Fundamental of Statistics

Author: Poulino Francis Deng Ajang.

Contact: +211918148552   +211925387833

Email: poulino.fashoda2008@gmail.com

PREFACE

Business Managements today make daily decisions on many issues, such as how much and where to produce, for which market, what prices to set, and how much stock to keep. Mathematical Statistics models can help to make the best decisions, among the possible alternatives.

The objective of this book Fundamentals of Statistics is to present a survey of selected Statistical methods, and some of their applications to managerial decision making, concerning production, inventory, distribution, and integrated supply statistics methods. Also included are lists of some examples in our daily life.

Statistics has been termed as the Science of Collecting and Analyzing Information in a better way for decision making which will give bright future. A problem in the real world is how to have correct information, usually in statistics or mathematical terms, and then mathematical techniques, together with data analysis and computational statistics, are applied, in order to find ways to do the job better. The word Statistics derives from the many successful applications of science to military operations in the 1940s. Nevertheless, since then, most statistics applications have been to peaceful activities, especially to business management, of which planning industrial production, and scheduling airlines, and other transportation, have been prominent. The name Statistics Management Science denotes the same discipline, with some emphasis on business management. Practitioners of Management will find many of these techniques relevant. The areas of Logistics, Supply Chain Management, Decision Sciences, and Manufacturing Management deal with similar applications of Statistics.

ABSTRACT

The book will give learners an overview of the statistical stepwise approach to understand ideally identify critical points for statistical taught and actions starting from definitions, scope of the statistics whereby applying thus role within the statistical systems to improve the understanding of the student.

The fundamental of statistics deals with simple introduction to statistics, starting with definition of statistics, data presentation, and measures of central tendency, variation, and statistical concepts in general, and fields of an application in other sciences and limitations of statistics, followed by various issues as regression, correlation, and moments. It is a continuation of introduction to statistics, is concerned with simple or elegant mathematical idealizations – the world is full of unpredictability, uncertainty, randomness. In the modern world, computers, and information technology to become widely to everybody, the importance of statistics is very well recognized by all the disciplines. Statistics has originated as a science of statehood and found applications slowly and steadily in Agriculture, Economics, Commerce, Biology, Medicine, Industry, Planning, education and so on.

Keywords: the statistical knowledge is widely used in many fields of live, whereby biostatistics become more interesting for biological factors, computerizing more knowledge by analytical test our life.

Corresponding author: poulino.fashoda2008@gmail.com

CHAPTER ONE

OVERVIEW, OBJECTIVES, DEFINITIONS, CONCEPTS AND SCOPE OF STATISTICS

Overview of Statistics

The statistics deals with simple introduction to statistics, starting with definition of statistics, data presentation, and measures of central tendency, variation, and statistical concepts in general, and fields of an application in other sciences and limitations of statistics, followed by various issues as regression, correlation and moments. It is a continuation of introduction to statistics, is concerned with simple or elegant mathematical idealizations – the world is full of unpredictability, uncertainty, randomness. In the modern world, computers, and information technology to become widely to everybody, the importance of statistics is very well recognized by all the disciplines. Statistics has originated as a science of statehood and found applications slowly and steadily in Agriculture, Economics, Commerce, Biology, Medicine, Industry, Planning, education and so on. As on date there is no other human occupation, where statistics cannot be applied.[1]

The word 'Statistics' and 'Statistical' are all derived from the Latin word Status, means a political state. The theory of statistics as a distinct branch of scientific method is of comparatively recent growth. Research particularly into the mathematical theory of statistics is rapidly proceeding and fresh discoveries are being made all over the world.

Objective of Statistics

The objective of the Statistics is to build skills of the student's concepts on statistics to create the fields of knowledge about the statistics and contribute into society as the solving their problem issues at the local and international levels.

Meaning of Statistics

Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting, and analyzing data as well as deriving valid conclusions and making reasonable

[1] Masembe Kabali 2011. Basic Business Statistics, third edition, P.O. Box 7453 Kapala – Uganda. Type set, printed and Bound by M. Kaabali and Basic Business Statistics Books, p 1.

decisions based on this analysis. Statistics is concerned with the systematic collection of numerical data and its interpretation.[2]

The word '**statistic**' used to refer to the following:

- Numerical facts, such as the number of people living in particular area.
- The study of ways of collecting, analyzing, and interpreting the facts.

Definitions of Statistics:

Different authors define statistics differently over a period:

First, in the olden days, statistics was confined to only state affairs but in modern days, it embraces almost every sphere of human activity. Therefore, several old definitions, which was confined to narrow field of enquiry, were replaced by more definitions, which are much more comprehensive and exhaustive.

Secondly, statistics has been defined in two different ways – Statistical data and statistical methods. The following are some of the definitions of statistics as numerical data.

- Statistics are the classified facts representing the conditions of people in a state. It can be listed in numbers or in tables of numbers or in any tabular or classified arrangement.
- Statistics are measurements, enumerations or estimates of natural phenomenon usually systematically arranged, analyzed, and presented as to exhibit important interrelationships among them.

Statistics is the methodology for collecting, analyzing, interpreting, and drawing conclusions from information. It is the methodology, which scientists and mathematicians have developed for interpreting and drawing conclusions from collected data.

Bowley defined statistics as the science of counting in one of the departments, obviously this is an incomplete definition as it considers only the aspect of collection and ignores other aspects such as analysis, presentation, and interpretation. Statistics may be rightly called the scheme of averages. This definition is also incomplete, as averages play an important role in understanding and comparing data and statistics provide more measures.

Croxton and Cowden defined statistics as the science of collection, presentation analysis and interpretation of numerical data from the logical analysis. The definition of statistics by Croxton

[2] N. G. Das 2009. Statistical Method the combined Edition (Volumes I & II). McGraw Hill Education (India) Private Limited, p, 1.

and Cowden is the most scientific and realistic one. According to this definition, there are four stages[3]:

- **Collection of Data:** It is the first step and this is the foundation upon which the entire data set. Careful planning is essential before collecting the data. There are different methods of collection of data such as census, sampling, primary, secondary, etc., and the investigator should make use of correct method.
- **Presentation of Data:** The mass data collected should be presented in a suitable, concise form for further analysis. The collected data may be presented in the form of tabular, diagrammatic, or graphic form. m
- **Analysis of Data:** The data presented should be carefully analyzed for making inference from the presented data such as measures of central tendencies, dispersion, correlation, regression etc.,
- **Interpretation of Data:** The final step is drawing conclusion from the data collected. A valid conclusion must be drawn because of analysis. A high degree of skill and experience is necessary for the interpretation.
- **Horace Secrist** defined statistics as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated, or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other.

Functions of Statistics

There are many functions of statistics. Let us consider the following five important functions.

- **Condensation**

  Condense means to reduce or to lessen. Condensation is mainly applied at embracing the understanding of a huge mass of data by providing only few observations. If in a particular class in **Manara Academic Secondary School (MASS) in Juba**, only marks in an examination are given, no purpose will be served, instead, if we are given the average mark in that examination, it serves the better purpose. Similarly, the range of marks is also another measure of the data. Thus, Statistical measures help to reduce the complexity of the data and consequently to understand any huge mass of data.

[3] Masembe Kabali 2011. Basic Business Statistics, third edition, P.O. Box 7453 Kampala – Uganda. Type set printed and Bound by M. Kaabali and Basic Business Statistics Books, p4.

- **Comparison**

  Classification and tabulation are the two methods that used to condense the data. They help us to compare data collected from different sources. Grand totals, measures of central tendency measures of dispersion, graphs and diagrams, coefficient of correlation…etc, provide ample scope for comparison.

- **Forecasting**

  Forecasting means to **predict** or to **estimate beforehand**. It is possible to predict or forecast the rainfall for the near future. In business also forecasting plays a dominant role in connection with production and product, sales, profits etc. The analysis of **time series** and **regression analysis** plays an important role in forecasting.

- **Estimation**

  One of the main objectives of statistics is drawn inference about a population from the analysis for the sample drawn from that population. The four major branches of statistical inference are:

  i. Estimation theory
  ii. Tests of Hypothesis
  iii. Non-Parametric tests
  iv. Sequential analysis.

  Estimation theory, estimate the unknown value of the population parameter based on the sample observations. Suppose we are given a sample of heights of hundred students in a school, based upon the heights of these 100 students, it is possible to estimate the average height of all students in that school.

- **Tests of Hypothesis**

  A statistical hypothesis is some statement about the probability distribution, characterizing a population based on the information available from the sample observations. In the formulation and testing of hypothesis, statistical methods are extremely useful. Whether **crop yield** has increased because of the use of **new fertilizer** or whether the new **medicine is effective** in eliminating a particular disease are some examples of statements of hypothesis and proper statistical tools test these.

Population

Population is a complete set of all possible observations of the type, which is to be investigated[4]. Total number of students studying in a school or college, total number of books in a library, total number of houses in a village or town are some examples of population.

Population and Sample.

Population can be characterized as the set of individual persons or objects in which an investigator is primarily interested during his or her research problem. Sometimes wanted measurements for all individuals in the population are obtained, but often only a set of individuals of that population are observed; such a set of an individual's constitutes a sample. This gives us the following definitions of population and sample.

Definition of Population

Population is the collection of all individual or items under consideration in a statistical study.

Definition of Sample

Sample is that part of the population from which information is collected. (Weiss, 1999)

Finite population and infinite population

A population is said to be finite if it consists of finite number of units. Numbers of workers in a factory, production of articles in a particular day for a company are examples of finite population. The total number of units in a population is called population size. A population is said to be infinite if it has infinite number of units.

For example, the number of stars in the sky, the number of people seeing the Television programmed etc.

Sampling

The theory of sampling has been developed recently but this is not new. In our everyday life, we have been using sampling theory as we have discussed in introduction. In all those cases, we believe that the samples give a correct idea about the population. Most of our decisions are based on the examination of a few items that is sample studies.

Sample

Sample is a portion chosen from the population. The number of units in a sample is called the Sample Size.

[4] N. G. Das 2009. Statistical Method the combined Edition (Volumes I & II). McGraw Hill Education (India) Private Limited, p,6.

Sampling Unit

The constituents of a population which are individuals to be sampled from the population and cannot be further subdivided for the sampling at a time are called sampling units. For example, to know the average income per family, the head of the family is a sampling unit. To know the average yield of rice, each farm owner's yield of rice is a sampling unit.

Sampling Frame

For adopting any sampling procedure, it is essential to have a list identifying each sampling unit by a number. Such a list or map is called sampling frame. A list of voters, a list of householders, a list of villages in a district, a list of farmers etc. are a few examples of sampling frame.

The field of statistics is concerned with the collection, description, and interpretation of data. (Data are numbers obtained through measurement) In the field of statistics, the term "statistic" denotes a measurement taken on a sample (as opposed to a population). In general, conversation, "statistics" also refers to data.

The field of statistics: The study and use of theory and methods for the analysis of data arising from random processes or phenomena. The study of how we make sense of data. With **Statistics**, you go from observed data to generalizations about how the world works. For example, if we observe that the seven hottest years on record occurred in the most recent decade, we may conclude (perhaps without justification) that there is global warming.

Statistical Methods can be used to find answers to the questions like:

- What kind and how much data need to be collected.
- How should we organize and summarize the data?
- How can we analyze the data and draw conclusions from it?
- How can we assess the strength of the conclusions and evaluate their uncertainty?

That is Statistics Provides Methods for

- **Design:** Planning and carrying out research studies.
- **Description:** Summarizing and exploring data.
- **Inference:** Making predictions and generalizing about phenomena represented by the data.

Statistics in practice is applied successfully to study the effectiveness of medical treatments, the reaction of consumers to television advertising, the attitudes of young people toward sex and marriage, and much more. It is safe to say that nowadays statistics is used in every field of science.

Statistics in Practice: Consider the following problems:

- **Agricultural problem:** Is new grain seed or fertilizer more productive?
- **Medical problem:** What is the right amount of dosage of drug to treatment?
- **Political science:** How accurate are the gall-ups and opinion polls?
- **Economics:** What will be the unemployment rate next year?
- **Technical problem:** How to improve quality of product?

Target Population and Samples.

Population and sample are two basic concepts of statistics. Population can be characterized as the set of individual persons or objects in which an investigator is primarily interested during his or her research problem. Sometimes wanted measurements for all individuals in the population are obtained, the area where you get these population is called the target population, but often only a set of individuals of that population are observed; such a set of individuals constitutes a sample. This gives us the following definitions of population and sample.

A population is the set of all measurements of interest to a researcher. Typically, the population is not observed, but we wish to make statements or inferences concerning it. Populations can be thought of as existing or conceptual. Existing populations are well–defined sets of data containing elements that could be identified explicitly.

Samples are observed sets of measurements that are subsets of a corresponding population.

Samples are used to describe and make inferences concerning the populations from which they arise.

Statistical methods are based on these samples having been taken at random from the population. However, in practice, this is rarely the case. We will always assume that the sample is representative of the population of interest.

Parameters and Statistics

A parameter is a characteristic of a population, and a statistic is a characteristic of a sample. Since samples are subsets of population, statistics provide estimates of the parameters. That is, when the parameters are unknown, they are estimated from the values of the statistics.

[N, μ, s, are the standard symbols for the size, mean, S.D, of population. n,$\overline{X}$, s, are the standard symbol for the size, mean, S.D of sample respectively].

Variables

Variable is any characteristic that varies from one individual member of the population to another. Examples of variables for humans are height, weight, and number of siblings, sex, marital status, and eye color. The first three of these variables yield numerical information (yield numerical measurements) and are examples of quantitative (or numerical) variables; last three yield non-numerical information (yield non-numerical measurements) and are examples of qualitative (or categorical) variables.

Types of Variables (Data).

Quantitative and qualitative variables.

Data are observations of random variables made on the elements of a population or sample.

- **Quantitative variable** is that which can be expressed numerical e.g., Age and weight. Quantitative variable is divided into discrete and continuous.
  - **a. Discrete quantitative variable.** Discrete quantitative variable is only assuming integer values i.e., whole number e.g., the number of people in the Republic of South Sudan is a discrete value.
  - **b. Continuous quantitative variable.** It is that which in theory assume any number over different range. E.g., age and weight of people.
- **Qualitative variable.** Is that which can be described in terms of a set of categories. E.g., sex and marital status.
- The word **data** is plural, **datum** is singular. A collection of data is often called a data set singular.
- **Rounding of data.**

Suppose you had 15.78 and you want to round it to one decimal place, this equals 15.8 and 15.72 equals 15.7. When rounding the number five (5), the rule is to look to the number preceding the 5 if it is odd, round the 5 to that number if it is even forgotten of the number 5.

Statistics and Other Sciences

Statistics is not a mere device for collecting numerical data, but as a means of developing sound techniques for their handling, analysing, and drawing valid inferences from them. Statistics is applied in every sphere of human activity – social as well as physical – like Biology, Commerce, Education, Planning, Business Management, Information Technology, etc. It is almost impossible to find a single department of human activity where statistics cannot be applied. We now briefly discuss the applications of statistics in other disciplines.

1. **Statistics and Economics**

   Statistical methods are useful in measuring numerical changes in complex groups and interpreting collective phenomenon. Nowadays the uses of statistics are abundantly made in any economic study. Both in economic theory and practice, statistical methods play an important role.

   Alfred Marshall said, “Statistics are the straw only which I like every other economist has to make the bricks”. It may also be noted that statistical data and techniques of statistical tools are immensely useful in solving many economic problems such as wages of the worker, prices of goods in the market, in the production farms, distribution of income and wealth and so on. Statistical tools like Index numbers, time series Analysis, estimation theory; Testing Statistical Hypothesis is extensively used in economics.

2. **Statistics and Industry**

   Statistics is widely used in many industries. In industries, control charts are widely used to maintain a certain quality level. In production engineering, to find whether the product is conforming to specifications or not, statistical tools, namely inspection plans, control charts, etc., are of extreme importance. In inspection plans, we must resort to some kind of sampling – a very important aspect of Statistics.

3. **Statistics and Commerce**

   Statistics are lifeblood of successful commerce. Any businessperson cannot afford to either by under stocking or having overstock of his goods. In the beginning, he estimates the demand for his goods and then takes steps to adjust with his output or purchases. Thus, statistics is indispensable in business and commerce. As so many multinational companies have invaded into our South Sudan economy, the size and volume of business is increasing. On one side, the stiff competition is increasing whereas on the other side, the tastes are changing, and new fashions are emerging. In this connection, market survey plays an important role to exhibit the present conditions and to forecast the likely changes in future.

4. **Statistics and Agriculture**

   Analysis of variance (ANOVA) is one of the statistical tools developed by Professor **R.A. Fisher**, plays a prominent role in agriculture experiments. In tests of significance based on small samples in agricultural farm, it can be shown that statistics is adequate to test the significant difference between two sample means. In analysis of variance, we are concerned with the testing of equality of several population means. For an example, five fertilizers are applied to five plots each of wheat and the yield of wheat on each of the plots is given. In such a situation, we are interested in finding out whether the effect of these fertilisers on the yield is significantly different or not. In other words, whether the samples are drawn from the same normal population or not. The answer to this problem is provided by the technique of ANOVA and it is used to test the homogeneity of several population means.

5. **Statistics and Education**

   Statistics is widely used in education. Research has become a common feature in all branches of activities. Statistics is necessary for the formulation of policies to start new course, consideration of facilities available for new courses etc. There are many people engaged in research work to test the past knowledge and evolve new knowledge. These are possible only through statistics.

6. **Statistics and Planning**

   Statistics is indispensable in planning. In the modern world, which can be termed as the "world of planning", almost all the organisations in the government are seeking the help of planning for efficient working, for the formulation of policy decisions and execution of the same. To achieve the above goals, the statistical data relating to production, consumption, demand, supply, prices, investments, income expenditure … etc and various advanced statistical techniques for processing, analysing and interpreting such complex data are of importance. In India, statistics play an important role in planning, commissioning both at the central and state government levels.

7. **Statistics and Medicine**

   In Medical sciences, statistical tools are widely used. To test the efficiency of a new drug or medicine, t - test is used or to compare the efficiency of two drugs or two medicines, t-test for the two samples is used. More and more applications of statistics are at present used in clinical investigation.

8. **Statistics and Modern Applications**

Recent developments in the fields of computer technology and information technology have enabled statistics to integrate their models and thus make statistics a part of decision-making procedures of many organisations. There are so many software packages available for solving design of experiments, forecasting simulation problems etc.

SYSTAT, a software package offers mere scientific and technical graphing options than any other desktop statistics package. SYSTAT supports all types of scientific and technical research in various diversified fields as follows:

- **Archelogy:** Evolution of skull dimensions
- **Epidemiology:** Tuberculosis
- **Statistics:** Theoretical distributions
- **Manufacturing:** Quality improvement
- **Medical research:** Clinical investigations.
- **Geology:** Estimation of Uranium reserves from ground water.

Limitations of Statistics

Statistics with all its wide application in every sphere of human activity has its own limitations. Some of them are given below.

1. **Statistics is not suitable to the study of qualitative phenomenon.**

Since statistics is a science and deals with a set of numerical data, it is applicable to the study of only these subjects of enquiry, which can be expressed in terms of quantitative measurements. In fact, qualitative phenomenon like honesty, poverty, beauty, intelligence … etc. Cannot be expressed numerically and any statistical analysis cannot be directly applied on these qualitative phenomena. Nevertheless, statistical techniques may be applied indirectly by first reducing the qualitative expressions to accurate quantitative terms. For example, the intelligence of a group of students can be studied based on their marks in a particular examination.

2. **Statistics does not study individuals.**

   Statistics does not give any specific importance to the individual items; in fact, it deals with an aggregate of objects. Individual items, when they are taken individually do not constitute any statistical data and do not serve any purpose for any statistical enquiry.

3. **Statistical laws are not exact.**

   It is well known that mathematical and physical sciences are exact. However, statistical laws are not exact and statistical laws are only approximations. Statistical conclusions are not universally true. They are true only on an average.

4. **Statistics table may be misused.**

   Only experts must use statistics; otherwise, statistical methods are the most dangerous tools on the hands of the inexpert. The use of statistical tools by the inexperienced and untraced persons might lead to wrong conclusions. Statistics can be easily misused by quoting wrong figures of data. As King says aptly "statistics are like clay of which one can make a God or Devil as one pleases".

5. **Statistics is only, one of the methods of studying a problem.**

   Statistical method does not provide complete solution of the problems because problems are to be studied taking the background of the countries culture, philosophy or religion into consideration. Thus, other evidence should supplement the statistical study.

Some Basic Ideas

1. What is statistics?

The term is somewhat loosely employed to cover two separate concepts:

a) Descriptive Statistics and

b) Analytical or inductive statistics.

Descriptive statistics covers the collection and summaries of numerical data.

However, it has some defects; Descriptive statistics are very often not exact or accurate in the arithmetic sense. Some data may be, for instance, Police Personnel, the number of Police Personnel and their salaries can be given precisely, but on the other hand, no one can state exactly how many people there are in Juba city. This is true even for the census day itself because errors due to omission and miscounting are inevitable.

Analytical statistics techniques that help decision makes to raise at national decision under uncertainty or is concerned with the process of drawing the National Strategic Plan, such as Annual National Budget, National Policies.

Comprehensive Agriculture Master Plan, (CAMP) South Sudan National Youth Policy, conclusions about specific characteristics of population based on sample information.[5]

More precisely, statistics is the scientific method for collecting, organizing, summarizing, presenting, and analysing data as well as drawing valid conclusions because of such analysis.

[5] Comprehensive Agriculture Master Plan, 2018-2025. (CAMP) South Sudan National Youth Policy, conclusions about specific characteristics of population based on sample information. P, 123

CHAPTER TWO

PRESENTATION OF DATA

Data can be presented in two forms:

Tabular Presentation.

1. Tabular Presentation or Tabulation may be defined as the systematic presentation of numerical data in rows or columns according to certain characteristics.
2. It expresses the data in concise and attractive form, which can be easily understood and used to compare numerical figures.
3. The advantages of a tabular presentation over the textual presentation are:
   a) It is concise.
   b) There is no repetition of explanatory matter.
   c) Comparisons can be made easily.
   d) The important features can be highlighted; and
   e) Errors in the data can be detected.

a) Table.

- An ideal statistical table should contain the following items:
  a) Table number: A number must be allotted to the table for identification, particularly when there are many tables in a study.
  b) Title: The title should explain what is contained in the table. It should be clear, brief, and set in bold type on top of the table. It should also indicate the time and place to which the data refer.
  c) Date: The date of preparation of the table should be given.
  d) Stubs, or Row designations: Each row of the table should be given a brief heading. Such designations of rows are called “stubs” or “stub items” and the entire column of stubs is called “stub column”.
  e) Column headings or Captions: Column designation is given on top of each column to explain to what the figures in the column refer. It should be clear and precise. This is called a **“Caption”, or, “Heading”.** Columns should be numbered if there are four or more columns.

Frequency Distributions:

Frequency distribution is a tabular arrangement of data by classes together with the corresponding class frequency, for example table.

Constructing a Frequency Table

One of the simplest ways to summarizing the data is by tabulation. **John Grauntin 1662** published his observation on bills of mortality, excerpts of which can be found in Newman 1956.[6]

Given a set of raw statistical data, there is no single grouped frequency distribution that is uniquely "correct" in summarizing its data values. There are many kinds of classes that can be set up to describe this data. Along as rules and practices for constructing frequency distributions discussed in the section below are adhered to, the resulting frequency distribution is normally acceptable. However, the following procedure gives a step-by step approach. There is a rule for doing this:

Step 1.

Find the range of the data (largest minus the smallest number). Calculate the range of the values covered by the data, ignoring (if necessary and only temporally) any extreme values at either end of the data set. The range is the numerical ranging from the lowest and the highest value of the field survey data.

Remark:

Identification of extreme values is a matter of judgment and or experience.

Step 2.

Divide this range into convenient number of classes, which is usually between 5 and 20 and calculate the class size. Divide the range obtains into 10 and adjust this value either upwards or downwards to obtain a standard class width for the distribution, which is appropriate for the data concerned. Class widths of five (or multiple, such as 10, 20, 30, or 50 and so on) are suitable and best since there are easily dealt with.

Step 3.

Determine the number of observations falling into each class. The class can now be constructed as inclusive or **open-ended** depending on the nature of the variable data. The first class should contain the lowest and the highest values (ignoring the extremes

[6] John Wiley, 2004. Biostatistics a Methodology for the Health Science, Second Edition.

values). When the extreme values are present, they can be taken account of by making the first and or last class open-ended. See the following table.

Table 2.1

The class 100-120 is called interval. The smaller number 100 is the lower-class limit, and the larger number 120 is the upper-class limit, class intervals can be opened or closed, e.g.

Classes	Frequency
Opened → 100 - 120	30
Closed → 121-141	25
Closed → 142-162	20
Closed → 163-183	15
Opened → 184-over	10
Total	**100**

The following table below shows the distribution of 100 Second year Students in Faculty of Economics and Social Studies, Upper Nile University, Malakal.

Table 2.2

Simple Frequency Distribution.

Daily number of car accident in Juba	Frequency (or No: of days.)
3	5
4	9
5	11
6	4
7	1
Total	**30**

Table 2.3

Groups or frequency distributions.

Age in a year	Frequency or No: of Persons.
15 – 19	37
20 – 24	81
25 – 29	43
30 – 34	24
35 – 44	9
45 – 59	6
Total	**200**

1. **Class interval (or class).**

   When many observations varying in wide range are available, these are usually classified in several groups according to the size of values. Each of these groups, defined by an interval, is called class interval, or simply class. Column (1), the class interval of ages (in year) is (15-19), (20-24), etc. There are six classes in the frequency distribution the last class is being 45-59.

When one end of a class is not specified, the class is called an open – end class. A frequency distribution may have either one or two open-end class. The necessity of open –end classes arises when there are relatively few observations, which are far apart from the rest. In such a case, it is not considered worthwhile to show several classes with zero frequencies (called empty class) before reaching a class with a very small frequency.

Table 2.4

Distribution of 100-second year student in the Faculty of Economics and Social Study, Department of Statistics and Demography, at Upper Nile University, by income in thousands of South Sudanese Pounds (SSP).'.

Class (Annual Income)	Frequency (No. of Students)
100-120	30
121-141	25
142-162	20
163-183	15
184–204	10
Total	**100**

Note: We can use the actual number when working with distribution, which is not very large e.g., not more than 35 or 40 but when the data is very large as in the College Board examinations, it is empirical to use intervals.

2. **Class frequency, Total frequency.**

   The number of observations falling within a class is called its class frequency, or simply frequency. The sum of all the class frequencies is called Total frequency, the class frequencies are 37, 81 etc. and the total frequency is 200. Total frequency shows the total number of observations considered in the frequency distribution.[7] See the below table number (2) are given in any class frequencies.

[7] N. G Das. 2009, Ibid. p, 71.

Table 2.5

Class limit, Class Boundary etc.

(Illustrated Data table).

Class Interval (1)	**Class Frequency (2)**	**Class Limits**		**Class Boundaries**		**Class Mark (7)**	**Width of Class (8)**	**Frequency Density (9)**	**Relative Frequency (10)**
		Lower (3)	**Upper (4)**	**Lower (5)**	**Upper (6)**				
15 – 19	37	15	19	14.5	19.5	17	5	7.4	0.185
20 – 24	81	20	24	19.5	24.5	22	5	16.2	.405
25 – 29	43	25	29	24.5	29.5	27	5	8.6	.215
30 – 34	24	30	34	29.5	34.5	32	5	4.8	.120
35 – 39	9	35	44	34.5	44.5	39.5	10	0.9	.045
40– 44	6	45	59	44.5	59.5	52	15	0.4	.030
Total	**200**	**-**		**-**		**-**	**-**	**-**	**1.000**

3. Class limitation

All recorded data or observation are discrete in character. For a discrete variable, the values themselves are isolate e g. records regarding the number of workers employed in Faculty of Economics and Social Studies in Upper Nile University, will show data such as 226, 105, 873, 66, 484, etc. (**No factional numbers are possible**). Although a continuous variable can theoretically take any value, all observation is rounded to a certain unit for convenience. See the above table number (3) and (4). There will be no change, but we consider them as lower-class limit and upper-class limit.

4. Class Mark

The values exactly at the middle of a class interval is called class mark or mid-value. It is the midpoint of the class obtained by adding the lower Limit to the upper limit and dividing by two. See the above table number (7) is =

$$15 + 19 = \frac{34}{2} = 17$$

$$20 + 24 = \frac{44}{2} = 22$$

$$25 + 29 = \frac{54}{2} = 27$$

Example:

From table 2.4

$$\frac{120+121}{2} = 120.5$$

Which is the upper-class boundary of the first- and lower-class boundary of second class.

$$\frac{141+142}{2} = 141.5$$

$$\frac{162+163}{2} = 162.5$$

$$\frac{183+184}{2} = 183.5$$

5. **Class boundaries:**

When measurements are taken on continuous variables, all data are recorded nearest to certain unit. Thus, if ages are recorded to nearest whole number of years, any age between 14.5 years and 15.5 years. Similarly, '19 years 'denotes an age between 18.5 years and 19.5 years. Hence, the class interval 15-19 includes all age Class boundary is the real class limit. If the incomes 100-120 thousand are recorded to, the nearest South Sudanese pounds then the class from 100-120 can include counting from 99.5.-120 -5 if we take the nearest point 99.5-120.5 becomes 100 -120 assuming that zero is an even number. Hence, the number 99.5 -120.5 are called class boundaries where 99.5 is the lower-class boundary and 120.5 is the upper-class boundary. Class boundaries are obtained by adding the upper Limit of one class to the lower Limit of the next higher class then dividing by two.

See the table 2.5 above column (5) and (6) or table 2.6 below, class boundaries maybe calculated from the class limit by applying the following rules.

Lower class boundary = lower class limit – 1/2d or 0.5

Upper class boundary = upper class limit + 1/2d or 0.5

Where d is the common difference between the upper-class limit of any class interval and the other lower-class limit of next class interval.

Table 2.6

Hence.

Class boundaries	**Frequency**
99.5 -120.5	30
120.5 -141.5	25
141.5 -162.5	20
162.5 -183.5	15
183.5 -204.5	10
Total	**100**

Note:

Class boundaries should not coincide with an observation.

6. **Width or Class Size**

   Width of size of a class is difference between the lower- and upper-class boundaries (not class limits). Width of class = upper class boundary – lower class boundary. It is the difference between the upper and the lower-class boundaries.

Example:

141.5 - 120.5 = 21 and is called the class size.

Alternatively, if the class size is equal then it can be obtained by subtracting any two successive lower or upper limits.

You can see the **Table 2.5 above (8)**.

7. **Frequency density.**

   Frequency density of a class is its frequency per unit width. It shows the concentration of frequency in class and is given by formula.

$$\text{frequency density} = \frac{class\ frequeny}{width\ of\ the\ classes}$$

See the **Table 2.5 (9)** above. The frequency density is used in drawing histogram when the classes are of unequal width.

Example:

The following set of data are the result of the field survey of the economics survey in Aweil West in August 2023, for the refugee and returnees during Sudan crisis.

Please construct a frequency table

10	24	23	21	18	17	17	13	14	17
19	25	20	19	21	15	18	16	25	18
21	20	18	19	13	11	22	14	19	27
20	22	15	20	24	12	26	24	19	23

Solution:

Step 1: Renge

For the range of the data in ascending order or descending order, from the lowest to the highest or from the highest to the lowest.

10 11 12 13 13 14 14 15 15 16 17 17 17 18 18 18 18 19 19 19

19 19 20 20 20 20 21 21 21 22 22 23 23 24 24 24 25 25 26 27

Step 2: Subtraction of the range, the lowest from the highest range is 27-10 = **17 ranges.**

Step 3: According to **Sturge's law** the number of grouping (i-e. classes) forming a frequency table is determined by the formula.

K = 1+3.222 (log N)

Whereby

K = number of groupings to the nearest number.

N = the total data in question.

From the data above N = 40.

∴ K=1 + 3.222 (log 40)

K = 1 + 3.222 (1.60206).

K = 1 + 5.1618373 = **6.162. is called classes.**

The number of classes in the drawing table will be six classes or rows.

Step 4: Class Size

To have the class size is the number of the range as per the result of your field data.

Since class size = $\frac{Range}{Class\backslash No.of grouping}$

$$\frac{17}{6.162} = 2.75 = 3$$

Therefore $\frac{17}{6}$ = 2.83 = 3

Hence class size = 3

Step 5. Frequency distribution

The frequency distribution table can now be formed using a tally chart. The following example demonstrates the use of this procedure. Determining the number of observations falling into each class.

Table 2.7 Distribution of 40 School pupils by age.

Class interval	Tally	Frequency
10 -12	///	3
13 -15	𝍸 /	6
16 -18	𝍸 ///	8
19 -21	𝍸 𝍸 //	12
22 -24	𝍸 //	7
25 -27	////	4
Total	**40**	**40**

Remarks:

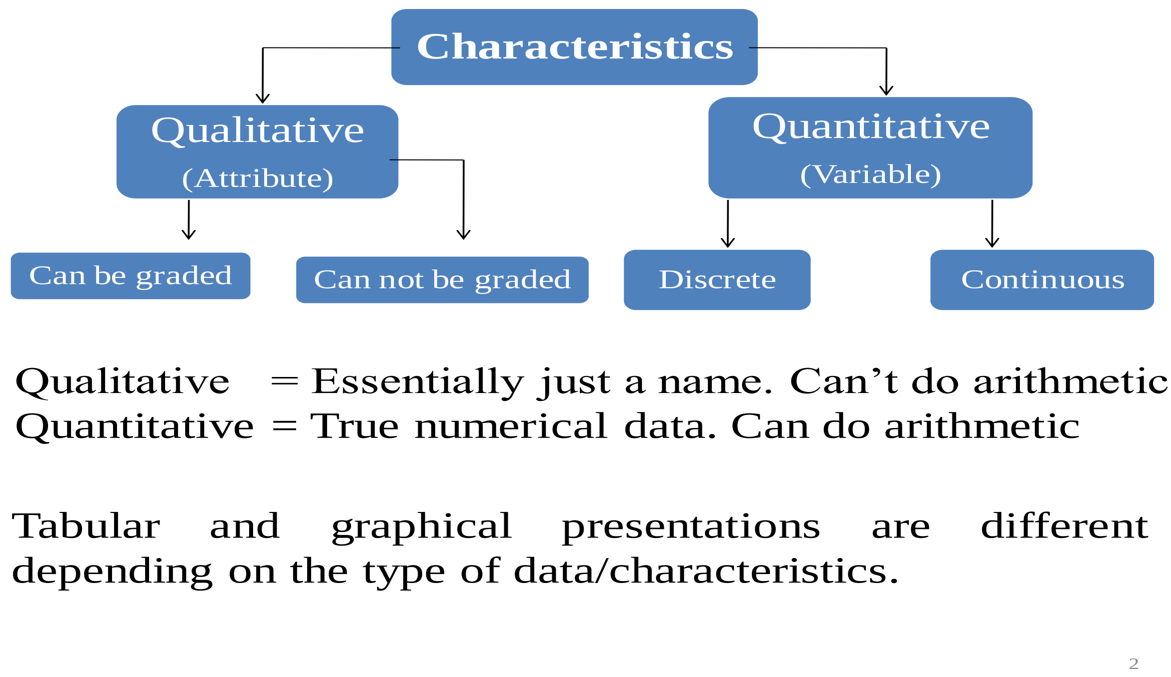
Table 2.8

Hence Frequency Distribution of 40 school pupils by age

Class Interval	Frequency
10-12	3
13-15	6
16-18	8
19-21	12
22-24	7
25-27	4
Total	**40**

(b) Graphs

A graph is a visual form of presentation of statistical data. A graph is more attractive than a table of figure or frequency. Even a common person can understand the message of data from the graph. **Charts and diagrams** are effective devices for vivid presentation of statistical data. The main objective of diagrammatic representation is to emphasise the relative position of different subdivision and not simply to record details. Comparisons can be made between two or more phenomena very easily with the help of a graph.



Characteristics

Qualitative
(Attribute)

Quantitative
(Variable)

Can be graded

Can not be graded

Discrete

Continuous

Qualitative = Essentially just a name. Can't do arithmetic
Quantitative = True numerical data. Can do arithmetic

Tabular and graphical presentations are different depending on the type of data/characteristics.

2

Graphical Presentation

1. Qualitative data may be presented graphically by using bar charts, pie diagrams, line diagrams, etc.
2. Column diagrams, frequency polygons, histograms, cumulative frequency diagrams, etc may present quantitative data graphically.
3. Graphics, such as maps, graphs, and diagrams, are used to represent large volume of data.
4. They are necessary:
5. The graphic method of the representation of data **enhances our understanding**.
6. It presents characteristics in a simplified way and makes the comparisons easy.
7. If the information is presented in tabular form or in a descriptive record, it becomes difficult to comprehend it.
8. Graphical form makes it possible to **easily draw visual impressions** of data and creates an imprint on mind for a longer time.

In many cases, a graphic presentation of frequency table gives more concise and clear information about a frequency distribution. There are three major types of graphic presentation: the histogram, the frequency polygon, and the frequency curve.

Type of charts and Diagrams.

1. line diagram, or Graph
2. Bar diagram
3. Pie diagram
4. Pictogram
5. Histogram, Frequency Polygon, Frequency Curve, Ogive, Relative frequency distribution and Cumulative frequency distribution.

Advantages of diagrams.

Diagrams are appealing to the eyes as well as to intellect, and are therefore, helpful in assimilating the data readily and quickly. Moreover, a chart can clarify a complex problem and reveal hidden facts, which are not apparent from the tabular form.[8] It is sometimes necessary in finding the trend in time series and in finding relation between several sets of observations. In some case, the graph may be used as a mean of checking mistakes.

Disadvantages of diagrams.

Charts do not show details, which is possible in the tabular presentation. Graphical presentation can reveal only the approximate position. In addition, graph and chart require much time to construct, whereas the desired information can be quick conveyed arranging the data in form of a table.

Histograms

A histogram is a bar chart or graph showing the frequency of occurrence of each value of the variable being analysed. In histogram, data are plotted as a series of rectangles. Class intervals are shown on the '**X-axis'** and the frequencies on the '**Y-axis'**. The height of each rectangle represents the frequency of the class interval. Each rectangle is formed with the other to give a continuous picture. Such a graph is also called staircase or block diagram. However, we cannot construct a histogram for distribution with open-end classes. It is also quite misleading if the distribution has unequal intervals and suitable adjustments in frequencies are not made. It

[8] N. G Das. 2009, Ibid. p, 27.

consists of a set of rectangles having base on the x-axis with Centre at the class mark and length equal to the class. The area of the rectangle is proportional to the class frequencies.

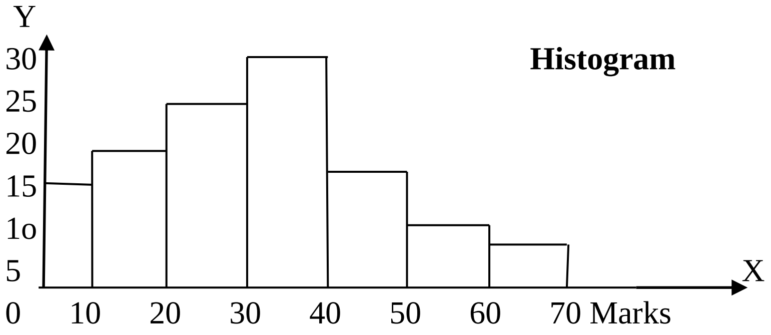
Example: (1)

With the help of following data, construct a histogram.

Marks obtained	0 – 10	10 – 20	20 – 30	30 - 40	40 – 50	50 – 60	60 – 70
Number of the Students:	16	20	25	30	18	10	8

Solution:

On the X-axis, we represent the class interval (**they are equal**) starting from zero (0) and ending to (70). Frequencies are to be on Y-axis – the ranges of frequencies are from (8) to 30. We can take six points on Y-axis representing equal distance of five (5) frequencies and draw the graph that is the rectangles.



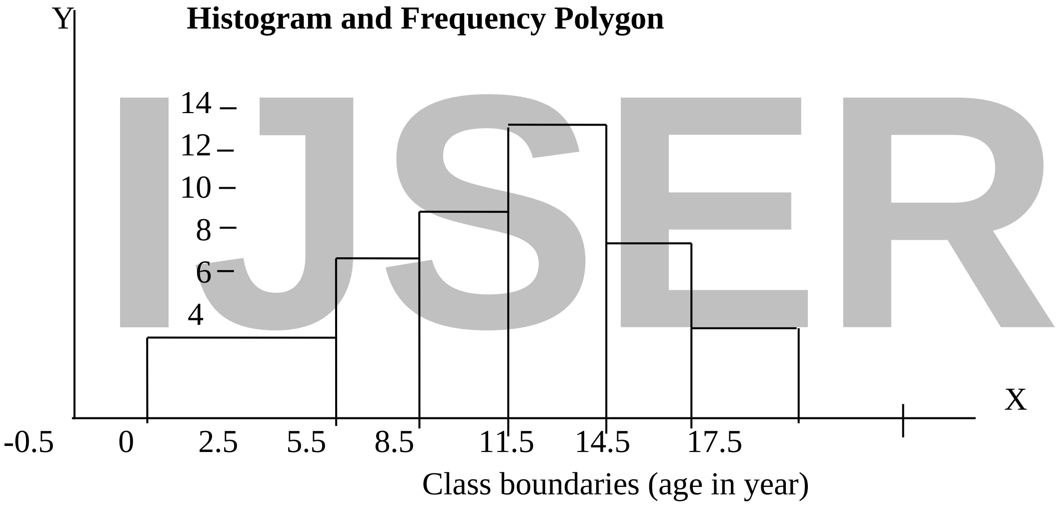
Histogram

Example: 2

Draw a histogram and frequency polygon of the distribution of 40 school pupils by age.

Table 2.9

Class Boundaries	Frequency	Class Mark
-0.5-2.5	3	1
2.5-5.5	6	4
5.5-8.5	8	7
8.5-11.5	12	10
11.5-14.5	7	13
14.5-17.5	4	16
Total	**40**	



Note: Since we don’t have a negative age, ignore the internal -0.5-0 but consider 0-2.5 years old.

Example: 3

With the help of following data, construct a histogram.

Income by SSP	0 – 3000	3000-4000	4000-5000	5000-6000	6000-7000	7000- 10,000
No: of Person	15	18	20	24	16	10

Solution:

Here the class intervals are not equal. Therefore, we will follow the directions for drawing a histogram for unequal class interval.

Adjusted frequency.

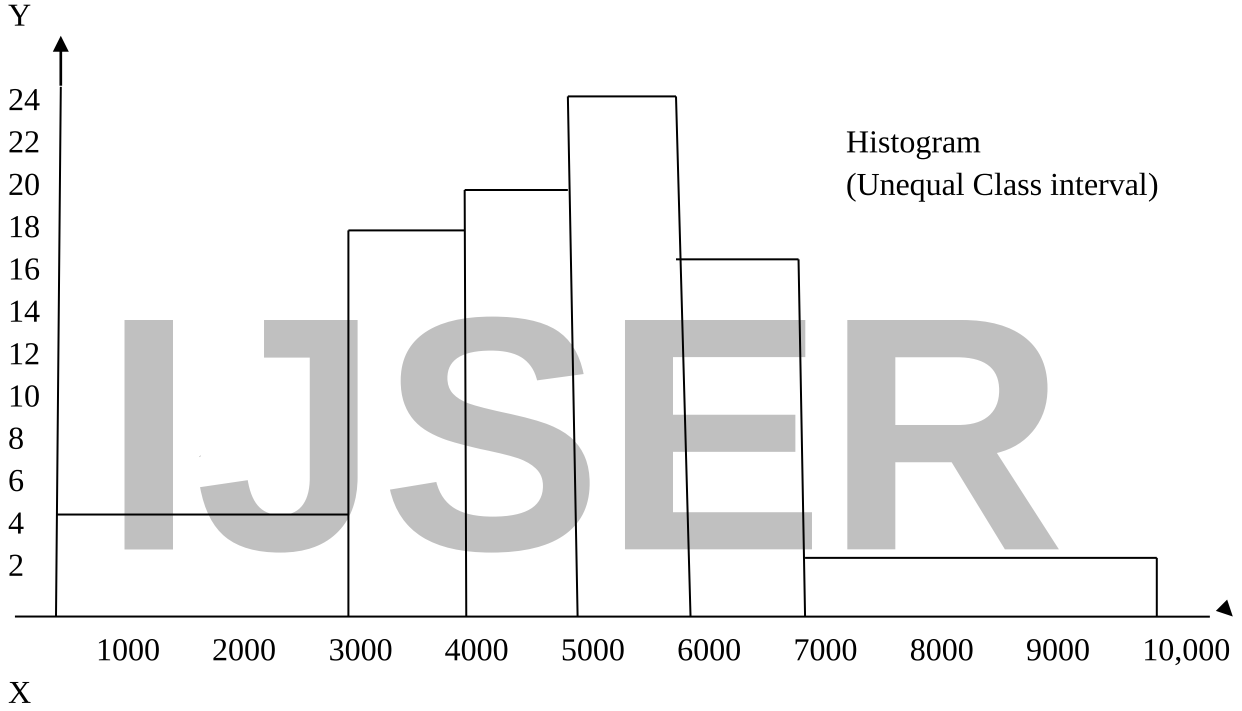
$$= \frac{frequency\ of\ the\ class\ x\ assumed\ uniform\ class\ interval}{Actual\ class\ interval}$$

First group – class interval = 3000

Normal class interval = 1000

Adjusted frequency for first group =$\frac{15\ x\ 1000}{3000} = 5$

Adjusted frequency for last group = $\frac{10\ x\ 1,0000}{3000} = 33.3$



Income in South Sudanese Pound SSP.

Example: 4

If we take, the frequency distribution likes.

Table 2.10

Class interval	Frequency
30 - 39	4
40 - 49	6
50 - 59	8
60 – 69	12
70 – 79	9
80 – 89	7
90 – 99	4

The first thing, we need to do is to enter the scale of the variable X (that is grades) on the Horizontal axis, since the data are discrete; there is a gap between the class interval 20 – 29 and 30 – 39. In such a case, the dividing point between the two intervals will be (29+30)/2 = 29.5 and similarly for the other dividing point by doing this, we avoid gaps between the bars.

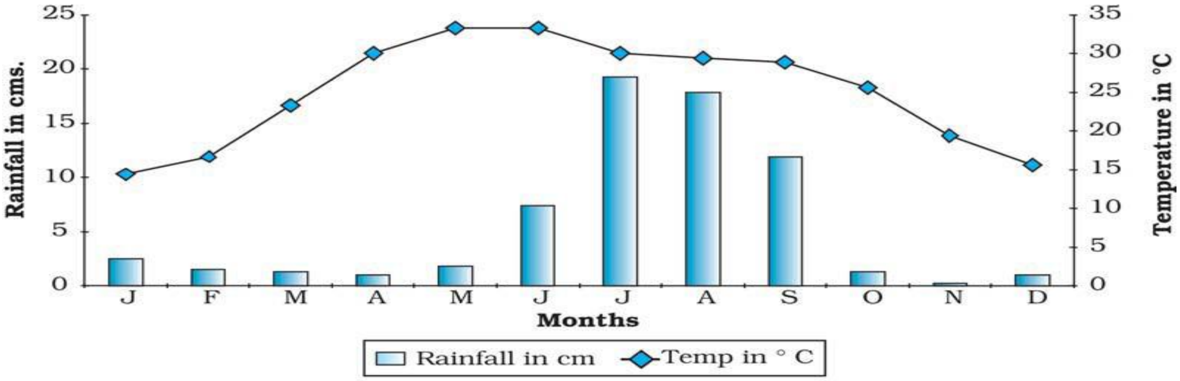
Line Diagram/Graph

- The line graphs are usually drawn to represent the time series data related to the temperature, rainfall, population growth, birth rates and the death rates. etc.
- Construction of a Line Graph
  - **First step**: Round the data to the appropriate level.
  - **Second step**: Draw X and Y-axis. Mark the time series variables (years/months) on the X-axis and the data quantity/value to be plotted on Y-axis.
  - **Third step:** Choose an appropriate scale to show data and label it on Y-axis. If the data involves a negative figure, then the selected scale should also show it.
  - **Fourth step:** Plot the data to depict year/month-wise values according to the selected scale on Y-axis, mark the location of the plotted values by a dot and join these dots by a free hand drawn line.

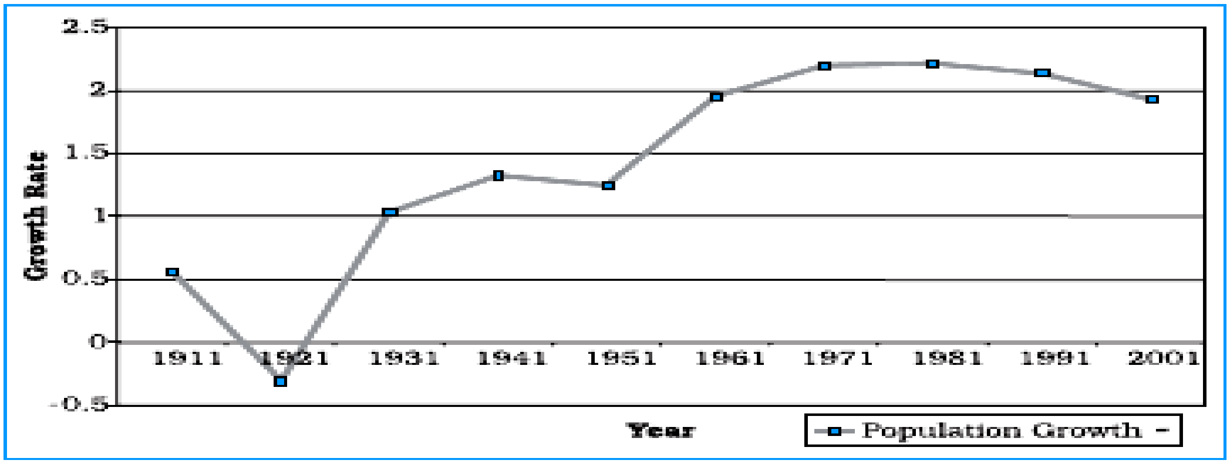
Line and Bar Graphs

The line and bar graphs as drawn separately may also be combined to depict the data related to some of the closely associated characteristics such as the climatic data of mean monthly temperatures and rainfall.

Example 4: Line and Bar Graph to Show Rainfall and Temperature in Different Months pera Year 2019in the Republic of South Sudan.

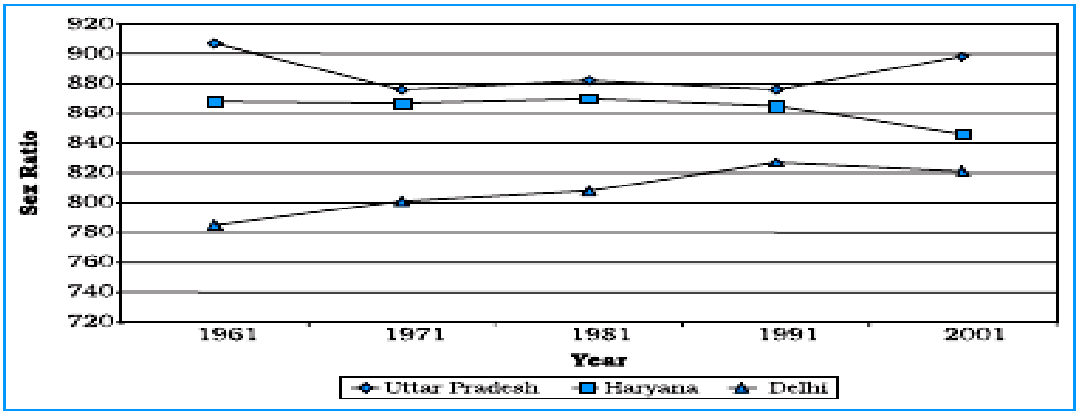


- **Line Diagram Showing Population Growth in East Africa Region**



- Polygraph (or Multiple Line Diagram) is a line-graph in which two or more than two variables are shown on a same diagram by different lines. It helps in comparing the data. Examples which can be shown as polygraph are:
  - The growth rate of different crops like rice in South Sudan, wheat, pulses in one diagram.
  - The birth rates and death rates in one diagram.
  - Sex ratio in different states of South Sudan take example of Upper Nile State, Central Equatoria State or countries in one diagram.
- Construction of a Polygraph
  - All steps of construction of polygraph are like that of line graph. However, different lines are drawn to indicate different variables.

A Polygraph Showing Sex Ratio in Uttar Predesh, Haryatta and New Delhi.



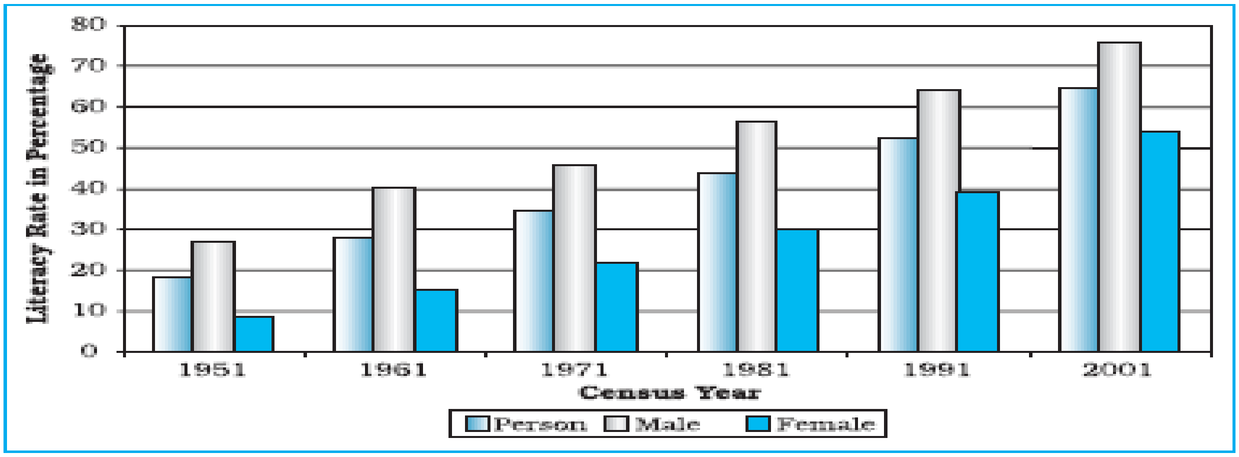
Multiple bar Diagram.

- Multiple bar diagrams are constructed to represent two or more than two variables for the purpose of comparison. For example, a multiple bar diagram may be constructed to show proportion of males and females in the total, rural and urban population in South Sudan or the share of canal, tube well and well irrigation in the total irrigated area in different states.
- Construction

(a) Mark time on X-axis and variable data on Y-axis as per the selected scale.

(b) Plot the data in closed columns.

Example 5: A Multiple bar Diagram Showing Literacy Rates by Gender in South Sudan.



Source: NBS 2010 during Sudan Household Survey.

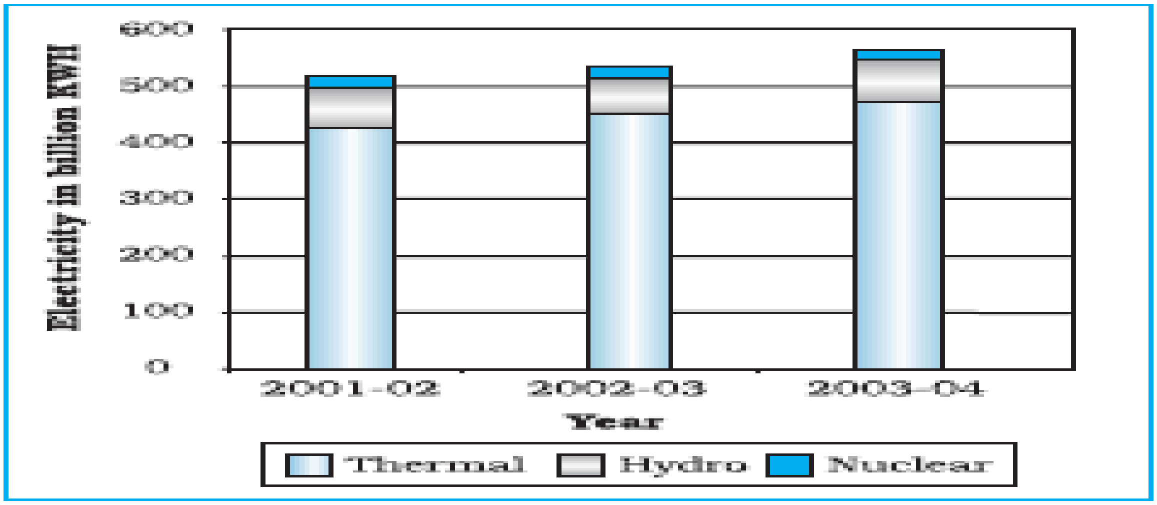
Compound or Divided Bar Diagram.

1. When different components are grouped in one set of variable or different variables of one component are put together, their representation is made by a compound (or Divided) bar diagram. In this method, different variables are shown in a single bar with different rectangles.
2. Construction

(a) Arrange the data in ascending or descending order.

(b) A single bar will depict the set of variables by dividing the total length of the bar as per percentage.

Example 6: Compound bar Diagram Showing Electricity Generation from Thermal, Hydro and Nuclear Projects.



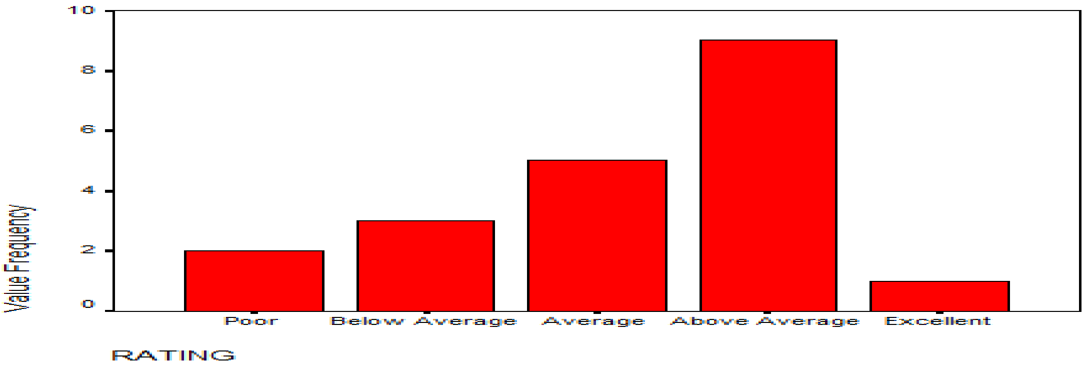
Source: NBS 2010 during Sudan Household Survey.

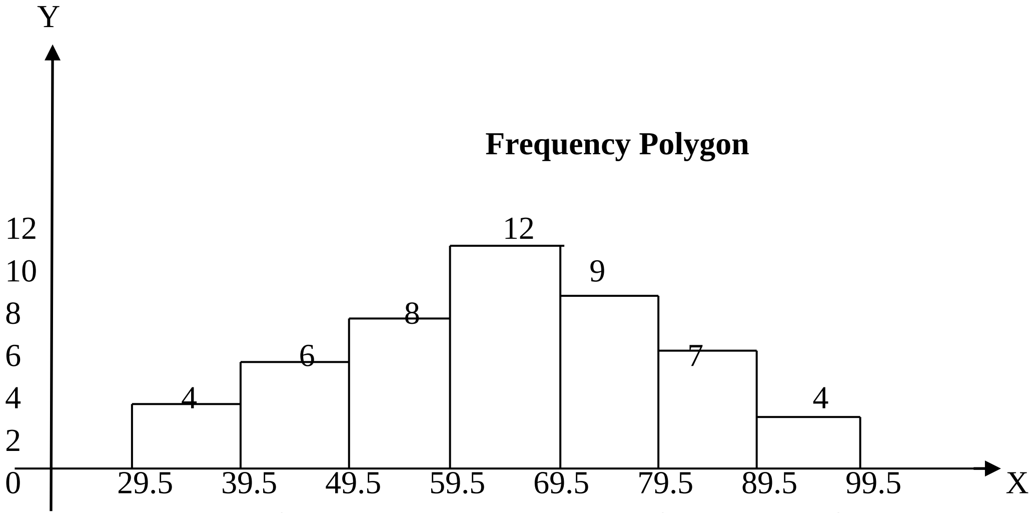
Table 2.11

Frequency distribution showing frequencies, relative frequencies, percentages etc.

Rating	Frequency	Relative Frequency	Percentage	Cumulative frequency
Poor	2	0.10	10.0	2
Below Average	3	0.15	15.0	5
Average	5	0.25	25.0	10
Above Average	9	0.45	45.0	19
Excellent	1	0.05	05.0	20
Total	20	1.00	100.0	---

Example: Bar Graph of Rating.





An alternative way of drawing the histogram is as follow: First, find the mid-points of the class intervals. In our present case they are $\frac{30 + 39}{2} = 34.5, 44.5, 54.5 \dots \dots .94.5$.

(ii)Frequency Polygon

Frequency polygon is a line graph of class marks plotted against class frequency. When the mid points of the class intervals are linked, we obtain a frequency polygon for example, the mid-point for the 40-49 class intervals is $\frac{40+49}{2} = 44.5$. this is consistent with the way in which the histogram has been draw. We note that the area under the frequency polygon is equal to the area under histogram. If we join the mid points in the above figure, we get a frequency polygon.[9]

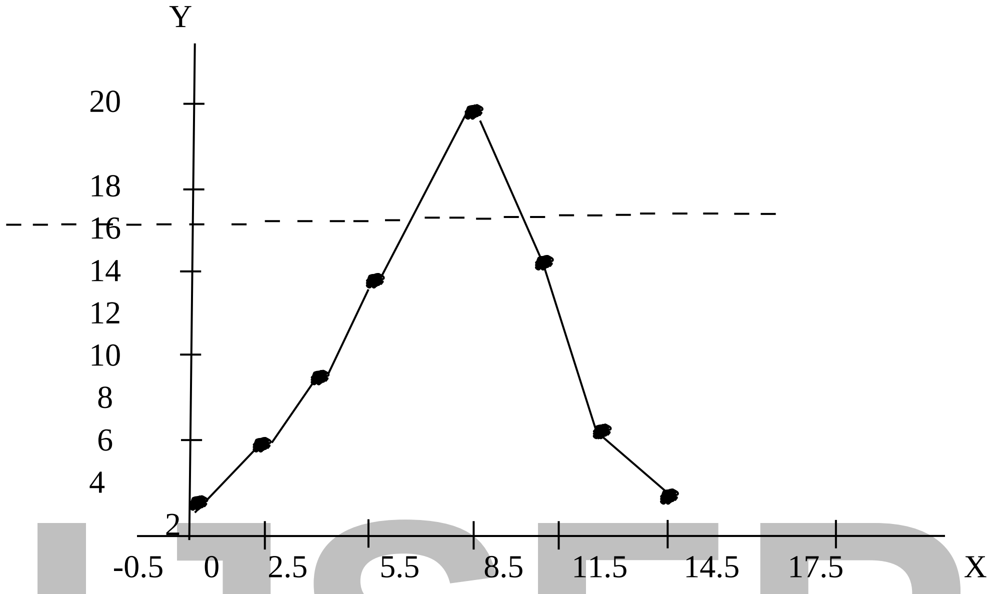
Illustration:

1- Draw a horizontal line and space out the numbers evenly along it, only enough spaces are needed.

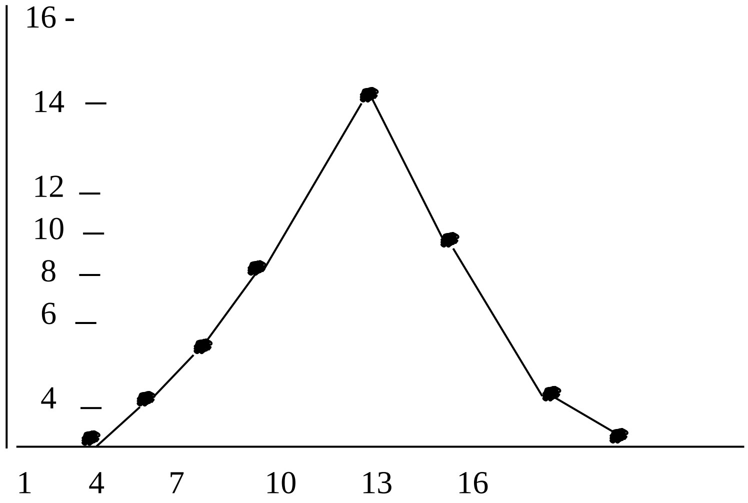
2- The boxes or the rectangles do not have to be fancy, but they should be reasonably equal in size.

[9] Satyabrata Pal: 2010, Statistics (I) for BBA Students p.52. as per West Bangal University. India, New Age International Publishers.

Frequency Polygon can be drawn using the class mark e.g.



Similarly, frequency Polygon can be drawn using the class mark or class density e.g.



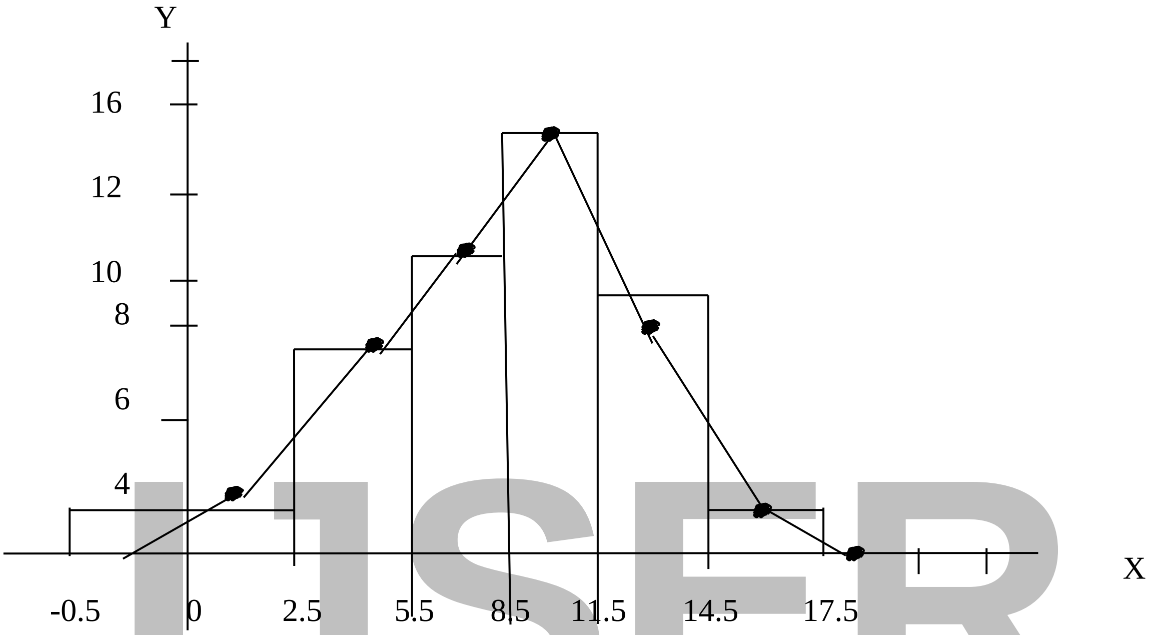
Converting Histogram into Frequency Polygon

1- Connect the midpoint of the top boxes in each column.

2- Also, connect the first and the last column to the baseline.

Example:

Histogram and frequency polygon of the distribution of 40 school pupils by age.[10]



Note: if you plotted both histogram and frequency polygon on the same graph, you should either used the class-mark or class boundaries but not both.

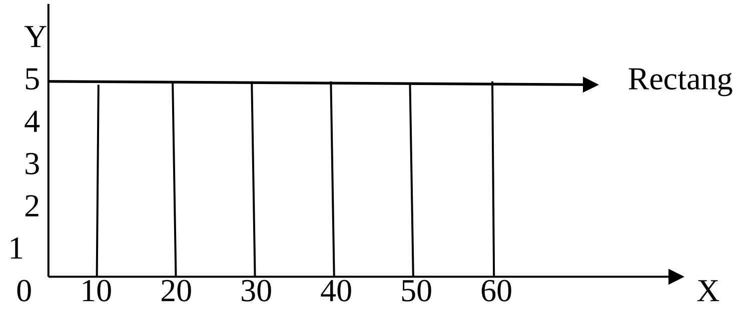
(iii) Frequency curve.

The frequency curve is a graphic presentation of a theoretical frequency distribution. Nevertheless, since it is difficult to give a full discussion of a theoretical frequency distribution at this point, we shall say a first approximation that a frequency curve is smoothed frequency polygon. Certain forms of frequency curve have been given specific names to correspond with specific types of frequency distribution. If a grades (marks) of the students of Faculty of Economic, first year are the same in each class, the shape of the distribution is rectangular and that is why the distribution is known as rectangular.

Table: 2.12

Class interval	Frequency
0 – 10	5
10 – 20	5
20 – 30	5
30 – 40	5
40 – 50	5
Total	**25**

[10] Masembe Kabali 2011. Basic Business Statistics, third edition, P.O. Box 7453 Kapala – Uganda. Type set, printed and Bound by M. Kaabali and Basic Business Statistics Books.



Normal distribution.

A normal distribution is perfectly symmetrical distribution about mean, with a frequency curve that is bell shaped.

Table 2.13

Class interval	Frequency
0 – 10	5
10 – 20	7
20 – 30	10
30 – 40	20
40 – 50	10
50 – 60	7
60 - 70	5

(iv) Ogives.

An ogive is a graph showing the commutative frequency less than any upper-class boundary plotted against the upper-class boundaries.

(v) Relative frequency Distribution

- The relative frequency of a class is the fraction or proportion of the total number of data items belonging to the class.
- It is the frequency of the class divided by the total frequency expressed as a percentage.

Relative frequency of a class = Frequency of the class/*n*.

Frequency distribution showing frequencies and relative frequencies.

Table 2.14

Rating	**Frequency**	**Relative Frequency**
Poor	**2**	**0.10**
Below Average	**3**	**0.15**

Average	**5**	**0.25**
Above Average	**9**	**0.45**
Excellent	**1**	**0.05**
Total	**20**	**1.00**

Frequency distribution showing frequencies, relative frequencies, percentages and cumulative frequency, etc.

Rating	Frequency	Relative Frequency	Percentage	Cumulative frequency
Poor	**2**	**0.10**	**10.0**	**2**
Below Average	**3**	**0.15**	**15.0**	**5**
Average	**5**	**0.25**	**25.0**	**10**
Above Average	**9**	**0.45**	**45.0**	**19**
Excellent	**1**	**0.05**	**05.0**	**20**
Total	**20**	**1.00**	**100.0**	**---**

Table 2.15

The calculation of the relative frequency distribution for the 40 school pupils from Fashoda Primary school by age distribution:

Class interval	Frequency	Relative Freq. Distribution
0-2	3	$\frac{3}{40} \times 100 = 7.5\%$
3-5	6	= 15%
6-8	8	20%
9-11	12	30%
12-14	7	17%
15-17	4	
Total	**40**	**100%**

Remarks:

The relative frequency distribution explains the proportions of the total number of observations that falls into that interval. Out of the 40 school pupils, 30% are pupils of age 9 -11 years old.

(vi). Cumulative Frequency Distribution: -

The total frequency of all values less than the upper-class boundary of a given class is called cumulative frequency up to and including that class. e.g.: 3+6+8=17 is the cumulative frequency up to and including the class (6-8).

The Cumulative Frequency Distribution of the 40 school pupils by age is shown below:

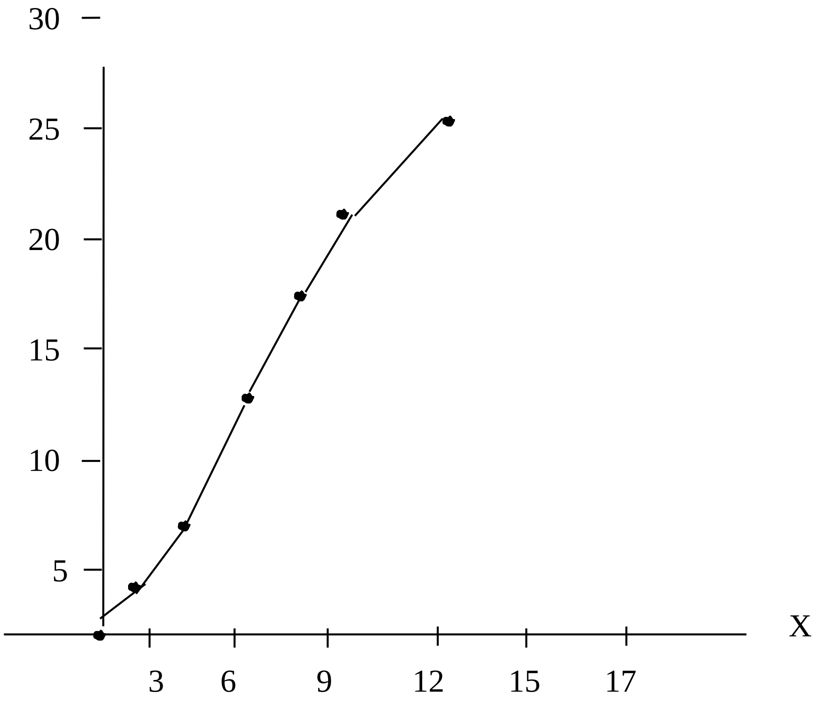
Table 2.16

Age (in years)	Cumulative Frequency
Less than 0	0
“ “ 3	3
“ “ 6	9
“ “ 9	17
“ “ 12	29
“ “ 15	36
“ “ 17	40

Remarks:

This is called less than cumulative distribution. Once data is graph, it is called an Ogive of less than cumulative frequency for the distribution of 40 school pupils by age.

An ogive of less than cumulative frequency for the distribution of 40 school pupils by age.

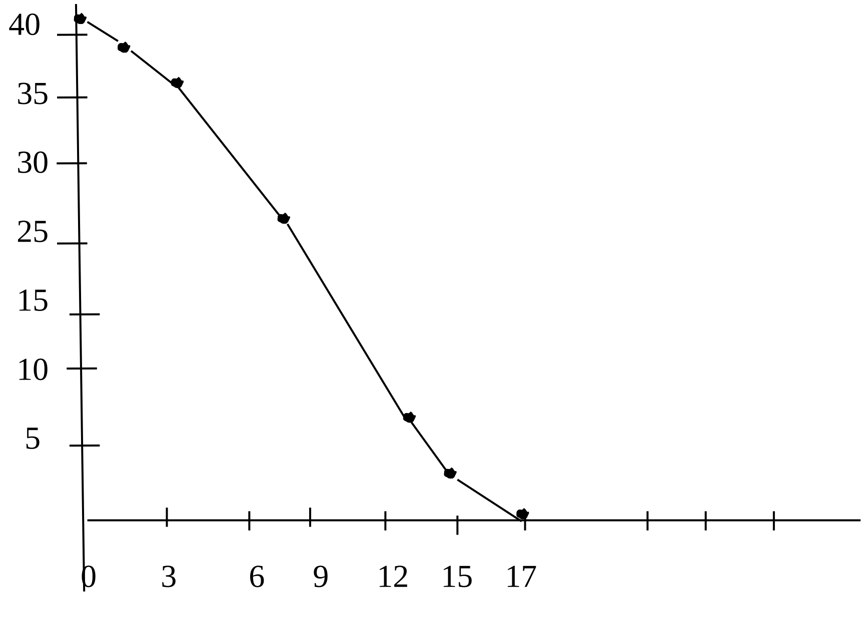


If we consider all values greater than or equal to the lower-class boundary of each class, we obtain "or more" cumulative frequency. The "or more" cumulative frequency for the distribution of 40 school pupils by age as shown below:

Table 2.17

Age (in years)	Cumulative frequency
0 or more	40
3	37
6	31
9	23
12	11
15	4
17	0

Note: once graph it is called an Ogive of frequency or more cumulative frequency, i.e.



Pie Diagram.

- Pie diagram is another graphical method of the representation of data. It is drawn to depict the total value of the given attribute using a circle. Dividing the circle into corresponding degrees of angle then represent the sub– sets of the data. Hence, it is also called as Divided.
- The angle of each variable is calculated using the following way.

Calculation of Angles.

(a) Calculate the degrees of angles for showing the given values assuming the total as 360.

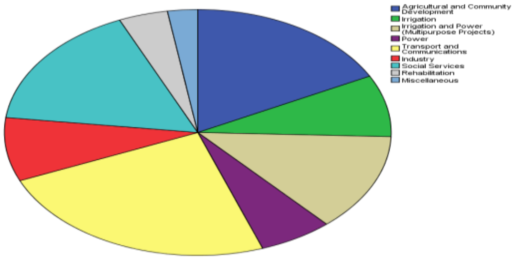
(b) It could be done by multiplying percentage with a constant of 3.6 as derived by dividing the total number of degrees in a circle by 100, i. e. 360/100.

(c) Plot the data by dividing the circle into the required number of divisions to show the share different regions/countries

Example

The following table gives the total outlay on rural development proposed in the first five-year plan and its breakdown into the major items. Give a suitable graphic presentation of the data.

Item	Amount (cost in SSP)
Agricultural and Community Development	360.43
Irrigation	167.97
Irrigation and Power (Multipurpose Projects)	265.90
Power	127.50
Transport and Communications	497.10
Industry	173.04
Social Services	339.81
Rehabilitation	85.00
Miscellaneous	51.99

Total **SPSS output**	**2068.78** 

Thematic Maps.

- Varieties of maps are drawn to understand the patterns of the regional distributions or the characteristics of variations over space these maps are known as the distribution maps or thematic maps.
- Requirements for Making a Thematic Map
  - **a.** State/District level data about the selected theme.
  - **b.** Outline map of the study area along with administrative boundaries.
  - **c.** Physical map of the region. For example, physiographic map for population distribution and relief and drainage map for constructing transportation map.
- Rules for Making Thematic Maps
  - (i) The drawing of the thematic maps must be carefully planned. The final map should properly reflect the following components:
  - **a.** Name of the area
  - **b.** Title of the subject-matter
  - **c.** Source of the data and year
  - **d.** Indication of symbols, signs, colors, shades, etc.
  - **e.** Scale
  - (ii) The selection of a suitable method to be used for thematic mapping.

Classification of Thematic Maps.

- There are three types of Thematic maps -

  (a) Dot maps

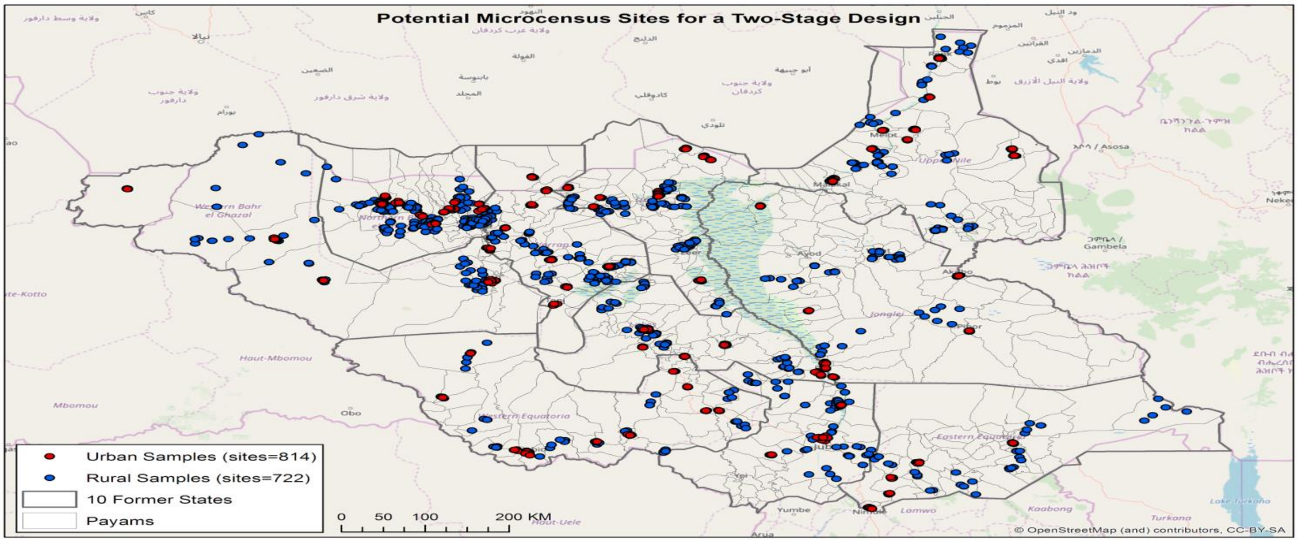
  (b) Choropleth maps

  (c) Isopleth maps

Dot Maps.

- The dot maps are drawn to show the distribution of phenomena such as population, cattle, types of crops, etc. The dots of same size as per the chosen scale are marked over the given administrative units to highlight the patterns of distributions.

- Requirement
  - (a) An administrative map of the given area showing state/district/block boundaries.
  - (b) Statistical data on selected theme for the chosen administrative units, i.e., total population, cattle etc.
  - (c) Selection of a scale to determine the value of a dot.
- Precaution
  - (a) The lines demarcating the boundaries of various administrative units should not be very thick and bold.
  - (b) All dots should be of same size.



Source: South Sudan National Bureau of Statistics 2021. [11]

Choropleth Map.

- The choropleth maps are also drawn to depict the data characteristics as they are related to the administrative units. These maps are used to represent the density of population, literacy/growth rates, sex ratio, etc.
- Requirement for drawing Choropleth Map
  - (a) A map of the area depicting different administrative units.
  - (b) Appropriate statistical data according to administrative units.
- Steps to be followed.
  - (a) Arrange the data in ascending or descending order.

[11] Charles Mona, 2021, South Sudan Population Estimation Survey 2021- Juba- South Sudan National Bureau of Statistics 2021

(b) Group the data into (5) categories to represent very high, high, medium, low and very low concentrations.

(c) The interval between the categories may be identified on the following formulae i.e., Range/5 and Range = maximum value – minimum value.

(d) Patterns, shades, or colour to be used to depict the chosen categories should be marked in an increasing or decreasing order.

Isopleth Map

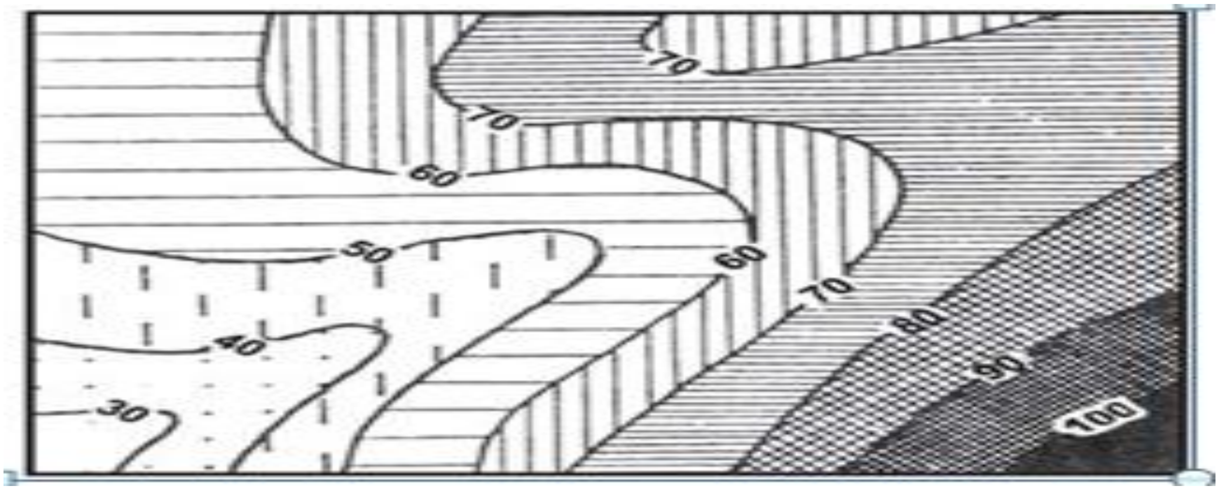
- Variations in the degrees of slope, temperature, occurrence of rainfall, may be represented by drawing the lines of equal values on a map. All such maps are termed as Isopleth Map. The word Isopleth is derived from Iso meaning equal and Plethrons meaning ***“entire number or quantity”*** or “the *whole aggregate”*. Thus, an imaginary line, which joins the places of equal values, is referred as Isopleth. The more frequently drawn isopleths include Isotherm (equal temperature), Isobar (equal pressure), Isohyets (equal rainfall), Isonephs (equal cloudiness), Isohels (equal sunshine), Isobaths (equal depths), Isohaline (equal salinity), etc.
- Requirements

  (a) Base line map depicting point location of different places.

  (b) Appropriate data of temperature, pressure, rainfall, etc. over a definite period.

  (c) Drawing instruments.

An Isopleth Map.



Box Plot.

- In descriptive statistics, a **box plot** or **boxplot** (also known as a **box-and-whisker diagram** or **plot**) is a convenient way of graphically depicting groups of numerical data through their five-number summaries:
  1) the smallest observation (sample minimum),
  2) lower quartile (Q1),
  3) median (Q2),
  4) upper quartile (Q3), and
  5) Largest observation (sample maximum).
- A box plot may also indicate which observations, if any, might be considered outliers.

Properties of Box Plot.

- Boxplot
  1) is non-parametric, and
  2) Does not make any assumptions of the underlying statistical distribution.
- The spacing's between the different parts of the box help to indicate the degree of dispersion (spread) and skewness in the data and identify outliers.
- Boxplots can be drawn either horizontally or vertically.

Figure 1.Boxplot with whiskers (Vertical).

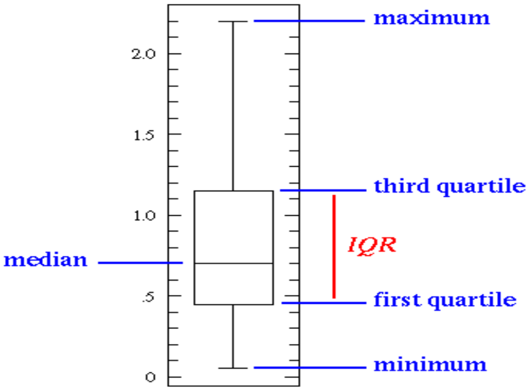
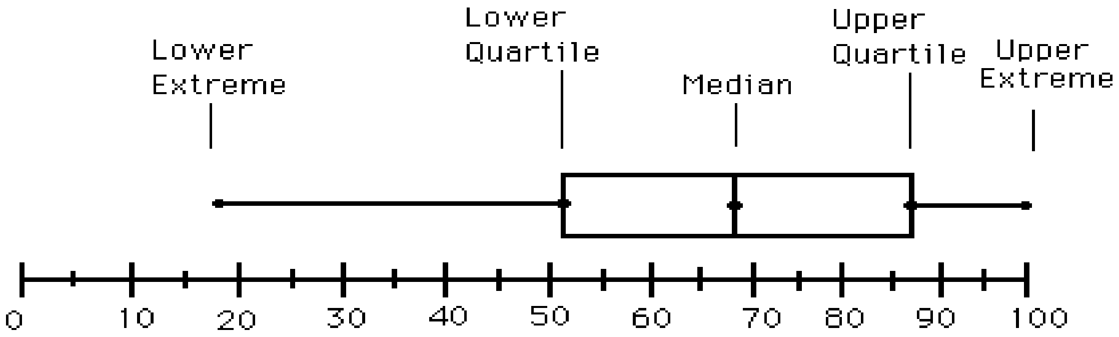


Figure 2. Boxplot with whiskers (Horizontal).



Descriptions.

- Box and whisker plots are uniform in their use of the box: the bottom and top of the box are always the 25th and 75th percentile (the lower and upper quartiles, respectively), and the band near the middle of the box is always the 50th percentile (the median). But the ends of the whiskers can represent several possible alternative values, as:
  - the minimum and maximum of all the data,
  - the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile,
  - one standard deviation above and below the mean of the data,
  - the 9th percentile and the 91st percentile, or
  - The 2nd percentile and the 98th percentile.

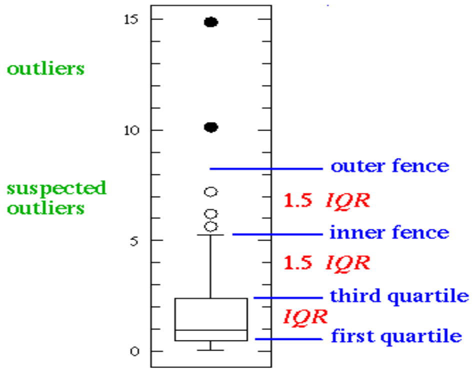
Outliers.

- John Tukey has provided a precise definition for two types of outliers:
  1) Outliers are either 3 × *IQR* or more above the third quartile or 3 × *IQR* or more below the first quartile.

  2) Suspected outliers are (i) either 1.5 × *IQR* or more above the third quartile or (ii) 1.5 × *IQR* or more below the first quartile.

- If either type of outlier is present the whisker on the appropriate side is taken to 1.5×*IQR* from the quartile (the "inner fence") rather than the max or min, and individual outlying data points are displayed as unfilled circles (for suspected outliers) or filled circles (for outliers). (The "outer fence" is 3 × *IQR* from the quartile.)

Outliers.



Outliers.

- If the data happens to be normally distributed, then *IQR* = 1.35 σ, where σ is the population standard deviation.
- Note that outliers are not necessarily "bad" data-points; indeed, they may well be the most important, most information rich, and part of the dataset. Under no circumstances should they be automatically removed from the dataset. Outliers may deserve special consideration: they may be the key to the phenomenon under study.

Summarizing Quantitative Data.

1. Frequency Distribution including Relative Frequency and Percent Frequency Distributions.
2. Dot Plot or Scatter Plot or Scatter Diagram.
3. Histogram.
4. Cumulative distribution diagram or Ogive.
5. Box Plot

   The first four topics will be explained on the white/black board and the last topic is discussed below.

CHPTER THREE.

MEASURES OF CENTRAL TENDENCY

(AVERAGES)

In addition to the graphical techniques encountered so far, it is often useful to obtain quantitative summaries of certain aspects of the data. Most simple summary measurements can be divided into two types.

firstly, quantities which are "typical" of the data, and **secondly,** quantities which summaries the variability of the data.

An average is sometimes called a measure of tendency because individual values of the variable usually cluster around it, Crum, and Smith. The former is known as measures of location and the latter as measures of spread. Suppose we have a sample of size n of quantitative data. We will denote the measurements by x_1 ,x_2,...x_n.

The average occupies an important place in statistics. Many other techniques of statistical analysis depend upon this measure. This is reason for Dr. Bowley defining statistics as the science of average.

Objective of Averages.

1. It determines a single figure of the whole series. It is a tool to represent salient feature of mass of complex data.
2. Averages are useful for comparison. The average of one group can be composed with average of other groups.
3. Averages are helpful for taking overview of statistical data, which ordinarily are not easily understood.
4. Averages are helpful for making decision in planning in various fields.
5. Averages are commonly used for the further treatment of statistical derivatives and series.

These values are typical representative of a set of data. The most common once are:

1- Arithmetic Mean.

2- Median

3- Mode

4- Geometric Mean and Harmonic Mean.

3.1 The Arithmetic Mean ($\bar{X}$)

The arithmetic mean is the most used and readily understood measure of central tendency. In statistics, the term an average refers to any of the measure of central tendency.

The arithmetic mean is defined as being equal to the sum of the numerical values of each observation divided by the total number of the observation. Symbolically it can be represented as:

$$\bar{\boldsymbol{X}} = \frac{\Sigma X}{N}$$

Whereby ΣX indicates the sum of the values of all the observations and N is the total number of observations.

It is the most frequently used average. Arithmetic Mean is the sum of the numbers included in the relevant set of data divided by the number of such numbers i.e the Arithmetic Mean of a set of any numbers.

$X_{1,}X_{2,}X_{3,}$ ....., $X_{N,}$ is defined

$\bar{X} = \frac{X_1+X_2+X_3+\ .....+\ X_n}{N} = \frac{\sum x}{N}$

The arithmetic mean in continuous series can be obtained by any of the following method.

- **I. Direct method.**
- **II. Short cut method.**
- **III. Deviation method.**

1) Direct method: $\bar{\mathrm{X}} = \frac{\sum \mathrm{fx}}{\sum \mathrm{f}}$

2) Short cut method: $\bar{\mathrm{X}} = \mathrm{A} + \frac{\sum \mathrm{fd}}{\sum \mathrm{f}}$

(A is assumed mean and d is the deviation from assumed mean that is = X-A)

3). Step Deviation method: $\bar{X} = A + \frac{\sum fd}{\sum f} \times \mathrm{C}$

Where C is size of class interval.

$\mathrm{d} = \frac{X-A}{C} = \frac{d}{C} \longrightarrow \quad \mathrm{fd} = \frac{1}{C}\,[\mathrm{f}X - \mathrm{Af}]$

$\sum \mathrm{fd} = \frac{1}{C} \sum \mathrm{f}\,(X\text{-}A)$

Example (a)

If you get grades of 170,285, and 1100 on three tests, your mean or average score is 185. You probably arrived at this answer by noticing that the first score was 15 points lower than the middle one, and the third score was 15 points higher. Therefore, the scores balance at 85. Technically

$\bar{X} = \frac{X}{N} = \frac{170 + 285 + 1100}{3} = \frac{565}{3} = 188.33$

If the numbers X$_1$, X$_2$, X3, …………. X$_N$ occurs with frequencies f$_1$, f2,…, f$_N$

Respectively, then $\bar{X}$ = is defined as:

$$\bar{X} = \frac{f_1X_{1+}f_2X_2 + f_3X_3 \ldots.. + f_NX_N}{f_1+f_2+f_3+f_N} = \frac{\begin{matrix} N \\ \mathit{\Sigma} f_1 X_1 \\ i=1 \end{matrix}}{\begin{matrix} \mathrm{N} \\ \mathit{\Sigma}\ f_1 \\ 1+1 \end{matrix}}$$

Where f$_1$, f$_2$, ...f$_n$ are the numbers of measurements in the class intervals and X$_1$, X$_2$,X$_3$, ….. X$_N$ are the midpoints of the first-class interval, 2nd and so on. This method of calculating $\bar{X}$is **called the long method.**

Example (b)

Suppose a man has 100 SSP, he gave out 40 SSP to his son and 60 SSP to his daughter. If the boy and the girl want to share the money equally then how much shall each one of them get. It is obvious that the amount the girl took (60 SSP) was more by 20 SSP than the 40 SSP taken by the boy. If the is to give her brother from her share additional 10 SSP, then both shall have 50 Pounds each. In this case, the "50" South Sudanese Pounds is called the arithmetic mean (average) for the numbers 60 and 40.

Algebraically this can be calculated as follows.

$$\frac{60 + 40}{2} = \frac{100}{2}$$

Hence

Arithmetic Mean= $\frac{Sum\ of\ values}{Number}$

Example:

If the scores in mathematics by student were 70 and 60: what is the mean score?

The mean score is $\frac{Sum\ of\ scores}{No.} = \frac{70+60}{2} = \frac{130}{2} = 65\ marks$

Example (c)

In four consecutive days **Obaj** was paid the following amount of money in South Sudan Pounds 220, 210, 215, 215 what is the average earning in these 4 days.

Solution

Total earning= 220 + 210 + 215 + 215 = 860

No: of days = 4 days

Therefore, the average earning = $\frac{860}{4}$ = 215 SSP

This means that, Obaj did earn on average = 215 SSP in those 4 days. If the distribution are discrete that is the variety is a completely number, but in the form of frequency distributions.

Example:

Table 3.1

X	**F**
10	8
12	10
14	6
16	4

In such a situation $\overline{\boldsymbol{X}} = \frac{\sum \mathrm{fX}}{\sum \mathrm{N}}$

Direct method of computing the Arithmetic Mean.

Example (d)

Kumar Radha has recorded the prices of meat in **Shimla Market** in five consecutive days in Rupees as shown below.

Day	1	2	3	4	5
Price (kg)	300	300	250	250	250

Find the average price of meat in five days.

Solution

Total price of meat = 300 + 300 + 250 + 250 + 250 =**1350**

No. of days = 5 days

Therefore $\frac{1350}{5}$ = **270** Rupees (***meaning that the average price of meat in five days***) N as 270 Rupees

Example:

For the distribution of 40 school students by age, calculate the arithmetic mean.

Table 3.2

Class interval	Frequency	*X*	*fX*
0-2	3	1	3
3-5	6	4	24
6-8	8	7	56
9-11	12	10	120
12-14	7	13	91
15-17	4	16	64
Total	40		358

How did I got X or X is coming from where? Where did x come from and how is it obtained?

X is called mid-point of the class interval; it is obtained by adding the two-class interval and divided in to two as follows = 0+2= 2/2 = $\underline{\underline{1}}$

$$\therefore \bar{X} = \frac{\overset{N}{\underset{i=1}{\Sigma}} f_1}{\overset{N}{\underset{i=1}{\Sigma}} f_1} = \frac{358}{40} = 8.95 \equiv 9\ years.$$

Important Properties of The Arithmetic Mean:

A. The total of set observations is equal to product of their number and the A.M.

(i) $\Sigma xi = n\bar{X}$ (ii) $\Sigma f ixi$ = N$\bar{X}$

If we assume that observation *xi* is all different, then the ***first relation*** implies that the simple sum is equal to the product of their numbers and simple A.M.

Second relation implies as:

Algebraic sum of the deviation of a set of numbers from their arithmetic mean is always zero.

i.e. $\sum_{i=1}^{\infty}(\mathrm{X} - \bar{\mathrm{X}}) = 0$

Proof $\sum_{i=1}^{N}(X_1 - \bar{X}) = \sum_{i=1}^{N} X - N\bar{X}$ -----------------(1)

But $\bar{X} = \frac{\sum X}{N}$

Cross-multiplying

N$\bar{X}$ = $\sum X$---------------------------- (2)

Substituting in (1)

$\therefore$ N$\bar{X}$ - N$\bar{X}$ = 0 proved

Example:

Supposing

$X_1 = 12, \; = \quad X_2 = 25,$ and $X_3 = 10$

$\sum_{i=1}^{N}(X - \bar{X}) \; = 0$

$\bar{X} = \frac{12 + 25 + 10}{3} = 15.6$

$\therefore \sum_{i=1}^{N=1}(X_1 - 15.6) = (X_1 - 15.6) + \; (X_2 - 15.6\,) + (X_3 - 15.6)$

$$\sum_{1+N}^{N-1} \; (X_1 - 15.6) = (12 - 15.6) + \; (25 - 15.6) + (10 - 15.6)$$

$$= -3.6 + 10.6 + (-5.6)$$

Hence $\sum_{i=1}^{N+1} \; (X_1 - 15.6) = 0$

The sum of the squares of the deviation of a set number $X_1 \, from$ any number a is minimum if and only if a = $\bar{X}$

Proof:

Let X$_1$ = X$_{2,}$ = X$_{3,}$ = 5, X, = 3, and $\bar{X}$ = 15.6 therefore, the sum of the squares of the Deviation $\bar{X}$ = 15.6 is

$$\sum (X - 15.6)^2 = \; (25 - 15.6)^2 + \; (10 - 15.6)^2$$

$$= (-3.6)^2 + 9.4^2 + -5.6^2 = 132.68$$

Example: one.

a. (Direct method).

Computation of simple Arithmetic average.

Table 3.3

S/N.	X
1	5
2	10
3	15
4	20
5	30

6	40
7	50
8	30
Total	**200**

Solution

$\boldsymbol{\bar{X}} = \frac{\sum X}{n} = \frac{200}{8} = \underline{\underline{\mathbf{25}}}$

b. Short cut Method

Table 3.4

S/N.	**X**	Deviation from Assumed Mean **(30) (dx)**
1	**5**	**-25**
2	**10**	**-20**
3	**15**	**-15**
4	**20**	**-10**
5	→ **30**	**0**
6	**40**	**+10**
7	**50**	**+20**
8	**30**	**0**
Total		**-40**

$\boldsymbol{\bar{X}}$ **= A +** $\frac{\sum dX}{N}$ $\boldsymbol{\bar{X}}$ **= 30 -** $\frac{40}{8}$ **= 30 – 5 = 25**

Direct Method: $\boldsymbol{\bar{X}} = \frac{\sum fX}{N} = \frac{\sum fX}{\sum f}$

Short cut Method: $\boldsymbol{\bar{X}}$ **= A +** $\frac{\sum f dX}{N}$

N represents the Total number of frequencies.

Find out the arithmetic mean from the following data.

Table 3.5

Marks	**N0: of Students**
10	2
15	4
20	6
25	8
30	10
Total	**30**

Solution.

a. **(Direct Method).**

(Computation of simple Arithmetic Average)

Table 3.6

Marks *X*	No: of Students = *f*	*fx*
10	2	*20*
15	4	*60*
20	6	*120*
25	8	*200*
30	10	*300*
Total	**30**	*700*

$\bar{\boldsymbol{X}} = \frac{\sum fX}{\sum f} = \frac{700}{30}$ = 23.33 Marks

b. **(Short cut Method).**

$\bar{\boldsymbol{X}}$ **= A +** $\frac{\sum f dX}{N}$

Table 3.7

Marks *X*	No: of Students = *F*	*dx Assumed Mean = 20*	*fdx*
10	2	-10	-20
15	4	-5	-20
20	6	0	0
25	8	5	40
30	10	10	100

Solution $\bar{\boldsymbol{X}}$ **= A +** $\frac{\sum f dX}{N}$ = 20 + $\frac{100}{30}$ = 20 + 3.33 = 23.33 *Ans*

From the following frequency distribution finds out mean wages of the workers.

Table 3.8

Wages	Number of workers
70 – 80	12
80 - 90	18
90 – 100	35
100 – 110	42
110 – 120	50
120 – 130	45
130 – 140	20
140 – 150	8

Solution.

a. **(Direct Method).**

(Computation of simple Arithmetic Average).

Table 3.9

Class interval	**m. v (x)**	**f**	**fx**
70 – 80	75	12	*900*
80 - 90	85	18	*1530*
90 – 100	95	35	*3325*
100 – 110	105	42	*4410*
110 – 120	115	50	*5750*
120 – 130	125	45	*5625*
130 – 140	235	20	*2700*
140 - 150	145	8	*1160*
Total		**∑f = 230**	**∑f = 25400**

$$\boldsymbol{\overline{X} = \frac{\sum fX}{\sum f} = \frac{25400}{230} = 110.43}$$

b. **(Short cut Method).**

$$\boldsymbol{\overline{X} = A + \frac{\sum fdX}{N}}$$

Table 3.10

Class interval	*m. v (x)*	*f*	*dx(105)*	*fdx*
70 – 80	75	12	-30	-360
80 - 90	85	18	-20	-360
90 – 100	95	35	-10	-350
100 – 110	105	42	0	0
110 – 120	115	50	10	500
120 – 130	125	45	20	900
130 – 140	235	20	30	600
140 - 150	145	8	40	320
Total		**∑f230**		**+2320 – 1070, ∑fdx = +1250**

$\bar{X}$ **= A +** $\frac{\sum f dX}{\sum f}$ = 105 + $\frac{1250}{230}$ = 105 + 5.434782609 = **110.4347826**

Alternative to Long Method or Deviation

c. **The Deviation Method:**

If A is any quested or assumed arithmetic mean and $d_1 = X_1 A\ are\ the$ Deviation of X_1 from A, Then

$\bar{X} = A + \frac{\sum d}{N} x\ i$ for a set of data and $\bar{X} = A + \frac{\sum fd}{\sum f} x\ i$ for a frequency table.

If the class size (i) is equal for all classes, then the deviations

m.v = X_i – A can be expressed as mean deviation or assumed mean some time:

$d_x = X_i - A = dx_i$

Where x_i = Positive or Negative integer

$\therefore \bar{X}$ = A + $\left(\frac{\sum f dx}{\sum f}\right) i$

Note: class size (*i*) = 10 and d, X = $\frac{d_x}{10}$

d. (**Calculating arithmetic mean using the deviation method**). $\bar{X}$ **= A +** $\frac{\sum f dX}{N} x\ i$

Table 3.11

Class interval	*m. v (x)*	*F*	*dx(105)*	*d'x(i =10)*	*fd'x*
70 – 80	75	12	-30	-3	-36
80 - 90	85	18	-20	-2	-36
90 – 100	95	35	-10	-1	-35
100 – 110	105	42	0	0	0
110 – 120	115	50	10	1	50
120 – 130	125	45	20	2	90
130 – 140	235	20	30	3	60
140 - 150	145	8	40	4	32
Total		$\sum$**f = 230**			$\sum$**fd'x = +125**

$\bar{X}$ = A + $\frac{\sum f dX}{\sum f} \times i$ = 105 + $\frac{125}{230}$ × 10 = 105 + 0.5434782609 × 10 = 105 + 5.434782609 = **110.43**

Example: Two

Calculate the arithmetic mean $\bar{X}$ from the following table using the deviation method.

Solution:

Table 3.12

Class-mark	F	$d_i = X_i - A$	fd'x
10	3	-90	-270
40	6	-60	-360
70	8	-30	-240
A→100	12	0	0
130	7	30	210
160	4	60	240
Total	40		-420

Let A = 100, and it is any number from X

Therefore, $\bar{X} = 100 + \left(\frac{-420}{40}\right) \times 30 = 100 + (-10.5) \times 30 = 100 - 315 = -215$

Disadvantages of Arithmetic Method ($\bar{X}$)

1. It is largely affected by the extreme values in the data.

It may not correspond to the actual value in the data, and this makes it appear unrealistic.

e.g for the data $2, 3, 4, 5, 6, 20$ $\bar{X} = \frac{40}{6} = 6.67$

2. When there are open-ended class assumptions must be made which may not be accurate.

The Median:

Median of set of observations is the middle-most value when the observations are arranged in order of magnitude. The number of observations smaller than median is the same number greater than it. Thus, median divides the observations into two equal parts. Median in a certain sense, the real measure of central tendency, as it gives the value of most central observations. It is unaffected by the presence of extremely large of small observation and can be calculated from frequency distributions with open-end classes. Median finds the largest application in psychological and achievement tests, e.g. to find the boy of average intelligence, the candidates may be ranked in order of intelligence and the median employed.

An important property of Median is that for any given set of observations the sum of absolute deviations from median is the least.

The median of a set of N numbers arranged in order of magnitude is the middle value or the arithmetic mean of the middle values.

Calculation of Median.

The median is calculated as follows:

(a). ***from simple series*** – The median given data are arranged in order of magnitude. If the number of observations be odd, the value of the middle-most item is the median. However, if the number be even, the arithmetic mean of two middle-most item is taken as median.

(b). ***from the simple frequency*** – the cumulative frequency (less than type) corresponding to each distinct value of the variable is calculated. If the total frequency be N, the value of variable corresponding to cumulative frequency (N+1)/2 gives the Median.

(c)***From grouped frequency distribution*** – Median from a grouped frequency distribution is that value which corresponds to cumulative frequency N/2 [Sometimes (N+1)/2 is used as in (b) above; but this procedure is not correct]. Median from a grouped frequency distribution can be calculated by any of the following methods.

(i) ***By the application of formula for median***: The cumulative frequency is calculated. The class, in which cumulative frequency N/2 lies, is called the median class. Now we apply the formula.

$$\text{Median} = L_i + \frac{\frac{N}{2} - F}{fm} \times \text{c}$$

Whereby

L_i = lower boundary of Median class.

N = total frequency.

F = cumulative frequency below l_1

f_m = frequency of median class;

C = width of median class or class size.

(ii) ***By the application of simple interpolation in a cumulative frequency distribution:*** if F_1 and F_2be the cumulative frequencies shown in the table which are just smaller than and just larger than N/2, and they correspond to the class boundaries $l_1$respectively, than.

$\frac{Median - Li}{l2 - Li} = \frac{\left(\frac{N}{2}\right) - F1}{F2 - F1}$

[**Note***:* *l*$_1$ and *l*$_2$ are lower and upper boundaries of median class, and $\frac{N}{2}$ lies between F1 and F2].

(iii). ***Graphic method:*** an approximate value of median can be obtained graphically from ogive, or cumulative frequency polygon. Draw a horizontal line from the point $\frac{N}{2}$ on the vertical scale showing the cumulative frequencies, until it meets the ogive (either less-than or more than type). From the point of intersection, a perpendicular is now drawn on the horizontal axis. The position of the foot of the perpendicular is read from the horizontal scale showing values of the variable, and this give the median.[12] If both Ogive (less-than and more than) are variable on the same graph paper, the position of the foot of the perpendicular drawn from the point of intersection of two ogives, give the median.

Advantages of Median.

1. Median is not difficult to understand, although it is notes popular as the arithmetic mean.
2. It is easy to calculate. Even when all the observations are not known, median can be calculated, provided the general location of all observations and values near to middle are available. Median can also be calculated without difficulty from grouped frequency distributions with classes of unequal width or with open-end classes.
3. Median is applicable to qualitative data in psychological and social studies, where numerical measurement may not be available, but it is possible to rank the objects in some order.

Disadvantages of Median.

1. For the calculation of median, the data must be arranged.
2. Unlike A.M or G.M., it cannot be treated algebraically. Given the median of several groups of observations, median of the composite group cannot be determined.
3. If is desired to give greater importance to large or small values, median is unsuitable.
4. The calculation of median from group frequency distribution is based on simple interpolation, which assumes that the observations in the median class are uniformly.

   The position (order) of the median is $\frac{N+1}{2}\ if\ N\ is\ odd\ and$

   $\frac{1}{2}\left[\frac{N}{2} + \frac{N+2}{2}\right]$ If N is even

[12] Ibis Das p.141

Example One:

For the numbers 16, 5, 4, 2, 9, 8. Determine the Median

Solution:

Step one: Arrange the data in ascending order or descending order.

2, 4, 5, 8, 9, 16

N = 6

Step Two:

Find out the position (order) of the Median. Since 7 is odd

$$\frac{7+1}{2}4^{Th}$$

∴ **Median = 8**

ExampleTwo:

For the set 5, 5, 7, 12, 15, 18
Calculate the Median
Solution:
Since N=8(even)
∴The order of the median

$\frac{1}{2}\left[\frac{N}{2}+\frac{N+2}{2}\right]=\frac{1}{2}[4+5]=\frac{9}{2}=4.5$

The median = $\frac{9+11}{2}=10$

Illustration

At evident from the above two examples, median

1/ Avoid giving full weight to extreme values.

2/ Represent the typical half $\left(1/_2\right)$ way point regardless of extreme or missing values/scores.

For frequency distribution the median is given by

$\bar{X}=L_n+\left[\frac{\frac{N}{2}-\sum f_1}{fmedian}\right]c$

Where:

$L_1=$ Lower class boundaries of the median class

$N=\sum f=Totalfrequency.$

$\sum f_1 = Sumoffrequenciesofclasseslowerthanthemedianclass.$

$fmediam$ = Frequency of the median class. $C = Classsize.$

Example:

From the below distribution of the Masses of 100 children in the primary school calculate the median.

Table 3.13: Masses of 100 children in a primary school.

Masses (kg)	No. of children (f)
8-10	6
	$\sum f_1$
11-13	18
median class ← 14-16	42 →
17-19	26
20-22	8
Total	**100**

Solution:

Order (position) of the median = $\frac{1}{2}\left[\frac{N}{2}+\frac{N+2}{2}\right] = \frac{1}{2}[50+51] = 50.5 = \underline{\underline{51}}$

Therefore $\boldsymbol{\bar{X} = I_n + \left[\frac{\frac{M}{2}-\sum f_1}{f(median)}\right] c}$

Whereby:

14 – 16 is the class containing the median, obtained by cumulating the

Frequency 5 + 18=23 + 42 = 65 where 51 is inclusive

$L_1 \frac{13+14}{2}$ = 135   N = 100 $\sum f_1 = 23 \ and \ C = 3$

Therefore $\bar{X}$ = 135 + $\left[\frac{50-23}{42}\right] \times 3 = 135 + 1.9286 = 15.43kg$

Example.

Find the median and median class of the given data below.

Class boundaries	15 – 25	25 – 35	35 – 45	45 – 55	55 – 65	65 – 75
Frequency	4	11	19	14	0	2

Solution:

First method

We first calculate the cumulative frequency against each class interval as shown below.

Table 3.14

Class boundaries	Frequency	Cumulative frequency
15 – 25	4	4
25 – 35	11	15 $\sum f_1$
(35 – 45)	19 f *median*	34
45 – 55	14	48
55 – 65	0	48
65 – 75	2	50 = N

First we need to find out samples which are given in the formula.

Where: $l_1 = \frac{35-35}{2} = \frac{70}{2}$ = 35, N = 50, $\sum f_1 = 4 + 11 = 15$, f *median* = 19, c = 10

$$\boldsymbol{\bar{X} = I_n + \left[\frac{\frac{M}{2} - \Sigma f_1}{f(median)}\right] c}$$

$$\boldsymbol{\bar{X}} \mathbf{= 35 +} \left[\frac{\frac{\mathbf{50}}{\mathbf{2}} \mathbf{- 15}}{\mathbf{19}}\right] \mathbf{10,} = 35 + \left[\frac{25-15}{19}\right] 10 = 35 + \left[\frac{10}{19}\right] 10 = 35 + [0.5263157895]10$$

35 + 5.2631578947 Median is = 40.26*Ans*.

Second method, interpolation.

Since$\frac{N}{2}$ = 25 lies between the cumulative frequency 15 and 34 the corresponding values of the variable. Median, must lie in the interval between 35 and 45. The median class is, therefore, (35 – 45).

Now applying simple interpolation.

$$\left[\frac{Median - 15}{45 - 35}\right] = \left[\frac{25 - 15}{35 - 15}\right]$$

$\left[\frac{Median-15}{19}\right] = \left[\frac{10}{19}\right]$. Median – 35 = $\left[\frac{10}{19}\right] \times 10$.

Median – 35 = $\frac{100}{19}$ = 5.2631578947

Median = 35 + 5.2631578947, Median is = 40.26

Example.

The following is the table which gives you the distribution of marks scared by some students in exams of statistics.

Marks	0 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 – 70	71 - 80
No. of Student	42	38	120	84	48	36	31

i. Find the median of the marks.

ii. Find the percentage of the failure if the minimum for pass is 35 marks.

Solution.

Note. Here class limits are given and therefore, we must find out class boundaries for the cumulative frequency distribution.

Table 3.15

Class (marks)	Class boundaries	Cumulative frequency
0 – 20	20.5	42
21 – 30	30.5	80
31 – 40	40.5	200f_1 = $\frac{N}{2}$ = 199.5
41 – 50	50.5	284
51 – 60	60.5	332
61 – 71	70.5	368
71 - 80	80.5	399 = N

Let us apply the rule.

(i). First method median.

$\bar{X} = I_n + \left[\frac{\frac{M}{2} - \Sigma f_1}{f(median)}\right] cl_1 = \frac{31+30}{2} = \underline{\underline{30.5}}$, N = 399, f_1 = 80, f_{median} = 120, c = 10

$\bar{X} = 30.5 + \left[\frac{\frac{399}{2} - 80}{120}\right] 10$

$\bar{X} = 30.5 + \frac{199.5 - 80}{120} \times 10$, = 30.5 + $\frac{119.5}{120} \times 10$, =30.5 + (0.9958 × 10), = 30.5 + 9.958 = $\underline{\underline{\mathbf{40.458}}}$

(ii). Second method interpolation.

$$\left[\frac{Median - l1}{l2 - l1}\right] = \left[\frac{\frac{N}{2} - f1}{f2 - f1}\right].$$

Where: l_1 = 30.5, l2 = 40.5, N = 399, f_1 = 80, f_2 = 200

$$\left[\frac{Median - 30.5}{40.5 - 30.5}\right] = \left[\frac{\frac{399}{2} - 80}{200 - 80}\right] \quad = \quad \left[\frac{Median - 30.5}{10}\right] = \left[\frac{199.5 - 80}{200 - 80}\right].$$

$\left[\frac{Median - 30.5}{10}\right] = \left[\frac{119.5}{120}\right]$x 10 $= median - 30.5$ $\left[\frac{119.5}{120}\right]$ x 10

Median = 30.5 + (0.9958 × 10). Median = 30.5 + 9.958 = **40.458**.

Remarks.

To find the percentage of the failure if the minimum for pass is 35 marks.

Now$f_1 = 80\,\frac{4}{10} \times 120$, 80 + 0.4 × 120, 80 + 48 = 128.

$\frac{128}{399}$ × 100 = 0.320802005 × 100 = 32.0802005 = **32%** **Ans.**

32% are the percentage of failure students.

Example (4).

Draw two ogives from the following data and find out the median of wages by using cumulative frequency.

Weekly wage in SSP	0 – 20	20 – 40	40 – 60	60 – 80	80 - 100
No: of workers	40	51	64	38	7

Solution.

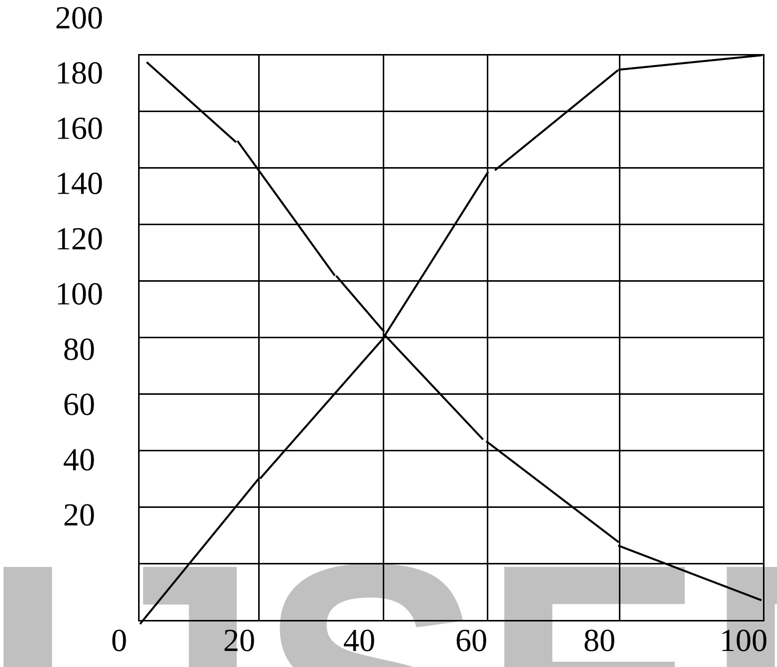
Let us draw the Ogive of less than type for this purpose, we have to constructs a cumulative frequency distribution.

Note. This is in the data class boundaries are given as shown below.

Class boundary.	0	20	40	60	80	100
Cumulative frequency.	0	40	91	155	193	200 N

The cumulative frequencies are plotted on graph paper against class boundaries and the ogive is drawn from the point $\frac{N}{2} = \frac{200}{2}$ = 100 on the vertical scale, horizontal line is drawn meeting the ogive from the point of intersection a perpendicular (*dotted line in the figure above*) is now drawn on the horizontal axis. The point showing the foot of the perpendicular is now read from the scale and is found to be approximately 34.

Thus, the median = 43.



Note:

Geometrically, the median is the value of X corresponding to the vertical line which divides a histogram into two equal parts having equal areas.

The Mode$(\overline{X})$:

Mode is a French word for fashion. It shows the score (value) that Occurs in most frequency or often in a distribution or it is that value which occurs with the greatest frequency. Mode of a given set of observation is that value which occurred with the maximum frequency. It is most typical or prevalent values, and at times represents the true characteristics of distribution as measure of central tendency. The Mode may exist, and even if it does exist, it may not be unique. Consider the following records of milk production (kg) in a farm over a period of one week. 104, 100, 99, 100, 101, 100, 102. What is the value, which is recorded for many times?

The value recorded many times is 100 kg. For it appear three times in the seven days of the week. Therefore, this value which is repeated for many is called the mode. Hence mode is any value (s) that occurs frequently or that value with the highest frequency.

Calculation of Mode.

From simple series

Example One:

The table below shows the distribution of marks scored by students in months.

Find the mode?

Table 3.16

Marks	Frequency
10	2
20	4
30	8
40	12
50	6
Total	**32**

The mode of the distribution is 40 marks.

Example Two:

The daily temperature record in the first week of February were.

(a) 21, 19, 24, 21, 15, 14, 21, centigrade and

(b) In the second week there was no mode

10, 19, 24, 25, 30, 35, 44,

Find the mode?

Solution:

(a) The mode is 21 centigrade in the first week and

(b) In the second week there was no mode

Example Three.

The table below shows the distribution of 150 infant according to age class. Find the mode.

Age	20	25	30	35	40
Freq.	14	20	45	51	20

For this table, the mode is 35 years.

Example 4

1/ the set 2, 2, 5, 1, 3, 7, 7, 7, 44. Has a Mode of 7

2/ the set 2, 3, 4, 4, 5, 7, 7, 8, have twos 4 and7.

3/ the set 6, 7, 8, 13, 2, have no Mode.

For the frequency distributions, the mode is calculated by a formula.

Mode $(\bar{X}) = L_1 + \left(\frac{\Delta_{1`}}{\Delta_1+\Delta_1}\right) c$

Where:

$\boldsymbol{L_1}$ **= Lower class boundary of the mode class**

Δ_1= Excess of the model frequency over frequency of the next lower class.

Δ_2 = Excess of model frequency over frequency of the next higher class

Example one:

From the data of table 3.17, compute the mode of the distribution.

Table 3.17

(a) Masses of 100 children in a primary school.

Mass (kg)	No. of Children
8-10	5
11-13	18 Δ_1
Model Class ← **14-16**	42 Δ_2
17-19	27
20-22	22
Total	**100**

Illustration:

The model class is the one corresponding to the largest frequency i.e.

14-16 correspond to 42.

$$L_1 = \frac{13+14}{2} = 135$$

$$\Delta_1 = 42 - 18 = 24$$

$$\Delta_2 = 42 - 27 = 15 \qquad \& \qquad C = 3$$

$$\therefore \text{ Mode } (\bar{X}) = 135 + \left(\frac{24}{24+15}\right)3$$

= 135 + 1.846 = 15.35Kg

Note:

A distribution may also have two or more modes for example, suppose the frequency corresponding to the class 20-22 is also 42 i.e.

Table 3.18

Mass	Frequency
8-10	5
11-13	18
14-16	42
17-19	27
20-22	42
Total	**134**

Then,

First mode = $L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) C$

Where:

$$L_1 = \frac{13 + 14}{2} = 135$$

$\Delta_{1 = 24}$

$\Delta_{2 = 15}$ & C=3

Therefore, Mode =$135 + \left(\frac{24}{39}\right) \times 3 = 1535\ kg$

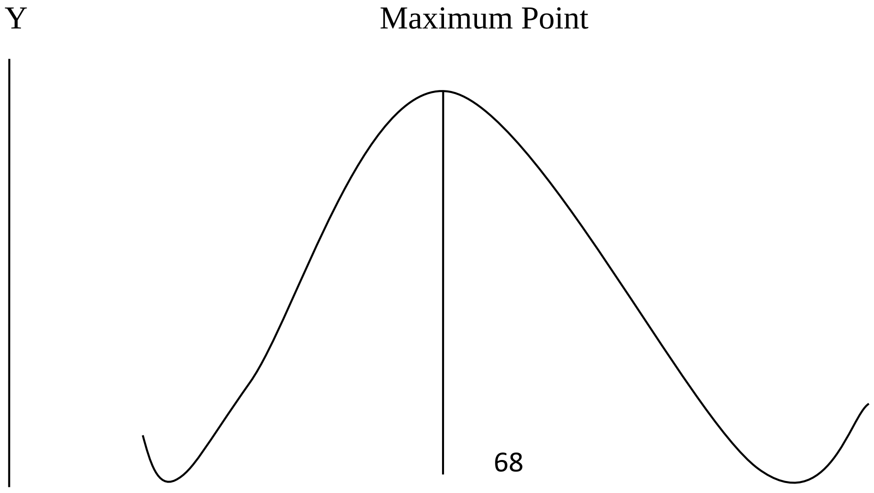
Second Mode = $L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) C$

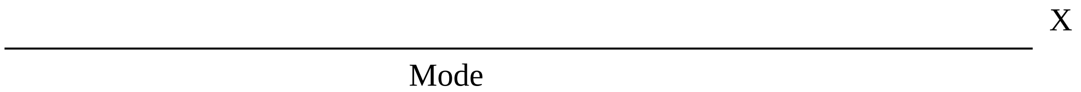
Whereby:

$$L_1 = \frac{19 + 20}{2} = 195\ C = 3, \Delta_1 = 42 - 27 = 15,\ \ \Delta_2 = 42 - 0 = 42.$$

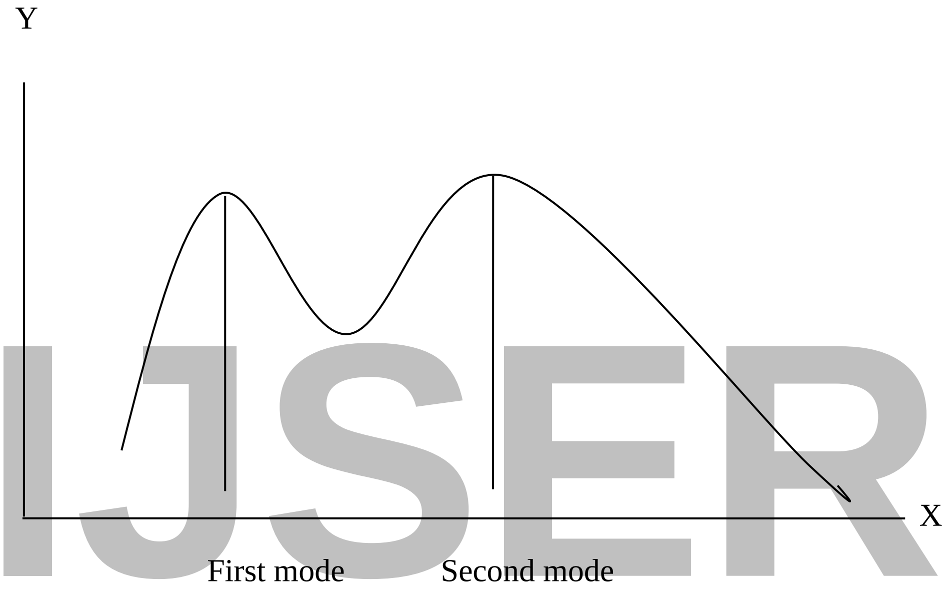
Hence Mode: for frequency curves the Mode is the value of x corresponding to the maximum point on the curve.

Example.





In case of two mode,



Empirical Relation between the Mean, Median and Mode.

For the unit-model frequency curves which are moderately skewed, we have:

(Mean – Mode) = 3 (Mean – Median)

Example:

For a distribution with Mean 6.5 and Median 5.2 find its mode.

Solution:

$$\left(\bar{X}-\hat{X}\right)=3\left(\bar{X}-\tilde{X}\right)$$

$$\left(6.5-\hat{X}\right)=3(6.5-5.2)=\therefore\ \hat{X}=2.6$$

3.5 <u>Minor Means:</u>

(i) The Geometric Mean (G)

The geometric mean of a set of number $X_{1,}X_{2\,,}X_3,\ldots,X_n\ is\ the\ N^{th}$root of the product and the numbers.

For the set of numbers C=$\sqrt[N]{X_{1,}X_2,\ldots.,X_N}$

Example:

Find the geometric mean of the numbers 2, 5, 8.

Solution:

$$C = \sqrt[3]{(2),(5),(8),} = \sqrt[3]{80} = 43$$

For the frequency distribution or group data

Geometric mean is the Anti – log of $\frac{\sum f \log X}{N}$

Example:

For the data below calculate.

The Geometric Mean.

Table 3.19

Class Boundaries	Freq.	X	Log x	F log x
0-3	3	1.5	0.17	0.51
3-6	6	4.5	0.65	3.90
6-9	13	7.5	0.87	11.31
9-12	6	10.5	1.02	6.12
12-15	2	13.5	1.13	2.26
Total	**30**			**24.1**

$\therefore \frac{\sum f \log X}{f} = \frac{24.1}{30} = 0.8$

Now taking the Anti- logarithm of 0.3

$\therefore$ Geometric mean = 6.3

(ii) The Harmonic Mean (H):

The harmonic mean of set of N number $X_1, X_2, X_3, \dots, X_N$ is the

Reciprocal of the arithmetic mean of the reciprocals of the numbers.

Symbolically:

$H = \frac{1}{\Sigma \frac{1}{X} N} = {}^{N}\!/_{\Sigma \frac{1}{X}}$ For a set data

And $\frac{1}{N} \Sigma \frac{f}{X}$ for grouped data.

Example:

For the previous example calculate. Harmonic mean both for the set of data and grouped data.

Solution:

(a) For a set of data H= ${}^{3}\!/_{\frac{1}{2} + \frac{1}{5} + \frac{1}{8}} = {}^{3}\!/_{0.825}$

$\therefore$ H = 3.6363636

(b) For a grouped data: H = $\frac{1}{N}\Sigma^{f}/_{X}$

Table 3.20

Class- boundaries	Freq.	X	$^{f}/_{X}$
30-33	3	1.5	2
33-36	6	4.5	1.3
36-39	13	7.5	1.7
39-42	6	10.5	0.6
42-45	2	13.5	0.2
Total	**30**		**5.8**

∴ H = $\frac{1}{30}(5.8) = 0.19$

(iii) Relationship between mean, geometric mean and harmonic mean $(\bar{X}, G, H)$

The Geometric mean of a set of positive numbers $X_{1,}X_{2,}X_{3,} \ldots. X_N$ *is* less than or equal to their arithmetic mean but is greater or equal to their harmonic mean i.e. $H \leq G \leq \bar{X}$.

Example:

For the numbers 2, 5 & 8

$$\bar{X} = {}^{\Sigma X}/_{N} = \frac{2+5+8}{3} = 5$$

$$G_1 = \sqrt[3]{(2)(5)(8) = 43}$$

$$H = {}^{N}/_{\Sigma\frac{1}{X}} = \frac{3}{\frac{1}{2} + \frac{1}{5} + \frac{1}{8}} = 3.63$$

Therefore H $< G\ \bar{X}\ Proved$

(iv) Quartiles and Percentage

Quartiles:

These are the division of the data into four (4) parts equal pans donated by $Q_{1,}Q_{2,}Q_{3,}$ with respective orders (position)

$\frac{N}{4}, \frac{2N}{4}, \frac{9N}{4}$ i.e $Q_1 \mid Q_2 \mid Q_3 \mid$

Deciles:

These are values which divided the data into ten (10) equal parts donated by $D_1, D_2, \ldots., D_9$ with respective orders (position).

$$\frac{N}{10}, \quad \frac{2N}{10}, \ldots., \quad \frac{9N}{10}$$

Percentile:

These are values which divide the data into hundred (100) equal parts, donated by $P_1, P_2, \ldots., P_{99}$ with respective orders (position).

$$\frac{N}{100}, \quad \frac{2N}{100}, \quad \frac{3N}{100} \cdots, \quad \frac{99N}{100}$$

Their median = $Q_{2,}Q_5$ *and* P_{50} *respectively*

Example:

For the Previous Date of Table 2.3

Calculate: $Q_1 = Q_2$ (ii) $D_5 D_5$ (iii) $P_{10} P_{50}$

Solution:

(i)The order of $Q_1 = \frac{N}{4} = \frac{30}{4} = 7.5$

Where N= Total frequency

$\therefore Q_1 = L_1 + \left(\frac{\frac{N}{4} - \sum f_1}{fQ_1}\right) c$

Now from Table 2.3

$L_1 = 3, \sum f_1 = 3 \quad FQ_1 = 6 \quad \therefore C = 3$

$\therefore Q_1 = 3 + \left(\frac{7.5-3}{6}\right) 3$ =3+2.25= 5.2

$Q_1 = L_1 + \left[\frac{\frac{3N}{4} - \sum f_1}{fQ_3}\right] c$

The order of $Q_1 = \frac{3N}{4} = \frac{3(30)}{4} = 2.25$

From Table 2.3

$L_1 = 9 \qquad \Sigma f = 3 + 6 + 13 = 22, \quad fQ_1 = 6, \qquad c = 3$

$$\therefore Q_1 = 9 + \left(\frac{22.5 - 22}{6}\right) 3 \quad = 9 + 0.25 = 9.25$$

(ii) The order of $D_1 = \frac{3N}{10} = \frac{3(30)}{10} = 9$

$$D_1 = L_1 + \left[\frac{\frac{3N}{10} - \sum f_1}{fD_1}\right] c$$

Where $L_1 = 3, \sum f_1 = 3, \qquad fD_1 = 6, \quad c = 3$

$$\therefore \ D_1 = 3 + \left(\frac{9-3}{6}\right) 3 \quad = 3 + 3 \quad = 6$$

$$D_1 = L_1 + \left[\frac{\frac{5N}{10} - \sum f_1}{fD_1}\right] c$$

Order of $\qquad D_1 = \frac{5N}{10} = \frac{(30)}{10} = 15$

From Table 2.3

$$L_1 = 6, \qquad \Sigma f_1 = 3 + 6 = 9, fD_1 = 13 \qquad C = 3$$

Therefore, $D_1 = 6 + \left(\frac{15-9}{13}\right) \times 3 \quad = 6 + 1.3846154 = 7.3846154$

(iii) $P_{10} = L_1 + \left[\frac{\frac{10N}{100} - \sum f_1}{fP_{100}}\right] \times C$

The Order of $P_{99} = \frac{10(30)}{100} = 3$

From Table 2.3

$$L_1 = 0, \qquad \sum f_1 = O, \qquad fP_{10} = 3, \quad C = 3$$

$$\therefore P_{10} = O + \left(\frac{3-0}{3}\right) 3$$

$$\therefore \ P_{90} = L_1 + \left[\frac{\frac{90N}{100} - \sum f_1}{fP_{90}}\right] C$$

The Order of $\qquad \therefore \ P_{90} = \frac{90(30)}{100} = 27$

From Table 2.3

$$L_1 = Q_1 \sum f_1 = 3 + 6 + 13 = 22, \qquad fP_{90} = 6 \qquad C = 3$$

$$\therefore \ P_{90} = 9 + \left(\frac{27-22}{6}\right) 3 = 9 + 2.5 = 11.$$

CHAPTER FOUR.

MEASURE OF DISPERSION (VARIATION)

A measure of central tendency locates a point of concentration, but it tells us nothing about the degree of concentration about the way the observations are dispersed throughout the distribution. However, the degree to which numerical date tends to spread about an average Value is called dispersion or variation these are:

- The Range:
  (a) Semi inter-quartile range (Quartile Deviation).
  (b) The 10-90 percentile Range
- The Mean Deviation
- The Standard Deviation
- The Variance and
- The Coefficient of Variation of this many possible measures of dispersion, only three (3) are in wide general use today and these are:
  1) The standard Deviation
  2) The Coefficient of Variation and
  3) The Range

The Range:

The Range is the first measure of dispersion. It is usually defined as the difference between the smallest and largest value of distribution. For frequency distribution, the range is given by:

1) Class mark of highest class minus, Class mark of the lowest Class
2) Upper class boundaries of the highest class minus lowest class boundary if there is lowest class.

Example: For the Class of Table 2.3

Class Boundaries	Class mark	Freq.
20 – 23	1.5	3
23– 26	4.5	6
26– 29	7.5	13
29– 32	10.5	6
32– 35	13.5	2
Total		**30**

Range = 13.5 - 1.5 = 12
15 – 0 = 15

(a) Quartile Deviation (Q):

This measures the dispersion of the part of any distribution lying between the two Quartile. I.e., upper and lower quartiles.

The semi-inter-quartile range (Q) or the Quartile deviation is donated by:

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

Example:

For the previous table 2.3 we have calculated $Q_1 = 5.25\ and$

$$Q_3 = 9.25\ then\ the\ sermi - inter - quartile\ range\ (Q) = \left(\frac{9.25 - 5.25}{2}\right) = 4$$

Remark:

The smaller results given by this formula is the dispersion of the middle half of the distribution about the median.

(b) <u>The 10-90 percentile Range:</u>

This is donated by $P_{90} - P_{10}$

But from Table 2.3

$$P_{90} = L_1 + \left[\frac{\frac{90N}{100} - \sum f_1}{fP_{90}}\right] C = 11.5$$

and

$$P_{10} = L_1 + \left[\frac{\frac{10N - \sum f_1}{100}}{f\ P_{10}}\right] C = 3$$

$$\therefore P_{90} - P_{10} = 11.5 - 3 = 8.5$$

The Mean Deviation

Mean Deviation measures the mean $(\bar{X}) of\ the\ sum\ of\ all$ the deviation of every item in the distribution from a central Value. It provides a useful method of comparing the relative tendency of the value.

The mean deviation (M.D) of a set of numbers $X_{1,}\ X_{2,} \ldots, X_N\ is$ defined as:

$$\text{M.D} = \frac{\overset{n}{\Sigma}|X - \bar{X}|}{N}$$

Example:

For the numbers 2, 3, 6, 8, 11

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$$

$$\therefore \text{M.D}\ \frac{|2-6|+|3-6|+|6-6|+|8-6|+|11-8|}{5} = {}^{14}\!/_{5} = 2.8$$

For grouped data, the Mean Deviation is given by M.D = $\frac{\Sigma f_1|X - \bar{X}|}{N}$

Example:

For the distribution of Table 2.3 Calculate mean Deviation.

Solution:

Class- Boundaries	F	X	fX	$f\|x-\bar{X}\|$
0 – 3	3	1.5	4.5	17.4
3 – 6	6	4.5	27	16.8
6 – 9	13	7.5	97.5	2.6
9 – 12	6	10.5	63.0	19.2
12 – 15	2	13.5	27	12.4
Total	**30**		**219**	**68.4**

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{219}{30} = 73$$

$\therefore \boldsymbol{M.D} = \frac{\Sigma f_1|X-\bar{X}|}{N} = \frac{63.4}{30} = 2.28$

The Standard Deviation

Today we come to one of the most important tools of statistics, one on which much of our letter work is based make sure you know how to compute readily before going on. Don't worry about later "Understanding" theσ, a comprehension of the use and significance of this statistic will grow on you as vague you use it. Standard deviation is the measure that statisticians use to study how spread out a data (score) is. It takes everything into account, just as the median did for the measure of central tendency.

Algebraically, the determining formula for the standard deviation is

$$S = \sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{\Sigma x^2}{N}}$$

Example:

For the number (s) 2, 3, 8 compute the Std Deviations

Solution:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{2+5+8}{3} = 5$$

σ or s = $\sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{(2-3)^2+(0)^2+(8-5)^2}{3}}$

Remark:

The smaller is the σ^2 or σ of a given distribution, the more concentrated are the observations and the larger the value of σ^2 or σ, the more the observations are dispersed. In the case of a frequency distribution:

σ or S = $\sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}}$

Example:
For the date of table 2.3. Calculate the standard deviation for the frequencies.

Class	f	X	fX	$f(X-\bar{X})2$
100- 130	3	1.5	4.5	100.92
130 -160	6	4.5	27.0	4704
160 -190	13	7.5	97.5	0.52
190 - 220	6	10.5	63.0	61.44
220 - 250	2	13.5	27.0	76.88
Total	**30**		**219**	**286.8**

Solution:

$$\bar{X} = \frac{219}{30} = 73$$

$$\text{S} = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}} = \sqrt{\frac{\Sigma f x^2}{N}}$$

$$S = \sqrt{\frac{286.8}{30}} \qquad = 3.09$$

Short Methods for Computing Standard Deviation.
The Standard Deviation can be given by the formula.

$S\ or\ \sigma = \sqrt{\frac{\sum X^2}{N}} - \bar{X}$ For a set of data

And

$\sigma = \sqrt{\frac{\sum f\ X^2}{N}} - \bar{X}^2$ For frequency distribution

Proof:

$$S = \sqrt{\frac{\sum(X-\bar{X})^2}{N}} = \sqrt{\frac{\sum X^2}{N}} - \bar{X}^2$$

LHS

$$S = \sqrt{\frac{\sum(\bar{X})^2}{N}}$$

Squaring the summation sign:

$$S = \frac{\sum (X - \bar{X})^2}{N} = \frac{\sum (X^2 + \bar{X} - 2\bar{X}\ X)}{N}$$

Introducing the summation sign:

$$\frac{\sum X^2 + N\bar{X}^2 - 2\ \bar{X}\sum X}{N}$$

This can be rewritten as:

$$\frac{\sum X^2}{N} + \frac{N\bar{X}^2}{N} - \frac{2\overline{X\ \sum X}}{N}$$

Since $\frac{\sum X}{N} = \bar{X}$

Therefore

$$S^2 = \frac{\sum X^2}{N} + \ \bar{X} - 2\bar{X}^2$$

S= $\sqrt{\frac{\sum X^2}{N} - \bar{X}^2}$

Proved

Now solving the data of table 2-3 by using this formula

Class boundaries	**Freq. (f)**	**x**	**fX^2**
0-3	3	1.5	6.75
3-6	6	4.5	121.50
6-9	13	7.5	731.50
9-12	6	10.5	661.50
12-15	2	13.5	364.50
Total	**30**		**1885.50**

$$S = \sqrt{\frac{\sum fX^2}{N} - \bar{X}^2} = \sqrt{\frac{1885.50}{30}} - (7.3)^2$$

$$= \sqrt{62.85 - 5329} = \sqrt{956} \therefore S = 3.09$$

<u>The Deviation Method:</u>

If we use the deviations d, = x$_{1,}$ -A where A is the assumed arithmetic mean than the standard deviation can be given by:

σ =$\sqrt{\frac{\sum d^2}{N}} - \left(\frac{\sum fd}{N}\right)^2$

Example:

For the numbers or set data 2, 5, 8, = 2 + 5 + 8 = 15, = $\frac{15}{3}$ = 5.   Let A = 5

$\therefore \Sigma d = (2-5) + (5-5) + (8-5) = -3 + 0 + 3$

-3 + 0 + 3 = 0

$\Sigma d^2 = (-3)^2 + (0)^2 + (3)^2 = 9 + 0 + 9 = 18$

$\therefore \sigma = \sqrt{\frac{18}{3} - \left(\frac{-3}{3}\right)^2} = \sqrt{6-1} = 2.236067977$

Example:

Calculate the Standard deviation for the previous table 2-3 by using the deviation method.

Solution:

x	**f**	**d $(x-7.5)$**	**d^2**	**f d**	**Fd^2**
1.5	3	-6	36	-18	108
A → 7.5	6	-3	9	-18	54
7.5	13	0	0	0	0
10.5	6	3	9	18	54
13.5	2	6	36	12	72
Total	**30**			**-6**	**288**

Hence $S = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2}$

Where C = class size

Example: Calculate the standard deviation for the previous table 2-3 by using the coding method

Solution:

X	Freq. (f)	D = (X$_i$,A)	U = d/c	U^2	fU	fU^2
1.5	3	-6	-2	4	-6	12
4.5	6	-3	-1	1	-6	-6
7.5 → A	13	0	0	0	0	0
10.	6	3	1	1	6	-6
13.5	2	6	2	4	4	8
Total	**30**				**-2**	**32**

$$S = \sqrt{\frac{32}{30} - \left(\frac{-2}{30}\right)^2} = \sqrt[3]{1.055} = 3.09$$

The Variance $(\mathbf{S^2}\text{or } \boldsymbol{\sigma^2})$:
This is the square of the standard deviation in symbols donated by $S^2$for the sample and σ^2 for the Population.

Since $$S = \sqrt{\frac{\sum_{i=1}^{\infty} (X-\bar{X})^2}{N}}$$

Therefore, $S^2 = \frac{\sum(X-\bar{X})^2}{N}$

Example:

For the prevision S example of table 2-3 the sample variance S = $(3.09)^2\ therefore\ S^2 = 9.5481$

Absolute and Relative Dispersion

Absolute dispersion is the actual dispersion as determined from the standard deviation, whereas relative dispersion is equal to absolute dispersion divided by an average. It is required for purpose of comparison.

Coefficient of Variation (v)

The coefficient of variation is used to measure the relative dispersion; it is given by ${}^{S}/_{\bar{X}}$ and expressed as a percentage. It is also a dimension- less quantity.

Example:

A manufacture of television (T.V) tubes has two types of tubes. Tube A and B with respective mean lifetime $\bar{X}$ A=1495 and $\bar{X}$: B = 1875 hours and the standard deviation $S_A = 280\ hr\ S_B = 310$ hr, which tube has greater:

1- Absolute dispersion
2- Relative dispersion

Solution:

1- Given $S_A = 280\ hr\ and\ S_B = 310. since\ absolute\ dispersion\ is\ the\ a$ctual dispersion as determined from the standard deviation, tube B has grater absolute dispersion.
2- For relative dispersion, we have to calculate coefficient of variation for each tube.

C.V = ${}^{S}/_{\bar{X}}$

C.V of Tube B=${}^{S_A}/_{X_B}$=${}^{280}/_{1494} \times 100$

$\therefore$ C.V=187%

C. V of tube B= ${}^{S_A}/_{SX_B}$= ${}^{310}/_{1875} \times 100$

$\therefore$ C.V=187%

Thus tube A has greater relative dispersion

Standardized variable (standard Scores)

This is given by Z $=\frac{X-\bar{X}}{S}$ and is called- standardized variable, the deviations are said to be expressed in standard unit or standard scores.[13] You can see the book of Probability for second year Upper Nile University. Z is a dimension-less quantity and is very useful in comparing scores from different distribution.

Example:

A student received a mark of 94 in the examination of introduction to economics for which the mean mark was 86 and the standard deviation was 20 Tone introduction to statistics for which the mean was 80 and standard deviation was 26, he received a mark of 82, in which course (Subject) was his relative standing better.

Solution:

Given: Introduction to economics $\bar{X} = 86, standard\ deviation = 20$ &

$$X = 94$$

Introduction to statistics $\bar{X} = 80, standard\ deviation = 26$ &

$$X = 82$$

For Comparing Scores from different distribution like this, we use the Standard Score

$$Z = \frac{X - \bar{X}}{S}$$

Therefore, the Student Standard Score:

(a) On Introduction to economic (Z)=$\frac{94-86}{20} = 0.4$

(b) On Introduction to Statistic (Z)=$\frac{82-80}{26} = 0.0769$

Hence, his relative standing in introduction to Economic is better than that in introduction to Statistics.

[13] Poulino F. D. Ajang, 2015, Lecture Note for Probability and Statistics for Second year undergraduate, Upper Nile University, p.69.

CHAPTER FIVE

MOMENT SKEWENESS AND KORTOSIS.

Moment for a set of data

Given n observation *x1, x2, ….. xn* and an arbitrary constant A,

$\frac{1}{n}\sum(x - A)$ is called 1st moment about A,

$\frac{1}{n}\sum(x - A)^2$ is called 2nd moment about A,

$\frac{1}{n}\sum(x - A)^3$ is called 3rd moment about A, and so on. Let us denote these moment successively by $m1', m2', m3', etc.$ (sometime these are also represented by $\mu1', \mu2', \mu3'\ etc$)

Then $m1' = \frac{\sum(x-A}{n} = \frac{\sum x - \sum A}{n} = \frac{\sum x - nA}{n} = \bar{X} - A$

The 1st moment about A is equals $(\bar{X} - A)$.

Moment about zero (i.e. when A = 0), and moments about mean (i.e. when A = $\bar{X}$) are particularly important.

1. **Moment about Zero (or Raw moments):**

1st moment about zero = $\frac{1}{n}\sum x = m1'\bar{X}$

2nd moment about zero = $\frac{1}{n}\sum x^2$

3rd moment about zero = $\frac{1}{n}\sum x^3$ and so on. Note that the 1st moment about zero is the mean $\overline{X\ .}$

$m1' = \bar{X}$.

2. **Moments about mean (or Central Moments).**

1stmoment about mean = $\frac{1}{n}\sum(x-\overline{X)} = 0$

2ndmoment about mean = $\frac{1}{n}\sum(x- \bar{X})^2 = \sigma 2$

3rd moment about mean = $\frac{1}{n}\sum(x- \bar{X})^3$

4thmoment about mean = $\frac{1}{n}\sum(x- \bar{X})^4$and so on.

These are usually denoted by m1, m2, m3, m4, etc. (sometime, these are represented by $\mu 1', \mu 2', \mu 3'\ etc$). Note that the 1st central moment is always zero, and 2nd central moment is variance $\sigma 2$ m1 = 0, m2 σ^2

From the second relation, we find that the standard deviation is square rood of the second central moment $m_{2.}$

The 3rd central moment m_3 is used to measure skewness and the 4th central moment m_4 to measure kurtosis. Higher order moments m_5, m_6 …etc. are seldom used.

In general, given n observation x1, x2, x3,…..xn, the r-th order, moments (r = 0, 1, 2, …) are defined as follows:

r-th moment of A: $mr' = \frac{1}{n}\sum(x\text{-}A)^{r}$

r – th raw moment: $mr' = \frac{1}{n}\sum x^{r}$

r-th central moment: $mr' = \frac{1}{n}\sum(x\text{-}\ \overline{X}\ )^{r}$

For a frequency distribution,

r-th moment about A: $\acute{m}r = \frac{1}{n}\sum f(x\text{-}A)^{r}$

r – th raw moment: $\acute{m}r = \frac{1}{n}\sum fx^{r}$

r-th central moment: $mr' = \frac{1}{n}\sum f(x\text{-}\ \overline{X}\ )^{r}$

When N = ∑f.

Note that moments about mean are written without dashes (ˏ), but moments about any others origin, i.e. non-central moments, with dashes.

There are important relations between central and non-central moment.

For example:

If non-central moments ($m1', m2', m3', etc.$ about any arbitrarily origin A are known, the central moments can be obtained by using the relations.

$m_2 = m'2 - m'1^{2}$

$m_3 = m'_3\ \text{-}3m'_2m'_1\text{+}2m'_1{}^{3}$

$m4 = m'_4\text{-}4m'_3m'_1\text{+}6m'2m'1^{2}\text{-}3m'_1{}^{4}$

In particularly, using the first two moments $m'_1$and m'_2, about an arbitrary origin A, the mean and variance may be obtained:

$\bar{X}$ **=** m'_1**+A, and σ^2 =** $m'2$ **-** $m'1$

Moments for Ungroup Data.

If X$_1$, X$_2$, X$_N$ are N values, assumed by the variable the quantity

$$\bar{X} = \frac{X'_1 + X'_2 + \cdots + X'_1}{N} = \frac{\sum X'_1}{N}$$

Moments is defined as the rth Moments about Zero where: r = 1, 2,…. And the first moment with r = 1 is the arithmetic mean $(\bar{X})$

Example:
Find the: 1st (ii) 2nd (iii) 3rd (iv) 4th moment about zero for the numbers 9, 5, 8, 6, 4,

Solution:

(i) The first moment with r = q is arithmetic

$$\therefore \bar{X} = \frac{\sum X^2}{N} = \frac{9+5+8+6+7+}{5} = 7$$

(ii) The second moment is

$$\bar{X} = \frac{\sum X^2}{N} = \frac{9^2 + 5^2 + 8^2 + 6^2 + 7^2}{5} = \frac{255}{5} = 51$$

(iii) The third moment about zero is

$$\bar{X} = \frac{\sum X^3}{N} = \frac{9^3 + 5^3 + 8^3 + 6^3 + 7^3}{5} = \frac{729 + 125 + 512 + 216 + 343}{5} = \frac{1924}{5} = 384.8$$

(iv) The fourth moment about zero is

$$\bar{X} = \frac{\sum X^4}{N} = \frac{2^4 + 5^4 + 8^4}{3} = 1579$$

Moment about the mean $(\bar{X})$ is defined as:

$$\text{M} = \frac{\overset{N}{\underset{1+1}{\Sigma}} \ (X_1 - \bar{X})^2}{N} = \frac{\sum (X - \bar{X})^2}{N}$$

When r =1 then M$_1$ = 1 and if r = 2 then M$_2$ = S^2 the variance

Example:
Find the first four moment S about the mean$(\bar{X})$ for the numbers 2, 3, 8

Solution:
The Firs moment about the mean $(\bar{X})$ *is*

$$M_1 = \frac{\sum (X - \bar{X})^1}{N}$$

We know that $\bar{X}$ = 5

$\therefore M_1 = \frac{(2-5)^1+(5-5)^1+(8-5)^1}{3} = \frac{-3+0+3}{3} = 0\ Proved$

The 2nd moment of the mean is:

$$M_2 = \frac{\sum(X-\bar{X})^2}{N} = \frac{(2-5)^2+(5-5)^2+(8-5)^2}{3} = \frac{-3^2+3^2}{3} = \frac{18}{3} = 6, the\ Variance$$

The 3rd moment about the mean is:

$$M_3 = \frac{\sum(X-\bar{X})^3}{N} = \frac{-3^3+3^3}{3} = \frac{-27+27}{3} = 0 \therefore M_3 = 0$$

The fourth moment about the mean is:

$$M_4 = \frac{\sum(X-\bar{X})^4}{N} = \frac{-3^4+0^4+3^4}{3} = \frac{81+81}{3} = 54 \ \therefore\ M_4 = 54$$

5-3 Moment about any origin A:

The rth moment about any origin A is defined as:

$$M_4 = \frac{\sum(X-A)^r \sum d'}{N} = \frac{\sum d'}{N}$$

Where d=X-A i.e. Deviation from the assumed mean

Example:

Find the first for moment for the number (S) 2, 5, 8 about the origin A:

Solution:

Let A = 4 assume mean. Therefore,

The first moment of any origin A is

$$M_1' = \frac{\sum(X-4)^1}{N} = \frac{(2-4)+(5-4)+(8-4)}{3} = \frac{-2+1+4}{3} = 1$$

The 2nd moment of any origin A

$$M_2^1 = \frac{\sum(X-4)^2}{3} = \frac{-2^2+1^2+4^2}{3} = 7$$

The 3rd moment of any origin A

$$M_3^1 = \frac{\sum(X-4)^3}{3} = \frac{-2^3+1^3+4^3}{3} = \frac{-8+1+64}{3}$$

The 4th moment of any origin A

$$M_4^1 = \frac{\sum(X-4)^4}{3} = \frac{-2^4+1^4+4^4}{3} = \frac{16+1+256}{3} = 91$$

$$\therefore M_4^1 = 91$$

Relationship Between moment S (M_r and M_4^1)

The relation between M_r and M_2^1 are given below: -

1) $M_1 = 0$
2) $M_2 = M_2^1 - M_2^1$
3) $M_3 = M_3^1 - 3M_2^1 M_2^1 + 2M_2^1$
4) $M_4 = M_4^1 - 4M_2^1 M_3^1 M_4^1 + 6M_1^{1\ 2} M_2^1 \text{-} 3M_2^1$

Example:

Find $M_2, M_3\ and\ M_4\ for\ the\ numbers$ 2,5,8 $by\ using\ the\ about\ rlation$.

Solution:

First Calculate $M_2^1 (moments\ about\ origin)\ In\ our\ previous\ caluation$

$$M_1^1 = 1, M_2^1 = 7, M_3^1 = 19\ and\ M_1^1 = 91$$

$$\therefore M_2 = M_2^1 - M_1^{,2}$$

$$M_2 = 7 - (1)^2 = 7 - 1 = 6$$

$$M_3 = M_3^1 - 3M_1^1 M_2^1 + 2M_1'^1$$

$$M_3 = 19 - 3(1)(7) + (1)^3 = 19 - 21 + 2 = 0$$

$$\therefore M_2 = 0$$

$$M_4 = M_4^1 - 4M_2^1 M_3^1 + 6M_1^2 M_2^1 - 3M_1'^1$$

$$= 91 - 4(1)(19) + 6(1)^3(7) - 3(1)^4$$

$$91 - 76 + 42 - 3 \quad .\therefore M_4 = 54$$

2- Moments for grouped Data:

The statistical definition of moment for frequency distribution or grouped data is given by:

(a) Moment S about zero

$$\bar{X} = \frac{\sum f\ X'}{N}$$

(b) Moment of the mean

$$M_, = \frac{\sum f(X - \bar{X})'}{N}$$

(c) Moment of any origin A:

$$M_2^1 = \frac{\sum f(X - \bar{X})'}{N} = \frac{\sum fd}{N}$$

Note: If the class size is equal for all class, we use coding method as:

$$M_2^1 = \frac{C \sum fu}{N}$$

Example one:

Find the first four moments.

(a) About zero.

(b) About the mean and

(c) About any origin A. For the previous distribution on table 2.3

Solution:

x	f	Fx	fx^2	fx^3	fx^4
1.5	3	4.5	6.75	10.125	15.1875
4.5	6	27	121.5	446.75	2460.375
7.5	13	97.5	731.25	5484.4	41132.813
10.5	6	63	661.5	6945.7	72930.375
13.5	2	27	364.5	4920.75	66430.125
Total	30	219	1885.75	17907.75	182968.88

Therefore, the first moment about zero.

$$\bar{X} = \frac{\sum fx}{N} = \frac{219}{30} = 7.3$$

Second moment about zero

$$\bar{X} = \frac{\sum fx^2}{N} = \frac{1885.75}{N} = 6285$$

Third Moment about zero

$$\bar{X}^3 = \frac{\sum fx^3}{N} = \frac{17997.75}{30} = 596.75$$

Fourth moment about zero

$$\bar{X}^4 = \frac{\sum fx^4}{N} = \frac{182968.88}{30} = 6098.9625$$

(b)Moment about mean.

X	f	fx	f(x-$\bar{X}$)	$(x-7.3)^2$	$f(x-7.3)^3$	$f(x-7.3)^4$
1.5	3	4.5	-17.4	100.92	-5853	3394.9
4.5	6	27	-16.8	47.04	-131.7	368.8
7.5	13	97.5	2.6	0.52	0.104	0.021
10.5	6	63	19.2	61.44	196.608	629.15
13.5	2	27	12.4	76.88	38.44	2955.3
Total	30	219	0	286.8	-481.848	7348.2

$$\bar{X} = \frac{\sum fx}{\sum f} = \frac{219}{30} 7.3$$

Therefore,

$$M_1 = \frac{0}{30} = 0$$

$$M_2 = \frac{286.8}{30} = 9.56$$

$$M_3 = \frac{-481.848}{30} = -16.08$$

$$M_4 = \frac{7348.2}{30} = 2449$$

(c) **Moment of any origin A**

x	f	f(x-7.5)	$f(x-7.5)^2$	f$(x-7.5)^3$	$f(x-7.5)^4$
1.5	3	-18	108	-648	3888
4.5	6	-18	54	-162	486
A ⟶ $\boxed{7.5}$	13	0	0	0	0
10.5	6	18	54	162	486
13.5	2	12	72	432	2592
Total	30	-6	288	-216	7452

Let A = 7.5

$$M_1' = \frac{-6}{30} = -02$$

$$M_2' = \frac{288}{30} = 9.6$$

$$M_3' = \frac{-216}{30} = -7.2$$

$$M_4' = \frac{7452}{30} = 248.5$$

Example Two:

Find (a) $M_{1,}'\ \ M_{2,}'\ M_3'\ M_4^1$ and

$$(b) M_1, M_2, M_3, and\ M_4\ for\ the\ following\ distribution$$

Table 5-1: Distribution of 53 fourth year students’ college of education U. of Juba.

Solution:

Class	No. of Students	X	d	U = d/c	fu	fu^2	fu^3	fu^4
4-6	11	5	6	-2	-22	14	-8.5	
7-9	12	8	3	-1	-12	12	-12	12
10-12	13	Let 11	9	0	0	0	0	4
13-15	14	14	3	1	14	14	14	14
16-18	03	17	6	2	6	12	24	45
Total	**53**				**-14**	**82**	**-62**	**25**

Since the class size is equal use coding method.

$\therefore$(a) $M_1' = \frac{C^1 \sum fu^1}{N} = \frac{3(-14)}{53}$ = -0.79

$$M_2' = \frac{C^2 \sum fu^2}{N} = \frac{9(82)}{53} = 139$$

$$M_3' = \frac{C^3 \sum fu^3}{N} = \frac{27(-62)}{53} = -31.58$$

$$M_4' = \frac{C^4 \sum fu^4}{N} = \frac{81(250)}{53} = 382.1$$

(b) By using the relationship between $M_1$and M'_1

$M_1 = 0$

$M_2 = M'_2 - M'^2 = 139 - (-079)^2 = 13.28$

$M_3 = M'_1 - 3M'_1M'_2 + 2M'_1$

$\therefore M_3 = -31.58 -3(-0.79)^2(13.9) +2(-0.79)^3 = 0.37$

$M_4 = M'_1 - 4M'_1M'_1 + 6M'_1M'_1-3M'_1 = 328.1-4(-0.79) (-3158) +6(-0.76)^2(139)-3(-0.79)^4$

$\therefore$ M4 = $\underline{\underline{33.2}}$

Moments in Dimension-less form:

To avoid using particular units, we define moments in dimension less form about the mean an:

$$a_r = \frac{M_r}{S^r} = \frac{M_r}{\left(\sqrt{M_2}\right)^r} = \frac{M_r}{\sqrt{M_2^r}}$$

Remarks:

In practice, the first four moments usually provide an ad3quate description a frequency distribution. Higher moments are rarely computed in practical problem.

The 1St moment estimates the average value.

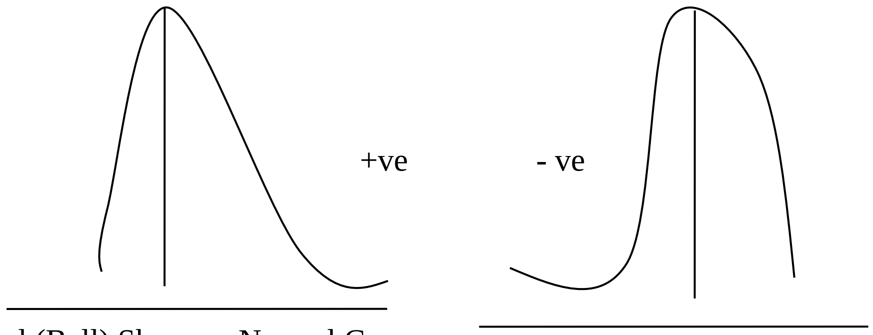
The 2nd moment measures the dispersion of the observation.

The 3rd moment evaluates the asymmetry of a distribution and

The 4th moment measures it relative height.

Measures of skewness:

Skewness is the degree of asymmetry or departure from symmetry of a distribution. A distribution may be symmetrical or a symmetrical, in which case it is skewed. Asymmetrical distributions are skewed either to the right (+vely) or to the left (-vely). i.e.



Symmetrical (Bell) Shape or Normal Curve

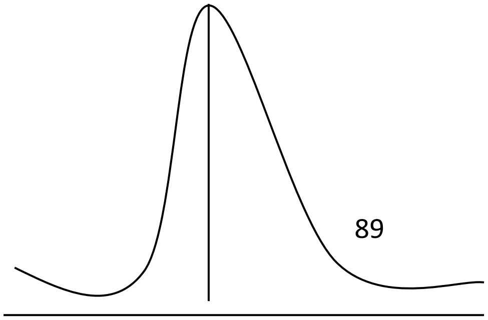


Illustration:

A right-skewed distribution is usually characterized by the fact that its longer tail is on, the right-hand side; most of the observation is then dispersed to the right of the mode. Similarly, a left –skewed distribution usually has its longer tail on the left-hand side, to the mode. Several measures are employing, but the two mains once are:

1- **<u>The third –moment measure (An important measure use)</u>**

This is given by:

$$a_3 = \frac{M_3}{\sigma^3} \ or \ \frac{M_3}{S^4} = \frac{M_3}{\sqrt{M_2^3}}$$

<u>Note: This measure is also called the Moment Coefficient of Skewness</u>

If $a_3 = 0$ the distribution is symmetrical (normal)

If a_3 is positive, the distribution is positively skewed.

And if a3 is negative, the distribution is negatively skewed.

2- **<u>The Pearsonian Measure of Skewness: -</u>**

In practice, the difference between the Mean and the Mode is used to measure skewness. The formula is.

$$\text{Skewness} = \frac{Mean - Mode}{S\tan dard\ deviation} = \frac{\bar{X} - Mode}{\sigma}$$

A second expression is that of:

$$\frac{3(Mean - Median)}{Std.Deviation}$$

We employ this empirical formula to avoid the use of Mode.

Note: -

Like the third moment measure of skewnes, this measure is+ve for a right skewed distribution –ve for a left –skewed distribution and zero to a symmetrical distribution.

<u>Example: -</u>

For a certain distribution, it was found that mean is 15, the mode is 15 and the std. Deviation 13

(i) Calculation the coefficient of skeweness.

(ii) Show what type of skewness it is

<u>Solution:</u>

(i) Using the Pearsonian Measure of Skewness,

$$\text{Skewness} \ = \frac{\bar{X} - Mode}{\sigma} = \frac{15 - 15}{3} = 0$$

(i) Since the coefficient of skewness is zero, therefore, the distribution is symmetrical (Normal).

<u>Remark</u>

As a general rule, a distribution is NOT considered to be markedly skewed as long as the pearsonian formula yields an absolute value less than one.

Example:

For the distribution of table 2-3, it is known that $M_2 = 9.56$ and $M_3 = -16.08$ find the moment coefficient of skeweness a_3

Solution:

a$_1\frac{M_1}{\sigma^1}\ But\ M_2 = \sigma^2\ or\ S^2\ i.e. variance. therefore, \sqrt{M_1}\ equals\ \sigma\ or\ S$

Hence

a$_3 = \frac{-16.08}{\sqrt{(9.56)^3}} = \frac{-16.08}{29.559}$

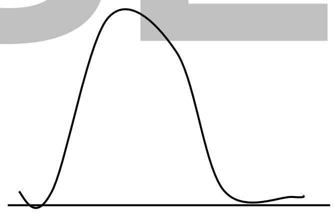
a$_3 = 0.5441$

The result confirms our suspicion of a –ve skewness in this distribution. If the distribution were not skewed all, a$_3$ would be zero. In general, distributions are not considered to be much skewed unless the absolute value of a$_3$ is at least 2.

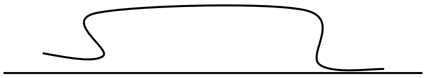
5-5 Measure of Kurtosis: -

Kurtosis is a Greek work referring to the relative height of a distribution i.e. its peakness. Thus is can be defined as the degree of peakness of a distribution and usually takes relatives to normal distribution. A didtribution is said to be Mesokurtic, if it has so-called "Normal" Kurtosis, Platykutic if its peak is abnormally fiat and Leptokurtic if its peak is abnormally high. i.e.

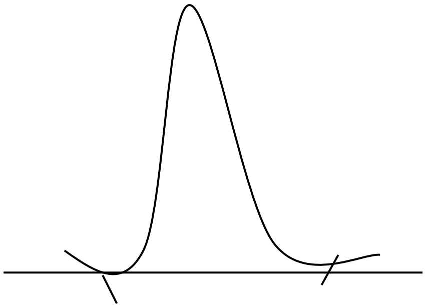
Mesokurtic



Platykurtic



Leptokurtic



Kurtosis is measured by moment coefficient of Kurtosis.

a$_4 = \frac{M_4}{\sigma^4}$

but $\sigma = \sqrt{M_2}$

Therefore, $a_4 = \frac{M_4}{\sqrt{(M_2)^4}}$

Remark:

(a) For a normal distribution (Mesokurtic)

$$a_4 = 3$$

(b) The more Platykurtic a distribution a_4 decrease below 3.

(c) The more Letokurtic a distribution, $a_4$excced 3
Alternative, if
$a_{4-3=0,}$ The distribution is Mesokurtic.
$a_4 - 3$= negative, the distribution is platy kurtic
$a_4 - 3$ = positive, the distribution is leptokurtic.

Example:

For a certain distribution, it was found that $M_2 = 8.5275$

$M_4 = 199.3759$ Calculate the moment coefficient of Kurtosis and show what type of Kurtosis it is.

Solution: -

$$a_4 = \frac{M_4}{\sigma^4} = \frac{M_4}{M_2^1} = \frac{199.3759}{(8.5275)^2}$$

Therefore, $a_4 = \underline{\underline{2.74}}$

Since a_4 decrease below 3, the distribution is more Platykurtic. Similarly

As $a_4 = 2.74 - 3 = 0.26, \mathit{the\ distribution\ is\ Platykurtic}.$

CHAPTER SIX
MULTIVARIATE ANALYSIS
(ANALYSIS OF VARIANCE)

The subject of multivariate analysis deals with the statistical analysis of the data collected on more than one (response) variable. These variables may be correlated with each other, and their statistical dependence is often considered when analyzing such data. In fact, this consideration of statistical dependence makes **multivariate analysis** somewhat different in approach and considerably more complex than the corresponding Univariate analysis, when there is only one response variable under consideration. Response variables under consideration are often described as random variables and since their dependence is one of the things to be accounted for in the analyses, these response variables are often described by their joint probability distribution. This consideration makes the modeling issue relatively manageable and provides a convenient framework for scientific analysis of the data. Multivariate normal distribution is one of the most frequently made distributional assumptions for the analysis of multivariate data. However, if possible, any such consideration should ideally be dictated by the particular context. Also, in many cases, such as when the data are collected on nominal or ordinal scales, multivariate normality may not be an appropriate or even viable assumption. In the real world, most data collection schemes or designed experiments will result in multivariate data. A few examples of such situations are given below.

1. During a survey of households, several measurements on each household are taken. These measurements, being taken on the same household, will be dependent. For example, the education level of the head of the household and the annual income of the family are related.
2. During a production process, several different measurements such as the tensile strength, brittleness, diameter, etc. are taken on the same unit. Collectively such data are viewed as multivariate data.
3. On a sample of 100 cars, various measurements such as the average gas mileage, number of major repairs, noise level, etc. are taken. Also, each car is followed for the first 50,000miles and these measurements are taken after every 10,000 miles. Measurements taken on the same car at the same mileage and those taken at different mileage are going

to be correlated. In fact, these data represent a very complex multivariate analysis problem.

4. An engineer wishes to set up a control chart to identify the instances when the production process may have gone out of control. Since an out-of-control process may produce an excessively large number of out of specification items, detection at an early stage is important. In order to do so, she may wish to monitor several process characteristics on the same units. However, since these characteristics are functions of process parameters (conditions), they are likely to be correlated leading to a set of multivariate data. Thus, many times, it is appropriate to set up a single (or only a few) multivariate control chart(s) to detect the occurrence of any out-of-control conditions. On the other hand, if several univariate control charts are separately set up and individually monitored, one may witness too many false alarms, which is clearly an undesirable situation.
5. A new drug is to be compared with a control for its effectiveness. Two different groups of patients are assigned to each of the two treatments, and they are observed weekly for next two months. The periodic measurements on the same patient will exhibit dependence and thus the basic problem is multivariate in nature. Additionally, if the measurements on various possible side-effects of the drugs are also considered, the subsequent analysis will have to be done under several carefully chosen models.
6. In a designed experiment conducted in a research and development center, various factors are set up at desired levels and a number of response variables are measured for each of these treatment combinations. The problem is to find a combination of the levels of these factors where all the responses are at their ‘optimum’. Since a treatment combination which optimizes one response variable may not result in the optimum for the other response variable, one has a problem of conflicting objectives especially when the problem is treated as collection of several univariate optimization problems. Due to dependence among responses, it may be more meaningful to analyze response variables simultaneously.
7. In many situations, it is more economical to collect a large number of measurements on the same unit but such measurements are made only on a few units. Such a situation is quite common in many remote sensing data collection plans. Obviously, it is practically impossible to collectively interpret hundreds of univariate analyses to come up with some

> definite conclusions. A better approach may be that of data reduction by using some meaningful approach. One may eliminate some of the variables which are deemed redundant in the presence of others. Better yet, one may eliminate some of the linear combinations of all variables which contain little or no information and then concentrate only on a few important ones. Which linear combinations of the variables should be retained can be decided using certain multivariate methods such as principal component analysis. Such methods are not discussed in this book, however.

The term "**multivariate statistics**" is appropriately used to include all statistics where there are more than two variables simultaneously analyzed. You are already familiar with bivariate statistics such as the Pearson product moment correlation coefficient and the independent groups *t*-test. A one-way ANOVA with 3 or more treatment groups might also be considered a bivariate design, since there are two variables: one independent variable and one dependent variable. Statistically, one could consider the one-way ANOVA as either a bivariate curvilinear regression or as a multiple regression with the K level categorical independent variable dummy coded into K-1 dichotomous variables.

Independent and Dependent Variables.

We shall generally continue to make use of the terms "independent variable" and "dependent variable," but shall find the distinction between the two somewhat blurred in multivariate designs, especially those observational rather than experimental in nature. Classically, the independent variable is that which is manipulated by the researcher. With such control, accompanied by control of extraneous variables through means such as random assignment of subjects to the conditions, one may interpret the correlation between the dependent variable and the independent variable as resulting from a cause-effect relationship from independent (cause) to dependent (effect) variable. Whether the data were collected by experimental or observational means is NOT a consideration in the choice of an analytic tool. Data from an experimental design can be analyzed with either an ANOVA or a regression analysis (the former being a special case of the latter) and the results interpreted as representing a cause-effect relationship regardless of which statistic was employed. Likewise, observational data may be analyzed with either an ANOVA or a regression analysis, and the results cannot be unambiguously interpreted with respect to causal relationship in either case. We may sometimes find it more reasonable to refer to “independent variables” as “predictors”, and “dependent variables” as “response-,” “outcome-,” or “criterion-variables.” For example, we may use SAT scores and high school GPA as predictor variables when predicting college GPA, even though we wouldn’t want to say that SAT causes college GPA. In general, the independent variable is that which one considers the causal variable, the prior variable (temporally prior or just theoretically prior), or the variable on which one has data from which to make predictions.

Descriptive and Inferential Statistics.

While psychologists generally think of multivariate statistics in terms of making inferences from a sample to the population from which that sample was randomly or representatively drawn, sometimes it maybe more reasonable to consider the data that one has as the entire population of interest. In this case, one may employ multivariate descriptive statistics (for example, a multiple regression to see how well a linear model fits the data) without worrying about any of the assumptions (such as homoscedasticity and normality of conditionals or residuals) associated with inferential statistics. That is, multivariate statistics, such as can be used as descriptive statistics. In any case, psychologists rarely ever randomly sample from some population specified a priori, but often take a sample of convenience and then generalize the results to some abstract population from which the sample could have been randomly drawn.

Rank-Data

I have mentioned the assumption of normality common to “parametric” inferential statistics. Please

note that ordinal data may be normally distributed and interval data may not, so scale of measurement is irrelevant. Rank-ordinal data will, however, be non-normally distributed (rectangular), so one might be concerned about the robustness of a statistic’s normality assumption with rectangular data. Although this is controversial issue, I am moderately comfortable with rank data when there are twenty to thirty or more ranks in the sample (or in each group within the total sample).

Multi-variate Analysis or Analysis of Variance or (ANOVA ANANLYSIS).

The analysis of variance has been defined as statistical technique for the “Separation of variable due to a group of causes from the variation due to other group”. Here we shall discuss the simplest use of this technique, namely testing whether the mean of number of populations are equal. The method is based upon an unused result that the equality of several population means can be tested by comparing the sample variable using F-distribution. It may be recalled that the t-statistical is used for testing whether two population means are equal.

The analysis of variance test may, therefore, be taken as extension of this test for the case of more than two means.

Sources of Variation

There are different sources of variation considering the examples which are following below:

As the Crown Motor-Cars Manufacture in South Sudan is producing three (3) types of the cars. We are requiring testing on the basis of these sample observations whether the mean of their populations is equal or not.

It will be noticed that there are differences in the observations "within" each sample or group. The variability occurs due to change, and it reflects the noticed differences that occur with each of the population from each of the samples have been taken. If we examine the mean of three samples, we see that another kind of variability is present. This variability is due to the combined effect of the natural differences and possible real differences "between" the difference's groups.

Our problem is to decide whether the differences among sample means is only due to change or whether the differences occur between the means of three populations (from which the sample have been drawn) are different. Hence, we now concerned with the problem to testing the hypothesis that the means of several population are equal rejection of the hypothesis will lead to be conclusion that at least one of the population means is different from others.

Necessary Assumption in the Analysis of Variances.

The following are some necessary assumptions in the analysis of variances or ANOVA analysis as follow:

1. The samples are independently drawn.
2. The populations are normally distributed, with a common variable.

3. The effects of various components are additive.

If the means of all the populations are equal, then the variability "between groups" would result only from chance and hence would be the same as the variability arising from "within groups". On the other hand, if the population means are not equal, the variability "*between groups*" would more than the variability "*within groups*". The measure of variability used in the analysis of the variance is called **"Mean Square".**

This is like a variance and is defined by:

Mean square $= \frac{Sum\ of\ Squared\ deviation\ from\ mean}{Degree\ of\ freedom}$

Note that in the t-test (14. 8. 6) for specified mean, the population variance σ^2 as estimated as the sum of squared deviation from mean divided by sample size minus one.

In t-test (14.8.9) for equality of two means, the common population variance was estimated as sum of the squared deviations of two groups of observations from the representative mean divided by the sum of sample size minus two.

These divisors are referred to as “degree of freedom”. The mean square appearing here are similar to the estimate variance cited above, but they relate to different sources of variation, and are based on different degrees of freedom. One mean square is used to measure variability “within the groups”. This based on the sum of square deviation of the observations within each group, the deviation being taken from the respective group means, and has degree of freedom “total sample size minus number of samples”. The sum of square within groups as it is called, when divided by the number of degrees of freedom provides the “mean square within group”. This mean square represents a measure of variability due to chance of experimental error.

The other mean square is used to measure group effect or possible differences existing are any, between the groups. This based on the sum of squared deviations of the individual sample or group means the deviation being taken from the grand mean of all observations being considered as one sample and has the degree of freedoms “***number of samples minus one***”. The sample means are weighted by the representative sample size. This sum of squares “between groups” when divided by these degrees of freedom provides the “mean square between groups”. If the means of all populations are equal, there is no group –effect and the mean square between groups will also represent variability due to chance alone.

Consequently, when the groups mean in the population are equal the mean square within groups and mean square between groups should not be much different, and their ratio should be close to one. Unusually large ratios would indicate that the groups’ means are not equal in the population.

Technique in one-way Analysis of Variance or

One-way ANOVA Analysis.

We have k independent random sample (or groups of observations), one group from each of k, Normal populations with the means μ_1, μ_2, ……..μ_k and a common variance σ^2 (unknown) once the basis of the data, it is required to test the "**null hypothesis"** that the population means are equal.

$$Ho\ (\mu_1 = \mu_2 = \ldots\ldots\ldots\ldots\mu_k)$$

Against the alternative hypothesis H_1 all μ_1 are not equal.

$$Xij = \mu_i + eij$$

Where xij denotes the *j-th* observations in the *i-th* groups, μ_i denoted mean of the *i-th* population and eij denotes error due to many unspecified causes.

This is called a linear model, since *xij* is assumed to be made up of this 'sum' of effects due to different components. It is assumed that eij are independent and individual distribution and distributed normal variates with mean zero (0) and variance σ2. This model may also be written in the form.

$$Xij = \mu + \alpha_i + eij.$$

Where μ denotes the general effect.

α_i denotes effects specifically to *i-th* population.

eij denotes error component.

The null hypothesis is equivalent to stating that there is not specifically or special effect due to any population.

$H_0 = (\alpha_1 = \alpha_2 = \ldots\ldots\ldots\ldots\alpha_k = 0)$.

Against the alternative hypothesis H1 (***all*** $\boldsymbol{\alpha_i}$ ***are not zero***).

Let introduce the following notating to you.

k = number of independent sample or groups.

ni = number of observations in the *i-th* sample.

N = total number of observations in all the samples = $\sum ni$.

Xij = the *j-th* observations in the *i-th* group.

Ti = Total of the *i-th* sample = ∑xij.

T = Grand total of all observations = ∑Ti ∑∑xij.

$\overline{X}$ = mean of the *i-th* sample = $\frac{Ti}{ni}$

$\overline{X}$ = Grand mean of all observations = $\frac{T}{N}$

It can be shown algebraically that:

$\sum_i\sum_j(xij\text{-}\overline{X}\ )^2\sum_i\sum_j(xij\text{-}\overline{X}\ )^2+\sum ni(\overline{X}i - \overline{X})^2$

Total SS = SSW + SSB.

The expression of the above formula is that, sum of squares (SS) of all sample observations about the grand mean and is called "***Total Sum of Square***" Total SS. The first term on the right is the sum of square of deviations of observations within each group about the respective group mean and it is called sum squares within groups, (SSW) or sum of square due to error (SSE).

The second term is the (weighted) Sum of Squares of group means about the grand mean and is called (***Sum of Squares Between Groups),*** (SSB).

The degrees of freedom can similarly be split up thus, we have:

Sum of Square: Total SS = SSW + SSB.

Degrees of Freedom: N-1 = (N-K) + (k-1).

The mean squares are now calculated no dividing the sum of square by corresponding degrees of freedom.
Mean Square within the groups (MSW) = SSW/ (N-k)
Mean Square between the groups (MSB) = SSB/ (K-1)
Assuming normal distribution in the population with a common variance (σ^2), it can be shown that the expected value of MSW is σ^2,

E (MSW) σ^2

Irrespective of whether the population mean is equal or not, but E(MSB) $\geq \sigma^2$

Where the sign of equality holds only when the population means are equal. Thus, when the null hypothesis H_0 is true, both the mean square MSB and MSW provide independent unbiased estimate of the same population variance σ^2

Hence, we have the test the statistics.

$F = \frac{MSB}{MSW}$

Which under H_0 follows F distribution with degrees of freedom (***K-1, N-k***). If the observations values of the statistics equal or exceeds the theoretical ***F values*** from statistical Tables at a specified level of significance[14], *we reject the null hypothesis and conclude that all the population mean are not equal.*

On other hand, if the observed values of statistics are less than theoretical value all the population means may be equal.[15]

The various sum of squares, degrees of freedom, means squares corresponding to the different source of variations, and the F values (both observed and theoretical) are shown in the form of a table known as ***"Analysis of Variance Table"*** as in the following example.

Steps in computation one way Analysis of Variance

(One Way ANOVA Analysis).

1. Reduce the sample observations by subtracting a suitable content.
2. From these reduced data obtain the following,

i. Total (Ti) for each group.

ii. Grand total *(T=∑Ti) i, e*, sum of group totals *Ti*,

iii. Total of the square of all figures ($\sum\sum xij^2$).

3. Calculate:

i. Correction Factor (CF) = $\frac{\text{T2}}{\text{N}}$.

ii. Total SS = $\sum\sum xij^2$- CF.

iii. SSB = $\sum\{\frac{\text{T}i2}{\text{n}i}\}$-CF.

iv. SSW = Total SS- SSB.

4. Write down the degrees of freedom (D.F).

   i. D.F for SSB = {K-1}.

   ii. D.F for SSW = {N-k}.

5. Calculate the mean square:

i. MSB = SSB/(k-1)

ii. MSW = SSW/(N-k)

6. Obtain the observed or calculated values value of F and dividing MSB by MSW.
   To find the calculate value.
   F = MSB/MSW
7. Consult the F-tables to find the theoretical value of F at (say) 5% level corresponding to the degrees of freedom {k-1,N-k}
8. If the observed of calculated values of F at {6} equal of exceeded the theoretical or table value of F {7}, we reject the null hypothesis and conclude that all the population mean are not equal.
   Otherwise, they may be taken to be equal.

[14] The theoretical F values are the values from found from F-Tables, t-Tables and Z- Tables.
[15] The observed values are the results of the calculation you have calculated from the data you have them.

Locating Unequal Parts of Means.

In the analysis of variance or ANOVA analysis, if the observed or calculated value of F is found to be significant, the calculated is that all the population means are not equal, this however does not rule out the possibility that some of them are equal, while other are different. This in such a case we may be interested to find which pair of means differ. This is real problem of {test for equality of two population means}.

Using t-distributions, if the number of observations in each sample is the same, say n, then the appropriate statistic for testing equality of means of μ_i and μ_j is:

$$t = \frac{\bar{X}i-\bar{X}j}{s\sqrt{\frac{1}{n}+\frac{1}{n}}} = \frac{\bar{X}i-\bar{X}j}{s\sqrt{\frac{2}{n}}}$$

Where, σ2 is the unbiased estimator of the common (but unknown) population variance σ^2.[16]

In the analysis of variance test {ANOVA}, σ^2 is estimated by the Means Square Error (MSE), i.e MSW I one-way classified data. $S^2 = MSE$

The population means μ_i and μ_j will be considered to be equal, at say 5% level of significance, if $|t| \leq t\ 0.025$, where t 0.025 denoted the upper 2.5% value of t-distribution with NM-k degrees of freedom.

$$\left|\frac{\bar{X}i-\bar{X}j}{s\sqrt{\frac{2}{n}}}\right| \leq t\ 0.025$$

$$\bar{X}i - \bar{X}\text{j} \leq s\sqrt{\frac{2}{n}}\ t\ 0.025.$$

Multiple both sides by n,

$$|T_i - T_j| \leq \sqrt[s]{2n}\ t\ 0.025$$

Where T_i and T_j represent the totals of the *i-th* and *j-th* sample respectively.

Therefore, in cases when the entire sample is of the same size (n), we calculate the "critical Difference" between sample totals.

Critical Difference (C.D) = $\sqrt[s]{2n}$ t 0.025

The difference between all pairs of sample $|T_i - T_j|$ are now compared with C.D. if the difference between any particular pair, say for samples 1 and 3, $|T_i - T_j|$, exceed the C.D, then, the population means µ1 and µ3 are considered to be variance in one-way ANOVA analysis with equal number of observations in each sample, which exams the theories discussed so for.

[16] Note that, this is the Degree of Freedom {d.f} of Error SS I,e the divisor used in MSE.

Example: (1)

A random sample of five motor-car types was taken from the Crown Auto to test each of three (3) brands manufactured by three companies. The lifetime of these types (as measured by the mileage run) is shown in below. One the basis of the data test whether the average lifetime of the 3 brands of types are equal or not.

Lifetime of Tyres Miles			
	A	**B**	**C**
Brands	35	32	34
	34	32	33
	34	31	32
	33	28	32
	34	29	33

Also test which pairs of tyres differ in quality.

Solution.
Reduce from each observation by 30.
Table calculation for analysis of variance. ANOVA.

Samples	1	2	3	
	5	2	4	
	4	2	3	
	4	1	2	
	3	-2	2	
	4	-1	3	
Total:	$T_1 = 20$	$T_2 = 2$	$T_3 = 14$	T = 36
Total of squares:	82	14	42	$\sum\sum xij^2 = 138$
Sample Size:	$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	N = 15

Correction Factor (CF) = $\frac{T^2}{N} = \frac{362}{15} = \frac{1296}{15} = \underline{\underline{86.4}}$

Total SS = $\sum\sum xij^2$- CF = 138 – 86.4 = $\underline{\underline{51.6}}$

SSB = $\sum \{\frac{Ti2}{ni}\}$-CF = $\{\frac{20,2}{5}+\frac{2,2}{5}+\frac{14,2}{5}\}$-86.4

= $(\frac{400+40+196}{5})$ - 86.4 = 120-86.4 = $\underline{\underline{33.6}}$

SSE = Total SS-SSB = 51.6-86.4 = 18.0.

Analysis of Variance Table. ANOVA Table.

Source of Variance	SS	d.f	m.s	Observed or calculate F	Tabulate F value
Between groups	33.6	2	16.8		F 0.5 = 3.89
Within groups (Error)	18.0	12	1.5	11.2	F 0.1 = 6.93
Total	51.6	14	-		-

Note:

1. Since the total number of observations in all samples is in the d.f for Total SS is 15 – 1 = 14.

   Since there are three groups (3) of observations, the d.f for SEE is the difference 14–2 =12.
2. The observed value of F is MSB/MSE = 16.8/1.5 = 11.2.
3. The theoretical value of tabulated value of F are obtained from statistical tables corresponding to d.f. (2, 12). It may be noticed that the first d.f is the divisor used in numerator and the second is the divisor used in the denominator for calculation of F.

Since the observed value of F is 11.2 exceed the 1% tabulated value is 6.93, we reject the null hypothesis of equality of population means and conclude that the average lifetime of (3) brands of Types are not equal. To test which brands of types differ in quality, we calculate the critical difference.

Here s = $\sqrt[s]{1.5}$, n = 5, t = 0.025 = 2.18 (for 12 d.f)

Critical Difference = $\sqrt{1.5} \times \sqrt{2} \times 5 \times 2.18 = \sqrt{15} \times 2.18 = 3.87 \times 2.18 = 8.44$

The sample totals (of the reduced observations) are $T_1 = 20$, $T_2 = 2$, $T_3 = 14$.

We have,

$|T_1 - T_2| = 18$

$|T_1 - T_3| = 6$

$|T_1 - T_3| = 12$

Comparing these figures with the critical difference 8.44, we find that brands A and B are different quality, and so are brands B and C. Brands A and C may be taken to be the same quality.

Example. (2)

A researcher from the Upper Nile University, Faculty of Agriculture, wished to study the effect of four (4) Fertilizers on the Yied of Crop. He divided the field into (24) plots and assigned each Fertilizer at random to (6) plots. Parts of his calculations are shown below:

Source of Variance	SS	d.f	m.s	Observed or calculate F	Tabulate F value F5%
Between groups (Fertilizers)	2940				F 0.5 = 3.10
Within groups (Error)					
Total	6212		-		-

a. Complete the above table by filling in the values marked.

b. Test at 5% level to see whether the Fertilizers differ significantly.

Solution.

Working Note:

a. i. Since 24 observations are available, one from each plot.

Total degrees of freedom (d.f) is (24 – 1) = 23.

Again, since 4 Fertilizers are being compared, thedf for Fertilizers is (4 – 1) = 3. Hence, the df for ‘within Group’ is (23 – 3) = 20.

ii. Sums of squares (SS) are additive. Therefore

Within Group SS = Total SS- Fertilizer SS = 6212-2940 = 3272

4. Mean squares (MS) are obtained on dividing SS by df:

MS between Fertilizers = 2940/3 = 980

MS within Group = 3272/20 = 163.6

$$F = \frac{\text{MS between Fertilizer}}{\text{MS within Group}} = \frac{980}{163.6} = 5.99$$

The completed table is as follows:

Table: Analysis of Variance Table.

Source	df	SS	MS	F	F5%
Between Group	3	2940	980	5.99	3.10
Within Group	20	3272	163.6		

Total	23	6212			

(d) Since the observed value of F is 5.99, that are larger than the 5% tabulated value is 3.10, it is significant at 5% level. We therefore, conclude that the fertilizers differ significantly.

Example. (3)

Each of the following sets observations is a random sample from a normal population.

Sets	Observations
I	249, 242, 247, 250, 252
II	251, 256, 255, 258
III	266, 261, 265, 264
IV	262, 260, 263, 262, 261, 264, 262

Test whether the population means are equal (assume that the population standard deviations are the same).

Solution

The observations, are reduce by 250, and also the preliminary results for calculating the various sum of squares, are shown in the following table.

	Sets				
	I	II	III	IV	
	-1	1	16	12	
	-8	6	11	10	
	-3	5	15	13	
	2		11		
	14				
	12				
Total	$T_1 = -10$,	$T_2 = 20$,	$T_3 = 56$,	$T_4 = 84$	Total = 150
Total of Sqs.	78,	126,	798,	1018	$\sum\sum xij^2 = 2020$
Sample Size	$n_1 = 5$,	$n_2 = 4$,	$n_3 = 4$,	$n_4 = 7$	N = 20

Note:

(i) This is a problem of analysis of variance in one-way classified data with unequal numbers of observations in the samples.

(ii) Degrees of freedom under different sources of variation and also the F-values are obtained in the same way as indicated in the previous (1).

C.F = $\frac{150^2}{20}$ = 1125

Total SSW = 2020 – 1125 = 895

SSB = $\frac{(-10)^2}{5}[+\frac{20^2}{4}\frac{56^2}{4}+\frac{84^2}{7}]$ – 1125 = (20 + 100 + 784 + 1008) – 1125 787

SSE = 895 – 787 = 108

The various sums of squares (S.S) along with the degree of freedom (d.f) are shown in the following table.

				F values	
Source of Variation	S.S	d.f	M.S	Calculated	Tabulated
Between sets Within sets (Error)	787 108	3 16	262.3 6.75	38.9	F %5 = 3.24 F %1 = 5.29
Total	895	19	-	-	-

Since the observed value is of F is larger than 1% tabulated value corresponding to d.f (3.16), we rejected the null hypothesis and conclude that, the means of normal population are not equal.

Technique in Two-Way of ANOV Analysis.

We have a random sample consisting of hk observation, which are classified according to two factors into h classes according to factor A, and into k classes according to factor B. There is one observation in each of line hk cells corresponding to a class of factor A, and simultaneously a class of factor B. These hk observations may therefore be arranged in form of two-way table with h rows and k columns.

The mathematical model is that any observation is made up of the sum 4 components:

$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$

where: μ is the general effect.

α_i is the effect special to the i-th class according to factor A,

β_i is the effect special to the j-th class according to factor B,

e_{ij} is the error component.

This is called a Linear Model, since any observation x_{ij} is supposed to be made up of the sum of effect of the various components. It is assumed that e_{ij} are independently distributed normal variables each with mean (0) and variance σ^2 (unknown).

In the two-ways of classified data or ANOVA analysis, the analysis of variance maybe used to decide on two types of problems simultaneously as following.

i. Whether there is any differential effect due to classification by factor A, Null hypotheses:
H0 ($\alpha_1 = \alpha_2 = \ldots\ldots.\alpha n = 0$).

ii. Whether there is any differential effect due to classification by factor B, Alternative hypotheses:
H1 ($\beta_1 = \beta_2 = \ldots\ldots\ldots\beta n = 0$)
We introduce the following notations:
h = Number of classes by factor A, (we can say numbers of rows).
k = Number of classes by factor B, (we can say numbers of columns).
N = hk = Total number of observations.
T_i = Total of the i-th row (i = 1, 2, …..,h).
T_j = Total of the j-th column (j = 1, 2, …….,k).
T = Grand total of all observations = $\sum Ti = \sum T'j = \sum\sum xij$
$\bar{X} = \frac{Ti}{k}$ = Mean of the i-th row.
$\bar{X}j = \frac{Tj}{h}$ = Mean of the j-th column.
$\bar{X} = \frac{T}{N}$ = Grand mean of all observations.
Algebraically, it can be shown that,
$\sum\sum(xij\text{-}\bar{X})^2 = k\sum(\bar{X}i - \bar{X})^2 + h\sum(\bar{X}i - \bar{X}) + \sum\sum(xij\text{-}\bar{X}i - \bar{X}j + \bar{X})^2$
i.e, Total sum of square.
= (SS between factor A) + (SS between B) + (SS due to Error).
or Total SS+SSB+SSE.
The degree of freedom (d.f) for the various SS can be also spilt up.
hk -1 = (h-1)+(k-1)+(h-1)(k-1).
[Note: it will be easier to calculate the d.f as follows:
Since there are hk observations in all the d.f for Total SS is one less, as (h-1); since there are h are classes according to factor A, the d.f for SSA if one less, as (h-1); similarly the d.f for SSB is one less, as (k-1). Again just as we have SSE = Total SS – SSA – SSB, similarly,
(d.f for SSE) = (d.f for Total SS) – (d.f for SSA) – (d.f for SSB).
= (hk-1) – (h-1) – (k-1) = (h-1) (k-1)]

Mean Square are now calculated as before, on dividing the sums of squares by the corresponding degrees of freedom.

Mean Square between classes by factor A(MSA) = SSA/(h-1).

Mean Square between classes by factor B(MSB) = SSB/(k-1).

Mean Square Error (MSE) = SSE/(h-1) (k-1).

Assuming normal distribution in the population with a constant variance σ^2, it can be shown that the expected value of MSE is σ^2.

E(MSE) = σ^2.

Irrespective of whether the null hypotheses H_{01} and H_{02} are true or not. However,

In E(MSE) $\geq \sigma^2$ and E(MSB) $\geq \sigma^2$.

But E(MSA) = σ^2 only if H_{01} is true; and E(MSB) = σ^2 only if H_{02} is true.

Thus when H_{01} is true the statistic.

$$F1 = \frac{MSA}{MSE}$$

Follows F distribution with the appropriate degree of freedom. Similarly, when H02 is true the statistic.

$$F2 = \frac{MSA}{MSE}$$

Follows F distribution with the appropriate degree of freedom.

Steps in Computation of Two-Way of ANOVA Analysis of Data.

1. Reduce the observation by subtracting a constant from each data.
2. From the reduced figures, obtain the following,
   - (i) Total *(T_i)* for each group classified according to factor A, i.e., for each row.
   - (ii) Total *(T'_j)*for each group classified according to factor B, i.e., for each column.
   - (iii) Grand total ($T = \sum T_i = \sum T'_j$) for all figures.
   - (iv) Total of the squares of all figures ($\sum\sum x_{ij}^2$).
3. Calculate the CF and Sums of Squares”
   - i. Correction Factor (CF) = $\frac{T^2}{N}$
   - ii. Total SS = ($\sum\sum x_{ij}^2$).
   - iii. SSA = $\sum\{\frac{Ti^2}{k}\}$ – CF.
   - iv. SSB = $\sum\{\frac{T\prime j^2}{h}\}$ – CF.
   - v. SSE = Total SS – SSA – SSB.

[**Note**: In the calculate of various SS, we have expression like$\frac{T^2}{N}$, $\frac{Ti^2}{k}$, $\frac{T\prime j^2}{h}$ in each of which the divisor is the number of observations included in the corresponding totals T, T_i, T'_j].

4. Write down the degree of freedom (D.F);

i. D.F for Total SS = hk – 1.
ii. D.F for SSA = h – 1.
iii. D.F for SSB = k – 1.
iv. D.F for SSE = (D.F for Total SS) minus D.F for *SSA* and *SSB*.

5. Calculate the Mean Squares:

   i. MSA = SSA / D.F.
   ii. MSB = SSB / D.F.
   iii. MSE = SSE / D.F.

6. Obtain the observed values of F on dividing *MSA* and MSB by *MSE*:

   i. F1 = MSA / MSE.
   ii. F2 = MSB / MSE.

7. Consult ***F-tables*** and obtain the tabulate values (theoretical values) of *F* at (say) 5% level for the appropriate degree of freedom.

[**Note**: Degree of Freedom for ***F*** are: divisor used for MS in numerator, followed by Divisor used for MS in denominator].

8. Conclusion.

(i) If the observed value of F_1 exceed the tabulate value (theoretical value), we conclude that classification according to factor A has a differential effect on the value of the variable. That is, the means of classes by factor A are significantly different, otherwise, there is no differential effect.

(ii) If the observed value of F_2exceed the calculated value, we concluded that classification according to factor B has a differential effect on the values.
If not, then there is no differential effect.

Example:

There are four experimenters' researchers from the Ministry of Energy and Dams of South Sudan determine the content of samples of Electricity Power line in Juba, each man taking a sample from each of 4 consignments. The results are given below.

(a) Perform an analysis of variance on these data and discuss whether there is any significant difference between consignments lines or between experimenters.

Experimenters	Consignment			
	I	II	III	IV
A	9	10	9	10

B	12	11	9	11
C	11	12	10	12

(b). Also, test at 5% level which part of experiments differ significantly, if any, [Given F 5% = 5.14 for d.f. (2, 6), F 5% = 4.76 for d.f. (3, 6), and t 25% = 2.45 for d.f .]

Solution

Each observation is reduced by 10, and shown below:

Calculation for Analysis of Variance.

Experimenters	Consignment				Total (T_i)
	I	II	III	IV	
A	-1	0	-1	0	-2
B	2	1	-1	1	3
C	1	2	0	2	5
Total T'_j)	**2**	**3**	**-2**	**3**	**T = 6**

Total of the Square of all figures

$\sum\sum xij2 = (-1)^2 + (0)^2 + (-1)^2 + \ldots\ldots\ldots,(0)^2 + (2)^2 = 18$

Correction Factor (CF) = $\frac{T^2}{N} = \frac{6^2}{12} = 3$

Total SS = $\sum\sum x_{ij}^2$ – CF = 18 -3 = 15.

SS between Experimenters = $\frac{T1^2 + \;\; T2^2 \;\; T3^2}{4}$ – CF = $\frac{(-2)^2 + \;\; 3^2 + 5^2}{4-3}$

SS between Consignments = $\frac{T1'^2 + \;\; T2'^2 T3'^2 \; T4'^2}{3}$ - CF

= $\frac{(2)^2 + \;\; 3^2 + (-2)^2 + 3^2}{3-3}$ = 5.67

SS due to Error = Total SS – (SS between experimenters) – (SS between consignments).

= 15 – 6.5 – 5.67 = 2.83

Analysis of Variance Table.

Source of Variation	**d.f**	**SS**	**MS**	**F(obs.)**	**F(tab.)**
Between Experiments	2	6.5	3.25	6.91	F.05 = 5.14
Between Consignments	3	5.67	1.89	4.02	F.05 = 4.76

Error	6	2.83	0.47	-	-
Total	**11**	**15.0**	**-**	**-**	**-**

a. Since the observation value of F for experiments is (6.91) is larger than the corresponding tabulated value for d.f (2, 6), it is significant at 5% level. We therefore, conclude that the mean power contents as determined by the (3) three experiments are not equal; i.e, there are significant difference between experimenters.

The observed value of F for consignments as (4.03) is less than the corresponding tabulated value or d.f. (3, 6), and hence is not significant at 5% level. We conclude that the power content of the (4) consignments may not be different from one another, i,e, there are no significant differences between consignments.

b. Critical Difference between totals for experimenters.

C.D. = $\sqrt{0.47} * \sqrt{2 * 4}$ *2.45 = 4.8

We have T_1 = (-2), T_2 = 3, T_3 = 5. The difference between T_1 and T_2, T_1 and T_3, T_2 and T_3 are respectively 5, 7 and 2. Since 5 and 7 are larger than C.D.; experimenters A and B and A & C differ significantly.

CHAPTER EVEN

SAMPE AND SAMPLING TECHNIQUE.

INTRODUCTION

In statistics, population represents the entire group of individuals in whom we are interested to study, generally it is costly and laborious to study the whole population and, in some cases, it might be impossible as well. That is why; most research studies involve the observation of chunk from some predefined population of interest. The chunk of observation is known **as sample**. In research studies, a sample individual is observed for exposure to certain risk factors, outcome, and related variables ultimately what we conclude from sample, is often generalized to whole population from where the sample is selected. However, reliability and accuracy of conclusion will depend upon, how well the sample is representative of the whole population, from where it was drawn. In this chapter, we will discuss the different terminology used in sampling, calculation of sampling size, different types of sampling methods with their salient features alone with their application in respective situation.

Common Terminology Used in Sampling

Universe, (whole population) we mean here by the universe, entire group study population is known *as universe or whole population*, in other words, population represents the complete set of individuals, objective or scores in which we are interested, population is often too large to cover in its entirety.

Sampling Unit

Sampling unity is each member of whole population is known as sampling unit.

Sampling Frame

The sampling frame is the list where all individuals from the whole population (universe) we want to study are drawn up in that list, is known as sampling frame.

Sample

Sample is a small representative part of the whole population.

Parameter

It is the quantity that describe a population character is known as parameters, thus it is the summary value of the population, denoted by **Greek letter μ** (remember parameter for population).

Statistics

Statistics is the summary value of sample denoted by the **Roman Letter $\bar{X}$** (*remember statistic for sample*). The value of a parameter is a fixed number. In contrast, since statistic depends upon a sample, the value of a statistic can varies from sample to sample. The ultimate goals in the field of statistics are to estimate a population parameter with the help of sample statistics.

Example:

The study was done in 2010, which was called *South Sudan health household survey* to measure the healthcare in the country before independence of South Sudan, in which there are 100 people was taken for that purpose, we have to select 20 persons so 100 people constitute the universe or whole population, each of the person being referred to be sampling unity, and the selected 20 persons are known as sample. Mean of South Sudan health household survey of 100 persons or the universe is called parameter and mean SSHHS of the selected 20 persons or the sample is known as statistic.

Rationale of Sampling

There are several reasons why samples are chosen for study, rather than the entire population.

(i) First, researchers want to minimize the financial costs of collecting the information, processing this information, and reporting on the results to the managers of the Ivory Bank, branch Juba. If a reasonable aperture of a population can be obtained by observing only a selection of it, the researcher economizes by choosing such a section of population.

(ii) The process of observing the entire population would take such a large amount of time and resources that the results would not be timely, and the observations mighty are less reliable.

(iii) The information collected from a sample is on accurate.

Some Terminologies

- **Target population.**

It is entire group of individuals or objects to which researchers are interested in generalizing the conclusion, e.g. Adult population (≥ 18 years at age).

- **Study population**

Subset of the target population and it's also known as accessible population.

e.g. Adult living in field productive areas of data collection in medical survey.

- **Study subject**

Samples drawn from the study population.
e.g. Selected adults upon whom the study is conducted.

Sample size calculation

Sample size calculation provides the number of study subjects needed to carry out the study. Without prior and proper estimation of appropriate sample size, the research may lead to an erroneous outcome. Sample size calculation solely depends upon the types of epidemiological study. For calculation of sample size for descriptive, case control cohort study and randomized controlled trial different formulas are used.

(A) Sample size calculation for descriptive study

In a descriptive study, prerequisites for calculation of sample size are:

(a) Known prevalence or standard deviation (S.D) from existing literature (previous studies) or from pilot study.
(b) Significant level.
(c) Desired Precision.
(d) Allowable error.

(i) For qualitative data

The formula used is = $\frac{\mathbf{Z\alpha^2pq}}{\mathbf{L^2}}$

Where zα = Standard normal deviation at a desired confidence level (95% or 99%).

p = previous prevalence, q = 100-p, L= allowable error 5%, 10% or 20% of p.

At 95% confidence level zα = 1.96 (sometimes 2 is considered instead of 1.96, so zα is considered as 4), while at 99% confidence level zα = 2.58.

Example:
If prevalence of low birth weight in Malakal Teaching Hospital was found to be 30% as the result of the high cases among the pregnancy women. Please, calculate the sample size, if allowable error is 10%.

Solution:
Here p = 30, q = 100 – p = 70, L = 10% of p = 10/100 x 30 =3
So the required sample size = $\frac{Za^2pq}{L^2} = \frac{4\times30\times70}{3\times3}$ **=933.333**

(ii) **For quantitative data**

The formula used is $\frac{Za^2\sigma^2}{L^2}$

Where za^2 = standard normal deviate at desired confidence level (95% or 99%), σ standard deviation (S.D); L= allowable error. This allowable error may be in absolute term or relative term. If taken as relative term, then it will be 5% or 10% of mean. At 95% confidence level zα = 1.96 (sometimes 2 is considered instead of 1.96, so za^2 is considered as 4), while at 99% considered level zα = 2.58.

Example:

If mean of systolic blood pressure of a population admitted in Juba Medical Complex is 130 mm of Hg and SD is 10 mm of Hg, calculate the required sample size if allowable error is ≠ 1 at 5% level. Here σ = 10mm of Hg, l = ±1.

Solution:

We used this formula = $\frac{Z\alpha^2\sigma^2}{L^2}$

So required sample size = $\frac{4\times100}{1\times1}$ = 400

(B) Sample Size Calculation for Analytical Studies

(Cases control, cohort, randomized controlled trial)

Here prerequisites are:

i. For case control study:
Anticipated probability of exposure for people with diseases and without diseases separately, anticipated add ratio. (These values are taken from previous studies), confidence level and precision.

ii. For cohort study: anticipated probability of diseases in people exposed to the factor of interest, anticipated probability of disease in people not exposed to the factor of interest anticipated relative risk (these values are taken from previous studies confidence level and precision.

iii. For randomized controlled trial (RCT):
zα = Z value for alpha error (**type I error**) zβ = Z value for beta error (**type II error**), standard deviation proportion of event and mean difference to be detected (RCTs) with one experimental group and one control group and considering only alpha error or both alpha and beta errors.

(1) RCTs using student's t-test and alpha error:

$$TheformulausedisN = \frac{(za)^2.\,2(s)2}{(d)^2}$$

(2) RCTs using student's t-test and considering alpha and beta error:

The formula used is N $\frac{(za+z\beta)^2.2(s)2}{(d)^2}$

Zα =Z value for alpha error (at 95% confidence level it is 1.96 in two talled);

Zβ = Z value for beta error (20% beta error and 80% power it is 0.84 in one talled); 8 = standard deviation, d = mean difference to be detected.

Example

The National Ministry of Health and drug authorities in Juba with collaboration with World Health Organization (WHO) has conducted Health Survey using randomized controlled trial (RCT) involving blood glucose reduction by hypoglycemic drugs in South Sudan, all the patients are diabetic. S = standard deviation, suppose = 10 mg% d = difference to be detected, suppose ≥ 5 mg %, Zα = 1.96 Zβ = 0.84. Find the sample size by using N= $\frac{(za+z\beta)^2.2(s)2}{(d)^2}$

Solution:
By using the above formula:

$1.96 + 0.84^{2}\,\frac{2.10^{2}}{5^{2}}$ = 63 – 73, SSS = 63

Thus 63 subjects will be selected per group × 2 groups = 126 subjects (total).

(3) RCTs using test at differences in proportion and considering both alpha and beta error.

The used in this N= $(Z\alpha + Z\beta)^{2(}\frac{P1Q1+P2Q2)}{(L)^{2}}$

Zα = Z value for alpha error (at 95% confidence level it is 1.96 in true tailed) Zβ = Z value for beta error (20% beta error and power it is 0.84 in one tailed);

P1 = proportion of success with new treatment. Q1 = 1-$p_{1.}$

P2 = proportion of success with standard treatment, Q2=1-p_2.

L= p1-p2.

Example:
The National Ministry of Finance and Economic Planning have chosen randomized controlled trial (RCTs) by field visited Nimule, customer check point to see the taxes collection procedures.
P1 = proportion of success the main system of paying taxes in Nimule through the Banks, suppose 80% = 0.8); Q_1 = 1- P_1 = 1 - 0.8 = (0.2);
P2 = proportion of success with standard checking (suppose 60% = 0.6); Q_2 = 1- p_2 =1-0.6 = (0.4);
L = p_1-p_2 (0.8-0.6) = 0.2

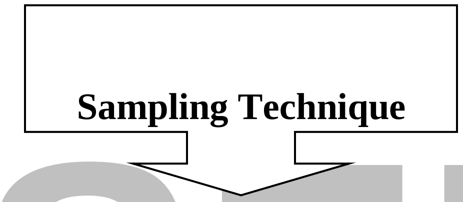
Zα = 1.96, Zβ = 0.84
Now the required sample size would be found using this formula $(Z\alpha + Z\beta)^{2(}\frac{P1q1+P2q2)}{(L)^{2}}$
$\frac{(1.96+0.84)^{2}.(0.8\times0.2+0.6\ \times0.4)}{(0.2)^{2}}$=78.4
78 subjects will be selected for group×2 groups = 156.8 are called subjects (total).

Large sample.
Sample size greater than or equal to 30 is called large sample.
Small sample.
Sample size less than or equal to 30 is called small sample.
Precision.
It reveals how well the repeated measurement agrees with each other. Epidemiologist refers to this as reproducibility, repeatability, consistence, or reliability. Precision means minimization of

error. Small sample size leaks precision and large the sample, the more precision is the estimate. High standard deviation indicates low precision, low standard deviation indicates high precision.

Sampling Technique

When we carry out the study or survey, we have two options with us based upon the feasibility, we may study all the study population, or we may select some subjects from the study population. When all the individuals in the population or accessible population are studied, this is known as complete enumeration or census method. But if only few individuals are selected from the study population, it requires sampling technique which is two types.



Non-Probability and Probability

I) Non-probability

Here selection does not depend upon any laws of probability. It is non-random sampling; samples are selected deliberated by the researcher on his own choice. The different types of non-probability sampling are as follows:

Purposive sampling, judgmental sampling, convenience sampling, self-selection sampling, snowball sampling, quota sampling etc.

A) Purposive sampling.

Here participant is purposively selected from whom information can be obtained easily. This is done to avoid people who might not be interested in participation in the study. This sampling is also known as judgmental sampling since it is based on the judgment of the researcher. This method is less time consuming and useful at times but can be.

B) Convenience sampling.

Here, participants are selected based on easy accessibility.

Example

In the state of Fashoda, some Primary Schools were selected to receive and welcome the new elected Governor based upon their location beside a main road, to sing the national Anthem.

C) Self- selection sampling.

Participants take part in the research on their own as a volunteer.

D) Snowball sampling.

This type of sampling is applied when the target population is hidden and/or hard to reach such as drug addicts, commercial sex worker, individuals with HIV/AIDS, etc, or any other problem in question is rare where study subjects are hard to find. In the process, one study subject is asked to identify persons with the same exposure in question for finding the next subject or persons. The researcher then goes to the identified persons and continues in the same way until the required sample size is obtained. This type of sampling is also referred to as network sampling.

E) Quota sampling

In quota sampling, researchers are given quotas to fill from different strata of population keeping the proportion of quota the same as observation or observed in the population.

Example:

In many villages on **Gokrial County**, the population are divided into two main paths believe Christian and Muslim as 60% Christian and respectively; and the researcher by the method of quota sampling can select participants by his own choice in the same ratio of 6:4, (Christian and Muslim). Inferences drawn from these non-probability samplings are not amenable to statistical analysis. Qualitatively, this sampling method is inferior to probability sampling.

II) **Probability Sampling**

This is superior to non-probability sampling. It obeys law of probability and is based on the concept of random selection. This type of sampling is also known as random sampling or chance sampling.

Types of probability sampling.

A) Simple random sampling.

This is the most common and simplest of the sampling method. The main benefit of simple random sampling is that each member of population (every sampling unity) has an equal chance of being chosen. This means that it guarantees that sample chosen is representative of the population. The method is applicable when population is small (Homogeneous) and readily available. Complete population list must be available to build up the sampling frame from where sampling unity are chosen. At first all the sampling unity are assigned with numbers. They sample can be selected either with replacement or without replacement.

i. Simple random sampling with replacement (SRSWR) sample random sampling is said to be "with replacement "when the sample numbers are drawn from the population one by one and after each drawing, the selected population unity is noted and then returned to the population before the next one is drawn. This means that at each stage of sampling process all the population unity (including those obtained in earlier drawings) are considered for section with equal probability. Thus, the population remains the same before each drawing and any of the population unity may appear more than once in the sample.

ii. Simple random sampling without replacement (SRSWOR) simple random sampling is said to be “without replacement “when either the sample members are drawn all at time or drawn one by one in each a manner that after each drawing the selected unity is not returned to the population when the next one is drawn. This means that when drawings are made one by one at each stage of the sampling process the population unity already chosen are not considered subsequent selections, but the drawing is made with equal probability only from this unity not selected in any of the earlier drawings. It is evident that in simple random sampling without replacement from a suit population, the size of population goes on diminishing as sampling process continues.

Let a simple random sample of n members be drawn from a finite population consisting of N members. (i.e sample size = n population = N) then total number of possible samples with distinct permutation of members. (i) in SRSWR is N^n (II) IN SRWOR is ${}_{P}^{N}n$ = N (N—1) (N—2) ……(N—n+1), among these the number of cases favorable to each group or distinct combination of members. (i) in the SRWR is not the same (ii) in SRWOR is the same n! For this reason, we generally consider.

(a) In SRWR N^n possible samples with distinct permutation each with a probability $\frac{1!}{N^n}$

(b) In SRWOR $^{N}C_{n}$ = n!$\frac{N!}{(N-n)^{!}}$ Possible sample with combination each with probability =

Random sampling is the simplest and most important among the various sampling techniques. It is free from the influence of human bias and hence called lottery sampling. Chance alone determines whether one unity or the other is selected. The selected random sample is facilitated by what is known as random numbers. Random sampling is the most appropriate cases when the population is homogeneous with respect to the characteristic under study. The theories of distribution and test of significance are based on random sampling only.

Simple sampling: this is a special case of simple random sampling in which the probability of section of any member remains constant throughout the sampling process irrespective of whether the member had been selected earlier or not. Therefore, SRWR will always give a simple sample from a finite or an infinite population. However, SRWOR gives a simple sample only when the population is infinite.

B) Purposive sampling

A sample, which is selected based on individual judgment of the sample, is called a purposive sample. There is no special technique for selecting a purposive sample, but the sampler picks out a typical or representative sample according to his own judgment. It all depends on the persona factor and chance is not allowed to pay at all. Consequently, there is much scope for bias and the degree of accuracy of the estimates is not known. Purposive sampling may be useful when the sample is small; but as the sample size increase the estimates become unreliable due to accumulation of bias.[17] The advantage purposive sampling that whereas a random sample may vary widely from the average purposive sample will not.

C) Stratified sampling

In stratified sampling, the population is subdivided into several parts called strata, and then a sub-sample is chosen from each of them, all the sub-sample combination to setter gives the stratified sample. If random sampling does the selection from strata, the method is known as stratified random sampling. Stratified sampling is generally used when the population is heterogeneous but can be sub divided into the strata within each of which the heterogeneity is not so prominent. Therefore, necessary for subdivision into strata called stratification. If a proper stratification can be made such that the strata differ from one another as much as possible, but there is much homogeneity within each of them, then a stratified sample of will yield better estimate them a random sample of the same size. This is because in the stratified sample the different sections of the population are suitably represented through the sub-samples, while random sampling some of these sections may be overrepresented, underrepresented, or even be omitted.

D) Systematic sampling

Systematic sampling involves the selection of sample unity at equal intervals, after all the unities in this population have been arranged in some order. If the population size is finite, the unity may be serially numbered and arranged. From the first k of these, single unity is chosen at random. This unity and every k-th unity thereafter constitutes sample. To obtain a systematic sample of 500 village out of 40,000 in South Sudan.

Example

One out of 80 on an average, all the villages must be numbered serially from the first % of this village is selected at random, suppose with the serial number 27. Then the villages with serial numbers 27, 107, 187, 267, 347……. Constitutes the systematic sample. If the characteristic under study is independent of the order of arrangement of unity, then a systematic sample is particularly equivalent to a random sample. The actual selection of sample is easier and quicker, systematic sampling is suitable when the unities are described on serial number cards, e.g. workers in the upper Nile University are listed on cards. Then the sample can be drawn from early by looking at the serial numbers. The sample may be biased it there are period features associated with the sampling interval.

[17] N. D Das 2009, Statistical Method, Values I and II. M.c Green Hill education, India, private limited, New Delhi, p, 493.

E) Multi-stage sampling

Multi-stage sampling refers to a sampling procedure, which is carried out in several stages. The population is first divided into large groups, called first stage unity. These first stage unity are again divided into smaller unities called second stage unity, the second stage unity is divided into three-stage unity, and so on, until we read the ultimate unity. Initially some of the first stage unity are selected from each of which some second stage unity are chosen and the process is carried on from stage to stage unity the selection of ultimate unity.

Example

To introduce a scheme on experimental bases in the village, we may have to select a few villages from the whole of the state. If we apply three stages sampling, sub-division may be used as first stage unity Anchall forming the second stage unity and then the village ultimate unity.

Multi –stage sampling enables existing division and subdivision of the population to be used as unity at various stages and permit the fieldwork to concentrate although a large area is covered. Another advantage is that the subdivisions into second stage unity need to be carried out for only those first stage unities, which are included in the sample. It, therefore, helps in survey of underdevelopment areas where no sampling frame is sufficiently detailed and accurate for subdivision of the natural unity into reasonably small stage sampling unity. Usually, considerable saving in cost is achieved through multi-stage sampling. However, this method is in general less accurate than any other sampling method using the same number of ultimate unities by some single stage process.

Example:

The following table gives the grades of 100 students in mathematics draw a random sample f size ten (10) from the group of students and estimate the mean grade from the sample.

75	85	80	86	76	65	75	76	72	86
66	63	60	69	80	66	87	73	58	78
86	79	95	84	41	76	78	74	74	56
68	79	73	72	73	87	77	60	87	40
66	81	84	72	63	59	76	52	57	78
75	74	98	64	45	68	57	79	79	83
60	52	63	80	94	34	78	64	58	56

Table of random sampling number to be supplied.

	83	82	80	88	68
	74	85	96	60	72
	50	80	66	96	80
	82	77	87	76	82
	79	92	80	65	90
	35	76	88	67	75

Solution:
Hence, the population size = 100 each of the population number is therefore allotted a two digited identify number 01, 02….99 and 00 (for the 100th number) as shown below:[18]
We now take down random numbers in-group of two starting from the digit 5 in the 2nd raw and 15th in column (this was down without looking at table) and moving horizontally using the random number table. The numbers obtained are 58, 99, 15, 75, 24, 58, rejected 99,00,85,66. The 7th number 58 is rejected, because it has already appeared one. The ten random numbers remaining and the grades having there as identify numbers are shown below.

Random numbers	58	99	83	15	75	24	92	00	85	66
Grades	87	58	79	68	90	58	52	56	83	59

The random sample of size ten is, therefore, composed of grade.

87, 58, 79, 68, 90, 58, 52, 83, 59.

The mean of the sample, when gives an estimate of population mean is $\overline{\times}$ = (87+58+ ….+83+59) Ho $=\frac{690}{10}$ = 69.

Incidentally, the population mean of 100 grades 72-66:

Example:
Draw a random sample of size 10 (without replacement) from the following data, stating clearly the procedure followed by you:

45	24	43	17	05	28	27	21	11	46	
33	26	24	14	34	21	25	48	35	38	
26	27	35	08	30	26	30	28	21	27	
20	13	23	36	38	20	25	31	24	18	12

[18] Ibid, p, 496.

You may use the random sampling numbers given below:

5967, 8941, 7989, 3335, 7577, 9735.

3042, 8409, 7053, 5364, 5872, 1143.

Solution:

There are altogether N = 41 member, and these are serial numbered horizontally from 1 to 41. Since the population size N = 41 is two digited, the identify numbers must also be tow digited. But there are 100 two digited numbers possible, as 00, 01, 02……99.

If we allot only one identify number to each member, many random numbers will be rejected before we get the required sample. Let us allot 2 identify numbers to get each member as follows:

Serial No	1	2	3	---	---	----	39	40	41
Identify	01	02	03	----	----	------	39	40	41
Numbers	42	43	41	----	-----	------	80	81	82

Note:

It will be notice that identify numbers up to 41 are the same as the corresponding serial number; and the identify numbers when divided by 41 leave remainders which are equal to the serial number. With identify number 82, however, the remainder is zero, and the last observers are referred to. Starting from the first of the given random sampling numbers and proceeding raw-wise, we now had taken groups of 2 consecutive digits.

59, 67, 89, (rejected) 41, 79, 89, (rejected) 33, 35, 75, 77, 97(“rejected) 35, 30, 42, 84, (rejected) 09, 70, etc.

The numbers greater than 82 have been rejected because the identify numbers are 01 to 82 only. (If 00 appeared it will be also have been rejected). Their remainders on division by 41 replace the numbers greater than 41 and we now have:

18, 26, 41, 38, 33, 35, 34, 36, 35 (rejected), 30, 1, 9, 20, etc. since the drawing are without replacement the second 35 occurring above has been rejected, of the remaining numbers, the first ten are:

18, 26, 41, 38, 33, 35, 34, 36, 30 AND 1

Note that some numbers all lie between 1 and 41, and none is rejected. The numbers having these as serial numbers are respectively. 48, 26, 12, 31, 23, 38, 36, 20, 27, 45. These observations comprise the required random sample of size 10.

Alternative method

This is a multiple of the previous method note that 100 identify numbers 00, 01, 02---- 99 and 00 can be allotted equal among 50 (slightly higher than 41) member. Let the population members be serially numbered horizontally from 1 - 41 as before and two digits random numbers are taken down. 50 divide these numbers and the remainders obtained. If a number is less than 50, the number it-self is taken as the remainders, any remainders between 42 and 49 and 00, if anyone retained only at the first appearance, the repetitions are rejected. Thus, the two digits random numbers.

59,67,89,41, 72, 89, 33, 35, 75, 77, 35, 30, 42, 84, 09 leave the remainders.

9, 17, 39, 41, 29, 39 (rejected because repeated) 33, 35, 25, 27, 47 (rejected because more than 41, 35, (rejected) because repeated 30, etc.

From the numbers of now remaining the first ten are only returned. 9, 17, 39, 41, 29, 33, 35, 25, 27, 30.

The numbers of the population that correspond to these as serial numbers as:

11, 25, 24, 12, 21, 23, 38, 30, 30, 97 constitute the required random sample.

Example:
Describe in detail how you will select without replacement, a random sample of 3 unities from a population of 121 unities using a procedure, which does not involve rejection of many random numbers.

Solution:
The method is described in the following steps:

1) All the 121 unities of the population are serially numbered from 1 to 121.
2) Let us now take a page of random sampling numbers and with closed eyes selected a digit from the page, starting from this proceeding horizontally, consecutive digits are taken down in groups of three because of largest serial number 121 is their digit given several 3-digit numbers.
3) 125 divide the random numbers and the remainders obtained. (Note that 1000 three-digit numbers are possible as 001,002---- 999 and 000; and 1000 is exactly divided 125). If a number is less than 125, that number it-self is taken as the remainder.
4) Any remainder between 122 and 124 are also 000, if any occurring, are rejected, if and remainder appears more than one, all subsequent repetition is rejected.
5) From the numbers now remaining only the first three numbers are taken and the population unities corresponding to these as serial numbers give the random sample.

Example

Draw a random sample of size 10 unity replacement from the following frequency distribution and compare the sample mean with the population mean.

Annual sales ssp.	30-39	40-49	50-59	60-69	70-79	80-89	90-99	100-101	Total
No, of firms	14	36	47	66	41	24	10	2	240

All 240 firms considered in the frequency distribution are serially numbered 14 observation in the class interval 30-39 are numbered 1 to 14, 36 in the next interval are numbered 15 to 50 and so on, show in the table below.

Are measured 15 to 50 and so on, as showing the table below.

Label Calculation of cumulative frequencies

Class intervals	Frequencies	Mid value	Cumulative frequency	Serial No. allow
30-39	14	34.5	14	1-14
40-49	36	44.5	50	15-50
50-59	47	54.5	97	51-97
60-69	66	64.5	163	98-163
70-79	41	74.5	204	164-204
80-89	24	84.5	229	205-228
90-99	10	94.5	238	229-238
100-109	2	104.5	240	239-240

Using the random sampling numbers, we start from (Row 6, column (1) a takedown 3 digit numbers proceeding horizontally. The numbers 250 and above are replaced by their remainders on division by 250 any number between 242-249 or 000, if occurring is now rejected.

Random number	489	572	665	890	501	154	786	475	888	750	147
Remainders	239	72	165	140	1	154	36	225	138	000	147

The remainder 000 is rejected, the first 10 at the numbers are now retained, and the corresponding to these as serial numbers constitute the sample of firms. Note that there is no necessity of rejecting any number that is repeated, because the sample is drawn with replacement.

The random sample is given by the mid- values of the classes in which the selectee firms are placed. They can be done by looking at the last column and then the 3rd column of table (A) above.

Serial No.	239	72	165	140	1	154	36	225	138	147	Total
Sample value	104.5	54.5	74.5	64.5	34.5	64.5	44.5	84.5	64.5	64.5	655.0

The sample means is 655÷10 = 65.5 (ssp). The mean calculated from the given frequency distribution. (Calculation not shown here) is 63.08 SSP sample mean ($\bar{x}$) = 65.5, population means μ = (63.08)

Standard Error (SE).

Standard error of a statistic is the standard deviation calculated from the sampling distribution of the statistic.[19]

When studying a population of universe, many different samplings can be chosen out of it if we calculate the sample mean; we could see that all sample means are different, through all the samples have been drawn from the same universe. Often the means of the samples follow a normal distribution, even if the original population value was not normal distribution such as a difference between sample and population value is measured by standard error. It is measured by statistic as standard error. It is an estimate of the standard deviation of the sampling distribution of a statistic.

Significant of standard error.

1. Standard error is used to construct confidence interval around a sample statistic.
2. Standard error is inversely related to the sample size. The larger of the study (sample size), the smaller the confidence interval and the greater the precision of the estimate.

Standard deviation versus standard error

S/N	Standard deviation (SD)	Standard error (SE)
1.	Standard deviation is used to measure the dispersion of the data in partier sample	Standard error is a measure of sample population.
2.	**Standard deviati**on measures the variability in the dataset	Standard error measures the precision of estimate of a population parameter provided by the sample means of proportion
3.	Standard deviation should be used when the purpose is to describe the data	Standard error should be used when the purpose is to describe the outcome of the study

Confidence limits

The upper and lower values of confidence interval are known as confidence limits.

Confidence interval

[19] Idranil, 1998, Bio-statistics, p, 85.

1. A confidence interval is the probable range for a population measure around a point estimate from the sample.
2. This is used to assess a single point estimate such as which is 95% confident that the true population value lies.
3. 95% confidence interval for population mean (proportion) = sample mean proportion ± 1.96 at less standard error. a 95% confidence interval for the population mean indicate that there is a 95% probability that the population mean lies within that confidence interval.
4. When the confidence interval is wider, the point estimate is likely to be less accurate and less reliable.
5. As the sample size increases, the confidence interval decreases in size.

90% confidence interval: sample statistic + 1.645 standard error
95% confidence interval sample ±1.96 standard error
99% confidence interval sample ±2.576 standard error

Testing of significance applying standard error

a- Standard error of the mean

I. For random sample of size n, it is denoted by the formula ($\frac{SD}{\sqrt{n}}$ or $\frac{\sigma}{\sqrt{n}}$ in some book for the sample mean in a random sample test of HIV of 100 students medicine aged 10-15 years,[20] mean weight was found to be 35kg with a SD of 5kg. what was the true mean of the universe from which the sample was drawn?

Solution:

Here $\frac{SD}{\sqrt{n}} = \frac{5}{\sqrt{100}} = \frac{5}{10} = 0.5$

Thus, mean of the universe will range from mean $\pm$ 2 SE = 35$\pm$2$\times$ 0.5. i.e from 34 to 36 kg; in 95% of times [2 $maybeusedinsteadof$ 1.96]

When either the population size is infinitely large, or the sample is drawn with replacement. We used the formula.

II. SE = $\frac{\sigma}{\sqrt{N.\sqrt{\frac{N-n}{N-1}}}}$

When The Population Size N infinity, but the sample is drawn without replacement.

Example

[20] Das p, 86.

A sample random sample of size 5 is drawn without replacement from a finite population consisting of 41 unities.[21] If the population standard deviation is 6.25, what is the standard error of sample mean?

Solution:

Here, n = n, N = 41, SD or σ = 6.25 using the formula of S. E of mean = $\frac{\sigma}{\sqrt{n \cdot \sqrt{\frac{N-n}{N-1}}}} = \frac{6.25\,.}{\sqrt{5}}$

$= \sqrt{\frac{41-5\,=}{41-1}}\frac{6.25\times 6}{{}^{10}\sqrt{2}}$= 2.65

B**) Standard error of Proportion**

If a random sample of size n is drawn from a population consisting of a proportion p of the units belonging to a certain category. (e.g.)

Proportion of detective in batch of articles, then standard error SE of sample proportion is given by.

(i) SE = $\sqrt{\frac{Pq}{n}}$ the formula

When either the population size is infinitely large, or the sample is drawn with replacement.

(ii) S.E = $\frac{P+q}{\sqrt{\frac{pq}{n}}} = \frac{1,}{\sqrt{\frac{N-n}{N-1}}}$ i.e. q = 1-p

When the population size N is finite, but the sample is without replacement. It may be noted that the formula of S.E in simple sampling are easier in sampling without population from a finite population S.E contain an error factor $\sqrt{\frac{N-n}{N-1}}$, which is called finite population correction (f.p.c)

Example

Proportion of Muslim who live in rank is 40%, now a random sample of 100 people was taken, where Muslim were 25%. How can we say that sample represents the true universe?

Solution:

N = 100, p = 25, q 75.

Here, SE = $\sqrt{\frac{pq}{n}} = \sqrt{\frac{25\times 75}{100}}$ = 4.33

Now if we take two SE on either side of 25, the value comes as 25±2 × 4.3 i.e. 16.4 to 33.6.

Thus, the sample having 25% Muslim does not incorporate the population value i.e. 40% at 95% of confidence link and is not representing the universe (we may use 2 instead of 1.96)?

[21] Ibid, p, 572.

Example

It has been found that 2% of the tools produced by a certain machine in Ogween Company are detective. What is the probability that in a shipment of 400 such tools, 3% or more will be detective? (Probability that the normal deviate lies between 0 and 1.43 is 0.4236).

Solution:

Since the sample size n = 400 is rare, the sample proportion (p) is a approximately normally distribution.

Mean = p = 2% = 0.02, q 1-p = 1 - 0.02 = 0.98, N = 400.

SE = $\sqrt{\frac{pq}{n}}$ = $\sqrt{\frac{0.02\times0.98}{400}}$ =0.007

Probability that the sample proportion (p) exceed 0.03 = area under the standard normal curve to the right of the ordinate at the standardized value Z = $\frac{0.03-0.02}{0.007}$ = 1.43. Area to the right 0) – (Area between 0 and 1.43) = 0.5 – 0.04236 = 0.764.

c) Standard Error of Difference between Two Mean

It is denoted by the formula.

SE = $\sqrt{\frac{(SD1)^2}{n1} \frac{(SD2)^2}{n2}}$

The significance of is difference is found by Z test.

Z test = $\frac{observationdifference}{standarderrorofdifference.}$

If the difference between the two means is more than twice (or 1.96) the SE of difference between the two means, than the difference between the means is significance.

Example.

The following data are given.

Primary school A:

Mean weight of student is 30 kg, standard deviation is 5kg (SD$_1$), Total students 50 (n$_1$).

Primary school B:

Mean weights of students are 28kg, standard deviation is 3kg (SD2) total student is 75 (n$_2$).

Is the difference between the two groups statistical significance.

Here standard error of difference between two means would be $\sqrt{\frac{(\sigma1)^2}{n1} \frac{(\sigma2)^2}{n2}}$ = $\sqrt{\frac{(5)}{50} +^2 \frac{(3)^2}{75}}$ = 0.78

So Z = $\frac{observeddifference}{standarderrorofdifferencebetweentwomeans.}$ = $\frac{30-28}{0.78}$ = 2.56

Since Z value is more than 2 (1.96), the difference of mean between the two groups is statistically significant at 95% confident unity.

D) **Standard Error of Difference between Two Proportions:**

It is donated by the formula,

The significance of difference is found by Z test = $\frac{\text{observed difference}}{\text{standard error of difference}}$

If the observed difference between the two groups is more than twice or 1.96 the SE of difference between the two proportions, then the difference is significant.

Example:

Cure rate of ARI I two different groups by two different drugs A and B.

Drug A:

Cure rate 70% (P) among 50 subjects (n_1); $q_1 = 100 - P_1 = (100 - 70) = 30$

Drug B

Cure rate 60% (p_1) among 75 subjects (n_2); $q_2 = (100 - p_{2)} = (100 – 60) = 40$

Here, standard error of difference between two proportions would be.

SE of proportion =

So, Z = observed difference standard error of different between two proportions

$\frac{70-60}{8-6}$ = 1.16

Since the Z value is less than two (2), the cure rate is not statistically significant in two groups of 95% confidence unity.

Null hypothesis.

Null hypothesis symbolized as Ho, is the hypothesis, which denoted that the sample or population being compared in an experiment or study, or test are similar. Any difference discerned is ascribed to chance and not to any other measurable factor.

Example

We Are Interested in comparing smoking rate in men and women in population. Then null hypothesis would be-

Ho: smoking rate are the same in men and women in the population.

H1: smoking rates are different in men and women in the population. Here, A_1 is the alternative hypothesis. Here we have not specified any direction for the difference in smoking rates i.e we have not specified whether men have higher or lower rates than women in the population. This leads to two-tailed test because we allow for either eventuality. In some cases we may carry out an on-tailed test in which a direction of effect is specified in H_1.

- The null hypothesis and the alternative hypothesis must be stated before initiation of the study.

- After data collection, relevant statistical test of significance is applied to determine whether to reject or accept null hypothesis.
- The level of confidence for rejecting null hypothesis is arbitrary. One conventional cut off for defining a significant difference is 5%; that is if the probability that the difference is due to chance is 5% or less than the null hypothesis is rejected and an alternative hypothesis is accepted.

The smaller the P values the greater evidence against null hypothesis. However, more test of statistical significance does not prove causality.

Null hypothesis is rejected when p $\leq$ 0.05

Null hypothesis is accepted when P 0.05

Statistical test of hypothesis:

1- Identified null hypothesis.
2- Determine the levels of Q and B error of inference.
3- Determine the best statistical test for the stated null hypothesis.
4- Perform the statistical test.
5- Conclusion according to the result.

By default, statistical packages report two tailed P values. This is because the most commonly used test statistical distributions (normal student's t) are symmetries about zero, and if necessary one tailed P value by two.

Sampling Error

An error occurs due to chance and is concerned with either incorrect acceptance or rejection of null hypothesis. There are two types of sampling error, Type I and type II. Type I error (α error)

- Also known as error an error of the first kind significance level of test or α error.
- It is the rejection of null hypothesis when it is true.
- We must decide the value of α error before we collect data; we usually assign a conventional value of 0.05, though we might choose a value such as 0.01. ***We reject null hypothesis if our p value is less than the significance level*** **i.e p ≤ α.**
- It is used to calculate sample size.

Type II error (β error)

- Also known as error of the second kind of β error.
- It is the acceptance of null hypothesis as true when it is the actually false. If a null hypothesis is true, there can be no type II error.
- Type II error is inversely related to type I error. The probability of type II error is equal to β.

- Type II error is used to determine the power of study, which equals 1 minus β. the power is therefore, the probability of rejecting the null hypothesis when it is false.
- The type II is a hidden error because it cannot be detected without a proper power analysis.

Decision

Null hypothesis	**Accept**	**Reject**
True	correct conclusion	type I (α) error
False	type II (B) error	correct conclusion

TYPE (I) ERROR AND TYPE (II) ERROR.

Type (I) Error	**Type (II) Error**
α error False position result (false arm) P values level of significance determine It should be less than 0.05 αerror decreases as p value decreases	B error False negative result (miss) Denoted power of study Power of test 1-β

According to statistical quality control

α = producer's risk. (The producer suffers because a lot with acceptable quality is rejected).

β = consumer's risk. (The consumer suffers because a lot with unacceptable).

Relationship between Type I and Type II Errors

Type I error is inversely related to type II error, so decreases in type I error will lead to increases in type II error. The probability of making a type I error can be decreased by setting a more stringent significance level. In other words one can set the level at 0.01 instead of 0.05; then there is only one chance in 100(1%) that the result termed significant could occur by chance alone.

Reduction of type I (α) error can be done by setting an acceptable level of (α) error conventionally 0.05 (1 in 20), though we might choose a value such as 0.01.

Reduction of type II (β) error

This is done by the following ways.

I. Increase sample size
II. Reduce variability of measurement (i.e) increased accuracy

III. Use a one tailed test (directional alternative hypothesis)

IV. Set less demanding type I error.

CHAPTER EIGHT

REGRETION ANALYSIS.

REGRESSION

The regression is represented estimation or prediction of the average value of one variable for a specific value of the other variable. The estimation is done by means of suitable equations, derived on the basis if available vicariate data, such an equation is known as a regression equation on its geometrical representation is called a regression curve. In linear regression (simple regression), the relationship between three variables is estimated to be linear. The estimate of Y (say, Y) is obtained from an equation of the form.

$\bar{Y} - \bar{Y} = byx\ (X - \bar{X})$ ____________ (i)

Moreover, estimation of X (say, X^{2}) from the equation (usually different from the farmer of the farm.

$X^{1} - \bar{X} = bxy\ (Y - \bar{Y})$ _____________________ (ii).

Equation (i) is known as Regression equation of Y and X, and equation (ii) as Regression equation of Y and X. the coefficient of byx appearing in the regression equation of Y on X is known as regression Coefficient of Y and X. similarly, bxy is called the regression coefficient of X and Y. the geometrical representation of linear regression equation (i) and (ii) are known as Regression lines, these lines are best fitting straight lines obtained by the method of at least squares.

Example:

Derived the regression equation of Y on X or X on Y. or obtained the equation of the two lines of regression for a vicariate distribution.

Solutions:

(1) Regression equations of Y on X.

The regression equation of Y on X is the equation is the equation of the best fitting straight line in form of Y = a+bx, obtained by the method of the last square.

Let (X, Y), (X_2, Y_2) ………………… (X_n, Y_n) be set of n pairs of observation and let us fit a straight line of the form Y = a + bx ________________ (i) to these data, (here X is considered to be the independent variable and Y is dependent variable, i.e. a value of Y is obtained when a value of Y is obtained when a value of X is given.

Applying the method of least squares the constant a and b are obtained by solving the normal equation:

$\sum y\ an + b\sum x$ = ________________________ (ii)

$\sum x\ y$ = a$\sum x$ + $b\sum x^2$ ____________________ (iii)

Dividing both sides of (ii) by n we set $\bar{Y}$ = a + b$\bar{X}$, so that a = $\bar{Y}$ - b$\bar{X}$. Substitute the Y in equation (i) Y – Y = b (X - $\bar{X}$) __________________ (iv).

Subtracting the first from the second,

n($\sum x\ y$) – ($\sum x$) ($\sum y$) = b {n $\sum x^2$ - ($\sum x^2$)}

:- $\frac{n\sum xy/n -(\sum x)(\sum x)}{n\ \sum x2-(\sum x2)}$

Dividing both numerator and denominator by n^2,

B = $\frac{n\sum xy/n -(\sum x/n)(\sum y/n)}{n\ \sum x2/n-(\sum xn2)} = \frac{cor(x,y)}{6\ X\ 2}$ ________________ (v)

Writing b with the usual subscripts, we have from (iv)

Y - $\bar{Y}$ = byx (x - $\bar{x}$) _____________ (vi)

Where byx = cor (x, y)/ $6x^2$. This is the required regression equation of Y on X since r = cor (x, y) (6x, 6y), we see that cor (X, Y) = r6X 6Y.

Substituting this:

byx $= \frac{cor\ (x,Y)}{6\ X\ 2} = \frac{r6y}{6x}$ ____________________ (vii)

(2) **Regression equation of X on Y,** if we fit an equation of the form X = a' + b'Y, assuming X is to be dependent variable and Y the independent variable, we obtained the regression equation of X on Y. applying the method of least squares, the normal quotations for determined a' and b' are $\sum x$ = a'n + b'$\sum y,$ $\sum x$ y = a'$\sum y + b'\sum y^2$ __________ (viii)

Processing the same way as before the regression equation of X on Y is found to be

X - $\bar{X}$ = bxy (Y -$\bar{Y}$) ________________________ (ix)

Where bxy = $\frac{cor\ (x,Y)}{6\ y2}$ = r$\frac{6y}{6x}$ ___________________ (x)

PROPERTIES OF LINEAR REGRESSION

The regression equation of Y on X is Y - Ȳ = bxy (X - X̄)

Where $\text{b}_{\text{xy}} = \boldsymbol{\frac{cor\ (x,Y)}{6\ y2}} = \mathbf{r}\boldsymbol{\frac{6y}{6x}}$

This equation is used to estimate the value of x is known. The regression equation of X on Y is

X - X̄ = bxy (Y - Ȳ), where bxy = $\frac{cor\ (x,Y)}{6\ y2}$ = r$\frac{6y}{6x}$, this equation is used to estimate X, when the value of Y is known. Substituting the values of byx and bxy, the regression equation above may also can be written as $\frac{y-\bar{y}}{6\ y}$ = r {$\frac{x-\bar{x}}{6x}$} and {$\frac{x-\bar{x}}{6x}$} = r {$\frac{y-\bar{y}}{6\ y}$}, respectively;
It will be noticed that both regression quotation are satisfied, when put X = X̄ and Y = Ȳ. Geometrically, this implies that both the regression lines pass through, and hence interest at, the point. The slopes of th regression lines are respectively byx and 1/bxy.
The two regression lines are usually different, however, since they must always interest at the point (X̄ - Ȳ), the line will coincide when their slopes are equal, i.e.

Byx = 1/bxy; or byx. Bxy = 1; i.e. or r = ±1

Where r = +1, we find that both regression equation reduce to $\boldsymbol{\frac{x-\bar{x}}{6x}}$**.** It is evident that the common regression line represented by this equation passes the point (X̄, Ȳ) with a positive slopes 6y/ 6x. Where r = -1, as in the previous case, the regression equation reduce to $\frac{y-\bar{y}}{6\ y}$ = r {$\frac{x-\bar{x}}{6x}$}

This common regression line passes through the point (X̄, Ȳ) with a negative slope (-6Y/ 6X) when r = 0, we find that regression equation reduce to $\frac{y-\bar{y}}{6\ y}$ = 0 and $\frac{\bar{\bar{x}}-\bar{x}}{6x}$. = 0 respectively.

That is the regression equation because Y = Ȳ and X = X̄. The regression line of y on X, y = ȳ, is now a straight line parallel to the X –axis; and the regression line of X on Y, as X = X̄ is parallel to Y –axis. The regression line are thus perpendicular to each other, nevertheless interacting at (X̄, Ȳ) as usual since now the regression equation of Y on X does not contain X, similarly, the regression equation of X on Y cannot be used to estimate X on the basis of Y.

Properties of Linear Regression

(1) There are two linear regression equations.

(i) Regression equation of Y on X; $Y = \bar{Y} = byx\ (X - \bar{X})$

(ii) Regression equation of X on Y; $X = \bar{X} = bxy\ (Y - \bar{Y})$

Where byx and bxy are respectively the regression coefficient of y on X and the regression coefficient of X on Y; $byx = \frac{Cor(x,y)}{6x2}. = r\ \frac{6y}{6x}$; $bxy = \frac{cor\ (x,y)}{6y2} = r\ \frac{6x}{6y}$.

(2) The product of the two-regression coefficient is equal to the same square of correlation coefficient. $Byx – Bxy = r^2$

(3) R, byx and bxy, all have the same sign, if the correlation coefficient is zero; the regression coefficient byx and bxy are also zero.

(4) The regression lines always interested at the point ($\bar{X}$, $\bar{Y}$), the slopes of the regression lines of Y on X and the regression line of X on Y are respectively bxy and 1/bxy.

(5) The angel between the two regression lines depends on the correlation coefficient r. when r = +1 or r = -1, they coincide. As r increases numerically from 0 to 1, the angle between the regression lines diminishes from 90^0 to 0^0.

(6) The two-regression equation are usually different, however when r= +1, they became identical and in this case; there is an exact linear relationship between the variables.

When r = 0, the regression equation reduce to $Y=\bar{Y}$ and $X = \bar{X}$, and neither Y nor X can be estimated from linear regression equations.

Example (1)

Find the regression of X on Y from the following data:

$\sum x = 24$, $\sum y = 44$, $\sum x\, y = 306$, $\sum x^2 = 164$ $\sum y^2 = 574$, n = 4

Find regression estimated value at X, when Y = 6.

Solution:

The regression equation of X on Y is; $X - \bar{X} = bxy\ (Y - \bar{Y})$ where $bxy = \frac{cor\ (x,y)}{6y2}$; hence

$\bar{X} = \frac{\sum x}{n} = \frac{24}{4} = 6$; $\bar{Y} = \frac{\sum y}{n} = \frac{44}{4} = 11$; $cor(x, y) = \frac{\sum xy}{n} – (\frac{\sum x}{n}) = \frac{306}{4} – (\frac{24}{4})\ (\frac{44}{4}) = 10.5$

Also: $6y^2 = \frac{574}{4} – (\frac{44}{4})^2 = 22.5$ **$:-\ bxy = \frac{10.5}{22.5} = 0.467$**

The regression equation of X on Y is then X – 6 = 0.467 (Y – 11); X = 0.467Y + 0.86

Where Y = 6, we have X = 0.467 X 6 + 0.86 = 3.7

Ans: X = 0.467Y + 0.86 = 3.7

Example (2)

The following data show the minimum temperature (degrees centigrade) on certain day ten states located throughout in South Sudan.

STATES	**U.N.S**	**U.S**	**N.B**	**W**	**W.B**	**W.E**	**C.E**	**J.S**	**L.S**	**E**
MAX TEMP.	29	23	25	15	27	29	24	31	22	28
MINI TEMP.	8	3	7	5	8	19	10	7	5	11

It is known that for high temperature given data:

n = 10 $\sum x^2 = 6595$, $\sum y^2 = 867$, $\sum x\, y = 2193$

Draw a scatter diagram and show regression line of Y on X there on.

Solutions:

With given value of (X, Y) as (29, 8), (23, 3) ………… points are plotted on a graph paper treating the pairs of value as co-ordinates. From the given data:

N= 10, $\sum x = 253$, $\sum y = 83$,

$\sum x^2 = 6595$, $\sum y^2 = 867$, $\sum xy = 2193$

With the results it is possible to find the regression equation of Y on X as Y - $\bar{Y}$ = bxy ($\bar{X}$ - $\bar{X}$), proceeding as in the previous example. $\bar{X}$ = 25.3, $\bar{Y}$ = 8.3 (cor(x, y) = 9.31, $6X^2$ = 19.41. hence bxy$= \frac{9.31}{19.41}$ = 0.48, and the regression equation is y – 8.3 = 0.48 (x – 25.3) or y = 0.48x – 3. 84 in order to show this line on the scatter diagram any two points satisfying this equations are points (20,5.76) and (30,10.56) will be the required regression line.

Example (3):

From the following results, obtain the two-regression equation and estimate the yield of crops when the rainfall is 22 Cms and the rainfall when the yield is 600 Kgs.

Subject	**Y (yield in Kg)**	**X (rainfall in Cm)**
Mean	508.4	26.7
S.D	36.8	4.6

Coefficient of correlation between yield and rainfall = 0.52

Solutions: for estimating yield (y) we have to use the regression equation of y on x, and for rainfall of y on x, (x) the regression equation of x on y.

Given $\bar{X}$ = 508.4, 6x = 4.6, 6y = 36.8, r = 0.52.

Byx = 0.52, X = $\frac{36.8}{46}$ = 4.16;   bxy = 0.52, X = $\frac{4.6}{36.8}$ = 0.65

The regression equation are therefore, y = 508.4 = 4.16 (x – 26.7) and x – 26.7 = 0.065 (y – 508.4) respectively which simplify to; y = 4.16x + 397.33 and x = 0.065y – 6.346; when x = 22, y = 4.16 × 22 + 397.33 = 488.8kg.

When y = 600, x = 0.065 × 600 – 6.346 = 32.7cm

Example (4) the following results were obtained from records ages, of husband (x) and stable blood pressure of wife (y) of a group of 10 moments.

Subject	X	Y
Mean	53	142
Variance	130	162

$\sum(x - \bar{x})$ (y - ȳ) = 1220

Find the appropriate regression equation and use it to estimate the blood pressure of a women whose age is (45)

Solution:

The appropriate equation is the pressure equation of Y and X as Y - $\bar{Y}$ = byx (X - $\bar{X}$),

Where byx = $\frac{cor\ (x,y)}{6\ X\ 2}$ but, cor(X, Y) = $\sum(x - \bar{x})$ (y - ȳ)/n = $\frac{1220}{10}$ = 122

So, that byx, = $\frac{1220}{130}$ = 0.94; the regression equation is therefore Y – 142 = 0.94 (X – 53) or,

Y = 0.94x + 92.18. When X = 45, Y = 0.94x +92.18 = 134.5

Ans: Y = 0.94x + 92.18.

Example (5)

X	41	45	50	68	47	77	90	100	80	100	40	43

Marks obtained by 12 students in the department of statistic and demography, faculty of economic and social studies in the test of statistics (x) and the test of accounting (y) are as follows. What is your estimate of the marks a student could have obtained in the accounting test if he or she obtained 60 in the test of statistic, but was he or she ill at the time of the test of accounting?

Solution:

For estimating Y, we use the regression equation of Y on X as X - $\bar{Y}$ = b_{yx} (X - $\bar{X}$); where

Y	60	63	60	48	85	56	53	91	74	95	65	43

X	Y	X = X – 60	Y = Y – 60	X^2	XY
41	60	-19	0	361	0
45	63	-15	3	225	-45
50	60	-10	0	100	0
68	48	8	-12	64	-96
47	85	-13	25	169	-132
77	56	17	-4	289	-68
90	53	30	-7	900	-210
100	91	40	31	1600	1240
80	74	20	14	400	280
100	98	40	38	1600	1520
40	65	-20	-17	400	-100
43	43	-17	-17	289	289
781	**796**	**61**	**76**	**6397**	**2485**

$b_{yx} = \frac{cor\ (x,y)}{6\ X\ 2}$. Therefore, we have to find $\bar{X}$, $\bar{Y}$, cor (X, Y) and $6X\ ^2$.

X	Y	X = X – 60	Y = Y – 60	X^2	XY
41	60	-19	0	361	0
45	63	-15	3	225	-45
50	60	-10	0	100	0
68	48	8	-12	64	-96
47	85	-13	25	169	-132
77	56	17	-4	289	-68
90	53	30	-7	900	-210
100	91	40	31	1600	1240
80	74	20	14	400	280
100	98	40	38	1600	1520
40	65	-20	-17	400	-100
43	43	-17	-17	289	289
781	**796**	**61**	**76**	**6397**	**2485**

Table: calculations for regression:

$X = \frac{781}{12} = \frac{\sum x}{n}$, $Y = \frac{\sum y}{n} = \frac{796}{12} = 66.33$; since the variable and covariance are unaffected by the changes of origin, $\text{Cov}(x,y) = \text{cov}(x,y) = \frac{\sum xx}{n} - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)$

$= \frac{2485}{12} - \left(\frac{61}{12}\right)\left(\frac{76}{12}\right) = \frac{25184}{144} =$ **174. 888; $6x^2$**$= \frac{\sum x}{n} - \left(\frac{\sum x}{n}\right)^2 - \frac{6397}{12} - \left(\frac{61}{12}\right)^2 = \frac{7304}{144}$ **= 50.722**

Therefore, $b_{yx} = \frac{25184/144}{73043/144} = \frac{25184}{73043}$ **= 0.345. The regression equation is y – y = byx ($\bar{x}$ _ x).**

= y _ 66.33 = 0.345 (x _ 65.08); i.e. y = 43.88 + 0.345x; when x = 60, y = 43.88+0.345×60 = 64.58 or 65

Example:

Derive the regression line, which you consider more important from the following series of observations:

Output in thousands	**5**	**7**	**9**	**11**	**13**	**15**
Profit per unit of output in (SSP)	**1.70**	**2.40**	**2.80**	**3.40**	**3.70**	**4.40**

Solution:

Let x represent output (thousands), and y represents profit per unit of output in SSP. It is considered that the regression line of y on x is more important, because, using the equation it will possible to

Find the expected profit based on output of the firm.

NOTE: [Her successive value of x are equidistant, Also since the number of pairs of observation is even, as 6, we apply the transformation: U = x - $\frac{[\textit{mean of two central values of } x]}{\frac{1}{2}[\textit{common different}]}$ for simplifying the calculations]

Table: 2 Calculation for regressions.

X	**Y**	**U = x – 10**	**U^2**	**UY**
5	1.70	-5	25	-8.50
7	2.40	-3	9	-7.20
9	2.80	-1	1	-2.80
11	3.40	+1	1	3.40
13	3.70	3	9	11.10
15	4.40	5	25	22.00
60	**18.40**	**0**	**70**	**18.00**

We are proceeding like previous example, $\bar{X} = \frac{60}{6} = 10$, $\bar{Y} = \frac{18.40}{6} = 3.06$

Cor (U,Y) = $\frac{\sum uy}{n} - (\frac{\sum u}{n})(\frac{\sum y}{n}) \frac{18.00}{6} - (\frac{0}{6}) \frac{18.40}{6} = 3.00$

$6X^2 = 6u^2 = \frac{\sum u2}{n} - \frac{\sum x2}{n} \frac{70}{6} - (\frac{0}{6})^2 = \frac{70}{6} = 11.6$;   byx = $\frac{18.00}{70} = 0.257$

The regression equation of y on x is; Y - $\bar{Y}$ = byx (X - $\bar{X}$); i.e. y = 3.06 = 0.257 (X – 10) or Y = 0.257x + 0.50.

Example: Find the equation of the line of the regression of x on y for the following data.

X	1.0	1.5	2.0	2.5	2.0	3.5	4.0
Y	5.3	5.7	6.3	7.2	8.2	8.7	8.4

Solution: the regression equation of X on Y is: - X - $\bar{X}$ = bxy (Y - $\bar{Y}$), where bxy = $\frac{cor\ (x,y)}{6y2}$ for simplifying the calculations, let us make a change of origin and scale for both the variances when n is add, u = $\frac{x-2.5}{0.5}$, V = $\frac{y-7.0}{0.1}$, U = X $\frac{central\ values\ of\ x}{common\ different}$.

Note; here, the values of x are equidistant and n s add so, we use the transformation as:

$\bar{X} = \frac{17}{7} = 2.5$, $\bar{Y} = \frac{49.8}{7} = 7.11$;   cor (u, v) = $\frac{\sum uv}{n} - (\frac{\sum u}{n})(\frac{\sum v}{n}) = \frac{172}{7} - (\frac{0}{7})(\frac{8}{7}) = \frac{172}{7} = 24.57$

$6r^2 = \frac{\sum uv2}{n} - (\frac{\sum u}{n})^2 \frac{1440}{7} - (\frac{8}{7})^2 = \frac{7916}{49} = 161.61$

Using the covariance of two different values; cor (x, y) = d.d' cor (u, v) = (0.5) (0.1) ($\frac{172}{7}$)

Table: Calculations for regression.

X	**Y**	**U**	**V**	**V^2**	**UV**
1.0	5.3	-3	-17	289	51
1.5	5.7	-2	-13	169	26
2.0	6.3	-1	-7	49	7
2.5	7.2	0	2	4	0
3.0	8.2	1	12	144	12
3.5	8.7	2	17	289	34
4.0	8.4	3	14	196	42
17.5	49.8	0	8	1140	172

Using this formulae $6y^2 = d^2.6^2v = (0.1)^2$ (7916/49)

Hence, bxy = $\mathrm{cor}\,\frac{(x,y)}{6y2} = \frac{0.5\ X\ 0.1\ X\ (\frac{172}{2}}{(0.1)2\ X\ (\frac{7916}{49})}$ **= 0.76;**

Therefore, the regression equation of X on Y is; X – 2.5 = 0.76 (y – 7.11); i.e. X = 0.76y – 2.90.

RANK CORRELATION:

The product – moment correlation coefficient (r) is calculated by using "values" of the variables, but many situations arise in which either precise measurement are not a variable, or the characters cannot be measured at all.

Example:

In order to find the extent of association between intelligence and efficiency in salesmanship for a group of salesman, we could attempts to measure the two qualities by allotting marks to each salesman. However, this method is open t many objections, and an exact measurement of the two qualities is not all possible.

Difficulties of this type however disappear if we arrange the individuals in order of merit of proficiency in the procession of the qualities, using numbers 1,2,3 ……… the individual are then said to be ranked and the number allotted to a particular candidate is called his rank. For each individual, we then have a pair of ranks, one in each character. The correlation coefficient between the two series of the ranks is called: Rank correlation coefficient.

It is given by the formula, R = 1 - $\frac{6\sum d2}{n3-n}$Where d represents the difference of the ranks of an individual in the two characters and n is the number of an individuals.

This formula is also known as spearman's formula for rank correlation coefficient. The rank correlation coefficient lies between -1 and +1. $-1 \leq R \leq +1$

It has the maximum value -1 when the ranks are just the opposite, i.e. individual with ranks 1,2,3…..n in one character have n, n – 1 ……3, 2, 1 in the other.

In the other calculations of the rank correlation coefficient from the given scores if several individual have the same score in any character, they must be allotted the same ranks and we are then concerned with what are known as tied ranks. In dealing with such cases, the usual way is to allot the average ranks to each of these individuals, and then calculates the products moment correlation coefficient from these ranks. However, in such case one way of correcting formula is to increase $\sum d^2$ by $\frac{(t3-t)}{12}$ in respect of each tie where t denotes the number of individuals involved in a tie whether in the first or second series. The modified formula for ranks correlation coefficient when there are ties, is then R' = $1 - 6\ \{\frac{(t3-t)_2}{12}\} + \sum d\,\frac{(t3-t)/\ 12)}{n2-1}$.

Using the following points:

1. The rank correlation coefficient R is used as a measurement of the degree of association between two attributes, where measurement on the characters are not available, but it is possible to rank the individual in some order without difficulty. (Note that the product-

moment correlation coefficient r is a measure of the degree of association between two variances). It is therefore applied in vocational and applied psychology for finding the degree of correspondence between abilities.

2. It is used in cases where exact or reliable measurement are not available. In educational tests it is known to all that the usual method of allotting the marks to the candidates, the product-moment correlation coefficient between the abilities, however, if the candidates are marked, the disappearance in the marks given by examiners would not seriously affect the value of rank correlation coefficient.
3. Even what exact measurements are available, the rank correlation coefficient R provides a quick estimate of the degree of association between the variances. The laborious calculations for the product-moment correlation coefficient r are replaced by ranking the individual and the using previous formula.

R' = 1 – 6 {$\sum d^2 + \sum d \frac{(t3-t)/12}{n2-1}$.

Example:

In a contest two judges ranked seven candidates in order of their references as in the following titles.

Candidates	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Ranks by Judges 1	2	1	4	5	3	7	6
Ranks by Judges 2	3	4	2	5	1	6	7

Calculate the rank correlation coefficient.

Solutions:

Table; Calculations for rank correlation coefficient.

Candidates	**Judge 1 X**	**Judge 2 Y**	**D = X – Y**	**D^2**
A	2	3	-1	1
B	1	4	-3	9
C	4	2	2	4
D	5	5	0	0
E	3	1	2	4
F	7	6	1	1
G	6	7	-1	1
Total	**-**	**-**	**0**	**20**

Here n = 7 ……= 20

R = 1 – 6 $\boldsymbol{\sum d \frac{(td2- =6\ X\ 20)}{7-7}}$ **= 0.64**

Example: In a certain examination, 20 students obtained the following marks in mathematics and physics. Find spearman, rank correlation coefficient.

St68udents (roll No)	1	2	3	4	5	6	7	8	9	10
Marks in mathematics	90	30	82	45	32	65	40	88	73	66
Marks in Physics	85	43	75	68	45	63	60	90	62	58

Solutions:

We have to obtain the ranks of students in the two subjects and calculate the rank correlation coefficient; in mathematics the students with Roll No.1 get the highest marks 90 and its given rank 1. Roll No. 8 with 88 marks has rank 2, and so on.

Table: Calculations for rank correlation coefficient.

Roll No. (1)	**Mathematics (2) X**	**(3) X**	**Physics (4) X**	**(5) X**	**D = X – Y (6)**	**D^2 (7)**
1	90	1	85	2	-1	1
2	30	10	42	10	0	0
3	82	3	75	3	0	0
4	45	7	68	4	3	9
5	32	9	63	5	1	1
6	65	6	60	7	1	1
8	88	2	90	1	1	1
9	73	4	62	6	-2	4
10	66	5	58	8	-3	9
Total	-	-	-	-	**0**	**26**

Applying this formula:

R = 1 - ……. = R – 1 - $\frac{6 X 26}{103-10}$ **= 1 -** $\frac{165}{990}$ **= 1 – 0.16666 = 0.833**

Note: it is interesting to note that spearman's rank correlation coefficient R = 0.84 calculate above is in good agreement with person's product moment correlation coefficient is r = 0.86 calculated from actual marks.

Example: deduce spearman's formula for rank correlation coefficient.

Solution:

Let (X, Y), (X_2, Y_2) …………… (X_n, Y_n) denote the rank of an individual in the two characters; spearman's rank correlation coefficient is the product moment correlation coefficient between these ranks treating them as values of variances.

It may be noted that X_1, X_2, …………X_n being ranks (and not the actual values) are only the numbers 1, 2 …….; arranged in some order, similarly (Y_1, Y_2….Y_n) are also the same numbers, but possibly in different order.

Since the x-series and the y-series comprise the same numbers, their means and variance will be equal, therefore $\bar{X} = \bar{Y} = \frac{n+1}{2}$;

Conclusion

The book is the written in sample words with combination of fundamental and principle of statistics applications, wide dentitions of statistics was introducing with the presentation of the data in graphs and table. Based on that outcome of this course or study at the end, student will be able to differentiate between single variable or more under the specific objectives, with the simple findings have special bias of positivity towards the link between statistics management and trade and business promotion in used of collected statistical tool in multivariate analysis, as per ANOVA ONE way or ANOVA TWO ways. These results reveal that there is significant relationship between the role of statistics on the one hand and trade and business promotion on the other hand, and that will lead you to take the right decision. Therefore, the importance course of statistics vehemently concludes that the role of statistics in daily life of the businesspeople, especially in economics, planning, agriculture, medicine and health science's trade and business promotion since statistics aids business decision making for short- or long-term policy intervention. Business promotion mix can be optimized with sound statistics that directs the trends of the business and so it is important to have sound and accurate statistics available when it comes to conducting trade and business promotion.

Reviewers:

1. Dr. David Abul, PhD, in Statistics and Population Studies, Kenyatta University – Nairobi Assistant Professor of Statistics in the Faculty of Human Development, Upper Nile University – Juba/ Malakal.
   Contact: +21122350036
   Email: davidabol36@gmail.com
2. Dr. Charles Tito, PhD in Mathematics' Assistant Professor of Mathematics and Dean of the Faculty of Education, Upper Nile University –Juba/ Malakal.
   Email: charlesthpo2013@gmail.com +211915580122 / +21124362414
3. Dr. Alex Yoro, PhD in Mathematical Statistics from Darussalam University, Former Dean of Post-graduate Upper Nile University – Juba/ Malakal.

BIBLIOGRAPHY.

1- American Statistical Association (2014). Curriculum guidelines for undergraduate programs in statistical science

2- Anthony O'Hagan, Kendall's Advanced Theory of Statistics, Vol. 2B (1994), Chapter 4.

3- Bethea, R.M., Duran, B.S., & Boullion, T.L. (1985). Statistical methods for engineers and scientists. New York: Marcel Dekker.

4- Box, G.E.P., Hunter, W.G., & Hunter, J.S. (1978). Statistics for experimenters, an introduction to design, data analysis and model building. New York: John Wiley & Sons.

5- Cobanovic, K., Nikolic-Djoric, E., & Mutavdzic, B. (1997). The Role and the importance of the application of statistical methods in agricultural investigations. Scientific Meeting with International Participation: "Current State Outlook of the Development of Agriculture and the Role of Agri-Economic Science and Profession" (Volume 26, pp. 457-470). Novi Sad: Agrieconomica.

6- Comprehensive Agriculture Master Plan, 2018-2025. (CAMP) South Sudan National Youth Policy, conclusions about specific characteristics of population based on sample information. P, 123, published by Japanese International Cooperation Agency, (JICA) in Juba- South Sudan.

7- D K Shangodoyin and D A Agunbiade. (1999). Fundamentals of Statistics and Database Management. Rasmed Publications, ISBN 978-34610-5-2.

8- Daniel Wavne W (1983). Business Statistic: Basic concepts and Methodology 3rd ed. Dallas Geneva: Houghton Mifflin Co.

9- Dodge, Y. (2006) The Oxford Dictionary of Statistical Terms, Oxford University Press. ISBN 0- 19-920613-9 Lund Research Ltd. "Descriptive and Inferential Statistics". Statistics.laerd.com. Retrieved 2014-03-23.

10- G. J. Feldman and R. D. Cousins, Phys Rev D57 (1998) 3873

11- Garson, G. D. (2012). Testing statistical assumptions. Asheboro, NC: Statistical Associates Publishing.

12- Govindarajulu, Zakkula (1999), Elements of Sampling Theory and Methods Prentice-Hall, New Jersey.

13- Hadzivukovic, S. (1991). Teaching Statistics to Students in Agriculture. In D. Vere-Jones (Ed.), Proceedings of the 3rd International Conference on Teaching Statistics (Volume1, pp. 399-401). Dunedin.

14- J. Neyman and E. S. Pearson, Phil. Trans. R. Soc. Ser. A 231 (1933) 289-337, reprinted in Breakthroughs in Statistics, Vol. 1, Kotz and Johnson eds., Springer 1992.

15- J. Neyman, Phil. Trans. R. *Soc.* Ser. A 236 (1937) 333, reprinted in A Selection of Early Statistical Papers on J. Neyman, Univ. of Cal. Press, Berkeley, 1967.

16- John Wiley, 2004. Biostatistics a Methodology for the Health Science, Second Edition.

17- Karl Pearson, Phil. Mag. Ser. 5 (1900) 157-175, reprinted in Breakthroughs in Statistics, Vol. 2, Kotz and Johnson eds., Springer 1992.

18- Kendall’s Advanced Theory of Statistics: In the Fifth Edition (1991) the authors are Stuart and Ord, and this material is at the beginning of chapter 23 in Volume 2. In the Sixth Edition (1999) the authors are Stuart, Ord and Arnold, and this material appears at the beginning of chapter 22 in Volume 2A.

19- Little, T.M., & Hills, F.J. (1975). Statistical methods in agricultural research. University of California.

20- Masembe Kabali 2011. Basic Business Statistics, third edition, P.O. Box 7453 Kapala – Uganda. Type set printed and Bound by M. Kaabali and Basic Business Statistics Books, p 1.

21- McLean, A. (1999). The predictive approach to teaching statistics. Working Paper 4/99, Monash University, Department of Econometrics and Business Statistics, Australia.

22- N. G. Das 2009. Statistical Method the combined Edition (Volumes I & II). McGraw Hill Education (India) Private Limited, p, 1.

23- Phillips, B., & Jones, P. (1991). Developing statistical concepts for engineering students using computer packages. In D. Vere-Jones (Ed.), Proceedings of the 3rd International Conference on Teaching Statistics (Volume 2, pp. 255-260). Dunedin.

24- Poulino F. D. Ajang, 2015, Lecture Note for Probability and Statistics for Second year undergraduate, Upper Nile University, p.69.

25- S. Ciampolillo, Il Nuovo Cimento 111 (1998) 1415

26- Satyabrata Pal: *2010,* Statistics (I) for BBA Students p.52. as per West Bangal University. India, New Age International Publishers.

27- Spiegel Murray, 1981. Theory and Problems of Statistic. New York: Mc Graw-Hill Book Co.

28- Steel, R.G.D., & Torrie, J.H. (1960). Principles and procedures of statistics. New York: McGraw Hill.

29- Summers George W., 1981. Student Supplement for basic statistics in business and economics 3rd ed. Belmont, Califomia: Wadsworth publishing Co.

30- The theoretical F values are the values from found from F-Tables, t-Tables and Z-Tables.

31- Townsed and Burke J. Paul, 1995. Using Statistics in Classroom instruction. New York: Macmillian Publishing Co., Inc. London: Collier Macmillian Publishers.

32- Wonnacott Tomas H., 1977. Introductory Statistic. 3rded. New York: John Wiley and Sons.