Techniques of Medical and Biological Statistics

Authors

Prof.Dr.
Sami Azeez Ababas Al-atbi
Biald Alrafain
University College

Assistant. Prof .Dr.
Lieth Abdulateef Majeed Al-Rubaie
Diyala University

Table of Contents

Introduction

We are pleased to introduce this book entitled "Techniques of Medical and Biological Statistics" to the students and researchers. The motivation for the publication of the book is to clarify the reality and importance of using statistics in scientific research.

Statistical methods are an essential and vital tool in scientific research. They help in the design of experiments, analysis and interpretation of data, and contribute to making appropriate decisions in the light of the researcher's results. The statistics are a way to read other research and the ability to distinguish new and stronger research. It is also a bright scientific method in which health institutions can be guided to assess and deliver the best health services to society.

Therefore, this book included seven chapters, each containing a number of items which makes it easier to teach scientific material.

In the first chapter we introduce and deal with the definition of statistics and some related concepts. The second chapter deals with data collection and the method of data collection either using the complex enumeration method or the sampling method, as well as the size and types of the sample, while the third chapter deals with the medical and biological data (classification and

presentation of medical and biological data) and its presentation in tabular and graphical terms.

Chapter 4 deals with measures of central tendency and desperation. Chapter 5 deals with simple, partial and multiple correlation analysis, as well as simple regression analysis and multiple regression. Chapter 6 deals with the statistical test of the hypotheses, the normal distribution, and the standard normal distribution as well as the scientific and non-scientific tests as well as ANOVA analysis.

Finally, Chapter 7 deals with the measure of statistical for the population and Bio statistics, and the Hospital statistics measures disease statistics.

We ask God that we have chosen the subject of this author and gave him the right to write and to be a great help to our students in their educational career.

God grants success.

Chapter One

The Concept of Statistics Science

Chapter One

The Concept of Statistics Science

1.1 The Concept of Statistics

The word "statistic" referred to the process of counting and inventory, and for that, statistic was called the science of counting. The word "statistic" was used in Europe to indicate calculations of the country in the affairs of war, taxation, population, births, deaths, production, etc.

More recently, the definition of statistics has evolved considerably. In particular in the twentieth century, it became an independent science of its own importance as a result of being a way method and mechanism used in all scientific specializations.

Generally, statistics is defined as the integrated science of mathematical procedures that is used to collect, organize, analyze, summarize and present data. Statistics can be applied in many disciplines, including economic, social, political and military fields, to explain data, and to draw conclusions from it, in order to make appropriate decisions, to formulate policy and to plan for future scientific research. Specifically, Medical Statistics and Biostatistics are the sciences that apply various statistical theories and techniques to medical and biological research.

Statistics is divided into two branches, which are:

1. Descriptive Statistics

It represents digital or computational statistical methods for data collection followed by classification, organization, scheduling and representation of the data graphically - whether from the sample or from the statistical population and then the results are drawn through some statistical measures and methods to make the appropriate decisions.

2. Inferential Statistics

It is a technique that allows the study of specific samples from the statistical population to draw conclusions or make inferences that can be attributed to the entire population. These samples must be representative of the population they are chosen from. The theory of probability is considered to be a key element in Statistical Inference, which deals with two main topics: Estimate and Test of Hypotheses.

From above, it is possible to say that the application of statistics includes the following:

1. Collecting data, information and facts regarding various phenomena, recording this data in digital form, classifying it into organized tables and representing it graphically.
2. To compare the phenomena to each other, study the relationship between them, and to use that knowledge to understand the reality of the phenomena and the laws that govern them.

3. Analyze data and draw conclusions that can then be used to conduct certain processes and make decisions to achieve specific programs or goals under uncertain circumstances.

1.2 Historical development of Statistics

Statistics is an ancient idea that dates far back into human history because the need has existed to obtain digital or descriptive information about civil societies since the existence of organized human societies. In addition, there are examples of statistical uses in ancient civilizations – Chinese, Indian, Babylonian, Egyptian and Assyrian- which were limited to collecting numerical information on the population (primarily men), cattle and agriculture. This data was recorded in special books and papers to which the state referred to as needed to conduct policy.

Moreover, the impact Arabs have had in statistics should be mentioned; the Arab thinker Ibn-Khaldun may be the first to have addressed population issues scientifically. He studied the construction of the state, its extension and its delay, and linked these to the growth of the population.

In the sixteenth century, because of the spread of gambling in Europe, gamblers sought mathematicians to obtain information about their chances of winning or losing. Among the most famous mathematicians are Pascal, Leibnitz, and Bernoulli, who is the founder of probability theory.

In the seventeenth century, between 1609-1681, the German scientist H. Conring had started teaching a new science called Staattkunde. Also, from 1620-1674, Englishman John Grant was the first to apply statistical methods to the field of infectious disease and microbiological studies. In 1719-1773, the German scientist G. Achenwall launched a new designation, that of statistics, for this science; he also published his book on the principles of statistics in European countries in 1749.

In 1666, the English scientist Gronix published a book in which he studied the population records of London, and from this calculated the death rate. In 1683, W. Petty wrote a book in which the quantitative methods were used and called it *Political Arithmetic*.

In 1812, statistics entered a new phase of its development when Laplace, in his book, *Analytic Theory of Probability*, explained the benefits and advantages that can be derived from the study of natural phenomena, and stated that their causes are so complicated that they cannot all be known. In addition, Laplace contributed to the consolidation of the concept of the general application of statistical methods, and proved that the theory of probability is a necessary tool to improve all kinds of human knowledge.

In the period 1796-1874, the Belgian astronomer A. Quetelet was drawn to statistics, especially the subject of probabilities,

which inspired him to demand to introduce an improvement to the process of census. Moreover, in 1840, he pointed to the use of natural distribution not only as a law of errors, but also as an explanation of the distribution of measurements. In addition, because of his initiative, the International Conference of Statistics was held in Brussels in 1853, which led to the "International Statistical Institute" that was founded in London in 1885.

In the modern era, the methods of statistical analysis have expanded from the end of the nineteenth century to today, such that they have reached all of the scientific fields. As a result, this led to the rapid and expansive development of statistical theory. In particular, in the period 1822-1911 the scientist F. Galton established a new branch of statistics called Biostatistics, which has now extended to the fields of therapeutics testing and therapeutics medicine.

The link between statistics and the economy has deepened, resulting in the emergence of a new science called Econometrics; some call it the economic statistic. Some of its pioneers were V. Pareto, A.A. Cournot, F. Divisia, and L. Walras,

who received a Nobel laureate in Economics Sciences in 1969.

The most renowned scientist of the twentieth century was the scientist R. A. Fisher in the period 1890-1962. During that period, his first contribution was his derivation of the exact distribution of

the correlation coefficient. In 1923 he developed the method of analysis of variance (ANOVA), and also presented his theory of maximum likelihood estimation. Thus, through Fisher's contributions, the application of statistics was extended to agriculture, economy, heredity, and so on.

Moreover, in the period 1891-1967, quality control methods that were developed through the contributions of W. Shewhart became one of the major statistical applications in the industry.

In 1944, the scientists John von Neumann and Oskar Morganstern shared in the publishing of a book called *Theory of Games and Economic Behavior*. It can be said that the works of J. von Neumann, E. Pearson and R. A. Fisher, and the experimental research that has emerged, have made statistics a powerful and effective tool in scientific and technical research and whose field of use continues to grow and develop.

In 1980 and beyond, the emergence of computers began to impact the progress of statistical work. However, the computers that were used in the end of the nineteenth century, which helped statisticians such as Pearson and Fisher in their research and in building their statistical tables, were not commonly used, as many statisticians still preferred to use logarithmic tables and Slide Rule instead.

With the use of computers, the completion of old operations required less time, and the development of new, more complicated operations became possible with many programs. SPSS in

particular is one of the most widely used statistical programs in various specializations such as medical, engineering and agricultural.

1.3 Statistics' relationship to other sciences

The field of statistics is the science of mutual relationships with other sciences. It affects and is affected by the domain of its continuous development through modern technological and scientific progress. Therefore, there is the possibility of applying its theories, principles and methods in all fields, as the various phenomena that can be expressed by data can be compiled, categorized, tabulated and graphically represented. Then, the results derived from the application of statistical methods can be used to make appropriate decisions.

To understand the relationship between statistics and other sciences, we will present some of these relationships:

1. The relation of statistics with pure mathematics:

The relationship between statistics and pure mathematics is very strong. Many statistical theories depend on mathematical methods and their development, and use of alternative methods of presentation and proof. Many of the probability distributions are formulated as mathematical functions with many variables; the mathematical processes of these functions use the theories of differentiation and integration, which can obtain the most accurate indicators and statistical measurements for crisis analysis and study.

2. The relationship of statistics with economics:

It is very difficult to separate statistics from economics - any economic study aimed at planning, estimating or predicting either private business or the national economy requires the availability of data and information on all the variables specified, which can be obtained by using statistical methodology. Hence, the science of statistics is an essential and necessary part of economics and its development, because any economic study depends on statistical methodology in its implementation. Also, statistical indicators and measurements have become a necessary tool in refining economic work whether it relates to supply and demand, production and consumption, market study, prices, wages, saving and investment, or any variable of the national economy in general, whether it is for planning or for the work of comparisons. Therefore, it can be said that economics is a field created and developed by statistics.

3. The relationship between statistics and the natural sciences:

Most of the natural science studies depend on the statistical method in the design and implementation of experiments. The theory of probability sampling also plays a prominent role in these fields, whether in chemistry, physics, agriculture, geography or others.

4. The relationship between statistics and medical sciences:

The science of statistics is applied in most medical fields, especially in the topics of disease comparison, treatment, and cause of disease, and measuring the efficiency of drugs and treatments used for many diseases. The theory of probability sampling also plays a large role in this area.

5. The relationship between statistics and demography:

Statistics help to study population development by studying the rates of births, deaths and migration. In addition, statistical methods are used to study the economic, social, occupational and educational parameters of the society.

6. The relationship of statistics to social sciences:

There have been great technological developments in all fields, including humanities, as well as the accompanying development of new ways to address social and psychological issues; the role of statistics in this progress cannot be denied. The statistical method, central limit theorem, correlation, regression, tests, etc. and basic applications are important in this area.

It is not strange to say that every researcher specializing in the field of social sciences must be familiar with statistical methods if he wants to raise his research and studies to the level of the spirit of the times.

Finally, it can be said that statistics is a science based on scientific methods, laws and multiple theories, which is the basis

for many other sciences as well, and which can be considered a tool for collecting data, and mathematical models to describe and interpret data.

1.4 The importance of statistics and its application areas:

Statistics is the scientific method necessary to investigate the facts of a phenomenon and draw conclusions from those facts. It also includes the theories of measurement and decision-making necessary to all economic, social and political fields, thus, it is one of the most important and accurate scientific research tools available. The importance of statistics is clarified by the following:

1. Its methods, laws and theories are essential tools in scientific research, where in it is possible to achieve the desired objectives through the accomplishment of such research.
2. Statistics play a prominent role in the planning and policy development of every country. By providing accurate data and statistical information, objectives can be met and this can enable workers to plan and follow-up on processes to achieve all stages of planning.
3. Statistics can provide analysis and prediction, a simplified digital or planning tool, and is internationally certified.

The following are the most important areas of statistical applications in which raw data can be processed, arranged, classified and judged:

1. Economic, administrative and applied research.
2. Biological and genetic research.
3. Applied Medical and Pharmaceutical Research
4. Environmental health research and health care policy management.
5. Applied engineering research.
6. Applied agricultural research.
7. Applied industrial research.
8. Sports and youth research.
9. Psychological and social research.
10. Geographical research.

1.5 The Scientific Method:

It is a set of techniques and methods designed to examine phenomena and existing knowledge, or to correct and modify information or old theories, and is summarized in the steps following:

1. The first step in the scientific method is to identify the research problem. Identify the subject of the problem and its basic variables and then formulate them in the format of a specific question (or specific questions).
2. Review previous research and studies relevant to the problem to be considered.

3. The formulation of hypotheses. Hypotheses are a set of initial solutions proposed by the researcher to the problem and are usually based on three cases:

a. Identify differences (such as, “There are differences between the ages of males and females”).

b. Determine the relationship (such as, “Is there a relationship between age and disease type?”).

c. Description (such as, “What are the migrants' attitudes towards returning to their home country?”).

1. Research design. The researcher must develop a research design that aims to answer the research problem or prove its hypotheses.
2. Testing of hypotheses. The researcher collects data for the benefit of the hypotheses. If the data supports the hypothesis, then the hypothesis may be correct, and if the data does not support the hypothesis, the researcher rejects it.
3. Data analysis. At this stage the researcher applies statistical methods appropriate to the nature of the data and according to the requirements of the hypotheses.
4. Interpretation of results. The researcher discusses the results of his research and interpretations. He can apply generalizations from these results to other, similar phenomena, and if these generalizations are validated, then these results may become a

scientific rule or law that can be used to predict the occurrence of the phenomenon in the future.

1.6 The Statistical Method:

It is a scientific method of dealing with special digital trends for some scientific and social phenomena. The main stages of the statistical method are:

1. Collection of digital data on phenomena or phenomena related to research.
2. Data tabulation and graphical representation.
3. Analysis of data and calculation of statistical indicators as estimates of the characteristics of the research community.
4. Analysis of research data and determination of results.
5. Interpretation and prediction of research results.

1.7 Research Design

After identifying the research problem, its matrix and its hypotheses, the researcher designs the research work plan in order to limit the appropriate curricula, identify the tools and techniques required, specify the requirements of the physical research and schedule the research process properly. It can be said that one of the design requirements is also that the researcher has strategic and tactical plans to help achieve the objectives of his research.

The strategic plan helps to identify the major stages of research, the main features of the information and data required at each stage, the type of tools needed to collect that data, and the

type of quantitative analysis to be undertaken. Tactical plans are created to address practical concerns that arise when collecting data and to act accordingly in these situations, whether they are expected or unexpected.

There are things to consider when designing the work plan, including:

1. **Identify research objectives:**

The research objectives are divided into scientific and practical objectives. The scientific objective is to add or modify theories, laws or scientific rules. The practical objective is to address the research problem through analysis of research data, determination of both the cause and its effects, and to provide the solutions and recommendations necessary to address them.

2. **Determine the actual implementation of the research requirements:**

The implementation of the research plan requires enumeration of physical factors as well as personnel - such as specialized doctors, pharmacists, technicians, analysts, workers, etc. - in such a way that ensures that the search is performed smoothly. Otherwise, the researcher may not be able to complete his research properly.

3. **Select the framework for research:**

It is essential that the researcher identify the type and nature of the research field or the statistical community that contains all

of the variables to be studied, as well as to determine the whereabouts of these variables and identify ways for the researcher to access the data required. The framework may consist of data in multiple formats such as a map of all the sites to be studied (hospitals, specialized health centers), or a group of lists containing the names and addresses of patients.

4. Determine the data collection method:

The technique used to collect the information needed depends both on the nature of the data, and the type of information that you intend to obtain from the data: distributions, relationships, associations, or differences between or within data sets. In general, there are two methods used to collect data: the comprehensive inventory method or the sampling method.

Chapter Two

Data collection

2. Data Collection

2.1 Sources of statistical data

2.2 The sources of medical data

2.3 Methods of data collection

2.3.1 Complete enumeration method

2.3.2 Enumeration partial method

2.3.3 Sampling Method

2.4 Methods of ensuring sample representation of the original population

Chapter Two

Data collection

2. Data Collection

The process of collecting data is considered one of the important stages on which statistical research is based on. Scientifically, collecting this data would inevitably lead to more accurate results in the analysis. To study data collection methods, it requires full knowledge of the following aspects:

First: Sources of statistical data

Second: Method of data collection

2.1 Sources of statistical data

It is not possible to complete any scientific research without sufficient data from the research subject and reliable sources so that the researcher can conduct statistical analysis of the data later. This data can be obtained from two primary sources:

1. Data and official statistics

This source of data can be found in published of the official and semi-official bodies and institutions of economic, social, commercial, agricultural, industrial and health data, as well as published by the competent government agencies, the Central Organization for Statistics and Information Technology, Health Institutions.As an example for these data, population, classification and distribution of the population according to geographical location, age, occupation, sex, scientific level, as

well as the income, births, marriages, divorce, etc. All data collected according to this method are called secondary data and their advantages save time, effort and financial costs.

2. Data and field statistics

Field researchers depend primarily on the primary sources of data in the natural and actual data sources, which the researcher collects the data by himself directly by an interview or indirectly using questionnaire given to the selected sample members. The data collected in this method is called the primary data which is characterized by accuracy and reliability, but it is difficult in that it needs time, effort and material cost.

2.2 The sources of medical data

As mentioned above, the sources of statistical data are primary sources and secondary sources in the same way. All the cases related to the sources of medical and health data also divided into two parts:

1. **Primary sources:** which are the sources through which the process of collecting data and information is controlled by the personal researcher. In accordance with the research requirements and objectives, the medical and health data collected from hospitals and health institutions through the direct viewer and experience or the interview or questionnaire.
2. **Secondary sources**: which are the sources that collect data and medical information which published in the form of books,

dictionaries, publications, scientific journals, document reports, medical atlases, letters, university papers, etc. These sources may be official, or semi-official local and international, among the most important which are the following:

1. Ministry of Health (Health Statistics Directorate).
2. Central Organization for Statistics and Information Technology.
3. The City of Medicine Foundation.
4. Health directorates in the provinces.
5. International organizations (World Health Organization, United Nations Children's Organization, Doctors Without Borders, Child Save Fund, International Federation of Red Cross and Red Crescent Societies)

2.3 Methods of data collection

The process of data collection requires determining the appropriate method for determining the method. This process is not easy and is a real problem that the researcher always has. At anyrate, some criteria must be taken to a choice of the appropriate method:

1. The degree of required accuracy: Sometimes the comprehensive inventory method used when we want accurate and comprehensive data, such as the case of research that relate to the lives of individuals and their health.

2. The homogeneity of the statistical units: The higher degree of homogeneity, especially in large societies, the sampling method is preferable.
3. The period allocated to the research: The more extended period used, the longer method of comprehensive or partial inventory used, in the opposite direction, using the sampling method in the case of short-term.
4. The availability of financial and human resources: When these resources are available in a sufficient way, then we can use the inventory method, but when it is not possible then we adopt the method of the sample. In general, in addition to the previous criteria, the used method of collecting depends on the nature of the objective of the research on the hand, and the size and the statistical population from the other side, in general. There are three stereotypes:

1. The method of comprehensive enumeration.
2. Partial inventory method.
3. Sampling method.

2.3.1 The method of comprehensive enumeration

It is a method of collecting data from all units of the statistical population. (The health institutions) without exception. The researcher seeks this method when he wishes to obtain detailed data on all the units of the population and when the

researcher is ignorant of the nature of the population, especially if it is not carried out by previous studies.

This method has the accuracy of results, inclusiveness, and impartiality, but the requirement of the physical, human and temporal needs may change the possibility of its use.

2.3.2 Partial inventory method

This method is used in many areas, especially to count and control the big population of the institutions and health centers, which have a large number. Their contribution in the provision of services is small, (such as health centers spread in residential neighborhoods, villages and rural areas).

According to this method, the statistical units are divided into units in which the studied phenomenon is concentrated, and then they are limited to a comprehensive (often a small number), whereas the remaining units are very small for their small contribution in the phenomenon, despite the enormous number of them, so we dispense with them and exclude them from the research by using the latest estimation methods.

2.3.3 Sampling method

It may be challenging to follow the method of comprehensive enumeration in the study of many phenomena, so the researcher can choose only small units of the statistical analysis population by analyzing and extracting their results and disseminating them to the original population.

This method is less accurate than the method of inventory, unless it is used in a best possible way, using an efficient scientific and accurate basis, as well as the cost of physical use, is much less than the comprehensive inventory method.

The success of using this technique depends on many factors:

1. Select the sample frame.
2. Specify the sample size.
3. Determining the type of sample size which will be used.

2.3.3.1 Select the sample frame

It is necessary before the start of the selection of the sample units by putting a plan of work containing all the units to be chosen from and determine the location of these units to determine the means to access them in the collection of data and the information required about it.

The frame may have multiple forms. It may be in the form of a map that includes all the sites to be studied (hospitals, specialized health centers), a list, a group of lists as many numbers of people, patients, hospitals, or a set of medical records. The researcher must verify the accuracy of the sample frame and its validity before choosing it, so that on the one hand it will represent the population in a good way and truthy manner, and on the other hand, it will match the goal of research and methodology.

2.3.3.2 Determine the size of the sample

After the modernization of the general framework of the population from which the sample will be selected, the researcher must be ready to select its units. But before the process of selection, he needs to determine the optimal size of the required sample ;choosing the right size is a critical factor, because if it is too small then it will not present the studied population. In contrast, if it is too big it will cost the researcher a lot of money, time and effort. In general there is no specific size or percentage of the size of the statistical population which can be adopted in all studies and research. There are indicators that can be guided to determine the size of sample as follows :

1. If the population under study is homogeneous, the size of the sample may be small, but if it is not homogeneous, the size of the sample should be large.
2. The degree of precision required for the research, where the size of the sample should be big whenever the degree of accuracy required high and vice versa.
3. The availability of material and human resources, as the researcher will be able to choose the size of a large sample and vice versa.
4. The nature of the research methodology that will follow is directly related to the size of the sample. The empirical approach, for example, requires a large size of the sample.

In general, random samples can be used to determine the sample size

$$\boldsymbol{n = \frac{t^2}{r^2 + \frac{1}{N}t^2}}$$

$\boldsymbol{n}$ = sample size required.

$\boldsymbol{t}$ = the values of the scheduler that correspond to the permissible error.

$\boldsymbol{r}$ = probability of error

$\boldsymbol{N}$ = number of statistical population units

Example

Find the size of the sample which will be used in the study of the factors affecting the fertility of the population in a province, the number of families is 249,361. If we want the error amount to be 0.04 of the standard deviation and the probability of error 0.05.

The solution

$$n = \frac{t^2}{r^2 + \frac{1}{N}t^2}$$

$$n = \frac{(1.96)^2}{(0.05)^2 + \frac{1}{24361} + (1.96)^2}$$

$n = \frac{3.841}{0.0025154} = 1526$ The Number of sample units to be withdrawn from the population.

Also, the sample can be obtained in fewer units when required by using Steven K. Thompson's equation as seen in the following formula:

$$\boldsymbol{n = \frac{N \times p(1-p)}{[N-1 \times (d^2 \div Z^2) + p(1-p)}}$$

When

n: Sample size

N: Population size

Z: Confidence level at 95 % = 1.96

d: Error proportion = 0.05

p: Probability =50%

Example: If the size of the population is (10000) after the application of the equation, the size of the sample required

$$n = \frac{10000 \times 0.50(1-0.50)}{[10000-1 \times (0.05^2 \div 1.96^2) + 0.50(1-0.50)}$$

$$n = \frac{10000 \times 0.25}{[9999 \times (0.00065) + 0.25}$$

$$n = \frac{2500}{6.74935} = 370$$

Number of samples to be taken from the population.

We can also use Robert Mason's equation to obtain the sample from the next formula:

$$\boldsymbol{n = \frac{N}{[(S^2 \times (N-1)) \div pq] + 1}}$$

When

n: Sample size

N: Population size

S: The ratio between the error on the standard score, i.e., the division of $\frac{0.05}{1.96}$

P: Availability ratio of the property, which is 0.50.

q: Remainder ratio of the property, which is 0.50.

Example:

If the size of the population is (12000) after the application of the equation, the size of the sample will be:

$$n = \frac{12000}{\left[\left(\frac{0.05}{(1.96)^2} \times 12000 - 1\right) \div 050 \times 0.50\right] + 1}$$

$$n = \frac{12000}{[(0.0255) \times (11999) \div 0.25] + 1}$$

$$n = \frac{12000}{[(0.00065) \times (11999)}$$

$$n = \frac{12000}{[(7.79935) \div 0.25] + 1}$$

$$n = \frac{12000}{32{,}1974}$$

$=373$ sample size.

2.3.3.3 Kinds of Samples

Before we review the different kinds of samples, it is necessary to determine the difference between the population and the sample taken from it.

Population

Vocabulary or units that share single or several attributes may be parts or objects or measurements.

The statistical population may be as specific **Finite population** as the number of doctors according to their specialties, the medical assistants according to their specialties, the number of hospitals, the capacity of each of them, etc. Population may be undefined **Infinite population** which is difficult and impossible to count the number of units, like the disease in a hospital, the number of insects carrying a disease, etc.

Sample

The sample in a small group of statistical population units is collected in different ways so that they represent the population if an honest way through the characteristics of that sample.

Figure (1) shows the difference between the population and the sample was taken from it.

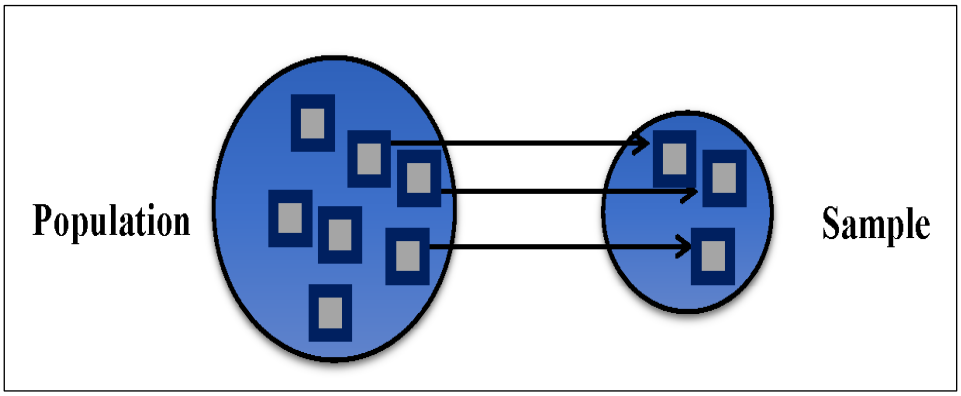


Figure (1) the difference between the population and the sample.

In general, the samples are divided according to the method of their choice of two types:

First: Potential Samples.

Second: Non-potential samples.

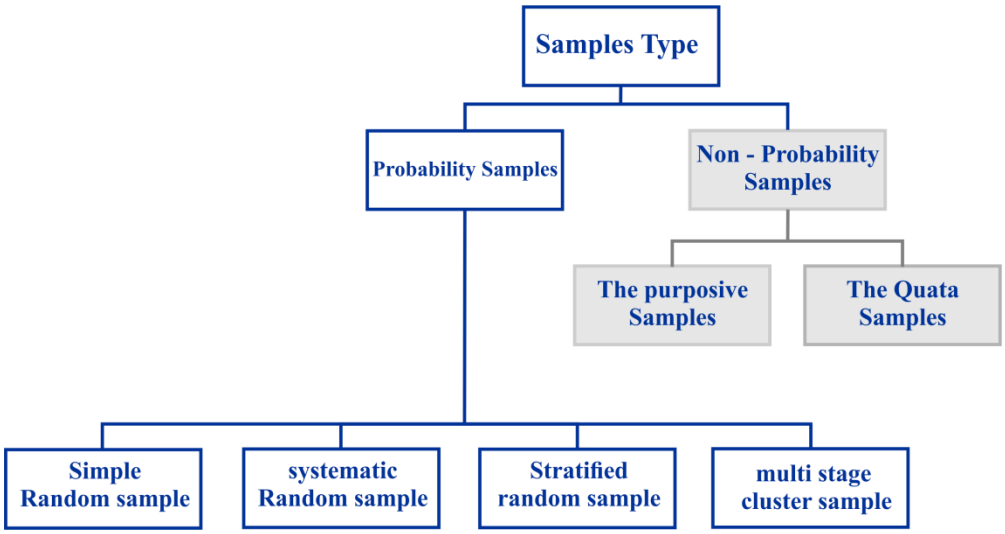


Figure (2) Samples types

2.3.3.3.1 Probability samples

They are samples whose selection depends on the probability theory. Each unit of the statistical population has the opportunity to appear in the selected sample equally in the case of simple random samples, or unequal in the case of class and cluster probability, etc. The most important types are:

1. Simple random sample

The units of this sample are chosen randomly from the effect of personal factors, so that all the units of inspection in the population have the same opportunity to appear, i.e., the

probability of selecting any unit is 1/N, and the sample can be chosen by returning the checked unit before drawing the sample,

In general, this sample is selected in one of two ways:

1. Lottery style: Under this method the names are written of patients or all units and numbers on small cards and then mixed with each other. Then we randomly withdraw some units that will represent the desired sample size.
2. The method of using tables of random numbers: Usually random selection cannot be made by lot, especially in the case of large statistical communities, so some of the statisticians have prepared special tables to make the selection process of random selection easier, they called the tables of random numbers.

To select a sample from a population by using the tables of random numbers, the following steps are required:

1. Give serial numbers to the original population members, starting from the one.
2. Adoption of the highest number will be given to the units of the original population, we can determine the value of random numbers if it is one or two or three or four digits and so on, For example, if the number of population unit is 99 patients, we will need a random number of two digits , if the number of members of the population 258 patients, we will need a random number of three digits and so on , so for example if we

have random numbers of two digits, then we combine each two columns in the table to create these numbers, And so we do the same if the numbers of digits increases.

3. After we have determined the number of random numbers that we need , we close our eyes and put our fingers on the point in the table and start moving from there geometrically fixed either up or down or to the right or left or diagonally and continue this process until we get the number required from the figures in the table, which are all within the figures given to the units of population

For example, if the number of units of population is (500) and we choose a sample of (50) units, then we take a random number of three digits, We set the starting point on the table heading towards the top or bottom, for example, and take the first 30 numbers that come to us which all less than (500). If it happens that we have reached the end of the table before we can select the required number for the sample, we will return this proses once again by using a new starting point.

2. Systematic random sample

Under this type of sampling, the original population is divided into equal groups. If the population, for example, consists of 1000 units and the wanted sample was 100 units, then the population is divided into 100/1000 or 10 units or equal categories groups. The first unit will be selected randomly from among the

units of the first group. We will assume the unit number is 50. Thus, the next units to be drawn within the sample are (50, 150, 250, 350... 950). The problem with this method is that choosing the sample is not pure random choosing of a random sample.

3. Stratified Random sample

Using this type of sample provides access to more accurate estimates than the estimates of the two previous methods. This type of sampling is used when the research population is composed of homogeneous groups or layers with specific characteristics or groups that are related to the phenomenon under study. A random sample shall be taken from each group in a way that considered the size which proportion with of size the original population. The adoption of this type of sampling is done under the following steps:

1. Select the layers or groups first, and then put each unit of a preview of the population in the appropriate class.
2. Determine the size of each class or group, then determine the size of the sample that will be drawn from each layer randomly.

If we assume that we are studying the reviews of patients who are asleep in five hospitals in Baghdad about the quality of health services provided to them by using a sample layer, for the population of 1300 beds and the required sample size is 10 %, and as follow:

a. The number of beds in the first hospital is 200
- The number of beds in the Second hospital is 240
- The number of bed in the third hospital is 300
- The number of beds in the fourth hospital is 350
- The number of beds in the fifth hospital is 210

b. Therefore, the size of the study population is $200 + 240 + 300 + 350 + 210 = 1300$

c. Determine the total size of the sample = $1300 * \frac{100}{10}$ patients

d. Determine the total sample size of each hospital as follows:
- The sample size of the first hospital is $200 * 10/100 = 20$
- The sample size of the second hospital is $240 * 10/100 = 24$
- The sample size of the third hospital is $300 * 10/100 = 30$
- The sample size of the fourth hospital is $350 * 10/100 = 35$
- The sample size of the fifth hospital is $210 * 10/100 = 21$

4. Multi-Stage Random Sample

When the researcher depends on this type of samples in his study, he must in the first stage, divide the original population into several groups and then select a random sample of them. In the second phase, he select a random sample from the selected sample n the first stage. So, until the researcher obtain the size of the sample based on the objective of the study on one hand and the size of the original population on the other.

This type of sample provides the researcher with sufficient flexibility to choose the required sample in more than one way, as well as the availability of a lot of time and effort and cost. It does not necessarily represent the population of the original study, especially in the inhomogeneity of the original study population

In general, in cluster sample, different groups or areas, should be chosen, such as health centers, pharmacies, laboratories, medical clinics, etc. These groups have the property that should behave in the same manner for all its units.

2.3.3.3.2 Non-probability Samples

In these types of samples, the units are selected according to the researchers. Personal judgment, which can not be isolated or measured, i.e., the original population units of the study do not give the same opportunity to appear in the selected sample, from these types which are:

1. Purposive sample

These are the samples whose units are deliberately selected from the researcher without others based on the characteristics and advantages that are consistent with the aim of the study and research.

For example, if a researcher wants to study the opinions of the voters, he will choose a sample of individuals who have at least some information about that person.

2. Quota Sample

According to this type of sample, the study population is divided into groups, categories or sectors within a given criterion. The researcher then selects the number of required units for each group or segment according to the conditions of the researcher

This type of sampling is of significant importance in public research because of its rapid implementation and low costs.

For example, if the researcher wanted to study the opinions of people about the high prices of doctors wages in Iraq, and he chooses a sample of 2000 people, the researcher will determine the size of the population of that province and then select the sample members from each province in a random way, by going to public places and interview people according to the specific number of each province, which constitute the total sample .

After reviewing the types of samples, it is necessary to note that if the method interview has been used, then the sample must be chosen carefully to represent the original population, knowing that many of the preview operations may resort to more than one type of data.

It is necessary for the researcher to know the locations of error in the selection of the sample, and the most important locations are:

1. Errors of prejudice: which are the result because of the way we choose the sample from the original population.

2. Errors caused by the size of the sample and called errors of stochastic.
3. Errors resulting from reactions of people to the measurement tools itself and called errors of the tool.

2.4 Ways to ensure the representation of the sample of the original population.

The selection of the sample correctly and accurately leads to the achievement of results almost very close to the actual reality; we can check the extent of representation of the sample to the original population through the following two methods:

1. Normal distribution method

Many quantitative data, if found in large numbers in a single set, such as data of height, weight, intelligence, age, blood pressure and other quantitative data taken from a particular population in a random way, if drawn in the form of a repetitive curve, we will notice that some of these data are larger than average, and others are less than average, while the minority is distributed on both sides. The shape of the resulting curve is shown in Fig. 3. This distribution is called natural distribution or Gaussian distribution, where the values of the mean, the median and the tendon in one point and all the division of this distribution into two halves completely identical, the characteristics of this curve are:

68% of the area under the curve falls within the range (±1) of standard deviation from the arithmetic mean.

95% of the area under the curve falls within the range (±2) of standard deviation from the arithmetic mean.

99% of the area under the curve falls within the range (±3) of standard deviation from the arithmetic mean.

To obtain an area below the curve of 95%, it is between the range (mean 1.96 standard deviation). However, If we want to get an area below the curve of 99%, then it is between the range (mean 2.58% standard deviation).

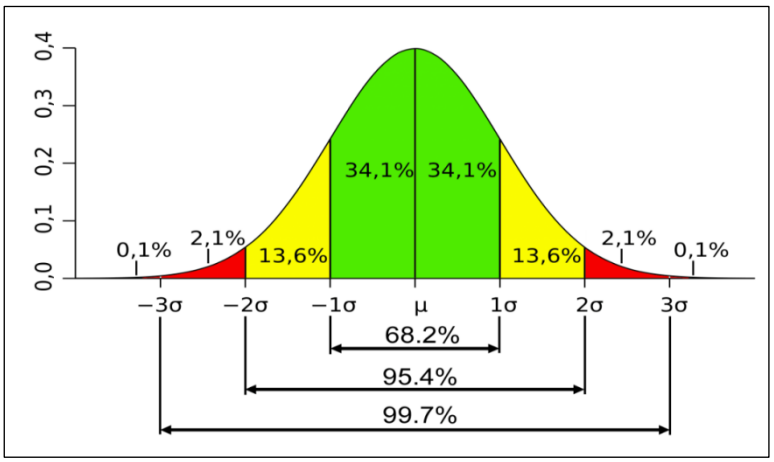


Figure (3) Normal distribution curve

To ensure the sample which is represented by the original population, follow the normal distribution method by determining the distribution of the selected sample. So if it is normally distributed then this indicates that the sample is representative of the original population. But if the distribution is not normal, it means that there is a bias in the selection of the sample, so the

sample will be unrepresentative of the original population under study and research.

2. Central Tendency Method

This method is used instead of the previous one when the original study population does not take the form of natural distribution initially. For example, the health situation for most people in the third world is low and therefore do not take the form of natural distribution. So in such cases usually the standards of centralism and dispersion are using, Such as the arithmetic mean and the standard deviation. Once the values are found by the selected sample, they are compared with the arithmetic mean and the standard deviation of the original population (this is usually published in the relevant scientific journals). If the results are close, then the selected sample is a representative of the original population. If there are significant differences, this indicates that there is a bias in the selected sample. Therefore, the results of this sample can't be adopted as a representative of the original study population.

Chapter Three
Classification and Presentation of Medical and Biological data

3. Classification and Presentation of Medical and Biological data
3.1 Statistical data types
3.1.1 Qualitative data
3.1.2 Quantitative data
3.1.2.1 Classification of quantitative data to Discrete data and Continuous data
3.1.2.2 Classification of quantitative data to temporal data and spatial data
3.1.2.3 Classification of quantitative data into grouped and ungrouped data
3.1.2.4 Classification of data into absolute and relative data
3.2 Presentation of statistical data
3.2.1 Data table presentation
3.2.1.1 Simple Frequency tables
3.2.1.1 Double frequency tables
3.2.1.3 Accumulative Frequency tables
3.2.1.4 Research Data matrix
3.2.2 Graphical presentation
3.2.2.1 Bar Charts
3.3.2.2 Histogram
3.3.2.3 Frequency polygon
3.3.2.4 Cumulative Frequency Polygon
3.3.2.5 The Pie Chart

Chapter Three

Classification and Presentation of Medical and Biological data

3. Classification and Presentation of Medical and Biological data

It has accompanied the development of the revolution in the quantity of medical and biological fields, especially in the past 20 years, has been accompanied by another parallel revolution of no less importance known as "data explosion" or information explosion. It is known that the information revolution is not limited to medical and health sciences, it is a general revolution in all kinds of science. Therefore, the plenty, diversity of medical, and health data have led to increased need for classification and presentation in a briefly manner that helps to understand and analyze them statistically, to identify and describe them and compare them with other data and come out with some statistical implications for the research and study community.

3.1 Statistical data type

Data is generally a set of facts or unorganized observations of those that can be recorded, measured and collected from the statistical sample or community. In addition, medical and health data have no advantages over other scientific data. So medical and health data like any other statistics, are divided into two basic types statistical data types are:

3.1.1 Qualitative data

This type of data is not quantifiable "digital" but can classify or arrange data in the form of categories or groups that can be measured by two main criteria:

1. Descriptive data measured by a nominal standard "Nominal Scales" like:

- Gender (male, Female).
- The social situation (Single Married, Separated, Widowed).
- The health situation (Excellent, good, bad).
- The education (Excellent, good, bad).
- The profession (Doctor, pharmacist, Analyst, employee).

2. Descriptive data measured by standard ordinal (Ordinal scales) like:

- The classification of illness (mild, moderate and sever)
- Educational level (Illiterate, elementary, secondary, Academic and postgraduate).
- The efficiency of health care professionals estimates (weak, accepted, average, good, very good and excellent).

3.1.2 Quantitative data

Quantitative data is information about quantities; that is information which can be measured and written with numeric values. For example of quantitative data are weight of a subject, Data of births, deaths and population, income, age, height, or weight… etc.

The quantitative data can be classified according to the following classifications:

3.1.2.1 Classification of quantitative data to Discrete data and Continuous data:-

1. Discrete data:

Discrete data is based on counts. It is data that does not enable the nature of the fragmentation of the unit of measurement in which it is measured, but rather moves at certain numbers. For example the number of students in a class, the number of patients, and the number of hospital, etc.

2. Continuous data:

A set of observations usually associated with physical measurement. Data can be indivisible units of measurement. For example length, weight, etc.

3.1.2.2 Classification of quantitative data to temporal data and spatial data

1. Temporal data:

Temporal data includes all data that are regulated within a specific period time, such as the number of patients in the summer, the quantities of medicines disbursed in July, and the number of patients hospitalized in the period 2000-2014.

From the above, it is clear that these data are closely related to the element of time and may be days, months, seasons, and years, etc.

2. Spatial data:

Spatial data includes all data that is regular within a specific spatial context; it can only be explained or understood through that framework. For example, the annual rate of the medicines production in the country, the annual rate of births and deaths etc.

3.2.1.3 Classification of quantitative data into grouped and ungrouped data:-

1. Ungrouped data:

Ungrouped data is preliminary data (raw data) collected directly from the statistical community or sample. They are of little use because they are raw data. It can be family budget forms or hospital admission forms and forms of births and deaths before they can classified into groups or groups can be a model for such data.

2. Grouped data:

Grouped data collected by a complete census, partial or sample inventory cannot be directly used to measure statistical measures. Therefore, it must treated by dividing these data into similar groups called classes. In each class, we assign the units to which they belong to, which are called Frequency. The output distribution from the download primary data source is called Frequency distribution and the table that includes this distribution is called Frequency table.

3.1.2.4 Classification of data into absolute and relative data

1. Absolute data:

Absolute data includes various types of quantitative data for which there is no absolute zero. For example, when the temperature of a city on March 14 is 10 $C^o$and March 15 is 5 C^o. This does not mean that the degree of the first day temperature is twice the degree of the second day temperature. Here, it can be said that zero temperatures make sense because there are temperatures below too.

2. Relative data:

Relative data includes various types of quantitative data that have absolute zero, Such as data on spaces, distances, etc. If the area of "A" hospital is 10000 $m^2$and the area of "B" hospital is 5000m^2, then the size of "A" hospital is twice the size of "B" hospital. Also, if the distance between the health center of the city and the residence of one of its doctors (1 km) and the second doctor's residence (2 km), the distance between the health center and the second doctor's residence is twice the distance from the first doctor's residence.

3.2 Presentation of statistical data:

The presentation and tabulation of statistical data is the second step after data collection in the concept of statistical analysis. The presentation of the data depends on the data type and nature of the facts required to be highlighted,

There are two basic ways of presenting and tabulating statistical data:

First: Tabular presentation.

Second: Graphical presentation.

3.2.1 Data table presentation

The order in which the downloaded data processed is called the table. In order for the statistical table to be clear, understandable and explaining the phenomenon studied should follow the following rules when design:

1. The table shall have a clear and explicit title.
2. The source from which the statistical data is taken should be mentioned.
3. Put the title column and row accurately.
4. Identify the appropriate units of quantitative measurement either in the main heading of the table or with the column heading and row.
5. Place notes at the bottom of the table when necessary, and when necessary mark with a special sign.

Frequency tables is a method used by the researcher to shorten the data that in itself represents a first step in the statistical analysis process.

The statistical tables vary widely, whether in terms of their characteristics, method of design or purpose to be achieved from them. We can distinguish between two types of statistical tables:

3.2.1.1 Simple Frequency tables

The data presented in the form of a frequency table and according to one attribute only. These data may be descriptive or quantitative.

A. Presentation descriptive data:

Example:

The following data represent the educational level of 30 staff members of a health center:

Secondary	academic	illiterate	secondary	academic
Illiterate	illiterate	secondary	elementary	secondary
Academic	secondary	illiterate	illiterate	elementary
Secondary	secondary	secondary	secondary	illiterate
Illiterate	elementary	academic	academic	elementary
Elementary	elementary	secondary	illiterate	secondary

1. Present the above data with a simple frequency distribution table.
2. Present the above data with a relative frequency distribution table.

Solution:

We schedule data taken from the first three columns representing the phenomenon-studied observation in ascending or descending order. The second for signs of download data, and the

third includes the frequency for each view "adjective". Thus, the frequency distribution schedule for the educational level will be as follows:

Download data table

Educational level	Download Data	Frequencies
academic	𝍸	**5**
secondary	𝍸 𝍸 |	**11**
elementary	𝍸 |	**6**
illiterate	𝍸 |||	**8**
Total		**30**

In order to obtain the simple frequency distribution table, we take the first and third columns according to Table (1).

Table (1) Simple repetitive distribution table

Educational level	frequencies
Academic	5
Secondary	11
elementary	6
Illiterate	8
Total	30

From Table (1) we can obtain the relative frequency distribution table, since the relative frequency:

Relative Frequency $= \frac{Partial\ frequency}{Total\ number\ of\ replicates}$

Table (2) the relative frequency distribution

Educational level	frequencies	relative frequency
academic	5	0.166
secondary	11	0.367
elementary	6	0.200
illiterate	8	0.267
Total	30	1.00

B. Presentation quantitative data:

Example:

The following data represent the age of 30 patients who were hospitalized during June 2014:

50	**59**	**47**	**53**	**56**	**55**	**51**	**60**	**50**	**68**
40	**53**	**54**	**59**	**57**	**52**	**59**	**48**	**61**	**54**
53	**45**	**58**	**54**	**43**	**63**	**53**	**51**	**70**	**57**

Show the above data in the distribution of relative frequency table

Solution: The presentation of the data in the distribution of frequency table requires the following steps.

1. **The Range "R":** Which represents the difference between the largest value and the smallest value in the distribution.
2. **Number of classes "K":** The determination of the number of classes depends on the size of the data and the objectives of the analysis and the researcher's experience, but can be determined according to the following formula:

$$K = 1 + 3.322 * (\text{Log } N)$$

3. **Class width "W":** It is sometimes called the class range and represents the class or distance capacity between the upper and lower limits of the class, and the length of the class is inversely proportional to the number of class. Whenever larger the class, the smaller the number and vice versa. The following relationship can be used to determine the length of the class:‹

$$W = \frac{Range}{Number\ of\ Classses}$$

4. **Lower and upper bounded classes:** Each class of frequency distribution has beginning and ending; the beginning represents the minimum of the class or smaller than a little, and the end means the upper limit of a class or greater than a little. It symbolizes the minimum with L and the upper limit with U.
5. **Center of classes "X_i":** The Center of classes represents the average for the upper and lower limits of each class, i.e. the class center of classes can be calculated by using the following formula: $\boldsymbol{X_i = \frac{L_i + U_i}{2}}$
6. **Frequency:** The frequency of the class is a part of the sample units that are characterized by being in numerical value - between the upper and lower limits for the class.

1. $R = 70 - 40 = 30$
2. $K = 1 + 3.322 * "Log30"$

   1+3.322 "1.477" = 1+ 4.90 = $5.9 \Longrightarrow 6$
3. $W = \frac{R}{K} = \frac{30}{6} = 5$

Then we prepare the required table.

Table (3) Frequency distribution of data

Classes	Frequency	relative frequency
40-45	2	0.066
45-50	3	0.100
50-55	12	0.400
55-60	8	0.267
60-65	3	0.100
65-70	2	0.066
Total	30	1.00

3.2.1.2 Double frequency tables

This a table combines two phenomena dependent on each other at the same time. The vertical axis represents the classes or groups of one of two phenomena. The horizontal axis represents the classes or groups of the other phenomenon. Finally, the squares resulting from the intersection of the axes of common frequencies between the two phenomena which are recorded.

Example:

The following symmetrical data represents the length (cm) and weight (kg) of a sample of twenty patients who were in a hospital.

length	**100**	**115**	**120**	**108**	**135**	**152**	**141**	**154**	**129**	**158**	**160**	**155**
weight	**30**	**39**	**38**	**33**	**40**	**49**	**41**	**49**	**47**	**54**	**52**	**50**
length	**146**	**139**	**147**	**138**	**144**	**127**	**109**	**118**				
weight	**48**	**45**	**43**	**41**	**37**	**41**	**48**	**45**				

Make a duplicate frequency table for the above data.

Solution:

The presentation of data in a frequency distribution table requires calculation of the range, number of classes and length of the class for both length and weight

For the length,

$R = 160 - 100 = 60$

$K = 1 + 3.332 * Log(20) = 5.6 \Rightarrow 6$

$W = \frac{R}{K} = \frac{60}{6} = 10$

And for the weight,

$R = 54 - 30 = 24$

$K = 1 + 3.332 * Log(20) = 5.6 \Rightarrow 6$

$W = \frac{R}{K} = \frac{24}{6} = 4$

Then we make a table with 36 squares through the intersection of the axes (6 × 6) and then divide the corresponding data pairs on the squares to get the table (4).

Table (4) Frequency distribution of patients by length and weight

Length / weight	110-100	120-110	130-120	140-130	150-140	160-150	total
34-30	2						2
38-34			1	1	1		3
42-38		1	2	1	1		5
46-42					1		1
50-46	1	1	1		1	1	5
54-50						4	4
total	3	2	4	2	4	5	20

3.2.1.3 Accumulative Frequency tables

Sometimes there may be a need to know the values or units of "views" that are less than or exceed a certain value, or to know the extent of the concentration or dispersion of the values of the phenomenon which are on two types:

1. Increasing cumulative frequency table:

In this case, we collect the frequencies starting from the top (the lower class) towards the lower (the upper class). The cumulative frequency of the first class is the same as the original frequency, because there is no previous frequency.

The second cumulative frequency of the second class will be the sum of the frequency of the first class and repeat the second class. Thus, we continue to add the frequency of the next class to the increasing cumulative frequency until we reach the last class, which is the cumulative frequency is the sum of the total frequencies.

2. Decreasing cumulative frequency table:

In this case, we collect the classes from the bottom (the upper class) upwards (the lower class) and the descending cumulative frequency of the first class (minimum). The sum is equal to the total frequencies. In addition, the descending cumulative frequency of the second category and the accumulated frequency of the first class will subtracted from the original class corresponding to it. Thus, we continue to descend until we reach

the last category class, which is the descending cumulative frequency that is equal to the original frequency and the following example shows the increasing and decreasing cumulative frequencies of the number of patients visiting a private hospital.

Table (5) simple frequency and accumulative increasing and decreasing distributions

simple frequency distribution table		**Accumulative Frequency tables**			
		increasing		**Decreasing**	
Classes	frequencies	Minimum limit for classes	Increasing frequency	Upper limits for classes	decreasing frequency
10-	3	Less than 10	0	More than 10	100
20-	4	Less than 20	3	More than 20	97
30-	11	Less than 30	7	More than 30	93
40-	20	Less than 40	18	More than 40	82
50-	36	Less than 50	38	More than 50	62
60-	18	Less than 60	74	More than 60	26
70-80	8	Less than 70	92	More than 70	8
		Less than 80	100	More than 80	0
total	100				

3.2.1.4 Research Data matrix

The matrix is the frame or shape that arranges data usually in a square or rectangle form which is divided into rows and columns and their number varies depending on the type and size of the entered data. Whenever the available and known data is small size the data become more private and vice versa whenever the available and known data is larger the data is provided back to the public.

The columns are usually used to enter data for the phenomena to be studied. For example, if we have data on population, males, females, births, deaths and income, and we want them to be included in the matrix for the purpose of study, then we will assign each column starting from the first column of the population and so on until we reach the last column that will be assigned to income. Therefore, these columns will represent the studied phenomena as variables because their value varies from region to region and from time to time.

The rows are representing views which are studying the phenomenon required which may be administrative units, cities, hospitals, laboratories and pharmacies, pharmacies and others.

Table (6) represents one of the forms of matrices

Varaibles / Cases	1	2	3	4	5	6
	Population million	Males million	Females million	Births Thousand	Mortality Thousand	Income rate A thousand dinars
The city (1)	8	4.5	3.5	18	3	6200
The city (2)	6	3.6	2.4	12	2	7500
The city (3)	4	2.7	1.3	10	5	4600
The city (4)	9	5.4	3.6	25	6	5800
The city (5)	3	1.6	1.4	6	4	5400
The city (6)	6.5	3.5	3	5	3	6300
The city (7)	4.25	2.25	2	6	2	5900
The city (8)	5	2.75	2.25	7	3	6100
The city (9)	7	3.5	3.5	9	4	5700
The city (10)	4.5	2.3	2.2	5	2	4900

3.2.2 Graphical presentation

The study of any phenomenon of a society may require a statistical representation of the data, which are often many, varied, and difficult to read, and explored if they remain in the form of statistical tables. In addition to get more clarity, we can transformed it into graphic forms through which to get conclusions and interpretations and predictions of the evolution of the phenomenon in the future.

Therefore, it is possible to say that the graphs are a visual translation of the statistical tables that enable us to highlight the characteristic or frequency of the evolution of the phenomenon, which is difficult to detect by tables only. In addition, the

graphical forms are a form of geographical expression through which figures and percentages are transformed into clear and specific information and facts that facilitate observation, comparison, analysis and extraction of results.

Moreover, the methods used to represent the statistical data graphically are many and varied, including the presentation using columns, lines or graphs and the statistical data which can be made in geometrical shapes such as ball and cube. These geometric or flat geometric shapes can be replaced by drawings and certain images of strong relevance to the particular phenomenon.

The following is a summary of the most important statistical data based on the phenomenon.

3.2.2.1 Bar Charts

Is a modern geometric shapes set of vertical rectangles, "horizontal" which represent an equal bases on the horizontal axis. The heights will vary depending on the different number of iterations represented by those columns or percentage. In addition, drawing it requires the following steps:

1. Draw a horizontal axes perpendicular one "X-axis" and the other vertical "y-axis."
2. Divide the horizontal axis into equal sections on a suitable scale, including all the data and the rectangle base representing the phenomenon.

3. Divide the vertical axis into sections with an appropriate scale that represents the frequencies so that each attribute corresponds to the corresponding frequency.

Example:

Table data (7) presents the age of a group of patients in a hospital who are required to be represented as columns.

Age Groups	Number of Patients
10-13	3
14-17	7
18-21	9
22-25	6
26-29	5
30-33	2
Total	32

Solution: When you follow the previous steps to draw graphs columns we can get the following figure:

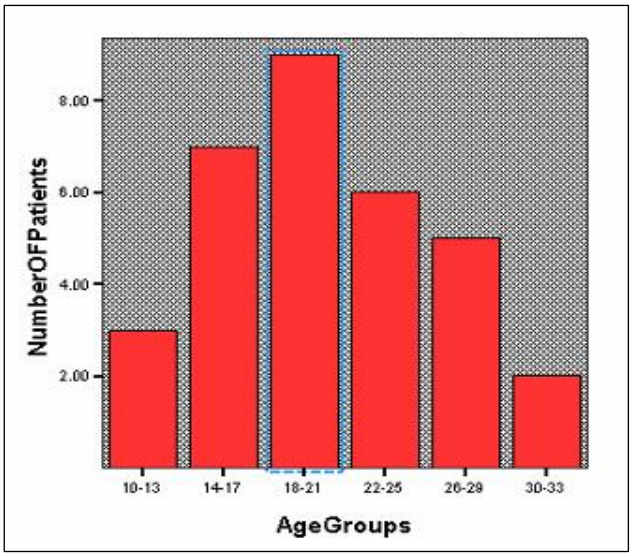


Figure (4) graphs columns for data table (7)

We use column graphics to represent two variables or more columns by using Clustered Bar. It is drawing a number of clustered bar column each of which represents one variable values as in the following example:

Table (8) Number of births of males and females for the years 2010-2014 to a hospital

Years	Male	Female
2010	120	150
2011	220	170
2012	175	200
2013	250	230
2014	260	210

The data above is required in the form of clustered bar column.

Solution:

When you follow the previous steps to draw graphs columns we can get the following figure:

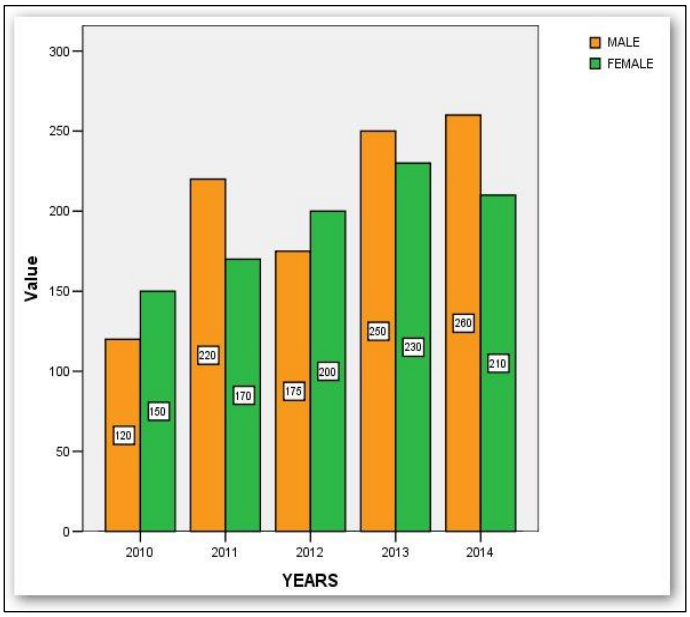


Figure (5) the clustered bar columns of births for males and females

3.3.2.2 Histogram

Is a type of bar chart, but there are no spaces between the bars. A histogram is a set of adjacent vertical rectangles; it represents the height of each rectangle assigned to a certain specific class. For Histogram chart requires the following steps:

1. Drawing two orthogonal axes, the vertical ones representing the frequencies and horizontal represents the classes.
2. Each class represents a rectangle whose height is a repeating class, and its base width is the length of the class.

3. Each rectangle starts from where the previous rectangle has ended.

The following diagram represents the histogram of table data (7).

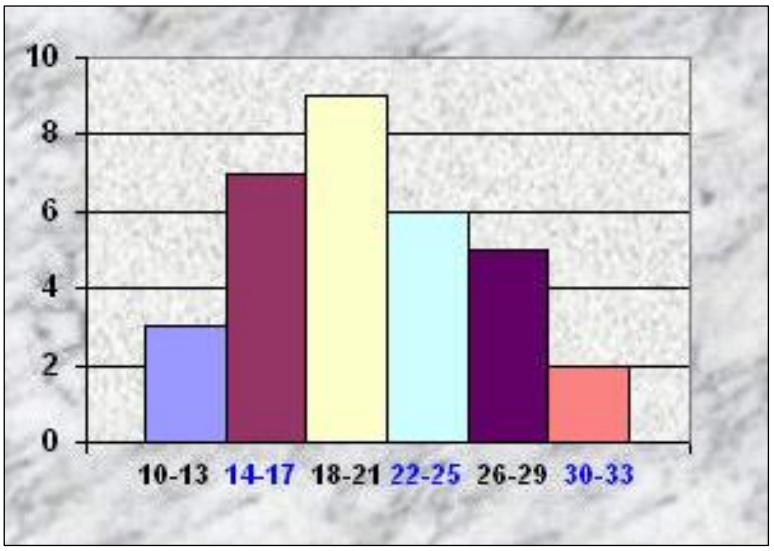


Figure (6) Histogram of the absolute data of the table (7)

3.3.2.3 Frequency polygon

Is a series of curved straight lines connecting between their respective points located above the center of a class at a height representing the frequency of that class. To represent the frequency distribution table data with frequency polygon requires drawing two axes perpendicular, horizontal for class centers and vertical for classes, and then pass on the points whose coordinates are the center of the classes and the frequencies, so that we get a curve line of graph representing the frequency polygon.

Example: Use table data (7) to draw the frequency polygon.

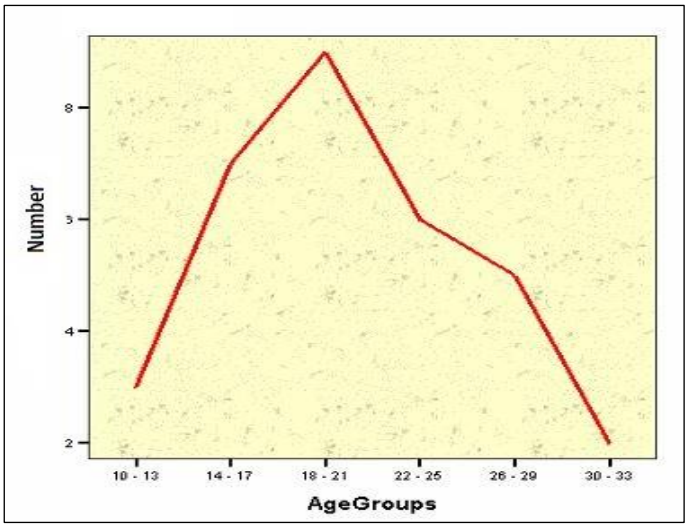


Figure (7) frequency polygon of table data (7)

We can draw a frequency polygon by using a histogram after taking the half-top of rectangles that represent the class centers by points and then connect those points with lines as in figure (8)

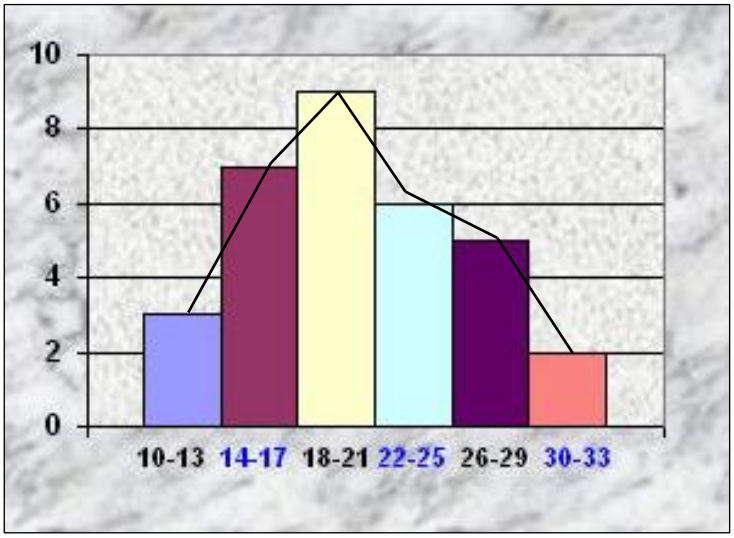


Figure (8) frequency polygon from histogram of table data (6)

3.3.2.4 Cumulative Frequency Polygon

Is a series of curved straight lines reaching between points located above the actual boundary of the classes represented on the horizontal axis and at the height of the cumulative frequencies rising or descending on the vertical axis.

A. Increasing Cumulative Frequency Polygon:

It is drawn by follow the following steps

1. Draw two axes perpendicular, one horizontal and the other vertical.
2. The horizontal axis divided into equal sections to represent all classes.
3. Divide the vertical axis into equal sections to represent all the cumulative frequency.
4. Determine the points on the shape, so that the X-coordinates of the upper boundary points of the classes and Y- coordinates have the cumulative frequencies corresponding to those classes.
5. Connect points that are determined on the shape with straight lines that will start from the lowest cumulative frequency to the largest cumulative frequency.

Example:

Draw the *Increasing Cumulative Frequency Polygon* by using table data (5).

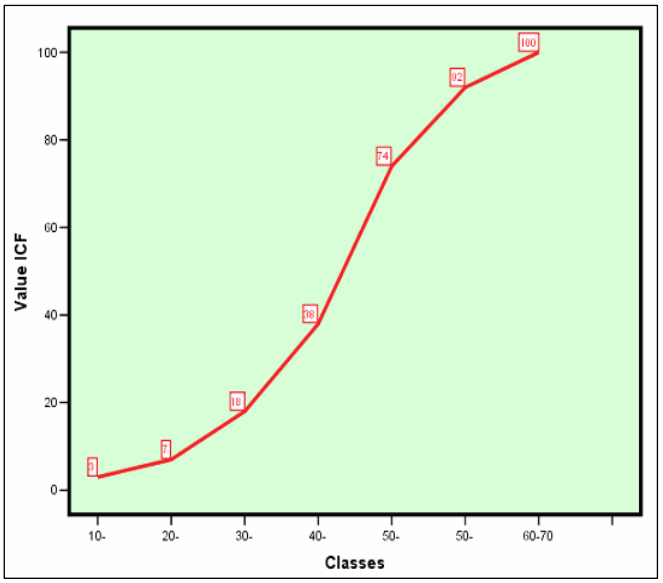


Figure (9) Cumulative Frequency Polygon

B. Decreasing Cumulative frequency Polygon:

The steps for drawing decreasing cumulative frequency Polygon are the same steps followed in increasing cumulative frequency polygon, except for cumulative frequencies; it will be here the decreasing cumulative frequency Polygon. Therefore, decreasing Cumulative Frequency Polygon will start from the highest point (Greater decreasing cumulative frequency) Correspond to the minimum classes to the lowest point (Less decreasing cumulative frequency) corresponds to the upper limit of classes.

Example:

Draw decreasing Cumulative frequency Polygon by using table data (5)

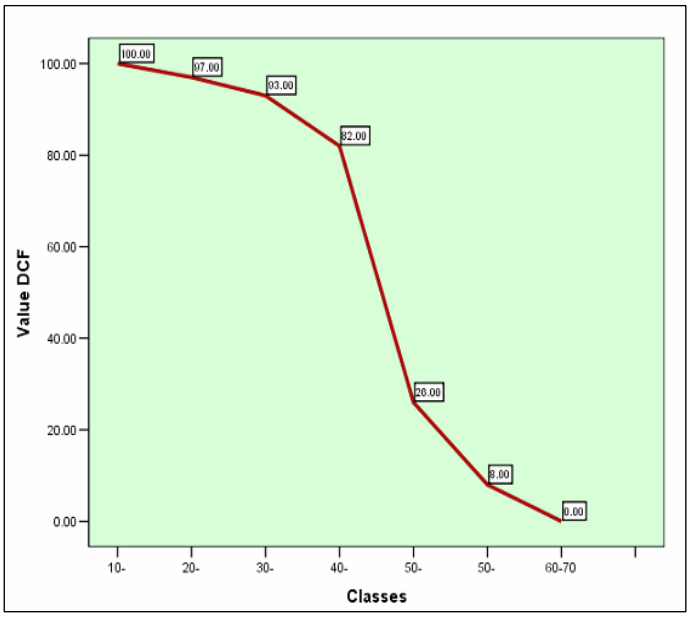


Figure (10) decreasing g Cumulative frequency Polygon

3.3.2.5 The Pie Chart

It is the geometric shape represents the total value of the phenomenon; they are divided into partial sectors that correspond with the values of the partial groups that make up the phenomenon. These sectors are distinguished from each other in different colors or shades for easy clarification.

The Pie Chart is often used in the following cases:

1. When the object of the Pie chart is to compare the different parts for the total sum.
2. Explain the relative development of parts of the phenomenon at different periods.
3. When the parts of the phenomenon are relatively few.

The descriptive data Pie chart is often used for metadata, and select the angle to each sector is defined as follows:

$$\frac{Partial\ data}{Total\ data} * 360$$

Example:

The following data represent the number of births classified according to the place of birth for the city of Baghdad in 2013. These data are required to represent the data in a Pie chart.

place of birth	Government hospital	private hospital	The house	Another place
Number of births	61230	58420	72325	18723

Solution:

1. Determine the amount of angle

place of birth	Number of births	Relative frequency	amount of angle
Government hospital	61230	0.29	0.29×360=104.4
private hospital	58420	0.28	0.28×360=100.8
The house	72325	0.34	0.34×360=122.4
Another place	18723	0.09	0.09×360=32.4
sum	210698	%100	360

2. Drawing the Pie chart:

Draw a circle divided into four parts, each birthplace has a part proportional to the amount of angle assigned to it as shown in Fig (11).

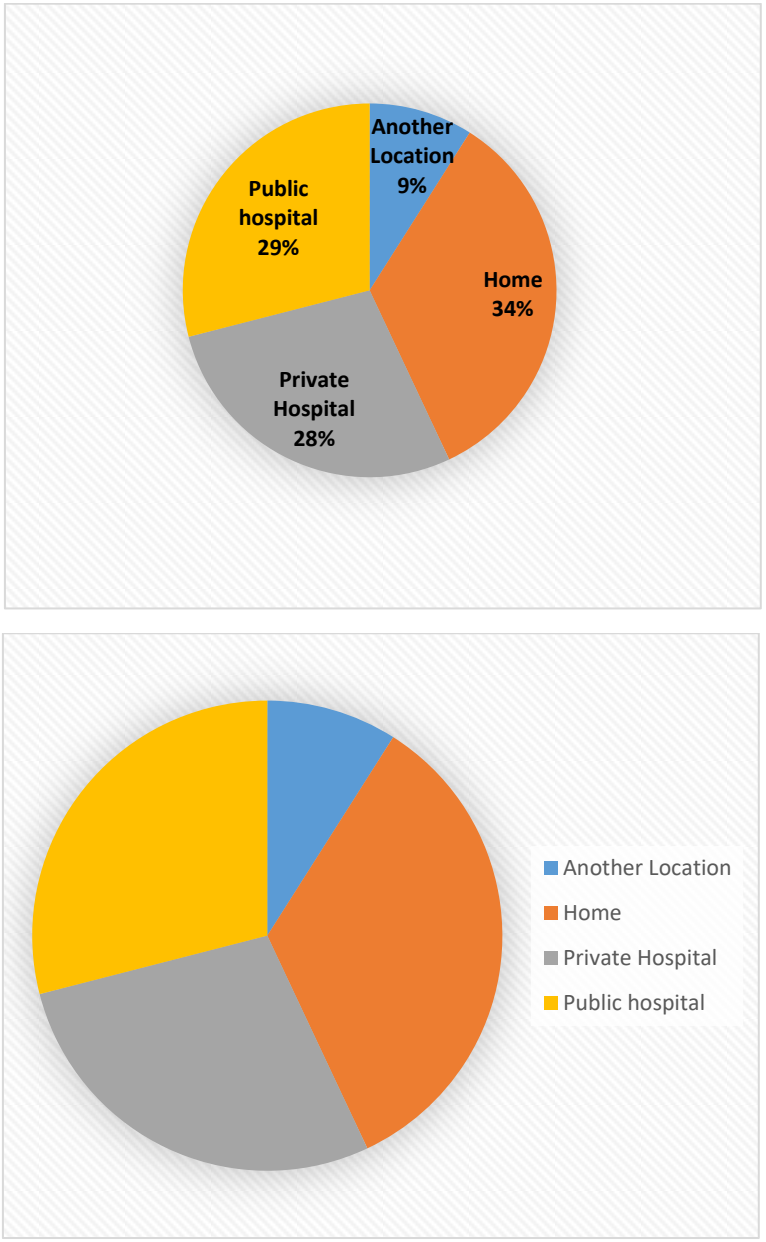


Figure (11) Pie chart

Chapter Four

Measures of central Tendency & Desperation

4. Measures of central Tendency & Desperation

4.1 Measures of Central Tendency

4.1.1 The Arithmetic Mean

4.1.2 The Geometric mean

4.1.3 The Harmonic Mean

4.1.4 The relationship between the arithmetic mean and the geometric and harmonic

4.1.5 The Weighted mean

4.1.6 The Median

4.1.7 The Mode

4.1.8 Relationship between the arithmetic mean and the Median and the Mode

4.2 Measures of Dispersion

4.2.1 The Range

4.2.2 Interquartile Range

4.2.3 The Mean Deviation

4.2.4 Standard Deviation

4.2.5 Variance

4.2.6 Standard Error

4.2.7 Coefficient of Variation

4.2.8 Standard Score

Chapter four

Measures of central Tendency & Desperation

4. Measures of Central Tendency & Desperation

The statistical data are two characteristics which are essential to help to give clear indications to describe these data: Central tendency and its measures of averages that have a great importance in the subject of inference characteristics by estimating the numerical values of some indicators of the study and research society, which are often not known.

For measures desperation, they measure the extent to which data values dispersed from the center, i.e., the degree to which the values of those data are dispersion from the center to form a clear idea of how homogeneous of those values.

4.1 Measures of Central Tendency

The representation of the tabular and graphic data are reliable statistical indicators to describe the phenomenon and presented it in a short and simple way. However, we always prefer to use the methods of quantitative measurement "numerical" to measure the data on a particular value so that it represents the best representation. This occurs when that value attracts the largest number of data values of the phenomenon and vice versa loses that value their importance if they moved away many data about them.

The measures of central tendency "averages" are among the most important numerical measures used for this purpose. The most important and most common arithmetic mean, Median, and Mode. There is also the geometric mean and the harmonic mean, but they are less widely used. Each of these averages has advantages and disadvantages depending on the nature of the data on the one hand and on the purpose of its use on the other.

4.1.1 The Arithmetic Mean

The mean (or average) is the most popular and well-known measure of central tendency. It is even the first statistical measures at all, the most important; because of the advantages and characteristics, and to entering into the calculation of many other statistical measures.

The arithmetic mean is generally defined as "the sum of all the values in the data set divided by the number of values in the data set" It is calculated from the grouped and ungrouped data.

The population's arithmetic mean using the Greek letter μ , and the sample by $\overline{X}$.

First: The arithmetic mean the arithmetic mean of the ungrouped data is calculated by using the following formula:

a. For population $$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

b. For the sample $$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Where:

$\sum X_i$ the sum of the values of Views.

N: the number of views of the statistical population.

n: the number of views for the sample.

Example:

The following data represent the weights of ten persons in kilograms and the calculation is required. Find the arithmetic mean (Average weight) for these *people* 55, 71, 68, 70, 56, 63, 67, 58, 55, 57 kg.

Solution:

$$\overline{X} = \frac{\sum x_i}{n} = \frac{55+71+68+70+56+63+67+58+55+57}{10} = \frac{620}{10} = 62Kg$$

i.e. that the average weight of these people is 62 kg.

Second: The arithmetic mean of the data classified is grouped using the following formula:

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{f_1 + f_2 + \cdots + f_k}$$

Where $\sum f_i x_i$ represents the sum of the multiplication of the frequency of each fi class in the corresponding xi center of class.

Example

The following frequency distribution represents 20 patients with pulmonary tuberculosis by age group.

Classes	10 -14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44
Frequencies	2	3	3	4	3	3	2

Find the arithmetic mean of the age of these patients.

Solution:

To find the arithmetic mean, it is necessary to find the class centers and then multiply the position of each class center in the corresponding frequency then we find the sum of the collection. Finally, we compensate for the data of the data to obtain the value of the arithmetic mean. Finally, compensates in the law of grouped data to obtain the value of the arithmetic mean.

Classes	Frequency Fi	Classes xi	Fixi
10- 14	2	$\frac{10+14}{2} = 12$	24
15-19	3	17	51
20-24	3	22	66
25-29	4	27	108
30-34	3	32	96
35-39	3	37	111
40-44	2	42	84
Total	20		540

Therefore, the arithmetic mean is equal:

$$\overline{X} = \frac{\sum_{i=1}^{7} f_i X_i}{\sum_{i=1}^{7}} = \frac{540}{20} = 27 \text{ Years}$$

Properties of the arithmetic mean:

1. To calculate the arithmetic mean we must have quantitative data, that is does not fit the arithmetic mean descriptive, nominal, or ordinal data.
2. The arithmetic mean depends on account of all the values of views which is not neglecting any of them.

3. The arithmetic mean is affected by extreme values and singular, so it is not suitable for skewed distributions.
4. The arithmetic mean cannot be calculated in the case of open tables, whether it is from one of the parties or both "Because it is not possible to find a class center".
5. The total deviations of values from their arithmetic mean is zero, i.e.$\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}}) = \mathbf{0}$.
6. The sum of squares of deviations of values from the arithmetic mean is minimal. i.e. $\sum(\boldsymbol{y_i} - \overline{\boldsymbol{y}})^{\mathbf{2}} = \mathbf{0}$, are less than the sum of squares of deviations of values for any value other than the arithmetic mean.
7. Add any fixed value to the arithmetic mean, subtracting it, multiplying it or dividing it which makes the arithmetic mean increases or decreases by that fixed value.
8. Cannot find the arithmetic mean of the graph on the opposite of the medium and the mode.

4.1.2 The Geometric mean

The geometric mean of a set of values $x_1, x_2, \dots, x_n$ is the nth root of the sum multiplying those values v and denoted by $\overline{G}$.

First: The geometric mean of ungrouped data

Which calculated by the following formula:

$$\overline{\boldsymbol{G}} = \sqrt[\boldsymbol{n}]{(\boldsymbol{x_1})\,(\boldsymbol{x_2})\dots(\boldsymbol{x_n})}$$

Because of the difficulty of calculating the normal calculations, we resort to the use of logarithms after the previous version which is converted to the following formula: $\boldsymbol{log\overline{G}} = \frac{\sum \mathbf{log}(\boldsymbol{x_i})}{\boldsymbol{n}}$

Example:

Find the geometric mean of the following values: 2, 4, 2, 16.

$\overline{G} = \sqrt[4]{(2)(4)(2)(16)}$

$Log\overline{G} = \frac{Log(2) + Log(4) + Log(2) + Log(16)}{4} = \frac{2.40824}{4} = 0.60206$ By finding the corresponding logarithm (Antilog), which symbolized by X^{10} or from the number of tables corresponding to the logarithms, we get the value of the geometric mean) $\therefore \overline{G} = 4$

Second: The geometric mean of grouped data

It will be calculated as follows: $\overline{G} = \sum^{f_i}\sqrt{(x_1{}^{F1})(x_2{}^{F2})...(x_n{}^{Fn})}$ and to simplify its calculation, this formula can be converted into a logarithm formula as follows:

$$Log\overline{G} = \frac{\sum FiLog(x_i)}{\sum F_i}$$

To compute the geometric mean of grouped data, the following steps can be taken:

1. Find class centers (xi).
2. Find logarithms of class centers (xi).
3. Multiply the logarithm of each class center with the corresponding frequency and then combine those values to get $\sum F_i Logx_i$.
4. Divide the sum result $\sum F_i Logx_i$ to the total of the frequencies to obtain the logarithm of the geometric mean.

5. Find the real value of the geometric mean by using the number tables corresponding to the logarithms or through the calculator from the base of the natural logarithm (Antilog) which denoted by 10^x.

Example

The following table represents the daily wages of 20 workers in a private hospital and is required to find the geometric mean of those wages.

Wage / Dollar	**0-10**	**10-20**	**20-30**	**30-40**
N0. of Workers	**5**	**8**	**3**	**4**

Solution:

Class	**Frequency**	**Class center x_i**	**Log x_i**	**F_i Log x_i**
0-10	5	$\frac{0+10}{2}=5$	0.699	3.495
10-20	8	15	1.176	9.408
20-30	3	25	1.397	4.191
30-40	4	35	1.544	6.176
Total	20			23.27

$$Log\overline{G}=\frac{23.27}{20}=1.1635$$

$\therefore \overline{G}=14.57\$$ Geometric mean of wage.

Geometric mean properties:

1. It gives more mild results than arithmetic mean.
2. It is not affected by extreme values but cannot be used with negative or zero values because the root of negative values

which is not valid. In addition, zero value eliminates the rest of the values; because the multiplication in zero is zero.

3. It is the most appropriate measures to calculate the average rates and rates of growth in production or population.
4. The geometric mean of a set of values which is always smaller than or equal to the arithmetic mean of that set of values.

4.1.3 The Harmonic Mean

Harmonic mean can be expressed as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations which is denoted by $\overline{H}$.

First: Harmonic mean of ungrouped data

If the variable (X) values are $x_1, x_2, \ldots, x_n$ as n represents the size of the group, the harmonic mean can be expressed as follows:

$$\overline{H} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x_i}}$$

Example:

Find the harmonic mean from the following values: 10, 20, 40, 50

Solution:

$$\overline{H} = \frac{4}{\frac{1}{10} + \frac{1}{20} + \frac{1}{40} + \frac{1}{50}} = \frac{4}{0.195} = 20.5$$

Second: Harmonic mean of the data grouped

If $x_1, x_2, \ldots, x_n$ represent the class centers in the frequency distribution table and are weighted by the corresponding frequencies $F_1, F_2, \ldots, F_n$, respectively. The harmonic mean can be expressed as follows:

$$\bar{H} = \frac{\sum f_i}{\sum \frac{f_i}{x_i}}$$

Example:

Find the harmonic mean of the repetitive distribution table:

Class	2.5-7.5	7.5-12.5	12.5-17.5	17.5-22.5
Frequencies	20	50	20	10

Solution:

Class	**Frequencies Fi**	**Class center xi**	$\frac{1}{x_i}$	$f\frac{1}{x}$
2.5-7.5	20	$\frac{2.5+7.5}{2}=5$	0.2	4
7.5-12.5	50	10	0.1	5
12.5-17.5	20	15	0.06	1.2
17.5-22.5	10	20	0.05	0.5
Total	100			10.7

$\overline{H} = \frac{100}{10.7} = 9.34$ The harmonic mean value.

Properties of the harmonic mean:

1. The harmonic medium is most used when a set of data is given to a fixed unit.

2. In general the measure by harmonic mean is complex. However, the nature of the data sometimes required to be used in cases where data are given in reverse, i.e., when the original values are not given directly, but in terms of other units such as "in the case of prices," or units of time "in the case of speed".

4.1.4 The relationship between the arithmetic mean and the geometric and harmonic

The arithmetic mean is always larger than the geometric mean and geometric mean is greater than the harmonic mean of the same values.

i.e. $\overline{H} < \overline{G} < \overline{X}$

Example:

Find the arithmetic mean and the geometric mean, and harmonic mean of the following values:

5, 10, 20, 40, 50.

For the arithmetic mean:

$$\overline{X} = \frac{5+10+20+40+50}{5} = \frac{125}{5} = 25$$

And for geometric mean: $\overline{G} = \sqrt[5]{(5)(10)(20)(40)(50)}$

$$Log\overline{G} = \frac{Log(5) + Log(10) + Log(20) + Log(40) + Log(50)}{5}$$

$$\therefore \overline{G} = 18$$

And for harmonic mean:

$$\overline{H} = \frac{5}{\frac{1}{5}+\frac{1}{10}+\frac{1}{20}+\frac{1}{40}+\frac{1}{50}} = \frac{5}{0.395} = 12.6$$

i.e. $\overline{H} < \overline{G} < \overline{X}$

$12.6 < 18 < 25$

4.1.5 The Weighted mean

When values are not equal in terms of importance to a wide range of items, the weighting of these values is required in proportion to the importance of each of them to become a computation of their arithmetic mean which is acceptable. This is known as the weighted mean, calculated as follows:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

First: For ungrouped data

Example:

Pharmacy sells three types of surgical thread at different prices, selling the first type 5 threads at a price of 12000 $, the second type sells 8 thread at 8000 $, and the third type sells 12 thread at a price of 4000 $. Find the weighted arithmetic mean sale price of a single thread of those threads.

Solution:

$$\overline{Y} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} = \frac{5\times1200 + 8\times800 + 12\times400}{5+8+12}$$

$$= \frac{6000 + 6400 + 4800}{25}$$

$$= \frac{17200}{25} = 688 \text{ \$ Per thread price}$$

Second: For grouped data

The frequency of the grouped values is actually reflects the amount relative importance "weight" of the center class corresponding to that frequency. The greater the frequency of class has increased the impact of this class on the calculated arithmetic mean. So, it can be said that the arithmetic mean of the grouped data $\rightarrow$ the weighted arithmetic mean.

$$\overline{Y} = \frac{\sum_{i=1}^{n} w_i f_i x_i}{\sum_{}^{n} w_i f_i}$$

Example:

The following frequency distribution represents the amount of production of a particular medicine for one production day for a pharmaceutical factory distributed according to the number of machines and the number of working hours specified for each machine in accordance with its design capacity.

Class Tons	Number of machines	The design capacity for each machine
0-2	2	5
2-4	3	6
4-6	6	4
6-8	4	5
8-10	1	4

Find the average machine productivity of the "weighted arithmetic mean"?

Class	Frequency fi	Weight wi	Class center Xi	wifi	wifixi
0-2	2	5	1	10	10
2-4	3	6	3	18	54
4-6	6	4	5	24	120
6-8	4	5	7	20	140
8-10	1	4	9	4	36
$\sum$	16			76	360

$$\overline{X}_w = \frac{\sum_{i=1}^{n} W_i F_i X_i}{\sum_{i=1}^{n} W_i F_i} = \frac{\sum_{i=1}^{5} W_i F_i X_i}{\sum_{i=1}^{5} W_i F_i} = \frac{360}{76} = 4.7 \text{ Tons.}$$

4.1.6 The Median

Is the value that lies in the middle of the values after ascending or descending order in the case of odd data, or the value of the arithmetic mean of the two values that are averaged in the case of even data.

First: Median for ungrouped data

It can calculated by using the following steps:

1. Order the values ascending or descending order.
2. Determine the rank of the median $\left(\frac{N+1}{2}\right)$, as N represents the number of values of the statistical population and n are used instead in the case of the sample.
3. If the number of values is odd, the median is the value that is equal to arrange $\frac{n+1}{2}$.

4. If the number of values even, the median is located between the value that order $\left(\frac{n}{2}\right)$ and the value that order $\left(\frac{n}{2}+1\right)$ i.e.

$$Median = \frac{The\ value\ that\ order\ \frac{n}{2} + The\ value\ that\ order\ \frac{n}{2}+1}{2}$$

Example:

The following data represent the weights of a group of patients in a hospital. The median value of these weights is required: 60, 100, 40, 50, 40, 30, and 80

Solution**:** Since the number of values is odd, the median value will be the value that $\frac{(n+1)}{2}$ order after the order of values is ascending or descending

40,40,30	(50)	100,80,60
60,80,100	(50)	30,40,40

The median order $\left(\frac{n+1}{2}\right) = \frac{7+1}{2} = 4$, i.e. the median value will be the fourth value of 50, whether in ascending or descending order.

Example:

The following data represent the number of people with hepatitis in a city and the median value of the Infected is required.

20,38,40,35,25,23

Solution:

Since the number of values is even, the median value will be the arithmetic mean of the first two values order by $\left(\frac{n}{2}\right)$ and the

second by the order $\left(\frac{n}{2}+1\right)$ after the order of values ascending or descending.

	30	
23 20	35 25	40 38
38 40	25 35	20 23

Median order of the first value $\frac{6}{2}=3$ the third value. Median order of the second value $\frac{6}{2}+1=4$ the fourth value.

That is, the median value $\frac{35+25}{2}=30$ is in the case of an infected ascending or descending order.

Second: Median for grouped data

To calculate the median of the frequency distributions we, need to use the following steps:

1. From the simple frequency distribution table, we will make ascending or descending cumulative table.
2. Determine the median order that is equal to$\frac{\sum f_i}{2}$.
3. Determine the class of the median which is the class where the value of the median is between its lowest and highest, that is, we find the class in which the reading in the order $\frac{\sum f_i}{2}$. This done by looking at the frequency cumulative ascending or descending column from two consecutive values the order of the median which is between them. These values correspond to two digits in the class boundary column and these numbers are: The upper and lower limits of the median class.

4. The following formula is used to calculate the median value.

$$med = L + \frac{n/2 - F}{f_m} C$$

Where:

L = the minimum of the median class in the case of the frequency cumulative ascending and the maximum frequency in the case cumulative descending.

n = Number of values in the data set.

F = Previous accumulated frequency of the median position value.

Fm = The original frequency of the median class.

C = length of the median class.

Example:

The following data represent the number of patients with typhoid fever in one of Baghdad hospitals distributed by age groups in 2014. The median age is required for patients with this disease.

Class	frequency fi	Minimum class	frequency cumulative ascending	The upper limits of the class	frequency cumulative descending
10-14	2	Less than 14	2	More than 10	20
15-19	3	Less than 19	5	More than 15	18
20-24	3	Less than 24	8 10→	More than 20	15
25-29	4	Less than 29	12	More than 25	12 10→
30-34	3	Less than 34	15	More than 30	8
35-39	3	Less than 39	18	More than 35	5
40-44	2	Less than 14	20	More than 40	2

In the case of frequency cumulative ascending

$$med = L + \frac{\frac{n}{2} - F}{F_m} C$$

$$= 25 + \frac{20/2 - 8}{4} 4 = 27 \text{ Year}$$

In the case of frequency cumulative descending

$$med = 29 + \frac{10 - 12}{4} 4$$

$$= 29 + \frac{-8}{4} = 27 \text{ Year}$$

The median can be graphically found for any of the cumulated frequencies by following these steps:

1. Represents the cumulative frequency ascending and descending graphically.
2. Determine the rank of the median by dividing the total of the frequencies by 2 and by our example, it is equal to 10.
3. We draw a straight line parallel to the horizontal axis from the median rank point on the vertical axis until it meets the cumulative frequency ascending and descending curve at point A.
4. Move from point A to the horizontal axis until it is connected to point B and the value on the horizontal axis corresponding to point B is the median value.

The following graphs clarify that:

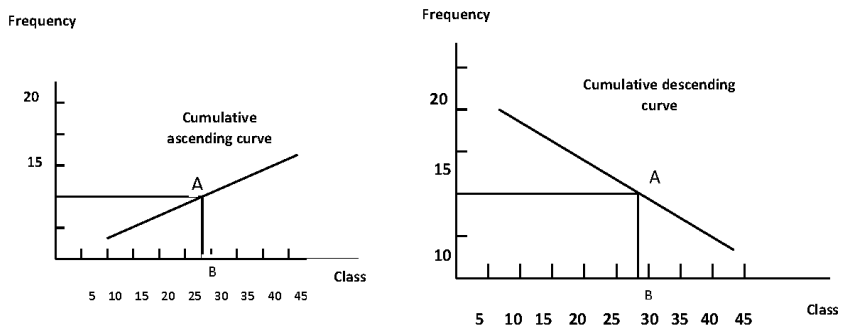


Figure (11) the value of the median graphically from the accumulated frequency ascending or descending.

The value of the median can be found in another way by drawing each of the two cumulative frequency ascending and

descending curves on the same one, and from the point of intersection we go down a column on the horizontal axis; to get the median value. As shown in the figure below:

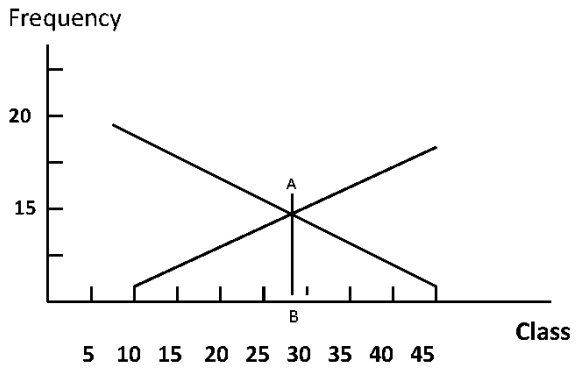


Figure (12) the value of the median graphically from the intersection of the ascending and descending frequency curves.

Median Properties:

1. Is not affected by extreme values or singular values but is influenced by the central values.
2. Used in skewed distributions.
3. It can be used in the case of open class from one end or both.
4. It can be calculated from the graph.

4.1.7 The Mode

One of the less accurate central tendency measures is often used for quick comparisons that do not require a high degree of accuracy.

The mode of a set of data values is the value that appears most often and the group may be Unmodal or Bimodal or may not

be a set of values. In the case of grouped data, the value of the mode is the center of the mode class corresponding to the largest frequency in the case of the symmetric distribution.

First: Mode for ungrouped data: It is calculated as follows

Mode= the most frequent value

Example:

The following data represent the age of the day for six children with diarrhea. Find a mode value of those ages:

30, 20, 25, 15, 10, 20.

Since the value 20 is repeated more than others, so 20 will be the only Mode value.

Example:

Find the value of the Mode if the age of the patients as follows: 20, 25, 5, 15, 24, 20, 10, and 5. We note that this group are two values Mode 20, 5

Example:

If the age of the patient is as follows:

12, 15, 4, 10, 8, 5

Because there is no repetition of any value from this group, there is no prognostic value for this disease.

Second: Mode for grouped data: It is calculated as follows**:**

$$\text{mod} = L + \frac{d_1}{d_1 + d_2} C$$

Where:

L = Minimal Mode of class "The class with the largest frequency".

d_1 = The difference between the frequency of the class Mode and the frequency of the previous class.

d_2 = The difference between the frequency of the Mode class and the frequency of the next class.

C = length of the Mode class.

Example:

Find a mode of the previous example data.

Solution:

Class C	Frequency f	
10-14	2	
15-19	3	
20-24	(3) ←	The previous frequency d1= 4 – 3
[25]- 29	[4] ←	Frequency mode class
30-34	(3) ←	Subsequent frequency d2=4 – 3
35-39	3	
40-44	2	
	20	

So, the mode is equal to:

$$Mod = L + \frac{d_1}{d_1 + d_2} c$$

$$= 25 + \frac{1}{1+1} 4$$

$$= 25 + \frac{4}{2} = 27 \ year$$

The Mode value can be calculated graphically from the histogram by following the steps below:

1. Representing the class of Mode and the preceding and subsequent class with a Histogram.
2. We connected the upper right head of the Mode class rectangle of the 'A' in the upper right head of the previous rectangle 'B' and connected upper left head of the Mode class rectangle of the 'c' in the upper left head of the subsequent rectangle 'D'.
3. Take down a column line from the intersection point on the horizontal axis, to get the value of the Mode as shown in the figure below.

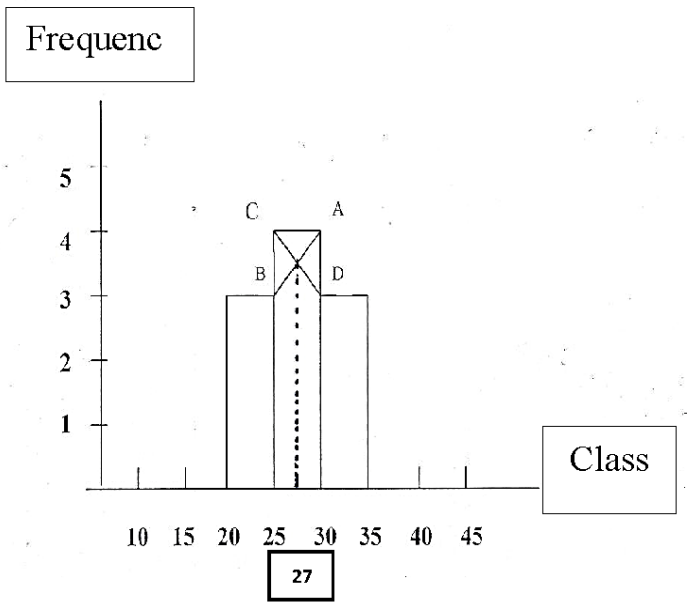


Figure (13) the value of the Mode in the graph method

The Mode properties:

1. It is easy to calculate it mathematically and graphically.
2. Is not affected by extreme values.
3. It can be calculated in the case of open class from one or both ends.
4. It can be used with quantitative and qualitative values.
5. Is affected by the length of the class in the distribution, as its value changes by changing the number of class in the frequency distribution; (because the change in the number of class changes the value of the class center containing the largest frequency).

4.1.8 Relationship between the arithmetic mean and the Median and the Mode

There is a relationship between the Arithmetic and Median and Mode and take this relationship four different cases:

1. In Symmetry and homogeneous shape of the curve completely, as in shape (14), where the three mean (arithmetic mean = med = Mode) i.e.

Mean = Med = Mod

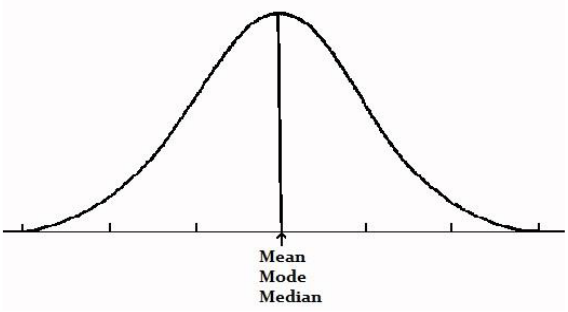
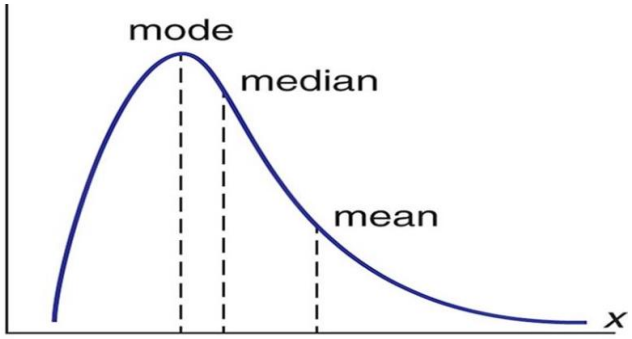


Figure (14) symmetric distribution

2. In the case that the curve is not asymmetrical i.e., Skewness and Skewness to the right, that is, the twist is positive (+) so the relationship is mean> med> mod.

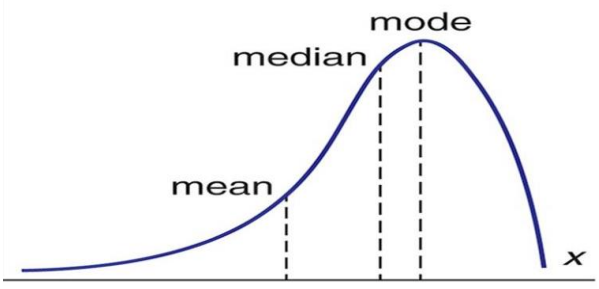
This means that the mean is greater than the median, and this is greater than the mode, as shown in Figure (15).



Figure(15) The relative position of the Mode, median, and Mean of the twisted curve to the right

3. In case the curve is not symmetric i.e., Skewness and Skewness overflowing towards the left. That is, the twisting is negative (-) so the relationship is mean <med <mod.

That is, the mean is smaller than the median and this is smaller than the Mode. As shown in Figure (16).



The relative position of the Mode, median, and Mean of the twisted curve to the left

4. If the distribution curve is Skewness moderately, the relationship is:

Mean - mod = 3(mean-med)

I.e., the arithmetic mean - the mode = 3 (the arithmetic mean - the median).

This relationship is not accurate but approximate.

4.2 Measures of Dispersion

Central tendency metrics may be insufficient to describe a set of data full description, some samples may be equal in their arithmetic mean, although the distribution of their data differs over their position. The following samples have a single arithmetic mean but are no doubt different from each other.

Sample (1): 10, 10, 10, 10, 10….. and its arithmetic mean = 10
Sample (2): 12, 11, 8, 9, 10……. its arithmetic mean = 10
Sample (3): 19, 2, 16, 10, 3..... its arithmetic mean = 10

The three groups have arithmetic mean of 10 but the first group all have equal values, i.e., it is equal to the arithmetic mean value, While the data in the second group are spread around this median as much as (the closer they are between them, the less dispersed). The data of the third group is spread around the arithmetic mean but in much more bigger that is (More divergence among them, the more dispersed).

So there was therefore a need to find measures to measure the degree of homogeneity of "convergence" or dispersion of the "divergent" data values from each other values.

These standards are called measures of dispersion that standardization of the degree to which quantitative data tend to spread out around the average value (one of the measures of central tendency) and often is the arithmetic mean or median value.

The value of dispersion or variance may be equal to zero if there is no difference between the data "That is, if all data are equal in value" while dispersion or variation is significant if the differences between the data and their value is less than their arithmetic mean. Therefore, dispersion or variation of values is a measure of the concentration of data around the average, or proximity to each other. It is known that the homogeneity of data within any statistical population or any sample of important measure that cannot be the researcher can dispense with it by any other measure of the average measure. Among the most famous and important Dispersion measures are:

1. The Range
2. Interquartile Range
3. The Mean Deviation
4. Standard Deviation
5. Variance
6. Standard Error
7. Standard Score

4.2.1 The Range

The range is one of the easiest Measures of dispersion in terms of definition and calculus and gives us a quick idea of how much data dispersed.

The range for a set of data is the difference between the largest value and the lowest value in the group and is denoted by R and calculated as follows:

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

First: In the case of ungrouped data

Example:

Calculate the range from the following data, which represents the number of surgeries performed during a month at city Hospital:

4 ،10 ،8 ،6 ،12 ،16

R=16-4 =12 surgical operation

Second: In the case of grouped data

The range here represents the difference between the upper limit of the upper class (the last) and the minimum of the minimum class (first).

Example:

Calculate the range of the following frequency distribution, which represents the number of traffic accidents for the month of July 2014 in the city of Baghdad.

Age categories	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30
Frequency	4	5	6	2	3

The solution:

The upper limit for the last class = 30

The minimum for the first class = 5

$\therefore R = 30 - 5 = 25$ Year

Range Properties:

1. It is easy to understand and calculated.
2. It is a measure approximation of the dispersion, although simple and easy to understand, and is used only in specific cases.

Disadvantages of Range:

1. Less efficiency if found in the sample abnormal values.
2. It does not depend on the calculation of all data. So excluding any single data in the sample does not affect the range value.
3. It is difficult to estimate from open frequency tables.

4.2.2 Interquartile Range

To eliminate the disadvantages of the absolute range, the group values can be arranged ascending or descending order, and we exclude the smallest quarter of the values "Q1" and the quarter top "Q3" and we are only sufficient with the middle of the set of values. Thus we eliminating of the extreme values. We then take the range of the middle values to measure their dispersion, as shown in the figure below:

25%	25%	25%	25%
Q_1	Q_2	Q_3	

Figure 17 shows the three Interquartile

In general, the Interquartile deviation is calculated according to the nature of the data.

First: In the case of ungrouped data

After the order of ascending or descending values, we find the lower and upper Interquartile by their order.

Lower interquartile order $\boldsymbol{Q_1 = \frac{No.\ of\ value+1}{4}}$

Top interquartile order $\boldsymbol{Q_3 = \frac{3(No.\ of\ value+1)}{4}}$

Interquartile Rang $\boldsymbol{Q_2 = \frac{Q_3-Q_2}{2}}$

Example:

Calculate the interquartile deviation from the following data:

164 ،176 ،172 ،167 ،169 ،160 ،174 ،168 ،165

Solution:

Ascending order values

1	2	3	4	5	6	7	8	9
160	164	165	167	168	169	172	174	179

Lower spring order$= \frac{No.value+1}{4} = \frac{9+1}{4} = \frac{10}{4} = 2.5$

The value of the lower interquartile Q1 = the second value of the ascending values $+\frac{1}{2}$ the difference between the second and third values

$$= 164 + \frac{1}{2}(165 - 164)$$

$$= 164 + 1 * \frac{1}{2}$$

164.5

Top interquartile order $= \frac{\mathbf{3(No.value+1)^3}}{\mathbf{4}} = \frac{\mathbf{3(9+1)^3}}{\mathbf{4}} = \frac{\mathbf{30}}{\mathbf{4}} = \mathbf{7.5}$

The value of the top interquartile Q3 = the seventh value of the ascending values + $\frac{1}{2}$ the difference between the seventh and eighth values.

$$172 + \frac{1}{2}(174 - 172)$$

$$= 172 + \frac{1}{2} \times 2$$

$$= 172 + 1 = 173$$

Interquartile deviation Q=

$$Q = \frac{Q_1 - Q_3}{2}$$

$$= \frac{164.5 - 173}{2}$$

$$= \frac{8.5}{2}$$

$$= 4.25$$

Second: grouped data:

Example:

The following table represents the number of deaths from pneumonia by an age group in Baghdad city for February 2013 finding the interquartile deviation of these deaths.

class	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	3	6	8	13	7	4	3

Solution:

class	Frequency	Accumulation Frequency tables
10 – 20	3	3
20 – 30	6	9
30 – 40	8	17
40 – 50	13	30
50 – 60	7	37
60 – 70	4	41
70 – 80	3	44
Total	44	

Lower interquartile order= $\frac{\mathbf{1}(\boldsymbol{Total\ Frequency}\ )}{\mathbf{4}} = \frac{\mathbf{1(44)}}{\mathbf{4}} = \mathbf{11}$

The value of the lower interquartile Q_1 =

Minimum for the lower spring class + $\frac{\textbf{Lower interquartile order – The previous ascending frequency of the lower interquartile order}}{\textbf{The original frequency of the lower interquartile class}}$ **∗ Length of Category**

$= 30 + 10 \times \frac{9-11}{8}$

=32.5

Top interquartile Order $= \frac{(\ Total\ Frequency)^3}{4}$

$= \frac{(44)3}{4} = \frac{132}{4} = 33$

The value of the Top interquartile $Q_3 =$

Minimum for the lower spring class + $\dfrac{\text{Lower interquartile order – The previous ascending frequency of the lower interquartile order}}{\text{The original frequency of the lower interquartile class}}$ **∗ Length of Category**

$$= 50 + \frac{(33-30)}{7} \times 10 = 54.28$$

Interquartile deviation $Q = \dfrac{Q_3 - Q_1}{2}$

$$= \frac{54.28 - 32.5}{2}$$

$$= \frac{21.78}{2} = 10.89$$ i.e. almost 11 dead

It should be noted here that this measure is easy to calculate and apply and is not affected by the extreme values but it is less accurate than the others of dispersion standards. Because it neglects 50% of the values when calculating it. It is also useful in measuring the dispersion of one or both open frequency distributions.

4.2.3 The Mean Deviation

The mean deviation is defined as the arithmetic mean of absolute values of deviations from the mean. Note that the absolute value of the deviation is the deviation of the value from the mean of the group by neglecting the accompanying algebraic sign, expressed in two vertical lines placed around the number. It is calculated as follows:

First: Ungrouped data

Formula for population $\boldsymbol{M.D} = \frac{\sum|x_i - \mu|}{N}$

Formula for sample $\boldsymbol{M.D} = \frac{\sum|x_i - \overline{x}|}{n}$

In general, the mean deviation is calculated by using the following steps:

1. Find the arithmetic mean.
2. We calculate the deviation values from the arithmetic mean by neglecting the signal.
3. Sum the deviations and divide them into their number (n) to get the mean deviation.

Example:

Find the Mean deviation from the following data.

8, 12, 9, 6, 15, 7, 13

Solution:

$\overline{x} = \frac{\sum x_i}{n} = \frac{70}{7} = 10$

1. We extract the arithmetic mean.
2. We calculate the deviations of values from their arithmetic mean.

x_i	$x_i - \bar{x}$	$\|x_i - \bar{x}\|$
8	-2	2
12	2	2
9	-1	1
6	-4	4
15	5	5
7	-3	3
13	3	3
$\sum$70	0	20

Then average mean deviation

$$\therefore M.D = \frac{\sum \left| x_i - \bar{x} \right|}{n} = \frac{20}{7} = 2.85$$

Second: For grouped data

For population $M.D = \dfrac{\sum f_i \left| x_i - \mu \right|}{\sum f_i}$

For sample $M.D = \dfrac{\sum f_i \left| x_i - \bar{x} \right|}{\sum f_i}$

Example:

Calculate the average deviation from the following frequency distribution, which represents the number of fractures that have been treated at the Children's Hospital in 2014.

class / age	4-6	6-8	8-10	10-12	12-14	14-16
Frequency	3	5	8	6	7	2

Solution:

class / age	frequency Fi	Class center xi	Fixi	$x_i - \overline{x}$	$\lvert x_i - \overline{x}\rvert$	$F_i\lvert x_i - \overline{x}\rvert$
4-6	3	5	15	5-	5	15
6-8	5	7	35	3-	3	15
8-10	8	9	72	1-	1	8
10-12	6	11	66	1	1	6
12-14	7	13	91	3	3	21
14-16	2	15	30	5	5	10
Total	31		309	0		75

$$\overline{X} = \frac{\sum f_i x_i}{\sum f_i} = \frac{309}{31} = 10$$

$M.D = \frac{\sum f_i|x_i - \overline{x}|}{\sum f_i} = \frac{75}{31} = 2.4$ Mean deviation value.

One of the disadvantages of mean deviation is met as a measure of dispersion that it is not subject to algebraic processing because it omits algebraic signals. So, it does not distinguish between negative and positive deviations, because it is statistically useless at all.

4.2.4 Standard Deviation

Most commonly used measure to describe variability deviation (SD). This measure is one of the most important measures of dispersion and the most common and widely used.

Standard Deviation is the squared differences of each observation from the mean. It is symbolized by (σ) in the case of the population and (S) in the case of the sample.

It is generally calculated by follow the following steps:

1. Find the arithmetic mean of the population or the sample.
2. Find deviations of values from the arithmetic mean.
3. Quadrature deviations of values from the arithmetic mean.
4. Sum the squares of deviations and find their mean and then their root to obtain the standard deviation.

The standard deviation is calculated for the data that is not grouped and grouped according to the following formulas:

First: For ungrouped data

Formula for Standard Deviation of the population

$$\boldsymbol{\sigma = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{N}}}$$

Or

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Formula for Standard Deviation of the sample:

$$\boldsymbol{S = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}}}$$

Or

$$\boldsymbol{S = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}}$$

Example:

Find the standard deviation from the following data:

3, 4, 9, 5, 7, 8, 6

Solution:

X_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
3	3-	9
4	2-	4
9	3	9
5	1-	1
7	1	1
8	2	4
6	0	0
Total 42	0	28

$$\overline{X} = \frac{\sum x_i}{n} = \frac{42}{7} = 6$$

$$S = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{28}{6}} = \sqrt{4.66}$$

$$\therefore S = 2.16$$

Another solution which is shorter than above,

Xi	Xi2
3	9
4	16
9	81
5	25
7	49
8	64
6	36
Total 42	280

$$S=\sqrt{\frac{\sum x_i{}^2-\frac{\left(\sum x_i\right)^2}{n}}{n-1}}=\sqrt{\frac{280-\frac{(42)^2}{7}}{6}}$$

$$S=\sqrt{\frac{280-252}{6}}=\sqrt{4.66}$$

$\therefore S=2.16$ The standard deviation value.

And we can get the value of variation by squaring the value of the standard deviation as $(2.16)^2$ which is equal to 4.66.

Second: For classified data

Formula for Standard Deviation of the population

$$\sigma=\sqrt{\frac{\sum f_i x_i{}^2-\frac{\left(\sum f_i x_i\right)^2}{\sum f_i}}{\sum f_i}}$$

Or

$$\sigma=\sqrt{\frac{\sum f_i\left(x_i-\mu\right)^2}{\sum f_i}}$$

Formula for Standard Deviation of the sample:

$$S=\sqrt{\frac{\sum f_i x_i{}^2-\frac{\left(\sum f_i x_i\right)^2}{\sum f_i}}{\sum f_i-1}}$$

Or

$$S=\sqrt{\frac{\sum f_i\left(x_i-\bar{x}\right)^2}{\sum f_i-1}}$$

Example:

Find the standard deviation from the following frequency distribution schedule, which represents 100 patients with atherosclerosis classified by age:

class / age	60-62	63-65	66-68	68-71	72-74
frequency	5	18	42	27	8

Solution:

class	Frequency Fi	Class center xi	Fixi	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$f_i(x_i - \overline{x})^2$
60-62	5	61	305	6.45-	41.602	208.012
63-65	18	64	1152	3.45-	11.902	214.245
66-68	42	67	2814	0.45-	0.202	8.505
69-71	27	70	1890	2.55	6.502	175.567
72-74	8	73	584	5.55	30.802	246.420
Total	100		6745			852.750

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{6745}{100} = 67.45$$

$$S = \sqrt{\frac{\sum f\left(x - \bar{x}\right)^2}{\sum f_i - 1}} = \sqrt{\frac{852.750}{99}}$$

$$S = \sqrt{8.613}$$

$\therefore S = 2.93$ The standard deviation value

Another solution which is shorter than above,

class	Frequency F_i	Class center x_i	x_i^2	$F_i x_i$	$f_i x_i^2$
60-62	5	61	3721	305	18605
63-65	18	64	4096	1152	73728
66-68	42	67	4489	2814	188538
71-69	27	70	4900	1890	132300
72-74	8	73	5329	584	42632
Total	100			6745	455803

$$S = \sqrt{\frac{\sum f_i x_i^{\,2} - \frac{\left(\sum f_i x_i\right)^2}{\sum f_i}}{\sum f_i - 1}}$$

$$S = \sqrt{\frac{455803 - \frac{(6754)^2}{100}}{100 - 1}} = \sqrt{8.6136}$$

$\therefore S = 2.93$ The standard deviation value.

4.2.5 Variance

Variation is one of the best and most widely used dispersion measures, especially in applied fields. The variation works to measure the mean dispersion of group values around its arithmetic mean.

The variance is defined as the sum of squares of deviations of values divided by their number, denoting the variance of the population with σ^2 and the variance of the sample with S^2, and calculated according to the nature of the data and according to the following:

First: Ungrouped data

For population $\boldsymbol{\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}}$

For sample $\boldsymbol{S^2 = \frac{\sum(x_i-\overline{x})^2}{n-1}}$

Example:

Find the following variations: 3, 7, 9, 5, 8, 4, 6

Solution:

X_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
3	3-	9
7	1	1
9	3	9
5	1-	1
8	2	4
4	2-	4
6	0	0
$\sum$42	0	28

$$\bar{x} = \frac{\sum x_i}{n} = \frac{42}{7} = 6$$

$$S^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}$$

Variance value. $S^2 = \frac{28}{6} = 4.66$

The root of the variance value, we can obtain the standard deviation value of $\sqrt{4.66}$, which is equal to 2.16.

Second: For grouped data

For population $\sigma^2 = \frac{\sum f_i (x_i - \mu)^2}{N}$

For sample $S^2 = \frac{\sum f_i \left(x_i - \bar{x}\right)^2}{\sum f_i - 1}$

Example:

Find the variance of the following frequency distribution which represents 13 patients with German measles.

class / age	4 - 8	8 -12	16-21	16-20	20-24
Frequency	3	2	2	4	2

Solution:

Class	Frequency Fi	Class center xi	Fixi	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$f_i(x_i - \overline{x})^2$
4 – 8	3	6	18	8-	64	192
8 -12	2	10	20	4-	16	32
16-21	2	14	28	0	0	0
16-20	4	18	72	4	16	64
20-24	2	22	44	8	64	128
Total	13		182	0		416

$$\overline{X} = \frac{\sum f_i x_i}{\sum f_i} = \frac{182}{13} = 14$$

$$S^2 = \frac{\sum f_i \left(x_i - \overline{x}\right)^2}{\sum f_i - 1} = \frac{416}{12} = 34.66$$

4.2.6 Standard Error

The researcher may take different random samples from a homogeneous statistical society. Their computational modes will certainly be different. So their standard deviations calculated from these samples will also be different as a result of the sampling error.

The standard deviation, which measures the dispersion of the arithmetic mean of these samples from the arithmetic mean of the statistical population from which these samples are taken, is called the "standard" standard error, which happens by coincidence, and not the researcher, and it is denoted by S. E,

calculated by dividing the standard deviation on the root Quadratic number of sample units i.e. $S.E = \frac{s}{\sqrt{n}}$

It should be noted here that the larger the samples taken from the population, the smaller the variance is expected to be between the rates of the samples of small sizes.

Example:

A medical researcher took two equal groups in their standard deviation of 2 and took samples from the first population was 16 and from the second population size 64, the standard error of the first population equals

$$S\bar{Y} = \frac{s}{\sqrt{n}} = \frac{2}{\sqrt{16}} = \frac{2}{4} = 0.5$$

And the standard error of the second society equals

$$S\bar{Y} = \frac{s}{\sqrt{n}} = \frac{2}{\sqrt{64}} = \frac{2}{8} = 0.25$$

From the above results, it is clear to us that the standard deviation of the sampling rates from the second population is equal to half the standard deviation of the sample rates of the first population.

So it can be said that the sample $(\bar{Y})$rate in the representation of the population rate "μ" accuracy depends heavily on increasing the size of the sample taken from that population.

The standard error has multiple uses, including:

1. The standard error is a measure of the degree of dependence on the sample mean. The lower the value of the standard error, the more reliable it is and vice versa.
2. The standard error is used in the fields of standardization, production quality control, and moral tests.
3. Gives the researcher a general idea of the average "population" under study and research.

4.2.7 Coefficient of Variation

It is defined as the standard deviation expressed as a percentage of the arithmetic mean, and is denoted by C.V and calculated according to the following formula:

For population $\boldsymbol{C.V = \frac{\sigma}{\mu} \times 100}$

For sample $\boldsymbol{C.V = \frac{S}{\bar{X}} \times 100}$

And there is another picture of the coefficient of variation if it is not possible to calculate the standard deviation (σ) and the arithmetic mean ($\bar{X}$) as in the case of open class or the existence of extreme anomalies, in which case the coefficient of variation can be calculated from the following formula:

$$\textbf{coefficient of variation } (\boldsymbol{C.V}) = \frac{\boldsymbol{standard\ deviation}}{\boldsymbol{Mean}} \times 100$$

The coefficient of variation is used to compare two or more groups. The smaller the coefficient of variation for a group, the more homogeneous the group values are and vice versa. This coefficient is the best types of dispersion measurements because it

is based on the best measures of central tendency on the one hand, and the best measurements of dispersion on the other hand. In addition, the resulting from his account, which is a percentage of an abstract unit of measurement, such as weights, sizes, lengths, ages, etc., help a researcher at the possibility of measuring the amount of dispersion and scattering qualities which are vary in units of measurement.

In general, it can be said that the upper limit of the coefficient of variation, which can be accepted in the field trials should be no more than 20% either in laboratory tests or experiments controlled. Then the value of this parameter does not exceed 15%.

Example:

The following data were extracted from a study on the effect of certain treatment on children infected with Giardia parasites, any of these measures for children infected are less dispersed and more responsive to treatment.

Children	measuring unit	Mean arithmetic $\overline{X}$	standard deviation σ
Age	year	8	1.4
Weight	K.g	22	3
Length	cm	70	15

Solution: By applying the formula of the coefficient of variation to the society σ we get

1. For age $C.V = \frac{1.4}{8} \times 100 = 17.5\%$
2. For weight $C.V = \frac{3}{22} \times 100 = 13.6\%$
3. For length $C.V = \frac{15}{70} \times 100 = 21.4\%$

From the above results, it can be said that weight is less dispersion and more responsive to treatment of age and height.

4.2.8 Standard Score

Is a quantitative expression that shows us the deviation of the degree of "observation" from the arithmetic mean by using the standard deviation as a measure. It determines the location of the raw grade of the arithmetic mean. The direction is determined by the sign "+ or -" if positive, it is higher than the arithmetic mean and otherwise if the signal is negative. The distance means greater value, the larger the value is away from the arithmetic mean and vice versa.

The Standard Score is used to compare two or more different values in terms of unit of measurement and is denoted by the letter Z and is calculated as the following:

$$\boldsymbol{Z = \left(\frac{x - \overline{x}}{S} \right)}$$

Example:

The grades of one of the students of the fifth stage of the College of Pharmacy for two academic years according to the following:

	Treatments	Industrial Pharmacy
Student degree X	82	88
The arithmetic mean of students $\bar{x}$	74	79
Standard Deviation of Students S	10	15

Find in which subjects was the student's achievement better for the stage level?

Standard Score treatments

$$Z = \frac{x - \overline{x}}{S} = \frac{82 - 74}{10} = 0.8$$

The standard score for industrial pharmaceuticals

$$Z = \frac{x - \overline{x}}{S} = \frac{88 - 79}{15} = 0.6$$

Since the value of Z for the treatment material is greater than the Z value of industrial pharmaceuticals. So, it can be said that the student's achievement of the treatment material better than the collection in industrial pharmacy. However it can be said that the student's achievement of the treatment material better than the result in industrial pharmacy.

Of all this, it can be said that the benefits of the standard that it gives us a picture of the location of the degree of the arithmetic mean and therefore we can identify the location of the student for his colleagues.

Chapter Five

Correlation Analysis and Regression

Chapter Five

Correlation Analysis & Regression

5. Correlation Analysis and Regression

In the previous chapters, we dealt with the different and statistical methods of data collection, classification, and the extraction of some descriptive statistical measures. Through these methods, it is possible to give a clear idea of the nature of such data, including averages and different dispersion measurements.

All of these methods are based on data collected from only one variable (Y) or (X), but in many cases, the researcher faces cases that require the study of two or more variables at the same time to identify the nature of the relationship that these variables related to their type and strength.

In general, the relationship is statistically measured by using two main methods: Correlation and Regression. These two methods are similar in so many ways, but at the same time are different in many ways. Correlation measures the degree and direction of the relationship between the two variables while regression examines the relationship between the variables through a mathematical equation by which one can interpret, estimate or predict one of the variables through the other.

5.1 Correlation Analysis

Correlation is a descriptive analysis tool to determine the relationship between two independent variables (x_1, x_2), each representing a particular phenomenon between one independent variable (X) and a dependent variable (Y) or between (Y) and a set of independent variables $(x_1 \ldots . x_n)$.

Moreover, correlation is to measure the degree of relationship between these variables, i.e. whether these variables are associated to a linear or non-linear relationship in one form or another, and determine the direction of that relationship in terms of whether it is reverse or positive. The basis for this relationship is called the correlation coefficient, which is a quantitative measure of the strength of this relationship. It is denoted in the case of the community by (ρ) and in the case of the sample by "r" and its value is between (0 and ± 1). In many applied aspects, we deal with the data of a sample drawn from the statistical society. Therefore, we will focus on calculating the correlation coefficient of the sample "r" as an estimation value of the social correlation coefficient taken from the sample.

The correlation coefficient in general focuses on two main points:

1. The type of relationship that takes three forms depending on the signal correlation coefficient is:

A. The positive relationship between two variables i.e. "$r > 0$" the increase of one of the variables (x) is accompanied by an increase in the second variable (y) or vice versa, the deficiency of one of the variables (x) is accompanied by a deficiency in the second variable (y).

B. The negative correlation between the two variables i.e. "$r < 0$" the increase in the first variable (x) is accompanied by a decrease in the second variable (y) or, conversely, any deficiency in the first variable (x) is accompanied by an increase in the second variable (y).

C. The loss of relationship between the two variables i.e. "$r = 0$" since any increase or decrease in the first variable "x" does not lead in any change in the second variable "y".

2. The strength of the relationship can be judged by how close or far from the value of r $(1\mp)$, since the correlation coefficient value is in the range $(-1 < r < 1)$. Moreover, the closer the value of the correlation coefficient from zero indicates the absence of a linear relationship between them, in the sense that there may be a relationship but it is not linear.

Some statisticians have rated the strength of the relationship as follows:

Positive correlation					Negative correlation				
Very strong	**strong**	**Average**	**Weak**	**Very weak**	**Very weak**	**Weak**	**Average**	**strong**	**Very strong**
0.9	**0.7**	**0.5**		**0.3**	**- 0.3**	**0.5 -**		**0.7-**	**0.9-**
$+1$ *Full correlation*				$N0$ *correlation*	-1 *Full correlation*				

Figure (26)

degrees of strength of the correlation coefficient

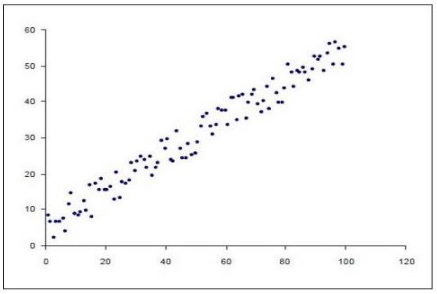
5.1.1 Scatter diagram to determine the nature of the direction of the correlation

The determination of the correlation and diagram of the scatter diagram depend on the existence of two variables, one of which is independent X and the other dependent "Y". For each value of variable X, there is a corresponding value of variable Y.

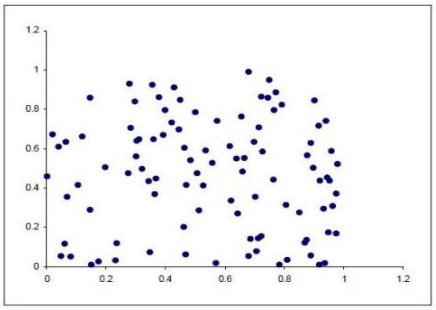
After all data pairs values of these two variables. Once values are paired from the data to this two-variable collection, we can graphically represent them in a Scatter diagram. A Scatter diagram allows a researcher to identify the nature of the relationship between these two variable through visual inspection. If the values are displayed in a “divergent” manner, this indicates a weak relationship between the two variables. If the values are represented as a straight line and can be defined as “convergent”, then there is a strong relationship between the two variables.

This information is highly important in statistical analysis and decision-making. When variables show a high degree of

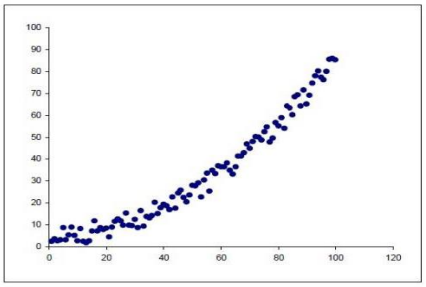
correlation. It can assumed that there is a relationship between the variables. In general, this relationship can take one of the following forms:



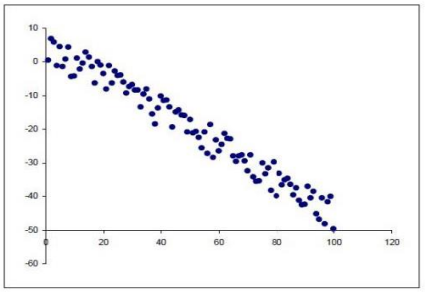
Positive linear correlation



No correlation



Non-linear correlation



Negative linear correlation

Figure 27 Examples of Scatter diagram

5.1.2 Types of correlation

The correlation that represents the phenomena that can be expressed quantitatively "numerically" can be divided into three types depending on the number of variables it includes:

1. Simple correlation:

This type of correlation is interested with studying the relationship between two variables, one being independent (x) and the other dependent (y).

2. Multiple correlation:

This type is interested with studying the relationship between more than two independent variables and a dependent variable.

3. Partial correlation:

This type of correlation is interested with studying the relationship between a pair of variables only among a group of other variables whose effect is stabilized by exclusion or isolation. The difference between simple correlation and partial correlation is that the first measures the strength and direction of the relationship between two variables within the effects of other variables, while the second measures the strength of the relationship and its direction between two variables after excluding the effect of other variables.

For example, if we have three variables Y, X_2, X_1, it is possible to measure the partial correlation between any two and isolate the effect of the third variable by using the partial correlation coefficient.

5.1.3 Measuring Correlation

After using the diagram of the propagation, the researcher can define the initial type of correlation between variables as

"positive or negative linear correlation or non-linear correlation". To measure the type and strength of this relationship, we use special measurements called correlation standards, which are divided into two major basic types:

5.1.3.1 Correlation coefficient of measured phenomena

This includes the study of the relationship between quantitative phenomena and "digital" which includes all phenomena that can be expressed in digital form, including height and weight, age, births and deaths, and other phenomena that can be expressed digitally. It is divided into the following types:

1. Simple correlation coefficient.
2. Multiple correlation coefficient.
3. Partial correlation coefficient.

5.1.3.1.1 Simple correlation coefficient

The main objective of the study of correlation analysis is to measure the linear correlation between quantitative variables with continuous variables. While the linear coefficient is a measurement of the strength of the linear relationship between the variables X and Y and measures the extent of the change in the value of Y if the increased value of X. When the value of Y increases by increasing the value of X, there will be a positive correlation, and when the value of Y decreases by increasing the value of X, then there is a negative correlation. While the value of Y does not affect the increase or decrease of the X value this

indicates that there is no correlation between the two variables. The Pearson correlation coefficient is one of the most important and most powerful correlations, especially when the relationship between the two variables is linear. This measurement is often used to study the relationship between treatment and disease, the relationship between age and blood pressure, the relationship between smoking and individual health, the relationship between child mortality and malnutrition and others.

The basic formula of Pearson correlation coefficient for the correlation of the sample is:

$$r = \frac{\sum Y_i X_i - \frac{(\sum Y_i)(\sum X_i)}{n}}{\sqrt{\sum Y_i^2 - \frac{\sum (Y_i)^2}{n}} \sqrt{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}}$$

Or

$$r = \frac{\sum Y_i X_i - n\overline{XY}}{\sqrt{\sum Y_i^2 - n\overline{Y}} \sqrt{\sum X_i^2 - n\overline{X}}}$$

1. **Calculate the simple correlation coefficient of the original data:**

Example:

The following data represents the number of doctors working in a hospital and the number of visits of doctors to their "patients Tour".

No of doctors (Y) :	**8**	**5**	**7**	**2**	**9**	**11**	**6**	**8**	**6**
No of visits (X) :	**6**	**2**	**5**	**4**	**5**	**6**	**3**	**5**	**4**

Find the simple correlation coefficient (Pearson) between the number of doctors and the number of visits to their "patients Tour".

Solution:

Yi	Xi	$YiXi$	Y_i^2	X_i^2
8	**6**	**48**	**64**	**36**
5	**2**	**10**	**25**	**4**
7	**5**	**35**	**49**	**25**
8	**4**	**32**	**64**	**16**
9	**5**	**45**	**81**	**25**
11	**6**	**66**	**121**	**36**
6	**3**	**18**	**36**	**9**
8	**6**	**48**	**64**	**36**
6	**4**	**24**	**36**	**16**
$\sum$68	**44**	**357**	**540**	**242**

$$r = \frac{\sum Y_i X_i - \frac{\left(\sum Y_i\right)\left(\sum X_i\right)}{n}}{\sqrt{\sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}}\sqrt{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}}$$

$$r = \frac{357 - \frac{(68)(44)}{9}}{\sqrt{540 - \frac{(68)^2}{9}}\sqrt{242 - \frac{(44)^2}{9}}}$$

$$r = \frac{357 - 332.44}{\sqrt{540 - 513.77}\sqrt{242 - 215.11}}$$

$$r = \frac{24.56}{\sqrt{26.23}\sqrt{26.89}} = \frac{24.56}{(5.121)(5.185)} = \frac{24.56}{26.55} = 0.92$$

There is a strong positive relationship between tour and doctors.

2. Calculate the simple correlation coefficient of the standard scores:

From the previous example data, the simple correlation can be calculated in this way and as follows:

Y	X	$Y - \bar{Y}$	$X - \bar{X}$	$(Y - \bar{Y})^2$	$(X - \bar{X})^2$	Zy	Zx	ZyZx
8	6	0.45	1.12	0.2025	1.2544	0.2451	0.6060	0.15
5	2	-2.55	- 2.88	6.5025	8.29	- 1.4115	- 1.5757	2.22
7	5	-0.55	0.12	0.3025	0.01	- 0.3066	0.0606	- 0.02
8	4	0.45	- 0.88	0.2025	0.77	0.2454	- 0.4848	- 0.12
9	6	1.45	1.12	2.1025	1.25	0.7978	0.6060	0.48
11	8	3.45	3.12	11.9025	9.73	1.9025	1.6969	3.23
6	3	-1.55	- 1.88	2.4025	3.53	- 0.8592	-1.0303	0.89
8	6	0.45	1.12	0.2025	1.25	0.2454	0.6060	0.15
6	4	-1.55	- 0.88	2.4025	0.77	- 0.8592	- 0.4845	0.42
∑68	44			26.2225	26.89			7.40

$$\bar{Y} = \frac{\sum yi}{n} = \frac{68}{9} = 7.55$$

$$\bar{X} = \frac{\sum xi}{n} = \frac{44}{9} = 4.88$$

$$sy = \sqrt{\frac{\sum(yi - \overline{y})^2}{n-1}} = \sqrt{\frac{26.2225}{8}} = \sqrt{3.2778} = 1.81$$

$$sx = \sqrt{\frac{\sum(xi - \overline{x})^2}{n-1}} = \sqrt{\frac{26.89}{8}} = \sqrt{3.361} = 1.83$$

$$r = \frac{ZyZx}{n-1} = \frac{7.40}{8} = 0.92$$

This shows that a strong and positive relationship between the number of doctors working in hospitals and the frequency of visits they make to their patients.

5.1.3.1.2 Multiple correlation coefficient

In most theoretical, applied and planning scientific studies, the correlation relationship is not dependent on two variables, one representing the dependent variable (Y) and the other representing the independent variable (X), but extends to a number of independent variables that affect one or the other dependent variable. For example, population fertility depends on the number of women of childbearing age, education, income, health status, profession, religion etc. Thus, the study of the relationship between the dependent variable (Y) and the set of independent variables at the same time is what is called the multiple correlation, which can be measured by the multiple correlation coefficient, which measures the strength of the relationship between more than two random variables related to the Multivariate distribution. This parameter is denoted by the letter

R. The calculation of the value of this parameter is an extension to the calculate of the value of the simple correlation coefficient (r) by adding other independent variables. Moreover, the correlation coefficient value is also between zero and $1\pm$.

The more its value is close to the one, indicates the strength of the relationship between the dependent variable and the independent variables conversely, if its value is close to zero. We will limit ourselves in our study of this parameter to the linear relationship between the three variables Y, X1 and X2 to get the following formulas:

$$r_{YX_1} = \frac{n\sum YX_1 - \sum Y \sum X_1}{\sqrt{N\sum Y_I^2 - \left(\sum Y_I\right)^2}\sqrt{n\sum X_1^2 - \left(\sum X_1\right)^2}}$$

$$r_{YX_2} = \frac{n\sum YX_2 - \sum Y \sum X_2}{\sqrt{n\sum Y_1^2 - \left(\sum Y_1\right)^2}\sqrt{n\sum X_2^2 - \left(\sum X_2\right)^2}}$$

$$r_{X_1X_2} = \frac{n\sum X_1X_2 - \sum X_1 \sum X_2}{\sqrt{n\sum X_1^2 - \left(\sum X_1\right)^2}\sqrt{n\sum X_2^2 - \left(\sum X_2\right)^2}}$$

Therefore, the final formula for calculating the multiple correlation coefficient is:

$$R_{YX_1X_2} = \sqrt{\frac{r^2{}_{YX_1} + r^2{}_{YX_2} - 2r_{YX_1}r_{YX_2}r_{X_1X_2}}{1 - r^2{}_{X_1X_2}}}$$

Example:

Find the value of the multiple correlation coefficient from the following table:

Y	X1	X_2
2	**1**	1
3	**6**	8
2	**5**	2
1	**7**	6
4	**10**	8

Solution:

It is necessary to obtain the multiple correlation coefficient between Y and variablesX_1 and X_2.

Y	X_1	X_2	YX_1	YX_2	X_1X_2	Y_i^2	X_1^2	X_2^2
2	1	1	2	2	1	4	1	1
3	6	8	18	24	48	9	36	64
2	5	2	10	4	10	4	25	4
1	7	6	7	6	42	1	49	36
4	10	8	40	32	80	16	100	64
$\sum$12	29	25	77	68	181	34	211	169

$$r_{YX_1} = \frac{n\sum YX_1 - \sum Y \sum X_1}{\sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}\sqrt{n\sum X_1^2 - \left(\sum X_1\right)^2}}$$

$$= \frac{5(77) - (12)(29)}{\sqrt{5(34) - (12^2)}\sqrt{5(211) - (29)^2}}$$

$$= \frac{383 - 348}{\sqrt{(170 - 144)}\sqrt{(1055 - 841)}} = \frac{37}{\sqrt{26}\sqrt{214}} = \frac{37}{74.5} = 0.49$$

$$r_{YX_2} = \frac{n\sum YX_2 - \sum Y \sum X_2}{\sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}\sqrt{n\sum X_2^2 - \left(\sum X_2\right)^2}}$$

$$= \frac{5(68) - (12)(25)}{\sqrt{5(34) - (12)^2}\sqrt{5(169) - (25)^2}}$$

$$= \frac{340 - 300}{\sqrt{(170 - 144)}\sqrt{(845 - 625)}} = \frac{40}{\sqrt{26}\sqrt{220}} = \frac{40}{75.6} = 0.53$$

$$r_{X_1X_2} = \frac{n\sum X_1X_2 - \sum X_1 \sum X_2}{\sqrt{n\sum X_1^2 - \left(\sum X_1\right)^2}\sqrt{n\sum X_2^2 - \left(\sum X_2\right)^2}}$$

$$r_{X_1X_2} = \frac{5(181) - (29)(25)}{\sqrt{\sum 5(211) - (29)^2}\sqrt{5(169) - (25)^2}}$$

$$= \frac{905 - 725}{\sqrt{(1055 - 841)}\sqrt{842 - 625}} = \frac{180}{\sqrt{214}\sqrt{220}} = \frac{180}{217} = 0.83$$

$$R_{YX_1X_2} = \sqrt{\frac{r^2{}_{YX_1} + r^2{}_{YX_2} - 2r_{YX_1}r_{YX_2}r_{X_1X_2}}{1 - r^2{}_{X_1X_2}}}$$

$$= \sqrt{\frac{(0.49)^2 + (0.53)^2 - 2(0.49)(0.53)(0.83)}{1 - (0.83)^2}}$$

$$= \sqrt{\frac{0.240 + 0.280 - 2(0.431)}{0.312}}$$

$$= \sqrt{\frac{0.52 - 0.431}{0.312}}$$

$$= \sqrt{\frac{0.089}{0.312}} = \sqrt{0.285} = 0.53$$

With the correlation coefficient's value of 0.53, as shown above, it can be said that there is a positive relationship between the dependent variable (Y) and the independent variables (X1 and X2) but a weak relationship.

5.1.3.1.3 Partial Correlation

We already know that multiple correlation is the study between the dependent variable and a set of independent variables at the same time. Partial correlation, however, examines the relationship between only two of these variables when the rest of the other variables are constant by isolating them or statistically excluding their effect.

For example, if we have a specific phenomenon (Y) and there is a set of independent variables (X_1, X_2, X_3, X_4) that have a partial effect on this phenomenon, the partial correlation between the two variables X_1 and X_3 will exclude the effect of the two variables X_2, X_4. The main purpose of this exclusion is to determine the nature of the relationship between them and the usefulness of the survival of one or both of them depending on their degree of influence on the Y variable.

The value of the partial correlation coefficient is calculated according to the following steps:

1. Find simple correlation coefficients (r) between the studied variables.

2. Find the partial correlation coefficient of the following formula:

A. The formula of the partial correlation coefficient between Y and X_1 with X_2 stability

$$R_{YX_1X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1-r^2_{X_1X_2}}\sqrt{1-r^2_{YX}}}$$

B. The formula of the partial correlation coefficient between Y and X_1 with X_2 stability

$$R_{YX_2X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{1-r^2_{X\ X_2}}\sqrt{1-r^2_{YX_1}}}$$

Example:

A researcher selected a sample of six workers working in one pharmaceutical factory and collected information about the monthly production of these six workers and the actual duration of service for them in the factory. The sample data is as follows:

Y amount of production (tons) per worker:	1	4	4	5	8	9
X_1 Design capacity of production (tons):	3	5	8	12	12	15
X_2 actual duration of service per worker (year):	2	4	6	8	9	9

Find partial correlation coefficients and indicate which ones are more effective on production.

Solution: The calculation of partial correlation coefficient requires first to calculate the simple correlation coefficients between the three variables, and this requires the following table:

Y	X1	X2	YX1	YX2	X1X2	Y2	X_1^2	Y_2^2
1	3	2	3	2	6	1	9	4
4	5	4	20	16	20	16	25	16
4	8	6	32	24	48	16	64	36
5	12	8	60	40	96	25	144	64
8	12	9	96	72	108	64	144	81
9	15	9	135	81	135	81	225	81
$\sum$31	55	38	346	235	413	203	611	282

$$r_{YX_1} = \frac{\sum X_1 Y - \frac{\sum X_1 \sum Y}{n}}{\sqrt{\sum X_1^2 - \frac{\left(\sum X_1\right)^2}{n}}\sqrt{\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}}}$$

$$r_{YX_1} = \frac{346 - \frac{(55)(31)}{6}}{\sqrt{611 - \frac{(55)^2}{6}}\sqrt{203 - \frac{(31)^2}{6}}} = \frac{62}{67.8} = 0.91$$

$$r_{YX_2} = \frac{\sum X_2 Y - \frac{\sum X_2 \sum Y}{n}}{\sqrt{\sum X_2^2 - \frac{\left(\sum X_2\right)^2}{n}}\sqrt{\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}}}$$

$$r_{YX_2} = \frac{235 - \frac{(38)(31)}{6}}{\sqrt{282 - \frac{(38)^2}{6}}\sqrt{203 - \frac{(31)^2}{6}}} = \frac{38.67}{42} = 0.92$$

$$r_{X_1X_2} = \frac{\sum X_1X_2 - \frac{\sum X_1 \sum X_2}{n}}{\sqrt{\sum X_1^2 - \frac{(\sum X_1)^2}{n}}\sqrt{\sum X_2^2 - \frac{(\sum X_2)^2}{n}}}$$

$$= \frac{413 - \frac{(55)(38)}{6}}{\sqrt{611 - \frac{(55)^2}{6}}\sqrt{282 - \frac{(38)^2}{6}}} = \frac{64.7}{66.4} = 0.97$$

After finding the simple correlation coefficients for the three variables of $r_{X_1} = 0.91$ ، $r_{X_2} = 0.92$ $r_{X_1X_2} = 0.97$ we start applying partial correlation law X_2 is fixed to $YX_1.X_2$ and that $X_1$is fixed to $YX_2.X_1$.

$$r_{YX_1X_2} = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{\sqrt{1-(r_{X_1X_2})^2}\sqrt{1-(r_{YX_2})^2}}$$

$$= \frac{0.91 - (0.92)(0.97)}{\sqrt{1-(0.97)^2}\sqrt{1-(0.92)^2}} = \frac{0.02}{0.097} = 0.20$$

$$R_{YX_2X_1} = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{\sqrt{1-(r_{X_1X_2})^2}\sqrt{1-(r_{YX_1})^2}}$$

$$= \frac{0.92-(0.91)(0.97)}{\sqrt{1-(0.97)^2}\sqrt{1-(0.91)^2}} = \frac{0.04}{0.100} = 0.4$$

From the results of the partial correlation above, it can be said that the second variable, "actual duration of service per worker ", is more important in influencing the amount of pharmaceutical production (Y) than the design power of the factory.

5.1.3.2 The correlation coefficient of unmeasured phenomena

There are some phenomena that cannot be measured quantitatively (digitally) and may be in the form of descriptions or in the form of ranking such phenomena, the social situation, the health status of the members of society, smoking, saturation, intelligence and others. These phenomena do not have a digital scale to measure and all we can do is to classify the members of society in terms of health status to, for example, the classes graded, either from top to bottom or vice versa. Another example might include the ranking of health from poor to excellent. Such an approach can repeated in cases with similar phenomena.

The most important of these measures are:

1. Spearmiain's Rank correlation coefficient.
2. Kendall's Rank correlation coefficient.

5.1.3.2.1 Spearmiain's Rank correlation coefficient

It is one of the most important and commonly used in the following cases:

1. If one or both variables are of the ordinal variables that can be arranged in ascending and descending order.
2. If one or both variables are not follow normal distribution, or in the case of non-scientific data, and are considered as an alternative to the Pearson correlation coefficient.

Spearman's work has been recognized in many studies that rely on non-quantitative data, and when used to produce quantitative data, the strength in the correlation measurement is not less than 0.91 of Pearson's strength.

Spearman devised an equation to calculate the correlation coefficient of ranks, which is symbolized by the symbol (r_s) and is stated as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

Where:

n = represents the number of rank pairs.

di = represents the difference between the levels of the first variable (Y) and the levels of the second variable (X). 1, 6 represent theoretical constants apart from n and di.

Example A)

The following table estimates the efficiency of the performance of five doctors working in the medical city of Baghdad and the duration of their employment:

Performance efficiency (y)	very good	Weak	Excellent	Good	Average	Acceptable
(X) Duration of service	15	12	20	5	10	8

Find the correlation coefficient between the efficiency of performance and the academic achievement value. What are the conclusion?

Solution:

Order estimates ascending as follows:

Y Performance efficiency	Weak	Acceptable	Average	Good	Very good	Excellent
Ranks	1	2	3	4	5	6
X Duration of service	5	8	12	10	15	20
Ranks	1	2	3	4	5	6

Y	X	rank y	rank x	di=yi-xi	di2
Very good	15	5	5	0	0
Weak	12	1	3	-2	4
Excellent	20	6	6	0	0
Good	5	4	1	3	9
Average	10	3	4	-1	1
Acceptable	8	2	2	0	0
$\sum$				0	14

$$r_s = 1 - \frac{6\sum di^2}{n(n^2-1)} = 1 - \frac{6(14)}{6(36-1)} = 1 - \frac{84}{210} = 1 - 0.4 = 0.6$$

That is, there is a positive and average correlation between efficiency of performance and duration of service.

Example B)

The following data represents the age and blood pressure of five people who visited a hospital in Baghdad.

Persons :	**1**	**2**	**3**	**4**	**5**
Blood Pressure:	**150**	**200**	**120**	**180**	**176**
Age:	**56**	**70**	**43**	**68**	**61**

Find: the correlation coefficient (Spearmiain) between age and blood pressure.

Solution:

We produce the following table to help the calculation of Spearmiain correlation coefficients.

Blood Pressure Y	**Age X**	**order X**	**order Y**	**di=yi-yi**	d_i^2
150	**56**	**4**	**3**	**1-**	**1**
200	**70**	**1**	**1**	**0**	**0**
120	**43**	**5**	**4**	**1-**	**1**
180	**68**	**2**	**2**	**0**	**0**
176	**61**	**3**	**5**	**2**	**4**
				0	**6**

$$r_s = 1 - \frac{6\sum di^2}{n(n^2-1)}$$

$$= 1 - \frac{6(6)}{5(25-1)}$$

$$= 1 - \frac{36}{120}$$

$$= 1 - 0.3$$

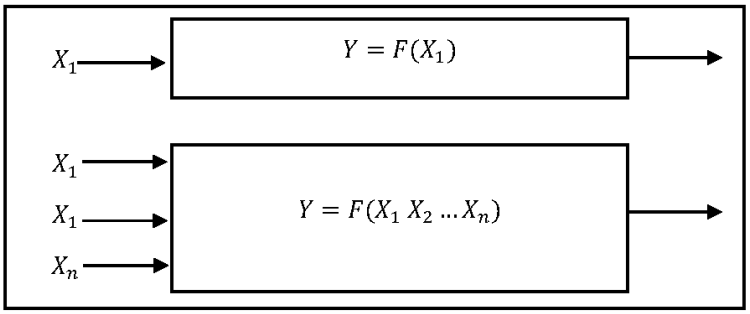
$$= 0.7$$

Since the Spearman correlation coefficient was 0.7, this indicates an average and positive relationship between age and blood pressure for these people.

5.2 Regression Analysis

Regression analysis is an analysis that enables us to find a mathematical equation to connect a one quantitative variable, which is the "dependent variable" with the quantitative independent variable, which is called a simple regression, or with several independent quantitative variables, known as a multiple regression. This relationship may be a linear or nonlinear "logarithmic, exponential, etc."

The general formula of this model that clarifies the relationship between the dependent variable and the independent variable or between the set of independent variables is: $\boldsymbol{y = f(x_I \ldots . x_n)}$



Where:

Y represents the dependent variable.

$\boldsymbol{X_1 \ldots . X_n}$ Represent independent variables ", which are the variables that cause or influence the change of the dependent variable.

If we give them a value (i.e, a value belonging to the set of real numbers), here the value of Y will be determined by the value of X. Therefore, the variable X is defined as the independent variable, while Y is the value of X, so Y is defined by the dependent variable (i.e., according to the X value).

The regression means looking for this relationship between variables X, Y, and that equation $Y = B_0 + B_1 X_1$ where B_0 and B_1 are constant values. B_0 represents the intersection coefficient (equation constant), which reflects the value of the dependent variable Y in the absence of the value of the independent variable, that is, X is equal to zero, B1 is the slope of the linear line "$B_0+B_1X_1$" and reflects the amount of change in Y if X changes in one unit.

5.2.1 The importance of regression analysis

1. The regression is used primarily for the purposes of Prediction, Planning and Estimation and it aims to predict the value of a given variable if the value of another variable associated with it such as population fertility prediction where the number of women of childbearing age is known.
2. Explain the ratio of variance in the dependent variable that can be explained by its relation to the independent variable.
3. Measuring the strength and type of relationship between the dependent variable and the independent variable in the simple

regression or with the set of independent variables in the case of multiple regression.

4. Identify and describe the community through the estimated equation of regression.
5. Finally, the importance of regression analysis is not limited to estimating the value of the dependent variable for the value of the independent at a specific place and time, but also in enabling us to go back to the past to know the conditions of certain phenomena.

5.2.2.1 Types of regression analysis

There are two types of regression analysis; the first is linear regression, which deals with the relationship between the variables in a linear equation, and the second is nonlinear regression, which is interested in studying the relationship between the variables in a curve form rather than a straight line.

Linear regression is the most common and widely used on two types: the first is simple linear regression from which we can predict the relationship between the dependent variable and how one variable affects it; the other multiple regression, is interested with the relationship between the dependent variable and several independent factors that affect it.

A detailed explanation of the linear method of regression analysis is provided here. If a nonlinear approach is needed, we can refer to several sources in statistics.

5.2.2.1 Simple linear Regression Analysis

The main objective of simple linear regression analysis is to study or analyze the effect of a quantitative variable on another quantitative variable. There are many examples of this, including:

1. Study of the effect of the increase in body weight on blood pressure.
2. Study of the effect of dust on people with respiratory diseases.
3. Study of the effect of education on population fertility.
4. Study of the effect of malnutrition on pregnant women on fetal health.
5. Study of the effect of smoking on the health of smokers.
6. Study of the impact of pollution on the health of the peoples.
7. Study of the effect of exercise on human health.
8. Study of the impact of fog on traffic accidents.
9. Study of the impact of wind on the spread of infectious diseases.
10. Study of the impact of increasing income on improving the health status of the population.

There are many other examples in various fields of social, health, economic and natural life.

5.2.2.1.1 Simple Regression Model

The researcher of simple regression analysis is interested in studying the relationship between two variables only one of which is dependent (Y) and the other independent (X).

The first step to study this relationship is to collect data and then graphically represent these two variables. If we assume that the corresponding values of variables X and Y are (Y_1, X_1), (Y_2, X_2) and… $(Y_N X_k)$, and we represent the values of the independent variable (X) on the horizontal axis and the values of the dependent variable (Y) on the vertical axis and we draw Scatter Diagram, the relationship between the two variables can be determined by tracing the points of the Scatter Diagram. It may be noted that the points in this figure represent a straight line or very close to the line, or may take one of the other forms shown in Figure (27). It is important to note that not all points are located on the straight line, since some points may deviate from this line for one reason or another. In these cases, these points are called extreme or abnormal points.

In order for this line to be straight or not and representative of the studied data, it must provide large number of point values and mediate the rest of them in an acceptable manner, i.e., the sum of squares of deviations of point values in the scatter diagram from their counterparts on the regression line less as possible.

From the above, the researcher will be able to form a clear idea of the relationship between these variables and their degree. If there is a linear or close relationship, the Linear Regression Model can take the form of a first-line linear equation that reflects the Y variable as a function in the independent variable (X).

$$Y = B_0 + B_1 X$$

Since it was not expected that all points values are located on the line completely, the linear relationship in the formula above should be modified to include the random error limit ε_i .

$$\boldsymbol{Y = B_0 + B_1 X + \varepsilon_i}$$

$$\begin{bmatrix} \textbf{Total} \\ \textbf{Variation in } y \end{bmatrix} = \begin{bmatrix} \textbf{Explained} \\ \textbf{variation} \end{bmatrix} + \begin{bmatrix} \textbf{Inexplained} \\ \textbf{variation} \end{bmatrix}$$

$$\begin{bmatrix} \textbf{Change} \\ \textbf{independent} \\ \textbf{Variable} \end{bmatrix} = \begin{bmatrix} \textbf{Change} \\ \textbf{independent} \\ \textbf{Variable} \end{bmatrix} + \begin{bmatrix} \textbf{Change in} \\ \textbf{random} \\ \textbf{Variable} \end{bmatrix}$$

That is, the first part of the changes in the dependent variable (Y) can be explained by changes in the independent variable (X), while the second part is explained by the random effect (ε_i) .

The equation $Y = B_0 + B_1 X + \varepsilon i$ confirms that (Y), the dependent variable, is a linear function in the independent variable (X). In other words, if we compare this equation with a graph, we will notice that it represents a straight line as shown in Figure (28) and that **B_0 and B_1** are the constants of the model as **B_0** represents the Intercept constant. That is, the graph represents the distance between zero and the point at which the straight line cuts **B_0+B_1** the axis of the dependent variable (Y) and thus represents the value of (Y) when the value of (X) is zero.

B1 is the slope of the straight line and is called the regression coefficient which represents the amount of change in (Y) when the

independent variable (X) changes in one unit, as shown in the following figure:

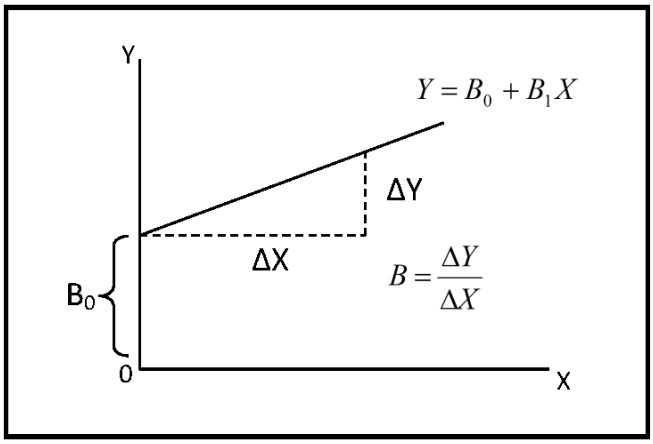


Figure (28) Graph of the simple linear regression equation

The $\boldsymbol{\varepsilon_i}$ expresses the random error, which represents the difference between the actual value and the estimated value $\hat{Y} = Y - (B_0 + B_1 X)$. If the actual value $(\boldsymbol{Y_i})$ is above the regression line, the difference $\varepsilon_i$is positive. If the actual value is below the regression line, then the difference is negative ei and the value of $\boldsymbol{\varepsilon_i}$ is zero when the actual and estimated values on the regression line are equal, as shown in the figure below:

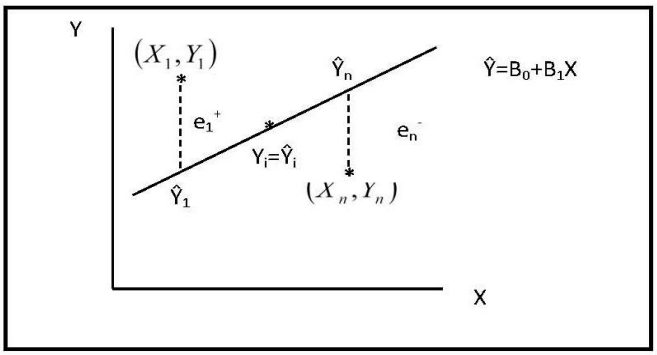


Figure (29) the random error diagram

We can measure the quality of regression line and compute the sum of square deviations of values.

$\sum\left(y-\hat{y}\right)^2=\sum ei^2=e_1^2+e_2^2+..e_n^2$ If the sum of the square deviations of values from the line zero is small, the line of observations is as good as possible.

5.2.2.1.2 Hypotheses simple linear regression analysis

Regression analysis is one of the most common and widely-used statistical methods by researchers for ease and relevance of analysis of many phenomena, especially within the medical and health fields. However, its use requires several assumptions that must be achieved to ensure a sound analytical result.

The most important of these assumptions are:

1. The dependent variable must be a quantitative variable; otherwise, the independent variables may be a combination of qualitative and quantitative variables.
2. The relationship between the dependent variable and the independent variable is a linear relationship, i.e., the change in the dependent variable is fixed, compared with the rate of change in the independent variable.
3. Both the dependent variable and the independent variable follow the Normal Distribution.
4. That follows the random error (ε_i) normal distribution with arithmetic mean equal to zero and variance equal to σ^2 with each value of the independent variable values.

5. The random error variation is homogeneous at each value of the independent variable.
6. There is no autocorrelation between Random Errors. This means that the phenomenon studied in time (t) is not affected by time (t+1) or (t-1), but this is rare in the applied side, however, this is rare on the practical side, as most data is usually influenced by past viewing and is influenced by subsequent viewing. On the other hand, this problem may occur due to the neglect of some independent variables from the studied relationship for one reason or another, or may occur as a result of the inaccurate formulation of the mathematical model of the problem being investigated.
7. The regression model data should be free of extreme values (Outliers). The more these values are multiplied in the estimation of the regression coefficients, the negative effect on these estimates, and in some anomalies, one extreme point may be able to completely change the direction of the regression line completely.
8. Independence of Residuals means that the rest to any point does not depend on the rest in the point or other points.

5.2.2.1.3 Estimation of simple linear Regression Equation

If we are interested in studying the relationship between two different variables, the first step to study this relationship is to collect data and then it represent graphically. Assume that the

pairs of values for variables Y and X are $(x_1, y_1)(x_2, y_2)...(x_n, y_n)$ and represent the values of y on the vertical axis and x values on the horizontal axis, and we draw the scatter diagram. We notice that the points in the scatter diagram take a linear direction. In order to reach the best line representing the relation between the two variables we follow the ordinary least squares method, which are characterized by their abilities as the best linear unbiased Estimator.

Which have:

1. Linearity.
2. Unbiasedness.
3. Have a smaller Variance.

Therefore, the method of the least squares is to find the estimated values of constants B_0 and B_1 by minimizing the total squares of errors to the lowest possible:

$$Min\sum_{i=1}^{n} ei^2 = Min\sum_{i=1}^{n} (Y_i - B_0 - B_1 X_i)^2$$

We can obtain this estimation of constants B_0, B_1 from the solution of the two natural equations in B_0, B_1:

1. $\sum Y = nB_0 + B_1 \sum X$
2. $\sum XY = B_0 \sum X + B_1 \sum X^2$

Thus, we obtain the values of B_0, B_1 as follows:

$$\hat{B}_1 = \frac{\sum XY - n\bar{x}\bar{y}}{\sum X^2 - n\left(\bar{x}\right)^2}$$

Or

$$\hat{B}_1 = \frac{\sum XY - \frac{\sum x \sum y}{n}}{\sum X^2 - \frac{\left(\sum x\right)^2}{n}}$$

And $\hat{B}_0$

$$\hat{B}_0 = \bar{Y} - \hat{B}\,\bar{X}$$

After the values of parameters B_0, and B_1 become known, we can predict the values of Y in the case of any change in the values of X, so the estimated regression equation is:

$$\hat{Y} = \hat{B} + \hat{b}_1\, X$$

Example:

The following data represents the number of open-heart operations that were made in a private hospital and the number of specialists involved the period from 2003 to 2012.

Number of Operation Y	**112**	**128**	**130**	**158**	**162**	**140**	**138**	**175**	**125**	**142**
Number of doctors X	**35**	**40**	**38**	**67**	**64**	**59**	**44**	**69**	**25**	**50**

Required:

1. Determine dependent variable and independent variable.
2. Drawing the scatter diagram.

3. Estimation of the regression equation.
4. How many operations are expected if the number of doctors in the hospital increases to 80?

Solution:

First: The dependent variable (Y) represents the number of open - heart operations that were made and the independent variable (X) represents the number of doctor's specializing in this field.

Second: The drawing of the scatter diagram requires the drawing of two vertical axes to represent the number of operations performed and a horizontal axis to represent the number of doctors, as illustrated in Figure (30).

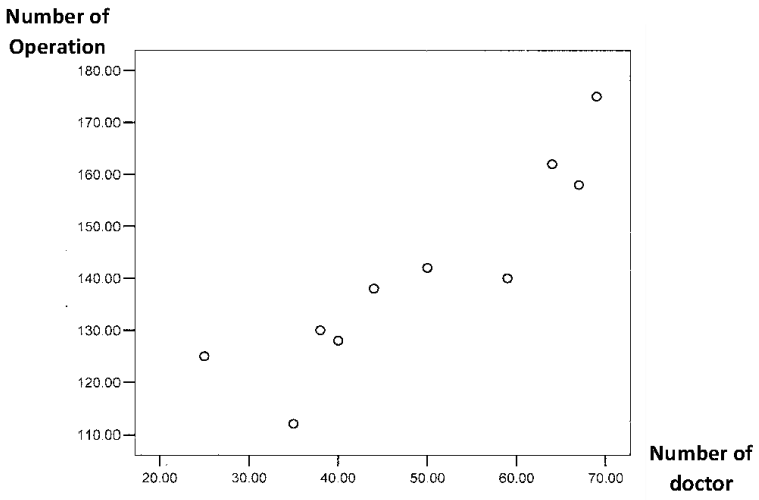


Figure (30) scatter diagram of variables Y and X

Third: Estimation the value of the regression equation:

To calculate the equations of regression equation B_0, B_1 requires the order of data in the following table through which the use of mathematical formulas of the regression we get those constants:

Year	Y_i	X_i	$\sum X_i Y_i$	$\sum X_i^2$	$\sum Y_i^2$
2001	112	35	3920	1225	12544
2002	128	40	5120	1600	16384
2003	130	38	4940	1444	16900
2004	138	44	6072	1936	19044
2005	158	67	10586	4489	24964
2006	162	64	10368	4096	26244
2007	140	59	8260	3481	19600
2008	175	69	12075	4761	30625
2009	125	25	3125	625	15625
2010	142	50	7100	2500	20164
$\sum$	1410	491	71566	26157	202094

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{1410}{10} = 141$$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{491}{10} 49.1$$

$$\hat{B} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X_i^2 - n(\bar{X})^2}$$

$$= \frac{71566 - 10(49.1)(141)}{26157 - 10(49.1)^2}$$

$$= \frac{71566 - 69231}{26157 - 24108.1}$$

$$= \frac{2335}{2048.9} = 1.140$$

$$\hat{B}_0 = \bar{Y} - \hat{B}\bar{X}$$

$$= 141 - (1.140)(49.1)$$

$$= 85.026$$

Therefore, the estimated regression equation is:

$$\hat{Y} = 85.026 + 1.140(X)$$

From the above equation it can be said that if the number of doctors does not increase, the number of operations will be 85,

and the greater the number of doctors increased by one doctor, it will lead to an increase in operations by one process.

Fourth The number of operations that is expected to take place in the hospital operations by increasing the number of doctors to 80 doctors is:

$\hat{Y} = 85.026 + 1.140\ (80)$

$\hat{Y} = 85.026 + 91.2$

$\hat{Y} = 176\ Operation$

5.2.2.1.4 Inference about goodness of fit regression line

An important topic in the regression analysis process is to find confidence intervals for each of the regression line constants B_0 and B_1, and test hypotheses about these constants through random error variation or standard deviation of random errors e_i. Which measures the degree of spread points in the Scatter diagram around the regression line and the R-square correlation, being a measure of the quality of success. Whereas R^2 shows the percentage of total changes in the dependent variable (Y) that can be illustrated by the independent variable (X).

It is known that the actual value of the dependent variable (Y) consists of two parts:

$$Y_i = \hat{Y}_i - \hat{e}_i$$

That is, the total changes in the $\left(Y_i - \hat{Y}\right)$ variable can be traced back to

$$\left(Y_i - \hat{Y}\right) \quad = \left(\hat{Y}_i - \overline{Y}\right) + \quad \hat{e}i$$

Difference because to random error + difference because to regression = total difference

And then reflecting the total sum of squares are the two components:

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\hat{e}\,i$$

Total sum of squares	=	**Regression sum of squares**	+	**Error sum of squares**
SST	=	**SSR**	+	**SSE**

From the above, two quality indicators can be obtained:

1. Determinant coefficient:

The determinant coefficient is expressed by the ratio of the total squares of the regression "SSR" to the sum of the squares of the total "SST" and also called "square simple correlation coefficient R - square" and denoted by R^2 and the value ranges between one and zero, and is calculated as follows:

$$R^2=\frac{SSR}{SST}=\frac{\sum\left(\hat{y}_i-\overline{y}\right)^2}{\sum\left(y_i-\overline{y}\right)^2}$$

$$=\frac{\hat{B}_1^2\sum\left(X_i-\overline{X}\right)^2}{\sum\left(Y_i-\overline{Y}\right)^2}=\frac{\hat{B}_1\left(\sum X_i^2-\frac{\left(\sum X_i\right)^2}{n}\right)}{\left(\sum Y_i^2-\frac{\left(\sum Y_i\right)^2}{n}\right)}$$

The data parameter of the previous example is:

$$R^2 = \frac{(1.140)^2\left(26157 - \frac{(491)^2}{10}\right)}{202094 - \frac{(1410)^2}{10}}$$

$$= \frac{1.2996(2049)}{202094 - 198810} = \frac{2662.88}{3284} = 0.81$$

This means that 81% of total changes in the number of operations (Y) is due to changes from the increase in the number of doctors (X), and 19% due to other factors and other random changes.

2. Standard Error of Estimate:

The standard error of measurement is the degree of spread of the real values of the dependent variable (Y) around the estimated regression line ($\hat{Y}$). In other words, it measures the real values of Y moving away or approaching from the estimated regression line. If the value is large, is significant, the deviations of the actual values of the dependent variable (Y) from the estimated values are large and the model is not efficient and vice versa. The closer the value is from zero, the closer the true value of Y is than the estimated regression line.

The standard error is calculated in the following formulas:

$$\mathbf{S_e} = \sqrt{\frac{\sum\left(Y_i - \hat{Y}\right)^2}{n-2}}$$

Or

$$\mathbf{Se} = \sqrt{\frac{Syy - bSxy}{n-2}}$$

From the previous example data, the value of the standard error Se can be found using the differences between the actual values (Y_i) and the estimated values $(\hat{Y}_i)$ as shown in table (17).

Table (17) The actual values, the estimated values and the total squares of the deviations of the error

Y_i	$\hat{Y}_i = 85.026 + 1.140(x)$	$Y_i - \hat{Y}_i$	$(y_i - \hat{y}_i)2$
112	**124.93**	**-12.93**	**167.18**
128	**130.63**	**-2.63**	**6.91**
130	**128.35**	**1.65**	**2.72**
138	**135.18**	**2.81**	**7.89**
158	**161.34**	**-3.377**	**11.55**
162	**157.98**	**4.02**	**16.16**
140	**152.28**	**-12.22**	**149.32**
175	**163.67**	**11.33**	**128.36**
125	**113.53**	**11.47**	**131.56**
142	**142.02**	**-0.02**	**0.0004**
1410		**0**	**621**

Therefore:

$$S_e = \sqrt{\frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

$$S_e = \sqrt{\frac{621}{10-2}}$$

$$S_e = \sqrt{77.625}$$

S_e = 8.81 The value of the standard error.

Moreover, the value of Se itself can be found from the previous example data shown below

$\sum Y_i^2 = 202094$, $\sum Y_i = 1410$, b=1.140, n=10 $\sum X_i^2 = 26157$,

$\sum X_i = 491$ ، $\sum X_i Y_i = 71566$

Using the following formula:

$$S_e = \sqrt{\frac{Syy - bSxy}{n-2}}$$

Where:

$$Syy = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}$$

$$= 202094 - \frac{(1410)^2}{10}$$

$$= 202094 - 198810$$

$$= 3284$$

$$Sxy = \sum X_i Y_i - \frac{\left(\sum X_i \sum Y_i\right)}{n}$$

$$= 71566 - \frac{(491)(1410)}{10}$$

$$= 71566 - 69231$$

$$= 2335$$

$$\therefore \quad S_e = \sqrt{\frac{3284 - (1.140)2335}{10-2}}$$

$$S_e = \sqrt{\frac{622.1}{8}}$$

$$S_e = \sqrt{77.7625}$$

$\therefore$ $S_e = 8.81$ The value of the standard error

5.2.2.2 Multiple Linear Regression

Our previous study was limited to studying the relationship between two variables, but the researcher often faces a phenomenon that is not related to one variable only, but is related to and influenced by several variables.

For example, the researcher's interest may be to study the relationship between the rate of population growth, the number of married women, the health care of pregnant women, the number of specialized doctors, education, occupation, income, and other independent variables.

In this case, the researcher may be interested in studying the relationship between the dependent variable its denoted (Y) and several independent variables which are denoted by $(X_1, X_2, ..., X_k)$.

Therefore, the simple linear regression model $\hat{Y} = B_0 + BX$ will expand to take the following formula

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + ... + B_kX_k$$

Where:

B_0 represents the Y-intercept, that is the value of Y when all x_i values are zero. B_1,B2….,B_k represent the partial regression coefficients that represent the change in Y for each unit of change in x_i. Therefore, in the case of multiple linear relationships and the existence of K of independent variables, the number of constants of this relationship will reach (k+1) i.e., (B_0, B_1B_2,… ,B_k), which requires the use of the following formulas:

$$\hat{B}_1 = \frac{\left(\sum X_{1i}Y_i\right)\left(\sum X_{2i}^2\right) - \left(\sum X_{2i}Y_i\right)\left(\sum X_{1i}X_{2i}\right)}{\left(\sum X_{1i}^2\right)\left(\sum X_{2i}^2\right) - \left(\sum X_{1i}X_{2i}\right)^2}$$

$$\hat{B}_2 = \frac{\left(\sum X_{2i}Y\right)\left(\sum X_{1i}^2\right) - \left(\sum X_{1i}X_{2i}\right)}{\left(\sum X_{1i}^2\right)\left(\sum X_{2i}^2\right) - \left(\sum X_{1i}X_{2i}\right)^2}$$

$$\hat{B}_0 = \overline{Y} - \hat{B}_1\ \overline{X}_1 - \hat{B}_2\ \overline{X}_2$$

We will limit our presentation of this topic to the linear relationship between only three variables, including one dependent variable (Y) and others independent X1 and X2 as the generalization of more than three variables requires some statistical methods and processes that may come out from the focus of our study now.

Example:

The following table represents blood pressure (Y) and age (X_1) year and Weight (X_2) kg for six people. Find the regression

equation Y (blood pressure) on X_1 and X_2 using ordinary least squares method OLS.

Table (18)

Blood pressure and independent variables age and weight

People	Blood Pressure Y	Age X	Weight X2
1	130	43	102
2	120	48	85
3	135	56	95
4	143	61	110
5	141	67	98
6	162	70	122
∑	831	345	612
Mean	138.5	57.5	102

Solution:

For estimating the value of regression equation constants $\hat{B}_0, \hat{B}_1, \hat{B}_2$ using the ordinary least squares method OLS a table can be prepared that contains all the required calculations for this:

Table (19)

Calculations required estimating Y regression equation constants on X1 and X2

Blood Pressure Y_i	**Age** X_{1i}	**Weights** X_{2i}	$(Y_i - \bar{Y})$	$(X_{1i} - \bar{X}_{1i})$	$(X_{2i} - \bar{X}_{2i})$	$Y_i X_{1i}$	$Y_i X_{2i}$	$X_{1i} X_{2i}$	Y_i^2	X_{1i}^2	X_{2i}^2
130	43	102	-8.5	-14.5	0	123.25	0	0	72.25	210.25	0
120	48	85	-18.5	-9.5	-17	175.75	314.5	161.5	342.25	90.25	289
135	56	95	-3.5	-1.5	-7	5.25	24.5	10.5	12.25	2.25	49
143	61	110	4.5	3.5	8	15.75	36	28	20.25	12.25	64
141	67	98	2.5	9.5	-4	23.75	-10	-38	6.25	90.25	16
162	70	122	23.5	12.5	20	293.75	470	250	552.25	156.25	400
$\sum Y_i$ 831	$\sum X_{1i}$ 345	$\sum X_{2i}$ 612	0	0	0	637.5	835	412	1005.5	561.5	818
$\bar{X}_1 = 138.5$	$\bar{X}_1 = 57.5$	$\bar{X}_1 = 102$									

From Table (19), regression equation constants can be estimated using the following formulas:

$$\hat{b}_1 = \frac{\left(\sum X_{1i}Y_i\right)\left(\sum X_{2i}^2\right) - \left(\sum X_{2i}Y_i\right)\left(\sum X_{1i}X_{2i}\right)}{\left(\sum X_{1i}^2\right)\left(\sum X_{2i}^2\right) - \left(\sum X_{1i}X_{2i}\right)^2}$$

$$= \frac{(637.5)(818) - (835)(412)}{(561.5)(818) - (412)^2} = \frac{177455}{289563} = 0.613$$

$$\hat{b}_2 = \frac{\left(\sum X_{2i}Y_i\right)\left(\sum X_{1i}^2\right) - \left(\sum X_{1i}Y_i\right)\left(\sum X_{1i}X_{2i}\right)}{\left(\sum X_{1i}^2\right)\left(\sum X_{2i}^2\right) - \left(\sum X_{1i}X_{2i}\right)^2}$$

$$= \frac{(835)(561.5) - (637.5)(412)}{(561.5)(818) - (412)^2} = \frac{2062025}{289563} = 0.712$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}_1 - \hat{b}_2 \bar{X}_2$$
$$= 138.5 - (0.613)(57.5) - (0.712)(102) = 30.626$$

Therefore, the estimated regression equation would be:

$$\hat{Y} = 30.626 + (0.613)(Age) + (0.712)(weights)$$

S.E	9.353	0.138	0.114
T	(3.275)	(4.444)	(6.232)

R= 0.99

R-square = 0.98

F= 73.169 D.F = (2,3) Sing = 0.003

It is clear from the above model that the signals of the parameters are consistent with the nature of the effect of these

variables (age, weight) on blood pressure. It can be said that the blood pressure of these six people will increase by 0.613 when the increase of life by one year and with the stable of weight, but when the weight is increase by 1k.g It will increase the pressure by 0.712. It is also evident from the model that if he passed the S.E test where we note that half the value of each parameter of the model parameter is greater than the value of S.E for each. We also note that the T-Test model, where we note that the calculated value t of the parameters of the model (B_0, age X_1, weight X_2) of (3.275, 4.444, 6.252) respectively is greater than the value of 3.182 with a significant level of 0.025 and the degree of freedom 3. In addition, the F-Test test confirms the importance and realism of the variables (age and weight) that the model provides and increases the confidence in it. The calculated F value of 37.169 is much higher than the planned value of 30.82 with a significant level of 0.01 and the freedom degree of (2,3).

Finally, to ascertain the strength of the relationship between the dependent variables (age, weight) and the dependent variable (blood pressure), we depend on the value of the R-Square, which is 0.98. This means that 98% of the changes in blood pressure are attributable to the factor Age and Weight and that 2% were attributed to other factors that the model was unable to enumerate.

Moreover, you can find out which of the two variables, Age and Weight, can be more influential on blood pressure for these peoples by performing a simple regression analysis as shown in Figure (31). We note that the R-square value of the weight

variable is 0.85 and is much larger than the R-Square value of the age variable. This means that weight has a greater effect on blood pressure than age.

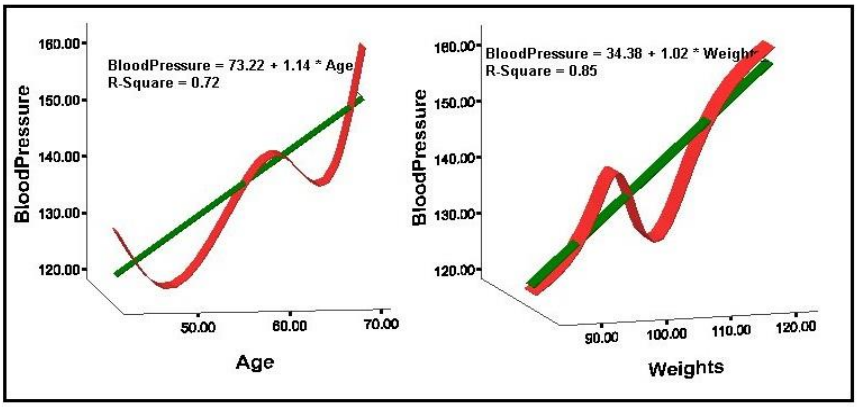


Figure (31)
the simple regression of the relationship between blood pressure and independent variables

Chapter Six

Statistical tests

Chapter Six

Statistical tests

Statistical hypothesis in general is an expression or conjecture that may be true or false around a parameter of society or about the probability distribution of society or about two or more parameters if the study is to compare two or more societies.

6.1 Test of Hypotheses

The process of dealing with assumptions and judging their credibility is called the hypothesis testing process, which is a methods of statistical inference that is based on a test of the hypothesis or claiming about a parameter of a society to prove the opposite. This means the validity of the claim is based on random sample data drawn from that community and that mechanism gives the hypothesis test the strength of avoiding bias and inaccuracy.

The statistical hypothesis, in general, is an expression or conjecture that may be true or false about a parameter of society about the probability distribution of society or about two or more parameters if the study is for the comparison of two or more societies.

6.1.1 Steps to Test Hypotheses

1. The formulation of hypotheses about the parameters of the societies under search and study includes:

A. The null hypothesis (Ho) is a hypothesis about the population parameter, which we are testing by using data from a random sample drawn from that population. This indicates that the difference between the population parameter statistic and the sample drawn from it is a result of the coincidence and denies the existence of a difference relationship or effect between them. The rejection of the null hypothesis leads to the acceptance of another hypothesis called the alternative hypothesis.

Examples of the null hypothesis:

- There is no statistically significant relation between height and intelligence.
- There is no statistically significant relationship between sex and academic achievement.
- There is no statistically significant relationship between male and female infection with a particular disease.

B. Alternative Hypothesis Hi This is the hypothesis that the researcher places as an alternative to the null hypothesis, and is accepted when the null hypothesis is rejected, because it will mean the existence of a statistically significant relation whether this relationship is inverse or positive between the observed variables of the selected sample of the statistical society.

Examples of the alternative hypothesis:

- There is a strong relationship between smoking and cancer.

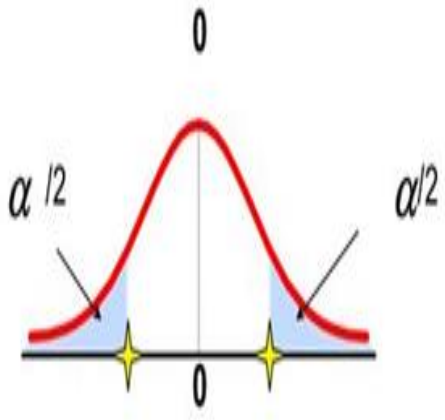
- There is a strong relationship between osteoporosis and sex.
- There is a strong relationship between pregnancy and health care.

2. Level of Significance When we accept the null hypothesis H_o, we accept the accuracy of 90%, 95% and 99%. These percentages are called confidence levels. From these levels, we find that our acceptance of the validity of the null hypothesis is accompanied by a certain error with each level of those levels, which is calculated (1-0.95 =0.5). This ratio is the area of an area under normal distribution curve representing the rejection area. The choice of any level of these levels depends on the type and nature of the study or research on the one hand and the areas of use of the results on the other. However, there is almost universal agreement to determine the use of a significant level of 0.05, 0.01.
3. Determination of critical areas (rejection zones) based on an appropriate distribution if P is an unknown parameter and Po is its assumed value in this case, the hypotheses can be formulated in one of three cases:

A. Tow-tailed test: is the test in which the alternative hypotheses do not show that the community parameter is larger or smaller than the sample of the selected sample, but rather that it is different. Therefore, the rejection zone will be on both ends of the distribution and the assumptions are formulated as follows:

$H_0: P = P_0$

$H_1: P \neq P_0$

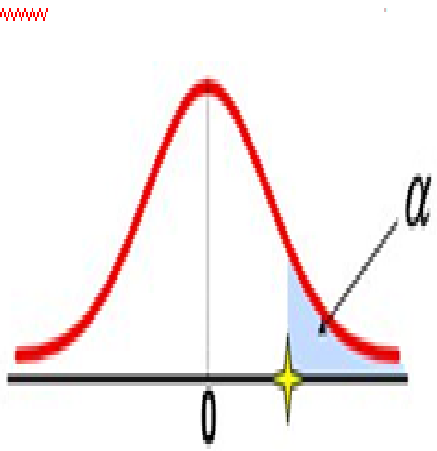


B. One-tailed test: Which is the test in which the alternative hypotheses show that the community parameter is larger or smaller than the sample statistics, and can be used to determine the direction as follows:

B_1- One-tailed test to the right: Meaning that rejection of the area will be on the right side of the distribution allowing us to take the null hypothesis and alternative hypothesis using the following formula:-

$H_0: P = P_0$

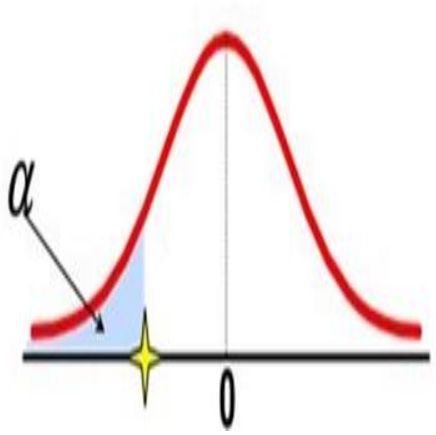
$H_1: P > P_0$



B_2- One-tailed test to the left: Meaning that rejection of the area will be on the left side of the distribution, allowing us to take the null hypothesis and alternative hypothesis using the following formula:-

$H_0: P = P_0$

$H_1: P < P_0$



4. Test Statistic Is the value of the random variable with the probability distribution known when the null hypothesis (Ho) is true, which depends on the calculation of the random sample data drawn from the community and when the value of the decision is to reject or not to reject (accept) null hypothesis.

In general, the set of possible test statistic values is divided into two non-overlapping groups:

A. Admissions area: An area containing all the values of a test statistic that leads to a rejection of the null hypothesis (Ho).

B. Rejection area: The area containing all the values of the test statistic that rejects the null hypothesis (Ho) which is also called the critical region.

The critical value here is the value that separates the rejection and acceptance areas, which are calculated through a certain distribution according to the test of interest and type of search.

5. Statistical decision-making to accept or reject hypotheses based on the test function value for the critical areas.

In general, the hypothesis is rejected by comparing the calculated value of the statistical test with the corresponding tabular value, so that the null hypothesis Ho is rejected and the alternative hypothesis is accepted and only if the calculated value is greater than the tabular value according to the level of significance of 0.05 or 0.01.

6.1.2 Justifications for hypothesis testing

The basic justification for the hypothesis test is that it is a way to prove the validity of the claim of a certain phenomenon, and the most important of these justifications are:

1. When the study variables or research are measured by nominal standards or education.
2. If the researcher does not know the distribution of community data from which the study sample was withdrawn.
3. When the researcher is not sure that the sample has been pulled equally.
4. When the sample size is small.
5. When the research or study groups are not homogeneous.

6.2 The Normal Distribution

Is a continuous distribution, which is a random variable that's continuously moderate because it consists of an infinite number of real values that can be arranged on a continuous scale. It is also known as the Gaussian distribution, relative to Carl Gauss, and was first published in 1733 AD.

The normal distribution is the most important distributions in statistics, is the basis for many statistical theories and plays a key role in the tests of statistical hypotheses and interval of confidence in others.

This can include many characteristics and features such as length, weight, age, level of intelligence and others, if measured and for a large number of observations; the distribution takes the form of normal distribution or close to it. For example, we note that the average age is between 60-75 years, but we find that a minority of individuals may live more than 75 years while another minority may not reach the age of 60 years and such cases apply to height, weight, intelligence level and other qualities.

In addition, normal distribution is used as a approximation of many probability distributions and in many statistical areas and statistical inference for many medical, social science research and other areas.

6.2.1 Normal Distribution Function

If X is assumed to be a randomly connected variable and its average μ and its variance is σ^2, then the random variable X follows a normal distribution of the two parameters μ and σ^2 if the probability function of the random variable X is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)}, \qquad -\infty < \times < \infty$$

Where:

μ = average community

σ^2 = the variance of the community, and is often written like $X \sim N(\mu, \sigma^2)$

e = a constant value of 2.7182

π = constant of 3.1414

6.2.2 Normal Distribution Characteristics

First. Normal distribution parameters:

Normal distribution depends on two parameters:

1. The location parameter μ is the one that determines the location of the distribution at the constant value of σ.
2. The measurement parameter σ determines the range of divergence (dispersion) curve around the center at the stability value of μ.

Second. Normal Distribution Curve:

1. The area under the curve is equal to the correct one.

2. The normal distribution curve is Bell-shaped, i.e. symmetric around the vertical axis passing through point X = μ, which is the average distribution. The area under the curve is divided into two symmetrical halves and is therefore the median.
3. The distribution has maximum end at point X = μ as well, so the mode is where the natural distribution of one point corresponds to the mode.
4. From the above it can be concluded that the X random variable that follows normal distribution is characterized by $MO = ME = \bar{X} = \mu$
5. Normal distribution has a twisting coefficient and is equal to zero.

Third. Empirical law for natural distribution:

The area under the natural curve under the Empirical Rule is divided into three sections as shown in the following table. These percentages are explain by the area below the curve as in Figure (32):

Restricted between	Approximate area
	Under the natural curve
$\mu - \sigma, \mu + \sigma$	68% of the total area
$\mu - 2\sigma, \mu + 2\sigma$	95% of the total area
$\mu - 3\sigma, \mu + 3\sigma$	99.7% of the total area

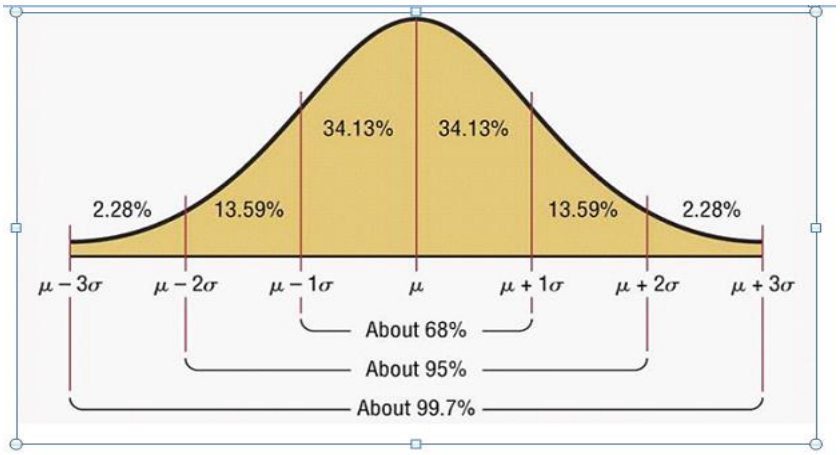


Figure (32) Natural distribution curve

6.3. The Standard Normal Distribution

Because it is difficult and unrealistic to set tables of natural curves for each value of μ and σ, in order to avoid the use of integration, one calculated table for the different areas of normal distribution with arithmetic mean of zero and with variation is one. This distribution is called the Standard Normal Distribution, which is a normal distribution, where μ = 0 and variance $\sigma^2 = 1$. The distribution is abbreviated as follows:

$$Z \sim N(0\,,1)$$

Because each natural distribution can be converted into a standard distribution by converting the actual variable (X) to the standard variable (Z) according to the following formula:

$$Z = \frac{(X - \mu)}{\sigma}$$

Z is called the standard score Z-score.

6.3.1. Standard Normal Distribution Characteristics

1. Mean Standard Normal Distribution of $\mu = 0$ and its $\sigma^2 = 1$ variation.
2. The area under the standard normal curve is equal to one.
3. The standard normal distribution curve is bell shaped, which is Symmetric around the vertical axis passing through point Z = 0, which is the average distribution, and then the area under the curve is divided into two symmetrical halves and is therefore the median.
4. The distribution has a maximum end at the point Z = 0 as well, so it is a mode where the normal distribution of the standard one top corresponds to the mode.
5. The random variable Z, which follows the standard normal distribution, is characterized by:

$$\boldsymbol{MO = ME = \bar{X} = 0}$$

6. Standard normal distribution has a twisting coefficient that is equal to zero.
7. The area is divided under the normal standard curve into three standard sections as in the form of (33):

– 68% of the total area is restrict between (-1 σ, 1 σ).
– 95% of the total area is restrict between (-2 σ, 2 σ).
– 99% of the total area is restrict between (-3 σ, 3 σ).

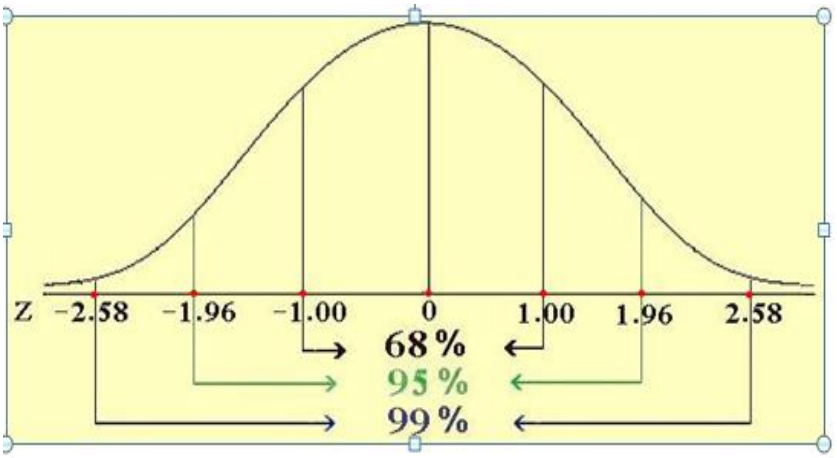
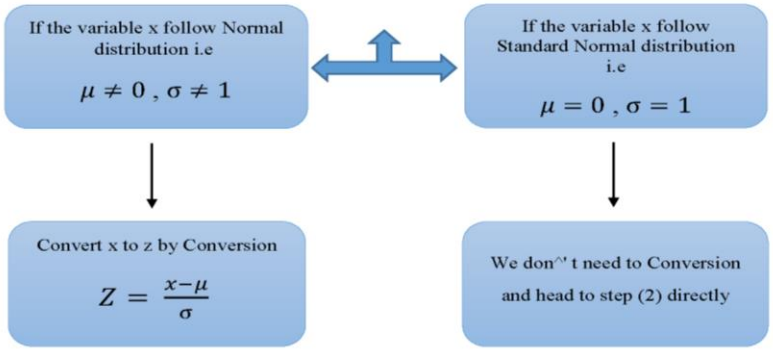


Figure (33) Standard normal distribution curve

6.3.2. Calculate the area under the normal distribution curve

Any area under the normal distribution curve is calculated by using Table (Z). There are three main steps to calculate this area in any manner that is given as follows:

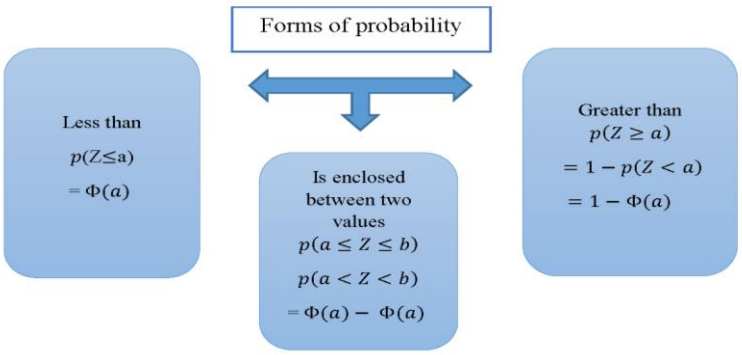
(1) Look at the random variable in question:



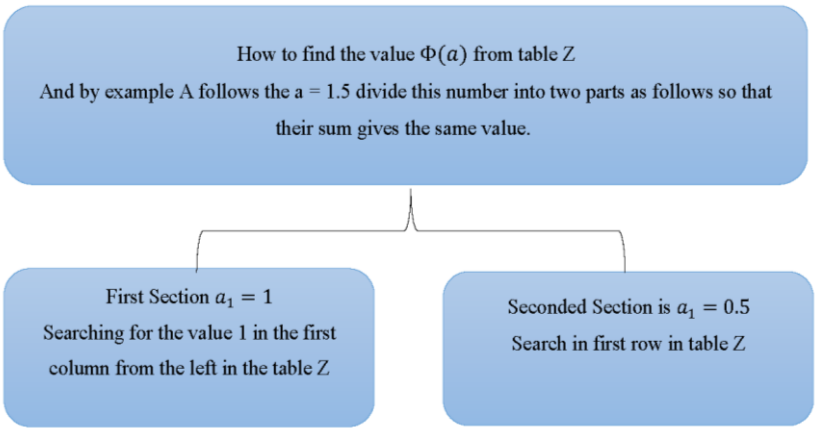
Where: μ = average normal distribution

σ = standard deviation of normal distribution

(2) Look at the required probability of the question on any picture is:



(3) We find $\Phi(a)$ value directly from Table Z and then substitute by its value in step (2) as required in the question



Note:

There are two tables for standard normal distribution (Z), one of which represents the table values in the case of a negative values and the other table in the case of a positive value.

Example A)

If the average blood pressure of 100 young men aged 18-24 years at Systolic Pressure is 115 mm mercury with a standard deviation of 20 mm mercury, find:

The number of young men between the ages of 18-24 years with blood pressure between the range 105-135 mm mercury outside of this range.

Solution:

The number of young people whose pressure is within the range 105-135 mm mercury will find a probability between the two values, i.e.

$P(105 < X < 135)$

$P\left(\frac{105-115}{20}\right) < Z < \left(\frac{135-115}{20}\right)$

$P\left(\frac{-10}{20}\right) < Z < (\frac{20}{20})$

$= P(-0.5 < Z < 1)$

$= P(Z < 1) - P(Z < -0.5)$

$= p(Z < 1) - P(Z > 0.5)$

We get value 1 from Z table

= 0.8413 – (1-0.6915)

= 0.8413 – 0.38413

= 0.4571

∴ 0.4571* 100 = 45.71

That is, the number of young people with blood pressure within the range of pressure (105-135) is approximately 46 young.

The number of young people with blood pressure outside this range will be 100 - 46 = 54

Example B)

Early detection costs for breast cancer in a hospital for a normal distribution of $115 and a standard deviation of $7. What

is the probability that one of these statements, the costs are between $ 104-122.

Solution:

If we assume that the cost per detection is X

Then X: $N(115\ ,7)$

And require P (104 < x < 122)

Converts X to normal standard and converts 104, 122 to standard values

$$Z_1 = \frac{(X-\mu)}{\sigma} = \frac{(104-115)}{7} = -1.57$$

$$Z_2 = \frac{(X-\mu)}{\sigma} = \frac{(122-115)}{7} = 1$$

∴P(-1.57 < Z<1)

= P(Z < 1) – P(Z < -1.57)

We get a value of 1 and 1.579 from a table Z

= 0.8413- (1- 0.9418)

= 0.8413 – 0.0582

= 0.7831

∴ The cost probability of an early breast cancer screening may be between $ 104 - $ 122 is 78%.

Example C)

If the number of surgeries in a hospital follows a probability distribution, with an average of 4 operations, and a standard deviation 2 within 64 weeks, what is the probability that the average number of surgeries is more than 4.2?

Solution:

Within 64 weeks, the possibility that the average number of surgeries in which more than 4.2 operation is $\mu = 4,\ \sigma = 2,\ n = 64$

$$P(\bar{X}) > 4.2 = P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > \frac{4.2 - 4}{\frac{2}{\sqrt{64}}}\right)$$

$$= P\left(Z > \frac{0.2}{0.25}\right) = P\,(Z > 0.8)$$

$$= 1 - p(Z < 0.8)$$

$= 1 - 0.7881$ We get a value of 0.8 from a table (Z)

= 0.2119 Probability value

Example D)

If the lengths of a group of 1000 person take the form of normal distribution with an arithmetic mean of 172 cm and a standard deviation of 5 cm, find

First: The number of persons whose length is within the range (170-175) cm.

Second: The percentage of persons whose length is less than or equal to 168 cm.

Solution:

First:

$$Z_1 = \frac{(X - \mu)}{\sigma} = \frac{(175 - 172)}{5} = 0.6$$

$$Z_2 = \frac{(X - \mu)}{\sigma} = \frac{(170 - 172)}{5} = -0.4$$

Area under curved = Area restricted between zero to + 0.6 area restricted between zero to - 0.4. Which we get from (Z) table as follows:

The area between zero and 0.6 is 0.7257- 0.5 = 0.2257

The area between zero and -0.4 is 0.6554 - 0.5 = 0.1554

The area under the curve is 0.1554 + 0.2257 = 0.3811

∴ The number of persons is 1000 x 0.3811 = 381 persons are located between the height 170-175 cm.

$$Z = \frac{(X-\mu)}{\sigma} = \frac{(168-172)}{5} = -0.8$$

Second:

The area required under the curve = the area restricted between zero and - 0.8

From the (Z) table we get a value of -0.8, which equals

0.2881- 0.7881 = - 0.5

The area required under the curve is equal∴

0.5 - 0.2881 = 0.2219

Hence the percentage of the number of persons whose length of each of them less than or equal to 168 equals

0.2119 × 100 = 21.19%

6.4 Binomial Distribution

A Binomial Distribution is a distribution of a random experiment with two outcomes, one resulting in success and one

in failure, with the main condition being that the probability of success is not affected by the repetition of the experiment.

In binomial experiments, the following four conditions must be consider:

1. There is a specific (known) number of the occurrences of the experiment.
2. The independence of the trials from each other, i.e., the probability of success P and hence the probability of failure q.
3. The sample space has two possible outcomes, only success or failure.
4. The Possibility of success is constant in all tests and does not differ from test to test.

The general formula for this distribution is

$F(x) = \binom{n}{x} p^x q^{n-x}$

Where:

X = Indiscrete random variable, which represents the number of times the occurrence of the event is to be a possibility of occurrence, i.e. X = 0, 1, 2, 3, ..., n

P = probability of success in each test

q = probability of failure in each test

Note that P + q = 1

Example:

If you know that the probability of death of a 60-year-old man because the bird flu vaccine is 0.025, if 100 men are vaccinated.

Find:

(1) average mortality among men (2) variance and (3) What is the probability that at most two of them will die (4) The probability of two men dying (5) finally, the probability that more than two men die?

Solution:

1. The mean is μ = NP = (100) (0.025) = 2.5, meaning that about three men are expected to die after vaccination.
2. The variance is equal to npq= (100) (0.025) (1-0.025) = 2.4375. This means that the number of deaths among men who obtained the vaccination against the bird flu virus is different from the average of deaths among them by 2.4 man.
3. The probability of death of two men at most = the probability of no death of any man (zero) + the probability of death of one man + the probability of death of two men.

$$P_r(0) = \binom{100}{0}(0.025)^0(0.975)^{100} = 0.0795$$

$$P_r(1) = \binom{100}{1}(\frac{0.025}{0.975})(0.0795) = 0.2038$$

$$P_r(2) = \frac{100!}{2(100-2)!}(0.025)^2(0.975)^{98}$$

$$= \frac{100!}{2!\,98!}\,(0.025)^2(0.975)^{98}$$

$$= 4950\,(0.000625)(0.083647) = 0.2587$$

Therefore, the probability of death of two men at most is 0.0795 + 0.2038 + 0.2587 = 0.542.

4. The Probability of death of two men = 0.542.
5. The probability of death of more than two men
= 1 - the probability of the death of at least two men
1 - 0.542 = 0.458.

6.5 Poisson Distribution

Poisson distribution is a discrete probability distribution named after its discoverer, Siméon Denis Poisson.

This distribution is used in experiments where the event of interest occurs an integer number of times within a specified unit of time (minute, hour, day, week, etc.) or unit size or a particular space to see how many times the event of interest in a specific number of tests. Thus, the Poisson variable takes a value from the values (0, 1, 2, 3, ... etc.) over a continuous period of time or in a connected space or area so that the distribution conditions require the following:

1. The probability of success is constant and the probability of failure in each trial is denoted by q, p respectively.
2. The probability of success is small and that is almost close to zero, while the probability of failure is almost approaching one.
3. The number of attempts is very large, where $\lambda = np =$ is a fixed amount. This distribution is often used to calculate rare events such as child suicide and traffic accidents and is currently widely used in communications technology, quality control, atmospheric monitoring, and in various fields of biology.

   And the general formula for such distribution

$$F_{(x)} = \frac{e^{-\lambda}\ \lambda^{x}}{X_i}$$

Where:

X = the number of successes 0, 1, 2, 3,… etc.

e = a constant value of 2.71828 (the base of the natural logarithm).

λ = average number of times of success in a given period or a certain space.

Example A)

A test vial contains forty bacterial cells in a dilute solution of 10 cm^3. If the solution is mixed well and 1 cm 3 is withdrawn, what is the probability of obtaining the following conditions 1, 2, 3, 4, 5.

Solution:

Median $\mu = NP = 40 \times \frac{1}{10} = 4$

$$P_{(x=1)} = \frac{e^{-4}\ 4^1}{1\,!} = \frac{(0.01831)(4)}{1} = 0.07326$$

$$P_{(x=2)} = \frac{e^{-4}\ 4^2}{2\,!} = \frac{(0.01831)(16)}{2\times1} = 0.14648$$

$$P_{(x=3)} = \frac{e^{-4}\ 4^3}{3\,!} = \frac{(0.01831)(64)}{3\times2\times1} = 0.19530$$

$$P_{(x=4)} = \frac{e^{-4}\ 4^4}{4\,!} = \frac{(0.01831)(256)}{4\times3\times2\times1} = 0.19530$$

$$P_{(x=5)}\ \frac{e^{-4}\ 4^5}{5\,!} = \frac{(0.01831)(1024)}{5\times4\times3\times2\times1} = 0.15624$$

X	**1**	**2**	**3**	**4**	**5**	**…. Total**
P(x)	**0.07326**	**0.14648**	**0.19530**	**0.19530**	**0.15624**	**…… 1**

Example B)

If the rate of infection of people with a blood pressure in a particular city is 0.005, then what is the probability that no one is

infected with this disease in one of its neighborhoods of 1000 people.

Solution:

The median is equal

$$\mu = 5 = \text{NP} = (100)\ (0.005)$$

$$\text{P}_{(x)} \frac{\text{e}^{-5}\ \ 1}{0\ \text{i}\ \ 5} = 0.006737$$

6.6. t.test

One of the most common tests used in statistics and its primary purpose is to compare two groups or between two independent or interconnected ones. This test is credited to William Sealey Gosset at the beginning of the 20th century. This test deals with quantitative data (interval or relative). The t distribution is a continuous distributions and is used when the variance of the community is unknown and is used with small samples of size ($n \leq 30$).

The general formula for this test is: $t = \frac{\bar{Y} - \mu}{S\bar{y}} = \frac{\bar{Y} - \mu}{\frac{S}{\sqrt{n}}}$

Where:

$\bar{Y}$ =is the sample average, and $\bar{X}$ may be used instead.

μ = mean community

S_y = standard error

S = standard deviation

n = sample size

6.6.1 Using the test t-test in estimating the confidence interval for the average community (μ)

The confidence interval represents the area of acceptance of the statistical hypotheses and outside this interval represents the rejection of statistical hypotheses.

It is possible to obtain the confidence of the average community μ of the statistical test by replacing the tabular t value with the calculated t value. The tabular t is the point between the acceptance and rejection of statistical hypotheses.

Example:

The daily check-ups of a random sample of six dentists of one of the specialized health centers were as follows:

The Doctor : 1	**2**	**3**	**4**	**5**	**6**
Daily checks : 11	**17**	**12**	**13**	**15**	**13**

Find: the interval of confidence with 95% probability of daily check-ups for all dentists' specialized centers, assuming that daily tests are distributed according to normal distribution or close to it.

Solution:

1. Calculation of arithmetic mean $\bar{X}$:

$$\bar{X} = \frac{11 + 17 + 12 + 13 + 15 + 13}{6} - \frac{81}{6} = 13.5$$

2. Calculation of standard deviation:

The doctor	Number of daily check-up X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	11	2.5-	6.25
2	17	3.5	12.25
3	12	1.5-	2.25
4	13	0.5-	0.25
5	15	1.5	2.25
6	13	-0.5	0.25
∑	81	0	23.5

$$33333\bar{x} = \frac{81}{6} = 13.5$$

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{N-1}}$$

$$= \sqrt{\frac{23.5}{5}} = 2.168$$

3. Find the confidence interval for the mean of the community (μ) (confidence interval) V = N -1=6-1= 5 d. F

– The required level of significance 0.05

Thus, according to the degree of freedom and the required significance level, we extract the t value from the t-test table which is equals to

$$t\left(v, \frac{\propto}{2}\right) = t(5, 0.025) = 2.571$$

Thus, the confidence interval for the average for all the dentists' center With a 95% confidence degree is

$$\mu = \bar{X} \mp t\left(v, \frac{\propto}{2}\right) \times \frac{S}{\sqrt{n}}$$

$$= 13.5 \mp 2.571 \times \frac{2.168}{\sqrt{6}}$$

$$= 13.5 \mp 2.571 \times 0.8852$$

$$= 13.5 \mp 2.2760$$

$$\therefore 11.224 < \mu < 15.776$$

Therefore, the examination of all the doctors' specialized center, from which the studied sample was withdrawn with 95% probability not more than 15.776 and not less than 11,224 examinations.

6.6.2 The use of the t-test in the statistical tests on the testing of means

A. One sample test i.e., the sample mean (X) test with the mean of the μ population from which the sample is taken to make sure that there is a significant difference between them with the assumption of equal variance for both. The null hypothesis is here

$H_0 : \mu = a$

Versus the alternative hypothesis

$1 - \ H_i : \mu \neq a$

$2 - \ H_i : \mu > a$

$3 - \ H_i : \mu < a$

Example:

The following data represent blood pressure readings of a random sample of 14 people with high blood pressure

200,	**207,**	**175,**	**180,**	**201,**	**212,**	**198**
202,	**197,**	**203,**	**205,**	**193,**	**210,**	**217**

What is needed is to test the hypothesis that mean blood pressure is 170.

Solution:

1. Assume the null hypothesis $H_0 : \mu = 170$
2. Alternative Hypothesis $H_i : \mu \neq 170$
3. We find the solution data from the following table.

$\mathbf{X_i}$	$\mathbf{X_i - \bar{X}}$	$\mathbf{(X_i - \bar{X})^2}$
198	**-2**	**4**
212	**12**	**144**
201	**1**	**1**
180	**-20**	**400**
175	**-25**	**625**
207	**7**	**49**
200	**0**	**0**
217	**17**	**289**
210	**10**	**100**
193	**-7**	**49**
205	**5**	**25**
203	**3**	**9**
197	**-3**	**9**
202	**2**	**4**
$\sum$ **2800**		$\sum$ **1708**

$$\bar{X} = \frac{2800}{14} = 200$$

$$S = \sqrt{\sum (X_i - \bar{X})^2} = \sqrt{\frac{1708}{13}} = 11.462$$

$$S_x = \frac{S}{\sqrt{n}} = \frac{11.462}{\sqrt{14}} = \frac{11.462}{3.741} = 3.06$$

$$\therefore t = \frac{\bar{X} - \mu}{S\bar{X}}$$

$$t = \frac{200 - 170}{3.0638} = 9.791$$

4. Making the statistical decision:

Since the level of t calculated value of 9.791 is greater than the scheduled value of 2.160 at a significant level of 0.05 and the degree of freedom 13. Therefore, we reject the null hypothesis and accept the alternative hypothesis that the blood pressure is not equal to 170.

B. Paired –sample T test:

The T-test of paired sample is used to test the difference between two quantitative variables and is often used in medical studies.

For example, blood pressure can be measured in a sample of individuals for the purpose of determining the effect of a particular treatment on this pressure. Treatment was given to all the sample. The blood pressure was then measured after a period of treatment and then, again, after treatment.

Thus, the T-test of the paired sample can be used to test that there is no difference between the mean of the two variables.

Example:

The following data was taken from (12) patients with high blood pressure after their review at a hospital. Their blood pressure was measured before and after treatment with the following results:

Before treatment	**129**	**193**	**140**	**160**	**132**	**135**	**151**	**136**	**160**	**134**	**139**	**144**
After treatment	**126**	**132**	**134**	**145**	**133**	**138**	**141**	**138**	**143**	**130**	**133**	**140**

Test whether there is a significant difference in the effect of treatment on people with blood pressure or not.

Solution:

1. Assume the null hypothesis $H_0: \mu_B = \mu_A$
2. Alternative Hypothesis $H_1: \mu_B \neq \mu_A$
3. Determination of the significant level 0.05
4. Determination of the rejection area since the specified level of significant is 0.05; the test will be the test of the two sides with a space of $\frac{1}{2} \propto$ i. e. 0.025 thus the $t_{0.025}$ ،11=2.201.
5. Find the solution data from the following table.

No.	Before treatment 1	After treatment 2	D difference between 1-2=3	2 D
1	129	126	3	9
2	139	132	7	49
3	140	134	6	36
4	160	145	15	225
5	132	133	-1	1
6	135	138	3	9
7	151	141	10	100
8	136	138	-2	4
9	160	143	17	289
10	134	130	4	16
11	139	133	6	36
12	144	140	4	16
			$\sum$ di = 72	$\sum D^2 = 790$

$$t = \frac{\bar{d} - \mu}{\frac{S\bar{d}}{n}}$$

$$\bar{d} = \frac{\sum d}{n} = \frac{72}{12} = 6$$

$$S^2 d = \frac{\sum D_i^2 - \frac{(\sum D_1)^2}{n}}{n-1}$$

$$S^2 d = \frac{790 - \frac{(72)^2}{12}}{11} = \frac{790 - 432}{11} = \frac{358}{11} = 32.545$$

$$S\bar{d} = \sqrt{\frac{32.454}{12}} = \sqrt{2.71208} = 1.6468$$

Then

$$t = \frac{6 - 0}{1.6468} = 3.643$$

6. Make decision:

With t calculated, we know that the amount of 3.643 greater than the value of (t), the scheduled amount of 2.201 with significant level of 0.05 and the degree of freedom (11). Therefore, we reject the null hypothesis in favor of the alternative hypothesis, which indicates and confirms that the treatment had a positive effect in reducing the pressure at the level of the reviewers.

C. Unpaired-Sample T test:

The purpose of this test is to know whether the difference between the two means taken from an independent community exists or not.

Example:

The following data shows the number of days that some patients spent in one hospital for the healing of the wounds sewn with natural and other synthetic thread. Assume that the data of the two samples are subject to a normal distribution of the same variation.

Synthetic X1	Natural X2
20	12
24	17
26	20
28	25
30	26
34	

Determine if this data are random samples is drawn from one population group.

Solution:

1. Assume the null hypothesis $H_0: \mu_1 = \mu_2$
2. Alternative Hypothesis $H_1: \mu_1 \neq \mu_2$
3. Determine the significant level $\propto = 0.05$.
4. Find the calculated t value by using this formula:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S^2P}{n_1} + \frac{S^2P}{n_2}}}$$

X_1	X_2	$X_i - \bar{X}1$	$X_i - \bar{X}2$	$(X_i - \bar{X}_1^2$	$(X_i - \bar{X}2)$
20	12	-7	-8	49	64
24	17	-3	-3	9	9
26	20	-1	0	1	0
28	25	1	5	1	25
30	26	3	6	9	3
34	-	7		49	143
$\sum$ 162	100			118	

$$\bar{X}_1 = \frac{162}{6} = 27 \quad \bar{X}_2 = \frac{100}{5} = 20$$

$$S_{X1} = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

$$= \sqrt{\frac{118}{6-1}}$$

$$= \sqrt{23.6}$$

$$= 4.857$$

$$SX_2 = \sqrt{\frac{134}{4}}$$

$$SX_2 = \sqrt{33.5}$$

$$SX_2 = 5.787$$

$$S^2P = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$S^2P = \frac{(6-1)23.59 + (5-1)33.48}{6+5-2}$$

$$S^2P = \frac{117.95 + 133.95}{9} = \frac{251.9}{9} = 28$$

$$t = \frac{(27-20) - (0)}{\sqrt{\frac{28}{6} + \frac{28}{5}}}$$

$$= \frac{7}{\sqrt{4.666 + 5.6}} = \frac{7}{\sqrt{10.266}} = \frac{7}{3.204} = 2.184$$

5. The decision:

Since the calculated t value of 2.184 is less than the tabular value which is $t_{0.025} = 2.262$. It accepts the null hypothesis (H_o) and we conclude that there are no significant differences. In the other words, we conclude that the two samples may have been taken from one population group.

6.7 Ch-Square X^2

Karl Pearson's first statistical measure in 1900 has since been expanded upon and has become a common statistical method in the analysis of numerical data, including discrete data. Examples might include the number of males and females, the number of smokers and non-smokers, the number of healthy and unhealthy

patients, the number of neighborhoods and number of deceased in a particular community or in a sample taken from it.

The overall objective of the use of this test is to make sure of the accuracy of the results we obtain from the statistical community compared to the results we get from the sample selected properly. The general idea of this test is to find the difference between the observed values and expected values and then test the extent of this difference.

The general formula for calculating this test:

$$\boldsymbol{X^2 = \sum \frac{(obsreved - expected)^2}{expected}}$$

Here we will focus on using chi-square for the following tests:

6.7.1 Using chi-square in the Test of Independence

This test is used to test hypotheses based on the existence of two types of attributes or characteristics of each random sample variable to determine whether there is a correlation between these two characteristics or whether they are independent.

Example:

A medical team wanted to know the relationship between the blood type and the severity of a particular disease of a sample of a community containing 750 people. After conducting tests, the team obtained the following results

Table (20) Classification of data by blood type and severity of infection

Blood type / Answer	A	B	AB	O	∑
Sound is not infected	250	100	60	200	610
Average infected	30	20	14	24	88
Severly infection	20	10	6	16	52
∑	300	130	80	240	750

Null hypothesis (H_o) The severity of the injury and the blood type are related.

Alternative Hypothesis H_i: The severity of injury and the blood type is independent.

Solution:

1. Find the expected frequencies for each class of the table

$$\frac{610}{750} \times 300 = 244$$

Expected frequencies for blood type A

$\frac{88}{750} \times 300 = 35$

$\frac{52}{750} \times 300 = 21$

$\frac{610}{750} \times 130 = 106$

Expected frequencies for blood type B

$\frac{88}{750} \times 130 = 15$

$\frac{52}{750} \times 130 = 9$

$\frac{610}{750} \times 80 = 65$

Expected frequencies for blood type AB

$\frac{52}{750} \times 80 = 6$

$\frac{610}{750} \times 240 = 195$

Expected frequencies for blood type O

$\frac{88}{750} \times 240 = 28$

$\frac{52}{750} \times 240 = 17$

Blood type / Answer	Ao	ae	Bo	Be	ABo	abe	Oo	Oe	∑
Sound is not infected	**250**	**244**	**100**	**106**	**60**	**65**	**200**	**195**	**610**
Average infected	**30**	**35**	**20**	**15**	**14**	**9**	**24**	**28**	**88**
Severly infected	**20**	**21**	**10**	**9**	**6**	**6**	**16**	**17**	**52**
∑	**300**		**130**		**80**		**240**		**750**

2. Calculate the value of chi-square

$$X^2 = \sum \frac{(0 - e)^2}{e}$$

$$= \frac{(250-244)^2}{244} + \frac{(30-35)^2}{35} + \frac{(20-21)^2}{21} + \frac{(100-106)^2}{106} + \frac{(20-15)^2}{15}$$

$$+ \frac{(10-9)^2}{9} + \frac{(60-65)^2}{65} + \frac{(14-9)^2}{9} + \frac{(6-6)^2}{6} + \frac{(200-195)^2}{195} + \frac{(24-28)^2}{28} +$$

$$\frac{(16-17)^2}{17}$$

$$= 0.1475 + 0.7142 + 0.0476 + 0.3396 + 1.666$$

$$+0.1111 + 0.3846 + 2.777 + 0 + 0.1282 + 0.5714$$

$$+0.0588 = 6.946$$

3. Statistical decision

To make a decision it is necessary to extract the X^2 value, which is scheduled degree of freedom (r-1) (c-1) which (number of rows - 1) multiplied by (number of columns - 1) which (4 – 1)(3 – 1) = 6 and the level of significance 0.05 amounts to 12.592. Hence, it is clear that the calculated value is less than the scheduled value, which means that the data is consistent with the hypothesis (H_1), which stated that the severity of the disease and the blood type is independent for this data.

6.7.2 Method of Contingency table (2 x 2) to calculate the value of the chi-square

This method of calculating the value of chi-square only depends on the values observed and from the following figure

may be reached chi-square by calculated using the methods in the following figure and formula

Variable	Data type I	Data type II	Total
Category I	**A**	**B**	**A+B**
Category II	**C**	**D**	**C+D**
Total	**A+B**	**B+D**	**A+B+C+D=N**

$$\therefore X^2 = \frac{N[(AD) - (BC]^2}{(A+B)(C+D)(B+D)(A+C)}$$

Example:

A medical team was commissioned to study the effects of smoking on the public health of smokers by taking a random sample of 250 people, half of them smokers and the other half of non-smokers. Through the tests conducted on the sample, the following results found:

Health variable / Variable smoking	Good		Not good		Total
Smokers	**A**	**78**	**B**	**47**	**125**
not smokers	**C**	**102**	**D**	**23**	**125**
Total	**180**		**70**		**250**

Required:

The smoking and public health variables are independent when the significant level of 0.05 is reached.

Solution:

1. Determine the hypothesis

H_0 = Smoking has a significant impact on the overall health of the sample

H_1 = Smoking has no significant effect on the public health of the sample

2. Find an X^2 value from the following formula

$$X^2 = \frac{N[(AD) - (BC]^2}{(A+B)(C+D)(B+D)(A+C)}$$

$$X^2 = \frac{250[(78)(23) - ((47)(102)]^2}{(125)(125)(70)(180)}$$

$$X^2 = \frac{250[(179)) - (479)]^2}{2825550000} = \frac{250(9000000)}{196875000}$$

$$X^2 = \frac{2250000000}{196875000}$$

$$X^2 = 11.428$$

3. Statistical decision:

The calculated value amount (11.428) is greater than the tabular value amount (3.841) with a significant level of 0.05 and the degree of freedom (2-1) (2-1) = 1. We accept hypothesis H_0 that said smoking has a significant impact on public health.

6.7.3 Use chi-square to test the quality of Goodness of fit

In many statistical studies, we need to know whether the values of the observations follow a specific distribution that was a natural distribution or a distribution of Poisson or binomial

distribution or any other distribution. Therefore, we can use chi-square to identify and detect the type and nature of the distribution to be able to identify the methods and statistical methods suitable for statistical analysis.

Example:

In a study related to female fertility, a scientific researcher took a sample of 140 married women according to the number of childbirths for each woman.

Required test the data matching below for Poisson distribution or not.

Number of births	**0**	**1**	**2**	**3**	**4**	**5**
Number of mothers	**4**	**15**	**20**	**36**	**40**	**25**

Solution:

We determine the hypothesis

H_0: The data viewing corresponds to Poisson distribution

H1: The data viewing does not match Poisson distribution:

Assuming that H_0 is correct so you must find the theoretical probabilities Pi and then find the expected values through:

1. Calculate the average number of births

$$\bar{X} = \frac{(0)4 + (1)15 + (2)20 + (3)36 + (4)40 + (5)25}{140}$$

$$= \frac{448}{140} = 3.2$$

2. Calculate the probability values for Poisson distribution of points 0, 1, 2, 3, 4, 5 as follows:

Points	Pr(X=r) Probability Value
0	Pro= $(2.718)^{-\overline{X}}$ Points
1	$p_{r1} = \overline{X} \times p_{ro}$
2	$p_{r2} = \frac{\overline{X}}{2} \times p_{r1}$
3	$p_{r3} = \frac{\overline{X}}{3} \times p_{r2}$
4	$p_{r4} = \frac{\overline{X}}{4} \times p_{r3}$
5	$p_{r5} = \frac{\overline{X}}{5} \times p_{r4}$

Where the constant 2.718 is the value of the exponential function symbolized by an exponential function (e)

3. Since the value of $\overline{X}$ is equal to 3.2, the probabilistic values calculated according to those previously mentioned. Then we complete the solution by extracting the expected values to find the value of X^2.

Number of births	Expected value 0	1X3	Probability Value Pr(X=r)	Expected value e	(o-e)	$\frac{(o-e)^2}{e}$
1	2	3	4	5	6	7
0	4	0	0.04076	140× 0.04076=6	-2	0. 666
1	15	15	0.13043	140× 0.13043=18	-3	0.5
2	20	40	0.20868	=29	-9	2.8
3	36	108	0.2370	=33	3	0.27
4	40	160	0.178809	=25	15	9
5	25	125	0.11403	=16	9	5
∑	140	448				17.736

4. Statistical decision:

Since the calculated X^2 value of 17.736 is greater than the scheduled value of 15.09 at a significant level of 0.01 and the degree of freedom of (N-1) i.e. (6-1) and is equal to 5, then we accept the hypothesis H_0 which means that the sample taken from the community matches the distribution of its Poisson.

6.8 Analysis of variance (ANOVA)

This is one of the statistical methods deductive performs the same purpose T- test and can be considered as an extension of it. The resulting value of the variance analysis called the (F value).

The analysis of variance is used to test the hypothesis of those averages that are more than equal or of different societies when those societies are normally distributed with equal variance.

6.8.1 Conditions of use of variance analysis:

1. Random of choice sampling.
2. Independence, i.e. the distribution of errors between samples must be random.
3. The Moderate distribution of variable scores for each sample.
4. Homogeneity of the variance of the communities from which independent random samples are taken.

6.8.2 Steps for analysis of variance:

1. Estimate of the community variation of variance between the samples averages MSA.

2. Estimate of the variation of the community of variation within the samples MSE.
3. Estimation ratio F-calculated.

$$F = \frac{MSA}{MSE}$$

4. Calculate the degrees of freedom (d. F) to determine the statistical significance of the F-calculated ratio and the corresponding proportion F-tabulated.

If the F-calculated value is greater than the scheduled value at the significance level and the degrees of limited freedom, the null hypothesis H_0 is equal to the mean of society rejects in benefit of the alternative hypothesis H_1. Table (21) shows the previous steps and the results of this test.

Table (21) Analysis of ANOVA variance

Sources of variance	Sum of square (ss)	Degrees of freedom d.f	Mean squares (mss)	F-calculated
Between Groups	$\mathbf{SSA = r \sum (\bar{X}_j - \bar{\bar{X}})^2}$	c-1	$\mathbf{MSA = \frac{SSA}{c-1}}$	$\mathbf{\frac{MSA}{MSE}}$
Within groups (Error)	$\mathbf{SSE = \sum \sum (\bar{X}_j - \bar{\bar{X}}\ j)^2}$	(r-1)c	$\mathbf{MSE = \frac{SSE}{(r-1)c}}$	
Total variance	$\mathbf{SST = \sum \sum (\bar{X}_j - \bar{\bar{X}})^2 = SSA + SSE}$	rc-1		

Where:

$\bar{X}_j$ =The average sample J is views of r that is equal to $(\sum iX_j)/r$
$\bar{\bar{X}}$ = Average mean (large mean) for all samples $(\sum i \sum j\, X_j)/rc$
SSA = the sum of squares that is explained by factor

$A = r\sum(\bar{X}_j - \bar{\bar{X}})^2$

SSE = the sum of squares that is explained by factor

$E = \sum\sum(\bar{X}_j - \bar{\bar{X}}\ j)^2$

SST = Total sum of squares $\sum\sum(X_j - \bar{X}_J)^2$ = SSA+SSA

d. f = Degrees of freedom (For numerator = C-1) , for the denominator = (r- 1) C

and (C) represents the number of samples and (r) is the number of views of the sample.

6.8.3 Calculate the variance when the sample size is equal

Example:

The following table represents a period of healing for the three groups (samples) of people with back spasms treated at the physiotherapy center in three different ways of physical therapy.

Test the hypothesis that there are no significant differences in the mean healing period due to the method of physical therapy or not.

Table (22) period of healing (per day) for patients with spasms treated according to the method of physical therapy

First method I	**5**	**6**	**9**	**7**	**8**
Second method II	**8**	**7**	**6**	**10**	**9**
Third method III	**10**	**9**	**7**	**11**	**8**

Solution:

1. Make the hypothesis

A. The null hypothesis $H_0: \mu_1 = \mu_2 = \mu_3$

B. Alternative Hypothesis $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

2. Calculate the mean of each method

First: 8+7+9+6+5=35 $\bar{X}_1 = \frac{\sum X_i}{n} = \frac{35}{5} = 7$

Second: 8+7+6+10+9= 40 $\bar{X}_2 = \frac{40}{5} = 8$

Third: 10+9+7+11+8=45 $\bar{X}_3 = \frac{45}{5} = 9$

Find the general arithmetic mean $\bar{\bar{X}}$

$$\bar{\bar{X}} = \frac{7+8+9}{3} = 8$$

3. Find the total variation squares between groups (samples) SSA

SSA = 5[(7-8)2 + (8-8)2 + (9-8)2]

= 5 [(1) + (0) + (1)]

= 10

4. Find the total variation squares between groups (samples)

SSE= $(5-7)^2 + (6-7)^2 + (9-7)^2 + (7-7)^2 + (8-7)^2 + (8-8)^2 (7-8)^2$ $+(6-8)^2 + (10-8)^2(9-8)^2+(10-9)^2+(9-9)^2+(7-9)^2+(11-9)^2+(8-9)^2$

SSE= 4+1+4+0+1+0+1+4+4+1+1+0+4+4+1=30

5. Find the sum of the total variation squares SST

SST = SSA + SSE

SST = 10+ 30 = 40

6. Finding degrees of freedom d. f

A. For the numerator (C-1) i.e. (3-1) = 2 i.e. number of treatment methods -1

B. For the denominator C (r-1) i.e. (5-1)3=12 i.e. the number of sample items -1

C. For total variation (n_i - 1) i.e. (15-1) =14

7. Find the average squares within groups (samples)

$$MSA = \frac{SSA}{C-1} = \frac{10}{2} = 5$$

8. Find the average squares within groups (samples)

$$MSE = \frac{SSA}{(r-1)c} = \frac{30}{12} = 2.5$$

9. Find the value of F- calculated $F = \frac{MSA}{MSE} = \frac{5}{2.5} = 2$

10. Statistical decision

Since the value of F –calculated and an amount of 2 is less than a value F- tabulated value of 3.89 at a significant level of 0.05 and the degree of freedom 12, 2 we accept the null hypothesis H_0 and reject the alternative hypothesis H_1. Because of this, there are no significant differences in the average period of healing is due to the natural treatment method. In other words, none of the three methods used in physical therapy does not affects the average duration of recovery for patients with back spasm.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig
Between Groups	**10.000**	**2**	**5.000**	**2.000**	**0.178**
Within Groups	**30.000**	**12**	**2.500**		
Total	**40.000**	**14**			

6.8.4 Calculate variance when sample size varies

Which is to follow the same steps above except for a simple adjustment as the sample size is equal to n_i instead of n, i.e. that the sum of the elements $n_i = n_1 + n_2 + \ldots + n_k$

Example:

The following data represents the sales payments to a pharmaceutical company that sold the same medicine packaged in three forms (Injection،tablet،Syrup) at the same price throughout 2014 and sales were normal distributed and have the same variation.

Test whether the average drug sales varies depended on the form of packaging or not.

Table (23)
Company sales payments of a drug in the three forms

Injection 1	Tablet 2	Syrup 3
60	75	40
40	60	35
48	70	55
52	55	30
	40	55
		25
		40
$\sum = 200$	300	280

Solution:

1. Make the hypothesis

A. The null hypothesis $H_0: \mu_1 = \mu_2 = \mu_3$

B. Alternative Hypothesis $H_1: \mu_1 \neq \mu_2 \neq \mu_3$

2. Total sales

$\sum_{i=1}^{k} \sum_{j=1}^{ni} X_{ij} = 60+40+48+52+75+60+70+55+40+40$

+35+55+30+55+25+40= 780

3. Total sales squares

$\sum_{i=1}^{k}\sum_{j=1}^{ni} X_{ij}^2 = (60)^2+(40)^2+ (48)^2+ (52)^2+ (75)^2+ (60)^2+ (70)^2+ (55)^2+ (40)^2+ (40)^2+ (35)^2+ (55)^2+ (55)^2+ (25)^2+ (40)^2 +(40)^2 += 40958$

4. Total sales square divided by the total sales of each form of packaging

$$\sum_{i=1}^{k} \frac{(\sum_{j=1}^{n_i} X)^2}{ni} = \frac{(200)^2}{4} + \frac{(300)^2}{5} + \frac{(280)^2}{7} = 39200$$

5. Calculation of the correction factor

$$\frac{(\sum_{i=1}^{k}\sum_{j=1}^{n_i} X)^2}{\sum n_i} = \frac{(780)^2}{16} = 38025$$

6. Total squares of total variance

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - (\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij})^2$$

7. Total squares of variation between groups (samples)

$$SSA = \frac{\sum_{i=1}^{k} X_i)^2}{n_i} - \frac{(\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij})^2}{\sum n_i}$$

SSA= 39200- 38025= 1175

8. Total squares of variation within groups (samples)

SSE = SST- SSA

SSE = 2933 -1175 =1758

9. Determination of degrees of freedom d. f

– For numerator (C-1) i.e. (3-1) = 2 i.e. number of columns-1

- For dominator (n_i-1) i.e. (16-3) =13 i.e. number of views - number of columns
- For total variance (n_i-1) i.e. (16-1) =15

10. Mean squares between groups (samples)

$$MSA = \frac{SSA}{C-1} = \frac{1175}{2} = 587.51$$

11. Mean squares within groups (samples)

$$MSE = \frac{SSA}{n_i - c} = \frac{1175}{13} = 135.230$$

12. Calculated F-calculated value $F = \frac{MSA}{MSE} = \frac{587.5}{135.230} = 4.344$

13. Statistical decision

Since the F-Calculated value amounting to 4.344 is greater than the F-tabulated value at a significant level of 0.05 and the freedom degree (2, 13), we reject the null hypothesis H_o and accept the alternative hypothesis H_1, i.e. the average sales of the drug varies depending on the form of packing.

Chapter Seven

Measures Statistics of Population and Biostatistics and Measures of Hospital and Disease Statistics

7.1. Statistical data required for the conduct of population and vital statistics

7.2. Population Estimation

7.3. Population Statistics Measures

7.4. Bio Statistics

7.4.1. The importance of using bio statistics

7.4.2. Bio Statistics Measures

7.4.2.1. Mortality statistics Measures

7.4.2.2. Population fertility Statistics Measures

7.5. Hospital Statistics Measures

7.6. Disease Statistics measures

Chapter Seven

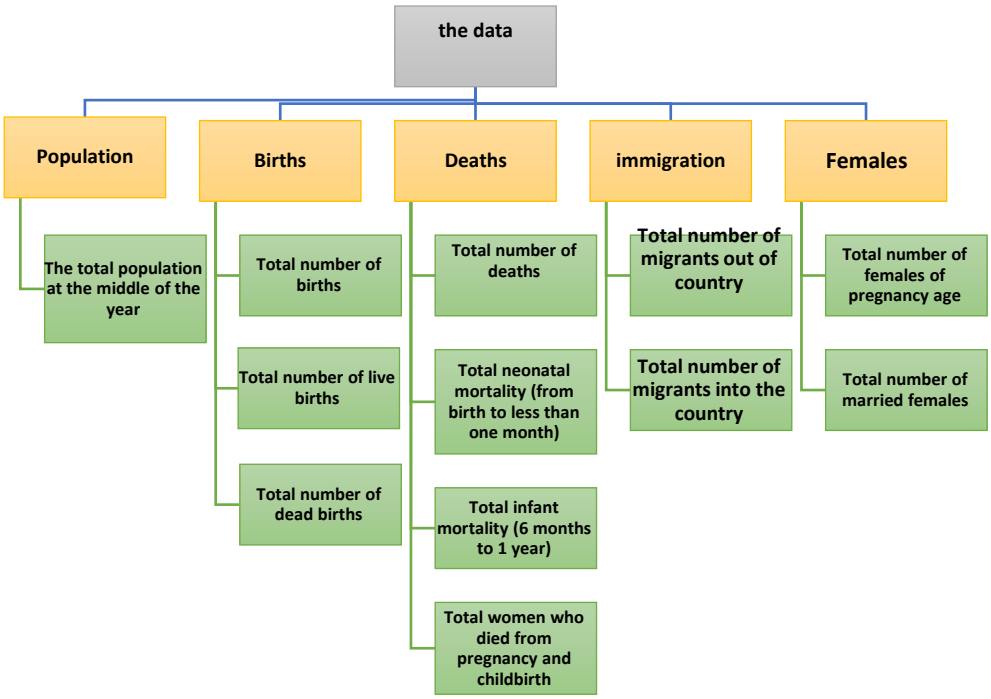
Measures Statistics of Population and Biostatistics and Measures of Hospital and Disease Statistics

We learned earlier, the State has used the word Statistic in the past to indicate the data and information it collects from society for war, taxation, production and other matters. Currently, it collects and uses information and statistics about society, especially the size of the society, of the vital statistics that occur in society (births, deaths, marriages, divorce, etc).

The attention to these statistics, once compiled and categorized, is necessary for every developing country as well as advanced, so that the governments of those countries can use those statistics to provide the best community services. This can only be done by using these statistics by certain standards in order to imagine reality and the extent to which this reality can be promoted in the future.

7.1. Statistical data required for the conduct of population and vital statistic

Include the following data



From the above data, we will be able to calculate the following population and bio-statistics measures:

First: Population estimates

Second: Measures of population statistics

Third: Measures of biostatistics statistics

7.2. Population Estimation

There are many ways to estimate the size of the population. The most widely used method is the linear way of estimating the population under the following formula:

$$P_t = (P)_n + P_o$$

$$P = \frac{P_n - P_o}{n}$$

Where,

P_t= Estimated population for the required period of time

P= the annual rate of population increase

N= the length of time (the difference between the two periods)

Po= Population at the beginning of the time period

Pn= Population at end of time period

Example: Baghdad's population was 6,700,000 in 2003 and 7,200,000 in 2014.

Find:

1- Percentage increase in population between2003-2014

2- Equation of linear population estimation

3- Estimated the population of Baghdad in 2010

4- Estimated the population of Baghdad in 2020

Solution:

1- $P = \frac{P_n - P_o}{n} = \frac{(7200000 - 6700000}{2014 - 2003}$

$= \frac{500000}{11} = 45454.545$ Population increase

2- $P_t = (P)_n + P_o$

$P_t = 45454.545 \times n + 6700000$

Note that n according to our example here, it is equal to

$2003 - 2010 = 7$

3- $P_{2010} = (45454.545) \times (2010 - 2003)\ 67000$

=7018182

Baghdad population estimated in 2010

4- $P_{2020} = (45454.545) \times n + 6700000$

$=(45454.545) \times 17 + 67000 = 7472727.265$

Baghdad population estimated in 2020

7.3. Population Statistics Measures

They are the main sources that provide basic data on population size and population structure that occur on them for analysis and planning at all levels.

The most important criteria and indicators of population statistics are:

1. The standard of natural increase of population at a given time= (Number of births + number of migrants within the country in the same time period) – (Number of deaths + number of migrants out of the country in the same time period)
2. Standard rate of natural increase of population $\frac{\text{Natural increase of the population}}{\text{Population in the middle of the year}} \times 1000$
3. The standard of the outcome of migration= number of immigrants to the country - number of migrants from the country
4. The standard of population increase = the natural increase of population + the outcome of migration
5. Standard rate of migration $\frac{\text{The outcome of migration in a given year}}{\text{The population in the middle of the same year}} \times 1000$
6. Standard rate of increase of population $\frac{\text{Increase of the population}}{\text{Population in the middle of the year}} \times 1000$

Example: The number of live births in a country was 285,000 in 2014, and the number of deaths was 120 people and the number of immigrants to the country's 167,000 people, while the number of migrants from the country of 85,000 people. With note, the population of that country is 8.5 million in mid-2014.

Calculate the following criteria:

1. The rate of natural increase of population
2. Migration rate
3. The rate of increase of population
4. Solution:

1. The rate of natural increase of population

$$\frac{\text{Number of live births} - \text{number of deaths}}{\text{Population in the middle of the year}} \times 1000$$

$$= \frac{120426-285000}{8500000} \times 1000$$

= 19.36 per 1,000 people

2. The rate of migration=

$$\frac{\text{Number of immigrants to country} - \text{Number of immigrants from the country}}{\text{The population in the middle of the year}} \times 1000$$

$$= \frac{85000-167000}{8500000} \times 1000$$

= 9.64 per 1,000 people

3. The rate of increase of population= Natural increase of population + outcome of migration

= 9.64 + 19.36 = 29 per 1,000 people

7.4. Bio Statistics

It is a record of the basic human life events from birth to death, and examines what happens in the case of the population (marriage, divorce, births, illness, death, immigration, etc.) and through which changes that occur in the population of a country or city. A community of human societies or any community of human societies.

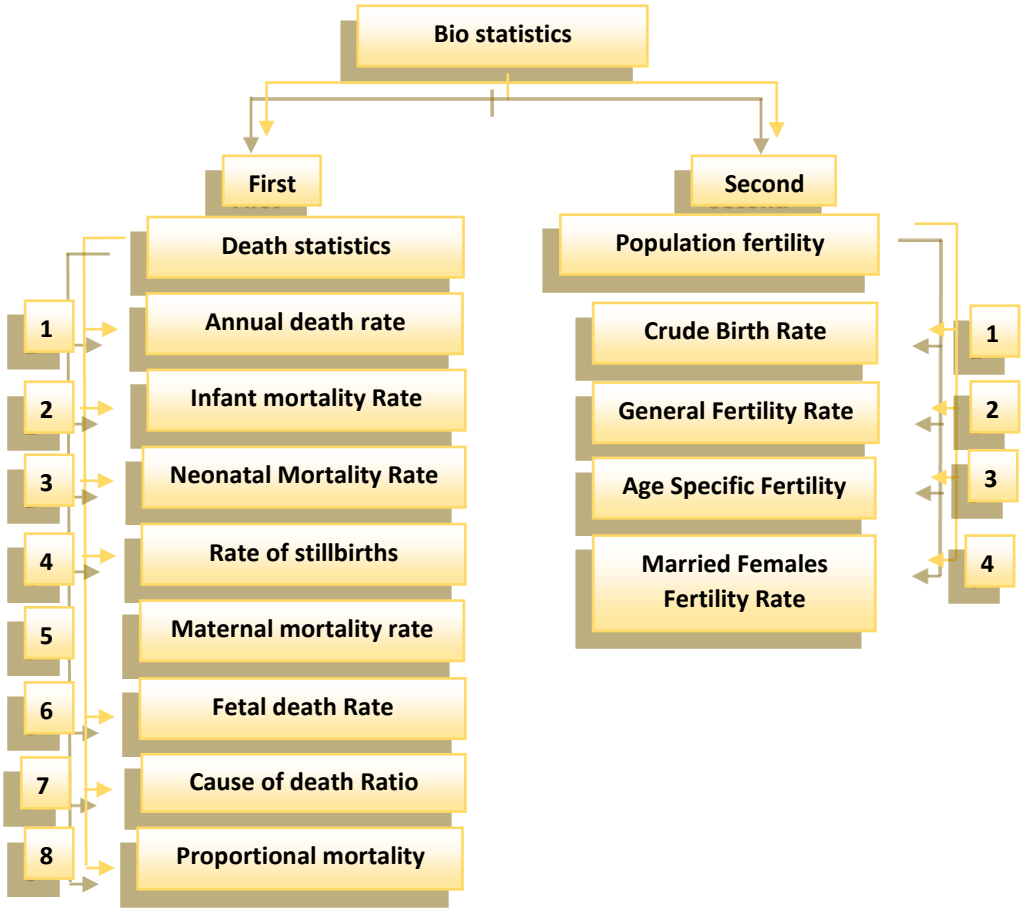
7.4.1. The importance of using bio statistics

1. Bio statistics are essential statistics through which population estimates can based mainly on the number of births and deaths as well as migration.
2. Bio statistics are a national registry of bio events that provide accurate statistics for decision makers and drawing public policies for all local and international users.
3. Bio statistics allow the peoples of different countries to obtain legal proof of their identity and to build a precise database of their characteristics.
4. In addition to the importance of bio statistics for governments and peoples, there is another importance which is the use of these statistics in scientific studies and research by academics and specialists.

7.4.2. Bio Statistics Measures

There are many measures used by statisticians to study population and biological phenomena to identify many health indicators that related to the natural increase of the population on

the one hand and the underlying causes leading to deaths on the other. The most important indicators and measures are:



7.4.2.1. Mortality statistics Measures

Mortality rates reflect the relative frequency of death within a particular community over a specified period time. The most important statistics are:

1. Annual Crude death Rate

$$= \frac{\text{Total number of deaths during the year}}{\text{Population in the middle of the year}} \times 1000$$

Example: The population of Iraq was 36,000,455 people in 2014 and the number of deaths in that year was 198820 people what the crude death rate for that year?

2. Annual crude death Rate:

$$= \frac{\text{Total deaths for 2014}}{\text{Total population of Iraq mid 2014}} \times 1000$$

$$= \frac{198820}{36000455} \times 1000$$

= 5.5 per 1,000 people

3. Infant mortality Rate**:**

$$= \frac{\text{Total Infant mortality in a given year}}{\text{Total number of live births in that year}} \times 1000$$

4. Neonatal Mortality Rate:

$$= \frac{Total\ number\ of\ neonatal\ deaths\ (less\ than\ one\ month)\ in\ a\ particular\ year}{\text{Total number of live births in that year}}$$

$\times$ 1000

5. Fetal death Rate:

$$= \frac{\text{Total Inumber of Fetal death during a given year}}{\text{Total number of live births in that year}} \times 1000$$

Maternity death Rate:

$$= \frac{\text{Total number of maternal deaths due to pregnancy or childbirth within one year}}{\text{Total number of live births in that year}}$$

$\times$ 1000

6. Fetal death Rate:

$$= \frac{\text{Total number of fetal cases during a given year}}{\text{Total number of live births in that year}} \times 1000$$

Example: The number of live babies born in a country is 320,000 children in 2014 and the number of dead births is 8350 children, and the number of mothers died due to pregnancy or childbirth in

the same year 12,800. Also the number of deaths of children under the age of less than one year was 6000 children, including 320 newborns aged less than 28 days in that year also, while the number of cases of fetal 80 cases in the same year.

Calculate:

a. Infant mortality Rate
b. Neonatal Mortality Rate
c. Foetal death Rate
d. Maternity death Rate
e. Fetal death Rate

Solution:

a. Infant mortality Rate

$$= \frac{\text{Number of infant mortality less than one year} - \text{Number of neonatal deaths less than 28 days}}{\text{Total number of live births in that year}} \times 1000$$

$$= \frac{320-6000}{320000} \times 1000$$

=17.75 per 1,000 people

b. Neonatal Mortality Rate

$$= \frac{Total\ number\ of\ \text{neonatal mortality}\ during\ a\ given\ year}{\text{Total number of live births in that year}} \times 1000$$

$$\frac{320}{320000} \times 1000 = 1 \text{ Per 1,000 people}$$

c. Foetal death Rate

$$= \frac{\text{Total number of Foetal deathduring the year}}{\text{Total number of live births in same year}} \times 1000$$

$$\frac{8350}{320000} \times 1000 = 26 \text{ Per 1,000 people}$$

d. Maternity death Rate

$$= \frac{\text{Total number of maternal deaths due to pregnancy or childbirth during the year}}{\text{Total number of live births in same year}} \times 1000$$

$$= \frac{12800}{320000} \times 1000 = 40 \text{ Per 1,000 people}$$

e. Fetal death Rate

$$= \frac{\text{Total number of fetal cases during a given year}}{\text{Total number of live births in that year}} \times 1000$$

$$\frac{80}{320000} \times 1000 = 0.25 \text{ Per 1,000 people}$$

7. ***Cause of death Ratio***

The main purpose of calculating this ratio is the importance and severity of a particular disease in causing death for all deaths and whatever their causes. It calculated as follows:

$$= \frac{\text{Total number of deaths due to a specific disease during a given year}}{\text{Total number of deaths in the same year}} \times 1000$$

Example: The number of deaths from cancer in Iraq was 4831 in 2014, with a total number of total death for the same year at 198820 people.

The mortality rate due to illness (eg cancer)$\frac{4831}{198820} \times 100 = 2.4\ \%$

7- Proportional mortality Ratio

This ratio used as a measure to compare total health status among different communities. Calculated according to the following formula:

$$= \frac{\text{Total number of deaths aged 50 years and more in a given year}}{\text{Total number of deaths in the same year}} \times 1000$$

Example: The number of Iraq's population aged 50 old and more 6,884,520 in 2014, while the deaths in those ages 48,860 people and the total death in Iraq for the same year amounted to 198,820 people.

Find:

a. Proportional mortality Ratio

b. Mortality rate aged 50 years and over

Solution:

Proportional mortality Ratio $= \frac{48860}{198820} \times 1000 = 24.5\%$

b- Mortality rate aged 50 years and over

$$= \frac{\text{Total number of deaths aged 50 years and more in a given year}}{\text{The total population of Iraq is 50 years old and over in the same year}} \times 1000$$

$$= \frac{48860}{6884520} \times 1000 = 7 \text{ per } 1{,}000 \text{ people}$$

7.4.2.2. Population fertility Statistics Measures:

Fertility knowledge for women of childbearing age (15-45 years) and live birth rate in society is very important for population studies and for health sector practitioners to plan for the provision of services and health care for mothers, pregnant women and children to ensure a healthy generation and the most important standards Population fertility is:

1. Crude Birth Rate

$$= \frac{\text{Total number of live births in given year}}{\text{Total population in the middle of the same year}} \times 1000$$

Example: The number of live births for 2014 in a given city was 2,580 while the population was 1,273,000 in the middle of that year. The crude live birth rate is required.

$= \frac{2580}{1273000} \times 1000$ Crude live birth rate for 2014

= 2 per 1,000 people

2. General Fertility Rate

$$= \frac{\text{Total number of live births in given year}}{\text{The total number of women of pregnancy age } (15-45) \text{ in the middle of the same year}} \times 1000$$

Example: The number of live births in 2014 in the city was 6,200 children and the number of women of pregnancy age in the middle of that year 58,920 women. Find the general fertility rate.

Solution:

$\frac{6200}{58920} \times 1000$ General Fertility Rate for 2014

= 105 per 1,000 people

3. Age Specific Fertility

$$= \frac{\text{Total number of live births of women in a certain age group}}{\text{The total number of women in that category in the middle of the year}} \times 1000$$

Example: The following frequency distribution shows age groups, number of women and number of live births for each age group for 2014.

class /year	number of live births	number of women
20 – 25	**2,400**	**58,400**
25 – 30	**7,100**	**78,000**

Find:

a. Fertility rate for the age group 20-25

b. Fertility rate for the age group 25-30

c. Fertility rate for the age group 20-30

Fertility rate for the age group (20-25) $= \frac{2400}{58408} \times 1000$

=41 per 1,000 people

d. Fertility rate for the age group (25-30) = $\frac{7100}{78000} \times 1000$

= 91 per 1,000 people

e. Fertility rate for the age group (20-30)= $\frac{7100+2400}{78000+58408} \times 1000$

$= \frac{9500}{136408} \times 1000$

$= 69.6$ per 1,000 people

4- Married Females Fertility Rate

$$= \frac{\text{Total number of live births in given year}}{\text{The total number of married women in the middle of the year}} \times 1000$$

Example: the number of live births in one city was 2,892 in 2014, the number of married women in the middle of that year was 320,000, and the fertility rate of married females was required.

Married Females Fertility Rate $= \frac{2892}{320000} \times 1000$

= 9 per 1,000 married women

7.5. Hospital Statistics Measures

Hospitals (government and private), health centers and outpatient clinics are important sources of health statistics to plan

and follow up to provide of the best health services to citizens on the basis of accurate scientific. These statistics include:

1. Statistics related to hospital beds

The hospital beds that are statistically defined as those in the various hospital wards, which will be ready to be occupied by patients for 24 hours or more.

The planning for the construction of any hospital and any city should be subject to urban planning to determine the appropriate geographical location of the hospital and to determine the size of the hospital (number of beds), which should be commensurate with the population of the current area and the future expansion of the population of that area. There are many ways to estimate the number of hospital beds including:

A. The method of the rate to be achieved per 1,000 people.

In order to calculate the number of beds for any hospital, it is necessary to determine the required rate of the bed per 1,000 resident divided according to different medical specialties, which depends mainly on the type of diseases prevalent in the region and the expected number of visitors to that hospital. For example, if three bed per 1,000 people are determined according to the following medical specialties:

Internal Medicine 0.7, General Surgery 0.7, Orthopedics and Fracture 0.2,, Obstetrics and Gynecology 0.35, Children 0.4, Radiation Therapy 0.02, Geriatric Diseases 0.2, Ear, Nose and

Throat 0.06, Eyes 0.07, Salient Diseases 0.1, Mental and Psychological 0.2.

$$\therefore \text{Number of beds} = \frac{\text{bed rate} \times \text{Population of the region}}{1000}$$

Example: The population of a city was 46,668 people and the health authorities of that city wanted to achieve three beds per 1,000 people, what is the number of beds required to be placed in that hospital?

$$\therefore \text{Number of beds} = \frac{3 \times 46668}{1000} = 140\ bed$$

Thus, it can be said that the number of beds that should be present in the city hospital 140 beds to achieve the proportion of three beds per 1,000 people.

This number increases as the percentage increases and vice versa.

B. Method of hospitalization rate and rate of stay

Example: The city's population reached 32,500 people and reached the rate of hospitalization from reviewers for the hospital was 20% and per patient stay rate of 6 per day in bed while the number of days of use of the bed in the year 300 days. Required to estimate the number of beds that need this hospital actually.

$$\text{Number of hospital beds required} = \frac{6 \times 32500 \times 20}{300 \times 100} = 130\ bed$$

Thus, the hospital needs 130 beds to meet the needs of the city's residents.

2. Average Length of Stay

Which is the number of days spent by the patient in the hospital between the entry and exit? The day of entry is calculated as the day of stay in the hospital while the day of departure is not counted. The duration of the stay is one day if the patient enters and leaves the same day.

This measure is one of the important statistical measures used to assess the efficiency of hospital performance in the area of services provided to patients and the ability of the hospital to use its available facilities. If the average length of stay in the hospital exceeds the average stay in similar hospitals that provide the same services. It indicates a defect what medical Procedures nursing or administrative followed in this hospital. Therefore, the hospital administration must take all measures that will avoid these mistakes to improve the hospital's attitude towards the best and most reliable services. This scale is calculated according to the following formula:

$$\text{Average length of stay} = \frac{\textit{Number of days} \times \textit{patients discharged from hospital,including those who died within a period of time}}{\text{Number of patients discharged (including deceased) during the same period of time}}$$

As in the following example:

Example: The total days of patients who get out of a hospital 70,534 day in 2014, and the number of those who get out of the hospital (including deceased) was 6,410. Find the average length of stay in the hospital?

$$\text{The average length of stay} = \frac{70534}{6410}$$

= 11 day

3. *Bed Occupancy Percentage*

This measure is one of the most common statistical measures which is used in hospitals, as this measure gives the hospital management the ability to identify the extent of congestion or lack of congestion hospital patients. Its height indicates that the hospital is very overcrowded and its administration must find ways to increase the number of beds in the hospital, while its decline provides the opportunity to provide its services and receive emergency cases not only for its geographical area, but also for its neighboring areas. This ratio is calculated according to the following formula:

$$\text{Bed Occupancy Percentage} = \frac{\textit{The total number of days of hospitalization (stay) in hospital during a specified period of time}}{\text{Number of beds available in the hospital x Number of days in the same period}} \; x\ 100$$

Example: The number of beds in Baghdad hospital was 350 beds, and the whole beds was busy during the month of July. The total number of patients hospitalized in the month was 8,867 patients. Required Calculate the Bed Occupancy Percentage per day.

Solution:

$$\text{Bed Occupancy Percentage} = \frac{8687}{31 \times 350} X100$$

$$= \frac{868700}{10850}$$

=80 % Average bed Occupancy Percentage per day

This Percentage allows the hospital to receive more patients.

Example: Diyala Hospital had 200 beds in 2014, and the total days of patients' stay in the hospital during the same year was 80,324. What is the occupancy Percentage of this bed per day?

$$\text{Bed Occupancy Percentage} = \frac{80324}{365\times200} \times 100$$

$$= 110\ \%$$

This Percentage indicates that the momentum is very high on this hospital and requires the management of the hospital and the competent health authorities to increase the number of hospital beds to be able to overcome this momentum.

4. *Bed Turnover Interval*

Is the period in which the bed remains vacant between the exit of the patient who was taking the bed and the entry of another patient in it to take the same bed is calculated as follows:

$$\text{Bed Turnover during the year} = \frac{\textit{Number of hospital beds}\times365-\text{Number of sick days}}{\text{The number of departures (deaths among them) for the same year}} \times 100$$

If the number of days of this period is zero, this means that the occupancy rate of the bed was 100% and is negative if the occupancy rate is more than 100%.

The existence of a long period of turnover indicates a surplus of beds or may indicate a defect in the system of admission to the hospital. While a very short or negative period of turnover indicates, the number of beds in this hospital is insufficient.

Example: The number of the bed of one of Baghdad's hospitals of 350 beds in 2014, and the number of sick days for patients entering was 88,688 days, while the number of departures from the hospital for the same year was 9,580, including the deceased. What is the turnover in this hospital in 2014?

$$\text{Bed Turnover Interval} = \frac{88688-365\times350}{9580} = 4 \text{ day}$$

This is a very short period of turnover, indicating that the hospital beds are not adequate for patients.

5. Bed Turnover Rate

This rate shows the average number of patients who have been in bed for a certain period of time, usually one year. This rate is inversely proportional to the rate of hospital stay in the sense that the higher the rate of stay, the lower the rate of turnover of the bed and vice versa. The high bed turnover rate leads to the optimum use of hospital beds on the one hand and their potential on the other hand and is calculated as follows:

$$\text{Bed turnover rate} = \frac{Total\ number\ of\ patients\ be\ out\ from\ hospital\ during\ a\ given\ year}{Total\ hospital\ beds\ for\ the\ same\ year}$$

Example: The number of beds in Baghdad hospital was 350 beds in 2014, and the total number of patients who were released from this hospital was 9,580 patients for the same year. Calculate the bed turnover rate at this hospital.

$$\text{Bed turnover rate} = \frac{9580}{350} = 27 \text{ Bed turnover rate in 2014}$$

This means that the average number of sick people who occupied the beds was 27 in 2014, i.e. 2.25 patients per month, which is a low turnover reflecting the lack of optimal use of the hospital's beds and facilities.

5. ***Cesarean Section Rate***

It is the ratio of the number of caesarean operations that take place in the hospital within a certain period of time is the total number of births in the hospital during that period. It is calculated according to the following formula:

$$\text{Cesarean Section Rate} = \frac{\text{Total number of cesarean sections performed during a certain period of time}}{\text{Total number of births during that period}} \times 100$$

Example: The number of births at the Central Children's Hospital reached 398 births during February of 2015, including 12 cesarean births. Find the rate of Caesarean section for February ?

$$\text{Caesarean section rate for February} = \frac{12}{398} \times 100$$

= 3% Caesarean section rate

7. ***Autopsy Rate***

It will calculated as follows:

$$\text{Autopsy Rate} = \frac{\text{Total number of autopy of deceased entrants within a certain time period}}{\text{Total number of deaths of patients entering during that period}} \times 100$$

8. ***Post- operation infection Rate***

$$\text{Postoperation infection Rate} = \frac{\text{Total number of wound infections after surgical operations over a period of time}}{\text{Total number of surgeries that were made during that period during}} \times 100$$

9. *Consultation Rate*

It is one of the important statistical measures used in evaluating the performance of medical work in hospitals. The high rate of counseling indicates the care of the physician to give sufficient opportunity for the patient for the diagnosis of accurate and safe treatment. It calculated according to the following formula:

$$\text{Consultation Rate} = \frac{\text{The total number of medical consultations that have actually been done to patients within a certain period of time}}{\text{The total number of departures for the same period of time}} \times 100$$

Example: The number of medical consultations during the month of March 2015 in a hospital 219 consultations, and total out of the hospital was 842. What is the rate of consultation in this hospital?

$$\text{Consultation Rate} = \frac{219}{842} \times 100 \qquad = 26\%$$

7.6. Disease statistics measures:

These measures are used by health sector workers to analyze and evaluate the health status of the community in order to develop plans and policies necessary to develop this sector in line with the needs of the community. One of these measures are:

1. Incidence Rate:

It is calculated according to the following formula:

$$\text{Incidence Rate} = \frac{\textit{Total number of new incidences with a particular disease during the year}}{\text{Total population in mid year}} \times 100$$

This measure enables health workers to identify the incidence of new infections (chronic diseases and infectious

diseases) and thus help decision-makers to take the necessary measures to prevent the spread of the disease or at least reduce its potential damage.

Example: The number of people infected with HIV was 51 in 2014 in a city with a population of 1,165,000 in the middle of the same year. What is the rate of infection with this disease?

Incidence rate = $\frac{51}{1.265.000} \times 1000$

= 0.040 per 1,000 people infected with HIV

2. Prevalence Rate

This measure used to measure the prevalence of a particular disease (especially epidemics and chronic diseases) in a country. Which calculated according to the following formula:

Prevalence rate =

$$\frac{\text{Total number of cases of a particular disease in a given year}}{\textit{Total population in the middle of the same year}} \times 1000$$

Example: Health authorities in one country after the spread of cholera in 2013 reported that the number of infected people was 8,710 people and the population of that country in the middle of the same year was 1,231,830 people.

Solution:

Prevalence rate = $\frac{8710}{1.231.830} \times 1000$

= 7 per 1,000 people

3. *Case fatality Rate*

This rate is used to know the results of the implementation of the governmental program to combat a particular disease by the

competent health authorities, which can be for a year or several years. The failure of this program is to fight this disease which may lead to death, so this rate can reveal to us the seriousness of this disease on the public health of the members of the community.

This rate is calculated according to the following formula:

Fatality Rate =

$$\frac{\text{Total number of deaths due to infection during a year}}{\textit{The total number of cases diagnosed with this disease for the same year}} \times 1000$$

Example: The number of deaths from hepatitis virus in Iraq was 176 in 2014 the total number of infections with the disease the same one year was 2,510 infected. Find the fatality rate of loss of this disease.

Fatality rate = $\frac{176}{2510} \times 1000$=70 deaths occur for every 1000 people with hepatitis

4. *Immaturity Ratio*

This ratio is mainly used to measure newborn babies, but each weighs less than 2500 grams, divided by the total number of children born alive in the middle of the same year. This ratio calculated according to the following formula:

Immaturity ratio =

$$\frac{\text{Total number of } \textit{children} \text{ born alive weighing less than 2500 during the year}}{\textit{Total number of children born alive at the middle of the same year}} \times 1000$$

Example: The health reports of a city in 2013 showed that there were 114,240 children born in the city and found that there were 14,510 children weighing no more than 2,500 grams.

Find the Immaturity Ratio.

Solution: Immaturity rate =

$\frac{14510}{114248} \times 1000$ = 127 per 1,000 children born alive in 2013

This ratio is relatively large and requires the competent health authorities in this city to discuss by all possible means the reasons behind this ratio to avoid continuation or at least to reduce them.

References

English sources

1. B.Burt Gerstman, Basic Biostatistics, Statistic For Public Health Practice, University San Jose, California, 2008.
2. Chap T.LE, Introductory Biostatistics, Published by Jhon Wiley& Sons, Inc, Hoboken, New Jersey, 2003.
3. E. pecicain, dv. O. tanasyiou, Econometrie, Academia de study economic, Bucuresti, Romania, 1989.
4. Gerald Van Belle and others, Biostatistics, A Methodology for the Health Sciences, University of Washington, published by Jhon Willy & Sons, Inc, Hoboken, New Jersey, 2004.
5. Larry Winner, Introduction to Biostatistics, University of Florida, July8, 2004.
6. Roobent R. Sokal and Fisames Rohlf, Introduction To Biostatistics, State University Of New Yourk, Dover publication, Inc, Mineola, New yourk, 2009.

Arabic sources translated to English

1. Abdul Basset Mohamed Hassan, Principles of Social Research, 11th addition, Wahab Library, Cairo, 1990.
2. Abdul Latif Hassan Shoman, Introduction to Statistics and Statistical Conclusion, Al- Genan for Publishing and Distribution, Amman, 2009.
3. Abdul Hamid Walid, Abdel Majeed Hamza, Samples, Printing books for printing and publishing, Mosul University, 1981.

4. Adas, Abdul Rahman, Principles of Statistics in Education and Statistics, Al-Aqsa, Jordan, Amman, 1978.
5. Adnan Shehab Hamad, Mahdi Mohsen Al-Alaq, Methods of sampling in the field, Arab Institute for Training and Research, Baghdad, 2001.
6. Abu Amma, Abdul Rahman bin Mohammed Suleiman, Hindi, Mahmoud Mohammed Ibrahim, Applied Statistics, Obeikan Library, Al-Riyadh, 2007.
7. Abu Saleh, Mohammed Subhi, Statistical Methods, Dar Al-Yazourdi Scientific for Publishing and Distribution, Amman, Jordan, 2009.
8. Al-Baldawi, Abdul-Hamid Abdel-Majid, Statistics for Economic Sciences and Business Administration, Published by Dar Wael for Publishing and Distribution , Amman, Jordan, 2009.
9. Al-Abbasi, Abdul Hamid Mohammed, Statistical method of analysis and interpretation using Computer and SPSS Program, Cairo University, institute of Statistical Studies and Research, Department of Biostatistics and Population, 2013.
10. Al-Atbi, Sami Azeez Abbas, Al-Hiti, Mohammed Yousef Hajim, Research methodology, concept, methods and writing, Al-Asdekaa Printer, Baghada, 2011.

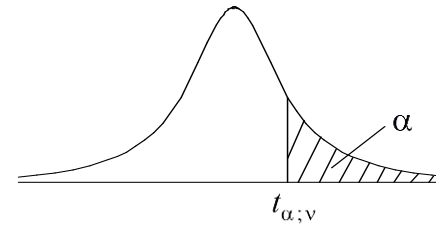
11. Al-Atbi, Sami Azeez Abbas, Al-Tai, Iyad Ashour, Statistics and modeling in geography, Akram Printing and Reproduction Press, Baghdad, 2013.
12. Al-Kade, Dalal, Suhaila Abdullah, Mahmoud al-Bayati, Statistics for managers and economists, Dar Al-Jamed for Publishing and Distribution, Amman, Jourdan, 2005.
13. Al-Nasser, Abdul Majeed Hamza and others, Entrance to Statistical Thought, Al- Thakera for dissemination and distribution, Iraq, Baghdad, 2012.
14. Al-Qurashi, Ihsan Kazem Sharif, Formal methods and informal methods in statistical tests, Al- Diwani Printer, Baghdad, 2007.
15. Al-Rubaie, Adnan Shukri, Introduction to Health and Life Statistics, 2nd addition, Baghdad, 1988.
16. Al-Rawee, Khasha Mahmoud,Entrance to Statistics, Mosul University, 1984.
17. Al-Saleh, Nasser Abdullah, Mohammed Mahmoud Syriani, Quantitative and Statistical Geography, Umm Al-Qura University, Mecca, 2nd addition, Obeikan Library, 1999.
18. Al-Safawi, Dia Younis, Statistic, Ministry of Higher Education and Scientific Research, Mosul University, 2008.
19. Berri, Adnan Majid Abdul Rahman, Hindi, Mahmoud Mohamed Ibrahim, Statistics and Probability, 4th addition, King Saud University, 2003.

20. Dominic Salvatore, Shum Series, Statistics and econometrics, Ain Shams University, Egypt, 1999.
21. Fathi Abdel Aziz Abu Radi, Quantitative Methods in Geography, Dar Al- Ma'arif Al- Gameea, Alexandria, 2002.
22. Hassan Ali Musa, Quantitative Methods in Geography, Damascus University, 2007.
23. Khalid Mohammed Daoud, Zaki Abdel-Elias, Statistical Methods of Agricultural Research, Ministry of Higher Education and Scientific Research, Mosul University, 1990.
24. Odda, Ahmed Odeh bin Abdul Hamid, Al-Kade, Mansour bin Abdul Rahman,Descriptive and deductive statistics, Al-Falah for Publishing and Distribution, Saudi Arab,2002.
25. Shehadeh, Numan, Statistical Analysis in Geography and Social Sciences, Dar Al-Safaa for Publishing and Distribution, Amman, Jordan, 2011.
26. Sharafuddin Khalil, Descriptive statistics, Library of the network of research and economic studies, 2012.
27. Shreiji, Abdul Razzaq, Khalid Al-Mulla, Descriptive statistics, Dar of science for millions, Lebanon, 1987.
28. Sahoki, Medhat, Karim Mohamed Wahib, Applications in designing and analyzing experiments, Baghdad University, 1990.
29. Tiba, Ahmed Abdel Samie, Principles of Statistics, Dar Al-Badaia, Jordan, 2007.

Table of the Student's t-distribution

The table gives the values of $t_{\alpha;\nu}$ where

$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



α ν	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Values of the Chi-squared distribution

DF	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548
7	0.989	1.690	9.803	12.017	14.067	16.013	16.622	18.475	20.278
8	1.344	2.180	11.030	13.362	15.507	17.535	18.168	20.090	21.955
9	1.735	2.700	12.242	14.684	16.919	19.023	19.679	21.666	23.589
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188
11	2.603	3.816	14.631	17.275	19.675	21.920	22.618	24.725	26.757
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.300
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32.000	34.267
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718
18	6.265	8.231	22.760	25.989	28.869	31.526	32.346	34.805	37.156
19	6.844	8.907	23.900	27.204	30.144	32.852	33.687	36.191	38.582
20	7.434	9.591	25.038	28.412	31.410	34.170	35.020	37.566	39.997
21	8.034	10.283	26.171	29.615	32.671	35.479	36.343	38.932	41.401
22	8.643	10.982	27.301	30.813	33.924	36.781	37.659	40.289	42.796
23	9.260	11.689	28.429	32.007	35.172	38.076	38.968	41.638	44.181
24	9.886	12.401	29.553	33.196	36.415	39.364	40.270	42.980	45.559
25	10.520	13.120	30.675	34.382	37.652	40.646	41.566	44.314	46.928
26	11.160	13.844	31.795	35.563	38.885	41.923	42.856	45.642	48.290
27	11.808	14.573	32.912	36.741	40.113	43.195	44.140	46.963	49.645
28	12.461	15.308	34.027	37.916	41.337	44.461	45.419	48.278	50.993
29	13.121	16.047	35.139	39.087	42.557	45.722	46.693	49.588	52.336
30	13.787	16.791	36.250	40.256	43.773	46.979	47.962	50.892	53.672
31	14.458	17.539	37.359	41.422	44.985	48.232	49.226	52.191	55.003
32	15.134	18.291	38.466	42.585	46.194	49.480	50.487	53.486	56.328
33	15.815	19.047	39.572	43.745	47.400	50.725	51.743	54.776	57.648
34	16.501	19.806	40.676	44.903	48.602	51.966	52.995	56.061	58.964
35	17.192	20.569	41.778	46.059	49.802	53.203	54.244	57.342	60.275
36	17.887	21.336	42.879	47.212	50.998	54.437	55.489	58.619	61.581
37	18.586	22.106	43.978	48.363	52.192	55.668	56.730	59.893	62.883
38	19.289	22.878	45.076	49.513	53.384	56.896	57.969	61.162	64.181
39	19.996	23.654	46.173	50.660	54.572	58.120	59.204	62.428	65.476
40	20.707	24.433	47.269	51.805	55.758	59.342	60.436	63.691	66.766
41	21.421	25.215	48.363	52.949	56.942	60.561	61.665	64.950	68.053
42	22.138	25.999	49.456	54.090	58.124	61.777	62.892	66.206	69.336
43	22.859	26.785	50.548	55.230	59.304	62.990	64.116	67.459	70.616
44	23.584	27.575	51.639	56.369	60.481	64.201	65.337	68.710	71.893

45	24.311	28.366	52.729	57.505	61.656	65.410	66.555	69.957	73.166
46	25.041	29.160	53.818	58.641	62.830	66.617	67.771	71.201	74.437
47	25.775	29.956	54.906	59.774	64.001	67.821	68.985	72.443	75.704
48	26.511	30.755	55.993	60.907	65.171	69.023	70.197	73.683	76.969
49	27.249	31.555	57.079	62.038	66.339	70.222	71.406	74.919	78.231
50	27.991	32.357	58.164	63.167	67.505	71.420	72.613	76.154	79.490
51	28.735	33.162	59.248	64.295	68.669	72.616	73.818	77.386	80.747
52	29.481	33.968	60.332	65.422	69.832	73.810	75.021	78.616	82.001
53	30.230	34.776	61.414	66.548	70.993	75.002	76.223	79.843	83.253
54	30.981	35.586	62.496	67.673	72.153	76.192	77.422	81.069	84.502
55	31.735	36.398	63.577	68.796	73.311	77.380	78.619	82.292	85.749
56	32.490	37.212	64.658	69.919	74.468	78.567	79.815	83.513	86.994
57	33.248	38.027	65.737	71.040	75.624	79.752	81.009	84.733	88.236
58	34.008	38.844	66.816	72.160	76.778	80.936	82.201	85.950	89.477
59	34.770	39.662	67.894	73.279	77.931	82.117	83.391	87.166	90.715
60	35.534	40.482	68.972	74.397	79.082	83.298	84.580	88.379	91.952
61	36.301	41.303	70.049	75.514	80.232	84.476	85.767	89.591	93.186
62	37.068	42.126	71.125	76.630	81.381	85.654	86.953	90.802	94.419
63	37.838	42.950	72.201	77.745	82.529	86.830	88.137	92.010	95.649
64	38.610	43.776	73.276	78.860	83.675	88.004	89.320	93.217	96.878
65	39.383	44.603	74.351	79.973	84.821	89.177	90.501	94.422	98.105
66	40.158	45.431	75.424	81.085	85.965	90.349	91.681	95.626	99.330
67	40.935	46.261	76.498	82.197	87.108	91.519	92.860	96.828	100.554
68	41.713	47.092	77.571	83.308	88.250	92.689	94.037	98.028	101.776
69	42.494	47.924	78.643	84.418	89.391	93.856	95.213	99.228	102.996
70	43.275	48.758	79.715	85.527	90.531	95.023	96.388	100.425	104.215
71	44.058	49.592	80.786	86.635	91.670	96.189	97.561	101.621	105.432
72	44.843	50.428	81.857	87.743	92.808	97.353	98.733	102.816	106.648
73	45.629	51.265	82.927	88.850	93.945	98.516	99.904	104.010	107.862
74	46.417	52.103	83.997	89.956	95.081	99.678	101.074	105.202	109.074
75	47.206	52.942	85.066	91.061	96.217	100.839	102.243	106.393	110.286
76	47.997	53.782	86.135	92.166	97.351	101.999	103.410	107.583	111.495
77	48.788	54.623	87.203	93.270	98.484	103.158	104.576	108.771	112.704
78	49.582	55.466	88.271	94.374	99.617	104.316	105.742	109.958	113.911
79	50.376	56.309	89.338	95.476	100.749	105.473	106.906	111.144	115.117
80	51.172	57.153	90.405	96.578	101.879	106.629	108.069	112.329	116.321
81	51.969	57.998	91.472	97.680	103.010	107.783	109.232	113.512	117.524
82	52.767	58.845	92.538	98.780	104.139	108.937	110.393	114.695	118.726
83	53.567	59.692	93.604	99.880	105.267	110.090	111.553	115.876	119.927
84	54.368	60.540	94.669	100.980	106.395	111.242	112.712	117.057	121.126
85	55.170	61.389	95.734	102.079	107.522	112.393	113.871	118.236	122.325
86	55.973	62.239	96.799	103.177	108.648	113.544	115.028	119.414	123.522
87	56.777	63.089	97.863	104.275	109.773	114.693	116.184	120.591	124.718
88	57.582	63.941	98.927	105.372	110.898	115.841	117.340	121.767	125.913
89	58.389	64.793	99.991	106.469	112.022	116.989	118.495	122.942	127.106
90	59.196	65.647	101.054	107.565	113.145	118.136	119.648	124.116	128.299
91	60.005	66.501	102.117	108.661	114.268	119.282	120.801	125.289	129.491
92	60.815	67.356	103.179	109.756	115.390	120.427	121.954	126.462	130.681

93	61.625	68.211	104.241	110.850	116.511	121.571	123.105	127.633	131.871
94	62.437	69.068	105.303	111.944	117.632	122.715	124.255	128.803	133.059
95	63.250	69.925	106.364	113.038	118.752	123.858	125.405	129.973	134.247
96	64.063	70.783	107.425	114.131	119.871	125.000	126.554	131.141	135.433
97	64.878	71.642	108.486	115.223	120.990	126.141	127.702	132.309	136.619
98	65.694	72.501	109.547	116.315	122.108	127.282	128.849	133.476	137.803
99	66.510	73.361	110.607	117.407	123.225	128.422	129.996	134.642	138.987
100	67.328	74.222	111.667	118.498	124.342	129.561	131.142	135.807	140.169

F Table for alpha=.05

DF2	1	2	3	4	5	6	7	8	9	10
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	1.7306	1.6717
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229	2.1768
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772

F Table for alpha=.01

DF2 /DF1	1	2	3	4	5	6	7	8	9	10
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801