Rothamsted Experimental Station for 14 years; there, he analyzed its immense data from crop experiments since the 1840s, and developed the analysis of variance (ANOVA). He established his reputation there in the following years as a biostatistician. He is known as one of the three principal founders of population genetics. He outlined Fisher's principle. His contributions to statistics include the maximum likelihood, the derivation of various sampling distributions, founding principles of the design of experiments, and much more.

1.3. STATISTICAL TERMS

Observation is a *value of something of interest* you're measuring or counting during a study or experiment: a person's height, a bank account value at a certain point in time, or number of animals. An observation is usually denoted by the letter *X*. Each of these observational units (X) represents data from a single observation.

A **population** is the collection or set of all objects or measurements that are of interest to the collector. Population refers to any well defined group of individuals who are being studied or the total numbers of observations of a particular type about which inferences are to be made. Total number of observations in the population is called *a population size* and denoted by ***N***.

For example, all university students in Astana could be a population. All patients of a hospital suffering from tuberculosis, patients suffering from breast cancer and treated a new drug or hemoglobin levels of all adult woman of a community may be considered as a population.

In real-world problems it is usually not possible to obtain information on the entire population. The primary objective of statistics is to collect and study a subset of the population, called a sample, to acquire information on some specific characteristics of the population that are of interest.

The **sample** is a subset of data selected from a population. The **size** of a sample is the number of elements in it. Researches want to find out specific feature about a population, but it is not possible to study every single individual in the population. They select a small number of individuals from the population, study them and use that information to draw conclusions about the whole population. Total number of observations in the sample is called *a* sample size and denoted by ***n***.

For example, if all university students in Astana is a population, then Astana medical university students is a sample.

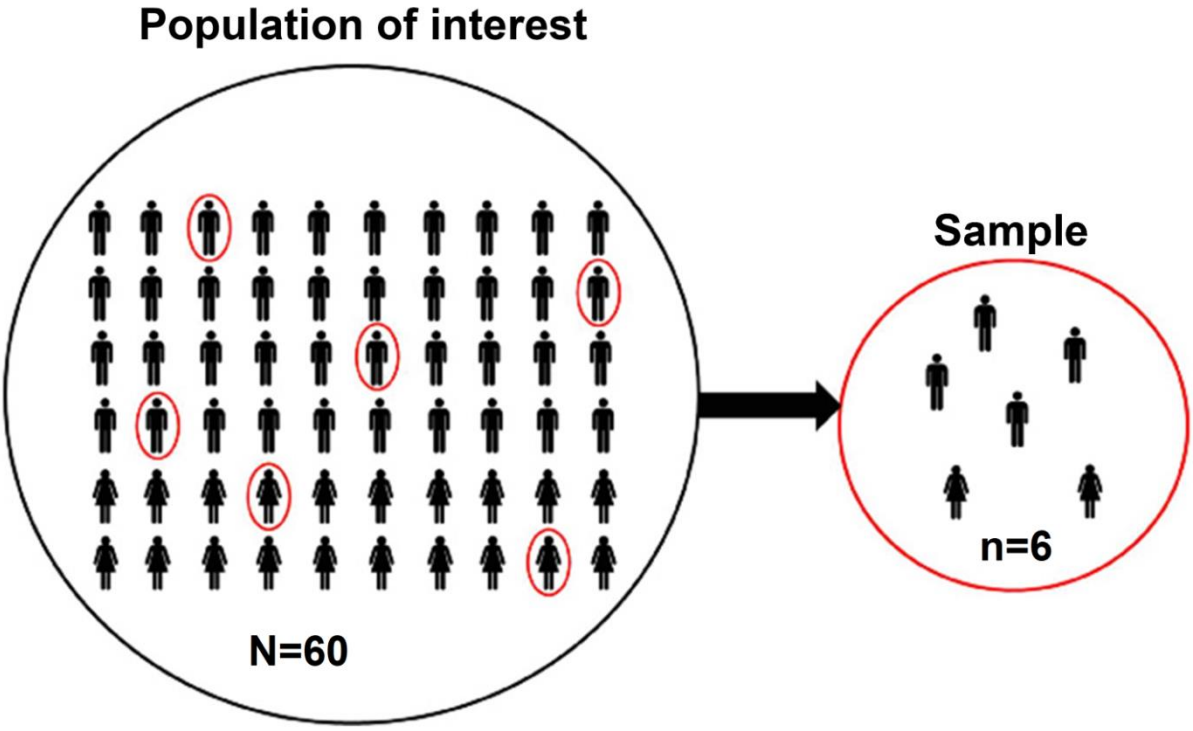


Figure 1. Population and sample

In any statistical analysis, it is important that we clearly define the target population. The population should be defined in keeping with the objectives of the study. When the entire population is included in the study, it is called a *census* study because data are gathered on every member of the population.

In general, it is usually not possible to obtain information on the entire population because the population is too large to attempt a survey of all of its members, or it may not be cost effective.

A small but carefully chosen sample can be used to represent the population. A sample is obtained by collecting information from only some members of the population. A good sample must reflect all the characteristics (of importance) of the population. Samples can reflect the important characteristics of the populations from which they are drawn with differing degrees of precision. A sample that accurately reflects its population characteristics is called a *representative* sample. A sample that is not representative of the population characteristics is called a *biased* sample. The reliability or accuracy of conclusions drawn concerning a population depends on whether or not the sample is properly chosen so as to represent the population sufficiently well.

There are many sampling methods available. We mention a few commonly used simple sampling schemes. The choice between these sampling methods depends on (1) the nature of the problem or investigation, (2) the availability of good sampling frames (a list of all of the population members), (3) the budget or

available financial resources, (4) the desired level of accuracy, and (5) the method by which data will be collected, such as questionnaires or interviews.

A sample selected in such a way that every element of the population has an equal chance of being chosen is called a *simple random sample*. Equivalently each possible sample of size n has an equal chance of being selected.

Parameters in statistics are an important component of any statistical analysis. A parameter is number that summarizes data for an entire population or some aspect of it. This means the parameter is a characteristic of a population. Statistics are numbers that summarize data from a sample. If you collect a sample and calculate the mean, this is the sample statistics. The statistic is a characteristic of a sample.

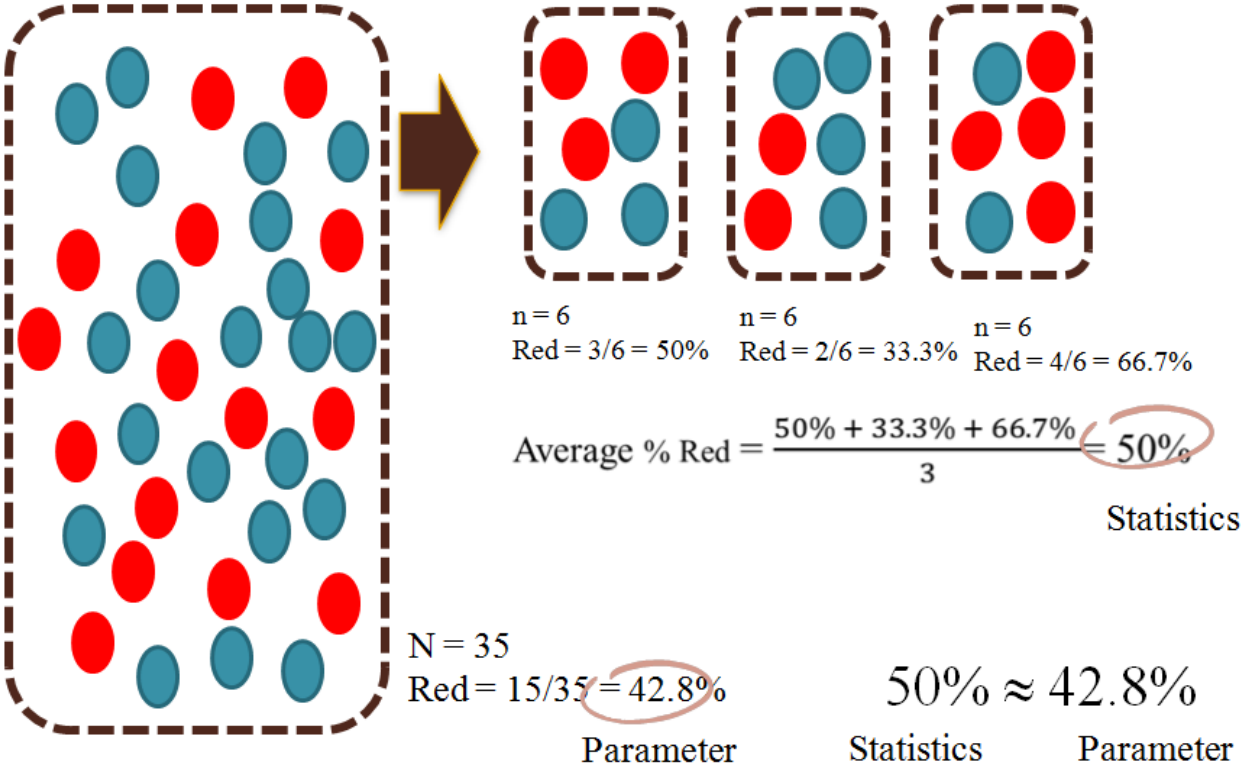


Figure 2. Parameter and Statistics

Variable is a quality or a characteristic which is being observed or measured and can vary from one individual to another. For example, animals of same species may differ in the length, weight, age, sex and etc.

The particular values of a variable are termed as variates. Variate is a quantity having a numerical value for each member of a group, especially one whose values occur according to a frequency distribution.

Random variable is a variable which possible values are numerical outcomes of a random phenomenon. Random values are denoted by capital Latin letters: A, B, C, ... X, Y. The study of random variables is the subject of statistical research.

The basis of biostatistics is the theory of probability. Probability theory is a science that studies the regularities of the behavior of random variables and

mass events. In probability theory, an event is a set of outcomes of an experiment to which a probability is assigned. There are three types of events: certain, impossible and random.

A certain event is an outcome event that is predetermined or predictable.

An impossible event is an outcome event that will not happen in a certain set of conditions.

A random event is an event that, with the implementation of a certain set of conditions, may or may not occur. Random events are denoted by capital Latin letters: A, B, C…

Examples of random events:

Event A - the presence of high blood pressure in the student during the examination.

Event B - low concentration of leukocytes in the blood.

Quantitative assessment of the possibility of the occurrence of this random event is **the *probability***. Probability is a subject that deals with uncertainty. In everyday terminology, probability can be thought of as a numerical measure of the likelihood that a particular event will occur. Probability values are assigned on a scale from 0 to 1, with values near 0 indicating that an event is unlikely to occur and those near 1 indicating that an event is likely to take place. A probability of 0.50 means that an event is equally likely to occur as not to occur. The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

The classical definition of the probability of the event (A) is the ratio of the number (m) of cases favorable to it, to the number of all possible outcomes (n).

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

Events and their probabilities:

- The probability of a certain event is 1.
- The probability of an impossible event is 0.
- The probability of a random event is $\mathbf{0 \leq P(A) \leq 1}$.

The normalization condition of probability of a random events says: $\sum_1^i \mathbf{P_i} = \mathbf{1}$.

1.3.1. STATISTICAL DATA

Statistical data is any record, descriptive or qualitative account or symbolic representation of any attribute, event or process expressed in quantitative form. The scientific record of the results or observations of an experiment or a series of experiments is also called data. Data are expressed on a numerical scale or are classified into categories that are numerically coded. For example, the heights and weights of students from a class represent the data. Thus, the statistical data are the set of facts about an attribute or character expressed in quantitative, qualitative or numerical form, i.e., it is set of values on one or more observational units.

Variable types can be distinguished based on their scale. Typically different statistical methods are appropriate for variables of different scales.

Table 1.Types of measurement scales

Scale	Characteristic Question	Examples
Nominal	Is A different than B?	Marital status Eye color Gender Religious affiliation Race
Ordinal	Is A bigger than B?	Stage of disease Severity of pain Level of satisfaction
Interval	By how many units do A and B differ?	Temperature SAT score
Ratio	How many times bigger than B is A?	Distance Length Time until death Weight

Table 2.Operations that make sense for variables of different scales

	Operations that make sense			
Scale	Counting	Ranking	Addition/ Subtraction	Multiplication/ Division
Nominal	√			
Ordinal	√	√		
Interval	√	√	√	
Ratio	√	√	√	√

Data can be classified as quantitative (numeric) or qualitative (categorical).

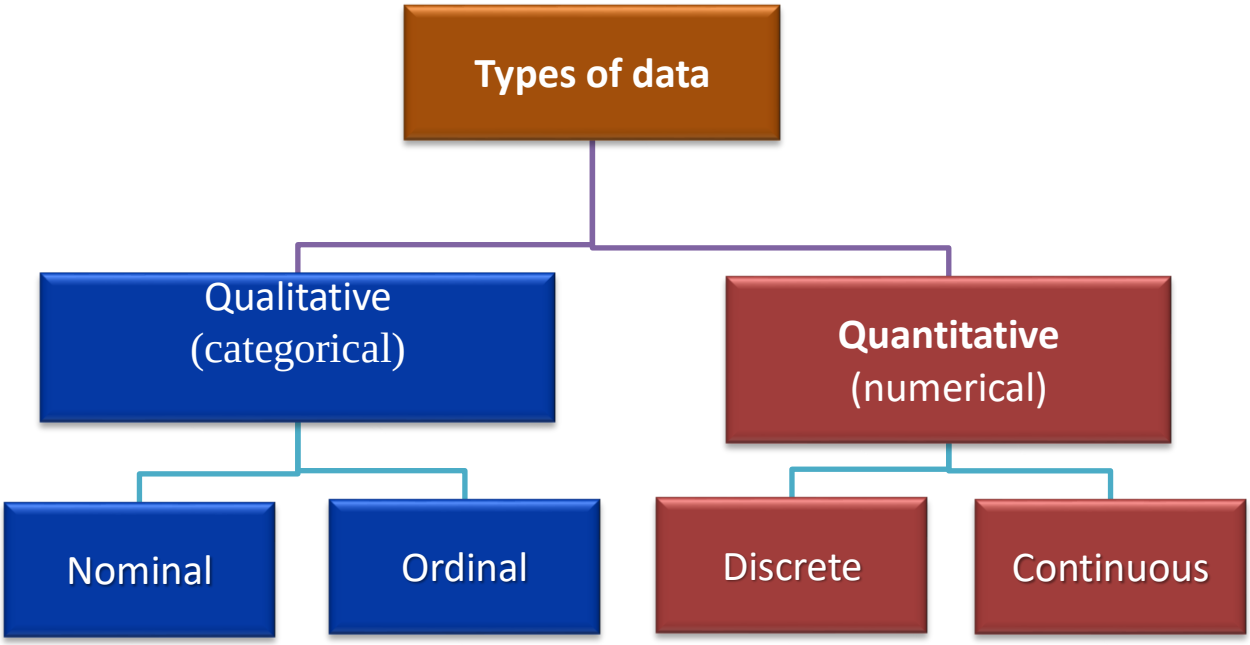


Figure 3. Classification of data

1. Qualitative data include observations that are nor numerical, but descriptive. They represent the number of individuals with the same characteristic or attribute and not the measurement of the attribute. It means qualitative data have only one variable, i.e. the number of individuals**.** There is no magnitude or size of the characteristic as the same cannot be measured, but the frequency of each type is determined and represented. The form a group or class such as young, old, infants, healthy, treated, non treated , or drug or placebo, etc.

Qualitative data can be *nominal* and *ordinal* types.

Table 3.Types of Qualitative data

Type	Feature	Examples
Nominal	Named categories; Unordered categories.	Gender, eye color, blood group types A, B, AB, O.
Ordinal	Same as nominal + ordered categories	Class grades (excellent, good, satisfied, unsatisfied), stages of cancer

Nominal comes from name. They can be classified or named into two or more qualitative categories. The categories don't have specified order and they cannot be ranked. There is no measure of distinction between the values. In medicine, nominal data are used to describe patient's condition.

Ordinal implies order and order means ranking. The variables measured or ordinal data have an inherent order in the relationship among different categories. In medicine, ordinal variables describe patient's condition or characteristic, status, attitude or behavior. Examples: educational level (as elementary, secondary or high), pain level (as severe, moderate or mild), stages of cancer (stage I, II, III or IV).

2. Quantitative data are represented in numbers. They have both frequency as well as magnitude. They deal with numbers and things you can measure objectively: dimensions such as height, width, and length.

Quantitative data can be discrete and continuous types.

Table 4.Types of Quantitative data

Type	Feature	Examples
Discrete	Countable only and takes only some isolated values	Numbers of family members, no. of workers in a factory, no. of persons suffering from a particular disease.
Continuous	Can take up infinite no. of values in a certain range	weight, height, blood sugar

Discrete variable can assume a countable number of values and can take on distinct and separate values. It is characterised by gaps or interruptions in the values or absence of values in between.

Examples: Number of children in a family, number of patients in a doctor's surgery, number of cars in a parking lot.

Continuous variable is one which can take any numerical value within a certain range. It doesn't possess any gaps or interruptions. Continuous variables are usually measurements and can take on an infinite number of possibilities.

Examples: The height of a child at various ages when he grows from 120 to 150 cm, assumes all possible values within this interval even fractional.

STUDY QUESTIONS:

1. The subject and objectives of biostatistics. Areas of application of statistical methods.
2. Population and Sample.
2. Classification, collection and presentation of data.
3. Qualitative data.
4. Quantitative data.
5. Types of scale:

• Nominal or categorical scale.
• Ordinal or evaluation scale.
• Interval scale.

• Ratio scale.

6. The frequency of event.

7. What is a probability of event?

8. For each of the following variables, indicate whether it is quantitative or qualitative and specify the measurement scale that is employed when taking measurements on each:

(a) Class standing of the members of this class relative to each other

(b) Admitting diagnosis of patients admitted to a mental health clinic

(c) Weights of babies born in a hospital during a year

(d) Gender of babies born in a hospital during a year

(e) Range of motion of elbow joint of students enrolled in a university health sciences curriculum

(f) Under-arm temperature of day-old infants born in a hospital

9. For each of the following situations, answer questions a through e:

(a) What is the sample in the study?

(b) What is the population?

(c) What is the variable of interest?

(d) How many measurements were used in calculating the reported results?

(e) What measurement scale was used?

Situation A. A study of 300 households in a small southern town revealed that 20 percent had at least one school-age child present.

Situation B. A study of 250 patients admitted to a hospital during the past year revealed that, on the average, the patients lived 15 miles from the hospital.

2. DESCRIPTIVE STATISTICS

The methods consisting mainly of organizing, summarizing, and presenting data in the form of tables, graphs, and charts are called descriptive statistics. The methods of drawing inferences and making decisions about the population using the sample are called inferential statistics. Inferential statistics uses probability theory.

A statistical inference is an estimate, a prediction, a decision, or a generalization about the population based on information contained in a sample.

The first step in looking at data is to describe the data at hand in some concise way. In smaller studies this step can be accomplished by listing each data point. In general, however, this procedure is tedious or impossible and, even if it were possible, would not give an overall picture of what the data look like.

Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data. The most recognized types of representations of data are:

I. Tabular presentation
II. Numerical presentation
III. Graphical presentation

2.1. TABULAR PRESENTATION OF DATA

The most important part of the statistical analysis is the construction of distribution series (structural grouping) in order to highlight the characteristic properties and patterns of the studied population. Depending on which characteristic (quantitative or qualitative) is taken as the basis for grouping the data, the types of distribution series are distinguished, respectively.

If the basis of the group is taken as a qualitative feature, such a number of distribution is called attributive (distribution by type of work, sex, profession, religion, nationality, etc.), and if quantitative characteristics are grouped, this distribution is called the variation distribution row. To construct a variation rows means to order the quantitative data set by values, and then count the numbers of units with these values.

There are two forms of the variation rows: variation rows and ranked variation rows.

Variation row is the distribution of individual units of the population in ascending or descending order.

For example, given a random value of X (x_1, x_2,…..x_{15}) - 15 measurements of blood pressure: 133, 102, 120, 102, 118, 108, 120, 155, 146, 135, 127, 107, 110, 155, 150.

1) The variation row (without repetition):
102, 108, 107, 110, 118, 120, 127, 133, 135, 146, 150, 155.

2) The ranked variation row (with repetitions):
102, 102, 108, 107, 110, 118, 120, 120, 127, 133, 135, 146, 150, 155, 155.

A table containing variates and their frequencies or relative frequencies is called a frequency distribution table.

A frequency (m_i) is the number of times a given datum occurs in a data set.

A relative frequency of an event (P_i) is defined as the number of times that the event occurs during experimental trials, divided by the total number of trials conducted. To find the relative frequencies, divide each frequency by the total number of observations in the sample. Relative frequencies can be written as fractions, percents, or decimals.

Cumulative frequency ($\boldsymbol{S_i}$) is the accumulation of the previous frequencies. To find the cumulative frequencies, add all the previous frequencies to the relative frequency for the current column.

Example:

Twenty students were asked how many hours they worked per day. Their responses, in hours, are listed below: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Table 5. Frequency distribution table

Data value, x_i	2	3	4	5	6	7	∑
Frequency , m_i	3	5	3	6	2	1	**20**
Relative frequency, $P_i=m_i/n$	0,150	0,250	0,150	0,300	0,100	0,050	**1**
Cumulative frequency, s_i	3	8	11	17	19	20	

The sum of the relative frequency column is 20/20 or 1.

2.2. GRAPHIC REPRESENTATION OF DATA

The source of our statistical knowledge lies in the data. Once we obtain the sample data values, one way to become acquainted with them is to display them in tables or graphically. Charts and graphs are very important tools in statistics because they communicate information visually. These visual displays may reveal the patterns of behavior of the variables being studied.

For example, in the medical sphere, graphical representations of data are used as statistical tools for everyday process management and improvements by decision makers to understand experiments, problems, and solutions. The purpose of this section is to introduce several tabular and graphical procedures commonly used to summarize both qualitative and quantitative data.

For a graphic representation of statistical distribution of the data is used polygon of frequencies, cumulate, ogive, for the interval variables - the histogram.

Frequency polygon is a graphical device for understanding the shapes of distributions. It is almost identical to a histogram, which is used to compare sets of data or to display a cumulative frequency distribution. It uses a line graph to represent quantitative data.

To create a frequency polygon, data values of the variables are placed on the X-axis and frequencies – on the Y-axis. The points are connected and the figure obtained by joining the straight lines is known as frequency polygon. The frequency polygon can be prepared with histogram or without histogram.

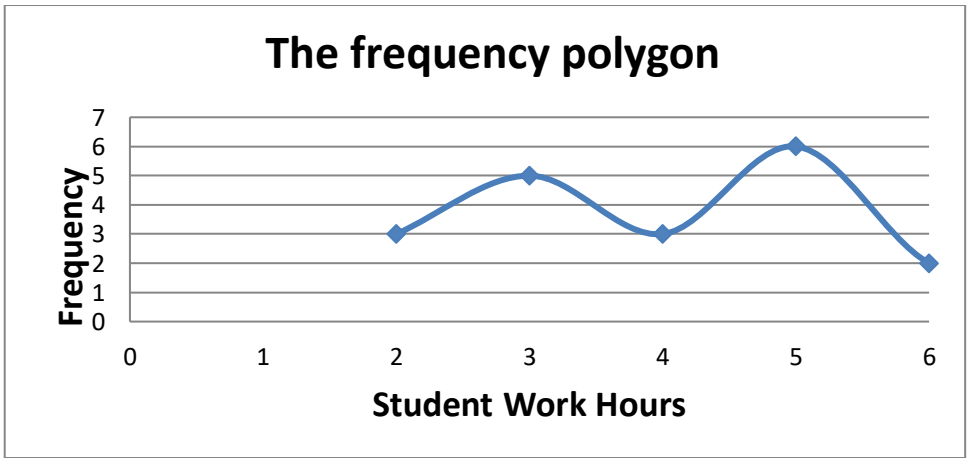


Figure 4. Frequency polygon graph

A cumulative frequency plot is a way to display cumulative information graphically. The basic idea is to assign to each individual value the sum of probabilities of all values that are no larger than the value being considered. It shows the number, percentage, or proportion of observations that are less than or equal to particular values. The values of a variable are usually plotted on the abscissa (x-axis), the cumulative frequencies on the ordinate (y-axis).

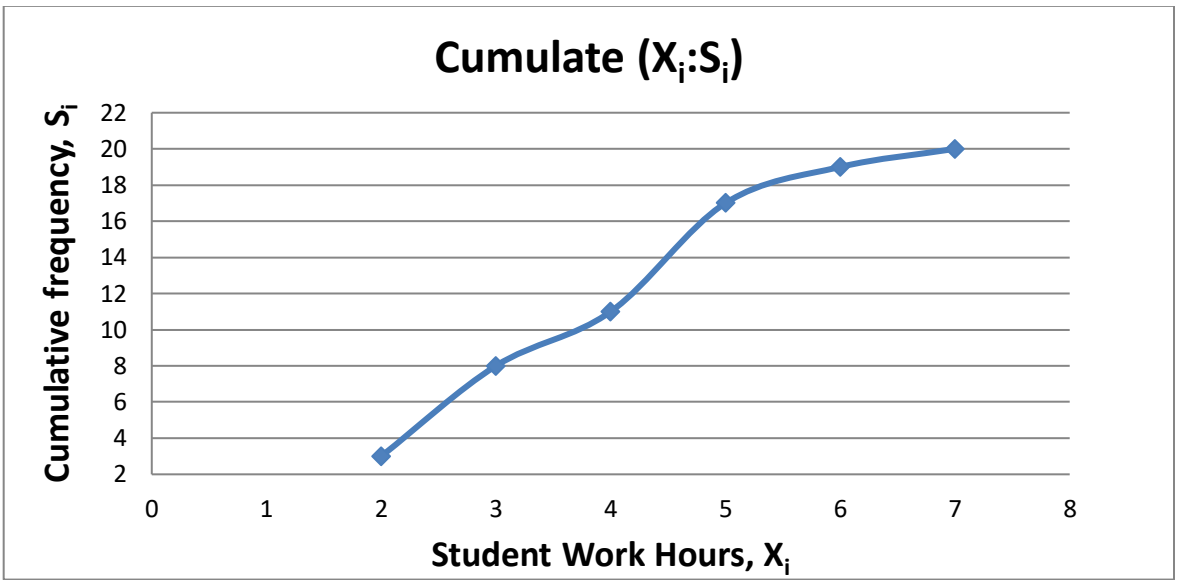


Figure 5. Cumulate graph

An ogive graph is a plot used in statistics to show cumulative frequencies. It allows us to quickly estimate the number of observations that are less than or equal to a particular value.

An ogive graph plots cumulative frequency on the y-axis and the values of a variable along the x-axis.

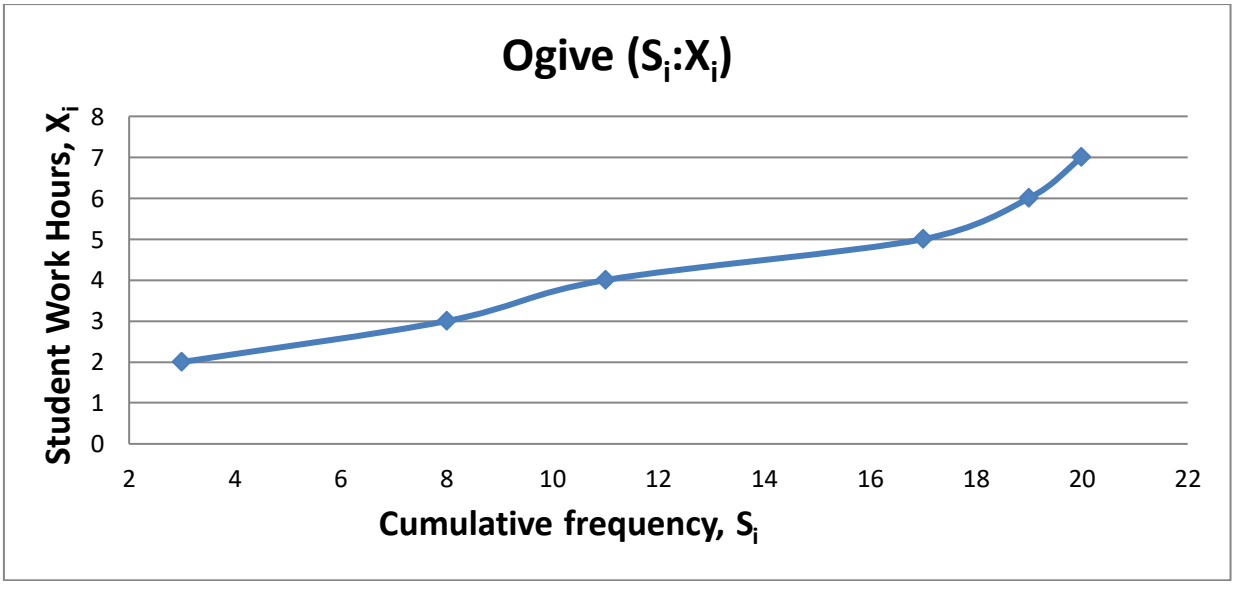


Figure 6. Ogive graph

2.3. NUMERICAL PRESENTATION OF DATA

As you saw, we can "describe" a mass of raw data by charting it, or arranging it in table form. These procedures will help us to make some sense of what initially might be a confusing picture, and hopefully to see patterns in the data. As you are about to see, however, it is often more useful to summarize the data numerically. There are two principal features of a set of data that can be summarized with a single numeric value:

First, a value around which the data has a tendency to congregate or cluster. This is called a summary measure of location, also known as measures of central tendency.

Second, a value which measures the degree to which the data are, or are not, spread out, called a summary measure of spread or dispersion.

With these two summary values you can then compare different sets of data quantitatively.

Measures of Central Tendency: Mean, Median, and Mode

There are several commonly used measures to describe the location or center of a population or sample. The most widely utilized measure of central tendency is the arithmetic mean or average.

The population mean is the sum of the values of the variable under study divided by the total number of objects in the population. It is denoted by a lower case μ ("mu"). Each value is algebraically denoted by X with a subscript denotation *i.*

The algebraic shorthand formula for a population mean is

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

∑ is the Greek capital sigma symbol and indicates summation. The subscript i =1 indicates to start with the first observation and the superscript N means to continue until and including the Nth observation. The subscript and superscript may represent other starting and stopping points for the summation within the population or sample.

Suppose that we have a sample with values $x_1, x_2, \ldots, x_n$. The sample mean is the sum of the values of the variable under study divided by the total number of objects in the sample. It is denoted by $\overline{X}$:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + ... + Xn}{n}$$

The symbol $\bar{x}$ (read "X bar") indicates that the observations of a subset of size *n* from a population have been averaged.

The Expected value (the mathematical expectation).

The expected value of a random variable is the mean or average value of the random variable over the population of units on which the random variable is defined.

For a random variable X, its expected value is usually denoted M(x). The expected value for a discrete variable can be computed from its probability distribution as the sum of multiplication of all possible values of variable X_i on probability of these values P_i.

$$M(X) = \sum_{i=1}^{n} x_i * P_i$$

Median (M_e). The *median* is the middle value of the sorted data set. The median is that value which divides the set into two equal parts. The observed values in one half are less than or equal to the median value and the values in the other half are greater or equal to the median value.

If the number of values is odd, the median will be the middle value when all values have been arranged in order of magnitude.

If the number of values is even, there is no single middle value. Instead, there are two middle values. In this case the median is taken the mean of these two middle values, when all values have been arranged in the order of their magnitudes.

For example, let's look at the data: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

First, we sort our data set sequentially: 2; 2; 2; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5; 5; 5; 5; 5; 6; 6; 7.

Next, we determine the total number of points in our data set (in this case, 20.)

Finally, since there is an even number of items in the data set, we compute the median by taking the average of the two middle numbers, (in this case, 10^{th} and 11^{th} points in data set: $M_e = \frac{x_{10} + x_{11}}{2} = \frac{4+4}{2} = 4$

The median is 4.

Let's take the same data set, but add a new number to it: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3; 7.

First, order the data from least to greatest: 2; 2; 2; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5; 5; 5; 5; 5; 6; 6;7;7.

Next, we determine the total number of points in our data set (in this case, 21.)

Finally, we determine the central position of data (in this case, the 11th position), and the number **4** in the central position is our median - 2; 2; 2; 3; 3; 3; 3; 3; 4; 4; **4**; 5; 5; 5; 5; 5; 5; 6; 6;7;7.

The Mode. The mode of values is that value which occurs most frequently. If all the values are different there is no mode; on the other hand, a set of values may have more than one mode.

For example: 2; 2; 2; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5; 5; 5; 5; 5; 6; 6; 7.

And the mode of this sample is 5, because number 5 occurs 6 times in our dataset.

MEASURES OF DISPERSION

The dispersion of a set of observations refers to the variety that they exhibit. A measure of dispersion conveys information regarding the amount of variability present in a set of data.

If all the values are the same, there is no dispersion; if they are not all the same, dispersion is present in the data. The amount of dispersion may be small when the values, though different, are close together.

The Range. One way to measure the variation in a set of values is to compute the range. The range is the difference between the largest and smallest value in a set of observations. We compute the range as follows:

$$R = X_{max} - X_{min}$$

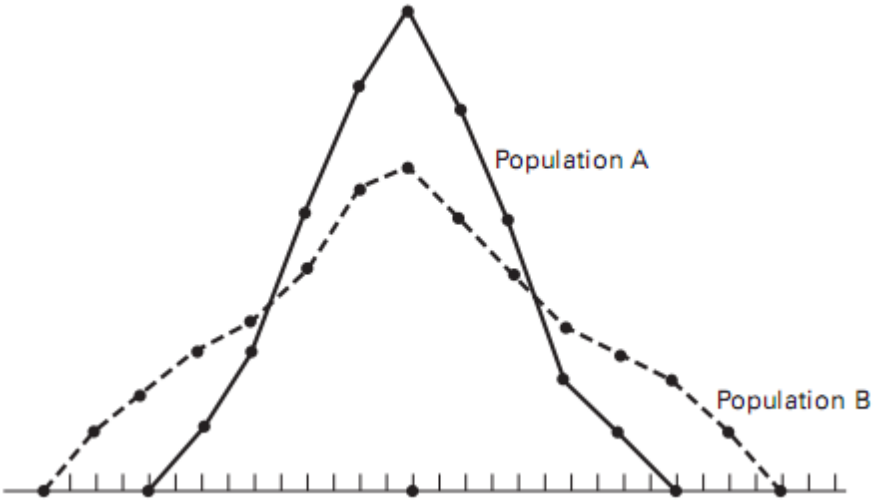


Figure 7. Two frequency distribution with equal means
but different amounts of dispersion

The Variance. When the values of a set of observations lie close to their mean, the dispersion is less than when they are scattered over a wide range. Since this is true, it would be intuitively appealing if we could measure dispersion relative to the scatter of the values about their mean. Such a measure is realized in what is known as the *variance*. In computing the sample variance, we subtract the mean from each of the values, square the resulting differences, and then add up the squared differences. This sum of the squared deviations of the values from their mean is divided by the sample size minus 1, to obtain the sample variance.

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

The reason for dividing by n - 1 rather than n, as we might have expected, is the theoretical consideration referred to as degrees of freedom. In computing the variance, we say that we have n – 1 **degrees of freedom**.

When we compute the population variance of N values, we subtract the mean μ from each of the values, square the resulting differences, and then add

up the squared differences and divide by N. If we let σ^2 denote the population variance, the formula is as follows:

$$\sigma^2 = D(x) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

Standard Deviation. The variance represents squared units and, therefore, is not an appropriate measure of dispersion when we wish to express this concept in terms of the original units. To obtain a measure of dispersion in original the standard units, we merely take the square root of the variance. The result is called the standard deviation. In general, the standard deviation of a sample is given by

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

The Coefficient of Variation. It is useful to relate the arithmetic mean and the standard deviation to each other because, for example, a standard deviation of 10 means something different conceptually if the arithmetic mean is 10 than if it is 1000. A special measure, the coefficient of variation, is often used for this purpose. The formula is given by

$$CV = \frac{s}{\bar{x}} * 100\%$$

The coefficient of variation gives an indication of the homogeneity of the population:

- *< 17% - absolutely homogeneous;*
- *17-33%% - fairly homogeneous;*
- *35-40%% – insufficiently homogeneous;*
- *40-60%% - this indicates a large fluctuation of the population.*

This measure remains the same regardless of what units are used because if the units change by a factor c, then both the mean and standard deviation change by the factor c; the CV, which is the ratio between them, remains unchanged.

The CV is most useful in comparing the variability of several different samples, each with different arithmetic means. This is because a higher variability is usually expected when the mean increases, and the CV is a measure that accounts for this variability. Thus, if we are conducting a study in which air pollution is measured at several sites and we wish to compare day-to-day variability at the different sites, we might expect a higher variability for the more highly polluted sites. A more accurate comparison could be made by comparing the CVs at different sites than by comparing the standard deviations.

Percentiles and Quartiles. The mean and median are special cases of a family of parameters known as location parameters. These descriptive measures are called location parameters because they can be used to designate certain positions on the horizontal axis when the distribution of a variable is graphed. In that sense the so-called location parameters “locate” the distribution on the horizontal axis. For example, a distribution with a median of 100 is located to the right of a distribution with a median of 50 when

the two distributions are graphed. Other location parameters include percentiles and quartiles. We may define a percentile as follows:

Given a set of ***n*** observations $x_1, x_2, \ldots x_n$, the $\boldsymbol{p}^{th}$ percentile ***P*** is the value of X such that ***p*** percent or less of the observations are less than ***P*** and (100-*p*) percent or less of the observations are greater than *P*. Subscripts on P serve to distinguish one percentile from another. The 10th percentile, for example, is designated the 70th is designated and so on. The 50^{th} percentile is the median and is designated The 25th percentile is often referred to as the first quartile and denoted The 50th percentile (the median) is referred to as the second or middle quartile and written and the 75th percentile is referred to as the third quartile,

When we wish to find the quartiles for a set of data, the following formulas are used:

$$\left.\begin{aligned} Q_1 &= \frac{n+1}{4}\text{th} && \text{ordered observation} \\ Q_2 &= \frac{2(n+1)}{4} = \frac{n+1}{2}\text{th} && \text{ordered observation} \\ Q_3 &= \frac{3(n+1)}{4}\text{th} && \text{ordered observation} \end{aligned}\right\}$$

Interquartile Range As we have seen, the range provides a crude measure of the variability present in a set of data. A disadvantage of the range is the fact that it is computed from only two values, the largest and the smallest. A similar measure that reflects the variability among the middle 50 percent of the observations in a data set is the interquartile range.

The interquartile range (IQR) is the difference between the third and first quartiles: that is

$$IQR = Q_3 - Q_1$$

A large IQR indicates a large amount of variability among the middle 50 percent of the relevant observations, and a small IQR indicates a small amount of variability among the relevant observations. Since such statements are rather vague, it is more informative to compare the interquartile range with the range for the entire data set. A comparison may be made by forming the ratio of the IQR to the range (R) and multiplying by 100. That is, 100(IQR/R) tells us what percent the IQR is of the overall range.

Skewness. Data distributions may be classified on the basis of whether they are symmetric or asymmetric. If a distribution is symmetric, the left half of its graph (histogram or frequency polygon) will be a mirror image of its right half. When the left half and right half of the graph of a distribution are not mirror images of each other, the distribution is asymmetric.

If the graph (histogram or frequency polygon) of a distribution is asymmetric, the distribution is said to be skewed. If a distribution is not symmetric because its graph extends further to the right than to the left, that is, if it has a long tail to the right, we say that the distribution is skewed to the right or is positively skewed. If a distribution is not symmetric because its graph extends

further to the left than to the right, that is, if it has a long tail to the left, we say that the distribution is skewed to the left or is negatively skewed.

A distribution will be skewed to the right, or positively skewed, if its mean is greater than its mode. A distribution will be skewed to the left, or negatively skewed, if its mean is less than its mode. Skewness can be expressed as follows:

$$Skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$$

1. In case of a symmetrical (normal) distribution mean, median and mode have the same value. Therefore, in a graph all three of them fall at the same position. It means the three values coincide (see figure 8) i.e. Mean (μ)=Median (μd)=Mode(μo).

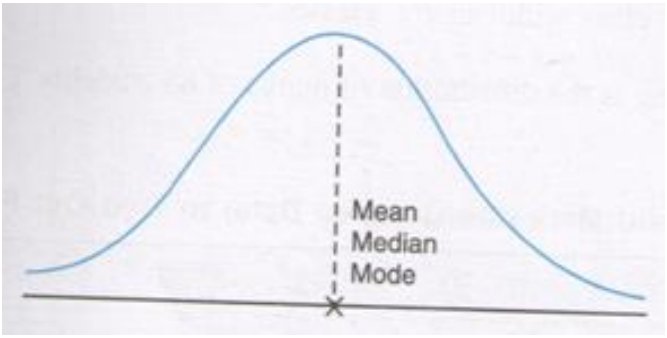


Figure 8. Graphic representation for a symmetrical distribution
Mean =Median =Mode.

In actual life most distributions are not symmetrical. In moderately asymmetrical distributions, the values of mean, median and mode present following empirical relationship:

Mode=3 Median – 2Mean

2. For a positively skewed distribution (figure 9) mean will be greater than median and median is greater than mode, i.e.,

M>Me>Mo or Mo<Me<M

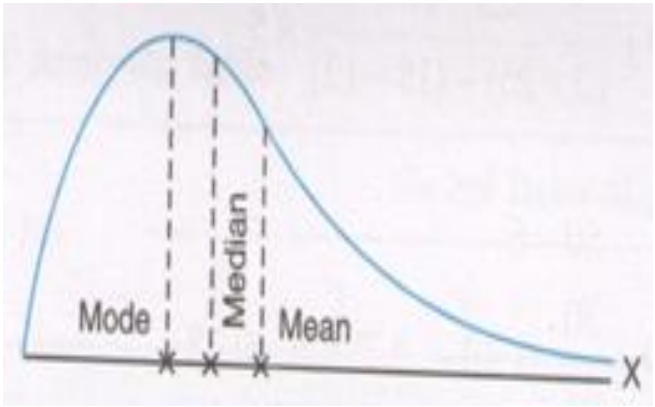


Figure 9. Relationship between Mean, Median and Mode in
Positively skewed curve.

3. For a negatively skewed distribution (figure 10) the order of the magnitude of the three averages will be reversed, i.e.,

Mo>Me>M or M<Me<Mo

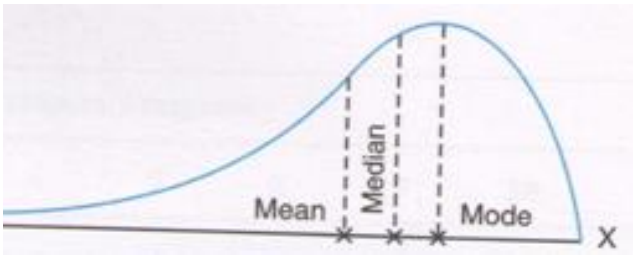


Figure 10. Relationship between Mean, Median and Mode in Negatively skewed curve.

Example: In a moderately asymmetrical distribution, the mode and mean are 64.2 and 67.4 respectively. Find the median.

Solution: The empirical relation between mean, median and mode:

Mode=3 Median – 2 Mean

Here, Mode=64.2 and Mean = 67.4

64.2 = 3 Median – 2 (67.4)

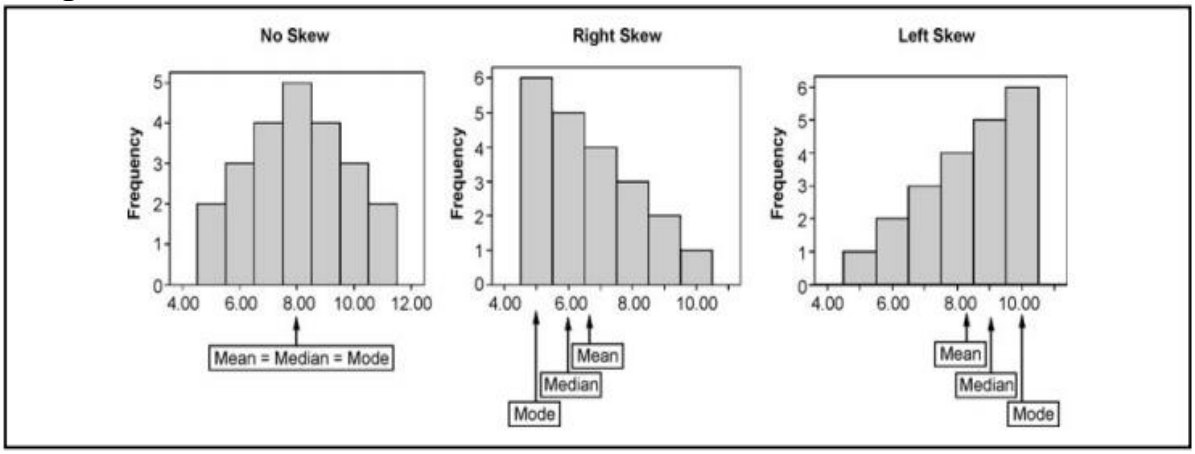
64.2 = 3 Median – 134.8

or -3 median = -134.8 -64.2

or Median = $\frac{199.0}{3} = 66.33$

Median=66.33

Most computer statistical packages include this statistic as part of a standard printout. A value of skewness 0 indicates positive skewness and a value of skewness 0 indicates negative skewness. An illustration of skewness is shown in Figure 10.



	No Skew	Right Skew	Left Skew
Mean	8.0000	6.6667	8.3333
Median	8.0000	6.0000	9.0000
Mode	8.00	5.00	10.00
Skewness	.000	.627	−.627

Figure 11. Three histograms illustrating skewness.

Kurtosis. Just as we may describe a distribution in terms of skewness, we may describe a distribution in terms of kurtosis.

Kurtosis is a measure of the degree to which a distribution is “peaked” or flat in comparison to a normal distribution whose graph is characterized by a bell-shaped appearance.

Kurtosis can be expressed as

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

The types of kurtosis are determined by the excess kurtosis of a particular distribution. The excess kurtosis can take positive or negative values as well, as values close to zero.

1. Mesokurtic

Data that follows a mesokurtic distribution shows an excess kurtosis of zero or close to zero. It means that if the data follows a normal distribution, it follows a mesokurtic distribution.

Figure 12. Mesokurtic distribution

2. Leptokurtic

Leptokurtic indicates a positive excess kurtosis. The leptokurtic distribution shows heavy tails on either side, indicating the large outliers. Outliers stretch the horizontal axis of the histogram graph, which makes the bulk of the data appear in a narrow (“skinny”) vertical range, thereby giving the “skinniness” of a leptokurtic distribution.

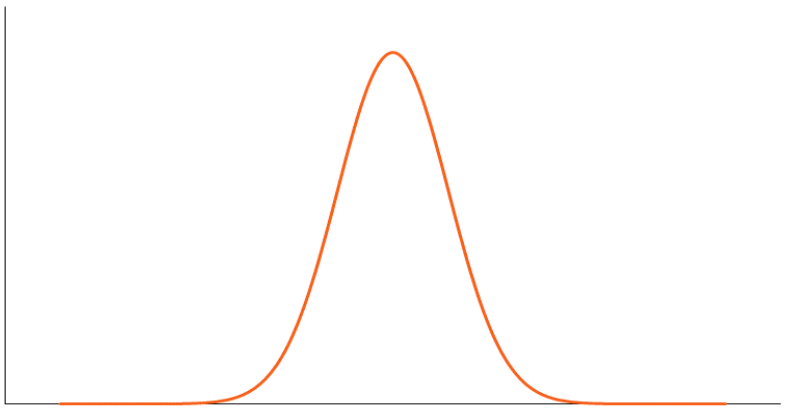


Figure 13. Leptokurtic distribution

3. Platykurtic

A platykurtic distribution shows a negative excess kurtosis. The kurtosis reveals a distribution with flat tails. Distribution is shorter, tails are thinner than the normal distribution. The peak is lower and broader than Mesokurtic, which means that data are light-tailed or lack of outliers. The reason for this is because the extreme values are less than that of the normal distribution. The flat tails indicate the small outliers in a distribution.

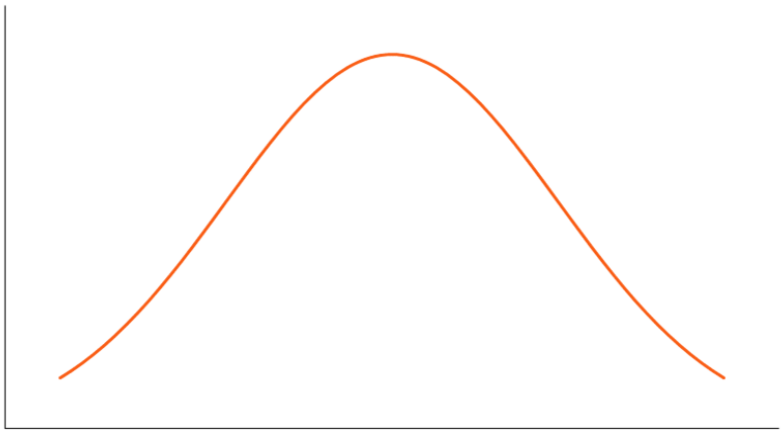


Figure 14. Platykurtic distribution

Graphs of distributions representing the three types of kurtosis are shown in Figure 15:

	Mesokurtic	Leptokurtic	Platykurtic
Mean	6.0000	6.0000	6.0000
Median	6.0000	6.0000	6.0000
Mode	6.00	6.00	6.00
Skewness	.000	.608	−1.158

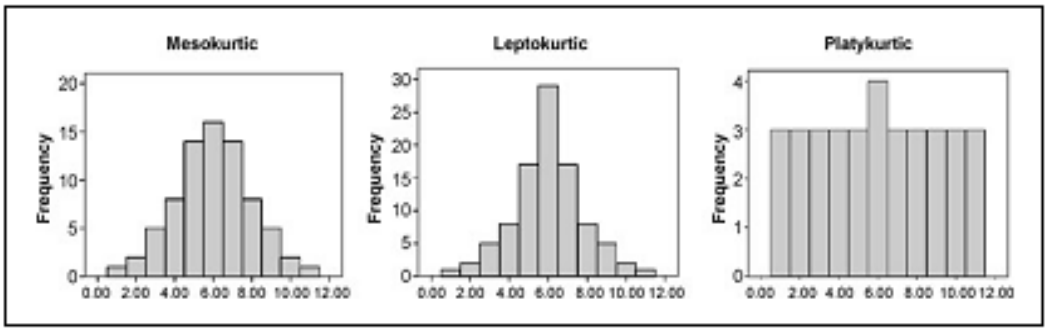


Figure 15. Three histograms representing kurtosis.

Box-and-Whisker Plots A useful visual device for communicating the information contained in a data set is the box-and-whisker plot. The construction of a box and-whisker plot (sometimes called, simply, a box plot) makes use of the quartiles of a data set and may be accomplished by following these five steps:

1. Represent the variable of interest on the horizontal axis.

2. Draw a box in the space above the horizontal axis in such a way that the left end of the box aligns with the first quartile Q_1 and the right end of the box aligns with the third quartile Q_3.

3. Divide the box into two parts by a vertical line that aligns with the median Q_2.

4. Draw a horizontal line called a whisker from the left end of the box to a point that aligns with the smallest measurement in the data set.

5. Draw another horizontal line, or whisker, from the right end of the box to a point that aligns with the largest measurement in the data set.

Examination of a box-and-whisker plot for a set of data reveals information regarding the amount of spread, location of concentration, and symmetry of the data.

The following example illustrates the construction of a box-and-whisker plot.

Example. Evans et al. (A-7) examined the effect of velocity on ground reaction forces (GRF) in dogs with lameness from a torn cranial cruciate ligament. The dogs were walked and trotted over a force platform, and the GRF was recorded during a certain phase of their performance. Table 5 contains 20 measurements of force where each value shown is the mean of five force measurements per dog when trotting.

Solution: The smallest and largest measurements are 14.6 and 44, respectively. The first quartile is the $Q_1=(20+1)/4=5.25^{th}$ measurement, which is 27.2+(0.25)(27.4-27.2)=27.25. The median is the $Q_2+(20+1)/4=10.5^{th}$ measurement or 30.7+(0.5)(31.5-30.7)=31.1; and the third quartile is the $Q_3+3(20+1)/4=15.75^{th}$ measurement, which is equal to 33.3+(0.75)(33.6-33.3)=33.525. The interquartile range is IQR=33.525-27.25=6.275. The range box-and-whisker plot is shown in Figure 16:

Examination of Figure 16 reveals that 50 percent of the measurements are between about 27 and 33, the approximate values of the first and third quartiles, respectively. The vertical bar inside the box shows that the median is about 31.

Table 6. GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	**24.3**	**24.9**	**27.0**	**27.2**	**27.4**	**28.2**	**28.8**	**29.9**	**30.7**
31.5	**31.6**	**32.3**	**32.8**	**33.3**	**33.6**	**34.3**	**369**	**38.3**	**44.0**

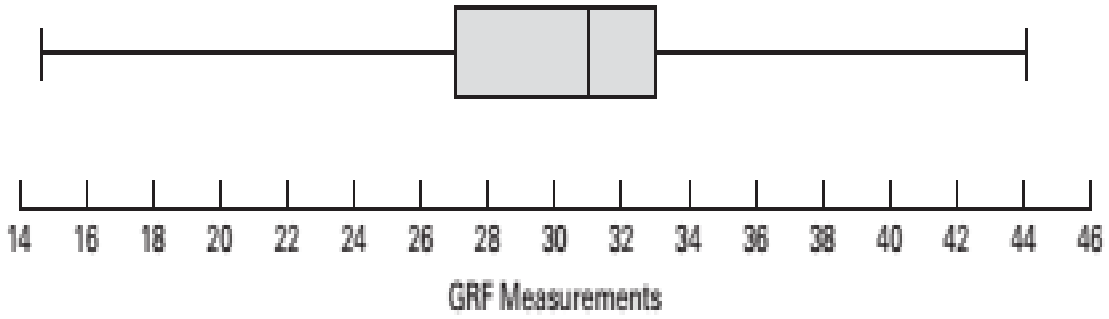


Figure 16. Box-and-whisker plot for Example

Many statistical software packages have the capability of constructing box-and whisker plots.

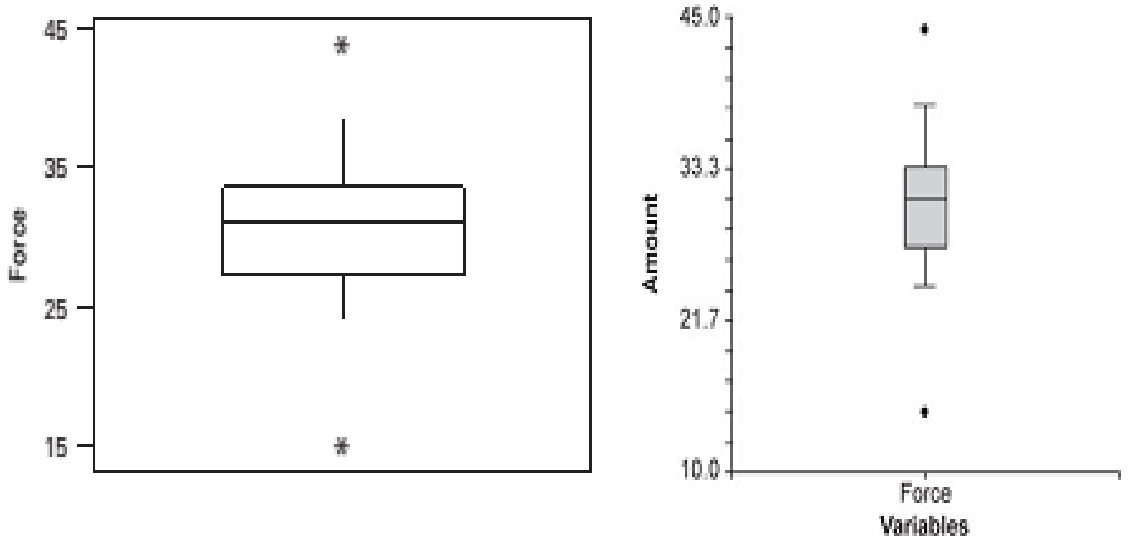


Figure 17. Box-and-whisker plot constructed by statistical software package from the data of Table 5.

The asterisks in Figure 17 alert us to the fact that the data set contains one unusually large and one unusually small value, called ***outliers.*** The outliers are the dogs that generated forces of 14.6 and 44. Box-and whisker plots may be displayed vertically as well as horizontally.

Table 7. Summary of formulae

Formula Number	Name	Formula
1	Class interval width using Sturges's rule	$w = \frac{R}{k}$
2	Mean of a population	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
3	Mean of a sample	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + ... + Xn}{n}$
4	Range	$R = X_{max} - X_{min}$
5	Sample variance	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
6	Population variance	$\sigma^2 = D(x) = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$
7	Standard deviation	$s = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
8	Coefficient of variation	$CV = \frac{s}{\bar{x}} * 100\%$
9	Quartile location in ordered array	$Q_1 = \frac{1}{4}(n+1)$ $Q_2 = \frac{1}{2}(n+1)$ $Q_3 = \frac{3}{4}(n+1)$
10	Skewness	$Skewness = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3}$
11	Kurtosis	$Kurtosis = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{ns^4} - 3$
Symbol Key	• C.V. – coefficient of variation • IQR – Interquartile range • k – number of class intervals • μ - population mean • N– population size • n – sample size • (n-1) – degrees of freedom • Q_1 – first quartile • Q_2 – second quartile - median • Q_1 – third quartile • R – range	

	• s – standard deviation • s^2 – sample variance • δ^2 – population variance • x_i – i^{th} data observation • x_L – largest data point • x_S – smallest data point • $\bar{x}$ - sample mean • w - class width

2.4. DESCRIPTIVE STATISTICS BY HAND IN EXCEL PROGRAM

Example: The height of 30 students on a college campus was recorded as follows:

125, 143, 143, 128, 135, 135, 149, 135, 135, 135, 140, 143, 125, 123, 125, 149, 150, 150, 128, 128, 128, 135, 140, 109, 114, 123, 140, 140, 140, 153.

Summarize this row data using descriptive statistics' methods.

Solution:

1. Construct the ranked variation row using Sort and Filter command in Excel:

109, 114, 123, 123, 125, 125, 125, 128, 128, 128, 128, 135, 135, 135, 135, 135, 135, 140, 140, 140, 140, 140, 143, 143, 143, 149, 149, 150, 150, 153.

2. Construct variation data row:

109, 114, 123, 125, 128, 135, 140, 143, 149, 150, 153.

3. Construct a frequency distribution table in Excel sheet by formulae:

Table 8. Frequency table for height of students

x_i	109	114	123	125	128	135	140	143	149	150	153	∑
m_i	1	1	2	3	4	6	5	3	2	2	1	**30**
p_i	0,033	0,033	0,067	0,100	0,133	0,200	0,167	0,100	0,067	0,067	0,033	**1**
S_i	1	2	4	7	11	17	22	25	27	29	30	

4. Make a graphical representation of data in Excel program: draw the polygon of frequencies, cumulates and ogive.

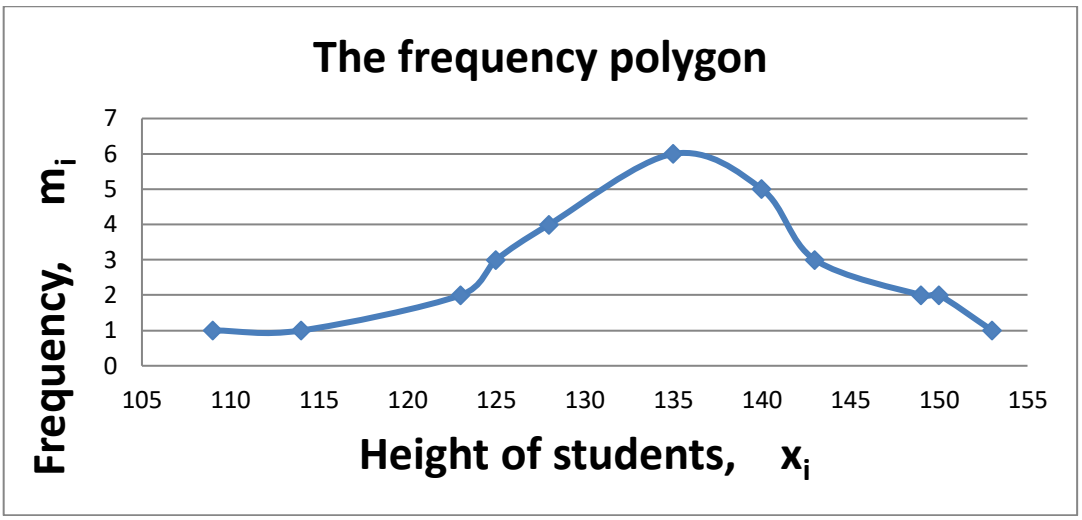


Figure 18. Frequency polygon

Select the range of cells with x_i and m_i values, then click Insert, and choose type of charts -Scatter –Scatter with smooth Lines and Markers. Then using Design, Layout tabs edit the chart settings: name of chart, name of axis.

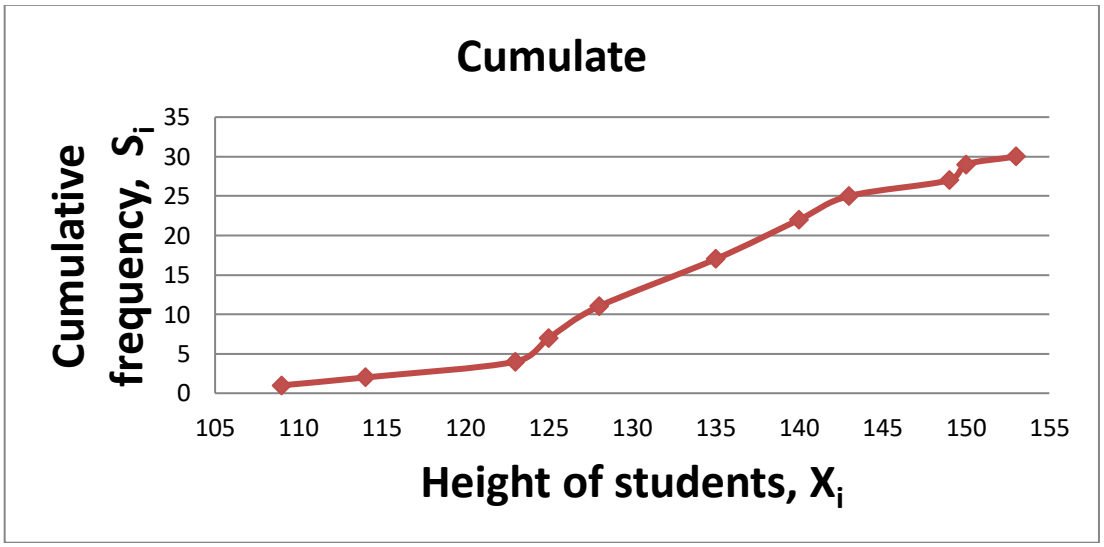


Figure 19. Cumulate

Select the range of cells with x_i and S_i values (select 2 rows by pressing CTRL key) , then click Insert, and choose type of charts -Scatter –Scatter with smooth Lines and Markers. Then using Design, Layout tabs edit the chart settings: name of chart, name of axis.

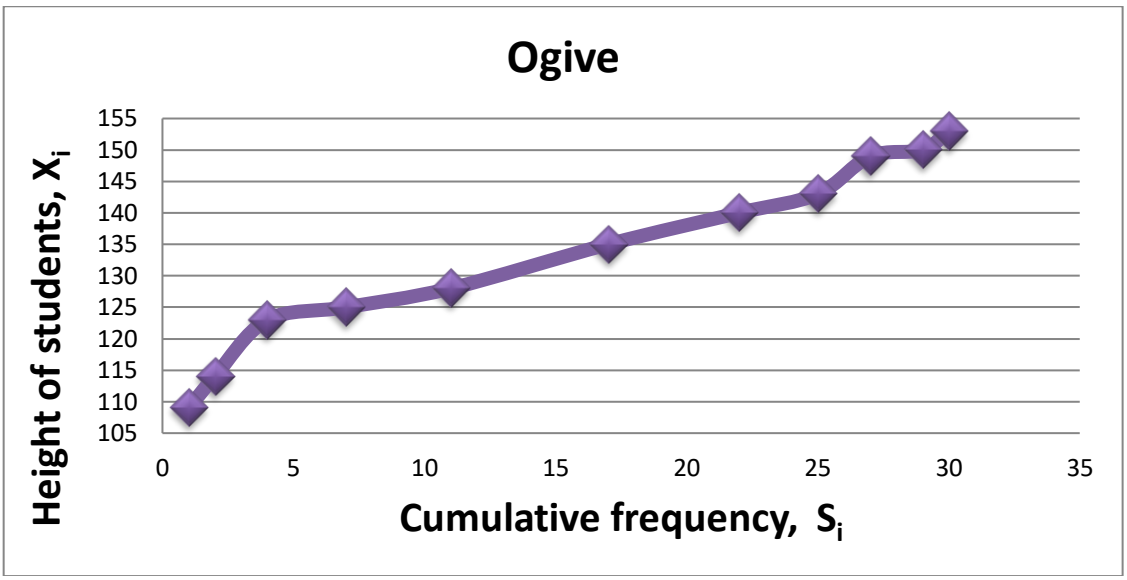


Figure 20. Ogive

Select the range of cells with x_i and S_i values (select 2 rows by pressing CTRL key) , then click Insert, and choose type of charts -Scatter –Scatter with smooth Lines and Markers. Then right click on the appeared graph and choose Select data option – Edit. In the Edit series box enter graph's name, in series X values select range of S_i values and in series Y values select range of x_i values. Then using Design, Layout tabs edit the chart settings: name of chart, name of axis.

5. Calculate the numerical characteristics of data: sample mean, expected value, median, mode, sample variance, sample standard deviation, range and coefficient of variation.

5.1. Sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

$$\bar{x} = \frac{109 + 114 + 123 + 123 + 125 + 125 + 125 + 128 + 128 + 128 + 128 + 135 + 135 + 135 + 135 + 135 + 135 + +140 + 140 + 140 + 140 + 140 + 143 + 143 + 143 + 149 + 149 + 150 + 150 + 153}{30}$$

$$= \frac{4046}{30} = \mathbf{134.87}$$

or

$$\bar{x} = \frac{109 * 1 + 114 * 1 + 123 * 2 + 125 * 3 + 128 * 4 + 135 * 6 + 140 * 5 + 143 * 3 + 149 * 2 + 150 * 2 + 153 * 1}{30}$$

$$= \frac{4046}{30} = \mathbf{134.87}$$

5.2. Expected value

$$M(x) = \mu = x_1P_1 + x_2P_2 + x_3P_3 + \cdots + x_nP_n = \sum_{i=1}^{n} x_iP_i$$

$$M(x) = 109 * 0.033 + 114 * 0.033 + 123 * 0.067 + 125 * 0.1 + 128 * 0.13 + 135 * 0.2 + 140 * 0.17 + 143 * 0.1 + 149 * 0.067 + 150 * 0.067 + 153 * 0.033 = \mathbf{134.87}$$

5.3. The median

$$Me = \frac{x_m + x_{m+1}}{2}$$

$$Me = \left|\frac{30}{2} = 15\right| = \frac{x_{15} + x_{16}}{2} = \frac{135 + 135}{2} = \mathbf{135}$$

5.4. The mode

$$Mo = 135$$

5.5. The sample Variance

$$S^2 = \frac{1}{n-1}\sum_{i+1}^{k}(x_i - \bar{x})^2\, m_i;$$

$$S^2 = \frac{1}{30-1}[(109 - \mathbf{134.87})^2 * 1 + (114 - \mathbf{134.87}3)^2 * 1 + (123 - \mathbf{134.87})^2 * 2 + (125 - \mathbf{134.87})^2 * 3 + (128 - \mathbf{134.87})^2 * 4 + (135 - \mathbf{134.87})^2 * 6 + (140 - \mathbf{134.87})^2 * 5 + (143 - \mathbf{134.87})^2 * 3 + (149 - \mathbf{134.87})^2 * 2 + (150 - \mathbf{134.87})^2 * 2 + (153 - \mathbf{134.87})^2 * 1] = \frac{\mathbf{3384.627}}{\mathbf{29}} \approx \mathbf{116.7}$$

5.6. The standard deviation:

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

$$s = \sqrt{116.7} = \mathbf{10.8}$$

5.7. The range

$$R = x_{max} - x_{min} = 153 - 109 = 44$$

5.8. The coefficient of variation

$$CV = \frac{s}{\bar{x}} * 100\% = \frac{10.8}{134.87} * 100\% = 8.02\%$$

Table 9. Summary statistics

Summary	
Mean	134,87
Expected Value	134,87
Median	135
Mode	135
Standard Deviation	10,80145
Sample Variance	116,6713
Range	44
CV	8.02%
Minimum	109
Maximum	153
Count	30

Analysis of results. The table shows that the sample mean = 134.87, median = 135, mode = 135. It can be considered almost equal, it means the three values coincide i.e., the data distribution is normal (symmetrical). The standard deviation = 10.8. The height of students in the distribution series may differ from the average by 10.8 cm. The coefficient of variation=8.02, it is less than 17%, it means the set is absolutely homogeneous.

2.4.1. DESCRIPTIVE STATISTICS BY DATA ANALYSIS PACKAGE IN EXCEL.

You can use the Analysis Toolpak add-in to generate **descriptive statistics**.

Example: The height of 30 students on a college campus was recorded as follows:

125, 143, 143, 128, 135, 135, 149, 135, 135, 135, 140, 143, 125, 123, 125, 149, 150, 150, 128, 128, 128, 135, 140, 109, 114, 123, 140, 140, 140, 153.

Summarize this row data using descriptive statistics' methods.

STEP 1. Type your data in Excel. Enter in the A column all values of students' height **X.**

	A
1	Height of students
2	109
3	114
4	123
5	123
6	125
7	125
8	125
9	128
10	128
11	128
12	128
13	135
14	135
15	135
16	135
17	135
18	135
19	140
20	140
21	140
22	140
23	140
24	143
25	143
26	143
27	149
28	149
29	150
30	150
31	153

Figure 21. Values of students' height

STEP 2. On the Data tab, in the Analysis group, click Data Analysis.

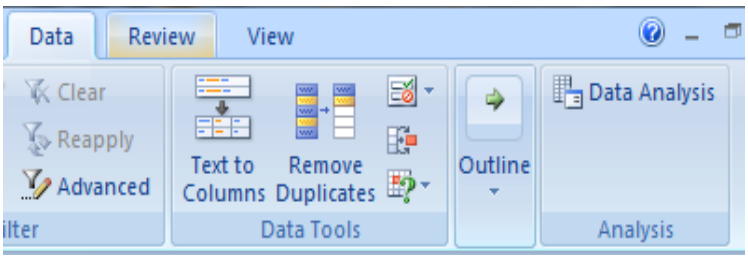


Figure 22. Data analysis command

Note: can't find the Data Analysis button? Add the Analysis ToolPak add-in from Excel Options.

Step 3. Descriptive Statistics and click OK.

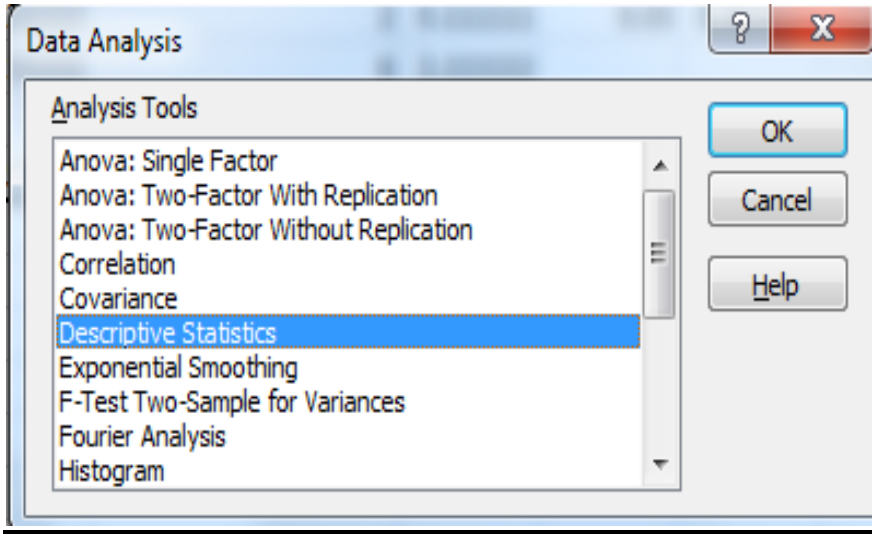


Figure 23. Descriptive statistics command

STEP 4. Select the range A2:A31 as the Input Range.

STEP 5. Select cell C1 as the Output Range. Tick the box Summary statistics, Confidence level for mean, Reliability level, K-th smallest, K-th largest.

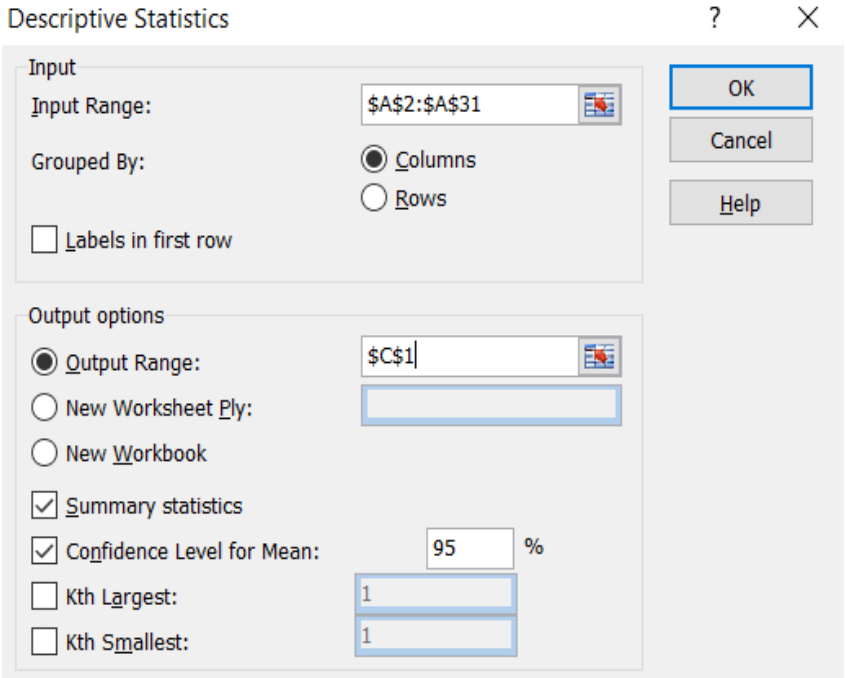


Figure 24. Descriptive statistics box

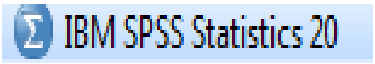
STEP 6. Click OK.

	A	B	C	D
1	Height of students		*Column1*	
2	109			
3	114		Mean	134.86667
4	123		Standard Error	1.9720655
5	123		Median	135
6	125		Mode	135
7	125		Standard Deviation	10.801447
8	125		Sample Variance	116.67126
9	128		Kurtosis	-0.1472775
10	128		Skewness	-0.3834821
11	128		Range	44
12	128		Minimum	109
13	135		Maximum	153
14	135		Sum	4046
15	135		Count	30
16	135		Confidence Level(95	4.0333267
17	135			
18	135			
19	140			
20	140			
21	140			
22	140			
23	140			
24	143			
25	143			
26	143			
27	149			
28	149			
29	150			
30	150			
31	153			

Figure 25. Result of descriptive statistics

Analysis of results. The table shows that the sample mean = 134.87, median = 135, mode = 135. It can be considered almost equal, it means the three values coincide i.e., the data distribution is normal (symmetrical). Skewness coefficient = -0,38. The obtained result indicates the presence of asymmetry that is insignificant in magnitude and negative in character. Since the magnitude of the skewness is negative, therefore, we are talking about left-side asymmetry. The coefficient of kurtosis = -0.147, which indicates a platykurtic distribution. The standard deviation = 10.8. The height of students in the distribution series may differ from the average by 10.8 cm. The coefficient of variation=8.02, it is less than 17%, it means the set is absolutely homogeneous.

2.4.2. DESCRIPTIVE STATISTICS BY SPSS STATISTICS PROGRAM

1. Run SPSS Statistics program. 

2. Choose "Enter data" -OK. The window of the SPSS Statistics program consists of two switchable tabs: "Data" and "Variables", which located on the bottom part of the window. The "Data" tab will be opened automatically. Enter the data in the first column.

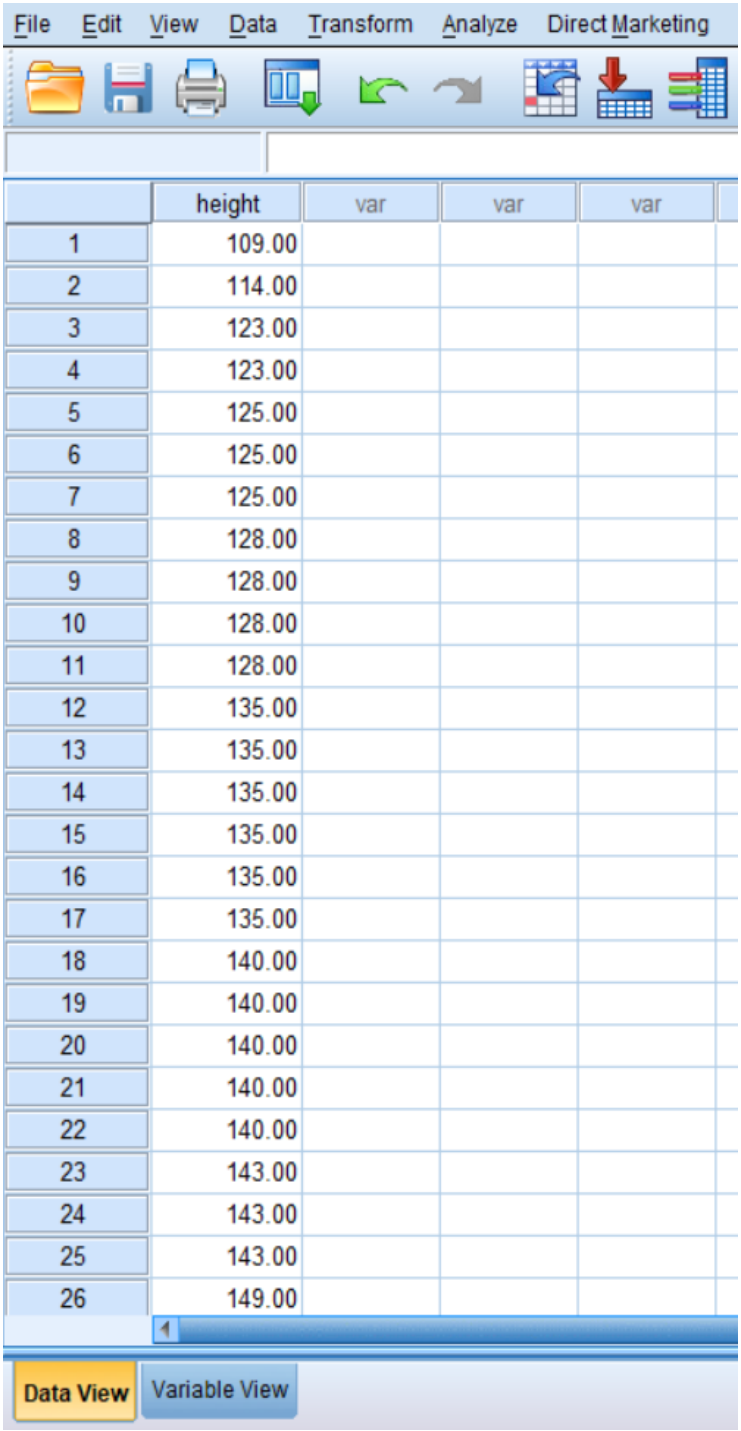


Figure 26. Data entry in data view section of SPSS Statistics program

Open Variables tab and specify format of the data:
In the NAME field type the name of variables “Height” without a space,
TYPE - numeric,
WIDTH - 8,
DECIMAL - 2,
MEASURE – Scale

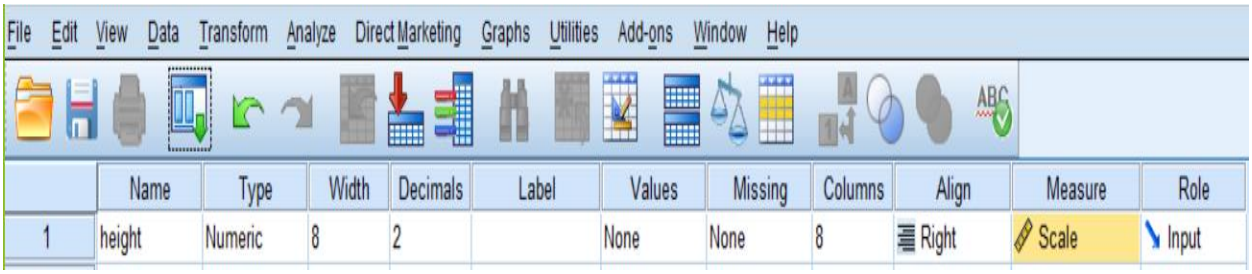


Figure 27. Specifying format of the data in SPSS Statistics program

Go to the "Data view" and click "ANALYZE" tab. From the drop-down window choose "Descriptive statistics" - "Frequencies". In the "Frequencies" window highlight the name of variable “Height” and using the button → transfer it to the Variables window.

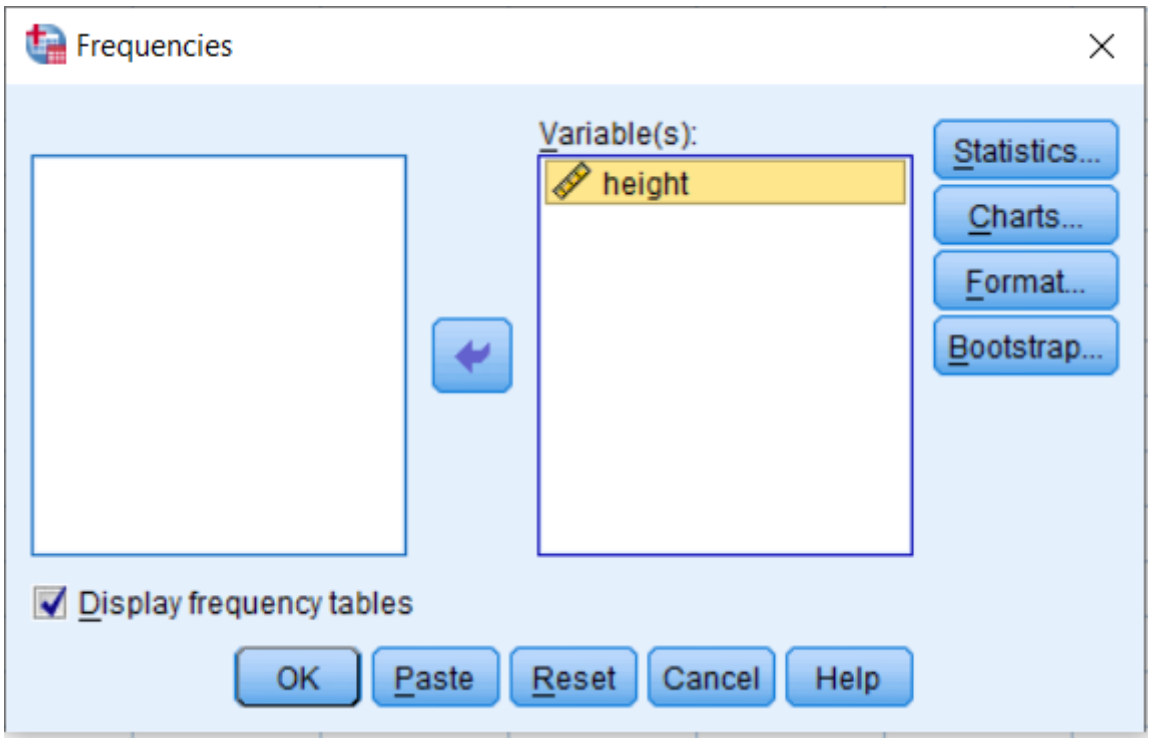


Figure 28. Frequencies box

Click "Statistics" command, in the opened "Frequencies: Statistics” window tick boxes: Variance, Standard deviation, Range, Minimum, Maximum, Mean, Median, Mode.

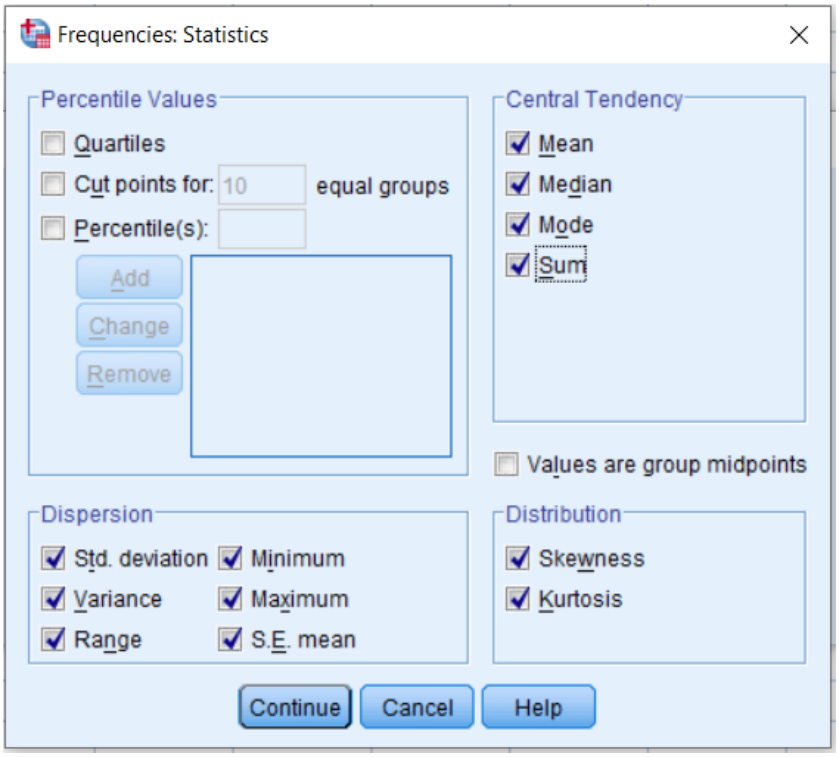


Figure 29. Frequencies: Statistics box

Then click Continue button - OK.

Statistics

height

N	Valid	30
	Missing	0
Mean		134.8667
Std. Error of Mean		1.97207
Median		135.0000
Mode		135.00
Std. Deviation		10.80145
Variance		116.671
Skewness		-.383
Std. Error of Skewness		.427
Kurtosis		-.147
Std. Error of Kurtosis		.833
Range		44.00
Minimum		109.00
Maximum		153.00
Sum		4046.00

height

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	109.00	1	3.3	3.3	3.3
	114.00	1	3.3	3.3	6.7
	123.00	2	6.7	6.7	13.3
	125.00	3	10.0	10.0	23.3
	128.00	4	13.3	13.3	36.7
	135.00	6	20.0	20.0	56.7
	140.00	5	16.7	16.7	73.3
	143.00	3	10.0	10.0	83.3
	149.00	2	6.7	6.7	90.0
	150.00	2	6.7	6.7	96.7
	153.00	1	3.3	3.3	100.0
	Total	30	100.0	100.0	

Figure 30. Summary statistics in SPSS Statistics program

Analysis of results. The table shows that the sample mean = 134.87, median = 135, mode = 135. It can be considered almost equal, it means the three values coincide i.e., the data distribution is normal (symmetrical). Skewness coefficient = -0.38. The obtained result indicates the presence of asymmetry that is insignificant in magnitude and negative in character. Since the magnitude of the skewness is negative, therefore, we are talking about left-side asymmetry. The coefficient of kurtosis = -0,147, which indicates a platykurtic distribution. The standard deviation = 10.8. The height of students in the distribution series may differ from the average by 10.8 cm. The coefficient of variation=8.02, it is less than 17%, it means the set is absolutely homogeneous.

STUDY QUESTIONS:

1. Statistical grouping and data summary.
2. Representativeness of the sample.
3. Creating of variation rows.
4. Tabular and graphical representation of variation rows.
5. Numerical representation. The numerical characteristics of discrete statistical distribution: sample mean, expected value, sample variance, sample standard deviation, median, mode, range and coefficient of variation.
6. Confidence interval.
7. Application of MS Excel and SPSS programs.

REVISIONARY PROBLEMS:

On the given samples perform the following calculations:

I. Construct a statistical and variation row.

II. Construct the table of frequencies and cumulative frequencies for bunched sampling.

III. Construct the polygon of frequencies and relative frequencies, cumulates and ogive.

IV. Calculate the numerical characteristics of discrete statistical distribution: sample mean, expected value, sample variance, sample standard deviation, median, mode, range and coefficient of variation.

Sample 1. **Age of patients admitted to hospital during a certain month. Summarize this data.**

Variant **№1**					
№	**Age**	№	**Age**	№	**Age**
1	**11**	12	**15**	23	**17**
2	**12**	13	**16**	24	**18**
3	**13**	14	**12**	25	**19**
4	**15**	15	**13**	26	**15**
5	**16**	16	**14**	27	**12**
6	**14**	17	**15**	28	**16**
7	**14**	18	**17**	29	**13**
8	**12**	19	**12**	30	**14**
9	**19**	20	**13**	31	**15**
10	**18**	21	**15**	32	**12**
11	**17**	22	**15**	33	**13**

Sample 2. **The weight of 3 year children was recorded in clinic. Summarize this data.**

Variant №2					
№	**weight**	№	**weight**	№	**weight**
1	**13**	11	**19**	21	**12**
2	**16**	12	**14**	22	**15**
3	**14**	13	**19**	23	**19**
4	**19**	14	**18**	24	**18**
5	**18**	15	**17**	25	**17**
6	**18**	16	**15**	26	**12**
7	**10**	17	**10**	27	**13**
8	**12**	18	**12**	28	**12**
9	**13**	19	**16**	29	**10**
10	**15**	20	**13**		

Sample 3. **The number of patients that visited emergency of hospital is given below.**

Variant №3					
№	**The number of patients**	№	**The number of patients**	№	**The number of patients**
1 day	**11**	11day	**11**	21day	**11**
2day	**13**	12day	**12**	22day	**13**
3day	**10**	13day	**13**	23day	**15**
4day	**15**	14day	**16**	24day	**16**
5day	**16**	15day	**15**	25day	**14**
6day	**14**	16day	**14**	26day	**15**
7day	**19**	17day	**17**	27day	**10**
8day	**18**	18day	**18**	28day	**12**
9day	**17**	19day	**19**	29day	**11**
10day	**12**	20day	**11**		

Sample 4. The weight of 26 students on a college campus was recorded as follows:

Variant №4					
№		№		№	
1	**55**	11	**73**	21	**83**
2	**67**	12	**55**	22	**83**
3	**60**	13	**56**	23	**85**
4	**45**	14	**56**	24	**78**
5	**70**	15	**67**	25	**70**
6	**78**	16	**78**	26	**72**
7	**77**	17	**90**		
8	**67**	18	**90**		
9	**67**	19	**56**		
10	**62**	20	**62**		

Sample 5. Weights of Malignant Tumours from 36 patients as recorded:

Variant №5							
№	**weight**	№	**weight**	№	**weight**	№	**weight**
1	**68**	11	**27**	23	**68**	33	**66**
2	**65**	12	**22**	22	**66**	32	**60**
3	**63**	13	**28**	23	**49**	33	**63**
4	**42**	14	**25**	24	**43**	34	**79**

5	**27**	15	**44**	25	**42**	**35**	**42**
6	**30**	16	**65**	26	**28**	**36**	**27**
7	**36**	17	**43**	27	**80**		
8	**28**	18	**25**	28	**65**		
9	**32**	19	**54**	29	**44**		
10	**76**	20	**56**	30	**46**		

Sample 6. The data belongs to the height of a particular species of grass in the college campus. Summarize this data:

Variant №6					
№	Height	№	Height	№	Height
1	**152**	11	**182**	21	**182**
2	**185**	12	**175**	22	**185**
3	**170**	13	**160**	23	**175**
4	**174**	14	**170**	24	**171**
5	**155**	15	**175**	25	**172**
6	**162**	16	**152**	26	**170**
7	**156**	17	**175**	27	**172**
8	**152**	18	**160**	28	**151**
9	**162**	19	**152**		
10	**163**	20	**157**		

Sample 7. Diastolic blood pressure in a population of 33 persons aged 20-45 years are prescribed in the table given below.

Variant №7							
№	**Diastolic blood pressure**	**№**	**Diastolic blood pressure**	**№**	**Diastolic blood pressure**	**№**	**Diastolic blood pressure**
1	**71**	11	**95**	21	**80**	31	**90**
2	**75**	12	**91**	22	**81**	32	**85**
3	**75**	13	**95**	23	**75**	33	**81**
4	**80**	14	**100**	24	**75**		
5	**80**	15	**100**	25	**75**		
6	**81**	16	**85**	26	**70**		
7	**70**	17	**70**	27	**90**		
8	**90**	18	**70**	28	**100**		
9	**70**	19	**75**	29	**95**		
10	**95**	20	**80**	30	**80**		

2.5. INTERVAL STATISTICAL DISTRIBUTION ROWS

Although a set of observations can be made more comprehensible and meaningful by means of an ordered data, further useful summarization may be achieved by grouping the data. To group a set of observations we select a set of contiguous, non overlapping intervals such that each value in the set of observations can be placed in one, and only one, of the intervals. These intervals are usually referred to as class intervals.

One of the first considerations when data are to be grouped is how many intervals to include. You may use a formula given by **Sturges** in the matter of deciding how many class intervals to use:

$$k = 1 + 3.3221 * lg\ n$$

Where k stands for the number of class intervals and n is the number of values in the data set under consideration.

2.5.1. TECHNIQUE OF CONSTRUCTING AN INTERVAL STATISTICAL DISTRIBUTION ROW

1. Sort the data from smallest value to largest value.
2. Determine the optimal width of the intervals. Class intervals generally should be of the same width. This width may be determined by dividing the range by k, the number of class intervals.

$$h = \frac{X\max - X\min}{1 + 3.32 \lg n}$$

3. Calculate the upper and lower bounds for the 1-st interval:

$a_1 = X_{min} - h/2$ (lower bounds)

$b_1 = a_1 + h$ (upper bounds).

4. Define the boundaries of all other intervals.

5. Distribute all values into intervals and count the number of values of the variable falling among the specified class intervals. (frequencies ***m)***.

6. Calculate the relative frequency for each class interval. $Pi = m_i / n$

It may be useful at times to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this information by dividing the number of values in the particular class interval by the total number of values. We may refer to the proportion of values falling within a class interval as the *relative frequency of occurrence* of values in that interval. A relative frequency may be interpreted also as the probability of occurrence within the given interval. This probability of occurrence is also called the *experimental probability* or the *empirical probability*.

7. Results are displayed in a table:

Table10. Interval statistical distribution rows

Interval X	$\|X_1, X_2\|$	$\| X_2, X_3\|$	…	$\|X_{i-1}, X_i\|$
m	m_1	m_2	…	m_i
P	P_1	P_2	…	P_i

A table which contains the intervals and their frequency (or relative frequency) is called a *frequency distribution* or a interval statistical distribution rows.

2.5.2. HISTOGRAM

We may display a frequency distribution (or a relative frequency distribution) graphically in the form of a *histogram*, which is a special type of bar graph.

When we construct a histogram the values of the variable under consideration are represented by the horizontal axis, while the vertical axis has as its scale the frequency (or relative frequency if desired) of occurrence. Above each class interval on the horizontal axis a rectangular bar is erected so that the height corresponds to the respective frequency when the class intervals are of equal width. The cells of a histogram must be joined and, to accomplish this, we must take into account the true boundaries of the class intervals to prevent gaps from occurring between the cells of our graph.

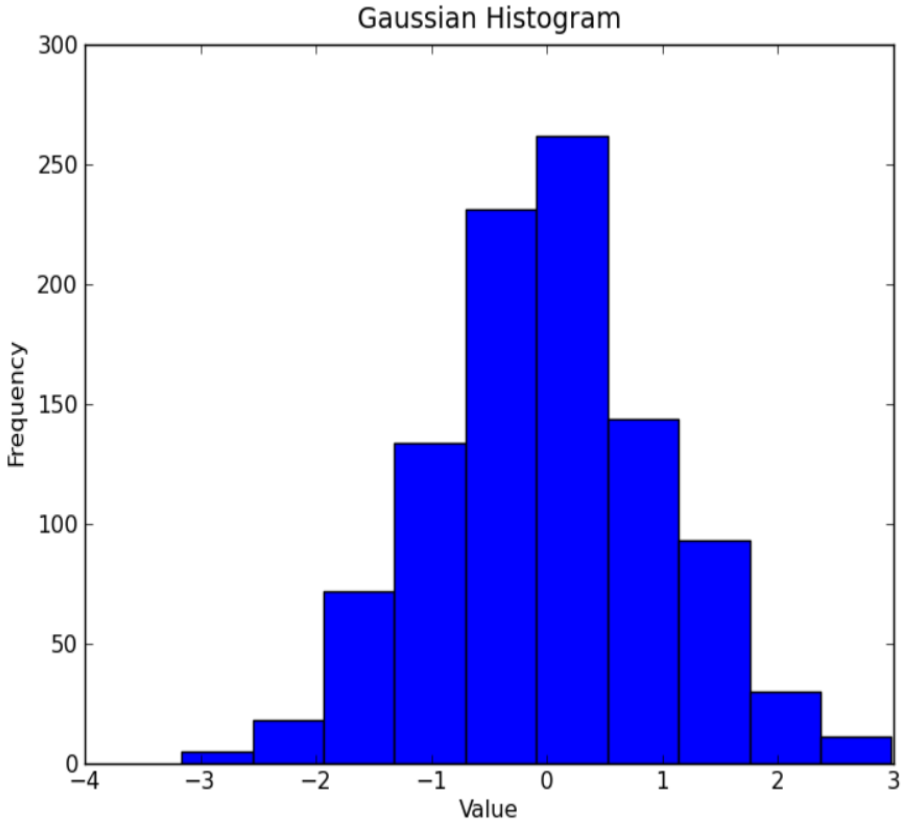


Figure 31. Histogram

3. THE NORMAL DISTRIBUTION

The most important distribution in all of statistics is the normal distribution. The formula for this distribution was first published by Abraham De Moivre (1667–1754) on November 12, 1733. Many other mathematicians figure prominently in the history of the normal distribution, including Carl Friedrich Gauss (1777–1855). *Karl Friedrich Gauss (1777-1855) was one of the foremost mathematicians of his time, advancing discoveries in number theory, algebra, geometry, probability, and other areas. The distribution is frequently called the Gaussian distribution in recognition of his contributions.

The normal density is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{[X-M(x)]^2}{2\sigma^2}}$$

where π and e are the familiar constants, 3.14159 . . . and 2.71828.. . , respectively, which are frequently encountered in mathematics. The two parameters of the distribution are μ, the mean, and σ the standard deviation.

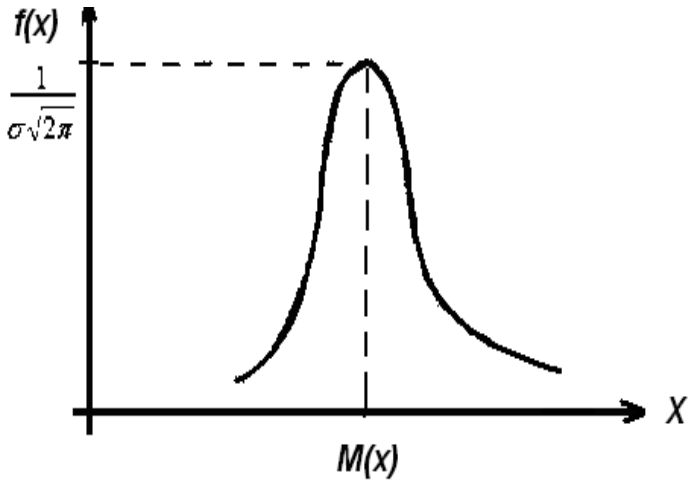


Figure 32. Normal distribution curve

3.1. CHARACTERISTICS OF THE NORMAL DISTRIBUTION

Probability distribution of a continuous variable.

1. It extends from minus infinity (-∞) to plus infinity (+∞).
2. It is unimodal, “bell-shaped”.
3. The mean, the median, and the mode are all equal.
4. It is symmetrical about its mean.
5. Data follow three sigma rule.

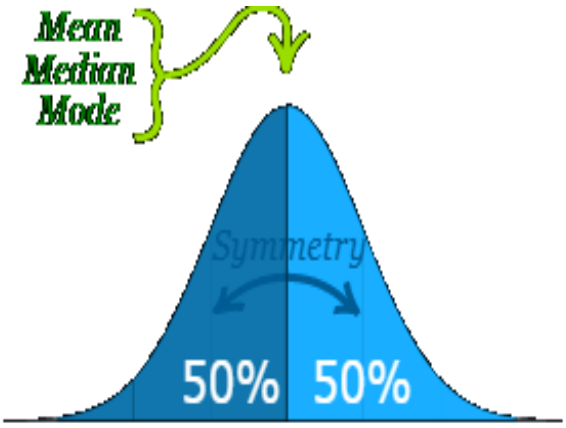


Figure 33. Symmetrical distribution of data

The normal distribution is completely determined by the parameters mean, μ and standard deviation, σ.

Different values of μ shift the graph of the distribution along the x-axis.

Three normal distributions with different means, but the same amount of variability.

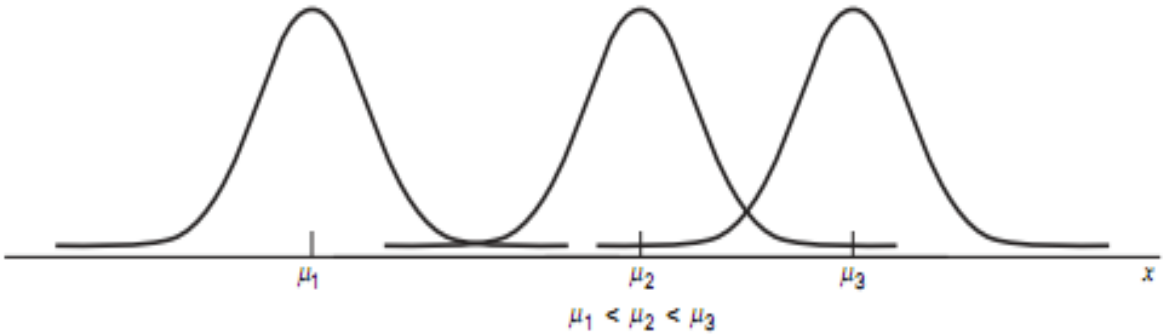


Figure 34. Three normal distributions with different means

Different values of σ determine the degree of flatness or peakedness of the graph of the distribution. Because of the characteristics of these two parameters, μ is often referred to as a location parameter and σ is often referred to as a shape parameter.

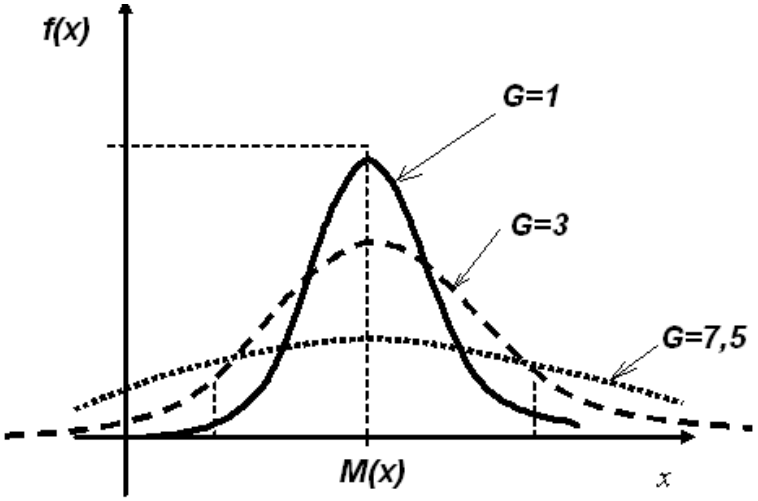


Figure 35. Three normal distributions with different standard deviations

For normal distribution:

• approximately 68% of observations fall within one standard deviation σ of the mean, in the interval (μ-σ, μ+σ)

• roughly 95% of observations fall within two standard deviations of the mean, in the interval (μ-2σ, μ+2σ)

• approximately 99,7% of observations fall within three standard deviations of the mean, in the interval (μ-3σ, μ+3σ).

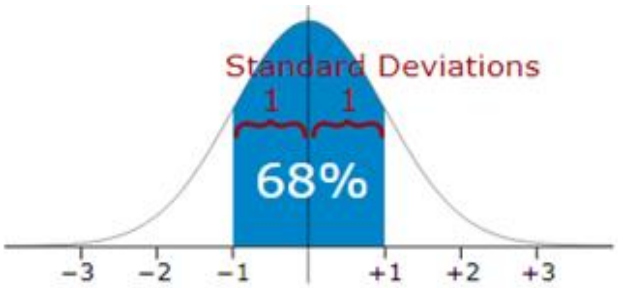


Figure 36. 68% of values are within 1 standard deviation of the mean

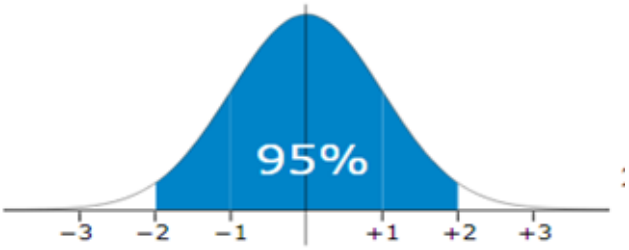


Figure 37. 95% of values are within 2 standard deviations of the mean

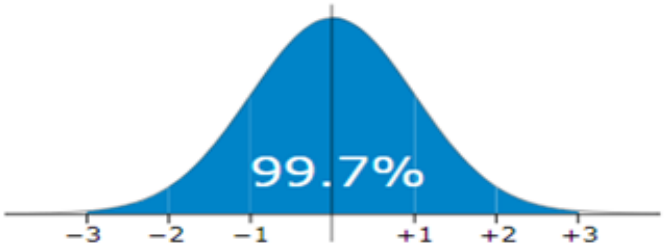


Figure 38. 99.7% of values are within 3 standard deviations of the mean

The three sigma rule states that, in a normal distribution, almost all the values remain within three standard deviations of the mean. Three sigma rule is also known as empirical rule.

3.2. ESTIMATION OF THE MEAN

In this section we consider how to use sample data to estimate unknown population parameters. Statistical inference uses sample data to form two types of estimators of parameters. A point estimate consists of a single number, calculated from the data that is the best single guess for the unknown parameter. An interval estimate consists of a range of numbers around the point estimate, within which the parameter is believed to fall.

3.2.1. POINT ESTIMATION

The object of point estimation is to calculate, from the sample data, a single number that is likely to be close to the unknown value of the population parameter. The available information is assumed to be in the form of a random sample X1,X2, . . . ,Xn of size n taken from the population. The object is to formulate a statistic such that its value computed from the sample data would reflect the value of the population parameter as closely as possible.

A point estimator of a unknown population parameter is a statistic that estimates the value of that parameter. A point estimate of a parameter is the value of a statistic that is used to estimate the parameter.

For instance, to estimate a population mean μ, perhaps the most intuitive point estimator is the sample mean:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Once the observed values x_1, x_2, . . . , x_n of the random variables Xi are available, we can actually calculate the observed value of the sample mean $\bar{X}$, which is called a point estimate of μ. A good point estimator of a parameter is one with sampling distribution that is centered on parameter, and has small standard error as possible. A point estimator is called unbiased if its sampling distribution centers on the parameter in the sense that the parameter is the mean of the distribution.

For example, the mean of the sampling distribution of the sample mean $\bar{X}$equals μ. Thus, $\bar{X}$ is an unbiased estimator of the population mean μ. A second preferable property for an estimator is a small standard error. An estimator whose standard error is smaller than those of other potential estimators is said to be efficient. An efficient estimator is desirable because, on the average, it falls closer than other estimators to the parameter. For example, it can be shown that under normal distribution, the sample mean is an efficient estimator, and hence has smaller standard error compared, e.g, to the sample median.

Point estimators of the population mean and standard deviation

The sample mean $\bar{X}$ is the obvious point estimator of a population mean μ. In fact, $\bar{X}$ is unbiased, and it is relatively efficient for most population distributions. It is the point estimator, denoted by ˆμ, used in this text:

$$\mu = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Moreover, the sample standard deviation s is the most popular point estimate of the population standard deviation. That is,

$$\sigma(x) = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

3.2.2. INTERVAL ESTIMATION

For point estimation, a single number lies in the forefront even though a standard error is attached. Instead, it is often more desirable to produce an interval of values that is likely to contain the true value of the unknown parameter. A confidence interval estimate of a parameter consists of an interval of numbers obtained from a point estimate of the parameter together with a percentage that specifies how confident we are that the parameter lies in the interval. The confidence percentage is called ***the confidence level.***

A confidence interval for a parameter is a range of numbers within which the parameter is believed to fall. The probability that the confidence interval contains the parameter is called the confidence coefficient. This is a chosen number close to 1, such as 0.95 or 0.99.

$$\bar{x} = \bar{x} \pm t * \frac{\sigma}{\sqrt{n}} \ ,$$

t - is a statistics of the distribution and $\frac{\sigma}{\sqrt{n}}$ is a standard error of the mean (sem), or the standard error (se). The standard error represents the estimated standard deviation obtained from a set of sample means.

If $$\Delta x = t * \frac{\sigma}{\sqrt{n}}$$

Then, $\overline{X} = \overline{X} \pm \Delta x$

3.2.3. TESTING THE DISTRIBUTION OF DATA FOR NORMALITY BY HAND IN EXCEL.

If the value of the random variable depends on many different factors, and if these factors have a different effect, then the random variable being considered is subject to the normal distribution law. In the given work, the measurement results are considered as a random quantity. There is reason to believe that the law of distribution of measurement results of some value will be normal. The purpose of this study is to test this hypothesis.

According to an individual task, make measurements of a given parameter. Write the measurement results to Table 1. Processing of measurement results is performed in the following order (see example):

1. Construct ranked variation data row.
2. Find the $\mathbf{X_{max}}$ and $\mathbf{X_{min}}$ values of the measurement results in the table.
3. Define width of the interval (**h**).
4. Determine the interval boundaries ($\mathbf{a_k}$ and $\mathbf{b_k}$)and write them to Table 2.
5. Define the mean value $\bar{X}$ of the interval.
6. Determine the frequency ($\mathbf{m_k}$) of the measured value at each interval.
7. Determine the probability of random variables at each interval ($\mathbf{P_k}$).
8. Determine the expected value **M(x)**

9. Determination of sample variance **D(x**).

10. Define the standard deviation **σ(x)**.

11. Calculate the normal deviation coefficient $\mathbf{t_k}$ for each interval.

12. Detection of the distribution function **f** $\mathbf{(t_k)}$ for each interval corresponding to $\mathbf{t_k}$ **.** (Appendix 2).

13. Calculate value of theoretical probability of distribution **P (X).**

14. Construct a histogram, theoretical and empirical distribution curves.

15. To define a confidence interval

16. Determine central tendencies (mean, median and mode) of data.

17. Checking the data for three sigma rules.

18. Do Conclusion.

Example2: Investigating dependence on mortality from cardiovascular mortality. Measurement results are given in Table 11.

Table 11. Measurement results

№	Age	№	Age	№	Age
1	23	**11**	54	**21**	41
2	60	**12**	73	**22**	66
3	35	**13**	55	**23**	46
4	75	**14**	56	**24**	68
5	38	**15**	54	**25**	73
6	65	**16**	59	**26**	54
7	68	**17**	28	**27**	37
8	58	**18**	62		
9	63	**19**	47		
10	52	**20**	64		

1. It is necessary to separate the range of values into intervals for the construction of a histogram

First of all, we will create a ranked variation data row:

23 28 35 37 38 41 46 47 52 54 54 54 55 56 58 59 60 62 63 64 65 66 68 68 73 73 75

2. We will Find the $\mathbf{X_{max}}$ and $\mathbf{X_{min}}$ values**:** $\mathbf{X_{max}}$=75**;** $\mathbf{X_{min}}$ =23.

3. We define the **h** – width of interval by the formula of Sturges:

4.

$$h = \frac{\text{Xmax} - \text{Xmin}}{1 + 3.321\ lgn} = \frac{75 - 23}{1 + 3.321\ lg27} = \frac{52}{5.75} = 9.037 \approx 9$$

where $\mathbf{X_{max}}$ and $\mathbf{X_{min}}$ are the maximum and minimum values of random variables**;** **n** is the sample size**.**

5. Determining of the interval boundaries (ak and bk).

The beginning of the first interval:

$$a_1 = \mathrm{X_{min}} - \frac{\mathrm{h}}{2} = 23 - \frac{9}{2} = 23 - 4{,}5 = 18{,}5 \approx 19$$

The end of the first interval:

$$b_1 = a_1 + h = 19 + 9 = 28$$

According the order, the beginning of the next interval is equal to the end of the previous interval. Consequently, we calculate the boundaries as follows:

$a_2 = b_1 = 28$ $\quad b_2 = a_2 + h = 28 + 9 = 37$

$a_3 = b_2 = 37$ $\quad b_3 = a_3 + h = 37 + 9 = 46$

$a_4 = b_3 = 46$ $\quad b_4 = a_4 + h = 46 + 9 = 55$

$a_5 = b_4 = 55$ $\quad b_5 = a_5 + h = 55 + 9 = 64$

$a_6 = b_5 = 64$ $\quad b_6 = a_6 + h = 64 + 9 = 73$

$a_7 = b_6 = 73$ $\quad b_7 = a_6 + h = 73 + 9 = 82$

We define the boundary value so that the value does not exceed $\mathbf{X_{max}}$. We will write the boundaries of obtained seven intervals in Table 12.

6. Define the mean value $\bar{X}$of the interval.

For each interval we calculate the average of the random variables:

$$\bar{X}_i = \frac{a_k + b_k}{2}$$

$$\bar{X}_1 = \frac{a_1 + b_1}{2} = \frac{19 + 28}{2} = 23{,}5$$

$$\bar{X}_2 = \frac{a_2 + b_2}{2} = \frac{28 + 37}{2} = 32{,}5$$

$$\bar{X}_3 = \frac{a_3 + b_3}{2} = \frac{37 + 46}{2} = 41{,}5$$

$$\bar{X}_4 = \frac{a_4 + b_4}{2} = \frac{46 + 55}{2} = 50{,}5$$

$$\bar{X}_5 = \frac{a_5 + b_5}{2} = \frac{55 + 64}{2} = 59{,}5$$

$$\bar{X}_6 = \frac{a_6 + b_6}{2} = \frac{64 + 73}{2} = 68{,}5$$

$$\bar{X}_7 = \frac{a_7 + b_7}{2} = \frac{73 + 82}{2} = 77{,}5$$

7. Determine the frequency (m_k) of the measured value at each interval (refer to Table 1).

The first $[a_2: b_2)$ = [28:37) interval contains only one number 23. Consequently, $\mathbf{m_1} = 1$;

The second $[a_2: b_2)$ = [28:37) interval contains two numbers 27, 35. Consequently, $\mathbf{m_2}=2$;

Similarly count to the rest of the intervals and we write the results in Table 2. Check the correctness of calculations by formula $\sum_i^k m_k = n = 27$.

8. Determine the probability of random variables at each interval ($\mathbf{P_k}$) by formula:

$$P_i = \frac{m_i}{n}$$

$$P_1 = \frac{m_1}{n} = \frac{1}{27} = 0{,}037$$
$$P_2 = \frac{m_2}{n} = \frac{2}{27} = 0{,}074$$
$$P_3 = \frac{m_3}{n} = \frac{3}{27} = 0{,}111$$
$$P_4 = \frac{m_4}{n} = \frac{6}{27} = 0{,}222$$
$$P_5 = \frac{m_5}{n} = \frac{7}{27} = 0{,}259$$
$$P_6 = \frac{m_6}{n} = \frac{5}{27} = 0{,}185$$
$$P_7 = \frac{m_7}{n} = \frac{3}{27} = 0{,}111$$

We write the results in Table 2. Check the correctness of calculations by formula: $\sum_i^k P_k = 1$.

$$\sum_i^k P_k = 0{,}037 + 0{,}074 + 0{,}111 + 0{,}222 + 0{,}259 + 0{,}185 + 0{,}111 = 0{,}999 \approx 1$$

9. Determine the expected value M(x)

$$\mathbf{M(x)} = \sum_i^k \bar{\boldsymbol{X}}_{\boldsymbol{i}} * P_k$$

where k is number of the intervals.

$$\begin{aligned}\mathbf{M(x)} &= \bar{X}_1 * P_1 + \bar{X}_2 * P_2 + \bar{X}_3 * P_3 + \bar{X}_4 * P_4 + \bar{X}_5 * P_5 + \bar{X}_6 * P_6 + \bar{X}_7 * P_7 \\ &= 23{,}5 * 0{,}037 + 32{,}5 * 0{,}074 + 41{,}5 * 0{,}111 + 50{,}5 * 0{,}222 \\ &+ 59{,}5 * 0{,}259 + 68{,}5 * 0{,}185 + 77{,}5 * 0{,}111 \\ &= 0{,}870 + 2{,}405 + 4{,}607 + 11{,}211 + 15{,}411 + 12{,}673 + 8{,}603 \\ &= 55{,}78 \approx \mathbf{56}\end{aligned}$$

M(x) *is constant for all intervals! And equals to 56!!!*

10. Determination of sample variance D(x).

$$\mathrm{D(x)} = \sum_{i}^{k} \overline{[X_i} - \mathrm{M(x)}]^{\,2} * P_k$$

$$= \overline{[X_1} - \mathrm{M(x)}]^{\,2} * P_1 + \overline{[X_2} - \mathrm{M(x)}]^{\,2} * P_2 + \overline{[X_3} - \mathrm{M(x)}]^{\,2} * P_3$$
$$+ \overline{[X_4} - \mathrm{M(x)}]^{\,2} * P_4 + \overline{[X_5} - \mathrm{M(x)}]^{\,2} * P_5 + \overline{[X_6} - \mathrm{M(x)}]^{\,2} * P_6$$
$$+ \overline{[X_7} - \mathrm{M(x)}]^{\,2} * P_7 = 193{,}390 \approx \mathbf{193}$$

11. Define the standard deviation σ(x).

$$\sigma(x) = \sqrt{\mathrm{D(x)}} = \sqrt{193} = 13{,}89 \approx 14$$

12. Calculate the normal deviation coefficient $\mathbf{t_k}$ for each interval.

$$\mathrm{tk} = \frac{|\bar{X}_i - \mathrm{M(x)}|}{\sigma}$$

$$\mathrm{t(x_1)} = \mathrm{t(23{,}5)} = \frac{|23{,}5 - 56|}{14} = 2{,}3$$

$$\mathrm{t(x_2)} = \mathrm{t(32{,}5)} = \frac{|32{,}5 - 56|}{14} = 1{,}7$$

$$\mathrm{t(x_3)} = \mathrm{t(41{,}5)} = \frac{|41{,}5 - 56|}{14} = 1{,}03$$

$$\mathrm{t(x_4)} = \mathrm{t(50{,}5)} = \frac{|50{,}5 - 56|}{14} = 0{,}4$$

$$\mathrm{t(x_5)} = \mathrm{t(59{,}5)} = \frac{|59{,}5 - 56|}{14} = 0{,}3$$

$$\mathrm{t(x_6)} = \mathrm{t(68{,}5)} = \frac{|68{,}5 - 56|}{14} = 0{,}9$$

$$\mathrm{t(x_7)} = \mathrm{t(77{,}5)} = \frac{|77{,}5 - 56|}{14} = 1{,}5$$

We write the results in Table 12.

13. Detection of the distribution function **f ($\mathbf{t_k}$)** for each interval using the Appendix 2.

$$\mathrm{f(tk\ )} = \frac{1}{\sigma} * \frac{1}{\sqrt{2\pi}} * e^{\frac{[X_i - M(x)]^2}{2\sigma^2}} = \frac{1}{\sigma} f(t)\text{-}$$

The formula for calculating F(t)

Since $\mathbf{t_k}$ is known, we can write f (t) from the table in Table 14:

f(2,3)=0,03
f(1,7)=0,09
f(1,03)=0,24
f(0,4)=0,37
f(0,3)=0,38
f(0,9)=0,27
f(1,5)=0,13

14. Calculate value of theoretical probability of distribution **P (X).**

$$\mathrm{P(x)} = \frac{1}{\sigma} * f(t_k) * h$$

$$\mathrm{P}(x_1) = \frac{1}{14} * 0{,}03 * 9 \approx 0{,}02$$

$$\mathrm{P}(x_2) = \frac{1}{14} * 0{,}09 * 9 \approx 0{,}06$$

$$\mathrm{P}(x_3) = \frac{1}{14} * 0{,}24 * 9 \approx 0{,}15$$

$$\mathrm{P}(x_4) = \frac{1}{14} * 0{,}37 * 9 \approx 0{,}24$$

$$\mathrm{P}(x_5) = \frac{1}{14} * 0{,}38 * 9 \approx 0{,}24$$

$$\mathrm{P}(x_6) = \frac{1}{14} * 0{,}27 * 9 \approx 0{,}17$$

$$\mathrm{P}(x_7) = \frac{1}{14} * 0{,}13 * 9 \approx 0{,}08$$

We write the results in Table 12.

Table 12. Results

h	ak	bk	mk	Pk= m/n	$\bar{X}$= (ak+bk)/2	$\bar{X}$*Pk	$\overline{(X}$- M(x))^2 *Pk	t_k	f(t)	p(x)
9	19	28	1	0,037	23,5	0,870	39,081	2,3	0,03	0,02
	28	37	2	0,074	32,5	2,405	40,867	1,7	0,09	0,06
	37	46	3	0,111	41,5	4,607	23,338	1,03	0,24	0,15
	46	55	6	0,222	50,5	11,211	6,7160	0,4	0,37	0,24
	55	64	7	0,259	59,5	15,411	3,173	0,3	0,38	0,24
	64	73	5	0,185	68,5	12,673	28,906	0,9	0,27	0,17
	73	82	3	0,111	77,5	8,603	51,310	0,13	0,13	0,08
	Сумма		**27**	**1**	**M(x) ≈56,000**		**D(x) ≈193**			
							σ(x) ≈14			

15. Construct a histogram, theoretical and empirical distribution curves in Excel.

We use the data in Table 2.

Histogram (a_k,b_k:Pk)

We construct an empirical curve using Pk - ($\bar{X}$:Pk);

We construct an theoretical curve using P(x) - ($\bar{X}$:P(x));

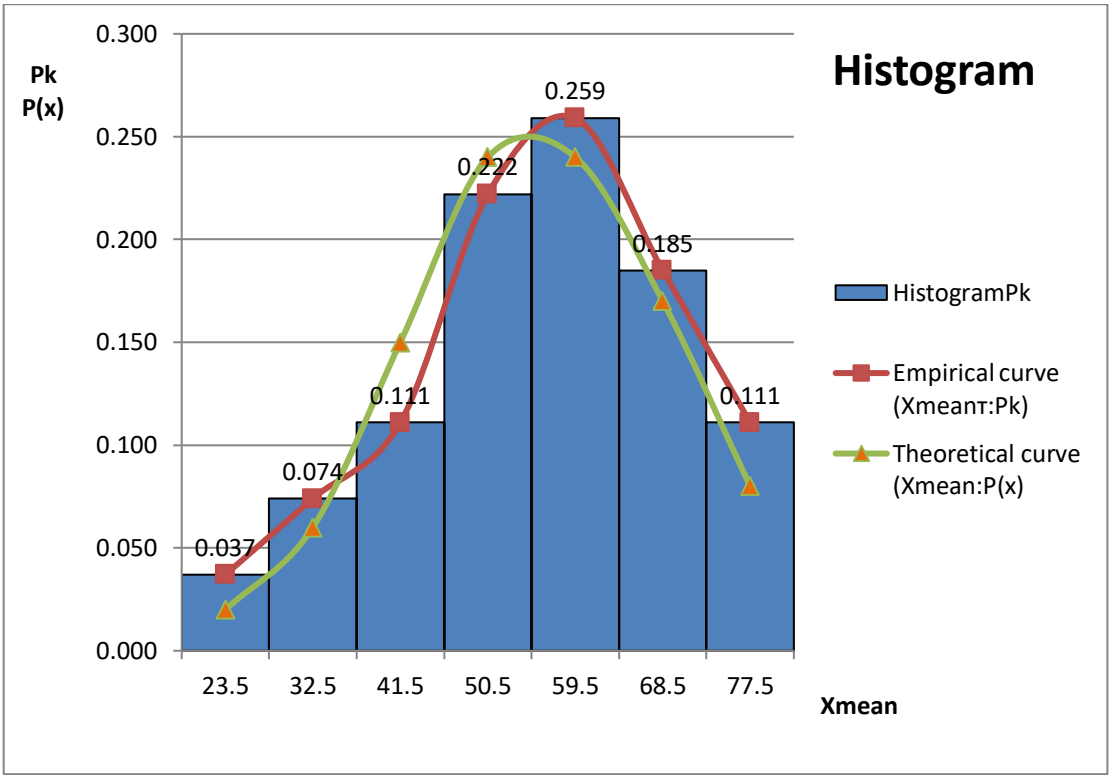


Figure 39. Histogram, empirical and theoretical curves

Comparing the histogram and theoretical curve, we can be sure that the investigated amount (Age) is distribute according to the normal distribution law. The probability of the random variable is 56.

$$\bar{X} = \mathrm{M(x)} = 56$$

16. To define a confidence interval by the formula:

$$\mathrm{X=M(x)\pm\Delta X, \text{ where } \Delta X = t * \frac{\sigma}{\sqrt{n}}.}$$

If the probability of a confidence is **P** = 0.95, the freedom level **k** = n-1 = 27-1 = 26 at n = 27, the normal deviation coefficient $t_{p.k}$ = 2,06 **(**Appendix 1. Normal deviation coefficient $\mathbf{t_{p.k}}$ **)**

$$= \frac{t_{p.k} * \sigma(x)}{\sqrt{n}} = \frac{2{,}06 * 14}{\sqrt{27}} = \frac{28{,}84}{5{,}2} = 5{,}5$$

Confidence Interval: $X = \mathbf{M(x)} \mp \Delta X$

that is, $X_1 = 56 - 5{,}5 = 50{,}5$ $X_2 = 56 + 5{,}5 = 61{,}5$

17. Determine central tendencies (mean, median and mode) of data by Data analysis tool:

Column1	
Mean	55
Standard Error	2.659852008
Median	56
Mode	54
Standard Deviatic	13.82099646
Sample Variance	191.019943
Kurtosis	-0.23004212
Skewness	-0.644281233
Range	52
Minimum	23
Maximum	75
Sum	1474
Count	27
Confidence Level(	5.467404051

Figure 40. Summary statistics by data analysis tool

18. Checking the data for three sigma rules.

Given dataset follows to 3 sigma rules: normal and actual values are close to each other.

(μ-σ, μ+σ)	Percentage of data in sigma area for normal distribution	normal			actual	Percentage of data in sigma area for given distribution
	68%	0	41	69	19	70%
	100	27				
1 sigma area						
		normal			actual	
(μ-2σ, μ+2σ)	95%	0	28	83	26	96%
2 sigma area	100	27				
					actual	
(μ-3σ, μ+3σ)	99.7%	0	14	97	27	100%
3 sigma area	100	27				

Figure 41. Normal and actual values distributions in three sigma areas

19. Conclusion:

Accordingly, the mortality rate as a result of cardiovascular disease depends on age and is distributed according to the law of normal distribution. We can conclude that the data are distributed by normal law, as the graph represents "Bell shaped". And also the mean, median, mode "(Mean(55)=Me(56)=Mo(54))" are almost equal and it follows the "three sigma law": 70% of data falls in the 1 sigma area, 96% of data falls in the 2 sigma area, 100% of data falls in the 3 sigma area. In 95 cases of 100 cases the mortality rate is 50.5 ÷ 61.5 years old due to cardiovascular disease.

Table 13. Normal deviation coefficient $t_{p.k}$

P/к	90%	95%	99%	99,9%
2	2.92	4.30	9.93	31.60
3	2.35	3.18	5.84	12.94
4	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.86
6	1.94	2.45	3.71	6.96
7	1.90	2.37	3.50	5.41
8	1.86	2.31	3.36	5.04
9	1.83	2.26	3.25	4.78
10	1.80	2.23	3.17	4.59
11	1.78	2.23	3.11	4.49
12	1.77	2.20	3.06	4.32
13	1.76	2.18	3.01	4.22
14	1.75	2.16	2.98	4.14
15	1.74	2.15	2.95	4.07
16	1.73	2.13	2.92	4.02
18	1.72	2.12	2.88	3.92
20	1.72	2.10	2.85	3.85
22	1.71	2.09	2.82	3.79
24	1.71	2.07	2.80	3.75
26	1.70	2.06	2.78	3.75
30	1.70	2.05	2.75	3.65
40	1.70	2.04	2.70	3.55
60	1.70	2.02	2.66	3.46
120	1.60	2.00	2.62	3.37
∞	1.96	1.98	2.58	3.00

Table 14.The value of the function $f(t) = \frac{1}{\sqrt{2*\pi}} * e^{\frac{t^2}{2}}$

t	f(t)	t	f(t)	t	f(t)	t	f(t)
0.0	0.3989	1.1	0.2179	2.2	0.0355	3.2	0.0024
0.1	0.3970	1.2	0.1942	2.3	0.0283	3.3	0.0017
0.2	0.3910	1.3	0.1714	2.4	0.0224	3.4	0.0012
0.3	0.3814	1.4	0.1497	2.5	0.0175	3.5	0.0009
0.4	0.3683	1.5	0.1295	2.6	0.0136	3.6	0.0006
0.5	0.3521	1.6	0.1109	2.7	0.0104	3.7	0.0004
0.6	0.3332	1.7	0.0940	2.8	0.0079	3.8	0.0003
0.7	0.3123	1.8	0.0790	2.9	0.0060	3.9	0.0002
0.8	0.2897	1.9	0.0656	3.0	0.0044	4.0	0.0001
0.9	0.2661	2.0	0.0540	3.1	0.0033		
1.0	0.2420	2.1	0.0440				

3.2.3.1 TESTING THE DISTRIBUTION OF DATA FOR NORMALITY BY DATA ANALYSIS PACKAGE IN EXCEL

This example teaches you how to create a histogram in Excel.

Step 1. First, enter the ranked variation row and the interval boundaries as shown below:

	A	B	C	D
1	Ranked variational row	h-width	ak	bk
2	23	9	18	28
3	28		28	37
4	35		37	46
5	37		46	55
6	38		55	64
7	41		64	73
8	46		73	82
9	47			
10	52			
11	54			
12	54			
13	54			
14	55			
15	56			
16	58			
17	59			
18	60			
19	62			
20	63			
21	64			
22	65			
23	66			
24	68			
25	68			
26	73			
27	73			
28	75			

Figure 42. Data entry in Excel

Step 2. On the Data tab, in the Analysis group, click **Data Analysis**.

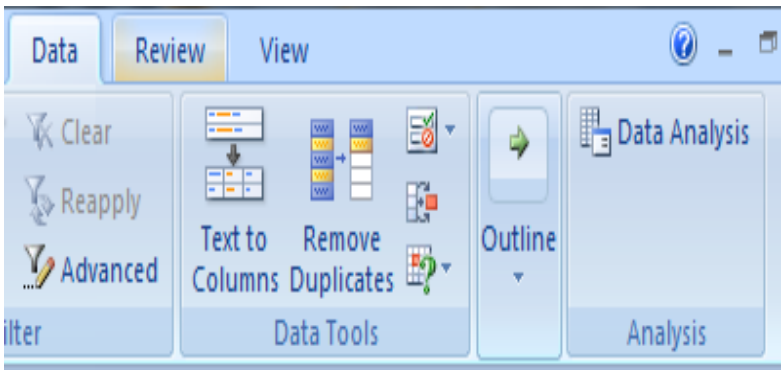


Figure 43. Data analysis command in Excel

Note: can't find the Data Analysis button? Add the Analysis ToolPak add-in from Excel Options.

Step 3. Select Histogram and click OK.

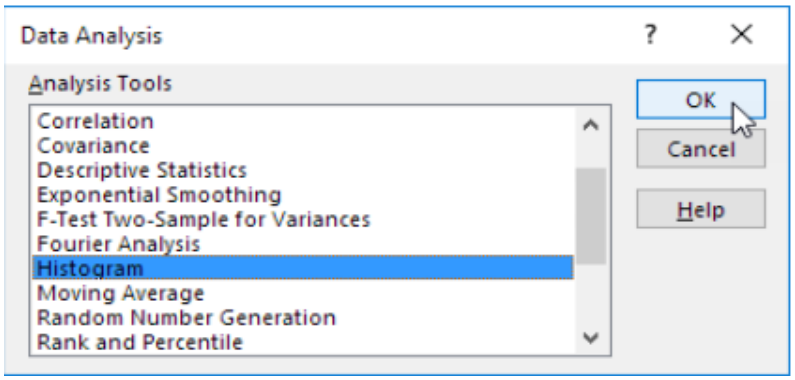


Figure 44. Histogram command in Data analysis

Step 4. Select the Input range A2:A28.

Step 5. Click in the Bin Range box and select the range C2:D8.

Step 6. Click the Output Range option button, click in the Output Range box and select any empty cell, for example, cell C11.

Step 7. Check Chart Output.

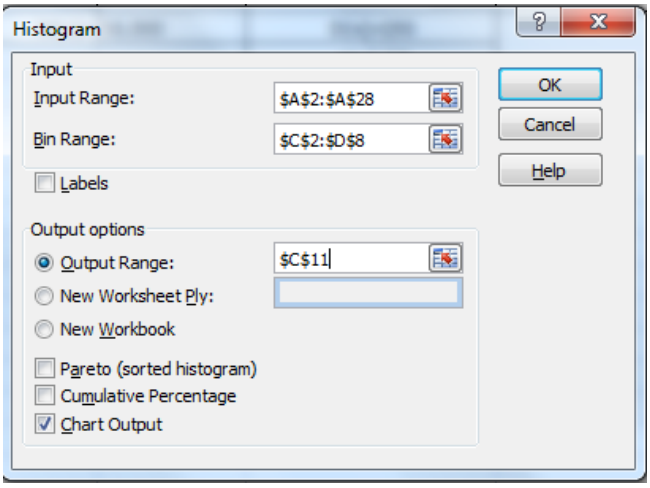


Figure 45. Histogram box

Step 8. Click OK.

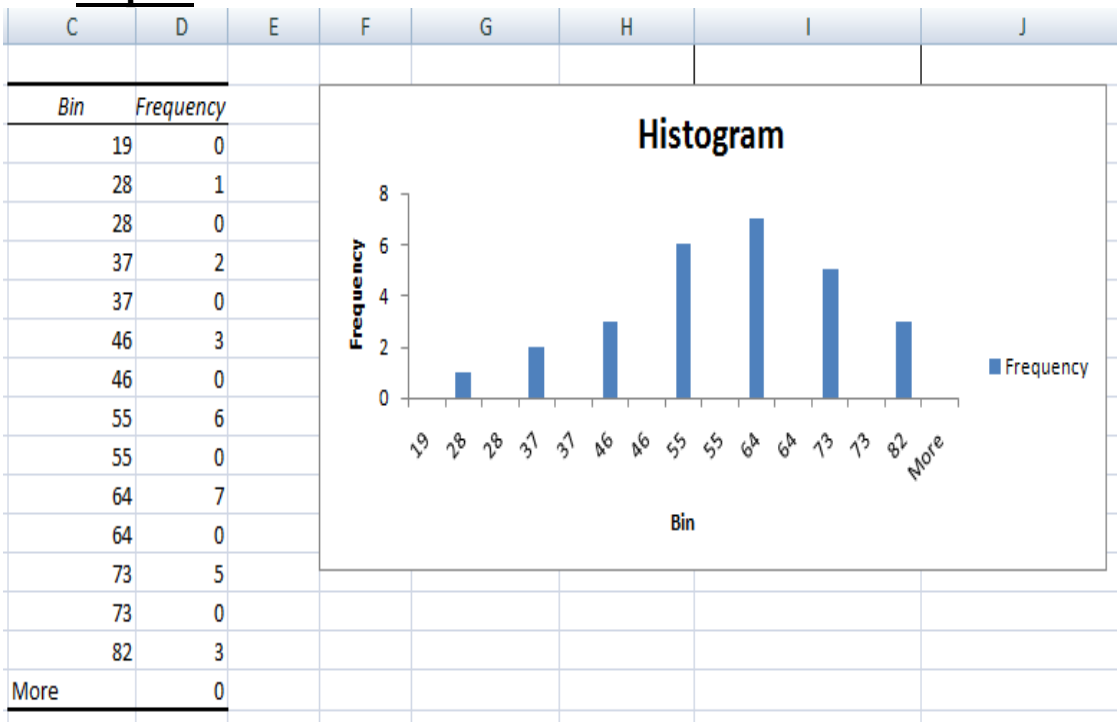


Figure 46. Results

Step 9. Click the legend on the right side and press Delete.

Step 10. Properly label your bins.

Step 11. To remove the space between the bars, right click a bar, click Format Data Series and change the Gap Width to 0%.

Step 12. To add borders, right click a bar, click Format Data Series, click the Fill & Line icon, click Border and select a color.

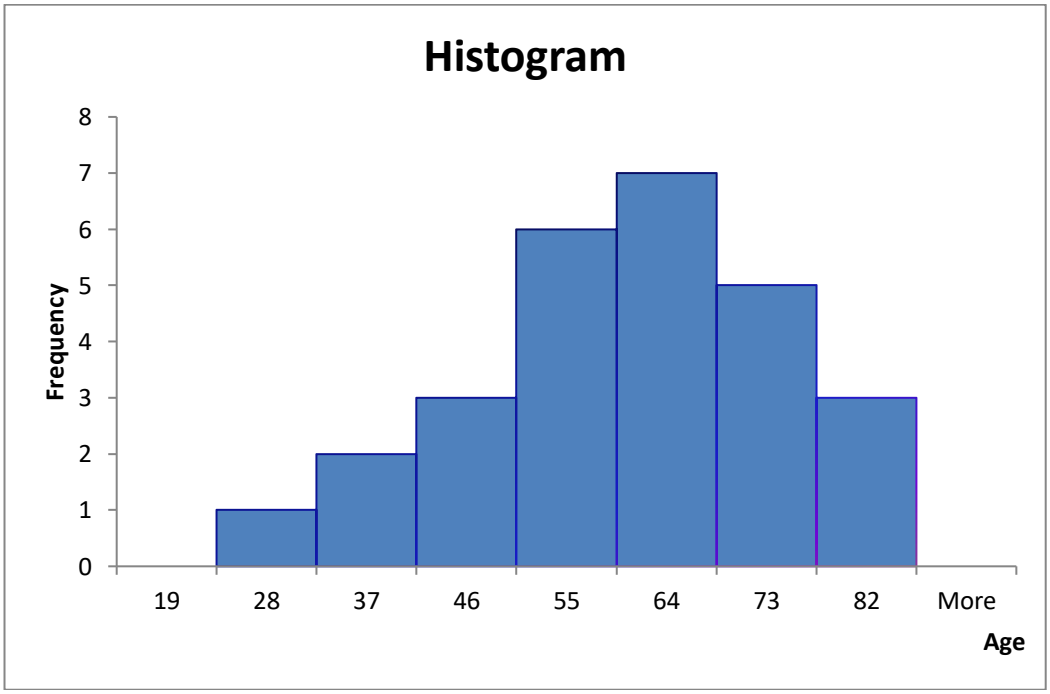
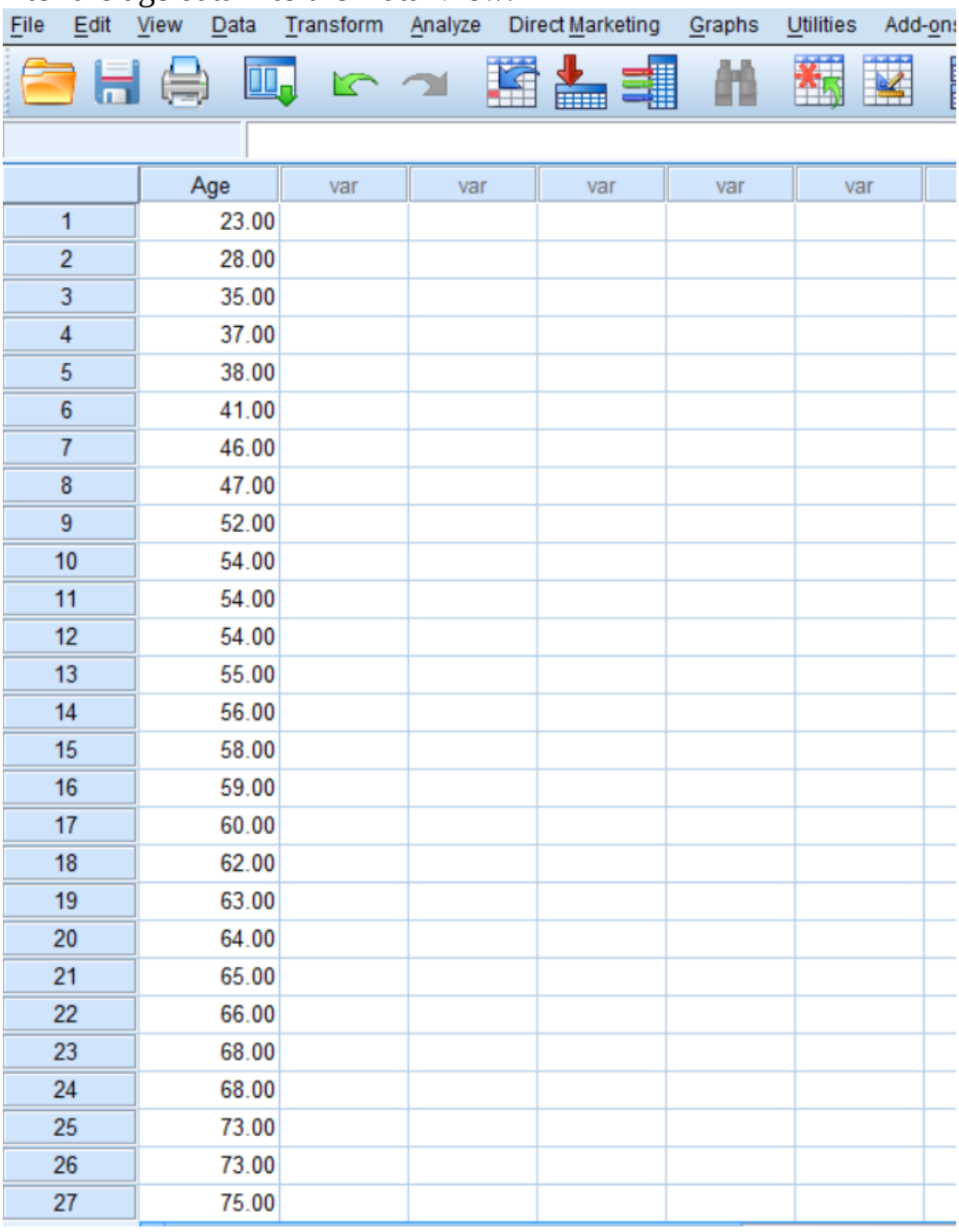


Figure 47. Histogram (Age, frequency)

3.2.3.2. TESTING OF DISTRIBUTION USING DESCRIPTIVE STATISTICS IN THE SPSS STATISTICS PROGRAM.

1. Run the SPSS program.
2. Enter the age data into the Data View.



	Age	var	var	var	var	var
1	23.00					
2	28.00					
3	35.00					
4	37.00					
5	38.00					
6	41.00					
7	46.00					
8	47.00					
9	52.00					
10	54.00					
11	54.00					
12	54.00					
13	55.00					
14	56.00					
15	58.00					
16	59.00					
17	60.00					
18	62.00					
19	63.00					
20	64.00					
21	65.00					
22	66.00					
23	68.00					
24	68.00					
25	73.00					
26	73.00					
27	75.00					

Figure 48. SPSS Program window - **Data view**

To get descriptive statistics for numeric variables Select Analyze from the menu. From the drop-down window choose "Descriptive statistics" - "Frequencies". In the "Frequencies" window highlight the name of variable “Age” and using the button  transfer it to the Variables window.

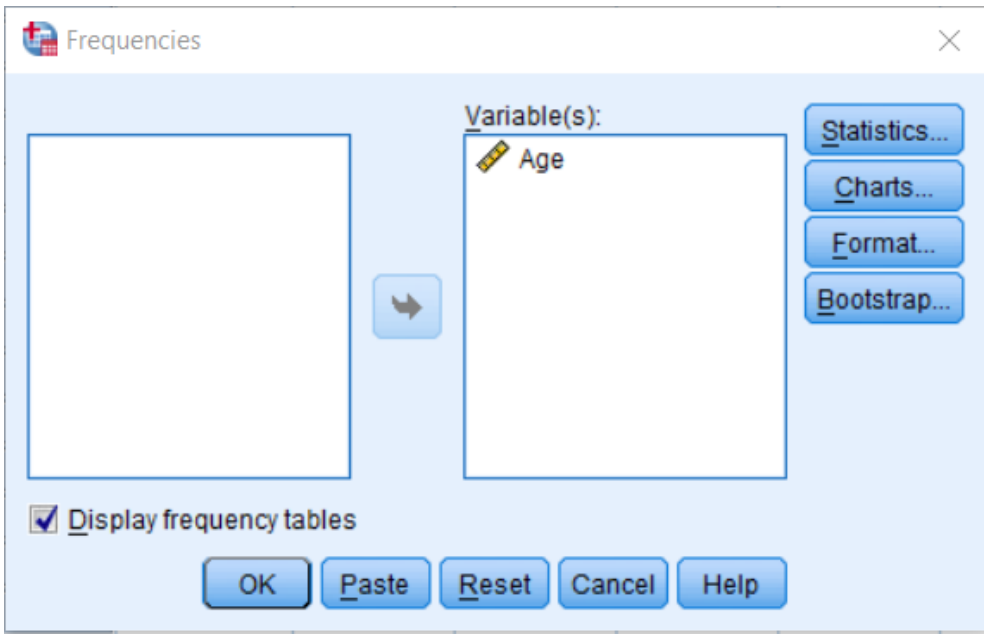


Figure 49. Frequencies box

Click the Statistics button.

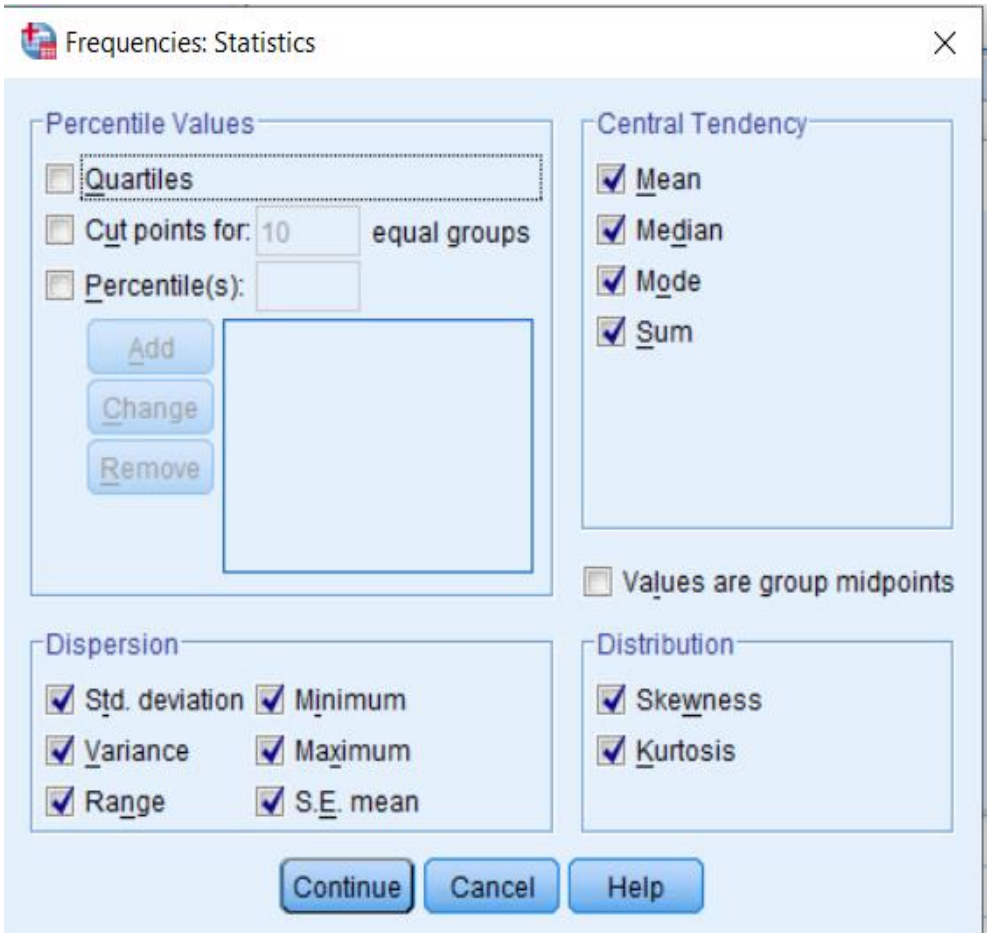


Figure 50. Frequency: Statistics dialog box.

For variable frequency, we define the following characteristics: mean, median, mode, standard deviation, variance, range, minimum, maximum, standard error, asymmetry and kurtosis. Proceed as follows:

In the Frequencies: Statistics dialog (Frequencies: Statistics) check the boxes of the desired characteristics

Return to the Frequencies window by clicking on Continue

In the Frequencies dialog box, click the Reset button to cancel the previous settings.

Transfer the variable frequency to the list of output variables.

In the Frequencies dialog, deactivate the Display frequency tables option. Click on the OK button.

The following results will appear in the Output window:

Statistics

Age

N	Valid	27
	Missing	0
Mean		54.5926
Std. Error of Mean		2.65985
Median		56.0000
Mode		54.00
Std. Deviation		13.82100
Variance		191.020
Skewness		-.644
Std. Error of Skewness		.448
Kurtosis		-.230
Std. Error of Kurtosis		.872
Range		52.00
Minimum		23.00
Maximum		75.00
Sum		1474.00

Figure 51. Calculation results Descriptive statistics.

The asymmetry coefficient is As =-0.644, the kurtosis Ex = -0.230. It can be seen that the distribution is only slightly shifted to the left and has a slightly flattened top in the vicinity of the mode, which suggests that the distribution is normal. For the final decision on the possibility of applying the criterion, we will check the normal distribution using a histogram and a normal distribution diagram, which in SPSS are divided into two types: P-P - normal probability graph and Q-Q graph - in the form of a scatterplot The Q-Q graph shows the dependence of the expected cumulative frequencies on the actual cumulative frequencies.

Select the Histograms item in the Date Editor: Graphics (Charts) dialog box.

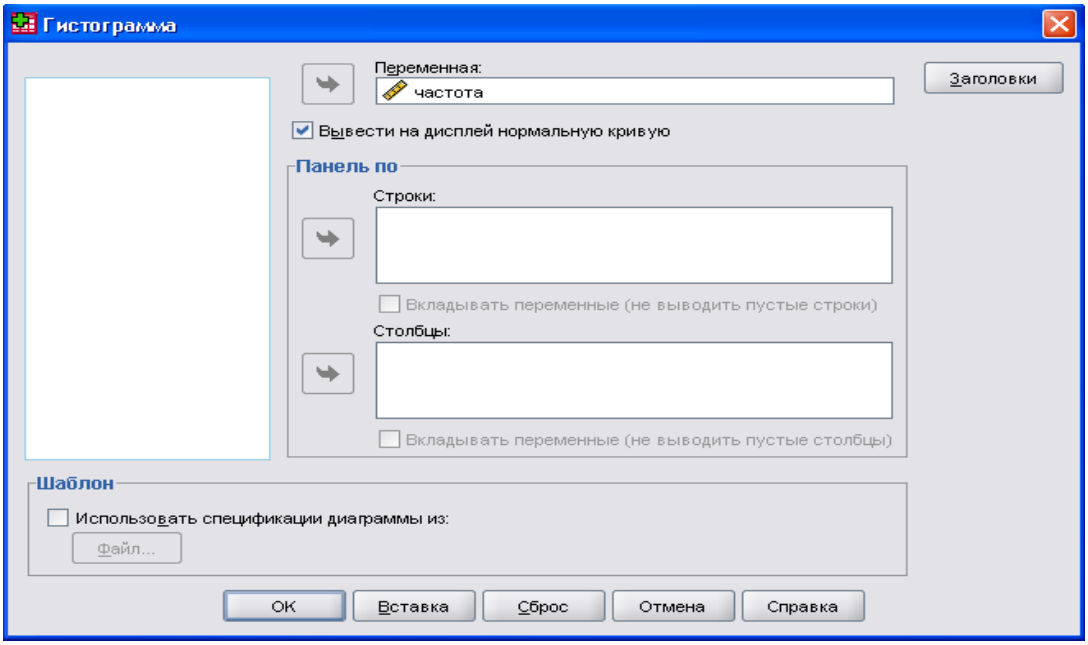


Figure 52. Histogram dialog box.

Transfer the variable frequency to the list of output variables.

Check the box Show normal curve (With normal curve) Click the OK button

Graph-Histogram will appear in the output window. The histogram in form is close to the law of normal distribution.

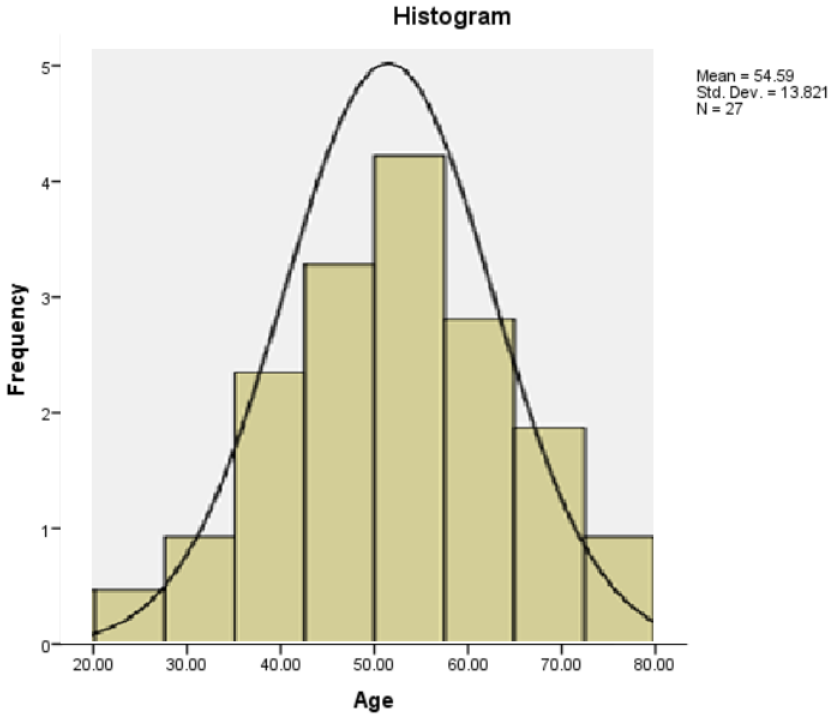


Figure 53. Histogram in SPSS Statistics program

STUDY QUESTIONS:

1. Continuous and discrete random variables.
2. Interval variation row.
3. Creation of a statistical interval row of distribution.
4. Numerical characteristics of the interval statistical row.
5. Sturges Formula.
6. Construction of the histogram.
7. Histogram of frequencies and relative frequencies.
8. Normal distribution of data.
9. The main properties of the normal distribution.
10. The three sigma rule.
11. Evaluation of the parameters of the general population using point and interval estimates.
12. Finding the confidence intervals with a given reliability.
13. Application of MS Excel and SPSS programs.

REVISIONARY PROBLEMS:

On the given samples do the following calculations:

I. Construct an interval statistical number of distribution.
II. Determine the numerical characteristics of the interval statistical row.
III. Construct a histogram, theoretical and empirical distribution curves.
IV. Check the sample for normality and draw a conclusion.
V. Find confidence intervals with a given reliability and Interpret the obtained results.
VI. Write a conclusion.

Sample 1. Investigating dependence on mortality from cardiovascular mortality. Measurement results are given in the table:

Variant №1							
№		№		№		№	
1	**35**	11	**48**	21	**61**	31	**71**
2	**25**	12	**49**	22	**62**	32	**72**
3	**35**	13	**85**	23	**63**	33	**74**
4	**37**	14	**54**	24	**65**	34	**75**
5	**40**	15	**58**	25	**66**	35	**77**
6	**42**	16	**57**	26	**67**	36	**78**
7	**43**	17	**58**	27	**68**	37	**85**
8	**84**	18	**58**	28	**69**		
9	**47**	19	**59**	29	**70**		
10	**47**	20	**60**	30	**71**		

Sample 2. These leukocyte counts in the blood of sarcomatous rats after treatment with leonin drug:

Variant №2							
№		№		№		№	
1	**14,5**	11	**20**	21	**21**	31	**24,3**
2	**17,5**	12	**25,3**	22	**25,7**	32	**26,6**
3	**18,5**	13	**27**	23	**27,4**	33	**27,6**
4	**18,5**	14	**28,7**	24	**28,8**	34	**29,1**
5	**19,8**	15	**35**	25	**30,6**	35	**30,6**
6	**19,8**	16	**32,3**	26	**35,7**	36	**34**
7	**37,5**	17	**38,2**	27	**39,6**		
8	**42**	18	**43,2**	28	**46**		
9	**47,9**	19	**48**	29	**49**		
10	**25,7**	20	**58,1**	30	**18,9**		

Sample 3. Systolic blood pressure for hypertension:

Variant №3									
№		№		№		№		№	
1	**135**	11	**148**	21	**148**	31	**148**	41	**149**
2	**145**	12	**152**	22	**153**	32	**154**	42	**155**
3	**156**	13	**157**	23	**159**	33	**190**	43	**162**
4	**165**	14	**163**	24	**163**	34	**167**	44	**190**
5	**168**	15	**168**	25	**168**	35	**169**	45	**172**
6	**172**	16	**173**	26	**173**	36	**173**		
7	**175**	17	**175**	27	**175**	37	**178**		
8	**181**	18	**183**	28	**184**	38	**187**		
9	**188**	19	**190**	29	**192**	39	**194**		
10	**199**	20	**168**	30	**204**	40	**207**		

Sample 4. Data of refractometric determination of protein concentration in blood in patients:

Variant №4							
№		№		№		№	
1	**16,5**	11	**17,7**	21	**18,3**	31	**19,8**
2	**19,8**	12	**20**	22	**21**	32	**24,3**
3	**25,5**	13	**25,7**	23	**25,3**	33	**27**
4	**27,4**	14	**27,6**	24	**28,7**	34	**28,8**
5	**29,1**	15	**25,3**	25	**27**	35	**27**
6	**32,5**	16	**32,7**	26	**33,7**	36	**34**
7	**37,5**	17	**38,2**	27	**39,6**	37	**40**
8	**42**	18	**43,2**	28	**46**	38	**47,2**
9	**47,9**	19	**48**	29	**49,5**		
10	**55,3**	20	**58,1**	30	**60,2**		

Sample 5. Data of weight measurements in 1 month old girls:

Variant №5					
№		№		№	
1	**2,35**	11	**3,48**	21	**3,49**
2	**3,55**	12	**3,68**	22	**3,68**
3	**2,69**	13	**3,72**	23	**3,79**
4	**3,55**	14	**3,91**	24	**3,93**
5	**3,95**	15	**3,97**	25	**3,97**
6	**3,98**	16	**4,01**		
7	**4,07**	17	**4,1**		
8	**4,15**	18	**4,15**		
9	**4,24**	19	**4,31**		
10	**4,37**	20	**4,72**		

Sample 6. Results of hemoglobin measurement in children aged 3 month:

Variant №6							
№		№		№		№	
1	**95**	11	**97**	21	**97**	31	**103**
2	**97**	12	**98**	22	**98**	32	**103**
3	**98**	13	**100**	23	**99**	33	**103**
4	**99**	14	**99**	24	**99**		
5	**99**	15	**100**	25	**100**		
6	**95**	16	**100**	26	**100**		
7	**100**	17	**100**	27	**100**		
8	**100**	18	**95**	28	**101**		
9	**101**	19	**101**	29	**102**		
10	**102**	20	**102**	30	**95**		

Sample 7. Data of weight measurements in 1 month old boys:

Variant №7					
№		№		№	
1	**2,55**	11	**3,42**	21	**3,43**
2	**3,65**	12	**3,68**	22	**3,68**
3	**2,69**	13	**3,72**	23	**3,79**
4	**3,55**	14	**3,91**	24	**3,93**
5	**3,95**	15	**3,95**	25	**3,98**
6	**3,98**	16	**4,01**	26	**4,02**
7	**4,07**	17	**4,1**	27	**2,12**
8	**4,15**	18	**4,15**	28	**4,23**
9	**4,24**	19	**4,31**	29	**3,93**
10	**4,37**	20	**4,72**		

CONCLUSION

This textbook can be used by students studying the course of biostatistics to fulfill practical, independent work and to prepare for exams. Basic concepts and methods of descriptive statistics are considered within the framework of this tutorial.

The topics and examples are grouped into sections and have a certain sequence for better perception and assimilation of the material. Each topic is accompanied by a brief summary of the theoretical material, specific medical problems are analyzed in detail, a calculation algorithm is given, and the use of Excel and SPSS Statistics computer programs is considered on specific examples.

Since the manual is written for students, at the end of the study of the topic, the student must find out to what extent he has mastered this material. For this purpose, the manual offers quiz for self-testing. If you receive positive answers at least 70%, the material can be considered assimilated.

Most of the methods are implemented in statistical software packages and users are freed from the need for mathematical calculations, but it is necessary to be able to correctly interpret the results. To this end, the manual provides examples of how to interpret the results.

MULTIPLE CHOICE QUESTIONS

1. An event that is influenced by chance and is not perfectly predictable is called…:

A. a likely event.
B. an impossible event.
C. a compound event .
D. a certain event.
E. a random event.

2. The addition of the probabilities for all random events must be

A. equal 1.
B. equal 0.
C. less than 1.
D. more than 0 and less than1.
E. does not exceed 1.

3. What is a probability?

A. the number of outcomes favorable to a given event divided by the total number of possible outcomes
B. a total number of outcomes in the given experience.
C. a number of occurrence all impossible events in the given experience.
D. a mass of the substance used during experiment.
E. a number of outcomes occurred simultaneously in the given experience.

4. The set of the variates which has been written down in the increasing order, is called:

A. a variation row.
B. a population.
C. a representative sample.
D. a not repeating sample.
E. a complete sample.

5. The subset randomly selected from the population so that can be made inferences about the population, is called:

A. a population.
B. a sample.
C. a random sample.
D. a repeated sample.
E. a complete sample.

6. From a sample of students in your statistics class, you collect the following data: the student's name, gender, age, and their grades. What scale is used to measure variable "gender"?

A. Nominal
B. Ordinal
C. Interval
D. Ratio
E. Numerical

7. From a sample of students in your statistics class, you collect the following data: the student's name, gender, age and their grades. What scale is used to measure variable-student's "grade" (A, B, C, D, or F)?

A. Nominal
B. Ordinal
C. Interval
D. Ratio
E. Dichotomous

8. Difference between the largest and the smallest observations in the data set is...

A. a median
B. a range
C. an average
D. a variance
E. a frequency

9. The sum of all previous frequencies accumulated up to the current point is…

A. a cumulative frequency
B. a mode
C. a probability
D. an expected value
E. a variance

10. A table which contains intervals as values and frequencies of values falling into this interval is named:

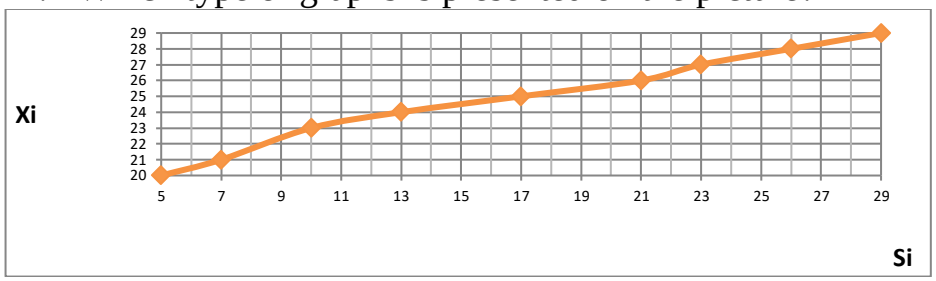
A. an interval series of distributions.
B. a variation row.
C. a random variables.
D. a ranked variation row.
E. a ratio series of distributions.

11. A value which is the middle number of the data set. It is found by numbering the values from least to greatest and finding the number that is positioned in the middle.

A. a mode.
B. a median.

C. a range.
D. a sample mean.
E. a variance.

12. Which type of graphs is presented on the picture?



A. polygon of frequency
B. ogive.
C. histogram.
D. cumulative curve.
E. polygon of relative frequency

13. What is the mode?

A. The most frequently occurring value in a set of observations.

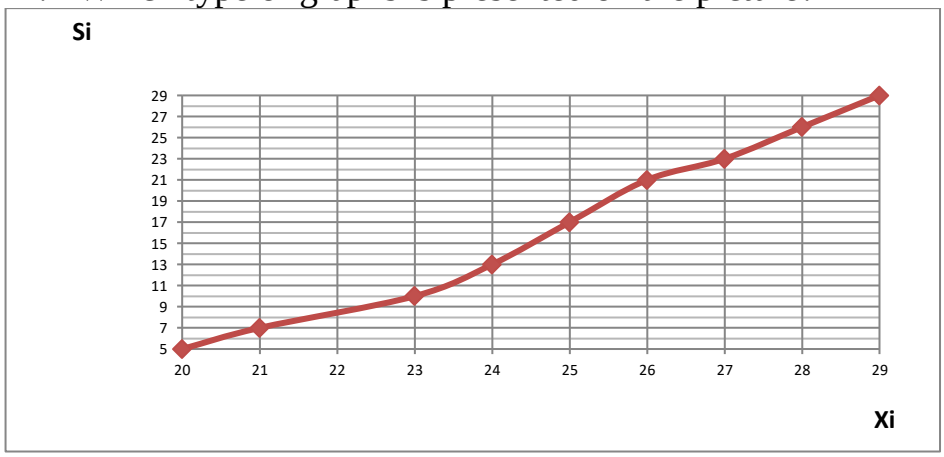
B. The value in a set of ranked observations that divides the data into two parts of equal size.

C. The average value calculated on the basis of a number set, each of them occurs only once in this data row.

D. The average value calculated on the basis of a number set, each of which meets several times in this data row.

E. The minimum value of a random variable which is defined on the basis of the analysis of all data

14. Which type of graphs is presented on the picture?



A. frequency polygon
B. ogive
C. histogram
D. cumulate

E. empirical curve

15. The coefficient of variation (CV) is defined by formula:

A. $\frac{\sigma}{\bar{X}} * 100\%$

B. $\sum_{i=1}^{n} X_i * P_i$

C. $\sqrt{D(x)}$

D. $= \frac{X\max - X\min}{1 + 3.32 \lg n}$

E. $= \frac{1}{n-1} \sum_{i=1}^{k} (X_i - \bar{x})^2$

16. An event, which has no chance of occurring is called…

A. likely event.
B. impossible event.
C. compound event .
D. certain event.
E. random event.

17. An event that is sure to happen is called…

A. a likely event.
B. an impossible event.
C. a compound event .
D. a random event.
E. a certain event.

18. Infinitely big set of all units of a specified group, is called

A. population.
B. sample.
C. random sample.
D. representative sample.
E. complete sample.

19. Variable , which is calculated by taking the differences between each number in the set and the mean, squaring the differences and dividing the sum of the squares by the number of values in the set, is …

A. a standard deviation.
B. a probability
C. a variance
D. a mean
E. a frequency

20. This is a numerical value used to indicate how widely individuals in a group vary. It is a measure of the spread of a random variable around its mean. It is called…

A. a size
B. a median
C. a mode
D. a variable
E. a variance

21. Data that can be counted. Typically it involves integers. Each value is separate and different. It is can only occur in certain values. What type of data is this?

A. Ordinal
B. Interval
C. Continuous
D. Nominal
E. Discrete data

22. According to the empirical rule, approximately what percentage of normally distributed data lies within three standard deviation of the mean?

A. about 68% of the data values
B. about 95% of the data values
C. about 99,7% of the data values
D. about 72% of the data values
E. about 78,7% of the data values

23. A Standard deviation is:

A. the square root of variance
B. the most frequently occurring value in a data set.
C. the number of times the event occurred in an experiment.
D. the number that is positioned in the middle.
E. ratio of the number of favorable cases to the total number of cases in a trial

24. A variable obtained by multiplying each possible value by its probability and summing all of those values ...

A. expected value
B. mode
C. median
D. variance
E. standard deviation

25. Gender, nationality, eye and hair color, patients' diagnosis are measured by scale:

A. nominal
B. ordinal
C. interval
D. relationship
E. rank

26. Stages of malignant disease (cancer) are recorded using symbols: 0, I, II, III, IV. What scale is used for this data?
A. Alphanumeric
B. Numerical
C. Ordinal
D. Nominal
E. Dichotomous

27. What type of Scale Example is shown below?
Where do you live?
- Suburbs
- City
- Town

A. Ordinal
B. Interval
C. Continuous
D. Dichotomous
E. Nominal

28. What type of Scale Example is shown below:
How much are you satisfied with our services?
- Very Unsatisfied – 1
- Unsatisfied – 2
- Neutral – 3
- Satisfied – 4
- Very Satisfied – 5

A. Ordinal
B. Interval
C. Continuous
D. Dichotomous
E. Nominal

29. Quantitative Discrete Variable takes exact number values and is often a result of counting. Which of the following options describes discrete variables?
A. The score on your last test
B. How long it takes to drive to Almaty
C. The distance from Astana to Almaty
D. The number of students in a university

E. The stages of disease

30. Make a ranked variation row of the following data set: 16; 18; 15; 26; 18; 85; 84; 35

A. 15; 16; 18; 18; 26; 35; 84; 85
B. 85; 84; 35; 26; 18; 16; 15
C. 15; 16; 85; 84; 35; 18; 26
D. 15; 18; 26; 35; 84; 85; 16
E. 85; 15; 16; 18; 26; 35; 84

31. Determine the mode of the given variation row:

Pulse rate, *x*	6	8	9	0	2	3	6
Number of examined people, *m*							

A. 66
B. 68
C. 69
D. 70
E. 72

32. Determine the mode of the given variation row: 16; 15; 19; 21; 15; 14; 13; 17; 18.

A. 15,5
B. 17
C. 20
D. 14
E. 15

33. Find the value of cumulative frequency for “36” from data set: 12, 23, 1, 20, 36, 20, 20.

A. 7
B. 3
C. 2
D. 1
E. 5

34. What is the frequency for variate “3” in the data set: 2,3,5,4,3,5,2,3,6,3,1,1?

A. 4
B. 3
C. 12
D. 2

E. 5

35. To find a median for given data set: 5; 6; 4; 3; 2; 9; 7
A. 5
B. 2
C. 3
D. 9
E. 6

ANSWER KEYS OF MCQ

№	Correct option	№	Correct option	№	Correct option
1	E	16	B	31	D
2	A	17	E	32	E
3	A	18	A	33	A
4	A	19	C	34	A
5	B	20	E	35	A
6	A	21	E		
7	B	22	C		
8	B	23	A		
9	A	24	A		
10	A	25	A		
11	B	26	C		
12	B	27	E		
13	A	28	A		
14	D	29	D		
15	A	30	A		

REFERENCES

1. Anders Kallen. Understanding biostatistics. Statistics in practice / Anders Kallen. - John Wiley & Sons, Ltd Publication, 2011. - 384 p.
2. Application of EXCEL and SPSS programs for statistical analysis of biomedical data / V.R Chudinovskikh, A.Sh. Kaipova., A.U. Altaeva, Zh.N. Abdikadyr - Astana: MUA, 2015. - 97 p.
3. Brian C. Cronk, How to use SPSS ® A Step-By-Step Guide to Analysis and Interpretation Tenth Edition / Brian C. Cronk. New York, NY 10017, 2018. - 179 p.
4. Chudinovskikh V.R, Abdikadyr Zh.N. The use of computer programs to test statistical hypotheses in biomedical research / V.R Chudinovskikh, Zh.N. Abdikadyr. - Astana, 2013. - 78 p.
5. Daniel, W. W. Biostatistics: A Foundation for Analysis in the Health Sciences / Daniel, W. W. - John Wiley & Sons. New York, New York. 2018. - 720 p.
6. Kandethody M. Ramachandran, Chris P. Tsokos. Mathematical statistics with applications / Kandethody M. Ramachandran, Chris P. Tsokos. - Elsevier Inc, 2009, - 824 p. - 978-0-12-374848-5
7. Koychubekov V..K. Biostatistics. / multimedia electronic textbook. / Kucherenko V.K. - Publishing house "Epigraph" - Almaty, 2016. - 60 p.
8. Kucherenko V.K. Application of statistical analysis methods. / Kucherenko V.K.. - Publishing house "GEOTAR-Media" - Moscow, 2011, - 256 p.
9. Oxford Handbook of Public Health Practice / Charles Guest, Walter Ricciardi, Ichiro Kawachi, Iain Lang. - Oxford University Press, 2013. - 1446 p.
10. Richard J. Rossi, Mathematical Statistics An Introduction to Likelihood / Richard J. Rossi. - Based Inference, John Wiley & Sons, Inc., 2018. - 423 p.
11. Rosner Bernard. Fundamentals of biostatistics.USA / Rosner Bernard Harvard University, 2017. - 927 p.
12. Thomas Glover and Kevin Mitchell. An Introduction to Biostatistics / Thomas Glover and Kevin Mitchell. Waveland Press, Inc., 2016. - 553 p. ISBN 1-4786-2779-1.
13. Veer Bala Rastogi. Biostatistics 3-d revised edition / Veer Bala Rastogi. India, Medtech, 2017. - 471 p.
14. Wayne W. Daniel. Biostatistics: A Foundation for Analysis in the Health Sciences / Wayne W. Daniel. Ninth Edition. 2009. - 956 p. - ISBN 978-0-470-10582-5.
15. William E. Using IBM SPSS Statistics for Research Methods and Social Science Statistics / William E. Fifth. - Edition, Sage Publications, Inc. 2015. - 147 p.