MASTER OF BUSINESS ADMINISTRATION

MBA-102

BUSINESS STATISTICS

Directorate of Distance Education
Guru Jambheshwar University of Science & Technology
HISAR-125001

CONTENTS

Lesson No.	Name of Topic	Page No.
1	An Introduction to Business Statistics	**3**
2	An Overview of Central Tendency	**17**
3	Dispersion and Skewness	**46**
4	Probability Theory	**80**
5	Probability Distributions-I	**113**
6	Probability Distributions-II	**138**
7	Sampling and Sampling Methods	**161**
8	Sampling Distributions	**182**
9	Statistical Estimation	**232**
10	Testing of Hypotheses	**254**
11	Non-Parametric Tests	**296**
12	Correlation Analysis	**320**
13	Regression Analysis	**358**
14	Analysis of Time Series	**404**
15	Index Number	**453**
16	Statistical Quality Control	**497**

Subject: Business Statistics	
Course code: MBA-102	**Author: Surinder Kundu**
Lesson: 01	**Vetter: Dr. B. S. Bodla**
AN INTRODUCTION TO BUSINESS STATISTICS	

STRUCTURE

1.1 Learning objective

1.2 Introduction

1.2.1 Meaning and Definitions of Statistics

1.2.2 Types of Data and Data Sources

1.2.3 Types of Statistics

1.2.4 Scope of Statistics

1.3 Importance of Statistics in Business

1.4 Limitations of statistics

1.5 Check your progress

1.6 Summary

1.7 Keywords

1.8 Self-Assessment Test

1.9 Answers to check your progress

1.10 References/Suggested Readings

1.1 LEARNING OBJECTIVE

After going through this lesson, you will be able to

- Understand the meaning of Statistics
- Explain the definition and nature of statistics
- Explain the importance of statistics in different fields and
- Explain the limitations of statistics.

1.2 INTRODUCTION

"A knowledge of statistics is like a knowledge of foreign language of algebra; it may prove of use at any time under any circumstance"- Bowley.

For a layman, 'Statistics' means numerical information expressed in quantitative terms. This information may relate to objects, subjects, activities, phenomena, or regions of space. As a matter of fact, data have no limits as to their reference, coverage, and scope. At the macro level, these are data on gross national product and shares of agriculture, manufacturing, and services in GDP (Gross Domestic Product). At the micro level, individual firms, howsoever small or large, produce extensive statistics on their operations. The annual reports of companies contain variety of data on sales, production, expenditure, inventories, capital employed, and other activities. These data are often field data, collected by employing scientific survey techniques. Unless regularly updated, such data are the product of a one-time effort and have limited use beyond the situation that may have called for their collection. A student knows statistics more intimately as a subject of study like economics, mathematics, chemistry, physics, and others. It is a discipline, which scientifically deals with data, and is often described as the science of data. In dealing with statistics as data, statistics has developed appropriate methods of collecting, presenting, summarizing, and analysing data, and thus consists of a body of these methods.

1.2.1 Meaning and definitions of Statistics

In the beginning, it may be noted that the word 'statistics' is used rather curiously in two senses plural and singular. In the plural sense, it refers to a set of figures or data. In the singular sense, statistics refers to the whole body of tools that are used to collect data, organise and interpret them and, finally, to draw conclusions from them. It should be noted that both the aspects of statistics are important if the quantitative data are to serve their purpose. If statistics, as a subject, is inadequate and consists of poor methodology, we could not know the right procedure to extract from the data the information they contain. Similarly, if our data are defective or that they are inadequate or inaccurate, we could not reach the right conclusions even though our subject is well developed.

A.L. Bowley has defined statistics as: (i) statistics is the science of counting, (ii) Statistics may rightly be called the science of averages, and (iii) statistics is the science of measurement of social organism regarded as a whole in all its manifestations. Boddington defined as: Statistics is the science of estimates and probabilities. Further, W.I. King has defined Statistics in a wider context, the science of

Statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis or enumeration or collection of estimates.

Seligman explored that statistics is a science that deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry. Spiegal defines statistics highlighting its role in decision-making particularly under uncertainty, as follows: statistics is concerned with scientific method for collecting, organising, summa rising, presenting and analyzing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis. According to Prof. Horace Secrist, Statistics is the aggregate of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose, and placed in relation to each other.

From the above definitions, we can highlight the major characteristics of statistics as follows:

(i) *Statistics are the aggregates of facts*. It means a single figure is not statistics. For example, national income of a country for a single year is not statistics but the same for two or more years is statistics.

(ii) *Statistics are affected by a number of factors.* For example, sale of a product depends on a number of factors such as its price, quality, competition, the income of the consumers, and so on.

(iii) *Statistics must be reasonably accurate.* Wrong figures, if analysed, will lead to erroneous conclusions. Hence, it is necessary that conclusions must be based on accurate figures.

(iv) *Statistics must be collected in a systematic manner.* If data are collected in a haphazard manner, they will not be reliable and will lead to misleading conclusions.

(v) *Collected in a systematic manner for a pre-determined purpose*

(vi) Lastly, Statistics should be placed in relation to each other. If one collects data unrelated to each other, then such data will be confusing and will not lead to any logical conclusions. Data should be comparable over time and over space.

1.2.2 Types of data and data sources

Statistical data are the basic raw material of statistics. Data may relate to an activity of our interest, a phenomenon, or a problem situation under study. They derive as a result of the process of measuring, counting and/or observing. Statistical data, therefore, refer to those aspects of a problem situation that can be measured, quantified, counted, or classified. Any object subject phenomenon, or activity that generates data through this process is termed as a variable. In other words, a variable is one that shows a

degree of variability when successive measurements are recorded. In statistics, data are classified into two broad categories: quantitative data and qualitative data. This classification is based on the kind of characteristics that are measured.

Quantitative data are those that can be quantified in definite units of measurement. These refer to characteristics whose successive measurements yield quantifiable observations. Depending on the nature of the variable observed for measurement, quantitative data can be further categorized as continuous and discrete data. Obviously, a variable may be a continuous variable or a discrete variable.

(i) **Continuous data** represent the numerical values of a continuous variable. A continuous variable is the one that can assume any value between any two points on a line segment, thus representing an interval of values. The values are quite precise and close to each other, yet distinguishably different. All characteristics such as weight, length, height, thickness, velocity, temperature, tensile strength, etc., represent continuous variables. Thus, the data recorded on these and similar other characteristics are called continuous data. It may be noted that a continuous variable assumes the finest unit of measurement. Finest in the sense that it enables measurements to the maximum degree of precision.

(ii) **Discrete data** are the values assumed by a discrete variable. A discrete variable is the one whose outcomes are measured in fixed numbers. Such data are essentially count data. These are derived from a process of counting, such as the number of items possessing or not possessing a certain characteristic. The number of customers visiting a departmental store everyday, the incoming flights at an airport, and the defective items in a consignment received for sale, are all examples of discrete data.

Qualitative data refer to qualitative characteristics of a subject or an object. A characteristic is qualitative in nature when its observations are defined and noted in terms of the presence or absence of a certain attribute in discrete numbers. These data are further classified as nominal and rank data.

(i) **Nominal data** are the outcome of classification into two or more categories of items or units comprising a sample or a population according to some quality characteristic. Classification of students according to sex (as males and females), of workers according to skill (as skilled, semi-skilled, and unskilled), and of employees according to the level of education (as matriculates, undergraduates, and post-graduates), all result into nominal data. Given any such basis of classification, it is always possible to assign each item to a particular class and make a summation

of items belonging to each class. The count data so obtained are called nominal data.

(ii) Rank data, on the other hand, are the result of assigning ranks to specify order in terms of the integers 1,2,3, ..., n. Ranks may be assigned according to the level of performance in a test. a contest, a competition, an interview, or a show. The candidates appearing in an interview, for example, may be assigned ranks in integers ranging from I to n, depending on their performance in the interview. Ranks so assigned can be viewed as the continuous values of a variable involving performance as the quality characteristic.

Data sources could be seen as of two types, viz., secondary and primary. The two can be defined as under:

(i) **Secondary data:** They already exist in some form: published or unpublished - in an identifiable secondary source. They are, generally, available from published source(s), though not necessarily in the form actually required.

(ii) **Primary data:** Those data which do not already exist in any form, and thus have to be collected for the first time from the primary source(s). By their very nature, these data require fresh and first-time collection covering the whole population or a sample drawn from it.

1.2.3 Types of statistics

There are two major divisions of statistics such as descriptive statistics and inferential statistics. The term **descriptive statistics** deals with collecting, summarizing, and simplifying data, which are otherwise quite unwieldy and voluminous. It seeks to achieve this in a manner that meaningful conclusions can be readily drawn from the data. Descriptive statistics may thus be seen as comprising methods of bringing out and highlighting the latent characteristics present in a set of numerical data. It not only facilitates an understanding of the data and systematic reporting thereof in a manner; and also makes them amenable to further discussion, analysis, and interpretations.

The first step in any scientific inquiry is to collect data relevant to the problem in hand. When the inquiry relates to physical and/or biological sciences, data collection is normally an integral part of the experiment itself. In fact, the very manner in which an experiment is designed, determines the kind of data it would require and/or generate. The problem of identifying the nature and the kind of the relevant data is thus automatically resolved as soon as the design of experiment is finalized. It is possible in the case of physical sciences. In the case of social sciences, where the required data are often collected through a questionnaire from a number of carefully selected respondents, the problem is not that simply

resolved. For one thing, designing the questionnaire itself is a critical initial problem. For another, the number of respondents to be accessed for data collection and the criteria for selecting them has their own implications and importance for the quality of results obtained. Further, the data have been collected; these are assembled, organized, and presented in the form of appropriate tables to make them readable. Wherever needed, figures, diagrams, charts, and graphs are also used for better presentation of the data. A useful tabular and graphic presentation of data will require that the raw data be properly classified in accordance with the objectives of investigation and the relational analysis to be carried out.

A well thought-out and sharp data classification facilitates easy description of the hidden data characteristics by means of a variety of summary measures. These include measures of central tendency, dispersion, skewness, and kurtosis, which constitute the essential scope of descriptive statistics. These form a large part of the subject matter of any basic textbook on the subject, and thus they are being discussed in that order here as well.

Inferential statistics, also known as inductive statistics, goes beyond describing a given problem situation by means of collecting, summarizing, and meaningfully presenting the related data. Instead, it consists of methods that are used for drawing inferences, or making broad generalizations, about a totality of observations on the basis of knowledge about a part of that totality. The totality of observations about which an inference may be drawn, or a generalization made, is called a population or a universe. The part of totality, which is observed for data collection and analysis to gain knowledge about the population, is called a sample.

The desired information about a given population of our interest; may also be collected even by observing all the units comprising the population. This total coverage is called census. Getting the desired value for the population through census is not always feasible and practical for various reasons. Apart from time and money considerations making the census operations prohibitive, observing each individual unit of the population with reference to any data characteristic may at times involve even destructive testing. In such cases, obviously, the only recourse available is to employ the partial or incomplete information gathered through a sample for the purpose. This is precisely what inferential statistics does. Thus, obtaining a particular value from the sample information and using it for drawing an inference about the entire population underlies the subject matter of inferential statistics. Consider a situation in which one is required to know the average body weight of all the college students in a given cosmopolitan city during a certain year. A quick and easy way to do this is to record the weight of only

500 students, from out of a total strength of, say, 10000, or an unknown total strength, take the average, and use this average based on incomplete weight data to represent the average body weight of all the college students. In a different situation, one may have to repeat this exercise for some future year and use the quick estimate of average body weight for a comparison. This may be needed, for example, to decide whether the weight of the college students has undergone a significant change over the years compared.

Inferential statistics helps to evaluate the risks involved in reaching inferences or generalizations about an unknown population on the basis of sample information. For example, an inspection of a sample of five battery cells drawn from a given lot may reveal that all the five cells are in perfectly good condition. This information may be used to conclude that the entire lot is good enough to buy or not.

Since this inference is based on the examination of a sample of limited number of cells, it is equally likely that all the cells in the lot are not in order. It is also possible that all the items that may be included in the sample are unsatisfactory. This may be used to conclude that the entire lot is of unsatisfactory quality, whereas the fact may indeed be otherwise. It may, thus, be noticed that there is always a risk of an inference about a population being incorrect when based on the knowledge of a limited sample. The rescue in such situations lies in evaluating such risks. For this, statistics provides the necessary methods. These centres on quantifying in probabilistic term the chances of decisions taken on the basis of sample information being incorrect. This requires an understanding of the what, why, and how of probability and probability distributions to equip ourselves with methods of drawing statistical inferences and estimating the degree of reliability of these inferences.

1.2.4 Scope of statistics

Apart from the methods comprising the scope of descriptive and inferential branches of statistics, statistics also consists of methods of dealing with a few other issues of specific nature. Since these methods are essentially descriptive in nature, they have been discussed here as part of the descriptive statistics. These are mainly concerned with the following:

(i) It often becomes necessary to examine how two paired data sets are related. For example, we may have data on the sales of a product and the expenditure incurred on its advertisement for a specified number of years. Given that sales and advertisement expenditure are related to each other, it is useful to examine the nature of relationship between the two and quantify the degree of that relationship. As this requires use of appropriate statistical methods, these falls under the purview of

what we call regression and correlation analysis.

(ii) Situations occur quite often when we require averaging (or totalling) of data on prices and/or quantities expressed in different units of measurement. For example, price of cloth may be quoted per meter of length and that of wheat per kilogram of weight. Since ordinary methods of totalling and averaging do not apply to such price/quantity data, special techniques needed for the purpose are developed under index numbers.

(iii) Many a time, it becomes necessary to examine the past performance of an activity with a view to determining its future behaviour. For example, when engaged in the production of a commodity, monthly product sales are an important measure of evaluating performance. This requires compilation and analysis of relevant sales data over time. The more complex the activity, the more varied the data requirements. For profit maximising and future sales planning, forecast of likely sales growth rate is crucial. This needs careful collection and analysis of past sales data. All such concerns are taken care of under time series analysis.

(iv) Obtaining the most likely future estimates on any aspect(s) relating to a business or economic activity has indeed been engaging the minds of all concerned. This is particularly important when it relates to product sales and demand, which serve the necessary basis of production scheduling and planning. The regression, correlation, and time series analyses together help develop the basic methodology to do the needful. Thus, the study of methods and techniques of obtaining the likely estimates on business/economic variables comprises the scope of what we do under business forecasting.

Keeping in view the importance of inferential statistics, the scope of statistics may finally be restated as consisting of statistical methods which facilitate decision-making under conditions of uncertainty. While the term statistical methods is often used to cover the subject of statistics as a whole, in particular it refers to methods by which statistical data are analysed, interpreted, and the inferences drawn for decision-making.

Though generic in nature and versatile in their applications, statistical methods have come to be widely used, especially in all matters concerning business and economics. These are also being increasingly used in biology, medicine, agriculture, psychology, and education. The scope of application of these methods has started opening and expanding in a number of social science disciplines as well. Even a political scientist finds them of increasing relevance for examining the political behaviour and it is, of

course, no surprise to find even historians statistical data, for history is essentially past data presented in certain actual format.

1.3 IMPORTANCE OF STATISTICS IN BUSINESS

There are three major functions in any business enterprise in which the statistical methods are useful. These are as follows:

(i) The planning of operations: This may relate to either special projects or to the recurring activities of a firm over a specified period.

(ii) The setting up of standards: This may relate to the size of employment, volume of sales, fixation of quality norms for the manufactured product, norms for the daily output, and so forth.

(iii) The function of control: This involves comparison of actual production achieved against the norm or target set earlier. In case the production has fallen short of the target, it gives remedial measures so that such a deficiency does not occur again.

A worth noting point is that although these three functions-planning of operations, setting standards, and control-are separate, but in practice they are very much interrelated.

Different authors have highlighted the importance of Statistics in business. For instance, Croxton and Cowden give numerous uses of Statistics in business such as project planning, budgetary planning and control, inventory planning and control, quality control, marketing, production and personnel administration. Within these also they have specified certain areas where Statistics is very relevant. Another author, Irwing W. Burr, dealing with the place of statistics in an industrial organisation, specifies a number of areas where statistics is extremely useful. These are: customer wants and market research, development design and specification, purchasing, production, inspection, packaging and shipping, sales and complaints, inventory and maintenance, costs, management control, industrial engineering and research.

Statistical problems arising in the course of business operations are multitudinous. As such, one may do no more than highlight some of the more important ones to emphasis the relevance of statistics to the business world. In the sphere of production, for example, statistics can be useful in various ways.

Statistical quality control methods are used to ensure the production of quality goods. Identifying and rejecting defective or substandard goods achieve this. The sale targets can be fixed on the basis of sale forecasts, which are done by using varying methods of forecasting. Analysis of sales affected against the targets set earlier would indicate the deficiency in achievement, which may be on account of several

causes: (i) targets were too high and unrealistic (ii) salesmen's performance has been poor (iii) emergence of increase in competition (iv) poor quality of company's product, and so on. These factors can be further investigated.

Another sphere in business where statistical methods can be used is personnel management. Here, one is concerned with the fixation of wage rates, incentive norms and performance appraisal of individual employee. The concept of productivity is very relevant here. On the basis of measurement of productivity, the productivity bonus is awarded to the workers. Comparisons of wages and productivity are undertaken in order to ensure increases in industrial productivity.

Statistical methods could also be used to ascertain the efficacy of a certain product, say, medicine. For example, a pharmaceutical company has developed a new medicine in the treatment of bronchial asthma. Before launching it on commercial basis, it wants to ascertain the effectiveness of this medicine. It undertakes an experimentation involving the formation of two comparable groups of asthma patients. One group is given this new medicine for a specified period and the other one is treated with the usual medicines. Records are maintained for the two groups for the specified period. This record is then analysed to ascertain if there is any significant difference in the recovery of the two groups. If the difference is really significant statistically, the new medicine is commercially launched.

1.4 LIMITATIONS OF STATISTICS

Statistics has a number of limitations, pertinent among them are as follows:

(i) There are certain phenomena or concepts where statistics cannot be used. This is because these phenomena or concepts are not amenable to measurement. For example, beauty, intelligence, courage cannot be quantified. Statistics has no place in all such cases where quantification is not possible.

(ii) Statistics reveal the average behaviour, the normal or the general trend. An application of the 'average' concept if applied to an individual or a particular situation may lead to a wrong conclusion and sometimes may be disastrous. For example, one may be misguided when told that the average depth of a river from one bank to the other is four feet, when there may be some points in between where its depth is far more than four feet. On this understanding, one may enter those points having greater depth, which may be hazardous.

(iii) Since statistics are collected for a particular purpose, such data may not be relevant or useful in other situations or cases. For example, secondary data (i.e., data originally collected by someone else) may not be useful for the other person.

(iv) Statistics are not 100 per cent precise as is Mathematics or Accountancy. Those who use statistics should be aware of this limitation.

(v) In statistical surveys, sampling is generally used as it is not physically possible to cover all the units or elements comprising the universe. The results may not be appropriate as far as the universe is concerned. Moreover, different surveys based on the same size of sample but different sample units may yield different results.

(vi) At times, association or relationship between two or more variables is studied in statistics, but such a relationship does not indicate cause and effect' relationship. It simply shows the similarity or dissimilarity in the movement of the two variables. In such cases, it is the user who has to interpret the results carefully, pointing out the type of relationship obtained.

(vii) A major limitation of statistics is that it does not reveal all pertaining to a certain phenomenon. There is some background information that statistics does not cover. Similarly, there are some other aspects related to the problem on hand, which are also not covered. The user of Statistics has to be well informed and should interpret Statistics keeping in mind all other aspects having relevance on the given problem.

Apart from the limitations of statistics mentioned above, there are misuses of it. Many people, knowingly or unknowingly, use statistical data in wrong manner. Let us see what the main misuses of statistics are so that the same could be avoided when one has to use statistical data. The misuse of Statistics may take several forms some of which are explained below.

(i) **Sources of data not given:** At times, the source of data is not given. In the absence of the source, the reader does not know how far the data are reliable. Further, if he wants to refer to the original source, he is unable to do so.

(ii) **Defective data:** Another misuse is that sometimes one gives defective data. This may be done knowingly in order to defend one's position or to prove a particular point. This apart, the definition used to denote a certain phenomenon may be defective. For example, in case of data relating to unemployed persons, the definition may include even those who are employed, though partially.

The question here is how far it is justified to include partially employed persons amongst unemployed ones.

(iii) Unrepresentative sample: In statistics, several times one has to conduct a survey, which necessitates to choose a sample from the given population or universe. The sample may turn out to be unrepresentative of the universe. One may choose a sample just on the basis of convenience. He may collect the desired information from either his friends or nearby respondents in his neighbourhood even though such respondents do not constitute a representative sample.

(iv) Inadequate sample: Earlier, we have seen that a sample that is unrepresentative of the universe is a major misuse of statistics. This apart, at times one may conduct a survey based on an extremely inadequate sample. For example, in a city we may find that there are 1, 00,000 households. When we have to conduct a household survey, we may take a sample of merely 100 households comprising only 0.1 per cent of the universe. A survey based on such a small sample may not yield right information.

(v) Unfair Comparisons: An important misuse of statistics is making unfair comparisons from the data collected. For instance, one may construct an index of production choosing the base year where the production was much less. Then he may compare the subsequent year's production from this low base. Such a comparison will undoubtedly give a rosy picture of the production though in reality it is not so. Another source of unfair comparisons could be when one makes absolute comparisons instead of relative ones. An absolute comparison of two figures, say, of production or export, may show a good increase, but in relative terms it may turnout to be very negligible. Another example of unfair comparison is when the population in two cities is different, but a comparison of overall death rates and deaths by a particular disease is attempted. Such a comparison is wrong. Likewise, when data are not properly classified or when changes in the composition of population in the two years are not taken into consideration, comparisons of such data would be unfair as they would lead to misleading conclusions.

(vi) Unwanted conclusions: Another misuse of statistics may be on account of unwarranted conclusions. This may be as a result of making false assumptions. For example, while making projections of population in the next five years, one may assume a lower rate of growth though the past two years indicate otherwise. Sometimes one may not be sure about the changes in business environment in the near future. In such a case, one may use an assumption that may turn out to be

wrong. Another source of unwarranted conclusion may be the use of wrong average. Suppose in a series there are extreme values, one is too high while the other is too low, such as 800 and 50. The use of an arithmetic average in such a case may give a wrong idea. Instead, harmonic mean would be proper in such a case.

(vii) Confusion of correlation and causation: In statistics, several times one has to examine the relationship between two variables. A close relationship between the two variables may not establish a cause-and-effect-relationship in the sense that one variable is the cause and the other is the effect. It should be taken as something that measures degree of association rather than try to find out causal relationship.

1.5 CHECK YOUR PROGRESS

There are some activities to check your progress. Answer the followings:

1. Statistics are affected to a marked extent by multiplicity of …… .
2. Statistics helps in the …….. of suitable policies.
3. Statistics bring definiteness and …… in conclusions by expressing them numerically.
4. Inferential statistics, also known as …… statistics.
5. Unwarranted conclusions may bring the result of making …….. assumptions.

1.6 SUMMARY

In a summarized manner, 'Statistics' means numerical information expressed in quantitative terms. As a matter of fact, data have no limits as to their reference, coverage, and scope. At the macro level, these are data on gross national product and shares of agriculture, manufacturing, and services in GDP (Gross Domestic Product). At the micro level, individual firms, howsoever small or large, produce extensive statistics on their operations. The annual reports of companies contain variety of data on sales, production, expenditure, inventories, capital employed, and other activities. These data are often field data, collected by employing scientific survey techniques. Unless regularly updated, such data are the product of a one-time effort and have limited use beyond the situation that may have called for their collection. A student knows statistics more intimately as a subject of study like economics, mathematics, chemistry, physics, and others. It is a discipline, which scientifically deals with data, and is often described as the science of data. In dealing with statistics as data, statistics has developed

appropriate methods of collecting, presenting, summarizing, and analyzing data, and thus consists of a body of these methods.

1.7 KEYWORDS

Statistics: It is a science that deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry.

Descriptive statistics: It deals with collecting, summarizing, and simplifying data, which are otherwise quite unwieldy and voluminous.

Inferential statistics: It goes beyond describing a given problem situation by means of collecting, summarizing, and meaningfully presenting the related data.

1.8 SELF-ASSESSMENT TEST

1. Define Statistics. Explain its types, and importance to trade, commerce and business.

2. "Statistics is all-pervading". Elucidate this statement.

3. Write a note on the scope and limitations of Statistics.

4. What are the major limitations of Statistics? Explain with suitable examples.

5. Distinguish between descriptive Statistics and inferential Statistics.

1.9 ANSWERS TO CHECK YOUR PROGRESS

1. Causes
2. Formulation
3. Precision
4. Inductive
5. False

1.10 REFERENCES/SUGGESTED READINGS

1. Gupta, S. P. : Statistical Methods, Sultan chand and Sons, New Delhi.
2. Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.
3. Hein, L. W. Quantitative Approach to Managerial Decisions, Prentice Hall, NJ.
4. Levin, Richard I. and David S. Rubin: Statistics for Management, Prentice Hall, New Delhi.
5. Lawrance B. Moore: Statistics for Business & Economics, Harper Collins, NY.
6. Watsman Terry J. and Keith Parramor: Quantitative Methods in Finance International, Thompson Business Press, London.

Subject: Business Statistics	
Course code: MBA-102	Author: Surinder Kundu
Lesson: 02	Vetter: Prof. M. S. Turan
AN OVERVIEW OF CENTRAL TENDENCY	

STRUCTURE

2.1 Learning Objectives

2.2 Introduction

- 2.2.1 Arithmetic Mean
- 2.2.2 Median
- 2.2.3 Mode
- 2.2.4 Relationships of the Mean, Median and Mode
- 2.2.5 The Best Measure of Central Tendency
- 2.2.6 Geometric Mean
- 2.2.7 Harmonic Mean
- 2.2.8 Quadratic Mean

2.3 Check your progress

2.4 Summary

2.5 Keywords

2.6 Self-Assessment Test

2.7 Answers to check your progress

2.8 References/Suggested Readings

2.1 LEARNING OBJECTIVES

After going through this lesson, you will be able to:

- Understand the calculations and main properties of measures of central tendency
- Understand the calculations and main properties of measures of mean
- Understand the calculations and main properties of measures of median
- Understand the calculations and main properties of measures of mode

- Understand the calculations and main properties of measures of harmonic mean
- Understand the calculations and main properties of measures of geometric mean

2.2 INTRODUCTION

The description of statistical data may be quite elaborate or quite brief depending on two factors: the nature of data and the purpose for which the same data have been collected. While describing data statistically or verbally, one must ensure that the description is neither too brief nor too lengthy. The measures of central tendency enable us to compare two or more distributions pertaining to the same time period or within the same distribution over time. For example, the average consumption of tea in two different territories for the same period or in a territory for two years, say, 2003 and 2004, can be attempted by means of an average.

2.2.1 Arithmetic Mean

Adding all the observations and dividing the sum by the number of observations results the arithmetic mean. Suppose we have the following observations:

10, 15, 30, 7, 42, 79 and 83

These are seven observations. Symbolically, the arithmetic mean, also called simply *mean* is

$\bar{x} = \sum x/n$, where $\bar{x}$ is simple mean.

$$= \frac{10+15+30+7+42+79+83}{7} = \frac{266}{7} = 38$$

It may be noted that the Greek letter μ is used to denote the mean of the population and n to denote the total number of observations in a population. Thus the population mean $\mu = \sum x/n$. The formula given above is the basic formula that forms the definition of arithmetic mean and is used in case of ungrouped data where weights are not involved.

Ungrouped Data-Weighted Average

In case of ungrouped data where weights are involved, our approach for calculating arithmetic mean will be different from the one used earlier.

Example 2.1: Suppose a student has secured the following marks in three tests:

Mid-term test	30
Laboratory	25
Final	20

The simple arithmetic mean will be $\frac{30+25+20}{3}=25$

However, this will be wrong if the three tests carry different weights on the basis of their relative importance. Assuming that the weights assigned to the three tests are:

Mid-term test	2 points
Laboratory	3 points
Final	5 points

Solution: On the basis of this information, we can now calculate a weighted mean as shown below:

Table 2.1: Calculation of a Weighted Mean

Type of Test	Relative Weight (w)	Marks (x)	(wx)
Mid-term	2	30	60
Laboratory	3	25	75
Final	5	20	100
Total	$\sum$ w = 10		235

$$\bar{x}=\frac{\sum wx}{\sum w}=\frac{w_1x_1+w_2x_2+w_3x_3}{w_1+w_2+w_3}$$

$$=\frac{60+75+100}{2+3+5}=23.5 \text{ marks}$$

It will be seen that weighted mean gives a more realistic picture than the simple or unweighted mean.

Example 2.2: An investor is fond of investing in equity shares. During a period of falling prices in the stock exchange, a stock is sold at Rs 120 per share on one day, Rs 105 on the next and Rs 90 on the third day. The investor has purchased 50 shares on the first day, 80 shares on the second day and 100 shares on the third' day. What average price per share did the investor pay?

Solution:

Table 2.2: Calculation of Weighted Average Price

Day	Price per Share (Rs) (x)	No of Shares Purchased (w)	Amount Paid (wx)
1	120	50	6000
2	105	80	8400
3	90	100	9000
Total	-	230	23,400

$$\text{Weighted average} = \frac{w_1x_1+w_2x_2+w_3x_3}{w_1+w_2+w_3}=\frac{\sum wx}{\sum w}$$

$$=\frac{6000+8400+9000}{50+80+100}=101.7 \text{ marks}$$

Therefore, the investor paid an average price of Rs 101.7 per share.

It will be seen that if merely prices of the shares for the three days (regardless of the number of shares purchased) were taken into consideration, then the average price would be

$$Rs.\frac{120+105+90}{3}=105$$

This is an unweighted or simple average and as it ignores the-quantum of shares purchased, it fails to give a correct picture. A simple average, it may be noted, is also a weighted average where weight in each case is the same, that is, only 1. When we use the term average alone, we always mean that it is an unweighted or simple average.

Grouped Data-Arithmetic Mean

For grouped data, arithmetic mean may be calculated by applying any of the following methods:

(i) Direct method, (ii) Short-cut method , (iii) Step-deviation method

In the case of direct method, the formula $\bar{x} = \sum fm/n$ is used. Here *m* is mid-point of various classes, *f* is the frequency of each class and *n* is the total number of frequencies. The calculation of arithmetic mean by the direct method is shown below.

Example 2.3: The following table gives the marks of 58 students in Statistics. Calculate the average marks of this group.

Marks	*No. of Students*
0-10	4
10-20	8
20-30	11
30-40	15
40-50	12
50-60	6
60-70	2
Total	58

Solution:

Table 2.3: Calculation of Arithmetic Mean by Direct Method

Marks	Mid-point m	No. of Students f	fm
0-10	5	4	20

10-20	15	8	120
20-30	25	11	275
30-40	35	15	525
40-50	45	12	540
50-60	55	6	330
60-70	65	2	130
			∑fm = 1940

Where, $\bar{x} = \frac{\sum fm}{n} = \frac{1940}{58} = 33.45$ marks or 33 marks approximately.

It may be noted that the mid-point of each class is taken as a good approximation of the true mean of the class. This is based on the assumption that the values are distributed fairly evenly throughout the interval. When large numbers of frequency occur, this assumption is usually accepted. In the case of short-cut method, the concept of arbitrary mean is followed. The formula for calculation of the arithmetic mean by the short-cut method is given below:

$$\bar{x} = A + \frac{\sum fd}{n}$$

Where A = arbitrary or assumed mean

f = frequency

d = deviation from the arbitrary or assumed mean

When the values are extremely large and/or in fractions, the use of the direct method would be very cumbersome. In such cases, the short-cut method is preferable. This is because the calculation work in the short-cut method is considerably reduced particularly for calculation of the product of values and their respective frequencies. However, when calculations are not made manually but by a machine calculator, it may not be necessary to resort to the short-cut method, as the use of the direct method may not pose any problem.

As can be seen from the formula used in the short-cut method, an arbitrary or assumed mean is used. The second term in the *formula* ($\sum fd \div n$) is the correction factor for the difference between the actual mean and the assumed mean. If the assumed mean turns out to be equal to the actual mean, ($\sum fd \div n$) will be zero. The use of the short-cut method is based on the principle that the total of deviations taken from an actual mean is equal to zero. As such, the deviations taken from any other figure will depend on how the assumed mean is related to the actual mean. While one may choose any value as assumed

mean, it would be proper to avoid extreme values, that is, too small or too high to simplify calculations. A value apparently close to the arithmetic mean should be chosen.

For the figures given earlier pertaining to marks obtained by 58 students, we calculate the average marks by using the short-cut method.

Example 2.4:

Table 2.4: Calculation of Arithmetic Mean by Short-cut Method

Marks	Mid-point m	f	d	fd
0-10	5	4	-30	-120
10-20	15	8	-20	-160
20-30	25	11	-10	-110
30-40	35	15	0	0
40-50	45	12	10	120
50-60	55	6	20	120
60-70	65	2	30	60
				∑fd = -90

It may be noted that we have taken arbitrary mean as 35 and deviations from midpoints. In other words, the arbitrary mean has been subtracted from each value of mid-point and the resultant figure is shown in column *d.*

$$\bar{x} = A + \frac{\sum fd}{n} = 35 + \left(\frac{-90}{58}\right)$$ = 35 -1.55 = 33.45 or 33 marks approximately.

Now we take up the calculation of arithmetic mean for the same set of data using the step-deviation method. This is shown in Table 2.5.

Table 2.5: Calculation of Arithmetic Mean by Step-deviation Method

Marks	Mid-point	f	d	d'= d/10	Fd'
0-10	5	4	-30	-3	-12
10-20	15	8	-20	-2	-16
20-30	25	11	-10	-1	-11
30-40	35	15	0	0	0
40-50	45	12	10	1	12
50-60	55	6	20	2	12
60-70	65	2	30	3	6
					∑fd' =-9

$$\bar{x} = A + \frac{\sum fd'}{n} \times C = 35 + \left(\frac{-9 \times 10}{58}\right)$$ = 33.45 or 33 marks approximately.

It will be seen that the answer in each of the three cases is the same. The step-deviation method is the most convenient on account of simplified calculations. It may also be noted that if we select a different arbitrary mean and recalculate deviations from that figure, we would get the same answer.

Now that we have learnt how the arithmetic mean can be calculated by using different methods, we are in a position to handle any problem where calculation of the arithmetic mean is involved.

Example 2.6: The mean of the following frequency distribution was found to be 1.46.

No. of Accidents	*No. of Days (frequency)*
0	46
1	?
2	?
3	25
4	10
5	5
Total	200 days

Calculate the missing frequencies.

Solution:

Here we are given the total number of frequencies and the arithmetic mean. We have to determine the two frequencies that are missing. Let us assume that the frequency against 1 accident is x and against 2 accidents is y. If we can establish two simultaneous equations, then we can easily find the values of X and *Y*.

$$\text{Mean} = \frac{(0.46)+(1.x)+(2.y)+(3.25)+(4.10)+(5.5)}{200}$$

$$1.46 = \frac{x+2y+140}{200}$$

$$x + 2y + 140 = (200)\ (1.46)$$

$$x + 2y = 152$$

$$x + y = 200 - \{46+25 + 10+5\}$$

$$x + y = 200 - 86 = 114$$

Now subtracting equation (ii) from equation (i), we get

$$x + 2y = 152$$

$$x + y = 114$$

$$- \quad - \quad -$$

$$y = 38$$

Substituting the value of $y = 38$ in equation (ii) above, $x + 38 = 114$

Therefore, $x = 114 - 38 = 76$

Hence, the missing frequencies are:

Against accident 1 : 76

Against accident 2 : 38

Characteristics of the Arithmetic Mean

Some of the important characteristics of the arithmetic mean are:

1. The sum of the deviations of the individual items from the arithmetic mean is always zero. This means I: $(x - \bar{x}) = 0$, where x is the value of an item and x is the arithmetic mean. Since the sum of the deviations in the positive direction is equal to the sum of the deviations in the negative direction, the arithmetic mean is regarded as a measure of central tendency.
2. The sum of the squared deviations of the individual items from the arithmetic mean is always minimum. In other words, the sum of the squared deviations taken from any value other than the arithmetic mean will be higher.
3. As the arithmetic mean is based on all the items in a series, a change in the value of any item will lead to a change in the value of the arithmetic mean.
4. In the case of highly skewed distribution, the arithmetic mean may get distorted on account of a few items with extreme values. In such a case, it may cease to be the representative characteristic of the distribution.

2.2.2 Median

Median is defined as the value of the middle item (or the mean of the values of the two middle items) when the data are arranged in an ascending or descending order of magnitude. Thus, in an ungrouped frequency distribution if the n values are arranged in ascending or descending order of magnitude, the median is the middle value if n is odd. When n is even, the median is the mean of the two middle values.

Suppose we have the following series: 15, 19,21,7, 10,33,25,18 and 5

We have to first arrange it in either ascending or descending order. These figures are arranged in an ascending order as follows: 5,7,10,15,18,19,21,25,33

Now as the series consists of odd number of items, to find out the value of the middle item, we use the formula Where $\frac{n+1}{2}$, Where *n* is the number of items. In this case, *n* is 9, as such $\frac{n+1}{2}$ = 5, that is, the size of the 5th item is the median. This happens to be 18. Suppose the series consists of one more items 23. We may, therefore, have to include 23 in the above series at an appropriate place, that is, between 21 and 25. Thus, the series is now 5, 7, 10, 15, 18, 19, and 21, 23, 25, 33. Applying the above formula, the median is the size of 5.5^{th} item. Here, we have to take the average of the values of 5th and 6th item. This means an average of 18 and 19, which gives the median as 18.5.

It may be noted that the formula $\frac{n+1}{2}$ itself is not the formula for the median; it merely indicates the position of the median, namely, the number of items we have to count until we arrive at the item whose value is the median. In the case of the even number of items in the series, we identify the two items whose values have to be averaged to obtain the median. In the case of a grouped series, the median is calculated by linear interpolation with the help of the following formula:

$$\mathrm{M} = l_1 \frac{l_2 + l_1}{f}(m - c)$$

Where *M* = the median

l_1 = the lower limit of the class in which the median lies

1_2 = the upper limit of the class in which the median lies

f = the frequency of the class in which the median lies

m = the middle item or *(n* + 1)/2th, where *n* stands for total number of items

c = the cumulative frequency of the class preceding the one in which the median lies

Example 2.7:

Monthly Wages (Rs)	*No. of Workers*
800-1,000	18
1,000-1,200	25
1,200-1,400	30
1,400-1,600	34
1,600-1,800	26
1,800-2,000	10
Total	143

In order to calculate median in this case, we have to first provide cumulative frequency to the table. Thus, the table with the cumulative frequency is written as:

Monthly Wages	Frequency	Cumulative Frequency
800 -1,000	18	18
1,000 -1,200	25	43
1,200 -1,400	30	73
1,400 -1,600	34	107
1,600 -1,800	26	133
1.800 -2,000	10	143

$$M = l_1 \frac{l_2 + l_1}{f}(m - c)$$

$$M = \frac{n+1}{2} = \frac{143+1}{2} = 72$$

It means median lies in the class-interval Rs 1,200 - 1,400.

$$\text{Now, } M = 1200 + \frac{1400-1200}{30}(72-43) = 1200 + \frac{200}{30}(29) = \text{Rs } 1393.3$$

At this stage, let us introduce two other concepts viz. quartile and decile. To understand these, we should first know that the median belongs to a general class of statistical descriptions *called fractiles.* A fractile is a value below that lays a given fraction of a set of data. In the case of the median, this fraction is one-half (1/2). Likewise, a quartile has a fraction one-fourth (1/4). The three quartiles Q_1, Q_2 and Q_3 are such that 25 percent of the data fall below Q_1, 25 percent fall between Q_1 and Q_2, 25 percent fall between Q_2 and Q_3 and 25 percent fall above Q_3 It will be seen that Q_2 is the median. We can use the above formula for the calculation of quartiles as well. The only difference will be in the value of m. Let us calculate both Q_1 and Q_3 in respect of the table given in Example 2.7.

$$Q_1 = l_1 \frac{l_2 - l_1}{f}(m - c)$$

$$\text{Here, } m \text{ will be } = \frac{n+1}{4} = \frac{143+1}{4} = 36$$

$$Q_1 = 1000 + \frac{1200-1000}{25}(36-18) = 1000 + \frac{200}{25}(18) = \text{Rs. } 1{,}144$$

In the case of Q_3, m will be 3 = $\frac{n+1}{4} = \frac{3\times 144}{4} = 108$

$$Q_1 = 1600 + \frac{1800-1600}{26}(108-107) = 1600 + \frac{200}{26}(1) = \text{Rs. } 1{,}607.7 \text{ approx}$$

In the same manner, we can calculate deciles (where the series is divided into 10 parts) and percentiles (where the series is divided into 100 parts). It may be noted that unlike arithmetic mean, median is not affected at all by extreme values, as it is a positional average. As such, median is particularly very useful when a distribution happens to be skewed. Another point that goes in favour of median is that it can be computed when a distribution has open-end classes. Yet, another merit of median is that when a distribution contains qualitative data, it is the only average that can be used. No other average is suitable in case of such a distribution. Let us take a couple of examples to illustrate what has been said in favour of median.

Example 2.8: Calculate the most suitable average for the following data:

Size of the Item	Below 50	50-100	100-150	150-200	200 or above
Frequency	*15*	*20*	*36*	*40*	*10*

Solution: Since the data have two open-end classes-one in the beginning (below 50) and the other at the end (200 and above), median should be the right choice as a measure of central tendency.

Table 2.6: Computation of Median

Size of Item	*Frequency*	*Cumulative Frequency*
Below 50	15	15
50-100	20	35
100-150	36	71
150-200	40	111
200 and above	10	121

Median is the size of $\frac{n+1}{2}$ th item = $\frac{121+1}{2} = 61^{st}$ item

Now, 61st item lies in the 100-150 class

Median = $1_1 = l_1 \frac{l_2 - l_1}{f}(m-c) = 100 + \frac{150-100}{36}(61-35) = 100 + 36.11 = 136.11$ approx.

Example 2.9: The following data give the savings bank accounts balances of nine sample households selected in a survey. The figures are in rupees.

745 2,000 1,500 68,000 461 549 3750 1800 4795

(a) Find the mean and the median for these data; (b) Do these data contain an outlier? If so, exclude this value and recalculate the mean and median. Which of these summary measures has a greater change when an outlier is dropped?; (c) Which of these two summary measures is more appropriate for this series?

Solution:

$$\text{Mean} = \text{Rs. } \frac{745+2,000+1,500+68,000+461+549+3,750+1,800+4,795}{9}$$

$$= \frac{\text{Rs}\,83,600}{9} = \text{Rs } 9,289$$

$$\text{Median} = \text{Size of } \frac{n+1}{2}\text{th item} = \frac{9+1}{2} = 5^{\text{th}} \text{ item}$$

Arranging the data in an ascending order, we find that the median is Rs 1,800.

(b) An item of Rs 68,000 is excessively high. Such a figure is called an 'outlier'. We exclude this figure and recalculate both the mean and the median.

$$\text{Mean} = \text{Rs. } \frac{83,600-68,000}{8} = \text{Rs } \frac{15,600}{8} = \text{Rs. } 1,950$$

$$\text{Median} = \text{Size of } \frac{n+1}{2}\text{item} = \frac{8+1}{2} = 4.5th \text{ item.} = \text{Rs. } \frac{1,500-1,800}{2} = \text{Rs. } 1,650$$

It will be seen that the mean shows a far greater change than the median when the outlier is dropped from the calculations.

(c) As far as these data are concerned, the median will be a more appropriate measure than the mean.

Example 2.10:

Suppose we are given the following series:

Class interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	6	12	22	37	17	8	5

We are asked to draw both types of ogive from these data and to determine the median.

Solution:

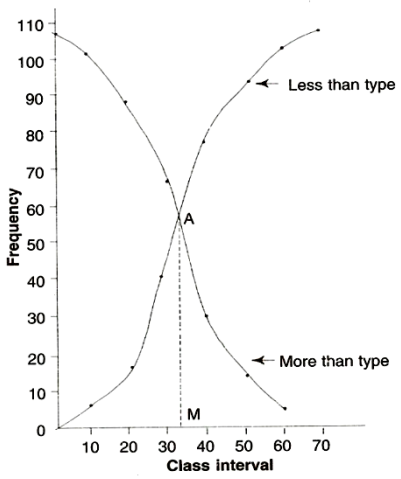
First of all, we transform the given data into two cumulative frequency distributions, one based on 'less than' and another on 'more than' methods.

Table A

	Frequency
Less than 10	6
Less than 20	18
Less than 30	40
Less than 40	77
Less than 50	94
Less than 60	102
Less than 70	107

Table B

	Frequency
More than 0	107
More than 10	101
More than 20	89
More than 30	67
More than 40	30
More than 50	13
More than 60	5



It may be noted that the point of intersection of the two ogives gives the value of the median. From this point of intersection A, we draw a straight line to meet the X-axis at M. Thus, from the point of origin

to the point at M gives the value of the median, which comes to 34, approximately. If we calculate the median by applying the formula, then the answer comes to 33.8, or 34, approximately. It may be pointed out that even a single ogive can be used to determine the median. As we have determined the median graphically, so also we can find the values of quartiles, deciles or percentiles graphically. For example, to determine we have to take size of $\{3(n + 1)\}/4 = 81^{st}$ item. From this point on the Y-axis, we can draw a perpendicular to meet the 'less than' ogive from which another straight line is to be drawn to meet the X-axis. This point will give us the value of the upper quartile. In the same manner, other values of Q_1 and deciles and percentiles can be determined.

Characteristics of the Median

1. Unlike the arithmetic mean, the median can be computed from open-ended distributions. This is because it is located in the median class-interval, which would not be an open-ended class.
2. The median can also be determined graphically whereas the arithmetic mean cannot be ascertained in this manner.
3. As it is not influenced by the extreme values, it is preferred in case of a distribution having extreme values.
4. In case of the qualitative data where the items are not counted or measured but are scored or ranked, it is the most appropriate measure of central tendency.

2.2.3 Mode

The mode is another measure of central tendency. It is the value at the point around which the items are most heavily concentrated. As an example, consider the following series: 8, 9, 11, 15, 16, 12, 15, 3, 7, 15

There are ten observations in the series wherein the figure 15 occurs maximum number of times three. The mode is therefore 15. The series given above is a discrete series; as such, the variable cannot be in fraction. If the series were continuous, we could say that the mode is approximately 15, without further computation. In the case of grouped data, mode is determined by the following formula:

$$\text{Mode} = l_1 + \frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \times i$$

Where, l_1 = the lower value of the class in which the mode lies

f_l = the frequency of the class in which the mode lies

f_o = the frequency of the class preceding the modal class

f_2 = the frequency of the class succeeding the modal class

i = the class-interval of the modal class

While applying the above formula, we should ensure that the class-intervals are uniform throughout. If the class-intervals are not uniform, then they should be made uniform on the assumption that the frequencies are evenly distributed throughout the class. In the case of inequal class-intervals, the application of the above formula will give misleading results.

Example 2.11: Let us take the following frequency distribution:

Class intervals (1)	*Frequency (2)*
30-40	4
40-50	6
50-60	8
60-70	12
70-80	9
80-90	7
90-100	4

We have to calculate the mode in respect of this series.

Solution: We can see from Column (2) of the table that the maximum frequency of 12 lies in the class-interval of 60-70. This suggests that the mode lies in this class-interval. Applying the formula given earlier, we get:

$$\text{Mode} = 60 + \frac{12-8}{12-8(12-8)+(12-9)} \times 10 = 60 + \frac{4}{4+3} \times 10 = 65.7 \text{ approx.}$$

In several cases, just by inspection one can identify the class-interval in which the mode lies. One should see which the highest frequency is and then identify to which class-interval this frequency belongs. Having done this, the formula given for calculating the mode in a grouped frequency distribution can be applied.

At times, it is not possible to identify by inspection the class where the mode lies. In such cases, it becomes necessary to use the method of grouping. This method consists of two parts:

(i) **Preparation of a grouping table:** A grouping table has six columns, the first column showing the frequencies as given in the problem. Column 2 shows frequencies grouped in two's, starting from the top. Leaving the first frequency, column 3 shows frequencies grouped in two's. Column 4 shows the frequencies of the first three items, then second to fourth item and so on. Column 5 leaves the first frequency and groups the remaining items in three's. Column 6 leaves the first two frequencies and then groups the remaining in three's. Now, the maximum total in each column is marked and shown either in a circle or in a bold type.

(ii) **Preparation of an analysis table**: After having prepared a grouping table, an analysis table is prepared. On the left-hand side, provide the first column for column numbers and on the right-hand side the different possible values of mode. The highest values marked in the grouping table are shown here by a bar or by simply entering 1 in the relevant cell corresponding to the values they represent. The last row of this table will show the number of times a particular value has occurred in the grouping table. The highest value in the analysis table will indicate the class-interval in which the mode lies. The procedure of preparing both the grouping and analysis tables to locate the modal class will be clear by taking an example.

Example 2.12:
The following table gives some frequency data:

Size of Item	Frequency
10-20	10
20-30	18
30-40	25
40-50	26
50-60	17
60-70	4

Solution:

Grouping Table

Size of item	1	2	3	4	5	6
10-20	10					
		28				
20-30	18					
			43			
30-40	25			53		
		51				
40-50	26				69	
			43			68
50-60	17			47		
		21				
60-70	4					

Analysis table

	Size of item				
Col. No.	10-20	20-30	30-40	40-50	50-60
1				1	
2			1	1	
3		1	1	1	1
4	1	1	1		
5		1	1	1	
6			1	1	1
Total	1	3	5	5	2

This is a bi-modal series as is evident from the analysis table, which shows that the two classes 30-40 and 40-50 have occurred five times each in the grouping. In such a situation, we may have to determine mode indirectly by applying the following formula:

Mode = 3 median - 2 mean

Median = Size of *(n* + l)/2th item, that is, 101/2 = 50.5th item. This lies in the class 30-40. Applying the formula for the median, as given earlier, we get

$$= 30 + \frac{40-30}{25}(50.5-28) = 30 + 9 = 39$$

Now, arithmetic mean is to be calculated. This is shown in the following table.

Class- interval	Frequency	Mid- points	d	d' = d/10	fd'
10-20	10	15	-20	-2	-20
20-30	18	25	-10	-I	-18
30-40	25	35	0	0	0
40-50	26	45	10	1	26
50-60	17	55	20	2	34
60-70	4	65	30	3	12
Total	100				34

Deviation is taken from arbitrary mean = 35

$$\text{Mean} = A + \frac{\sum fd'}{n} \times i = 35 + \frac{34}{100} \times 10 = 38.4$$

Mode = 3 median - 2 mean = (3 x 39) - (2 x 38.4) = 117 -76.8 = 40.2

This formula, Mode = 3 Median-2 Mean, is an empirical formula only. And it can give only approximate results. As such, its frequent use should be avoided. However, when mode is ill defined or the series is bimodal (as is the case in the present example) it may be used.

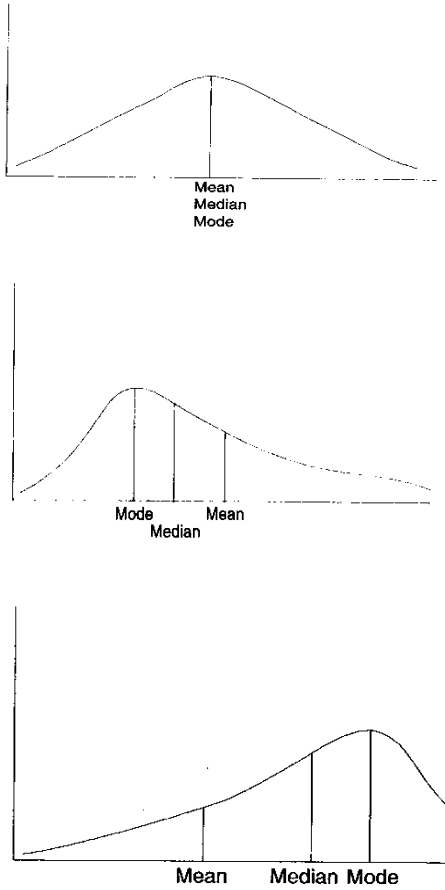
2.2.4 Relationships of the Mean, Median and Mode

Having discussed mean, median and mode, we now turn to the relationship amongst these three measures of central tendency. We shall discuss the relationship assuming that there is a unimodal frequency distribution.

(i) When a distribution is symmetrical, the mean, median and mode are the same, as is shown above in the following figure. In case, a distribution is skewed to the right, then mean> median> mode. Generally, income distribution is skewed to the right where a large number of families have relatively low income and a small number of families have extremely high income. In such a case, the mean is pulled up by the extreme high incomes and the relation among these three measures is as shown in Fig. 6.3. Here, we find that mean> median> mode.

(ii) When a distribution is skewed to the left, then mode> median> mean. This is because here mean is pulled down below the median by extremely low values. This is shown as in the figure.

(iii) Given the mean and median of a unimodal distribution, we can determine whether it is skewed to the right or left. When mean> median, it is skewed to the right; when median> mean, it is skewed to the left. It may be noted that the median is always in the middle between mean and mode.



2.2.5 The best Measure of Central Tendency

At this stage, one may ask as to which of these three measures of central tendency the best is. There is no simple answer to this question. It is because these three measures are based upon different concepts. The arithmetic mean is the sum of the values divided by the total number of observations in the series. The median is the value of the middle observation that divides the series into two equal parts. Mode is the value around which the observations tend to concentrate. As such, the use of a particular measure will largely depend on the purpose of the study and the nature of the data; For example, when we are interested in knowing the consumers preferences for different brands of television sets or different kinds of advertising, the choice should go in favour of mode. The use of mean and median would not be

proper. However, the median can sometimes be used in the case of qualitative data when such data can be arranged in an ascending or descending order. Let us take another example. Suppose we invite applications for a certain vacancy in our company. A large number of candidates apply for that post. We are now interested to know as to which age or age group has the largest concentration of applicants. Here, obviously the mode will be the most appropriate choice. The arithmetic mean may not be appropriate as it may be influenced by some extreme values. However, the mean happens to be the most commonly used measure of central tendency as will be evident from the discussion in the subsequent chapters.

2.2.6 Geometric Mean

Apart from the three measures of central tendency as discussed above, there are two other means that are used sometimes in business and economics. These are the geometric mean and the harmonic mean. The geometric mean is more important than the harmonic mean. We discuss below both these means. First, we take up the geometric mean. Geometric mean is defined at the *nth* root of the product of *n* observations of a distribution.

Symbolically, GM = $\sqrt[n]{x_1 \ldots x_2 \ldots . x_n \ldots}$ If we have only two observations, say, 4 and 16 then GM = $\sqrt{4 \times 16} = \sqrt{64} = 8.$ Similarly, if there are three observations, then we have to calculate the cube root of the product of these three observations; and so on. When the number of items is large, it becomes extremely difficult to multiply the numbers and to calculate the root. To simplify calculations, logarithms are used.

Example 2.13: If we have to find out the geometric mean of 2, 4 and 8, then we find

$$\text{Log GM} = \frac{\sum \log x_i}{n} = \frac{Log2 + Log4 + Log8}{3} = \frac{0.3010 + 0.6021 + 0.9031}{3}$$

$$= \frac{1.8062}{3} = 0.60206$$

GM = Antilog 0.60206 = 4

When the data are given in the form of a frequency distribution, then the geometric mean can be obtained by the formula:

$$\text{Log GM} = \frac{f_1 . \log x_l + f_2 . \log x_2 + \ldots + f_n . \log x_n}{f_1 + f_2 + \ldots\ldots fn} = \frac{\sum f . \log x}{f_1 + f_2 + \ldots\ldots fn}$$

Then, GM = Antilog *n*

The geometric mean is most suitable in the following three cases:

1. Averaging rates of change.
2. The compound interest formula.
3. Discounting, capitalization.

Example 2.14: A person has invested Rs 5,000 in the stock market. At the end of the first year the amount has grown to Rs 6,250; he has had a 25 percent profit. If at the end of the second year his principal has grown to Rs 8,750, the rate of increase is 40 percent for the year. What is the average rate of increase of his investment during the two years?

Solution:

$$\text{GM} = \sqrt{1.25 \times 1.40} = \sqrt{1.75}. = 1.323$$

The average rate of increase in the value of investment is therefore 1.323 - 1 = 0.323, which if multiplied by 100, gives the rate of increase as 32.3 percent.

Example 2.15: We can also derive a compound interest formula from the above set of data. This is shown below:

Solution:

Now, 1.25 x 1.40 = 1.75. This can be written as $1.75 = (1 + 0.323)^2$.

Let $P_2 = 1.75$, $P_0 = 1$, and $r = 0.323$, then the above equation can be written as $P_2 = (1 + r)^2$ or $P_2 = P_0 (1 + r)^2$.

Where *P2* is the value of investment at the end of the second year, P_0 is the initial investment and *r* is the rate of increase in the two years. This, in fact, is the familiar compound interest formula. This can be written in a generalised form as $P_n = P_0(1 + r)^n$. In our case *Po* is Rs 5,000 and the rate of increase in investment is 32.3 percent. Let us apply this formula to ascertain the value of *Pn*, *that* is, investment at the end of the second year.

$$P_n = 5{,}000\ (1 + 0.323)^2 = 5{,}000 \text{ x } 1.75 = \text{Rs } 8{,}750$$

It may be noted that in the above example, if the arithmetic mean is used, the resultant figure will be wrong. In this case, the average rate for the two years is $\frac{25+40}{2}$ percent per year, which comes to 32.5.

Applying this rate, we get $P_n = \frac{165}{100}$ x 5,000 = Rs 8,250. This is obviously wrong, as the figure should have been Rs 8,750.

Example 2.16: An economy has grown at 5 percent in the first year, 6 percent in the second year, 4.5 percent in the third year, 3 percent in the fourth year and 7.5 percent in the fifth year. What is the average rate of growth of the economy during the five years?

Solution:

Year	*Rate of Growth (percent)*	*Value at the end of the Year x (in Rs)*	*Log x*
1	5	105	2.02119
2	6	106	2.02531
3	4.5	104.5	2.01912
4	3	103	2.01284
5	7.5	107.5	2.03141
			$\sum$ log X = 10.10987

$$GM = \text{Antilog}\left(\frac{\sum \log x}{n}\right) = \text{Antilog}\left(\frac{10.10987}{5}\right) = \text{Antilog } 2.021974 = 105.19$$

Hence, the average rate of growth during the five-year period is 105.19 - 100 = 5.19 percent per annum. In case of a simple arithmetic average, the corresponding rate of growth would have been 5.2 percent per annum.

Discounting

The compound interest formula given above was

$$P_n = P_0(1+r)^n \quad \text{This can be written as } P_0 = \frac{P_n}{(1+r)^n}$$

This may be expressed as follows:

If the future income is P_n rupees and the present rate of interest is 100 *r* percent, then the present value of *P* n rupees will be P_0 rupees. For example, if we have a machine that has a life of 20 years and is

expected to yield a net income of Rs 50,000 per year, and at the end of 20 years it will be obsolete and cannot be used, then the machine's present value is

$$\frac{50{,}000}{(1+r)^n}+\frac{50{,}000}{(1+r)^2}+\frac{50{,}000}{(1+r)^3}+\ldots\ldots\ldots\ldots\frac{50{,}000}{(1+r)^{20}}$$

This process of ascertaining the present value of future income by using the interest rate is known as discounting.

In conclusion, it may be said that when there are extreme values in a series, geometric mean should be used as it is much less affected by such values. The arithmetic mean in such cases will give misleading results. Before we close our discussion on the geometric mean, we should be aware of its advantages and limitations.

Advantages of G.M.

1. Geometric mean is based on each and every observation in the data set.
2. It is rigidly defined.
3. It is more suitable while averaging ratios and percentages as also in calculating growth rates.
4. As compared to the arithmetic mean, it gives more weight to small values and less weight to large values. As a result of this characteristic of the geometric mean, it is generally less than the arithmetic mean. At times it may be equal to the arithmetic mean.
5. It is capable of algebraic manipulation. If the geometric mean has two or more series is known along with their respective frequencies. Then a combined geometric mean can be calculated by using the logarithms.

Limitations of G.M.

1. As compared to the arithmetic mean, geometric mean is difficult to understand.
2. Both computation of the geometric mean and its interpretation are rather difficult.
3. When there is a negative item in a series or one or more observations have zero value, then the geometric mean cannot be calculated.

In view of the limitations mentioned above, the geometric mean is not frequently used.

2.2.7 Harmonic Mean

The harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocals of individual observations. Symbolically,

$$\text{HM} = \frac{n}{1/x_1 + 1/x_2 + 1/x_3 + \ldots + 1/x_n} = \text{Re}ciprocal\frac{\sum 1/x}{n}$$

The calculation of harmonic mean becomes very tedious when a distribution has a large number of observations. In the case of grouped data, the harmonic mean is calculated by using the following formula:

$$\text{HM} = \text{Reciprocal of } \sum_{i-1}^{n}\left(f_i \times \frac{1}{x_i}\right)$$

or

$$\frac{n}{\sum_{i-1}^{n}\left(f_i \times \frac{1}{x_i}\right)}$$

Where *n* is the total number of observations.

Here, each reciprocal of the original figure is weighted by the corresponding frequency *(f)*. The main advantage of the harmonic mean is that it is based on all observations in a distribution and is amenable to further algebraic treatment. When we desire to give greater weight to smaller observations and less weight to the larger observations, then the use of harmonic mean will be more suitable. As against these advantages, there are certain limitations of the harmonic mean. First, it is difficult to understand as well as difficult to compute. Second, it cannot be calculated if any of the observations is zero or negative. Third, it is only a summary figure, which may not be an actual observation in the distribution.

It is worth noting that the harmonic mean is always lower than the geometric mean, which is lower than the arithmetic mean. This is because the harmonic mean assigns lesser importance to higher values. Since the harmonic mean is based on reciprocals, it becomes clear that as reciprocals of higher values are lower than those of lower values, it is a lower average than the arithmetic mean as well as the geometric mean.

Example 2.17: Suppose we have three observations 4, 8 and 16. We are required to calculate the harmonic mean. Reciprocals of 4,8 and 16 are: $\frac{1}{4}, \frac{1}{8}, \frac{1}{16}$ respectively

$$\text{HM} = \frac{n}{1/x_1 + 1/x_2 + 1/x_3} = \frac{3}{1/4 + 1/8 + 1/16}$$

$$= \frac{3}{0.25+0.125+0.0625} = 6.857 \text{ approx.}$$

Example 2.18: Consider the following series:

Class-interval	2-4	4-6	6-8	8-10
Frequency	20	40	30	10

Solution:

Let us set up the table as follows:

Class-interval	Mid-value	Frequency	Reciprocal of MV	f x 1/x
2-4	3	20	0.3333	6.6660
4-6	5	40	0.2000	8.0000
6-8	7	30	0.1429	4.2870
8-10	9	10	0.1111	1.1111
			Total	20.0641

$$= \frac{\sum_{i-1}^{n}\left(f_i \times \frac{1}{x_i}\right)}{n} = \frac{100}{20.0641} = 4.984 \text{ approx.}$$

Example 2.19: In a small company, two typists are employed. Typist A types one page in ten minutes while typist B takes twenty minutes for the same. (i) Both are asked to type 10 pages. What is the average time taken for typing one page? (ii) Both are asked to type for one hour. What is the average time taken by them for typing one page?

Solution: Here Q-(i) is on arithmetic mean while Q-(ii) is on harmonic mean.

$$\text{M} = \frac{(10\times 10)+(20\times 20)(\text{min}utes)}{10\times 2(pages)} = 15 \text{ minutes}$$

$$\text{HM} = \frac{60\times(\text{min}utes)}{60/10+60/20(pages)} = \frac{120}{\frac{120+60}{20}} = \frac{40}{3} = 13\text{min}utes \text{ and } 20 \text{ seconds.}$$

Example 2.20: It takes ship A 10 days to cross the Pacific Ocean; ship B takes 15 days and ship C takes 20 days. (i) What is the average number of days taken by a ship to cross the Pacific Ocean? (ii) What is

the average number of days taken by a cargo to cross the Pacific Ocean when the ships are hired for 60 days?

Solution: Here again Q-(i) pertains to simple arithmetic mean while Q-(ii) is concerned with the harmonic mean.

$$M = \frac{10+15+20}{3} = 15 \text{ days}$$

$$HM = \frac{60 \times 3(days)\_}{60/10+60/15+60/20} = \frac{180}{\frac{360+240+180}{60}} = 13.8 \text{ days approx.}$$

2.2.8 Quadratic Mean

We have seen earlier that the geometric mean is the antilogarithm of the arithmetic mean of the logarithms, and the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. Likewise, the quadratic mean (Q) is the square root of the arithmetic mean of the squares. Symbolically,

$$Q = \sqrt{\frac{x_1^2 + x_2^2 + ...+_n^2}{n}}$$

Instead of using original values, the quadratic mean can be used while averaging deviations when the standard deviation is to be calculated. This will be used in the next chapter on dispersion.

Relative Position of Different Means

The relative position of different means will always be:

$Q > \bar{x} > G > H$ provided that all the individual observations in a series are positive and all of them are not the same.

Composite Average or Average of Means

Sometimes, we may have to calculate an average of several averages. In such cases, we should use the same method of averaging that was employed in calculating the original averages. Thus, we should calculate the arithmetic mean of several values of *x,* the geometric mean of several values of GM, and the harmonic mean of several values of HM. It will be wrong if we use some other average in averaging of means.

2.3 CHECK YOUR PROGRESS

1. When we use the term average alone, we always mean that it is an ……. or simple average.
2. When a distribution is to …… the left, then mode> median>mean.

3. When a distribution is, the mean, median and mode are the same
4. When there is a negative item in a series or one or more observations have zero value, then the …….. cannot be calculated.
5. The quadratic mean (Q) is the …… of the arithmetic mean of the squares.

2.4 SUMMARY

It is the most important objective of statistical analysis is to get one single value that describes the characteristics of the entire mass of cumbersome data. Such a value is finding out, which is known as central value to serve our purpose.

2.5 KEYWORDS

Arithmetic mean: Adding all the observations and dividing the sum by the number of observations results the arithmetic mean

Median: It is defined as the value of the middle item (or the mean of the values of the two middle items) when the data are arranged in an ascending or descending order of magnitude.

Mode: It is the value at the point around which the items are most heavily concentrated.

Geometric mean: It is defined at the *nth* root of the product of *n* observations of a distribution.

Harmonic mean: It is defined as the reciprocal of the arithmetic mean of the reciprocals of individual observations.

2.6 SELF-ASSESSMENT TEST

1. What are the desiderata (requirements) of a good average? Compare the mean, the median and the mode in the light of these desiderata? Why averages are called measures of central tendency?
2. "Every average has its own peculiar characteristics. It is difficult to say which average is the best." Explain with examples.
3. What do you understand .by 'Central Tendency'? Under what conditions is the median more suitable than other measures of central tendency?
4. The average monthly salary paid to all employees in a company was Rs 8,000. The average monthly salaries paid to male and female employees of the company were Rs 10,600 and Rs 7,500 respectively. Find out the percentages of males and females employed by the company.

5. Calculate the arithmetic mean from the following data:

Class	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
Frequency	*2*	*4*	*9*	*11*	*12*	*6*	*4*	*2*

6. Calculate the mean, median and mode from the following data:

Height in Inches	Number of Persons
62-63	2
63-64	6
64-65	14
65-66	16
66-67	8
67-68	3
68-69	1
Total	50

7. A number of particular articles have been classified according to their weights. After drying for two weeks, the same articles have again been weighed and similarly classified. It is known that the median weight in the first weighing was 20.83 gm while in the second weighing it was 17.35 gm. Some frequencies *a* and *b* in the first weighing and *x* and *y* in the second are missing. It is known that $a = 1/3x$ and $b = 1/2\ y$. Find out the values of the missing frequencies.

Class	*Frequencies*	
	First Weighing	*Second Weighing*
0- 5	a	z
5-10	b	y
10-15	11	40
15-20	52	50
20-25	75	30
25-30	22	28

8. Cities A, Band C are equidistant from each other. A motorist travels from A to B at 30 km/h; from B to C at 40 km/h and from C to A at 50 km/h. Determine his average speed for the entire trip.
9. Calculate the harmonic mean from the following data:

Class-Interval	2-4	4-6	6-8	8-1
Frequency	20	40	30	10

10. A vehicle when climbing up a gradient, consumes petrol @ 8 km per litre. While coming down it runs 12 km per litre. Find its average consumption for to and fro travel between two places situated at the two ends of 25 Ian long gradients.

2.7 ANSWERS TO CHECK YOUR PROGRESS

1. Unweighted
2. Skewed
3. Symmetrical
4. Geometric mean
5. Square root

2.8 REFERENCES/SUGGESTED READINGS

7. Levin, Richard I. and David S. Rubin: Statistics for Management, Prentice Hall, New Delhi.
8. Watsman Terry J. and Keith Parramor: Quantitative Methods in Finance International, Thompson Business Press, London.
9. Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.
10. Hein, L. W. Quantitative Approach to Managerial Decisions, Prentice Hall, NJ.

Subject: Business Statistics	
Course code: MBA-102	Author: Surinder Kundu
Lesson: 03	Vetter: Prof. M. S. Turan
DISPERSION AND SKEWNESS	

STRUCTURE

3.1 Learning objectives

3.2 Introduction

- 3.2.1 Meaning and Definition of Dispersion
- 3.2.2 Significance and Properties of Measuring Variation
- 3.2.3 Measures of Dispersion
- 3.2.4 Range
- 3.2.5 Interquartile Range or Quartile Deviation
- 3.2.6 Mean Deviation
- 3.2.7 Standard Deviation
- 3.2.8 Lorenz Curve

3.3 Skewness: Meaning and Definitions

- 3.3.1 Tests of Skewness
- 3.3.2 Measures of Skewness

3.4 Moments and Kurtosis

3.5 Check your progress

3.6 Summary

3.7 Keywords

3.8 Self-Assessment Test

3.9 Answers to check your progress

3.10 References/Suggested Readings

3.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Impart the knowledge of measures of dispersion
- Impart the knowledge of measures of skewness
- Distinguish between average, dispersion, skewness, moments and kurtosis

3.2 INTRODUCTION

In the previous chapter, we have explained the measures of central tendency. It may be noted that these measures do not indicate the extent of dispersion or variability in a distribution. The dispersion or variability provides us one more step in increasing our understanding of the pattern of the data. Further, a high degree of uniformity (i.e. low degree of dispersion) is a desirable quality. If in a business there is a high degree of variability in the raw material, then it could not find mass production economical. Suppose an investor is looking for a suitable equity share for investment. While examining the movement of share prices, he should avoid those shares that are highly fluctuating-having sometimes very high prices and at other times going very low. Such extreme fluctuations mean that there is a high risk in the investment in shares. The investor should, therefore, prefer those shares where risk is not so high.

3.2.1 Meaning and Definitions of Dispersion

The various measures of central value give us one single figure that represents the entire data. But the average alone cannot adequately describe a set of observations, unless all the observations are the same. It is necessary to describe the variability or dispersion of the observations. In two or more distributions the central value may be the same but still there can be wide disparities in the formation of distribution. Measures of dispersion help us in studying this important characteristic of a distribution.

Some important definitions of dispersion are given below:

1. "Dispersion is the measure of the variation of the items." -A.L. Bowley

2. "The degree to which numerical data tend to spread about an average value is called the variation of dispersion of the data."-Spiegel

3. Dispersion or spread is the degree of the scatter or variation of the variable about a central value." -Brooks & Dick

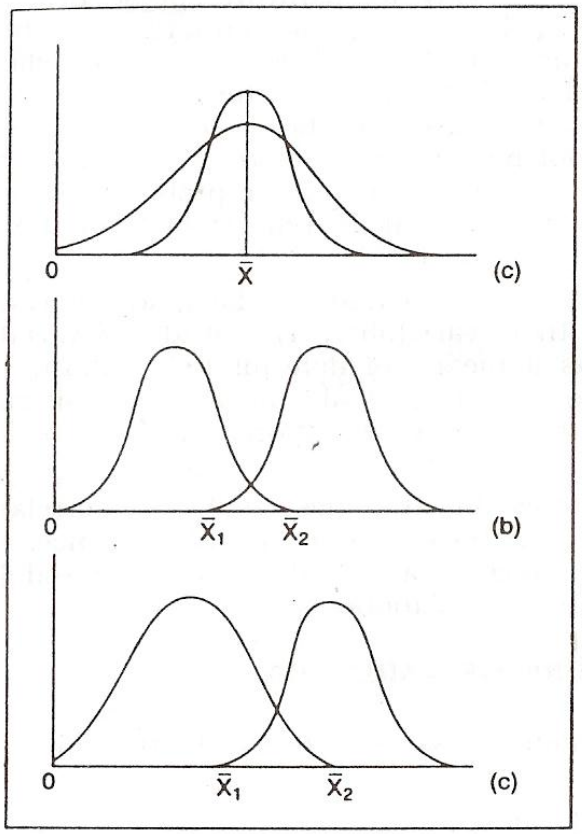
4. "The measurement of the scatterness of the mass of figures in a series about an average is called

measure of variation or dispersion." -Simpson & Kajka

It is clear from above that dispersion (also known as scatter, spread or variation) measures the extent to which the items vary from some central value. Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of the second order. An average is more meaningful when it is examined in the light of dispersion. For example, if the average wage of the workers of factory A is Rs. 3885 and that of factory B Rs. 3900, we cannot necessarily conclude that the workers of factory B are better off because in factory B there may be much greater dispersion in the distribution of wages. The study of dispersion is of great significance in practice as could well be appreciated from the following example:

	Series A	**Series B**	**Series C**
	100	100	1
	100	105	489
	100	102	2
	100	103	3
	100	90	5
Total	500	500	500
$\overline{X}$	100	100	100

Since arithmetic mean is the same in all three series, one is likely to conclude that these series are alike in nature. But a close examination shall reveal that distributions differ widely from one another. In series A, (In Box-3.1) each and every item is perfectly represented by the arithmetic mean or in other words none of the items of series A deviates from the arithmetic mean and hence there is no dispersion. In series B, only one item is perfectly represented by the arithmetic mean and the other items vary but the variation is very small as compared to series C. In series C. not a single item is represented by the arithmetic mean and the items vary widely from one another. In series C, dispersion is much greater compared to series B. Similarly, we may have two groups of labourers with the same mean salary and yet their distributions may differ widely. The mean salary may not be so important a characteristic as the variation of the items from the mean. To the student of social affairs the mean income is not so vitally important as to know how this income is distributed. Are a large number receiving the mean income or are there a few with enormous incomes and millions with incomes far below the mean? The three figures given in Box 3.1 represent frequency distributions with some of the characteristics. The



two curves in diagram (a) represent two distractions with the same mean $\overline{X}$, but with different dispersions. The two curves in (b) represent two distributions with the same dispersion but with unequal means $\overline{X}_1$ and $\overline{X}_2$, (c) represents two distributions with unequal dispersion. The measures of central tendency are, therefore insufficient. They must be supported and supplemented with other measures.

In the present chapter, we shall be especially concerned with the measures of variability or spread or dispersion. A measure of variation or dispersion is one that measures the extent to which there are differences between individual observation and some central or average value. In measuring variation we shall be interested in the amount of the variation or its degree but not in the direction. For example, a measure of 6 inches below the mean has just as much dispersion as a measure of six inches above the mean.

Literally meaning of dispersion is 'scatteredness'. Average or the measures of central tendency gives us an idea of the concentration of the observations about the central part of the distribution. If we know the average alone, we cannot form a complete idea about the distribution. But with the help of dispersion, we have an idea about homogeneity or heterogeneity of the distribution.

3.2.2 Significance and Properties of Measuring Variation

Measures of variation are needed for four basic purposes:

1. Measures of variation point out as to how far an average is representative of the mass. When dispersion is small, the average is a typical value in the sense that it closely represents the individual value and it is reliable in the sense that it is a good estimate of the average in the corresponding universe. On the other hand, when dispersion is large, the average is not so typical, and unless the sample is very large, the average may be quite unreliable.

2. Another purpose of measuring dispersion is to determine nature and cause of variation in order to control the variation itself. In matters of health variations in body temperature, pulse beat and blood pressure are the basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production efficient operation requires control of quality variation the causes of which are sought through inspection is basic to the control of causes of variation. In social sciences a special problem requiring the measurement of variability is the measurement of "inequality" of the distribution of income or wealth etc.
3. Measures of dispersion enable a comparison to be made of two or more series with regard to their variability. The study of variation may also be looked upon as a means of determining uniformity of consistency. A high degree of variation would mean little uniformity or consistency whereas a low degree of variation would mean great uniformity or consistency.
4. Many powerful analytical tools in statistics such as correlation analysis. the testing of hypothesis, analysis of variance, the statistical quality control, regression analysis is based on measures of variation of one kind or another.

A good measure of dispersion should possess the following properties:

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be rigidly defined.
4. It should be based on each and every item of the distribution.
5. It should be amenable to further algebraic treatment.
6. It should have sampling stability.
7. Extreme items should not unduly affect it.

3.2.3 Measures of Dispersion

There are five measures of dispersion: Range, Inter-quartile range or Quartile Deviation, Mean deviation, Standard Deviation, and Lorenz curve. Among them, the first four are mathematical methods and the last one is the graphical method. These are discussed in the ensuing paragraphs with suitable examples.

3.2.4 Range

The simplest measure of dispersion is the range, which is the difference between the maximum value and the minimum value of data.

Example 3.1: Find the range for the following three sets of data:

Set 1:	05	15	15	05	15	05	15	15	15	15
Set 2:	8	7	15	11	12	5	13	11	15	9
Set 3:	5	5	5	5	5	5	5	5	5	
5										

Solution: In each of these three sets, the highest number is 15 and the lowest number is 5. Since the range is the difference between the maximum value and the minimum value of the data, it is 10 in each case. But the range fails to give any idea about the dispersal or spread of the series between the highest and the lowest value. This becomes evident from the above data.

In a frequency distribution, range is calculated by taking the difference between the upper limit of the highest class and the lower limit of the lowest class.

Example 3.2: Find the range for the following frequency distribution:

Size of Item	Frequency
20- 40	7
40- 60	11
60- 80	30
80-100	17
100-120	5
Total	**70**

Solution: Here, the upper limit of the highest class is 120 and the lower limit of the lowest class is 20. Hence, the range is 120 - 20 = 100. Note that the range is not influenced by the frequencies. Symbolically, the range is calculated b the formula L - S, where L is the largest value and S is the smallest value in a distribution. The coefficient of range is calculated by the formula: (L-S)/ (L+S). This is the relative measure. The coefficient of the range in respect of the earlier example having three sets of data is: 0.5.The coefficient of range is more appropriate for purposes of comparison as will be evident from the following example:

Example 3.3: Calculate the coefficient of range separately for the two sets of data given below:

Set 1	8	10	20	9	15	10	13	28
Set 2	30	35	42	50	32	49	39	33

Solution: It can be seen that the range in both the sets of data is the same:

Set 1 28 - 8 = 20

Set 2 50 - 30 = 20

Coefficient of range in Set 1 is:

$$\frac{28 - 8}{28+8} = 0.55$$

Coefficient of range in set 2 is:

$$\frac{50 - 30}{50 +30} = 0.25$$

LIMITATIONS OF RANGE

There are some limitations of range, which are as follows:

1. It is based only on two items and does not cover all the items in a distribution.
2. It is subject to wide fluctuations from sample to sample based on the same population.
3. It fails to give any idea about the pattern of distribution. This was evident from the data given in Examples 1 and 3.
4. Finally, in the case of open-ended distributions, it is not possible to compute the range.

Despite these limitations of the range, it is mainly used in situations where one wants to quickly have some idea of the variability or' a set of data. When the sample size is very small, the range is considered quite adequate measure of the variability. Thus, it is widely used in quality control where a continuous check on the variability of raw materials or finished products is needed. The range is also a suitable measure in weather forecast. The meteorological department uses the range by giving the maximum and the minimum temperatures. This information is quite useful to the common man, as he can know the extent of possible variation in the temperature on a particular day.

3.2.5 Inter-quartile Range or Quartile Deviation

The interquartile range or the quartile deviation is a better measure of variation in a distribution than the range. Here, avoiding the 25 percent of the distribution at both the ends uses the middle 50 percent of the distribution. In other words, the interquartile range denotes the difference between the third quartile and the first quartile. Symbolically,

$$\text{Interquartile range} = Q_3 - Q_1$$

Many times the interquartile range is reduced in the form of semi-interquartile range or quartile deviation as shown below:

Semi interquartile range or Quartile deviation = $(Q_3 – Q_I)/2$

When quartile deviation is small, it means that there is a small deviation in the central 50 percent items. In contrast, if the quartile deviation is high, it shows that the central 50 percent items have a large variation. It may be noted that in a symmetrical distribution, the two quartiles, that is, Q3 and QI are equidistant from the median. Symbolically,

$$M\text{-}Q_I = Q_3\text{-}M$$

However, this is seldom the case as most of the business and economic data are asymmetrical. But, one can assume that approximately 50 percent of the observations are contained in the interquartile range. It may be noted that interquartile range or the quartile deviation is an absolute measure of dispersion. It can be changed into a relative measure of dispersion as follows:

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

The computation of a quartile deviation is very simple, involving the computation of upper and lower quartiles. As the computation of the two quartiles has already been explained in the preceding chapter, it is not attempted here.

MERITS OF QUARTILE DEVIATION

The following merits are entertained by quartile deviation:

1. As compared to range, it is considered a superior measure of dispersion.
2. In the case of open-ended distribution, it is quite suitable.
3. Since it is not influenced by the extreme values in a distribution, it is particularly suitable in highly skewed or erratic distributions.

LIMITATIONS OF QUARTILE DEVIATION

1. Like the range, it fails to cover all the items in a distribution.
2. It is not amenable to mathematical manipulation.
3. It varies widely from sample to sample based on the same population.
4. Since it is a positional average, it is not considered as a measure of dispersion. It merely shows a distance on scale and not a scatter around an average.

In view of the above-mentioned limitations, the interquartile range or the quartile deviation has a limited practical utility.

3.2.6 Mean Deviation

The mean deviation is also known as the average deviation. As the name implies, it is the average of absolute amounts by which the individual items deviate from the mean. Since the positive deviations from the mean are equal to the negative deviations, while computing the mean deviation, we ignore positive and negative signs. Symbolically,

$$MD = \frac{\sum |x|}{n}$$

Where MD = mean deviation, |x| = deviation of an item from the mean ignoring positive and negative signs, *n* = the total number of observations.

Example 3.4:

Size of Item	Frequency
2-4	20
4-6	40
6-8	30
8-10	10

Solution:

Size of Item	Mid-points (m)	Frequency (f)	fm	d from $\bar{x}$	f |d|
2-4	3	20	60	-2.6	52
4-6	5	40	200	-0.6	24
6-8	7	30	210	1.4	42
8-10	9	10	90	3.4	34
	Total	**100**	**560**		**152**

$$\bar{x} = \frac{\sum fm}{n} = \frac{560}{100} = 5.6$$

$$MD(\bar{x}) = \frac{\sum f\,|d|}{n} = \frac{152}{100} = 1.52$$

MERITS OF MEAN DEVIATION

1. A major advantage of mean deviation is that it is simple to understand and easy to calculate.

2. It takes into consideration each and every item in the distribution. As a result, a change in the value of any item will have its effect on the magnitude of mean deviation.
3. The values of extreme items have less effect on the value of the mean deviation.
4. As deviations are taken from a central value, it is possible to have meaningful comparisons of the formation of different distributions.

LIMITATIONS OF MEAN DEVIATION

1. It is not capable of further algebraic treatment.
2. At times it may fail to give accurate results. The mean deviation gives best results when deviations are taken from the median instead of from the mean. But in a series, which has wide variations in the items, median is not a satisfactory measure.
3. Strictly on mathematical considerations, the method is wrong as it ignores the algebraic signs when the deviations are taken from the mean.

In view of these limitations, it is seldom used in business studies. A better measure known as the standard deviation is more frequently used.

3.2.7 Standard Deviation

The standard deviation is similar to the mean deviation in that here too the deviations are measured from the mean. At the same time, the standard deviation is preferred to the mean deviation or the quartile deviation or the range because it has desirable mathematical properties.

Before defining the concept of the standard deviation, we introduce another concept viz. variance.

Example 3.5:

X	X-μ	$(X-\mu)^2$
20	20-18=12	4
15	15-18= -3	9
19	19-18 = 1	1
24	24-18 = 6	36
16	16-18 = -2	4
14	14-18 = -4	16
108	**Total**	**70**

Solution:

$$\text{Mean} = \frac{108}{6} = 18$$

The second column shows the deviations from the mean. The third or the last column shows the squared deviations, the sum of which is 70. The arithmetic mean of the squared deviations is:

$$\frac{\sum(x-\mu)^2}{N} = 70/6=11.67 \text{ approx.}$$

This mean of the squared deviations is known as the variance. It may be noted that this variance is described by different terms that are used interchangeably: the variance of the distribution X; the variance of X; the variance of the distribution; and just simply, the variance. Symbolically,

$$\text{Var (X)} = \frac{\sum(x-\mu)^2}{N}$$

It is also written as $$\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$$

Where σ^2 (called sigma squared) is used to denote the variance.

Although the variance is a measure of dispersion, the unit of its measurement is (points). If a distribution relates to income of families then the variance is $(Rs)^2$ and not rupees. Similarly, if another distribution pertains to marks of students, then the unit of variance is $(marks)^2$. To overcome this inadequacy, the square root of variance is taken, which yields a better measure of dispersion known as the standard deviation. Taking our earlier example of individual observations, we take the square root of the variance

$$\text{SD or } \sigma = \sqrt{Variance} = \sqrt{11.67} = 3.42 \text{ points}$$

$$\text{Symbolically, } \sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}$$

In applied Statistics, the standard deviation is more frequently used than the variance. This can also be written as:

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

We use this formula to calculate the standard deviation from the individual observations given earlier.

Example 7.6:

X	X^2
20	400
15	225
19	361
24	576
16	256
14	196
108	2014

Solution:

$$\sum x_i^2 = 2014 \qquad \sum x_i = 108 \qquad N = 6$$

$$\sigma = \sqrt{\frac{2014 - \frac{(108)^2}{6}}{6}} \quad \text{Or,} \quad \sigma = \sqrt{\frac{2014 - \frac{11664}{6}}{6}}$$

$$\sigma = \sqrt{\frac{\frac{12084 - 11664}{6}}{6}} \quad \text{Or,} \quad \sigma = \sqrt{\frac{\frac{420}{6}}{6}}$$

$$\sigma = \sqrt{\frac{70}{6}} \quad \text{Or,} \quad \sigma = \sqrt{11.67}$$

$$\sigma = 3.42$$

Example 3.7:

The following distribution relating to marks obtained by students in an examination:

Marks	Number of Students
0- 10	1
10- 20	3
20- 30	6
30- 40	10
40- 50	12
50- 60	11
60- 70	6
70- 80	3
80- 90	2
90-100	1

Solution:

Marks	Frequency (f)	Mid-points	Deviations (d)/10=d'	Fd'	fd'^2
0- 10	1	5	-5	-5	25
10- 20	3	15	-4	-12	48
20- 30	6	25	-3	-18	54
30- 40	10	35	-2	-20	40
40- 50	12	45	-1	-12	12
50- 60	11	55	0	0	0
60- 70	6	65	1	6	6
70- 80	3	75	2	6	12
80- 90	2	85	3	6	18
90-100	1	95	4	4	16
Total	**55**		**Total**	**-45**	**231**

In the case of frequency distribution where the individual values are not known, we use the midpoints of the class intervals. Thus, the formula used for calculating the standard deviation is as given below:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{K} fi(m_i - \mu)^2}{N}}$$

Where m_i is the mid-point of the class intervals μ is the mean of the distribution, *fi* is the frequency of each class; N is the total number of frequency and K is the number of classes. This formula requires that the mean μ be calculated and that deviations (m_i - μ) be obtained for each class. To avoid this inconvenience, the above formula can be modified as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{K} fid_i^2 \left(\sum_{i=1}^{K} fd_i\right)}{N}}$$

Where C is the class interval: f_i is the frequency of the *i*th class and d_i is the deviation of the of item from an assumed origin; and N is the total number of observations.

Applying this formula for the table given earlier,

$$\sigma = 10\sqrt{\frac{231}{55} - \left(\frac{-45}{55}\right)^2}$$

$$= 10\sqrt{4.2 - 0.669421}$$

$$= 18.8 \text{ marks}$$

When it becomes clear that the actual mean would turn out to be in fraction, calculating deviations from the mean would be too cumbersome. In such cases, an assumed mean is used and the deviations from it are calculated. While mid-point of any class can be taken as an assumed mean, it is advisable to choose the mid-point of that class that would make calculations least cumbersome. Guided by this consideration, in Example 3.7 we have decided to choose 55 as the mid-point and, accordingly, deviations have been taken from it. It will be seen from the calculations that they are considerably simplified.

USES OF THE STANDARD DEVIATION

The standard deviation is a frequently used measure of dispersion. It enables us to determine as to how far individual items in a distribution deviate from its mean. In a symmetrical, bell-shaped curve:

(i) About 68 percent of the values in the population fall within: $\pm$ 1 standard deviation from the mean.

(ii) About 95 percent of the values will fall within $\pm$2 standard deviations from the mean.

(iii) About 99 percent of the values will fall within $\pm$ 3 standard deviations from the mean.

The standard deviation is an absolute measure of dispersion as it measures variation in the same units as the original data. As such, it cannot be a suitable measure while comparing two or more distributions. For this purpose, we should use a relative measure of dispersion. One such measure of relative dispersion is the coefficient of variation, which relates the standard deviation and the mean such that the standard deviation is expressed as a percentage of mean. Thus, the specific unit in which the standard deviation is measured is done away with and the new unit becomes percent. Symbolically,

$$\text{CV (coefficient of variation)} = \frac{\sigma}{\mu} \text{x} 100$$

Example 3.8: In a small business firm, two typists are employed-typist A and typist B. Typist A types out, on an average, 30 pages per day with a standard deviation of 6. Typist B, on an average, types out 45 pages with a standard deviation of 10. Which typist shows greater consistency in his output?

Solution:

Coefficient of variation for $A = \frac{\sigma}{\mu} \text{x} 100$ Or

$$A = \frac{6}{30} \text{x} 100 \text{ Or } 20\% \text{ and}$$

Coefficient of variation for $B = \frac{\sigma}{\mu} \text{x} 100$

$$B = \frac{10}{45} \text{x} 100 \text{ or } 22.2\ \%$$

These calculations clearly indicate that although typist B types out more pages, there is a greater variation in his output as compared to that of typist A. We can say this in a different way: Though typist A's daily output is much less, he is more consistent than typist B. The usefulness of the coefficient of variation becomes clear in comparing two groups of data having different means, as has been the case in the above example.

Standardized Variable, Standard Scores

The variable Z = (x - $\bar{x}$)/s or (x - μ)/μ, which measures the deviation from the mean in units of the standard deviation, is called a standardised variable. Since both the numerator and the denominator are in the same units, a standardised variable is independent of units used. If deviations from the mean are given in units of the standard deviation, they are said to be expressed in standard units or standard scores. Through this concept of standardised variable, proper comparisons can be made between individual observations belonging to two different distributions whose compositions differ.

Example 3.9: A student has scored 68 marks in Statistics for which the average marks were 60 and the standard deviation was 10. In the paper on Marketing, he scored 74 marks for which the average marks were 68 and the standard deviation was 15. In which paper, Statistics or Marketing, was his relative standing higher?

Solution: The standardised variable Z = (x - $\bar{x}$) ÷ s measures the deviation of x from the mean x in terms of standard deviation s. For Statistics, Z = (68 - 60) ÷ 10 = 0.8

For Marketing, Z = (74 - 68) ÷ 15 = 0.4

Since the standard score is 0.8 in Statistics as compared to 0.4 in Marketing, his relative standing was higher in Statistics.

Example 3.10: Convert the set of numbers 6, 7, 5, 10 and 12 into standard scores:

Solution:

X	$\mathbf{X^2}$
6	36
7	49
5	25
10	100

12	144
$\sum X = 40$	$\sum X^2 = 354$

$$\bar{x} = \sum x \div N = 40 \div 5 = 8$$

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{\left(\sum X\right)^2}{N}}{N}} \quad \text{or, } \sigma = \sqrt{\frac{354 - \frac{(40)^2}{5}}{5}} = \sqrt{\frac{354-320}{5}} = 2.61 \text{ approx.}$$

$$Z = \frac{x - \bar{x}}{\sigma} = \frac{6-8}{2.61} = \text{ -0.77 (Standard score)}$$

Applying this formula to other values:

(i) $\frac{7-8}{2.61}$ = -0.38 (ii) $\frac{5-8}{2.61}$ = -1.15 (iii) $\frac{10-8}{2.61}$ = 0.77 (iv) $\frac{12-8}{2.61}$ = 1.53

Thus the standard scores for 6, 7,5,10 and 12 are -0.77, -0.38, -1.15, 0.77 and 1.53, respectively.

3.2.8 Lorenz Curve

This measure of dispersion is graphical. It is known as the Lorenz curve named after Dr. Max Lorenz. It is generally used to show the extent of concentration of income and wealth. The steps involved in plotting the Lorenz curve are:

1. Convert a frequency distribution into a cumulative frequency table.
2. Calculate percentage for each item taking the total equal to 100.
3. Choose a suitable scale and plot the cumulative percentages of the persons and income. Use the horizontal axis of X to depict percentages of persons and the vertical axis of Y to depict percent ages of income.
4. Show the line of equal distribution, which will join 0 of X-axis with 100 of Y-axis.
5. The curve obtained in (3) above can now be compared with the straight line of equal distribution obtained in (4) above. If the Lorenz curve is close to the line of equal distribution, then it implies that the dispersion is much less. If, on the contrary, the Lorenz curve is farther away from the line of equal distribution, it implies that the dispersion is considerable.

The Lorenz curve is a simple graphical device to show the disparities of distribution in any phenomenon. It is, used in business and economics to represent inequalities in income, wealth, production, savings, and so on.

Figure 3.1 shows two Lorenz curves by way of illustration. The straight line AB is a line of equal distribution, whereas AEB shows complete inequality. Curve ACB and curve ADB are the Lorenz curves.

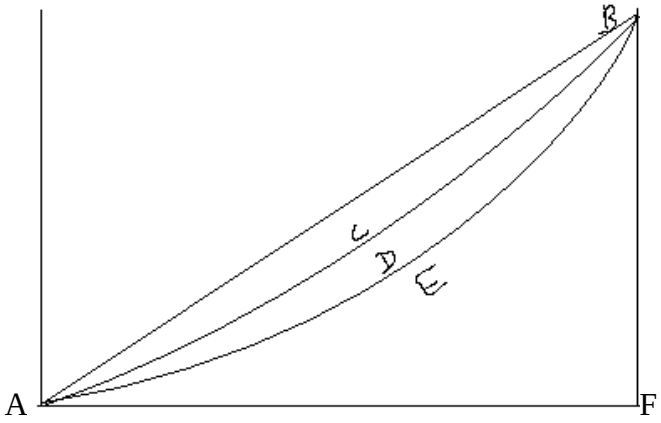
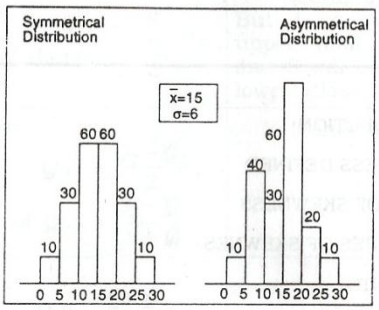


Figure 3.1: Lorenz Curve

As curve ACB is nearer to the line of equal distribution, it has more equitable distribution of income than curve ADB. Assuming that these two curves are for the same company, this may be interpreted in a different manner. Prior to taxation, the curve ADB showed greater inequality in the income of its employees. After the taxation, the company's data resulted into ACB curve, which is closer to the line of equal distribution. In other words, as a result of taxation, the inequality has reduced.

3.3 SKEWNESS: MEANING AND DEFINITIONS

In the above paragraphs, we have discussed frequency distributions in detail. It may be repeated here that frequency distributions differ in three ways: Average value, Variability or dispersion, and Shape. Since the first two, that is, average value and variability or dispersion have already been discussed in previous chapters, here our main spotlight will be on the shape of frequency distribution. Generally, there are two comparable characteristics called skewness and kurtosis that help us to understand a distribution. Two distributions may have the same mean and standard deviation but may differ widely in their overall appearance as can be seen from the following:



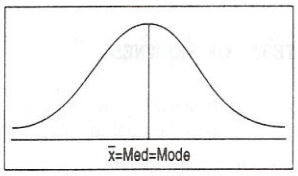
In both these distributions the value of mean and standard deviation is the same ($\overline{X}$ = 15, σ = 5). But it does not imply that the distributions are alike in nature. The distribution on the left-hand side is a

symmetrical one whereas the distribution on the right-hand side is symmetrical or skewed. Measures of skewness help us to distinguish between different types of distributions.

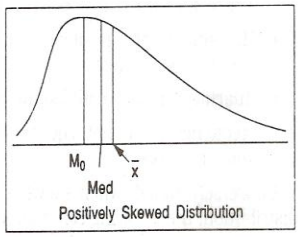
Some important definitions of skewness are as follows:

1. "When a series is not symmetrical it is said to be asymmetrical or skewed."
   -Croxton & Cowden.
2. "Skewness refers to the asymmetry or lack of symmetry in the shape of a frequency distribution." - Morris Hamburg.
3. "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness."-Simpson & Kalka
4. "A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other-to left or right." - Garrett

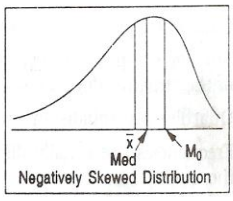
The above definitions show that the term 'skewness' refers to lack of symmetry" i.e., when a distribution is not symmetrical (or is asymmetrical) it is called a skewed distribution. The concept of skewness will be clear from the following three diagrams showing a symmetrical distribution. a positively skewed distribution and a negatively skewed distribution.



1. **Symmetrical Distribution.** It is clear from the diagram (a) that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the centre point of the curve.



2. **Asymmetrical Distribution.** A distribution, which is not symmetrical, is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed as would be clear from the diagrams (b) and (c).



3. **Positively Skewed Distribution.** In the positively skewed distribution the value of the mean is maximum and that of mode least-the median lies in between the two as is clear from the diagram (b).
4. **Negatively Skewed Distribution.** The following is the shape of negatively skewed distribution. In

a negatively skewed distribution the value of mode is maximum and that of mean least-the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the high-value end of the curve (the right hand side) than they are on the low-value end. In the negatively skewed distribution the position is reversed, i.e. the excess tail is on the left-hand side. It should be noted that in moderately symmetrical distributions the interval between the mean and the median is approximately one-third of the interval between the mean and the mode. It is this relationship, which provides a means of measuring the degree of skewness.

3.3.1 Tests of Skewness

In order to ascertain whether a distribution is skewed or not the following tests may be applied. Skewness is present if:

1. The values of mean, median and mode do not coincide.
2. When the data are plotted on a graph they do not give the normal bell-shaped form i.e. when cut along a vertical line through the centre the two halves are not equal.
3. The sum of the positive deviations from the median is not equal to the sum of the negative deviations.
4. Quartiles are not equidistant from the median.
5. Frequencies are not equally distributed at points of equal deviation from the mode.

On the contrary, when skewness is absent, i.e. in case of a symmetrical distribution, the following conditions are satisfied:

1. The values of mean, median and mode coincide.
2. Data when plotted on a graph give the normal bell-shaped form.
3. Sum of the positive deviations from the median is equal to the sum of the negative deviations.
4. Quartiles are equidistant from the median.
5. Frequencies are equally distributed at points of equal deviations from the mode.

3.3.2 Measures of Skewness

There are four measures of skewness, each divided into absolute and relative measures. The relative measure is known as the coefficient of skewness and is more frequently used than the absolute measure of skewness. Further, when a comparison between two or more distributions is involved, it is the relative measure of skewness, which is used. The measures of skewness are: (i) Karl Pearson's

measure, (ii) Bowley's measure, (iii) Kelly's measure, and (iv) Moment's measure. These measures are discussed briefly below:

KARL PEARON'S MEASURE

The formula for measuring skewness as given by Karl Pearson is as follows:

$$\text{Skewness} = \text{Mean} - \text{Mode}$$

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

In case the mode is indeterminate, the coefficient of skewness is:

$$Sk_p = \frac{\text{Mean} - (3\,\text{Median} - 2\,\text{Mean})}{\text{Standard deviation}}$$

$$Sk_p = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

Now this formula is equal to the earlier one.

$$\frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}} \qquad \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

Or 3 Mean - 3 Median = Mean - Mode

Or Mode = Mean - 3 Mean + 3 Median

Or Mode = 3 Median - 2 Mean

The direction of skewness is determined by ascertaining whether the mean is greater than the mode or less than the mode. If it is greater than the mode, then skewness is positive. But when the mean is less than the mode, it is negative. The difference between the mean and mode indicates the extent of departure from symmetry. It is measured in standard deviation units, which provide a measure independent of the unit of measurement. It may be recalled that this observation was made in the preceding chapter while discussing standard deviation. The value of coefficient of skewness is zero, when the distribution is symmetrical. Normally, this coefficient of skewness lies between ± 1. If the mean is greater than the mode, then the coefficient of skewness will be positive, otherwise negative.

Example 3.11: Given the following data, calculate the Karl Pearson's coefficient of skewness: $\sum x = 452$ $\sum x^2 = 24270$ Mode = 43.7 and N = 10

Solution:

Pearson's coefficient of skewness is:

$$Sk_P = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

Mean ($\bar{x}$)= $\frac{\sum X}{N} = \frac{452}{10} = 45.2$

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

$$(\sigma) = \sqrt{\frac{24270}{10} - \left(\frac{452}{10}\right)^2} = \sqrt{2427 - (45.2)^2} = 19.59$$

Applying the values of mean, mode and standard deviation in the above formula,

$$Sk_p = \frac{45.2 - 43.7}{19.59} = 0.08$$

This shows that there is a positive skewness though the extent of skewness is marginal.

Example 3.12: From the following data, calculate the measure of skewness using the mean, median and standard deviation:

X	10 - 20	20 - 30	30 - 40	40 - 50	50-60	60 - 70	70 - 80
f	18	30	40	55	38	20	16

Solution:

x	***MVx***	d_x	***f***	fd_x	fd_x^2	***cf***
10 - 20	15	-3	18	-54	162	18
20 - 30	25	-2	30	-60	120	48
30 - 40	35	-1	40	-40	40	88
40-50	45=a	0	55	0	0	143
50 - 60	55	1	38	38	38	181
60 - 70	65	2	20	40	80	201
70 - 80	75	3	16	48	144	217
		Total	**217**	**-28**	**584**	

a = Assumed mean = 45, *cf* = Cumulative frequency, *dx* = Deviation from assumed mean, and *i* = 10

$$\bar{x} = a + \frac{\sum fdx}{N} \times i = 45 - \frac{28}{217} \times 10 = 43.71$$

Median= $l_1 + \frac{l_2 - l_1}{f_1}(m - c)$ Where m = $(N + 1)/2^{th}$ item = $(217 + 1)/2 = 109^{th}$ item

$$\text{Median} = 40 - \frac{50-40}{55}(109-88) = 40 + \frac{10}{55} \times 21 = 43.82$$

$$\text{SD} = \sqrt{\frac{\sum fd_x^2}{\sum f} - \left(\frac{\sum fd_x}{\sum f}\right)^2} \times 10 = \sqrt{\frac{584}{217} - \left(\frac{-28}{217}\right)^2} \times 10 = \sqrt{2.69\text{-}0.016} \times 10 = 16.4$$

Skewness = 3 (Mean - Median)

= 3 (43.71 - 43.82)

= 3 x -0.011 = -0.33

Coefficient of skewness = Skewness / Standard Deviation or SD

= -0.33/16.4 = -0.02

The result shows that the distribution is negatively skewed, but the extent of skewness is extremely negligible.

Bowley's Measure

Bowley developed a measure of skewness, which is based on quartile values. The formula for measuring skewness is:

$$\text{Skewness} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

Where Q_3 and Q_1 are upper and lower quartiles and M is the median. The value of this skewness varies between ± 1. In the case of open-ended distribution as well as where extreme values are found in the series, this measure is particularly useful. In a symmetrical distribution, skewness is zero. This means that Q_3 and Q_1 are positioned equidistantly from Q_2 that is, the median. In symbols, $Q_3 - Q_2 = Q_2 - Q_1'$ In contrast, when the distribution is skewed, then $Q_3 - Q_2$ will be different from $Q_2 - Q_1'$ When $Q_3 - Q_2$ exceeds $Q_2 - Q_1'$ then skewness is positive. As against this; when $Q_3 - Q_2$ is less than $Q_2 - Q_1'$ then skewness is negative. Bowley's measure of skewness can- be written as:

$$\text{Skewness} = (Q_3 - Q_2) - (Q_2 - Q_1)$$

or $$Q_3 - Q_2 - Q_2 + Q_1$$

or $$Q_3 + Q_1 - 2Q_2 \ (2Q_2 \text{ is } 2M)$$

However, this is an absolute measure of skewness. As such, it cannot be used while comparing two distributions where the units of measurement are different. In view of this limitation, Bowley suggested a relative measure of skewness as given below:

$$\text{Relative Skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

$$= \frac{Q_3 - Q_2 - Q_2 - Q_1}{Q_3 - Q_2 + Q_2 - Q_1}$$

$$= \frac{Q_3 - Q_1 - 2Q_2}{Q_3 - Q_1}$$

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

Example 3.13: For a distribution, Bowley's coefficient of skewness is - 0.56, $Q_1=16.4$ and Median=24.2. What is the coefficient of quartile deviation?

Solution:

Bowley's coefficient of skewness is: $Sk_B = \frac{Q_3 - Q_1 - 2M}{Q_3 - Q_1}$

Substituting the values in the above formula,

$$Sk_B = \frac{Q_3 + 16.4 - (2 \times 24.2)}{Q_3 - 16.4} = -0.56 = \frac{Q_3 + 16.4 - 48.4}{Q_3 - 16.4}$$

or - 0.56 $(Q_3-16.4)$ = Q_3-32

or - 0.56 Q_3 + 9.184 = Q_3-32

or - 0.56 Q_3 - Q_3 = -32 - 9.184

- 1.56 Q_3 = - 41.184

$$Q3 = \frac{-41.184}{1.56} = 26.4$$

Now, we have the values of both the upper and the lower quartiles.

Coefficient of quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{26.4 - 16.4}{26.4 + 16.4} = \frac{10}{42.8} = 0.234$ Approx.

Example 3.14: Calculate an appropriate measure of skewness from the following data:

Value in Rs	Frequency
Less than 50	40
50 - 100	80
100 - 150	130
150 – 200	60
200 and above	30

Solution: It should be noted that the series given in the question is an open-ended series. As such, Bowley's coefficient of skewness, which is based on quartiles, would be the most appropriate measure of skewness in this case. In order to calculate the quartiles and the median, we have to use the cumulative frequency. The table is reproduced below with the cumulative frequency.

Value in Rs	Frequency	Cumulative Frequency
Less than 50	40	40
50 - 100	80	120
100 - 150	130	250
150 - 200	60	310
200 and above	30	340

$$Q_1 = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$

Now $m = (\frac{n+1}{4})$ item $= \frac{341}{4} = 85.25$, which lies in 50 - 100 class

$$Q_1 = 50 + \frac{100-50}{80}(85.25-40) = 78.28$$

$M = (\frac{n+1}{4})$ item $= \frac{341}{4} = 170.25$, which lies in 100 - 150 class

$$M = 100 + \frac{150-100}{130}(170.5-120) = 119.4$$

$$Q_3 = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$$

$m = 3(341) \div 4 = 255.75$

$$Q_3 = 150 + \frac{200-150}{60}(255.75-250) = 154.79$$

Bowley's coefficient of skewness is:

$\underline{Q_3 + Q_1 - 2M} = \underline{154.79 + 78.28 - (2 \times 119.4)} = \underline{-5.73}$

$Q_3 - Q_1$ 154.79 -78.28 76.51 = - 0.075 approx.

This shows that there is a negative skewness, which has a very negligible magnitude.

Kelly's Measure

Kelly developed another measure of skewness, which is based on percentiles. The formula for measuring skewness is as follows:

$$\text{Coefficient of skewness} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \text{ Or, } \frac{D_1 + D_9 - 2M}{D_9 - D_1}$$

Where *P* and *D* stand for percentile and decile respectively. In order to calculate the coefficient of skewness by this formula, we have to ascertain the values of 10th, 50th and 90th percentiles. Somehow, this measure of skewness is seldom used. All the same, we give an example to show how it can be calculated.

Example 3.15: Use Kelly's measure to calculate skewness.

Class Intervals	*f*	*cf*
10 - 20	18	18
20 - 30	30	48
30- 40	40	88
40- 50	55	143
50 - 60	38	181
60 – 70	20	201
70 - 80.	16	217

Solution: Now we have to calculate P_{10} P_{30} and P_{90}.

$P_{10} = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$, where m = $(n + 1)/10^{th}$ item = $\frac{217+1}{10} = 21.8th$ item

This lies in the 20 - 30 class. So,

$$P_{10} = 20 + \frac{30-20}{30}(21.8 - 18) = 20 + \frac{10 \times 3.8}{30} = 21.27approx$$

P_{50} (median): where m = $(n + 1)/2^{th}$ item = $\frac{217+1}{2} = 109^{th}$ item. This lies in the class 40 - 50. Applying the above formula:

$$P_{50} = 40 + \frac{50-40}{55}(109-88) = 40 + \frac{10\times 21}{55}\times 21 = 43.82 approx$$

$P_{90:}$ here m = 90 (217 + 1)/100th item = 196.2th item. This lies in the class 60 - 70. Applying the above formula:

$$P_{90} = 60 + \frac{70-60}{20}(196.2-181) = 60 + \frac{10\times 15.2}{20} = 67.6 approx$$

Kelley's skewness

$$Sk_K = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} = \frac{67.6 - (2 \times 43.82) + 21.27}{67.6 - 21.27} = \frac{88.87 - 87.64}{46.63} = 0.027$$

This shows that the series is positively skewed though the extent of skewness is extremely negligible. It may be recalled that if there is a perfectly symmetrical distribution, then the skewness will be zero. One can see that the above answer is very close to zero.

3.4 MOMENTS AND KURTOSIS

MOMENTS

In mechanics, the term *moment* is used to denote the rotating effect of a force. In Statistics, it is used to indicate peculiarities of a frequency distribution. The utility of moments lies in the sense that they indicate different aspects of a given distribution. Thus, by using moments, we can measure the central tendency of a series, dispersion or variability, skewness and the peakedness of the curve. The moments about the actual arithmetic mean are denoted by μ. The first four moments about mean or *central moments* are as follows:

First moment	μ_1	=	$\frac{1}{N}\sum\left(x_1 - \bar{x}\right)$
Second moment	μ_2	=	$\frac{1}{N}\sum\left(x_1 - \bar{x}\right)^2$
Third moment	μ_3	=	$\frac{1}{N}\sum\left(x_1 - \bar{x}\right)^3$
Fourth moment	μ_3	=	$\frac{1}{N}\sum\left(x_1 - \bar{x}\right)^4$

These moments are in relation to individual items. In the case of a frequency distribution, the first four moments will be:

First moment	μ_1	=	$\frac{1}{N}\sum fi\left(x_1-\bar{x}\right)$
Second moment	μ_2	=	$\frac{1}{N}\sum fi\left(x_1-\bar{x}\right)^2$
Third moment	μ_3	=	$\frac{1}{N}\sum fi\left(x_1-\bar{x}\right)^3$
Fourth moment	μ_3	=	$\frac{1}{N}\sum fi\left(x_1-\bar{x}\right)^4$

It may be noted that the first central moment is zero, that is, μ= 0. The second central moment is $\mu_2=\sigma$, indicating the variance. The third central moment μ_3 is used to measure skewness. The fourth central moment gives an idea about the Kurtosis.

Karl Pearson suggested another measure of skewness, which is based on the third and second central moments as given below:

$$\beta_1=\frac{\mu_3^2}{\mu_2^3}$$

Example 3.16: Find the (a) first, (b) second, (c) third and (d) fourth moments for the set of numbers 2,3,4,5 and 6.

Solution:

(a) $\bar{x}=\frac{\sum x}{N}=\frac{2+3+4+5+6}{5}=\frac{20}{5}=4$

(b) $\bar{x}=\frac{\sum x^2}{N}=\frac{2^2+3^2+4^2+5^2+6^2}{5}$

$=\frac{4+9+16+25+36}{5}=18$

(c) $\bar{x}=\frac{\sum x^3}{N}=\frac{2^3+3^3+4^3+5^3+6^3}{5}$

$=\frac{8+27+64+125+216}{5}=88$

(d) $\bar{x} = \frac{\sum x^4}{N} = \frac{2^4 + 3^4 + 4^4 + 5^4 + 6^4}{5}$

$= \frac{16+81+256+625+1296}{5} = 454.8$

Example 3.17: Using the same set of five figures as given in Example 3.7, find the (a) first, (b) second, (c) third and (d) fourth moments about the mean.

Solution:

$m_1 = (x - \bar{x}) = \frac{\sum(x-\bar{x})}{N} = \frac{(2-4)+(3-4)+(4-4)+(5-4)+(6-4)}{5}$

$= \frac{-2-1+0+1+2}{5} = 0$

$m_2 = (x-\bar{x})^2 = \frac{\sum(x-\bar{x})^2}{N} = \frac{(2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2}{5}$

$= \frac{(-2)^2 + (\_1)^2 + 0^2 + 1^2 + 2^2}{5}$

$= \frac{4+1+0+1+4}{5} = 2$. It may be noted that m_2 is the variance

$m_3 = = (x-\bar{x})^3 = \frac{\sum(x-\bar{x})^3}{N} = \frac{(2-4)^3+(3-4)^3+(4-4)^3+(5-4)^3+(6-4)^3}{5}$

$= \frac{(-2)^3 + (\_1)^3 + 0^3 + 1^3 + 2^3}{5} = \frac{-8-1+0+1+8}{5} = 0$

$m_4 = = (x-\bar{x})^4 = \frac{\sum(x-\bar{x})^4}{N} = \frac{(2-4)^4+(3-4)^4+(4-4)^4+(5-4)^4+(6-4)^4}{5}$

$= \frac{(-2)^4 + (\_1)^4 + 0^4 + 1^4 + 2^4}{5}$

$= \frac{16+1+0+1+016}{5} = 6.8$

Example 3.18: Calculate the first four central moments from the following data:

Class interval	50-60	60-70	70-80	80-90	90-100
Frequency	5	12	20	7	6

Solution:

Class Interval	f	MV	d from 75	d/10	fd	fd^2	fd^3	fd^4
50- 60	5	55	-20	-2	-10	20	-40	80
60- 70	12	65	-10	-1	-12	12	-12	12
70- 80	20	75	0	0	0	0	0	0
80- 90	7	85	10	1	7	7	7	7
90-100	6	95	20	2	12	24	48	96
Total	50				-3		-4	195

$$\mu_1' = \frac{\sum fd \times i}{N} = \frac{-3 \times 10}{50} = -0.6$$

$$\mu_2' = \frac{\sum fd^2 \times i}{N} = \frac{63 \times 10}{50} = 12.6$$

$$\mu_2' = \frac{\sum fd^3 \times i}{N} = \frac{-4 \times 10}{50} = 0.8$$

$$\mu_2' = \frac{\sum fd^4 \times i}{N} = \frac{195 \times 10}{50} = 19$$

Moments about Mean

$\mu_1 = \mu_1' - \mu_1' = -0.6-(-0.6) = 0$

$\mu_2 = \mu_2' - \mu_1'^2 = 10-(-0.6)^2 = 10-3.6 = 6.4$

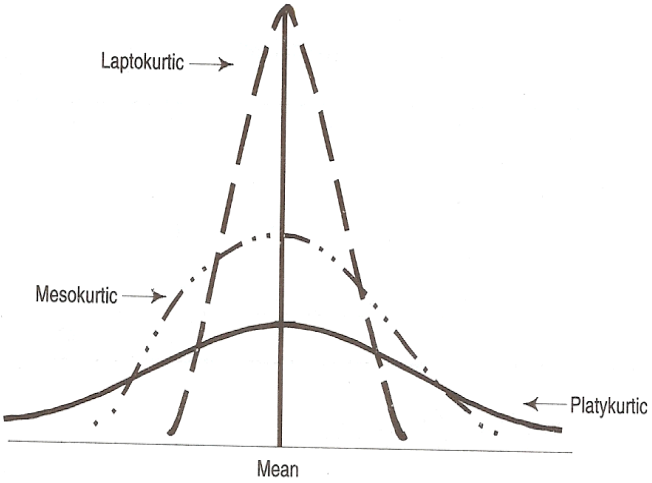
$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -0.8-3(12.6)(-0.6)+2(-0.6)^3 = -0.8 + 22.68 + 0.432 = 22.312$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2\mu_1'^2 - 3\mu_1'^4 = 19 + 4(-0.8)(-0.6) + 6(10)(-0.6)^2 - 3(-0.6)^4 = 19 + 1.92 + 21.60 - 0.3888$

$= 42.1312$

KURTOSIS

Kurtosis is another measure of the shape of a frequency curve. It is a Greek word, which means bulginess. While skewness signifies the extent of asymmetry, kurtosis measures the degree of peakedness of a frequency distribution. Karl Pearson classified curves into three types on the basis of the shape of their peaks. These are mesokurtic, leptokurtic and platykurtic. These three types of curves are shown in figure below:



It will be seen from Fig. 3.2 that mesokurtic curve is neither too much flattened nor too much peaked. In fact, this is the frequency curve of a normal distribution. Leptokurtic curve is a more peaked than the normal curve. In contrast, platykurtic is a relatively flat curve. The coefficient of kurtosis as given by Karl Pearson is $\beta_2=\mu_4/\mu_2^2$. In case of a normal distribution, that is, mesokurtic curve, the value of $\beta_2=3$. If β_2 turn out to be > 3, the curve is called a leptokurtic curve and is more peaked than the normal curve. Again, when $\beta_2 < 3$, the curve is called a platykurtic curve and is less peaked than the normal curve. The measure of kurtosis is very helpful in the selection of an appropriate average. For example, for normal distribution, mean is most appropriate; for a leptokurtic distribution, median is most appropriate; and for platykurtic distribution, the quartile range is most appropriate.

Example 3.19: From the data given in Example 3.18, calculate the kurtosis.

Solution: For this, we have to calculate β_2 This can be done by using the formula $\beta_2=\mu_4/\mu_2^2$. In the preceding example, values of μ_4 and μ_2 are given. Hence, $\beta_2 = 42.1312 \div (6.4)^2 = 1.03$. As $\beta_2 . < 3$, the distribution is platykurtic.

Another measure of kurtosis is based on both quartiles and percentiles and is given by the following formula:

$$K = \frac{Q}{P_{90} - P_{10}}$$

Where K = kurtosis, $Q = ½ (Q_3 – Q_1)$ is the semi-interquartile range; P_{90} is 90th percentile and P_{10} is the 10th percentile. This is also known as the percentile *coefficient of kurtosis.* In case of the normal distribution, the value of K is 0.263.

Example 3.20: From the data given below, calculate the percentile coefficient of kurtosis.

Daily Wages in Rs.	**Number of Workers**	**cf**
50- 60	10	10
60-70	14	24
70-80	18	42
80 - 90	24	66
90-100	16	82
100 -110	12	94
110 - 120	6	100
Total	100	

Solution: It may be noted that the question involved first two columns and in order to calculate quartiles and percentiles, cumulative frequencies have been shown in column three of the above table.

$Q_1 = l_1 \frac{l_2 - l_1}{f_1}(m - c)$, where $m = (n + 1)/4^{th}$ item, which is = 25.25^{th} item. This falls in 70 - 80 class interval. So, Q_1

$$Q_1 = 70 + \frac{80-70}{18}(25.25-24) = 70.69$$

$Q3 = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$, where m = 75.75. This falls in 90 - 100 class interval. So, Q_3

$$Q_3 = 90 + \frac{100-90}{16} (75.75 - 66) = 96.09$$

$P_{I0} = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$, where m = 10.1. This falls in 60 - 70 class interval. So, P_{10}

$$P_{10} = 60 + \frac{70-60}{14} (10.01 - 10) = 60.07$$

$P_{90} = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$, where m = 90.9. This falls in 100 - 110 class interval. So, P_{90}

$$P_{90} = 100 + \frac{110-100}{12} (90.9 - 82) = 107.41$$

$$K = \frac{Q}{P_{90} - P_{10}} = \frac{1/2(Q_3 - Q_1)}{P_{90} - P_{10}} = \frac{½\ (96.09\text{-}\ 70.69)}{107.41\text{-}60.07} = 0.268$$

It will be seen that the above distribution is very close to normal distribution as the value of K is 0.268, which is extremely close to 0.263.

3.5 CHECK YOUR PROGRESS

1. Since measures of dispersion give an average of the differences of various items from an average, they are also called averages of the …… .
2. The standard deviation is preferred to the …… or the quartile deviation or the range because it has desirable mathematical properties.
3. If deviations from the mean are given in units of the ……, they are said to be expressed in standard units or standard scores.
4. Lorenz is generally used to show the extent of …… of income and wealth.
5. Karl Pearson classified curves into three types on the basis of the shape of their peaks. These are mesokurtic, …… and platykurtic.

3.6 SUMMARY

The average value cannot adequately describe a set of observations, unless all the observations are the same. It is necessary to describe the variability or dispersion of the observations. In two or more distributions the central value may be the same but still there can be wide disparities in the formation of distribution. Therefore, we have to use the measures of dispersion.

Further, two distributions may have the same mean and standard deviation but may differ widely in their overall appearance in terms of symmetry and skewness. To distinguish between different types of distributions, we may use the measures of skewness.

3.7 KEYWORDS

Dispersion: It measures the extent to which the items vary from some central value.

Range: It is the difference between the maximum value and the minimum value of data.

Interquartile range: It denotes the difference between the third quartile and the first quartile.

Skewness: It refers to lack of symmetry" i.e., when a distribution is not symmetrical (or is asymmetrical) it is called a skewed distribution.

Moments: In Statistics, it is used to indicate peculiarities of a frequency distribution.

Kurtosis: It is another measure of the shape of a frequency curve. It is a Greek word, which means

bulginess.

3.8 SELF ASSESSMENT TEST

1. What do you mean by dispersion? What are the different measures of dispersion?
2. "Variability is not an important factor because even though the outcome is more certain, you still have an equal chance of falling either above or below the median. Therefore, on an average, the outcome will be the same." Do you agree with this statement? Give reasons for your answer.
3. Why is the standard deviation the most widely used measure of dispersion? Explain.
4. Define skewness and Dispersion.
5. Define Kurtosis and Moments.
6. What are the different measures of skewness? Which one is repeatedly used?
7. Measures of dispersion and skewness are complimentary to one another in understanding a frequency distribution." Elucidate the statement.
8. Calculate Karl Pearson's coefficient of skewness from the following data:

Weekly Sales (Rs lakh)	Number of Companies
10-12	12
12 – 14	18
14 – 16	35
16 - 18	42
18 – 20	50
22-24	30
24-26	8

9. For a distribution, the first four moments about zero are 1,7,38 and 155 respectively. (i) Compute the moment coefficients of skewness and kurtosis. (ii) Is the distribution mesokurtic? Give reason.
10. The first four moments of a distribution about the value 4 are 1, 4, 10 and 45. Obtain various characteristics of the distribution on the basis of the information given. Comment upon the nature of the distribution.
11. Define kurtosis. If β_1=1 and β_2 =4 and variance = 9, find the values of β_3 and β_4 and comment upon the nature of the distribution.
12. Calculate the first four moments about the mean from the following data. Also calculate the values of β_1 and β_2

Marks	0-10	10 – 20	20-30	30 – 40	40 – 50	50 - 60	60 - 70

No. of students	5	12	18	40	15	7	3

3.9ANSWERS TO CHECK YOUR PROGRESS

1. Second order
2. Mean deviation
3. Standard deviation
4. Concentration
5. Leptokurtic

3.10 REFERENCES/SUGGESTED READINGS

1. Levin, Richard I. and David S. Rubin: Statistics for Management, Prentice Hall, New Delhi.
2. Watsman terry J. and Keith Parramor: Quantitative Methods in Finance International, Thompson Business Press, London.
3. Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.
4. Hein, L. W. Quantitative Approach to Managerial Decisions, Prentice Hall, NJ.

Subject: Business Statistics	
Course Code: MBA-102	Author: Anil Kumar
Lesson: 04	Vetter: Dr. Karam Pal
PROBABILITY THEORY	

STRUCTURE

4.1 Learning Objectives

4.2 Introduction

- 4.2.1 Some Basic Concepts
- 4.2.2 Approaches to Probability Theory
- 4.2.3 Probability Rules
- 4.2.4 Bayes' Theorem

4.3 Some Counting Concepts

4.4 Check your Progress

4.5 Summary

4.6 Keywords

4.7 Self-Assessment Test

4.8 Answers to check your progress

4.9 References/Suggested Readings

4.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- **Understand the concept of probability**
- **Appreciate the relevance of probability theory in decision-making under conditions of uncertainty**
- **Understand and use the different approaches to probability as well as different probability rules for calculating probabilities in different situations.**

4.2 INTRODUCTION

Life is full of uncertainties. 'Probably', 'likely', 'possibly', 'chance' *etc.* is some of the most commonly used terms in our day-to-day conversation. All these terms more or less convey the same sense - *"the situation under consideration is uncertain* and commenting *on the future with certainty is impossible*". Decision-making in such areas is facilitated through formal and precise expressions for the uncertainties involved. For example, product demand is uncertain but study of demand spelled out in a form amenable for analysis may go a long to help analyze, and facilitate decisions on sales planning and inventory management. Intuitively, we see that if there is a high chance of a high demand in the coming year, we may decide to stock more. We may also take some decisions regarding the price increase, reducing sales expenses *etc.* to manage the demand. However, in order to make such decisions, we need to quantify the chances of different quantities of demand in the coming year. Probability theory provides us with the ways and means to quantify the uncertainties involved in such situations.

A probability is a quantitative measure of uncertainty - a number that conveys the strength of our belief in the occurrence of an uncertain event.

Since uncertainty is an integral part of human life, people have always been interested - consciously or unconsciously - in evaluating probabilities.

Having its origin associated with gamblers, the theory of probability today is an indispensable tool in the analysis of situations involving uncertainty. It forms the basis for inferential statistics as well as for other fields that require quantitative assessments of chance occurrences, such as quality control, management decision analysis, and almost all areas in physics, biology, engineering and economics or social life.

4.2.1 Some Basic Concepts

Probability, in common parlance, refers to the chance of occurrence of an event or happening. In order that we are able to compute it, a proper understanding of certain basic concepts in probability theory is required. These concepts are an *experiment,* a *sample space,* and an *event.*

EXPERIMENT

*An **experiment** is a process that leads to one of several possible **outcomes.** An **outcome** of an experiment is some observation or measurement.*

The term experiment is used in probability theory in a much broader sense than in physics or chemistry. Any action, whether it is the drawing a card out of a deck of 52 cards, or reading the temperature, or measurement of a product's dimension to ascertain quality, or the launching of a new product in the market, constitute an experiment in the probability theory terminology.

The experiments in probability theory have three things in common:

- there are two or more outcomes of each experiment
- it is possible to specify the outcomes in advance
- there is uncertainty about the outcomes

For example, the product we are measuring may turn out to be undersize or right size or oversize, and we are not certain which way it will be when we measure it. Similarly, launching a new product involves uncertain outcome of meeting with a success or failure in the market.

A single outcome of an experiment is called a ***basic outcome*** or an ***elementary event***. Any particular card drawn from a deck is a basic outcome.

SAMPLE SPACE

*The **sample space** is the universal set S pertinent to a given experiment. It is the set of all possible outcomes of an experiment.*

So each outcome is visualized as a sample point in the sample space. The sample spaces for the above experiments are:

Experiment	Sample Space
Drawing a Card	{all 52 cards in the deck}
Reading the Temperature	{all numbers in the range of temperatures}
Measurement of a Product's Dimension	{undersize, outsize, right size}
Launching of a New Product	{success, failure}

EVENT

An event, in probability theory, constitutes one or more possible outcomes of an experiment.

*An **event** is a subset of a sample space. It is a set of basic outcomes. We say that the event occurs if the experiment gives rise to a basic outcome belonging to the event.*

For the experiment of drawing a card, we may obtain different events A, B, and C like:

A: The event that card drawn is king of club

B: The event that card drawn is red

C: The event that card drawn is ace

In the first case, out of the 52 sample points that constitute the sample space, only one sample point or outcome defines the event, whereas the number of outcomes used in the second and third case is 13 and 4 respectively.

4.2.2 Approaches to Probability Theory

Three different approaches to the definition and interpretation of probability have evolved, mainly to cater to the three different types of situations under which probability measures are normally required. We will study these approaches with the help of examples of distinct types of experiments.

Consider the following situations marked by three distinct types of experiments. The events that we are interested in, within these experiments, are also given.

Situation I:

Experiment: Drawing a Card Out of a Deck of 52 Cards

Event A: On any draw, a king is there

Situation II

Experiment: Administering a Taste Test for a New Soup

Event B: A consumer likes the taste

Situation III

Experiment: Commissioning a Solar Power Plant

Event C: The plant turns out to be a successful venture

Situation I : THE CLASSICAL APPROACH

The first situation is characterized by the fact that for a given experiment we have a sample space with equally likely basic outcomes. When a card is drawn out of a well-shuffled deck, every one of the cards (the basic outcomes) is as likely to occur as any other. This type of situations, marked by the presence of ***"equally likely"*** outcomes, gave rise to the ***Classical Approach*** to the probability theory. In the Classical Approach, probability of an event is defined as the *relative size* of the event with respect to the size of the sample space. Since there are 4 kings and there are 52 cards, the size of A is 4 and the size of the sample space is 52. Therefore, the probability of A is equal to 4/52.

The rule we use in computing probabilities, assuming equal likelihood of all basic outcomes, is as follows:

Probability of the event A:

$$P(A) = \frac{n(A)}{N(S)} \qquad \ldots\ldots\ldots(6\text{-}1)$$

where $n(A)$ = the number of outcomes favorable to the event A

$n(S)$ = total number of outcomes

Situation II : THE RELATIVE FREQUENCY APPROACH

If we try to apply the classical definition of probability in the second experiment, we find that we cannot say that consumers will equally like the taste of the soup. Moreover, we do not know as to how many persons have been tested. This implies that we should have the past data on people who were administered the soup and the number that liked the taste. In the absence of past data, we have to undertake an experiment, where we administer the taste test on a group of people to check its effect.

The ***Relative Frequency Approach*** is used to compute probability in such cases. As per this approach, the probability of occurrence of an event is given by the observed relative frequency of an event in a very large number of trials. In other words, the probability of occurrence of an event is the ratio of the number of times the event occurs to the total number of trials. The probability of the event B:

$$P(B) = \frac{n}{N} \qquad \ldots\ldots\ldots(6\text{-}2)$$

Where n = the number of times the event occurs

N = total number of trials

It is appreciated in this approach that, in order to take such a measure, we should have the soup tested for a large number of people. In other words, the total number of trials in the experiment should be very large.

Situation III : THE SUBJECTIVE APPROACH

The third situation seems apparently similar to the second one. We may be tempted here to apply the Relative Frequency Approach. We may calculate the probability of the event that the venture is a success as the ratio of number of successful ventures to the total number of such ventures undertaken *i.e.* the relative frequency of successes will be a measure of the probability.

However, the calculation here presupposes that either

(a) it is possible to do an experiment with such ventures, or

(b) that past data on such ventures will be available

In practice, a solar power plant being a relatively new development involving the latest technology, past experiences are not available. Experimentation is also ruled out because of high cost and time involved, unlike the taste testing situation. In such cases, the only way out is the ***Subjective Approach*** to probability. In this approach, we try to assess the probability from our own experiences. We may bring in any information to assess this. In the situation cited, we may, perhaps, look into the performance of the commissioning authority in other new and related technologies.

Therefore the Subjective Approach involves personal judgment, information, intuition, and other subjective evaluation criteria. A physician assessing the probability of a patient's recovery and an expert assessing the probability of success of a merger offer are both making a personal judgment based upon what they know and feel about the situation. The area of subjective probability - which is relatively new, having been first developed in the 1930*s* - is somewhat controversial. One person's subjective probability may very well be different from another person's subjective probability of the same event. We may note here that since the assessment is a purely subjective one, it will vary from person to person and, therefore, subjective probability is also called ***Personal Probability.***

Three Approaches – A Comparative View

As already noted, the different approaches have evolved to cater to different kinds of situations. So these approaches are not contradictory to one another. In fact, these complement each other in the sense that where one fails, the other becomes applicable. These are identical inasmuch as probability is defined as a ratio or a weight assigned to the occurrence of an event. However, in contrast to the Subjective measure of the third approach, the first two approaches - Classical and Relative Frequency - provide an objective measure of probability in the sense that no personal judgment is involved.

We can bring out the commonality between the Classical Approach and the Relative Frequency Approach with the help of an example. Let us assume that we are interested in finding out the chances of getting a head in the toss of a coin. By now, you would have come up with the answer by the Classical Approach, using the argument, that there are two outcomes, heads and tails, which are equally likely. Hence, given that a head can occur only once, the probability is ½ : Consider the following alternative line of argument, where the probability can be estimated using the Relative Frequency

Approach. If we toss the coin for a sufficiently large number of times and note down the number of times the head occurs, the proportion of times that a head occurs will give us the required probability.

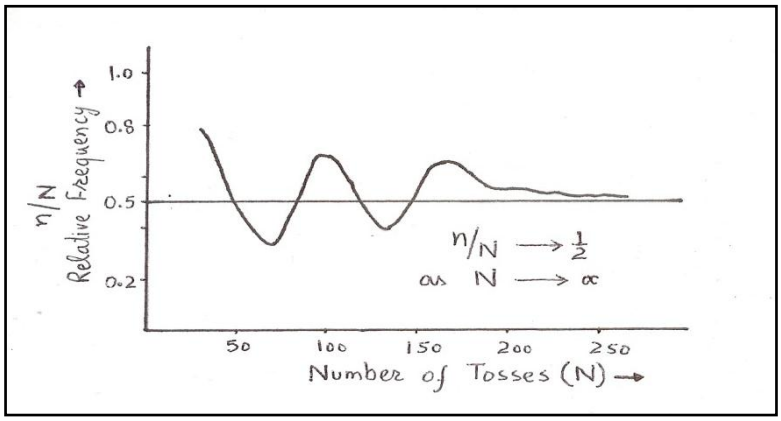


Figure 4-1 ***P(H) = n/N→1/2*** **as** ***N→α***

Thus, given our definition of the approaches, we find both the arguments to be valid. This brings out, in a way, the commonality between the Relative Frequency and the Classical Approach. The difference, however, is that the probability computed by using the Relative Frequency Approach will be tending to be ½ with a large number of trials; moreover an experiment is necessary in this case. In comparison, in the Classical Approach, we know apriori that the chances are ½ , based on our assumption of "equally likely" outcomes.

Example 4-1

A fair coin is tossed twice. Find the probabilities of the following events:

(a) A, getting two heads

(b) B, getting one head and one tail

(c) C, getting at least one head or one tail

(d) D, getting four heads

Solution: Being a Two-Trial Coin Tossing Experiment, it gives rise to the following $O^n = 2^n = 4$, possible equally likely outcomes:

HH HT TH TT

Thus, for the sample space $N(S) = 4$

We can use the Classical Approach to find out the required probabilities.

(a) For the event A, the number of favourable cases are:

$n(A) = 1$ { HH }

So the required probability

$$P(A) = \frac{n(A)}{N(S)} = \frac{1}{4}$$

(b) For the event B, the number of favourable cases are:

$n(B) = 2$ { HT, TH }

So the required probability

$$P(B) = \frac{n(B)}{N(S)} = \frac{2}{4} = \frac{1}{2}$$

(c) For the event C, the number of favourable cases are:

$n(C) = 4$ { HH, HT, TH, TT }

So the required probability

$$P(C) = \frac{n(A)}{N(S)} = \frac{4}{4} = 1$$

(d) For the event D, the number of favourable cases are:

$n(D) = 0$

So the required probability

$$P(D) = \frac{n(D)}{N(S)} = \frac{0}{4} = 0$$

It may be noted that the occurrence of C is certainty, whereas D is an impossible event.

Example 4-2

A newspaper boy wants to find out the chances that on any day he will be able to sell more than 90 copies of *The Times of India*. From his dairy where he recorded the daily sales of the last year, he finds out that out of 365 days, on 75 days he had sold 80 copies, on 144 days he had sold 85 copies, on 62 days he had sold 95 copies and on 84 days he had sold 100 copies of *The Times of India*. Find out the required probability for the newspaper boy.

Solution: Taking the Relative Frequency Approach, we find:

Sales(Event)	No. of Days (Frequency)	Relative Frequency
80	75	75/365
85	144	144/365
95	62	62/365
100	84	84/365

Thus, the number of days when his sales were more than 90 = (62 + 84) days = 146 days

So the required probability

$$P(Sales > 90) = \frac{n}{N} = \frac{146}{365} = 0.4$$

Probability Axioms

All the three approaches to probability theory share the same basic axioms. These axioms are fundamental to probability theory and provide us with unified approach to probability.

The axioms are:

(a) The probability of an event A, written as *P(A),* must be a number between zero and one, both values inclusive. Thus

$$0 \leq P(A) \leq 1 \qquad \ldots\ldots\ldots(6\text{-}3)$$

(b) The probability of occurrence of one or the other of all possible events is equal to one. As *S* denotes the sample space or the set of all possible events, we write

$$P(S) = 1. \qquad \ldots\ldots\ldots(6\text{-}4)$$

Thus in tossing a coin once; *P(a head or a tail) = 1.*

(c) If two events are such that occurrence of one implies that the other cannot occur, then the probability that either one or the other will occur is equal to the sum of their individual probabilities. Thus, in a coin-tossing situation, the occurrence of a head rules out the possibility of occurrence of tail. These events are called ***mutually exclusive events***. In such cases then, if A and B are the two events respectively, then

$$P(A \text{ or } B) = P(A) + P(B)$$

$$i.e.\ P(Head \text{ or } Tail) = P(Head) + P(Tail)$$

It follows from the last two axioms that if two mutually exclusive events form the sample space of the experiment, then

$$P(A \text{ or } B) = P(A) + P(B) = 1; \text{ thus } P(Head) + P(Tail) = 1$$

If two or more events together define the total sample space, the events are said to be collectively exhaustive.

Given the above axioms, we may now define probability as a function, which assigns probability value *P* to each sample point of an experiment abiding by the above axioms. Thus, the axioms themselves define probability.

Interpretation of a Probability

From our discussion so far, we can give a general definition of probability:

Probability is a measure of uncertainty. The probability of event A is a quantitative measure of the likelihood of the event's occurring.

We have also seen that 0 and 1, both values inclusive, sets the range of values that the probability measure may take. In other words $0 \leq P(A) \leq 1$

When an event cannot occur (impossible event), its probability is zero. The probability of the empty set is zero: $P(\Phi) = 0$. In a deck where half the cards are red and half are black, the probability of drawing a green card is zero because the set corresponding to that event is the empty set: there are no green cards.

Events that are certain to occur have probability 1.00. The probability of the entire sample space *S* is equal to 1.00: $P(S) = 1.00$. If we draw a card out of a deck, 1 of the 52 cards in the deck will certainly be drawn, and so the probability of the sample space, the set of all 52 cards, is equal to 1.00.

Within the range of values 0 to 1, the greater the probability, the more confidence we have in the occurrence of the event in question. A probability of 0.95 implies a very high confidence in the occurrence of the event. A probability of 0.80 implies a high confidence. When the probability is 0.5, the event is as likely to occur as it is not to occur. When the probability is 0.2, the event is not very likely to occur. When we assign a probability of 0.05, we believe the event is unlikely to occur, and so on. Figure 4-2 is an informal aid in interpreting probability.

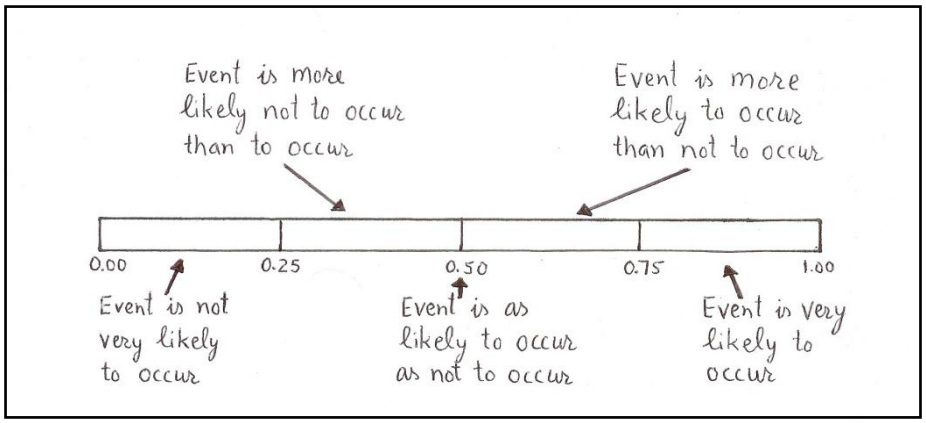


Figure 4-2 Interpretation of a Probability

Note that probability is a measure that goes from 0 to 1. In everyday conversation we often describe probability in less formal terms. For example, people sometimes talk about **odds.** If the odds are 1 to 1, the probability is $\frac{1}{1+1} i.e. \frac{1}{2}$; if the odds are 1 to 2, the probability is $\frac{1}{1+2} i.e. \frac{1}{3}$; and so on. Also, people sometimes say, "The probability is 80 percent." Mathematically, this probability is 0.80.

4.2.3 Probability Rules

We have seen how to compute probabilities in certain situations. The nature of the events was relatively simple, so that direct application of the definition of probability could be used for computation. Quite often, we are interested in the probability of occurrence of more complex events. Consider for example, that you want to find the probability that a king or a club will occur in a draw from a deck of 52 cards. Similarly, on examining couples with two children, if one child is known as a boy, you may be interested in the probability of the event of both the children being boys. These two situations, we find, are not as simple as those discussed in the earlier section. As a sequel to the theoretical development in the field of probability, certain results are available which help us in computing probabilities in such situations. Now we will explore these results through examples.

THE UNION RULE

A very important rule in probability theory, the ***Rule of Unions*** (also called ***Addition Theorem)*** allows us to write the probability of the union of two events in terms of the probabilities of the two events and the probability of their intersection.

Consider two events A and B defined over the sample space S, as shone in Figure 6-3.

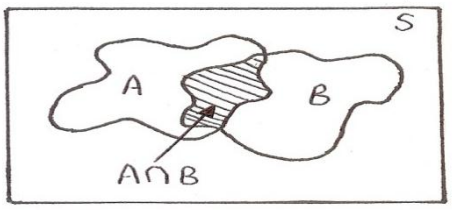


Figure 4-3 Two Overlapping Events A and B

We may define

$$P(A \cup B) = \frac{n(A \cup B)}{N(S)}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{N(S)}$$

$$= \frac{n(A)}{N(S)} + \frac{n(B)}{N(S)} - \frac{n(A \cap B)}{N(S)}$$

$$= P(A) + P(B) - P(A \cap B)$$

Thus, the rule of unions is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \text{……(6-5)}$$

The probability of the intersection of two events $P(A \cap B)$ is called their ***joint probability.*** The meaning of this rule is very simple and intuitive: When we add the probabilities of A and B, we are measuring, or counting, the probability of their intersection *twice—once* when measuring the relative size of A within the sample space and *once* when doing this with B. Since the relative size, or probability, of the intersection of the two sets is counted twice, we subtract it once so that we are left with the true probability of the union of the two events.

The rule of unions is especially useful when we do not have the sample space for the union of events but do have the separate probabilities.

Example 4-3

A card is drawn from a well-shuffled pack of playing cards. Find the probability that the card drawn is either a club or a king.

Solution: Let A be the event that a club is drawn and B the event that a king is drawn. Then,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

= 13/52 + 4/52 – 1/52

= 16/52

= 4/13

Example 4-4

Suppose your chance of being offered a certain job is 0.45, your probability of getting another job is 0.55, and your probability of being offered both jobs is 0.30. What is the probability that you will be offered at least one of the two jobs?

Solution: Let A be the event that the first job is offered and B the event that the second job is offered. Then,

$$P(A) = 0.45 \qquad P(B) = 0.55 \qquad \text{and } P(A \cap B) = 0.30$$

So, the required probability is given as:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

= 0.45 + 0.55 – 0.30

= 0.70

Mutually Exclusive Events

When the sets corresponding to two events are disjoint (*i.e.*, have no intersection), the two events are called ***mutually exclusive*** (see Figure 6-4).

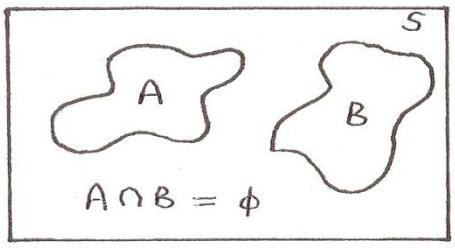


Figure 4-4 Two Mutually Exclusive Events A and B

For mutually exclusive events, the probability of the intersection of the events is zero. This is so because the intersection of the events is the empty set, and we know that the probability of the empty set is zero.

For mutually exclusive events A and B:

$$P(A \cap B) = 0 \qquad \text{……(6.6)}$$

This fact gives us a special rule for unions of mutually exclusive events. Since the probability of the intersection of the two events is zero, there is no need to subtract $P(A \cap B)$ when the probability of the union of the two events is computed. Therefore, For mutually exclusive events A and B:

$$P(A \cup B) = P(A) + P(B) \qquad \text{……(6.7)}$$

This is not really a new rule since we can always use the rule of unions for the union of two events: If the events happen to be mutually exclusive, we subtract zero as the probability of the intersection.

Example 4-5

A card is drawn from a well-shuffled pack of playing cards. Find the probability that the card drawn is either a king or a queen.

Solution: Let A be the event that a king is drawn and B the event that a queen is drawn. Since A and B are two mutually exclusive events, we have,

$$P(A \cup B) = P(A) + P(B)$$

$$= 4/52 + 4/52 = 8/52 = 2/13$$

We can extend the Rule of Unions to three (or more) events. Let A, B, and C be the three events defined over the sample space *S*, as shown in Figure 6-5 Then, the Rule of Unions is

$$P(A \cup B \cup C) =$$

$$P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C) \qquad \text{……(6.8)}$$

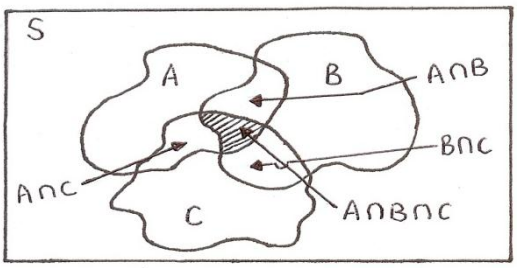


Figure 4-5 Three Overlapping Events A, B and C

When the three events are mutually exclusive (see Figure 8-6), the Rule of Unions is

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \qquad \ldots\ldots\ldots\ldots (6.9)$$

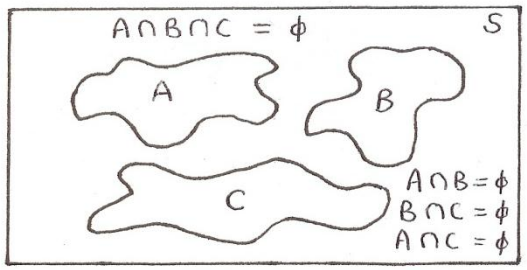


Figure 4-6 Three Mutually Exclusive Events A, B and C

Example 4-6

A card is drawn from a well-shuffled pack of playing cards. Find the probability that the card drawn is

(a) either a heart or an honour or king

(b) either an ace or a king or a queen

Solution: (a) Let A be the event that a heart is drawn, B the event that an honour is drawn and C the event that a king is drawn. So we have

n(A) = 13 *n(B) = 20* *n(C) = 4*

$n(A \cap B) = 5$ $n(B \cap C) = 4$ $n(A \cap C) = 1$

and $n(A \cap B \cap C) = 1$

The required probability (using *Eq. (6.8*) is

$P(A \cup B \cup C)$ = 13/52 + 20/52 + 4/52 – 5/52 – 4/52 – 1/52 +1/52

= 28/52

= 7/13

(b) Let A be the event that an ace is drawn, B the event that a king is drawn and C the event that a queen is drawn. So we have

n(A) = 4 *n(B) = 4* *n(C) = 4*

Since A, B and C are mutually exclusive events, the required probability (using *Eq. (6.9*) is

$P(A \cup B \cup C)$ = 4/52 + 4/52 + 4/52

= 12/52

= 3/13

THE COMPLEMENT RULE

The ***Rule of Complements*** defines the probability of the complement of an event in terms of the probability of the original event. Consider event A defined over the sample space *S*. The complement of set A, denoted by $\overline{A}$, is a subset, which contains all outcomes, which do not belong to A (see Figure 6-7).

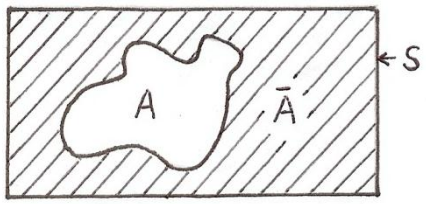


Figure 4-7 Complement of an Event

In other words $A + \overline{A} = S$

so $P(A + \overline{A}) = P(S)$

or $P(A) + P(\overline{A}) = 1$

or $P(\overline{A}) = 1 - P(A)$ ……*(6.10)*

Eq. (8.10) is our Rule of Complements. As a simple example, if the probability of rain tomorrow is 0.3, then the probability of no rain tomorrow must be 1 - 0.3 = 0.7. If the probability of drawing a king is 4/52, then the probability of the drawn card's not being a king is 1 - 4/52 = 48/52.

Example 4-7

Find the probability of the event of getting a total of less than 12 in the experiment of throwing a die twice.

Solution: Let A be the event of getting a total 12.

Then we have,

A = {6,6} and *P(A)* = 1/36

The event of getting a total of less than 12 is the complement of A, so the required probability is

$P(\overline{A}) = 1 - P(A)$

$P(\overline{A}) = 1 - 1/36$

$P(\overline{A}) = 35/36$

THE CONDITIONAL PROBABILITY RULE

As a measure of uncertainty, probability depends on information. We often face situations where the probability of an event A is influenced by the information that another event B has occurred. Thus, the probability we would give the event "Xerox stock price will go up tomorrow" depends on what we know about the company and its performance; the probability is *conditional* upon our information set. If we know much about the company, we may assign a different probability to the event than if we know little about the company. We may define the probability of event A *conditional upon* the occurrence of event B. In this example, event A may be the event that the stock will go up tomorrow, and event B may be a favorable quarterly report.

Consider two events A and B defined over the sample space *S*, as shown in Figure 4-8

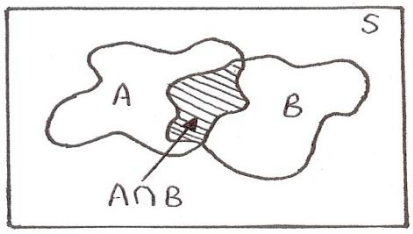


Figure 4-8 Conditional Probability of Event A

Thus, the probability of event A given the occurrence of event B is

$$P(A/B) = \frac{n(A \cap B)}{n(B)}$$

$$P(A/B) = \frac{n(A \cap B)/N}{n(B)/N}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \qquad \ldots\ldots\ldots\ldots(6.11)$$

The vertical line in $P(A/B)$ is read *given,* or *conditional upon.*

Therefore, the probability of event A given the occurrence of event B is defined as the probability of the intersection of A and B, divided by the probability of event B.

Example 4-8

For an experiment of throwing a die twice, find the probability:

(a) of the event of getting a total of 9, given that the die has shown up points between 4 and 6 (both inclusive)

(b) of the event of getting points between 4 and 6 (both inclusive), given that a total of 9 has already been obtained

Solution: Let getting a total 9 be the event A and the die showing points between 4 and 6 (both inclusive) be the event B

Thus, *N(S)* = 36 and A = {(3,6) (4,5) (5,4) (6,3)}

B = {(4,4) (4,5) (4,6) (5,4) (5,5) (5,6) (6,4) (6,5) (6,6)}

and $P(A \cap B)$ = {(4,5) (5,4)}

So *n(A)* = 4 *n(B)* = 9 $n(A \cap B)$ = 2

So the required probabilities are

(a) $$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A/B) = \frac{2/36}{9/36}$$

$$P(A/B) = \frac{2}{9}$$

(b) $$P(B/A) = \frac{P(B \cap A)}{P(A)}$$

$$P(B/A) = \frac{2/36}{4/36}$$

$$P(B/A) = \frac{1}{2}$$

THE PRODUCT RULE

The ***Product Rule*** (also called ***Multiplication Theorem)*** allows us to write the probability of the simultaneous occurrence of two (or more) events.

In the conditional probability rules

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B/A) = \frac{P(B \cap A)}{P(A)}$$

$A \cap B$ or $B \cap A$ is the event A and B occur simultaneously. So rearranging the conditional probability rules, we have our ***Product Rule***

$P(A \cap B) = P(A/B).P(B)$ and $P(A \cap B) = P(B/A).P(A)$ *……(6.12)*

The Product Rule states that the probability that both A and B will occur simultaneously is equal to the probability that B (or A) will occur multiplied by the conditional probability that A (or B) will occur, when it is known that B (or A) is certain to occur or has already occurred.

Example 4-9

A box contains 10 balls out of which 2 are green, 5 are red and 3 are black. If two balls are drawn at random, one after the other without replacement, from the box. Find the probabilities that:

(a) both the balls are of green color

(b) both the balls are of black color

(c) both the balls are of red color

(d) the first ball is red and the second one is black

(e) the first ball is green and the second one is red

Solution: (a) $P(G_1 \cap G_2) = P(G_2/G_1).P(G_1)$

$$= \frac{1}{9} x \frac{2}{10}$$

$$= \frac{1}{45}$$

(b) $P(B_1 \cap B_2) = P(B_2/B_1).P(B_1)$

$$= \frac{2}{9} x \frac{3}{10}$$

$$= \frac{1}{15}$$

(c) $P(R_1 \cap R_2) = P(R_2 / R_1).P(R_1)$

$$= \frac{4}{9} x \frac{5}{10}$$

$$= \frac{2}{9}$$

(d) $P(R_1 \cap B_2) = P(B_2 / R_1).P(R_1)$

$$= \frac{3}{9} x \frac{5}{10}$$

$$= \frac{1}{6}$$

(e) $P(G_1 \cap R_2) = P(R_2 / G_1).P(G_1)$

$$= \frac{5}{9} x \frac{2}{10}$$

$$= \frac{1}{9}$$

Example 4-10

A consulting firm is bidding for two jobs, one with each of two large multinational corporations. The company executives estimate that the probability of obtaining the consulting job with firm A, event A, is 0.45. The executives also feel that if the company should get the job with firm A, then there is a 0.90 probability that firm B will also give the company the consulting job. What are the company's chances of getting both jobs?

Solution: We are given *P(A)* = 0.45. We also know that *P(B* / A) = 0.90, and we are looking for $P(A \cap B)$, which is the probability that both A and B will occur.

So $P(A \cap B) = P(B / A).P(A)$

$$P(A \cap B) = 0.90x0.45$$

$$= 0.405$$

Independent Events

Two events are said to be ***independent*** of each other if the occurrence or non-occurrence of one event in any trial does not affect the occurrence of the other event in any trial. Events A and B are ***independent*** of each other if and only if the following three conditions hold:

Conditions for the independence of two events A and B:

$$P(A/B) = P(A) \quad \text{……(6.13a)}$$

$$P(B/A) = P(B) \quad \text{……(6.13b)}$$

and $$P(A \cap B) = P(A).P(B) \quad \text{……(6.14)}$$

The first two equations have a clear, intuitive appeal. The top equation says that when A and B are independent of each other, then the probability of A stays the same even when we know that B has occurred - it is a simple way of saying that knowledge of B tells us nothing about A when the two events are independent. Similarly, when A and B are independent, then knowledge that A has occurred gives us absolutely no information about B and its likelihood of occurring.

The third equation, however, is the most useful in applications. It tells us that when A and B are independent (and only when they are independent), we can obtain the probability of the joint occurrence of A and B (*i.e.* the probability of their intersection) simply by multiplying the two separate probabilities. This rule is thus called the ***Product Rule for Independent Events***.

As an example of independent events, consider the following: Suppose I roll a single die. What is the probability that the number 5 will turn up? The answer is 1/6. Now suppose that I told you that I just tossed a coin and it turned up heads. What is now the probability that the die will show the number 5? The answer is unchanged, 1/6, because events of the die and the coin are independent of each other. We see that $P(6/H) = P(6)$, which is the first rule above.

The rules for union and intersection of two independent events can be extended to sequences of more than two events.

Intersection Rule

The probability of the intersection of several independent events A_1, A_2, ……is just the product of separate probabilities *i.e.*

$$P(A_1 \cap A_2 \cap A_3) = P(A_1).P(A_2).P(A_3)..... \quad \text{……(6.15)}$$

Union Rule

The probability of the union of several independent events A_1, A_2, ……is given by the following equation

$$P(A_1 \cup A_2 \cup A_3 \cup .....) = 1 - P(\overline{A_1}).P(\overline{A_2}).P(\overline{A_3})..... \quad \text{……(6.16)}$$

The union of several events is the event that at least one of the events happens.

Example 4-11

A problem in mathematics is given to five students A, B,C, D and E. Their chances of solving it are 1/2, 1/3, 1/3, 1/4 and 1/5 respectively. Find the probability that the problem will

(a) not be solved

(b) be solved

Solution: (a) The problem will not be solved when none of the students solve it. So the required probability is:

$$\begin{aligned} P(problem will not be solved) &= P(\overline{A}).P(\overline{B}).P(\overline{C}).P(\overline{D}).P(\overline{E}) \\ &= (1-1/2).(1-1/3).(1-1/3).(1-1/4).(1-1/5) \\ &= 2/15 \end{aligned}$$

(b) The problem will be solved when at least one of the students solve it. So the required probability is:

$$\begin{aligned} P(A \cup B \cup C \cup D \cup E) &= 1 - P(\overline{A}).P(\overline{B}).P(\overline{C}).P(\overline{D}).P(\overline{E}) \\ &= 1 - 2/15 \\ &= 13/15 \end{aligned}$$

4.2.4 Bayes' Theorem

As we have already noted in the introduction, the basic objective behind calculating probabilities is to help us in making decisions by quantifying the uncertainties involved in the situations. Quite often, whether it is in our personal life or our work life, decision-making is an ongoing process. Consider for example, a seller of winter garments, who is interested in the demand of the product. In deciding on the amount he should stock for this winter, he has computed the probability of selling different quantities and has noted that the chance of selling a large quantity is very high. Accordingly, he has taken the decision to stock a large quantity of the product. Suppose, when finally the winter comes and the season ends, he discovers that he is left with a large quantity of stock. Assuming that he is in this business, he feels that the earlier probability calculation should be updated given the new experience to help him decide on the stock for the next winter.

Similar to the situation of the seller of winter garment, situations exist where we are interested in an event on an ongoing basis. Every time some new information is available, we do revise our odds

mentally. This revision of probability with added information is formalised in probability theory with the help of famous Bayes' Theorem. The theorem, discovered in 1761 by the English clergyman Thomas Bayes, has had a profound impact on the development of statistics and is responsible for the emergence of a new philosophy of science. Bayes himself is said to have been unsure of his extraordinary result, which was presented to the Royal Society by a friend in 1763 - after Bayes' death. We will first understand *The Law of Total Probability*, which is helpful for derivation of Bayes' Theorem.

The Law of Total Probability

Consider two events A and B. Whatever may be the relation between the two events, we can *always* say that the probability of A is equal to the probability of the intersection of A and B, plus the probability of the intersection of A and the complement of B (event $\overline{B}$).

$$P(A) = P(A \cap B) + P(A \cap \overline{B})$$

or $$P(A) = P(A/B).P(B) + P(A/\overline{B}).P(\overline{B}) \quad \ldots\ldots\ldots\ldots (6.17)$$

The sets B and $\overline{B}$ form a ***partition*** of the sample space. A partition of a space is the division of the sample space into a set of events that are mutually exclusive (disjoint sets) and cover the whole space. Whatever event B may be, either B or $\overline{B}$ must occur, but not both. Figure 6-9 demonstrates this situation and the law of total probability.

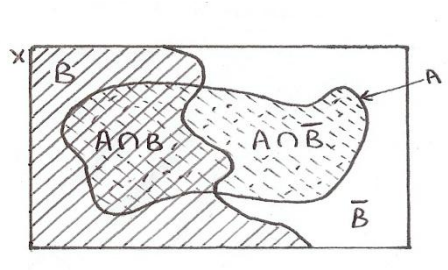


Figure 4-9 Total Probability of Event A

The law of total probability may be extended to more complex situations, where the sample space *X* is partitioned into more than two events. Say, we have partition of the space into a collection of *n* sets B_1, B_2,……B_n .The law of total probability in this situation is:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

or $$P(A) = \sum_{i=1}^{n} P(A/B_i).P(B_i) \qquad \text{……(6.18)}$$

Figure 8-10 shows the partition of a sample space into five events B_1, B_2, B_3, B_4 and B_5 ; and shows their intersections with set A.

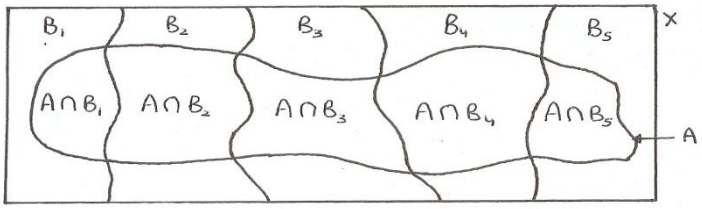


Figure 4-10 Total Probability of Event A

We can demonstrate the rule with a more specific example. Let us define A as the event that an honour card is drawn out of a deck of 52 cards (the honour cards are the aces, kings, queens, jacks and 10). Letting H, C, D, and S denote the events that the card drawn is a heart, club, diamond, or spade, respectively, we find that the probability of an honour card is:

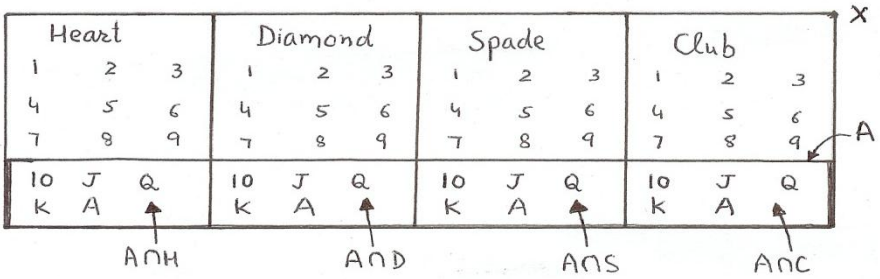


Figure 4-11 Total Probability of Event A: An Honour Card

$$P(A) = P(A \cap H) + P(A \cap C) + P(A \cap D) + P(A \cap S)$$

= 5/52 + 5/52 + 5/52 + 5/52

= 20/52

= 5/13

which is what we know the probability of an honour card to be just by counting 20 honour cards out of a total of 52 cards in the deck. The situation is shown in Figure 6-11.

As can be seen from the figure, the event A is the set addition of the intersections of A with each of the four sets H, D, C, and S.

Example 4-12

A market analyst believes that the stock market has a 0.70 probability of going up in the next year if the economy should do well, and a 0.20 probability of going up if the economy should not do well during the year. The analyst believes that there is a 0.80 probability that the economy will do well in the coming year. What is the probability that stock market will go up next year?

Solution: Let U be the event that the stock market will go and W is the event that the economy will do well in the coming year.

Then

$$\begin{aligned} P(U) &= P(U/W).P(W) + P(U/\overline{W}).P(\overline{W}) \\ &= (0.70)(0.80) + (0.20)(0.20) \\ &= 0.56 + 0.04 \\ &= 0.60 \end{aligned}$$

BAYES' THEOREM

We will now develop the Bayes' theorem. Bayes' theorem is easily derived from the law of total probability and the definition of conditional probability.

By definition of conditional probability, we have

$$P(B/A) = \frac{P(B \cap A)}{P(A)} \qquad \ldots\ldots\ldots(6.19)$$

By product rule, we have

$$P(B \cap A) = P(A \cap B) = P(A/B).P(B) \qquad \ldots\ldots\ldots\ldots(6.20)$$

Substituting *Eq.(6.19)* in *Eq.(6.20)*, we have

$$P(B/A) = \frac{P(A/B).P(B)}{P(A)} \qquad \ldots\ldots\ldots\ldots(6.21)$$

By the law of total probability, we have

$$P(A) = P(A/B).P(B) + P(A/\overline{B}).P(\overline{B})$$

Substituting this expression for *P(A)* in the denominator of *Eq.(6.21)*, we have the Bayes' theorem

$$P(B/A) = \frac{P(A/B).P(B)}{P(A/B).P(B) + P(A/\overline{B}).P(\overline{B})} \qquad \ldots\ldots\ldots\ldots(6.22)$$

Thus the theorem allows us to reverse the conditionality of events: we can obtain the probability of B given A from the probability of A given B(and other information). As we see from the theorem, the probability of B given A is obtained from the probabilities of B and $\overline{B}$ and from the conditional probabilities of A given B and A given $\overline{B}$.

The probabilities *P(B) and P(* $\overline{B}$ *)* are called ***prior probabilities*** of the events B and $\overline{B}$; the probability *P(B /A)* is called the ***posterior probability*** of B. It is possible to write Bayes' theorem in terms of $\overline{B}$ and A, thus giving the posterior probability of $\overline{B}$, *P(* $\overline{B}$ /A). Bayes' theorem may be viewed as a means of transforming our prior probability of an event B into a posterior probability of the event B - posterior to the known occurrence of event A.

The Bayes' theorem can be extended to a partition of more than two sets. This is done by using the law of total probability involving a partition in sets B_1, B_2, …… B_n .The resulting form of Bayes' theorem is:

$$P(B_i / A) = \frac{P(A/B_i).P(B_i)}{\sum_{i=1}^{n} P(A/B_i).P(B_i)} \qquad \text{……(6.23)}$$

The theorem gives the probability of one of the sets in the partition B_i, given the occurrence of event A.

Example 4-13

An Economist believes that during periods of high economic growth, the Indian Rupee appreciates with probability 0.70; in periods of moderate economic growth, it appreciates with probability 0.40; and during periods of low economic growth, the Rupee appreciates with probability 0.20.During any period of time the probability of high economic growth is 0.30; the probability of moderate economic growth is 0.50 and the probability of low economic growth is 0.20. Suppose the Rupee value has been appreciating during the present period. What is the probability that we are experiencing the period of (a) high, (b) moderate, and (c) low, economic growth?

Solution: Our partition consists of three events: high economic growth (event H), moderate economic growth (event M) and low economic growth (event L). The prior probabilities of these events are:

P(H) = 0.30 *P(M)* = 0.50 *P(L)* = 0.20

Let A be the event that the rupee appreciates. We have the conditional probabilities

P(A / H) = 0.70 *P(A / M)* = 0.40 *P(A / L)* = 0.20

By using the Bayes' theorem we can find out the required probabilities

P(H /A), P(M / A) and *P(L / A)*

(a) *P(H /A)*

$$P(H/A) = \frac{P(A/H).P(H)}{P(A/H).P(H) + P(A/M).P(M) + P(A/L).P(L)}$$
$$= \frac{(0.70)(0.30)}{(0.70)(0.30) + (0.40)(0.50) + (0.20)(0.20)}$$
$$= 0.467$$

(b) *P(M /A)*

$$P(M/A) = \frac{P(A/M).P(M)}{P(A/H).P(H) + P(A/M).P(M) + P(A/L).P(L)}$$
$$= \frac{(0.40)(0.50)}{(0.70)(0.30) + (0.40)(0.50) + (0.20)(0.20)}$$
$$= 0.444$$

(c) *P(L /A)*

$$P(L/A) = \frac{P(A/L).P(L)}{P(A/H).P(H) + P(A/M).P(M) + P(A/L).P(L)}$$
$$= \frac{(0.20)(0.20)}{(0.70)(0.30) + (0.40)(0.50) + (0.20)(0.20)}$$
$$= 0.089$$

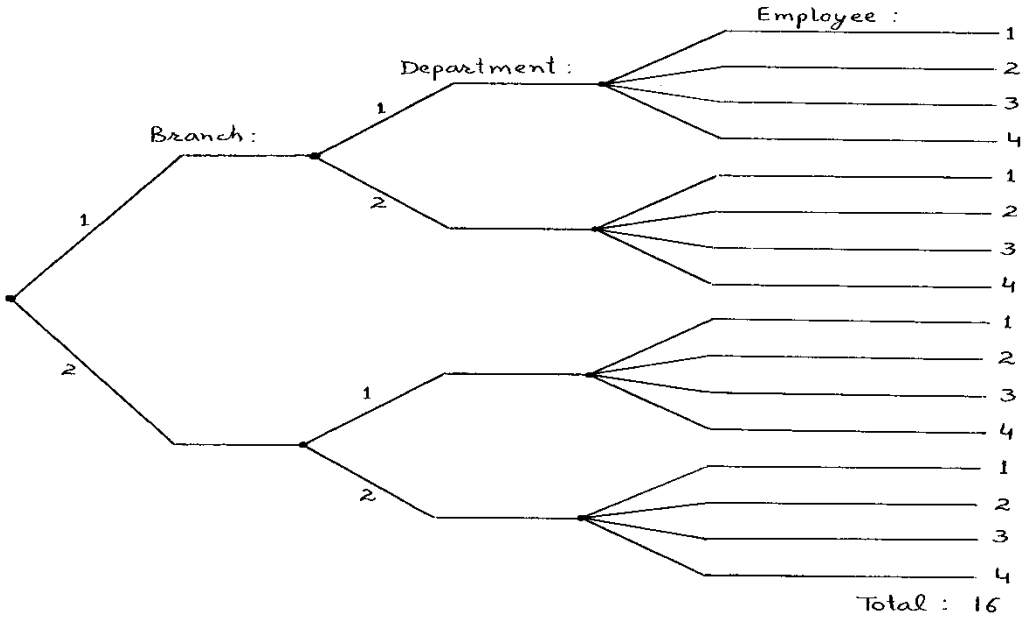
4.3 SOME COUNTING CONCEPTS

If there are *n* events and event *i* can occur in N_i possible ways, then the number of ways in which the sequence of *n* events may occur is

$$N_1.\ N_2.\ N_3 \ldots\ldots\ldots N_n \qquad \ldots\ldots\ldots (6.24)$$

Suppose that a bank has two branches, each branch has two departments, and each department has four employees. Then there are (2)(2)(4) choices of employees, and the probability that a particular one will be randomly selected is 1/(2)(2)(4) = 1/16.

We may view the choice as done sequentially: First a branch is randomly chosen, then a department within the branch, and then the employee within the department. This is demonstrated in the tree diagram in Figure 4-12.



For any positive integer *n,* we define ***n* factorial** as

$n(n-1)(n-2)\ldots\ldots 1$ ……*(6.25)*

We denote *n* factorial by *n*!. The number *n*! is the number of ways in which *n* objects can be ordered. By definition, 0! = 1.

For example, 5! is the number of possible arrangements of five objects. We have 5! = (5) (4) (3) (2) (1) = 120. Suppose that five applications arrive at a center on the same day, all written at different times. What is the probability that they will be read in the order in which they were written? Since there are 120 ways to order five applications, the probability of a particular order (the order in which the applications were written) is 1/120.

Permutations are the possible ordered selections of *r* objects out of a total of *n* objects. The number of permutations of *n* objects taken *r* at a time is denoted by ${}^nP_r$

$$ {}^nP_r = \frac{n!}{(n-r)!} \qquad \ldots\ldots\ldots(6.26)$$

Suppose that 4 people are to be randomly chosen out of 10 people who agreed to be interviewed in a market survey. The four people are to be assigned to four interviewers. How many possibilities are there? The first interviewer has 10 choices, the second 9 choices, the third 8, and the fourth 7. Thus,

there are (10)(9)(8)(7) = 5,040 selections. We can see that this is equal to *n(n - l)(n - 2) ……. (n - r +* 1), which is equal to $^{n}P_{r} = \frac{n!}{(n-r)!}$!.

If choices are made randomly, the probability of any predetermined assignment of 4 people out of a group of 10 is 1/5,040.

Combinations are the possible selections of *r* items from a group of *n* items regardless of the order of selection. The number of combinations is denoted by $^{n}C_{r}$ and is read *n choose r*. We define the number of combinations of *r* out of *n* elements as

$$^{n}C_{r} = \frac{n!}{r!(n-r)!} \qquad \text{……..(6.27)}$$

Suppose that 3 out of the 10 members of the board of directors of a large corporation are to be randomly selected to serve on a particular task committee. How many possible selections are there? Using *Eq. (6.27),* we find that the number of combinations is $^{n}C_{r} = \frac{n!}{r!(n-r)!}$ = 10!/(3!7!) = 120.

If the committee is chosen in a truly random fashion, what is the probability that the three-committee members chosen will be the three senior board members? This is 1 combination out of a total of 120, so the answer is 1/120 = 0.00833.

4.4 CHECK YOUR PROGRESS

1. Subjective Approach involves personal judgment, information, intuition, and other …… criteria.
2. Rule of Unions allows us to write the probability of the union of two events in terms of the probabilities of the two events and the probability of their ……. .
3. The Rule of Complements defines the probability of the …… of an event in terms of the probability of the original event.
4. The Product Rule allows us to write the probability of the …… occurrence of two (or more) events.
5. Bayes' theorem, discovered in 1761 by the English clergyman …… .

4.5 SUMMARY

In life, we may also take some decisions regarding the price increase, reducing sales expenses *etc.* to manage the demand. However, in order to make such decisions, we need to quantify the chances of

different quantities of demand in the coming year. Probability theory provides us with the ways and means to quantify the uncertainties involved in such situations. Three different approaches to the definition and interpretation of probability have evolved, mainly to cater to the three different types of situations under which probability measures are normally required. There are different rules to find the probability like union rule, complement rule and product rule. Every time some new information is available, we do revise our odds mentally. This revision of probability with added information is formalized in probability theory with the help of famous Bayes' Theorem.

4.6 KEYWORDS

Experiment: It is a process that leads to one of several possible outcomes. An outcome of an experiment is some observation or measurement.

Sample space: It is the universal set S pertinent to a given experiment. It is the set of all possible outcomes of an experiment.

Event: It is a subset of a sample space. It is a set of basic outcomes. It can say that the event occurs if the experiment gives rise to a basic outcome belonging to the event.

Classical Approach: In this, probability of an event is defined as the *relative size* of the event with respect to the size of the sample space.

Relative Frequency Approach: It is used to compute probability in such cases. As per this approach, the probability of occurrence of an event is given by the observed relative frequency of an event in a very large number of trials.

Permutations: It is the possible ordered selections of *r* objects out of a total of *n* objects.

Combinations: It is the possible selections of *r* items from a group of *n* items regardless of the order of selection.

4.7 SELF-ASSESSMENT TEST

1. Explain what you understand by the term ‘probability’. How the concept of probability is is relevant to decision making under uncertainty?
2. What are different approaches to the definition of probability? Are these approaches contradictory to one another? Which of these approaches you will apply for calculating the probability that:
    (a) A leap year selected at random, will contain 53 Monday.
    (b) An item, selected at random from a production process, is defective.

(c) Mr. Bhupinder S. Hooda will win the assembly election from Kiloi.

3. With the help of an example explain the meaning of the following:
   (a) Random experiment, and sample space
   (b) An event as a subset of sample space
   (c) Equally likely events
   (d) Mutually exclusive events.
   (e) Exhaustive events
   (f) Elementary and compound events.
4. A proofreader is interested in finding the probability that the number of mistakes in a page will be less than 10. From his past experience he finds that out of 3600 pages he has proofed, 200 pages contained no errors, 1200 pages contained 5 errors, and 2200 pages contained 11 or more errors. Can you help him in finding the required probability?
5. State and develop the Addition Theorem of probability for:
   (a) mutually exclusive events
   (b) overlapping events
   (c) complementary events
6. Explain the concept of conditional probability with the help of a suitable example.
7. State and develop the Multiplication Theorem of probability for:
   (a) dependent events
   (b) independent event
8. State the Bayes' Theoram of probability. Using an appropriate example, develop the Bayesian probability rule and generalize it.
9. What do you understand by permutations and combinations?
   (a) In how many ways we can select three players out of 12 players of the Indian Cricket team, for playing in the World XI team?
   (b) In how many ways can a sub-committee of 2 out of 6 members of the executive committee of the employees' association be constituted?
10. What is the probability that a non leap year, selected at random, will contain
    (a) 52 Sundays? (b) 53 Sundays? (c) 54 Sundays?

11. A card is drawn at random from well shuffled deck of 52 cards, find the probability that
    (a) the card is either a club or diamond
    (b) the card is not a king
    (c) the card is either a face card or a club card.
12. From a well-shuffled deck of 52 cards, two cards are drawn at random.
    (a) If the cards are drawn simultaneously, find the probability that these consists of *(i)* both clubs, *(ii)* a king and a queen, *(iii)* a face card and a 8.
    (b) If the cards are drawn one after the other with replacement. Find the probability that these consists of *(i)* both clubs, *(ii)* a king and a queen, *(iii)* a face card and a 8.
13. A problem in mathematics is given to four students A, B,C, and D their chances of solving it are 1/2 , 1/3, 1/4 and 1/5 respectively. Find the probability that the problem will
    (a) be solved
    (b) not be solved
14. The odds that A speaks the truth are 3:2 and the odds that B does so are 7:3. In what percentage of cases are they likely to
    (a) contradict each other on an identical point?
    (b) agree each other on an identical point?
15. Among the sales staff engaged by a company 60% are males. In terms of their professional qualifications, 70% of males and 50% of females have a degree in marketing. Find the probability that a sales person selected at random will be
    (a) a female with degree in marketing
    (b) a male without degree in marketing
16. A and B play for a prize of Rs. 10,000. A is to throw a die first and is to win if he throws 1: If A fails, B it to throw and is to win if he throws 2 or 1. If B fails, A is to throw again and to win if he throws 3, 2 or 1: and so on. Find their respective expectations.
17. A factory has three units A, B, and C. Unit A produces 50% of its products, and units B and C each produces 25% of the products. The percentage of defective items produced by A, B, and C units are 3%, 2% and 1%, respectively. If an item is selected at random from the total production of the factory is found defective, what is the probability that it is produced by:
    (a) Unit A (b) Unit B (c) Unit C

4.8 ANSWERS TO CHECK YOUR PROGRESS

1. Subjective evaluation
2. Intersection
3. Complement
4. Simultaneous
5. Thomas Bayes

4.9 REFERNCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.

Subject: Business Statistics	
Course Code: MBA-102	Author: Anil Kumar
Lesson: 05	Vetter: Dr. Karam Pal
PROBABILITY DISTRIBUTIONS-I	

STRUCTURE

5.1 Learning Objectives

5.2 Introduction

- 5.2.1 Discrete Probability Distribution
- 5.2.2 Bernoulli Random Variable

5.3 The Binomial Distribution

- 5.3.1 Conditions for a Binomial Random Variable
- 5.3.2 Binomial Probability Function
- 5.3.3 Characteristics of a Binomial Distribution
- 5.3.4 Importance of the Binomial Distribution
- 5.3.5 Fitting a Binomial Distribution

5.4 The Poisson Distribution

- 5.4.1 Characteristics of Poisson Distribution
- 5.4.2 Role of the Poisson Distribution
- 5.4.3 Fitting a Poisson Distribution

5.5 Check your Progress

5.6 Summary

5.7 Keywords

5.8 Self-Assessment Test

5.9 Answers to check your progress

5.10 References/Suggested Readings

5.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understand the concept of random variable and discrete probability distributions.
- Appreciate the usefulness of probability distributions in decision-making
- Identify situations where Binomial and Poisson probability distributions can be applied.

5.2 INTRODUCTION

In many situations, our interest does not lie in the outcomes of an experiment as such; we may find it more useful to describe a particular property or attribute of the outcomes of an experiment in numerical terms. For example, out of three births; our interest may be in the matter of the probabilities of the number of boys. Consider the sample space of 8 equally likely sample points.

GGG	GGB	GBG	BGG
GBB	BGB	BBG	BBB

Now look at the variable ***"the number of boys out of three births".*** This number varies among sample points in the sample space and can take values *0,1,2,3,* and it is random –given to chance.

"A random variable is an uncertain quantity whose value depends on chance."

A random variable may be…

- ***Discrete*** if it takes only a countable number of values. For example, number of dots on two dice, number of heads in three coin tossing, number of defective items, number of boys in three births and so on.
- ***Continuous*** if can take on any value in an interval of numbers (*i.e.* its possible values are unaccountably infinite). For example, measured data on heights, weights, temperature, and time and so on.

A random variable has a probability law - a rule that assigns probabilities to different values of the random variable. This probability law - the probability assignment is called the ***probability distribution*** of the random variable. We usually denote the random variable by *X*. In this lesson, we will discuss discrete probability distributions. Continuous probability distributions will be discussed in the next lesson.

5.2.1 Discrete Probability Distribution

The random variable *X* denoting *"the number of boys out of three births",* we introduced in the introduction of the lesson, is a discrete random variable; so it will have a discrete probability distribution. It is easy to visualize that the random variable *X* is a function of sample space. We can see the correspondence of sample points with the values of the random variable as follows:

BBB *(X=0)*	GGB	GBG *(X=1)*	BGG
GBB	BGB *(X=2)*	BBG	BBB *(X=3)*

The correspondence between sample points and the value of the random variable allows us to determine the probability distribution of *X* as follows:

P(X=0) = 1/8 since one out of 8 equally likely points leads to *X* = 0

P(X=1) = 3/8 since three out of 8 equally likely points leads to *X* = 1

P(X=2) = 3/8 since three out of 8 equally likely points leads to *X* = 2

P(X=3) = 1/8 since one out of 8 equally likely points leads to *X* = 3

The above probability statement constitute the probability distribution of the random variable *X = number of boys in three births*. We may appreciate how this probability law is obtained simply by associating values of *X* with sets in the sample space. (For example, the set GGB, GBG, BGG leads to *X = 1*). We may write down the probability distribution of *X* in table format (see Table 5-1) or we may plot it graphically by means of probability Histogram (see Figure 5-1*a*) or a Line chart (see Figure 5-1*b*).

Table 5-1

Probability Distribution of the Number of Boys out of Three Births

No. of Boys X	Probability $P(X)$
0	1/8
1	3/8
2	3/8
3	1/8

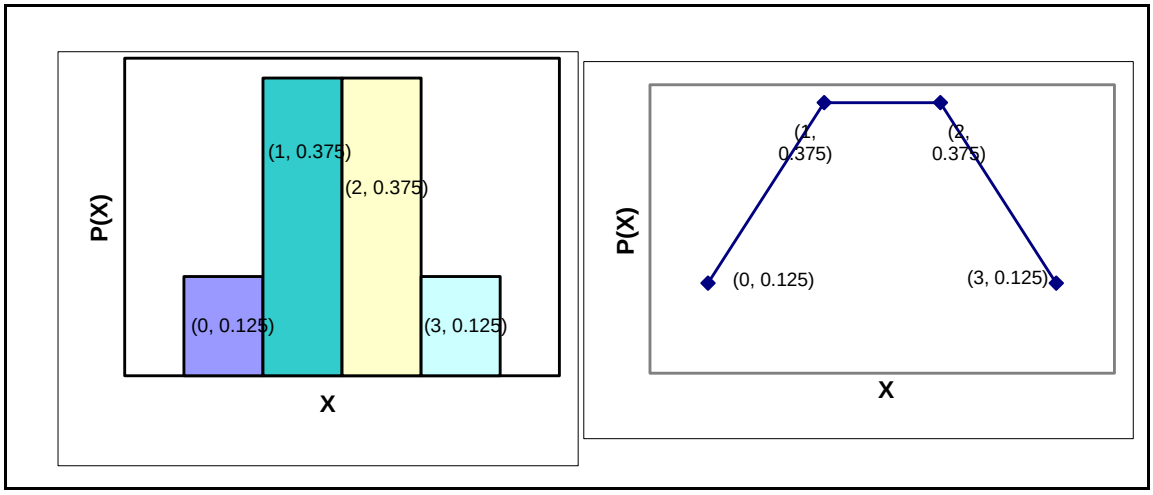


Figure 5-1 Probability Distribution of the Number of Boys out of Three Births

The probability distribution of a discrete random variable X must satisfy the following two conditions:

1. $P(X = x) \geq 0$ for all values x
2. $\sum_{all\,x} P(X = x) = 1$

These conditions must hold because the $P(X = x)$ values are probabilities. First condition specifies that all probabilities must be greater than or equal to zero, as we know from Lesson 6.

For the second condition, we note that for each value x, $P(x) = P(X = x)$ is the probability of the event that the random variable equals x. Since by definition all x means all the values the random variable X may take, and since X may take on only one value at a time, the occurrences of these values are mu-

tually exclusive events, and one of them must take place. Therefore, the sum of all the probabilities $P(X = x)$ must be 1.00.

Cumulative Distribution Function

The probability distribution of a discrete random variable lists the probabilities of occurrence of different values of the random variable. We may be interested in *cumulative* probabilities of the random variable. That is, we may be interested in the probability that the value of the random variable is *at most* some value *x.* This is the sum of all the probabilities of the values *i* of *X* that are less than or equal to *x.*

The *cumulative distribution function* (also called *cumulative probability function)* $F(X = x)$ of a discrete random variable *X* is

$$F(X = x) = P(X \leq x) = \sum_{all i \leq x} P(i)$$

For example, to find the probability of at most two boys out of three births, we have

$$F(X = 2) = P(X \leq 2) = \sum_{all i \leq 2} P(i)$$

$$= P(X = 0) + P(X = 1) + P(X = 2)$$

$$= 1/3 + 3/8 + 3/8$$

$$= 7/8$$

Expected Value and Variance of a Discrete Random Variable

The expected value of a discrete random variable *X* is equal to the sum of all values of the random variable, each value multiplied (weighted) by its probability.

$$\mu = E(X) = \sum_{all x} x.P(x)$$

The variance of a discrete random variable is given by

$$\sigma^2 = V(X) = E[(X - \mu)^2] = \sum_{all x} (x - \mu)^2 . P(x)$$

In the same way we can calculate the other summary measures *viz.* skewness, kurtosis and moments.

Probability Distributions are Theoretical Distributions

Consider a random variable *X* that measures the *"number of heads"* in a three-trial coin tossing experiment. The probability distribution of *X* will be

X	:	0	1	2	3
P(X)	:	1/8	3/8	3/8	1/8

Now imagine this experiment is repeated 200 times, we may expect 'no head' and 'three heads' will each occur 25 times; 'one head' and 'two heads' each will occur 75 times. Since these results are what we expect on the basis of theory, the resultant distribution is called a ***theoretical or expected distribution.***

However, when the experiment is actually performed 200 times, the results, which we may actually obtain, will normally differ from the theoretically expected results. It is quite possible that in actual experiment 'no head' and 'three heads' may occur 20 and 28 times respectively and 'one head' and 'two heads' may occur 66 and 86 times respectively. The distribution so obtained through actual experiment is called the ***empirical or observed distribution.***

In practice, however, assessing the probability of every possible value of a random variable through actual experiment can be difficult, even impossible, especially when the probabilities are very small. But we may be able to find out what type of random variable the one at hand is by examining the causes that make it random. Knowing the type, we can often approximate the random variable to a standard one for which convenient formulae are available.

The proper identification of experiments with certain known processes in Probability theory can help us in writing down the probability distribution function. Two such processes are the ***Bernoulli Process*** and the ***Poisson Process.*** The standard discrete probability distributions that are consequent to these processes are the ***Binomial*** and the ***Poisson*** distribution. We will now look into the conditions that characterize these processes, and examine the standard distributions associated with the processes. This will enable us to identify situations for which these distributions apply.

Let us first study the *Bernoulli random variable,* named so in honor of the mathematician Jakob Bernoulli (1654-1705). It is the building block for other random variables and the resulting distributions we will study in this lesson.

5.2.2 Bernoulli Random Variable

Suppose an operator uses a lathe to produce pins, and the lathe is not perfect in the sense that it does not always produce a good pin. Rather, it has a probability *p* of producing a good pin and (*1 - p*) of producing a defective one. Let us denote a good pin as *"success"* and a defective pin as *"failure".*

Just after the operator produces one pin, it is inspected; let *X* denote the *"number of good pins produced" i. e. "the number of successes".*

Now analyzing the trial- ***"inspecting a pin"*** and our random variable ***X-"number of successes"***, we note two important points:

- The trial-*"inspecting a pin"* has only two possible outcomes, which are mutually exclusive. Such a trial, whose outcome can only be either a success or a failure, is a ***Bernoulli trial.*** In other words, the sample space of a Bernoulli trial is *S* = {success, failure}.
- The random variable, *X*, that measures number of successes in one Bernoulli trial, is a ***Bernoulli random variable.*** Clearly, *X is* 1 if the pin is good and 0 if it is defective.

It is easy to derive the probability distribution of Bernoulli random variable

X	*:*	0	1
P(X)	*:*	*p*	*1-p*

If *X* is a Bernoulli random variable, we may write

$X \sim BER(p)$

Where ~ is read as "is distributed as" and BER stands for Bernoulli.

A Bernoulli random variable is too simple to be of immediate practical use. But it forms the building block of the ***Binomial random variable***, which is quite useful in practice. The binomial random variable in turn is the basis for many other useful cases, such as ***Poisson random variable.***

5.3 THE BINOMIAL DISTRIBUTION

In the real world we often make several trials, not just one, to achieve one or more successes. Let us consider such cases of several trials.

Consider *n* number of *identically and independently distributed* Bernoulli random variables X_1, X_2 ……, X_n. Here, identically means that they all have the same *p,* and independently means that the value of one *X does* not in any way affect the value of another. For example, the value of X_2 does not affect the value of X_3 or X_8 and so on. Such a sequence of identically and independently distributed Bernoulli variables is called a ***Bernoulli Process.***

Suppose an operator produces *n* pins, one by one, on a lathe that has probability *p* of making a good pin at each trial, the sequence of numbers (1 or 0) denoting the good and defective pins produced in each of

the *n* trials is a Bernoulli process. For example, in the sequence of nine trials denoted by 001011001, the third, fifth, sixth and ninth are good pins, or successes. The rest are failures.

In practice, we are usually interested in the total number of good pins rather than the sequence of *1's* and *0's.* In the example above, four out of nine are good. In the general case, let *X* denote the total number of good pins produced in *n* trials. We then have $X = X_1 + X_2 + \ldots\ldots + X_n$ where all $X_i \sim BER(p)$ and are independent.

The random variable that counts the number of successes in many independent, identical Bernoulli trials is called a Binomial Random Variable.

5.3.1 Conditions for a Binomial Random Variable

We may appreciate that the condition to be satisfied for a binomial random variable is that ***the experiment should be a Bernoulli Process.***

Any uncertain situation or experiment that is marked by the following three properties is known as a Bernoulli Process:

- There are only two mutually exclusive and collectively exhaustive outcomes in the experiment *i.e.* S = {success, failure}
- In repeated trials of the experiment, the probabilities of occurrence of these events remain constant
- The outcomes of the trials are independent of one another

The probability distribution of Binomial Random Variable is called the ***Binomial Distribution***

5.3.2 BINOMIAL PROBABILITY FUNCTION

Now we will develop the distribution of our Binomial random variable. To describe the distribution of Binomial random variable we need two parameters, *n* and *p* we write $X \sim B(n, p)$ to indicate that *X* is Binomially distributed with *n* number of independent trials and *p* probability of success is each trial. The letter *B* stands for binomial.

Let us analyze the probability that the number of successes *X* in the *n* trials is exactly *x* (obviously number of failures are *n-x*) *i.e.* $X = x$ and $x = 0,1,2,\ldots\ldots\ldots\ n$; as *n* trials are made, at the best all *n* can be successes.

Now we know that there are ${}^nC_x$ ways of getting *x* successes out of *n* trials. We also observe that each of these ${}^nC_x$ possibilities has $p^x(1-p)^{n-x}$ probability of occurrence corresponding to *x* successes and (*n-x*) failures. Therefore,

$$P(X = x) = {}^{n}C_x \, p^x \, (1-p)^{n-x} \qquad \text{for } x = 0,1,2,\ldots\ldots, n$$

This equation is the Binomial probability formula. If we denote the probability of failure as q then the Binomial probability formula is

$$P(X = x) = {}^{n}C_x \, p^x \, q^{n-x} \qquad \text{for } x = 0,1,2,\ldots\ldots, n$$

We may write down the Binomial probability distribution in table format (see Table 5-2)

Table 5-2

Binomial Distribution of *X*

$X = x$	$P(X = x)$
0	${}^{n}C_0 \, p^0 \, q^n$
1	${}^{n}C_1 \, p^1 \, q^{n-1}$
…	…
x	${}^{n}C_x \, p^x \, q^{n-x}$
…	…
…	…
n	${}^{n}C_n \, p^n \, q^0$

Each of the term for $x = 0,1,2,\ldots\ldots, n$ correspond to the Binomial expansion of $(p + q)^n$

5.3.3 Characteristics of a Binomial Distribution

1. Expected Value or Mean

The expected value or the mean, denoted by μ, of a Binomial distribution is computed as $E(X) = \mu = \sum_{x=0}^{n} x.P(x)$. An evaluation of μ will show that $\mu = n\,p$.

2. Variance

The variance, denoted by σ^2, of a Binomial distribution is computed as

$$V(X) = \sigma^2 = E[(X - \mu)^2]$$

$$= \sum_{x=0}^{n} (x - \mu)^2 . P(x)$$

An evaluation of σ^2 will show that $\sigma^2 = n\,p\,q$

3. Moments about the Origin

The rth moment about the origin denoted by m_r^0, of a Binomial distribution is computed as: $m_r^0 = \sum_{x=0}^{n} x^r.P(x)$. For example:

(a) First moment about the origin will be $m_1^0 = \sum_{x=0}^{n} x.P(x) = np = \mu.$

(b) Second moment about the origin will be $m_2^0 = \sum_{x=0}^{n} x^2.P(x) = n(n-1)p^2 + np.$

4. Moments about the Mean

The rth moment about the mean denoted by m_r^μ, of a binomial distribution is computed as: $m_r^\mu = \sum_{x=0}^{n} (x-\mu)^r.P(x)$. For example,

(a) First moment about the mean will be $m_1^\mu = \sum_{x=0}^{n} (x-\mu)^1.P(x) = 0$

(b) Second moment about the mean will be $m_2^\mu = \sum_{x=0}^{n} (x-\mu)^2.P(x) = npq = \sigma^2$

(c) Third moment about the mean will be $m_3^\mu = \sum_{x=0}^{n} (x-\mu)^3.P(x) = npq(q-p)$

(d) Fourth moment about the mean will be $m_4^\mu = \sum_{x=0}^{n} (x-\mu)^4.P(x) = 3(npq)^2 + npq(1-6pq)$

5. Skewness

To bring out the skewness of a Binomial distribution we can calculate, moment coefficient of skewness, γ_1

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{(m_3^\mu)^2}{(m_2^\mu)^3}} = \frac{m_3^\mu}{\left(\sqrt{m_2^\mu}\right)^3} = \frac{npq(q-p)}{\left(\sqrt{npq}\right)^3} = \frac{q-p}{\sqrt{npq}}$$

Evaluating $\gamma_1 = \frac{q-p}{\sqrt{npq}}$ we note:

- the Binomial distribution is skewed to the right *i.e.* has positive skewness when $\gamma_1 > 0$, which is so when $p < q$
- the Binomial distribution is skewed to the left *i.e.* has negative skewness when $\gamma_1 < 0$, which is so when $p > q$
- the Binomial distribution is symmetrical *i.e.* has no skewness when $\gamma_1 = 0$, which is so when $p = q$. Thus, n being the same, the degree of skewness in a Binomial distribution tends to vanish as p approaches ½ *i.e.* as $p \to$ ½
- for a given value of p, as n increases the Binomial distribution moves to the right, flattens and spreads out As $n \to \infty$, $\gamma_1 \to 0$, the distribution tends to be symmetrical.

6. Kurtosis

A measure of kurtosis of the Binomial distribution is given by the moment coefficient of kurtosis γ_2

$$\gamma_2 = \beta_2 - 3 = \frac{m_4^{\mu}}{\left(m_2^{\mu}\right)^2} - 3 = \frac{3n^2p^2q^2 + npq(1-6pq)}{n^2p^2q^2} - 3 = \frac{1-6pq}{npq}$$

Evaluating $\gamma_2 = \frac{1-6pq}{npq}$ we note

- the Binomial distribution is leptokurtic when $\gamma_2 > 0$, which is so when $6pq < 1$.
- the Binomial distribution is platykurtic when $\gamma 2 < 0$, which is so when 6pq >1.
- the Binomial distribution is mesokurtic when $\gamma 2 = 0$, which is so when 6pq =1.

7. Normal approximation of the Binomial distribution

If n is large and if neither of p or q is too close to zero, the Binomial distribution can be closely approximated by a Normal distribution with standardized variable $Z = \frac{X - np}{\sqrt{npq}}$.

8. Poisson approximation of the Binomial distribution

Binomial distribution can reasonably be approximated by the Poisson distribution when n is infinitely large and p is infinitely small *i. e.* when $n \to \infty$ *and* $p \to 0$.

5.3.4 Importance of Binomial Distribution

The binomial probability distribution is a discrete probability distribution that is useful in describing an enormous variety of real life events. For example, a quality control inspector wants to know the probability of defective light bulbs in a random sample of 10 bulbs if 10% of the bulbs are defective. He can quickly obtain the answer from tables of the binomial distribution. The binomial distribution can be used when:

- *The outcome or results of each trial in the process are featured as one of two types of possible outcomes. In other words, they are attributes.*
- *The possibility of outcome of any trial does not change and is the independent of the results of previous trials.*

5.3.5 Fitting a Binomial Distribution

When a binomial distribution is to be fitted to observe data, the following procedure is adopted:

1. *Determine the value of p and q. if one of these values is known the other can be found out by the simple relationship p = (1 – q), and q = (1 – p). When p and q are equal, the distribution is symmetrical. For p and q may be interchanged without alternating the value of any terms and consequently terms equidistant from the two ends of the series are equal. If p and q are unequal, the distribution is skew. If p is less than ½ , the distribution is positively skewed and when p is more than ½ the distribution is negatively skewed.*
2. *Expand the binomial $(q + p)^n$. The power n is equal to one less than the number of terms in the expected binomial. Thus when two coins are tosses (n = 2) there will be three terms in the binomial. Similarly, when four coins are tossed (n = 4) there will be five terms and so on.*
3. *Multiply each term of the expended binomial by N (the total frequency) in order to obtain the expected frequency in each category.*

The following examples illustrate the procedure:

Example 5-1

Assuming the probability of male birth as ½, find the probability distribution of number of boys out of 5 births.

(a) Find the probability that a family of 5 children have

(i) at least one boy

(ii) at most 3 boys

(b) Out of 760 families with 5 children each find the expected number of families with (i) and (ii) above

Solution: Let the random variable X measures the number of boys out of 5 births. Clearly X is a binomial random variable. So we apply the Binomial probability function to calculate the required probabilities.

$X \sim B(5, ½)$

$P(X = x) = {}^nC_x\, p^x\, q^{n-x}$ *for* $x = 0, 1, 2, 3, 4, 5$

The probability distribution of X is given below

X = x	***:***	0	1	2	3	4	5
P(X = x)	***:***	1/32	5/32	10/32	10/32	*5/32*	1/32

(a) The required probabilities are

(i) $P(X \geq 1) = 1 - P(X = 0)$

$= 1 - 1/32$

$= 31/32$

(ii) $P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$= 1/32 + 5/32 + 10/32 + 10/32$

$= 26/32$

(b) Out of 760 families with 5 children, the expected number of families with

(i) at least one boy $= 760 * P(X \geq 1)$

$= 760 * 31/32$

$= 730$

(ii) at most 3 boys $= 760 * P(X \leq 3)$

$= 760 * 26/32$

$= 720$

Example 5.2 Eight coins are tossed at a time 256 times. Number of heads observed at each throw is recorded and the result are given below. Find the expected frequencies. What are the theoretical values

of mean and standard deviation? Calculate also the mean and standard deviation of the observed frequencies.

Number of heads at a throw	Frequency	Number of heads at a throw	Frequency
0	2	5	56
1	6	6	32
2	30	7	10
3	52	8	1
4	67		

Solution: The chance of getting a head in a single throw of one coin is ½ , hence p = ½ , q = ½ , n = 8, N = 256.

By expanding 256(½ + ½)^8 we shall get the expected frequencies of 1, 2, ……, 8 heads (successes).

Number of Heads (X)	Frequency = N × ^nC_r q^{n-r} p^r
0	256 (½)^8 = 1
1	256 × $^8C_1$ (½)^7 (½)^1 = 8
2	256 × $^8C_2$ (½)^6 (½)^2 = 28
3	256 × $^8C_3$ (½)^5 (½)^3 = 56
4	256 × $^8C_4$ (½)^4 (½)^4 = 70
5	256 × $^8C_5$ (½)^3 (½)^5 = 56
6	256 × $^8C_6$ (½)^2 (½)^6 = 28
7	256 × $^8C_7$ (½)^1 (½)^7 = 8
8	256 × $^8C_8$ (½)^0 (½)^8 = 1
n = 8	Total (N) = 256

The mean of the above distribution is np = 8 × ½ = 4. The standard deviation is:

$$Standard\ Deviation = \sqrt{npq} = \sqrt{\frac{1}{2} \times \frac{1}{2} \times 8} = \sqrt{2} = 1.414$$

These are the mean and standard deviation of the expected frequency distribution. The mean and the standard deviation of the observed frequency distribution shall be:

X	f	d	fd	fd^2
0	2	-4	-8	32
1	6	-3	-18	54
2	30	-2	-60	120
3	52	-1	-52	52
4	67	0	0	0
5	56	+1	+56	56
6	32	+2	+64	128
7	10	+3	+30	90
8	1	+4	+4	16
	N = 256		$\sum fd = 16$	$\sum fd^2 = 548$

$$\bar{X} = A + \frac{\sum fd}{N} = 4 + \frac{16}{256} = 4.0625$$

$$\sigma = \sqrt{\frac{fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{548}{256} - \left(\frac{16}{256}\right)^2} = \sqrt{2.141 - 0.04} = \sqrt{2.137} = 1.462$$

5.4 THE POISSON DISTRIBUTION

Poisson Distribution was developed by a French Mathematician Simeon D Poisson (1781-1840). If a random variable *X* is said to follow a ***Poisson Distribution***, then its probability distribution is given by

$$\boldsymbol{P(X = x) = \frac{e^{-\mu}\mu^x}{x!} \qquad x = 0,1,2,\ldots\ldots}$$

Where, *x* is the number of successes

μ is the mean of the Poisson distribution and

e = 2.71828 (the base of natural logarithms)

The random variable *X* counts the number of successes in **Poisson Process**. A Poisson process corresponds to a Bernoulli process under the following conditions:

- the number of trials *n*, is infinitely large *i.e.* $n \to \propto$
- the constant probability of success p, for each trial is infinitely small i.e. $p \to 0$ (obviously $q \to 1$)
- $np = \mu$ is finite

We can develop the Poisson probability rule from the Binomial probability rule under the above conditions.

Let us consider a Bernoulli process with *n* trials and probability of success in any trial

$p = \frac{\mu}{n}$, where $\mu \geq 0$. Then, we know that the probability of *x* successes in *n* trials is given by

$$P(X = x) = {}^{n}C_{x}\left(\frac{\mu}{n}\right)^{x}\left(1-\frac{\mu}{n}\right)^{n-x} = \frac{n!}{x!(n-x)!}\left(\frac{\mu}{n}\right)^{x}\left(1-\frac{\mu}{n}\right)^{n-x}$$

$$= \frac{n[n-1][n-2].....[n-(x-1)]}{x!}\left(\frac{\mu}{n}\right)^{x}\left(1-\frac{\mu}{n}\right)^{n-x}$$

$$= \frac{\mu^{x}}{x!}\left[\frac{n}{n}.\frac{n-1}{n}.\frac{n-2}{n}.....\frac{n-(x-1)}{n}\right]\left(1-\frac{\mu}{n}\right)^{n-x}$$

$$= \frac{\mu^{x}}{x!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right).....\left(1-\frac{x-1}{n}\right)\left(1-\frac{\mu}{n}\right)^{n}\left(1-\frac{\mu}{n}\right)^{-x}$$

Now if $n \to \propto$, then the terms, $\left(1-\frac{1}{n}\right);\left(1-\frac{2}{n}\right);.....;\left(1-\frac{x-1}{n}\right)$ *and* $\left(1-\frac{\mu}{n}\right)^{-x}$ will all be tending to 1 and $\left(1-\frac{\mu}{n}\right)^{n} \to e^{-\mu}$ if $n \to \propto$ Thus we have $\boldsymbol{P(X = x) = \frac{e^{-\mu}\mu^{x}}{x!}}$ ***where, x*** *= 0, 1, 2,……..*

This equation is the probability distribution function of Poisson distribution.

Thus, we have seen that to describe the distribution of Poisson random variable we need only one parameterμ, we write If $X \sim POI(\mu)$, Then $\boldsymbol{P(X = x) = \frac{e^{-\mu}\mu^{x}}{x!}}$ ***x*** *= 0, 1, 2,……..* We may write down the Poisson probability distribution in table format (see Table 5-3)

Table 5-3

Poisson Distribution of *X*

X = x	***P(X = x)***

0	$e^{-\mu}$
1	$\mu e^{-\mu}$ or $\mu P(X = 0)$
2	$\frac{\mu^2}{2!}e^{-\mu}$ or $\frac{\mu}{2} P(X = 1)$
…	…
…	…
x	$\frac{\mu^x}{x!}e^{-\mu}$ or $\frac{\mu}{x} P(X = x-1)$
…	…
…	…

Poisson distribution may be expected in situations where the chance of occurrence of any event is small, and we are interested in the occurrence of the event and not in its non-occurrence. For example, number of road accidents, number of defective items, number of deaths in flood or because of snakebite or because of a rare disease *etc.* In these situations, we know about the occurrence of an event although its probability is very small, but we do not know how many times it does not occur. For instance, we can say that two road accidents took place today, but it is almost impossible to say as to how many times, accident fails to take place. The reason is that the number of trials is very large here and the nature of event is of rare type. The Poisson random variable *X*, counts the number of times a rare event occurs during a fixed interval of time or space.

5.4.1 Characteristics of a Poisson Distribution

1. Expected Value or Mean

The expected value or the mean, denoted by μ, of a Poisson distribution is computed as $E(X) = \mu = \sum_{allx} x.P(x)$ An evaluation of *mean* will show that it is always μ itself.

2. Variance

The variance, denoted by σ^2, of a Poisson distribution is computed as $V(X) = \sigma^2 = E[(X - \mu)^2] = \sum_{all\,x}(x-\mu)^2.P(x)$ An evaluation of σ^2 will show that $\sigma^2 = \mu$

3. Moments about the Origin

The rth moments about the origin denoted by m_r^0, of a Poisson distribution is computed as:

$$m_r^0 = \sum_{all\,x} x^r.P(x)$$

For example, (a) First moment about the origin will be

$$m_1^0 = \sum_{all\,x} x.P(x)$$

$$= \mu$$

(b) Second moment about the origin will be

$$m_2^0 = \sum_{all\,x} x^2.P(x)$$

$$= \mu + \mu^2$$

4. Moments about the Mean

The rth moments about the mean denoted by m_r^μ, of a Poisson distribution is computed as:

$$m_r^\mu = \sum_{all\,x}(x-\mu)^r.P(x)$$

For example, (a) First moment about the mean will be

$$m_1^\mu = \sum_{all\,x}(x-\mu)^1.P(x)$$

$$= 0$$

(b) Second moment about the mean will be

$$m_2^\mu = \sum_{all\,x}(x-\mu)^2.P(x)$$

$$= \sigma^2$$

$$= \mu$$

(c) Third moment about the mean will be

$$m_3^{\mu} = \sum_{all\,x} (x-\mu)^3 . P(x)$$

$$= \mu$$

(d) Fourth moment about the mean will be

$$m_4^{\mu} = \sum_{all\,x} (x-\mu)^4 . P(x)$$

$$= 3\mu^2 + \mu$$

5. Skewness

To bring out the skewness we can calculate, moment coefficient of skewness, γ_1

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{(m_3^{\mu})^2}{(m_2^{\mu})^3}} = \frac{m_3^{\mu}}{\left(\sqrt{m_2^{\mu}}\right)^3} = \frac{1}{\sqrt{\mu}}$$

Evaluating $\gamma_1 = \frac{1}{\sqrt{\mu}}$ we note that Poisson distribution is always skewed to the right *i.e.* has positive skewness which is so as it is a distribution of rare events.

The degree of skewness in a Poisson distribution decreases as the value of μ increases.

6. Kurtosis

A measure of kurtosis of the Poisson distribution is given by the moment coefficient of kurtosis γ_2

$$\gamma_2 = \beta_2 - 3 = \frac{m_4^{\mu}}{\left(m_2^{\mu}\right)^2} - 3 = \frac{1}{\sqrt{\mu}}$$

Evaluating $\gamma_2 = \frac{1}{\sqrt{\mu}}$ we note that the Poisson distribution is leptokurtic.

7. Poisson approximation of the Binomial distribution

Poisson distribution can reasonably approximate Binomial distribution when *n* is infinitely large and *p* is infinitely small *i. e.* when

$$n \to \propto \; and \; p \to 0$$

5.4.2 Role of Poisson Distribution

The Poisson distribution is used in practice in a wide variety of problems where there are infrequently occurring events with respect to time, area, volume or similar units. Some practical situations in which Poisson distribution can be used are given below:

1. It is used in quantity control statistics to count the number of defects of an item.
2. It is used in biology and physics to count the number of bacteria and to count the number of particles emitted from a radioactive substances.
3. It is used in insurance problems to count the number of causalities.
4. It is used in call center or telephonic companies in waiting-time problems to count the number of incoming telephones calls or incoming customers.
5. It is used in count the number of traffic arrivals such as trucks at terminals, airplanes at airports, ships at docks and so forth.
6. It is used in determining the number of deaths in a district in a given period say a year, by a rare diseases.
7. It is used in count the number of typographical errors per page in typed material, number of deaths as a result of road accident etc.
8. It is used in dealing with the inspection of manufactured products with the probability that any one piece is defective is very small and the lots are very large.

In general, the Poisson distribution explains the behaviour of those discrete variants where the probability of occurrence of the event is small and the total number of possible cases is sufficiently large.

5.4.3 Fitting of a Poisson Distribution

The process of fitting a Poisson distribution is very simple. We have just to obtain the value of m, i.e., the average occurrence, and calculate the frequency of 0 successes. The other frequencies can be very easily calculated as follows:

$N(P_0) = Ne^{-m}$

$N(P_1) = N(P_0) \times m/1$

$N(P_2) = N(P_1) \times m/2$

$N(P_3) = N(P_2) \times m/3$

$N(P_4) = N(P_3) \times m/4$

Example 5.3 *At a parking place the average number of car-arrivals during a specified period of 15 minutes is 2. If the arrival process is well described by a Poisson process, find the probability that during a given period of 15 minutes*

(a) no car will arrive

(b) atleast two cars will arrive

(c) atmost three cars will arrive

(d) between 1 and 3 cars will arrive

Solution: Let X denote the number of cars arrivals during the specified period of 15 minutes. So

$X \sim POI(\mu)$

We apply the Poisson probability function $P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$ $x = 0,1,2,\ldots\ldots$ to calculate the required probabilities.

(a) P(no car will arrive) $= P(X = 0) = \frac{e^{-2}2^0}{0!}$

$= 0.1353$

(b) P(atleast two cars will arrive) $= P(X \geq 2)$

$= 1-[P(X = 0) + P(X = 1)]$

$= 1-[\frac{e^{-2}2^0}{0!} + \frac{e^{-2}2^1}{1!}]$

$= 1-[0.1353 + 0.2707]$

$= 1 - 0.4060$

$= 0.5740$

(c) P(atmost three cars will arrive) $= P(X \leq 3)$

$$= \sum_{x=0}^{3} \frac{e^{-2} 2^{x}}{x!}$$

$$= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$= 0.8571$$

(d) P(between 1 and 3 cars will arrive) = $P(1 \leq X \leq 3)$

$$= P(X \leq 3) - P(X = 0)$$

$$= \sum_{x=0}^{3} \frac{e^{-2} 2^{x}}{x!} - \frac{e^{-2} 2^{0}}{0!}$$

$$= 0.8571 - 0.1353$$

$$= 0.7218$$

5.5 CHECK YOUR PROGRESS

1. The Binomial distribution is skewed to the right *i.e.* has …… when $\gamma_1 > 0$, which is so when $p < q$.
2. The degree of skewness in a Poisson distribution ……. as the value of μ increases.
3. Poisson distribution can reasonably approximate …… when *n* is infinitely large and *p* is infinitely small.
4. If *n* is large and if neither of *p* or *q* is too close to zero, the Binomial distribution can be closely approximated by a …… with standardized variable.
5. The random variable that counts the number of successes in many independent, …… is called a Binomial Random Variable.

5.6 SUMMARY

A random variable has a probability law - a rule that assigns probabilities to different values of the random variable. This probability law - the probability assignment is called the probability distribution of the random variable. The probability distribution of a discrete random variable lists the probabilities of occurrence of different values of the random variable. The proper identification of experiments with certain known processes in Probability theory can help us in writing down the probability distribution function. Two such processes are the Bernoulli Process and the Poisson Process. The standard discrete probability distributions that are consequent to these processes are the Binomial and the Poisson distribution. There are different characteristics of both.

5.7 KEYWORDS

Discrete: It takes only a countable number of values.

Continuous: It take on any value in an interval of numbers (*i.e.* it's possible values are unaccountably infinite).

Binomial Random Variable: The random variable that counts the number of successes in many independent, identical Bernoulli trials is called a Binomial Random Variable.

Poisson distribution: It may be expected in situations where the chance of occurrence of any event is small, and we are interested in the occurrence of the event and not in its non-occurrence.

5.8 SELF-ASSESSMENT TEST

1. Explain what you understand by random experiment and a random variable. Briefly explain the following:
   a. Discrete and continuous random variables
   b. Discrete probability distribution.
2. "Binomial random variable measures the number of successes in a Bernoulli Process". Explain this statement. Also develop and generalize Binomial probability rule with the help of an example.
3. State the important properties of a Binomial distribution. Give examples of some of the important area where Binomial distribution is used.
4. Under what condition can the Poisson distribution approximate Binomial distribution? Develop the Poisson probability rule from the Binomial probability rule under these conditions.
5. List some of the important areas where Poisson distribution is used. Also state the important properties of a Poisson distribution.
6. On an average a machine produces 20 % defective item find the probability that a random sample of 4 items consists of
   (a) none to four defective items (b) atleast 3 defective items
   (c) almost 2 defective items.
   Out of 200 samples of 4 items, find the expected number of samples with (a), (b), and (c) above
7. A gardener knows from his personal experiences that 2% of seedlings fail to service on transplantation. Find the mean, standard deviation and moment coefficient of skewness of the distribution of rate of failure to service in a sample of 400 seedlings.

8. If the sum of mean and variance of a binomial distribution of 5 trials is 7/5, find the binomial distribution.
9. The mean and variance of a binomial distribution are 2 and 1.5 respectively. Find the probability of
   (a) 2 successes (b) atleast 2 successes (c) at most 2 successes.
10. 150 random samples of 4 units each are inspected for number of defective item. The results are:

Number of defective items	:	0	1	2	3	4
Number of Samples	:	28	62	46	10	4

    Fit a binomial distribution to the observed data.
11. The probability that a particular injection will have reaction to an individual is 0.002. Find the probability that out of 1000 individuals (a) no, (b) 1, (c) at least 1, and (d) almost 2; individuals will have reaction from the injection.
12. In a razor blades manufacturing factory, there is small chance of 1/500 for any blade to be defective. The blades are supplied in packets of 10. Find the approximate number of packets containing (a) no, (b) 1, and (c) 2 defective blades in a consignment of 10,000 packets.
13. If $P(x = 1) = P(x = 2)$, for a distribution of Poisson random variable X. Find the mean of the distribution.
14. The distribution of typing mistakes committed by a typist is given below:

Number of mistakes *(X)*	:	0	1	2	3	4	5
Number of pages *(f)*	:	142	156	67	27	5	1

    Fit a Poisson distribution and find the expected frequencies.

5.9 ANSWERS TO CHECK YOUR PROGRESS

1. Positive skewness
2. Decreases
3. Binomial distribution
4. Normal distribution
5. Identical Bernoulli trials

5.10 REFERNCES/SUGGESTED READINGS

9. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.

10. Statistics for Management *by* G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
11. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
12. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
13. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
14. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
15. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
16. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.

Subject: Business Statistics	
Course Code: MBA-102	Author: Anil Kumar
Lesson: 06	Vetter: Dr. Karam Pal
PROBABILITY DISTRIBUTIONS-II	

STRUCTURE

6.1 Learning Objectives

6.2 Introduction

6.2.1 Continuous Probability Distribution

6.2.2 The Normal Distribution

6.3 The Standard Normal Distribution

6.3.1 The Standard Area Table

6.3.2 Finding Probabilities of Standard Normal Distribution

6.3.3 Finding the Value of Z given a Probability

6.4 The Transformation of Normal Random Variables

6.5 Check your Progress

6.6 Summary

6.7 Keywords

6.8 Self-Assessment Test

6.9 Answers to check your progress

6.10 References/Suggested Readings

6.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understand the concept of continuous random variable and Normal distribution
- Appreciate the usefulness of normal distribution in decision-making
- Identify situations where normal probability distribution can be applied.

6.2 INTRODUCTION

We have learnt that a probability distribution is basically a convenient representation of the different values a random variable may take, together with their respective probabilities of occurrence. In the last lesson, we have examined situations involving discrete random variables and the resulting discrete probability distributions. Consider the following random variables that we have taken up in the last lesson:

1. Number of Successes (X_1) in a Bernoulli's Process
2. Number of Successes (X_2) in a Poisson Process

In the first case, Binomial random variable X_1 could take only finite number of integer values; *0,1,2…n;* whereas in the second case, Poisson random variable X_2 could take an infinite number of integer value; *0,1,2,3……* The random variables X_1 and X_2 are discrete, in the sense that they could be listed in a sequence, finite or infinite. In contrast to these, let us consider a situation, where the variable of interest may take any value within a given range. Suppose we are planning for measuring the variability of an automatic bottling process that fills *½-liter* (500 cm^3) bottles with cola. The variable, say *X*, indicating the deviation of the actual volume from the normal (average) volume can take any real value - positive or negative; integer or decimal. This type of random variable, which can take an infinite number of values in a given range, is called a ***continuous random variable***, and the probability distribution of such a variable is called a ***continuous probability distribution.*** The concepts and assumption inherent in the treatment of such distributions are quite different from those used in the context of a discrete distribution. In the present lesson, after understanding the basic concepts of continuous distributions, we will discuss Normal distribution - an important continuous distribution that is applicable to many real-life processes.

6.2.1 Continuous Probability Distribution

Consider our planning for measuring the variability of the automatic bottling process that fills *½-liter* (500cm^3) bottles with cola. The random variable *X* indicates 'the deviation of the actual volume from the normal (average) volume.' Let us, for some time, measure our random variable *X* to the nearest one cm^3.

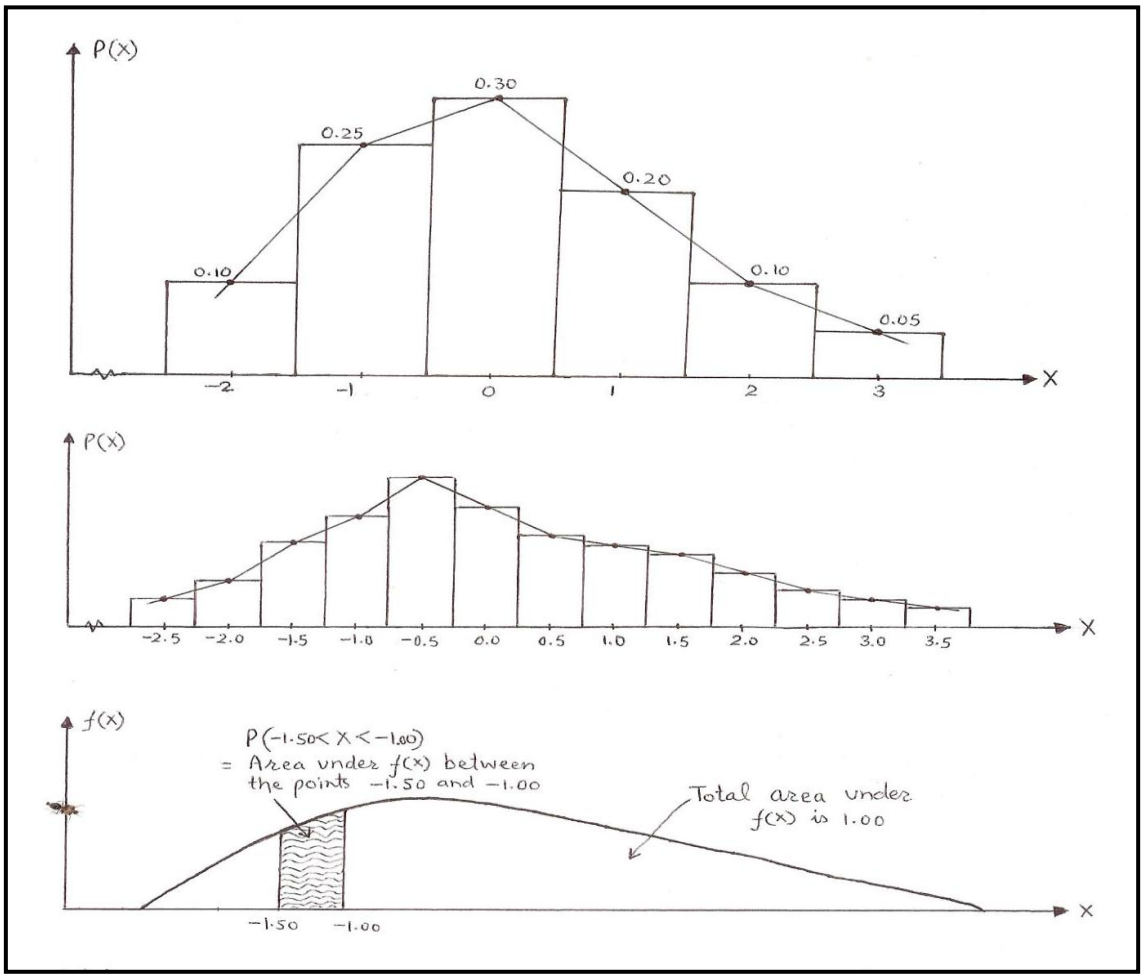


Figure 6-1 Histograms of the Distribution of X *as Measurements is refined to Smaller and Smaller Intervals of Volume, and the Limiting Density Function* f(x)

Suppose Figure8-1*a* represent the histogram of the probability distribution of *X*. The probability of each value of *X* is the area of the rectangle over the value. Since the rectangle will have the same base, the height of each rectangle is proportional to the probability. The probabilities also add to 1.00 as required for a probability distribution.

Volume is a continuous random variable; it can take on any value measured on an interval of numbers. Now let us imagine the process of refining the measurement scale of *X* to the nearest *1/2* cm^3, the nearest *1/8* cm^3… and so on. Obviously, as the process of refining the measurement scale continues, the number of rectangles in the histogram increases and the width of each rectangle decreases. The probability of each value is still measured by the area of the rectangle above it, and the total area of all rectangles remains 1.00. As we keep refining our measurement scale, the discrete distribution of *X* tends

to a continuous probability distribution. The step like surface formed by the tops of the rectangles in the histogram tends to a smooth function. This function is denoted by *f(x)* and is called the ***probability density function*** of the continuous random variable *X*. The density function is the limit of the histograms as the number of rectangles approaches infinity and the width of each rectangle approaches zero. The density function of the limiting continuous variable *X* is shown in Figure 6-1 *i.e.* the values *X* can assume between the intervals –2.00 to –3.00 approaches infinity. The probability that *X* assumes a particular value (Say *X* = 1.5) approaches zero. Probabilities are still measured as areas under the curve. The probability that deviation will be between –1.50 and –1.00 is the area under *f(x)* between the points *x* = -1.50 and *x* = -1.00. Let us now make some formal definitions.

A continuous random variable is a random variable that can take on any value in an interval of numbers.

The probabilities associated with a continuous random variable *X* are determined by the ***probability density function*** of the random variable. The function, denoted by *f(x),* has the following properties:

1. *f(x)* = 0 for all *x*
2. The probability that *X* will be between two numbers *a* and *b* is equal to the area under *f(x)* between *a* and *b.*

$$P(a < X < b) = \int_a^b f(x).dx$$

3. The total area under the entire curve of *f(x)* is equal to 1.00.

$$P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x).dx = 1.00$$

When the sample space is continuous, the probability of any single given value is zero. For a continuous random variable, therefore, the probability of occurrence of any given value is zero. We see this from property 2, noting that the area under a curve between a point and itself is the area of a line, which is zero. ***For a continuous random variable, non-zero probabilities are associated only with intervals of numbers.***

We define the cumulative distribution function *F(x)* for a continuous random variable similarly to the way we defined it for a discrete random variable: *F(x) is* the probability that *X is* less than (or equal to) *x.*

Thus, the ***cumulative distribution function*** of a continuous random variable:

F(x) = P(X = x) = area under *f(x)* between the *smallest* possible value of *X* (often $-\infty$) and point *x*

$$= \int_{-\infty}^{x} f(x).dx$$

The cumulative distribution function *F(x) is* a smooth, non-decreasing function that increases from 0 to 1.00.

The expected value of a continuous random variable *X*, denoted by *E(X)*, and its variance, denoted by *V(X)*, require the use of calculus for their computation. Thus

$$E(X) = \int_{-\infty}^{\infty} x.f(x).dx$$

$$V(X) = \int_{-\infty}^{\infty} [x - E(x)]^2.f(x).dx$$

6.2.2 The Normal Distribution

The Normal Distribution is the most versatile of all the continuous probability distributions. It is being widely used in all data-based research in the field of agriculture, trade, business and industry. It is found to be useful in characterizing uncertainties in many real-life processes, in statistical inferences, and in approximating other probability distributions.

A large number of random variables occurring in practice can be approximated to the normal distribution.

A random variable that is affected by many independent causes, and the effect of each cause is not overwhelmingly large compared to other effects, closely follow a normal distribution.

The lengths of pins made by an automatic machine; the times taken by an assembly worker to complete the assigned task repeatedly; the weights of baseballs; the tensile strengths of a batch of bolts; and the volumes of cola in a particular brand of canned cola - are good examples of normally distributed random variables. All of these are affected by several independent causes where the effect of each cause is small. This knowledge helps us in calculating the probabilities of different events in varied situations, which in turn is useful for decision-making.

In many real life situations, we face the problem of making statistical inferences about processes based on limited data. Limited data is basically a sample from the full body of data on the process. Irrespective of how the full body of data is distributed, it has been found that the Normal Distribution can be used to characterize the sampling distribution of many of the sample statistics. (we will see it in next few lessons). This helps considerably in Statistical Inferences.

Finally, the Normal Distribution can be used to approximate certain probability distributions. This helps considerably in simplifying the probability calculations.

Probability Density Function

If X is normally distributed with mean μ and variance σ^2, we write $X \sim N(\mu, \sigma^2)$ and the probability density function $f(x)$ is given by the formula:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < +\infty$$

In the equation e is the base of natural logarithm, equal to 2.71828.... By substituting desired values for μ and σ, we can get any desired density function. For example, a distribution with mean 80 and standard deviation 5 will have the density function.

$$f(x) = \frac{1}{\sqrt{2\pi}5} e^{-\frac{1}{2}\left(\frac{x-100}{5}\right)^2} \qquad -\infty < x < +\infty$$

This function when plotted (see Figure 6-2) will give the famous bell-shaped mesokurtic normal curve.

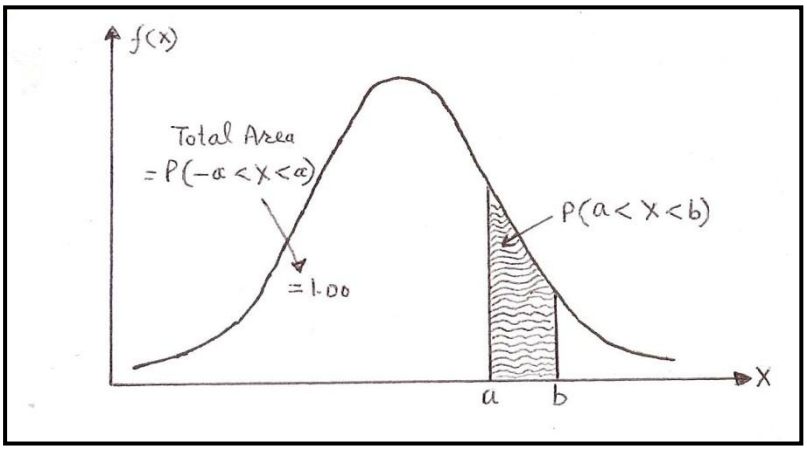


Figure 6-2 A Normal Distribution with μ = 80 and σ = 5

Many mathematicians have worked on the mathematics behind the normal distribution and have made many independent discoveries. In the initial stages, the normal distribution was developed by Abraham De Moivre (1667-1754). His work was later taken up by Pierre S Laplace (1949-1827). But the discovery of equation for the normal density function is attributed to Carl Friedrich Gauss (1777-1855), who did much work with the formula. In science books, this distribution is often called the Gaussian distribution.

We will now examine the properties of the Normal distribution.

Properties of Normal Distribution

1. The normal curve is not a single curve representing only one continuous distribution. Obviously, it represents a family of normal curves; since for each different value of μ and σ, there is a specific normal curve different in its positioning on the *X-axis* and the extent of spread around the mean. Figure 8-3 shows three different normal distributions – with different shapes and positions.

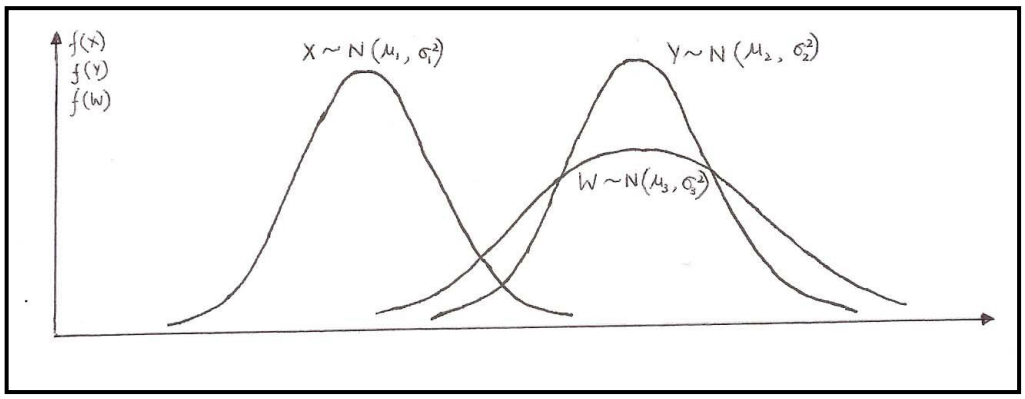


Figure 8-3 Three Different Normal Distribution

2. The normal curve is bell-shaped and perfectly symmetric about its mean. As a result 50% of the area lies to the right of mean and balance 50% to the left of mean. Perfect symmetry, obviously, implies that mean, median and mode coincide in case of a normal distribution. The normal curve gradually tapers off in height as it moves in either direction away from the mean, and gets closer to the *X-axis.*
3. The normal curve has a (relative) kurtosis of 0, which means it has average peakedness and is mesokurtic.

4. Theoretically, the normal curve never touches the horizontal axis and extends to infinity on both sides. That is the curve is asymptotic to *X-axis.*
5. If several independent random variables are normally distributed, then their sum will also be normally distributed. The mean of the sum will be the sum of all the individual means, and by virtue of the independence, the variance of the sum will be the sum of all the individual variances.

   If $X_1, X_2, \ldots\ldots\ldots X_n$ are independent normal variables, the their sum S will also be a normal variable with

   $$E(S) = E(X_1) + E(X_2) + \ldots\ldots\ldots E(X_n)$$

   and $$V(S) = V(X_1) + V(X_2) + \ldots\ldots\ldots V(X_n)$$
6. If a normal variable X under goes a linear change in scale such as $Y = aX + b$, where a and b are constants and $a \neq 0$; the resultant Y variable will also be normally distributed with mean = $a\,E(X) + b$ and Variance = $a^2\,V(X)$

We can combine the above two properties.

If $X_1, X_2, \ldots\ldots\ldots X_n$ are independent random variables that are normally distributed, then the random variable Q defined as

$Q = a_1X_1 + a_2X_2 + \ldots a_nX_n + b$ will also be normally distributed with

$$E(Q) = a_1E(X_1) + a_2E(X_2) + \ldots\ldots\ldots a_nE(X_n) + b$$

and $$V(Q) = a_1^2\,V(X_1) + a_2^2\,V(X_2) + \ldots\ldots\ldots a_n^2\,V(X_n)$$

Let us see the application of this result with the help of an example.

Example 6-1

A cost accountant needs to forecast the unit cost of a product for the next year. He notes that each unit of the product requires 8 labor hours and 5 kg of raw material. In addition, each unit of the product is assigned an overhead cost of Rs 200. He estimates that the cost of a labor hour next year will be normally distributed with an expected value of Rs 45 and a standard deviation of Rs 2; the cost of raw material will be normally distributed with an expected value of Rs 60 and a standard deviation of Rs 3. Find the distribution of the unit cost of the product. Find its expected value and variance.

Solution: Since the cost of labor L may not influence the cost of raw material M, we can assume that the two are independent. This makes the unit cost of the product Q a random variable. So if

$L \sim N(45, 2^2)$ and $M \sim N(60, 3^2)$

Then, $Q = 8L + 5M + 200$ will follow normal distribution with

Mean = $E(Q) = 8E(L) + 5E(M) + 200$

$= 8(45) + 5(60) + 200$

$= 950$

Variance = $V(Q) = 8^2V(L) + 5^2V(M)$

$= 80(4) + 25(9)$

$= 625$

So $Q \sim N(950, 25^2)$

7. Same important area relationships under normal curse are

   Area between $\mu - 1\sigma$ and $\mu + 1\sigma$ is about 0.6826

   Area between $\mu - 2\sigma$ and $\mu + 2\sigma$ is about 0.9544

   Area between $\mu - 3\sigma$ and $\mu + 3\sigma$ is about 0.9974

   Area between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95

   Area between $\mu - 2.58\sigma$ and $\mu + 2.58\sigma$ is 0.99

Importance of Normal Distribution

The normal distribution has long occupied a central place in the theory of statistics. Its importance will be clear from the following points:

1. It has the remarkable property stated in the Central Limit Theorem. According to this theorem as the sample size n increases, the distribution of mean, $\bar{X}$ of a random sample taken from any population approaches to a normal distribution (with mean μ and standard deviation $\sigma/\sqrt{n}$). Thus, if samples of large size, n are drawn from a population that is not normally distributed, nevertheless the successive sample means will form themselves a distribution that is approximately normal. The Central Limit Theorem also applies to the distribution of Median and standard deviation but not range.
2. As n becomes large the normal distribution serves as a good approximation of many discrete distributions such as Binomial or Poisson model whenever the exact discrete probability is difficult to obtain or impossible to calculate accurately.
3. In theoretical statistics many problems can be solved only under the assumptions of a normal population.

4. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate.
5. The normal distribution is used extensively in statistical quality control in on industry in setting up of control limits.

6.3 THE STANDARD NORMAL DISTRIBUTION

There are infinitely many possible normal random variables and the resulting normal curves for different values of μ and σ^2. So the range probability P(a < X < b) will be different for different normal curves. We can make use of integral calculus to compute the required range probability

$$P(a < X < b) = \int_a^b f(x).dx$$

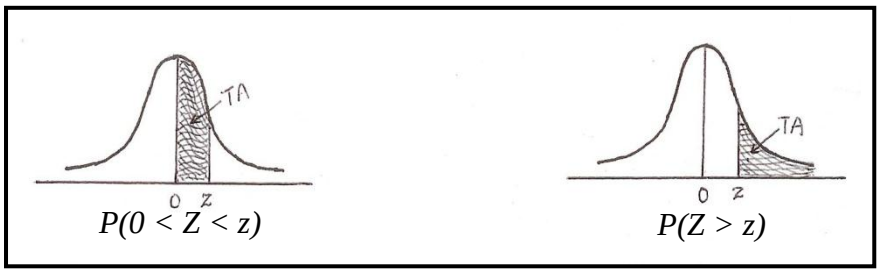
It may be appreciated that we can simplify this process of computing range probabilities to a great extent by tabulating the range probabilities. Since it is not practicable and indeed impossible to have separate probability tables for each of the infinitely many possible normal curves, we select one normal curve to serve as a ***standard***. Probabilities associated with the range of values of this standard normal random variable are tabulated. A special transformation then allows us to apply the tabulated probabilities to *any* normal random variable. The standard normal random variable is denoted by a special name, *Z* (rather than the general name *X* we use for other random variables).

We define the standard normal random variable Z as the normal random variable with mean = 0 and standard deviation = 1. We say

$$Z \sim N(0,1^2)$$

6.3.1 Standard Area Tables

The probabilities associated with standard normal distribution are tabulated in two ways – say Type I and Type II tables, as shown in Figure 6-4. Type I Tables give the area between μ = 0 and any other *z* value, as shown by vertical hatched area in Figure 8-4*a*. The hatched area shown in figure is *P (0 < Z < z).*



P(0 < Z < z) *P(Z > z)*

Figure 6-4 Standard Area Tables

Type II Tables give the area towards the tail–end of the standard normal curve beyond the ordinate at any particular *z* value. The hatched area shown in Figure 8-4*b* is *P (Z > z).*

As the normal curve is perfectly symmetrical, the areas given by Type 1 Tables when subtracted from 0.5 will provide the same areas as given by Type II Tables and vice-versa.

i.e *P (0 < Z < z) = 0.5 - P (Z > z).*

6.3.2 Finding Probabilities of the Standard Normal Distribution

We will now illustrate the use of standard normal area tables for calculating the range probabilities. Probability of intervals is areas under the density curve $f(z)$ over the intervals in question.

Example 6-2

Find the probability that the value of the standard normal random variable will be…

(a) between 0 and 1.74 (b) less than -1.47

(c) between 1.3 and 2 (d) between -1 and 2

Solution: (a) *P(Z* is between 0 and 1.74*)*

That is, we want *P(0 < Z <* 1.74*).* In Figure 8-4*a*, substitute 1.74 for the point *z* on the graph. We are looking for the table area in the row labeled 1.7 and the column labeled 0.04. In the table, we find the probability 0.4591.Thus

P (0 < Z < 1.74) = 0.4591

(b) *P(Z* is less than -1.47*)*

That is, we want *P(Z <* -1.47*).* By the symmetry of the normal curve, the area to the left of -1.47 is exactly equal to the area to the right of 1.47. We find

P(Z < -1.47) = *P(Z >1.47)*

= 0.5000 - 0.4292

= 0.0808

(c) *P(Z* is between 1.3 and 2*)*

That is, we want *P(1.3 < Z <* 2*)*. The required probability is the area under the curve between the two points 1.3 and 2. The table gives us the area under the curve between 0 and 1.3, and the area under the curve between 0 and 2. Areas are additive; therefore,

P(1.30 < Z< 2) = TA(for 2.00) - TA(for 1.30)

= *P(0 < Z <* 2*)* - *P(0 < Z <* 1.3*)*

= 0.4772 - 0.4032

= 0.0740

(d) *P(Z* is between -1 and 2*)*

That is, we want *P(-1< Z <* 2*)*. The required probability is the area under the curve between the two points -1 and 2. The table gives us the area under the curve between 0 and 1, and the area under the curve between 0 and 2. Areas are additive; therefore,

P(-1 < Z< 2) = *P(-1 < Z <* 0*)* + *P(0 < Z <* 2*)*

= *P(0 < Z <* 1*)* + 0.4772

= 0.3413 + 0.4772

= 0.8185

In cases, where we need probabilities based on values with greater than second-decimal accuracy, we may use a linear interpolation between two probabilities obtained from the table.

Example 6-3

Find P(0 ≤ Z ≤ 1.645)

Solution: *P(0 ≤ Z ≤* 1.645*)* is found as the midpoint between the two probabilities *P(0 ≤ Z ≤* 1.64*)* and *P(0 ≤ Z ≤* 1.65*)*. So

P(0 ≤ Z ≤ 1.645*)* = ½[*P(0 ≤ Z ≤* 1.64*)* + *P(0 ≤ Z ≤* 1.65*)*]

= ½[0.4495 + 0.4505]

= 0.45

6.3.3 Finding Values of Z Given a Probability

In many situations, instead of finding the probability that a standard normal random variable will be within a given interval; we may be interested in the reverse: finding an interval with a given probability. Consider the following examples.

Example 6-4

Find a value *z* of the standard normal random variable such that the probability that the random variable will have a value between 0 and *z is* 0.40.

Solution: We look inside the table for the value closest to 0.40. The closest value we find to 0.40 is the table area 0.3997. This value corresponds to 1.28 (row 1.2 and column .08).

So for $P(0 < Z < z) = 0.40$; $z = 1.28$

Example 6-5

Find the value of the standard normal random variable that cuts off an area of 0.90 to its left.

Solution: Since the area to the left of the given point *z* is greater than 0.50, *z must be on the right side of 0.* Furthermore, the area to the left of *0* all the way to $-\infty$ is equal to 0.50. Therefore, TA = 0.90 - 0.50 = 0.40. We need to find the point *z* such that TA = 0.40.

We find that for TA = 0.40; $z = 1.28$.

Thus $z = 1.28$ cuts off an area of 0.90 to the left of standard normal curve.

Example 6-6

Find a 0.99 probability interval, symmetric about 0, for the standard normal random variable.

Solution: The required area between the two *z* values that are equidistant from 0 on either side is 0.99. Therefore, the area under the curve between *0* and the positive *z* value is TA = 0.99/2 = 0.495. We now look in our normal probability table for the area closest to 0.495. The area 0.495 lies exactly between the two areas 0.4949 and 0.4951, corresponding to $z = 2.57$ and $z = 2.58$, Therefore, a simple linear interpolation between the two values gives us $z = 2.575$. The answer, therefore, is $z = \pm 2.575$.

So for $P(-z < Z < z) = 0.99$; $z = 2.575$

6.4 THE TRANSFORMATION OF NORMAL RANDOM VARIABLES

The importance of the standard normal distribution derives from the fact that any normal random variable may be transformed to the standard normal random variable. If we want to transform X, where $X \sim N(\mu, \sigma^2)$, into the standard normal random variable $Z \sim N(0, 1^2)$, we can do this as follows:

$$Z = \frac{X - \mu}{\sigma}$$

We move the distribution from its center of μ to a center of 0. This is done by subtracting μ from all the values of X. Thus, we shift the distribution μ units back so that its new center is 0. To make the standard deviation of the distribution equal to 1, we divide the random variable by its standard deviation σ. The area under the curve adjusts so that the total remains the same. All probabilities (areas under the curve) adjust accordingly. Thus, the transformation from X to Z is achieved by first subtracting μ from X and then dividing the result by σ.

Example 6-7

If $X \sim N(50, 8^2)$, find the probability that the value of the random variable X will be greater than 60

Solution:

$$P(X > 60) = P\left(\frac{X - \mu}{\sigma} > \frac{60 - \mu}{10}\right)$$

$$= P\left(Z > \frac{60 - 50}{10}\right)$$

$$= P(Z > 1)$$

$$= P(Z > 0) - P(0 < Z < 1)$$

$$= 0.5000 - 0.3413$$

$$= 0.1587$$

Example 6-8

The weekly wage of 2000 workmen is normally distribution with mean wage of Rs 70 and wage standard deviation of Rs 5. Estimate the number of workers whose weekly wages are

(a) between Rs 70 and Rs 71 (b) between Rs 69 and Rs 73 (c) more than Rs 72 (d) less than Rs 65

Solution: Let X be the weekly wage in Rs, then

$X \sim N(70, 5^2)$

(a) The required probability to be calculated is $P(70 < X < 71)$

So $$P(70 < X < 71) = P\left(\frac{70-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{71-\mu}{\sigma}\right)$$

$$= P\left(\frac{70-70}{5} < Z < \frac{71-70}{5}\right)$$

$$= P(0 < Z < 0.2)$$

$$= 0.0793$$

So the number of workers whose weekly wages are between Rs 70 and Rs 71

$$= 2000 \times 0.0793$$

$$= 159$$

(b) The required probability to be calculated is $P(69 < X < 73)$

So $$P(69 < X < 73) = P\left(\frac{69-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{73-\mu}{\sigma}\right)$$

$$= P\left(\frac{69-70}{5} < Z < \frac{73-70}{5}\right)$$

$$= P(-0.2 < Z < 0.6)$$

$$= P(-0.2 < Z < 0) + P(0 < Z < 0.6)$$

$$= P(0 < Z < 0.2) + P(0 < Z < 0.6)$$

$$= 0.0793 + 0.2257$$

$$= 0.3050$$

So the number of workers whose weekly wages are between Rs 69 and Rs 73

$$= 2000 \times 0.3050$$

$$= 68$$

(c) The required probability to be calculated is $P(X > 72)$

So $$P(X > 72) = P\left(\frac{X-\mu}{\sigma} > \frac{72-\mu}{\sigma}\right)$$

$$= P\left(Z > \frac{72-70}{5}\right)$$

$$= P(Z > 0.4)$$

$$= 0.5 - P(0 < Z < 0.4)$$

$= 0.5 - 0.1554$

$= 0.3446$

So the number of workers whose weekly wages are more than Rs 72

$= 2000 \times 0.3446$

$= 689$

(d) The required probability to be calculated is $P(X < 65)$

So $P(X < 65) = P(\frac{X-\mu}{\sigma} < \frac{65-\mu}{\sigma})$

$= P(Z < \frac{65-70}{5})$

$= P(Z < -1.0)$

$= P(Z > 1.0)$

$= P(Z > 0) - P(0 < Z < 1.0)$

$= 0.5 - 0.3413$

$= 0.1567$

So the number of workers whose weekly wages are less than Rs 65

$= 2000 \times 0.1567$

$= 313$

The Inverse Transformation

The transformation $Z = \frac{X-\mu}{\sigma}$ takes us from a random variable X with mean μ, and standard deviation σ to the standard normal random variable. We also have an opposite, or inverse, transformation, which takes us from the standard normal random variable Z to the random variable X with mean μ and standard deviation σ. The inverse transformation is given as

$$X = \mu + Z\sigma$$

We use the inverse transformation when we want to get from a given probability, the value or values of a normal random variable X.

Example 6-9

The amount of fuel consumed by the engines of a jetliner on a flight between two cities is a normally distributed random variable X with mean μ = 5.7 tons and standard derivation σ = 0.5 tons. Carrying

too much fuel is inefficient as it slows the plans. If, however, too little fuel is loaded on the plane, an emergency landing may be necessary. What should be the amount of fuel to load so that there is 0.99 probability that the plane will arrive at its destination without emergency landing?

Solution: Given that $X \sim N(5.7, 0.5^2)$,

We have to find the value x such that

$$P(X < x) = 0.99$$

or $$P\left(\frac{X-\mu}{\sigma} < z\right) = 0.99$$

or $$P(Z < z) = 0.99$$

$$= 0.5 + 0.49$$

$$= 0.5 + P(0 < Z < z)$$

From the table, value of z is 2.33

So $$x = \mu + z\sigma$$

$$x = 5.7 + 2.33x0.5$$

$$x = 6.865$$

Therefore, the plane should be loaded with 6.865 tons of fuel to give 0.99 probability that the fuel will last throughout the flight.

Example 6-10

Monthly sale of beer at a bar is believed to be approximately normally distributed with mean 2450 units and standard 400 units. To determine the level of orders and stock, the management wants to find two values symmetrically on either side of mean, such that the probability that sales of beer during the month will be between the two values is

(a) 0.95 (b) 0.99

Find the required values.

Solution: Let X be the monthly sale of beer, then

$$X \sim N(2450, 400^2),$$

(a) We have to find the values x_1 *and* x_2 such that

$$P(x_1 < X < x_2) = 0.95$$

or $P(\frac{x_1-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{x_2-\mu}{\sigma}) = 0.95$

or $P(z_1 < Z < z_2) = 0.95$

We know $P(-1.96 < Z < 1.96) = 0.95$

So $z_1 = -1.96$ and $z_2 = 1.96$

Using the inverse transformation,

$x_1 = \mu + z_1\sigma$ and $x_2 = \mu + z_2\sigma$

$x_1 = 2450 + (-1.96)400$ $x_2 = 2450 + (1.96)400$

$x_1 = 1666$ $x_2 = 3234$

Therefore, the management may be 95% sure that sales in any given month will be between 1666 and 3234 units.

(b) We have to find the values x_1 *and* x_2 such that

$P(x_1 < X < x_2) = 0.99$

or $P(\frac{x_1-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{x_2-\mu}{\sigma}) = 0.99$

or $P(z_1 < Z < z_2) = 0.99$

We know $P(-2.58 < Z < 2.58) = 0.99$

So $z_1 = -2.58$ and $z_2 = 2.58$

Using the inverse transformation,

$x_1 = \mu + z_1\sigma$ and $x_2 = \mu + z_2\sigma$

$x_1 = 2450 + (-2.58)400$ $x_2 = 2450 + (2.58)400$

$x_1 = 1418$ $x_2 = 3482$

Therefore, the management may be 99% sure that sales in any given month will be between 1418 and 3482 units.

We can summarize the procedure of obtaining values of a normal random variable, given a probability, as:

- draw a picture of the normal distribution in question and the standard normal distribution
- in the picture, shade in the area corresponding to the probability

- use the table to find the z value (or values) that gives the required probability
- use the transformation from Z to X to get the appropriate value (or values) of the original normal random variable

6.5 CHECK YOUR PROGRESS

1. The probabilities associated with a continuous random variable X are determined by the …… of the random variable.
2. For a continuous random variable, non-zero probabilities are associated only with …… of numbers.
3. In science books, normal distribution is often called the …….. .
4. The normal curve has a (relative) kurtosis of 0, which means it has …… and is mesokurtic.
5. We define the standard normal random variable Z as the ……. random variable with mean = 0 and standard deviation = 1.

6.6 SUMMARY

A continuous random variable is a random variable that can take on any value in an interval of numbers. The Normal Distribution is the most versatile of all the continuous probability distributions. It is being widely used in all data-based research in the field of agriculture, trade, business and industry. It is found to be useful in characterizing uncertainties in many real-life processes, in statistical inferences, and in approximating other probability distributions. A large number of random variables occurring in practice can be approximated to the normal distribution. A random variable that is affected by many independent causes, and the effect of each cause is not overwhelmingly large compared to other effects, closely follow a normal distribution. It may be appreciated that we can simplify this process of computing range probabilities to a great extent by tabulating the range probabilities. Since it is not practicable and indeed impossible to have separate probability tables for each of the infinitely many possible normal curves, we select one normal curve to serve as a standard. Probabilities associated with the range of values of this standard normal random variable are tabulated. The importance of the standard normal distribution derives from the fact that any normal random variable may be transformed to the standard normal random variable.

6.7 KEYWORDS

Continuous random variable: The variable, say *X*, indicating the deviation of the actual volume from the normal (average) volume can take any real value - positive or negative; integer or decimal. This type of random variable, which can take an infinite number of values in a given range, is called a continuous random variable.

Continuous probability distribution: The probability distribution of continuous random variable is called a continuous probability distribution.

Standard Normal Distribution: It define the standard normal random variable Z as the normal random variable with mean = 0 and standard deviation = 1.

6.8 SELF-ASSESSMENT TEST

1. Define continuous probability distribution. State the properties of the probability density function of a continuous random variable.
2. (a) Define normal random variable. State the probability density function of a normal random variable.
   (b) List down important properties of a normal curve.
3. Discus the role of normal distribution in statistical theory.
4. What do you mean by standard normal variable? Bring out the need for having a standard normal curve.
5. Find the probability that a standard normal variable will have a value
   (a) less than –8 (b) between -0.01 and 0.05
6. A sensitive measuring device is calibrated so that errors in the measurements it provides are normally distributed with mean 0 and variance 1.00. Find the probability that a given error will be between -2 and 2.
7. The deviation of a magnetic needle from the magnetic pole in a certain area in northern Canada is a normally distributed random variable with mean 0 and standard deviation 1.00. What is the probability that the absolute value of the deviation from the north pole at a given moment will be more than 2.4?
8. Find two values of the standard normal random variable, *z* and *-z*, such that
   (a) the two corresponding "tail areas" of the distribution add to 0.01.

(b) each tail have an area of 0.05

9. Let *X be* a normally distributed random variable with mean $\mu = 16$ and standard deviation $\sigma = 3$. Find

   (a) $P(8 < X < 18)$ (b) $P(16 < X < 18)$ (c) $P(X > 14)$

10. For a normally distributed random variable with mean -44 and standard deviation 16, find the probability that the value of the random variable will be

    (a) above 0 (b) -8 (c) below 0

11. A normal random variable has mean 0 and standard deviation 4. Find the probability that the random variable will be…

    (a) above 2.5 (b) between 2 and 3 (c) below 1

12. The time it takes an international telephone operator to place an overseas phone call is normally distributed with mean 45 seconds and standard deviation 8 seconds.

    (a) What is the probability that my call will go through in less than 1 minute?

    (b) What is the probability that my call will get through in less than 40 seconds?

    (c) What is the probability that I will have to wait more than 70 seconds for my call to go through?

13. The number of votes cast in favor of a controversial proposition is believed to be approximately normally distributed with mean 8,000 and standard deviation 1,000. The proposition needs at least 9,322 votes in order to pass. What is the probability that the proposition will pass? (Assume numbers are on a continuous scale.)

14. A manufacturing company regularly consumes a special type of glue purchased from a foreign supplier. From past experience, the materials manager notes that the company's demand for glue during the uncertain lead-time is normally distributed with a mean of 187.6 gallons and a standard deviation of 12.4 gallons. The company follows a policy of placing the order when the glue stock falls to a predetermined value, called "re-order point". It the demand during lead-time exceeds the reorder level, the glue would go 'stock-out' and production process would have to stop.

    (a) If the re-order point is kept at 187.6 gallons, what is the probability that a stock-out condition would occur?

    (b) If the reorder point is kept at 200 gallons, what is the probability that a stock-out condition would occur?

(c) If the company wants to be 95% confident that the stock-out condition will not occur, what should be the reorder point? The reorder point minus the mean demand during lead-time is known as the "safety stock." What is the safety stock in this case?

(d) If the company wants to be 99% confident that the stock-out condition will not occur, what should be the reorder point? What is the safety stock in this case?

15. If X is a normally distributed random variable with mean 125 and standard deviation 44, find a value x such that the probability that X will be less than x is 0.66.
16. For a normal random variable with mean 8.5 and standard deviation 0.4, find a point of the distribution such that there is a 0.95 probability that the value of the random variable will be above it.
17. For a normal random variable with mean 29,500 and standard deviation 48, find a point of the distribution such that the probability that the random variable will exceed this value is

    (a) 0.03 (b) 0.25
18. Find two values of the normal random variable with mean 80 and standard deviation 5 lying symmetrically on either side of the mean and covering an area of 0.98 between them.
19. For $X \sim N(32, 7^2)$, find two values x_1 and x_2, symmetrically lying on each side of the mean, with

    (a) $P(x_1 < X < x_2) = 0.99$ (b) $P(x_1 < X < x_2) = 0.95$
20. The results of a given selection test exercise are summarized as

    (i) cleared with distinction = 8%

    (ii) cleared without distinction = 60%

    (iii) those who failed = 30%.

    A candidate gets failed if he/she obtains less than 40% marks, while one must obtain at least 75% marks to pass with distinction. Determine the mean and standard deviation of the distribution of marks, assuming the same to be normal.
21. The demand for gasoline at a service station is normally distributed with mean 27,009 gallons per day and standard deviation 4,530. Find two values that will give a symmetric 0.95 probability interval for the amount of gasoline demanded daily.
22. The percentage of protein in a certain brand of dog food is a normally distributed random variable with mean 11.2 % and standard deviation 0.6 %. The manufacturer would like to state on the

package that the product has a protein content of at least x_1 % and no more than x %. He wants the statement to be true for 99% of the packages sold. Determine the values x_1 and x_2.

6.9 ANSWERS TO CHECK YOUR PROGRESS

1. Probability density function
2. Intervals
3. Gaussian distribution
4. Average peakedness
5. Normal

6.10 REFERENCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.

Subject: Business Statistics	
Course Code: MBA-102	**Author: Dr. Pardeep Gupta**
Lesson: 07	**Vetter: Dr. B.S. Bodla**

SAMPLING AND SAMPLING METHODS

STRUCTURE

7.1 Learning Objectives

7.2 Introduction

7.2.1 Census Vs. Sampling method

7.2.2 Definitions

7.3 Probability samples vs. non-probability samples

7.3.1 Probability sampling methods

7.3.2 Non-probability sampling methods

7.3.3 Determination of Sample size

7.4 Sampling and Non-sampling errors

7.5 Check your Progress

7.6 Summary

7.7 Keywords

7.8 Self-Assessment Test

7.9 Answers to check your progress

7.10 References/Suggested readings

7.1 LEARNING OBJECTIVES

After going through this lesson, the students will be able to:

Understand various terms associated with sampling

Understand various methods of probability and non-probability Sampling

Understanding of how to determine sample size.

7.2 INTRODUCTION

Sampling is the procedure or process of selecting a sample from a population. A sampling can also be defined as the process of drawing a sample from a population and of compiling a suitable statistic from such a sample in order to estimate the parameter of the parent population and to test the significance of the statistic computed from such sample. When secondary data are not available for the problem under study, a decision may be taken to collect primary data by using any of the methods discussed in this lesson. The required information may be obtained by following either the census method or the sample method.

7.2.1 Census Vs. Sampling Method

Sample is a part of the population from which it is selected. The process of selecting a sample is known as sampling. Thus, the sampling theory is a study of relationship that exists between the population and the samples drawn from the population. The complete enumeration, popularly known as census, may not be feasible either due to non-availability of time or because of high cost involved. Therefore, it becomes essential to draw inferences for the population on the basis of sample information. Thus, sampling helps us to get as much information as possible of the whole universe. The sampling also helps us in determining the reliability of the estimates. This can be done by drawing samples from the same parent population and comparing the results obtained from different samples.

In a survey of the entire population, data is collected from every elementary unit of the population. Suppose, one is studying the wage structure of the coal mining industry in the country, then one approach is to collect the data on wages of every worker in the coal industry. From this data, one can calculate the various characteristics of the population, such as average wage, the range and the variance, etc. This is referred as census survey. The advantages of the census approach are

(i) every unit of the population is considered and the respective data on the various characteristics are compiled,

(ii) the analysis made on the basis of census data is very accurate and reliable, and

(iii) in one time studies of special importance, only census method is adopted in order to get accurate and reliable data. The data collected by this method becomes a data base for all future studies. This is one of the reasons why population data are collected once in a decade by the census method.

Although there are many advantages with the census method, the cost, effort and the time required to conduct census survey is very large, unless the population is very small, and in many cases it is so prohibitive that one rarely uses this method in surveys.

Sampling involves an examination of a small portion of the elementary units in a population. Although, a census operation gives a more reliable data, sampling method is more desired when

(i) the population is very large, i.e., infinite and it would be impossible to conduct census surveys;

(ii) when quick results are required it would be appropriate to conduct sample surveys rather than census surveys;

(iii) in studies involving destruction of the elementary units under study, it would only be appropriate to go for sample testing. Items such as light bulbs and ammunition often must be destroyed as a part of testing process;

(iv) cost of conducting surveys would be very prohibitive in census method, and therefore, it is advisable to carry out a sample survey, and lastly; and

(v) some times accuracy may be lost because of the large size of the population. Sampling involves a small portion of the population and therefore, would involve very few people for conducting surveys and for data collection and compilation. This would not be so in the census method and the chances of committing errors would increase.

As the sampling involves less time and money, it would be possible to give attention to different characteristics of the elementary units. A sample using same money and time can produce a detailed study of lesser number of units. The process of sampling involves selecting a sample, collecting all relevant information, and finally drawing conclusions about the population from which the sample has been drawn.

7.2.2 Definitions

The surveys are concerned with the attributes of certain entities, such as business enterprises, human beings, etc. The attributes that are the object of the study are known as characteristics and the units possessing them are called the *elementary units.*

The aggregate of elementary units to which the conclusions of the study apply is termed as *population/universe,* and the units that form the basis of the sampling process are called sampling units. The sampling unit may be an elementary unit.

The sample is defined as an aggregate of sampling units actually chosen in obtaining a representative subset from which inferences about the population are drawn. *The frame*— a list or directory, defines all the sampling units in the universe to be covered. This frame is either constructed for the purpose of a particular survey or may consist of previously available description of the population; the latter is the commonly used method. For example, telephone directory can be used as a frame for conducting opinion surveys in a city or locality.

In order that, sampling results reflect the characteristics of the population, it is necessary that the sample selected for study should be

(i) Truly representative, i.e., the selected sample truly represent the universe so that the results can be generalised;

(ii) Adequate, i.e., the size of the sample or the sample size should be adequate enough to represent the various characteristics of the universe;

(iii) Independent, i.e. the elementary units selected should be independent of one another and all units of the population should have the same chance of being selected in the sample; and lastly

(iv) Homogeneous, i.e., there should not be any basic difference between the characteristics of the units in the sample and that of the population. This means that if two or more samples are drawn from the same population, the results should be more or less identical.

7.3 PROBABILITY SAMPLES VS. NON-PROBABILITY SAMPLES

A probability sample is one for which the inclusion or exclusion of any individual element of the population depends upon the application of probability methods and not on a personal judgement. It is so designed and drawn that the probability of inclusion of an element is known. The essential feature of drawing such a sample is the randomness. As against the probability sample, we have a variety of other samples, termed as judgement samples, purposive samples, quota samples, etc. These samples have one common distinguishing feature: personal judgement rather than the random procedure to determine the composition of what is to be taken as a representative sample. The judgement affects the choice of the individual elements. All such samples are non-random, and no objective measure of precision may be attached to the results arrived at.

In a probability sampling, it is possible to estimate the error in the estimates and they can be minimized also. It is also possible to evaluate the relative efficiency of the various probability sampling designs.

Probability sampling does not depend upon the detailed information about population for its effectiveness. However, probability sampling requires a high level of skill and experience for its use. It also requires sufficient time and money to execute.

Non-probability sampling is a procedure of selecting a sample without the use of probability or randomisation. It is based on convenience, judgement, etc. The major difference between the two approaches is that it is possible to estimate the sampling variability in the case of probability sampling while it is not possible to estimate the same in the non-probability sampling. The classification of various probability and non-probability methods are shown in Fig. 7.1.

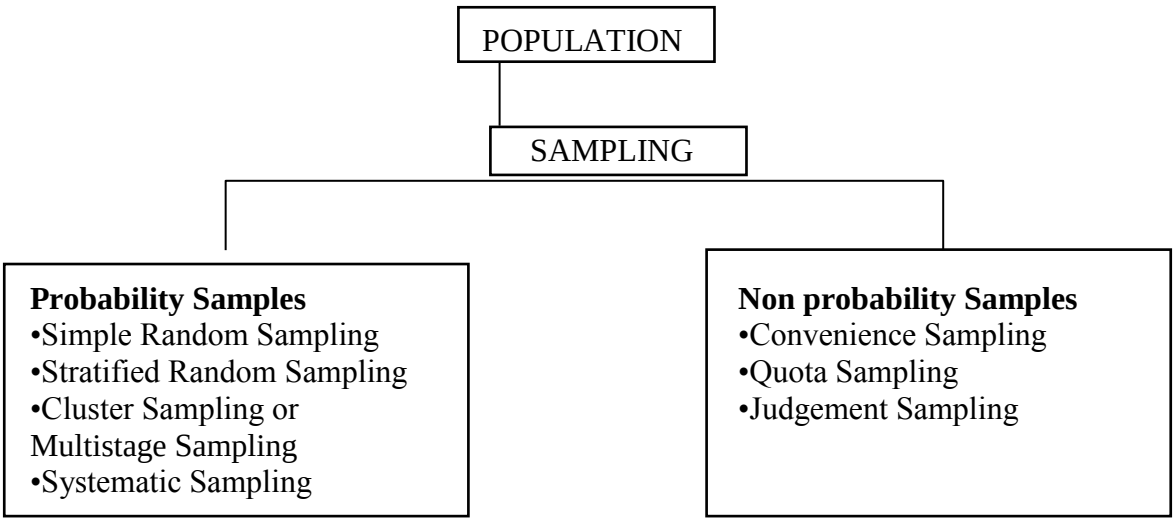


Fig. 7.1 Classification of sampling schemes

7.3.1 Probability Sampling Methods

The various probability sampling methods are described as under:

(a) Simple Random Sampling Method

In simple random sampling, drawing of elements from the population is random and the choice of an element is made in such a way that every element has the same probability of being chosen. When the sample is so selected, every possible set of elements has the same chance of being drawn. With N, population size, fairly large, the number of such possible sets of size n is of course very large. This number is given by ${}^{N}C_{n}$. Of course, it is unnecessary in a specific case to compute the number of

possible sets of stated size that might be drawn from a given population, but the process of sample selection should be such that the probability of selection is the same for every such set.

The objective is to achieve randomness in drawing the individual elements of a sample for ensuring that all possible samples have the same chance of being selected. If we are to draw from a population containing N elementary units, the elementary unit also being a sampling unit, it is necessary that each of the N units should be individually numbered or otherwise distinctively designed. One of the approaches for drawing random sample of size n from a population of N units is to draw n cards from N cards which are numbered from 1 to N and mixed thoroughly. The sample size n, thus drawn, would constitute a simple random sample (SRS). Another popular method of selecting a random sample is by lottery method. In this method all the elements are named or numbered on a small slip of paper of identical shape and size. These slips are folded identically and mixed up well in a container. Number of slips of desired sample size is selected blindly from this container. Thus, the selection of elementary units depends purely on chance and no personal bias exists. We shall illustrate this method of selection of a sample with the following example: Suppose the warden of a student's hostel with 200 occupants wants to constitute a welfare committee with the members randomly selected. The lottery method of selecting these five members from a group of 200 would be first to prepare 200 slips of identical shape and size and write the name of each student on a slip. Fold these 200 slips identically and mix them well in a container. Then select five folded slips, from the container at random. The five students so selected would constitute a welfare committee of the hostel.

There are, however, some difficulties in these procedures. For, if N is large, the task becomes physically difficult. So it is desirable to use better methods for ensuring randomness. One such method is the use of random number tables.

Use of random number tables

If the N elements of a total population are numbered serially from 1 to N, a random sample may be most readily and reliably drawn by using a table of random numbers. Such tables enable us to select n numbers at random from the full list of serial numbers from 1 to N. In a random number table, digits in each column are in random order and so are the digits in each row. As the arrangement is random in all directions, it makes no difference where we begin in our selection of random numbers from such a table. However, the column arrangement is generally found more convenient for references.

Several random number tables are available for use. These numbers have been adequately tested for randomness. Among them, the most popular ones are:

(i) Tippett's (1927) 10,400 sets of four-digited random numbers;

(ii) Fisher and Yates (1938) table of random numbers with 1,500 sets of ten-digited random numbers; and

(iii) Rand Corporation (1955) table of random numbers of 2,00,000 sets of five-digited random numbers.

Tippet's table of random numbers is most popularly used in practice. Given below are the first forty sets from Tippet's table as an illustration of the general appearance of random numbers:

2952	6641	3992	9792	7969	5911	3170	5624
4167	9524	1545	1396	7203	5356	1300	2693
2670	7483	3408	2762	3563	1089	6913	7691
0560	5246	1112	6107	6008	8125	4233	8776
2754	9143	1405	9025	7002	6111	8816	6446

Tippett's numbers have been subjected to numerous tests and used in many investigations and their randomness has been well established for all practical purposes. An example to illustrate how Tippett's table of random numbers may be used is given below.

Suppose ten numbers from out of 0 and 80 are required. We start anywhere in the table and write down the numbers in pairs. The table can be read horizontally, vertically, diagonally or in any methodical way. Starting with the first and reading horizontally first we obtain 29, 52, 66, 41, 39, 92, 97, 92, 79, 69, 59, 11, 31, 70, 56, 24, 41, 67 and so on. Ignoring the numbers greater than 80, we obtain for one purpose ten random numbers, namely 29, 52, 66, 41, 39, 79, 69, 59, 11 and 31.

The sampling procedure described above is quite satisfactory for a small population. With a large population, the process of identification of numbers to each elementary sampling unit becomes very prohibitive with respect to both time and money. Moreover, the population is often geographically spread out or composed of clearly identified strata possessing unique characteristics. Whenever any of the above situations arise, alternative sampling schemes that are sophisticated combinations of simple random sampling provide significantly better results for the same expenditure and time. As a result, the simple random sampling method is not very frequently used in practice. However, the simple random sampling scheme is the basis of any other probabilistic sampling schemes.

(b) Stratified Random Sampling Method

In simple random sampling, the population to be sampled is treated as homogeneous and the individual elements are drawn at random from the whole universe. However, it is often possible and desirable to classify the population into distinctive classes or strata and then obtain a sample by drawing at random the specified number of sampling units from each of the classes thus constructed. This may be desirable because of our interest in the distinct classes of the universe as a whole.

In stratified random sampling, the population is sub-divided into strata before the sample is drawn. Strata are so designed that they should not overlap. A sample of specified size is drawn at random from the sampling units that make up each stratum. If a given stratum is of our interest, the corresponding sub-sample provides the basis for estimates concerning the attributes of the population stratum, or sub-universe from which it is drawn. The total of sub-samples constitutes the aggregate sample on which estimates of attributes of the entire population are based.

Stratified samples may be either proportional or non-proportional. In a proportional stratified sampling, the number of elements to be drawn from each stratum is proportional to the size of that stratum compared with the population. For example, if a sample size of 500 elementary units have to be drawn from a population with 10,000 units divided in four strata in the following way:

	Population size	**Sample size**
Stratum I =	2000	500 × 0.2 = 100
Stratum II =	3000	500 × 0.3 = 150
Stratum III =	4000	500 × 0.4 = 200
Stratum IV =	1000	500 × 0.1 = 50
	Total 10000	500

Thus, the elements to be drawn from each stratum would be 100, 150, 200 and 50 respectively. Proportional stratification yields a sample that represents the population with respect to the proportion in each stratum in the population. Proportional stratified sampling yields satisfactory results if the dispersion in the various strata is of proportionately the same magnitude. If there is a significant difference in dispersion from stratum to stratum, sample estimates will be much more efficient if non-proportional stratified random sampling is used. Here, equal numbers of elements are selected from

each stratum regardless of how the stratum is represented in the population. Thus, in the earlier example, an equal number, i.e., 125, of elementary units will be drawn to constitute the sample.

A sample drawn by stratified random sampling scheme ensures a representative sample as the population is first divided into various strata and then a sample is drawn from each stratum. Stratified random sampling also ensures greater accuracy and it is maximum if each stratum is formed in such a way that it consists of uniform or homogeneous items. Compared with a simple random sample, a stratified sample can be more concentrated geographically, i.e., the elementary units from different strata may be selected in such a way that all of them are located in one geographical area. This would also reduce both time and cost involved in data collection. However, care should be exercised in dividing the population into various strata. Each stratum must contain, as far as possible, homogeneous units, as otherwise the reliability of the results would be lost.

In conclusion, stratification is an effective sampling device to the extent that it creates classes that are more homogeneous than the total. When this can be done, the classes are distinguished that differ among themselves in respect of a stated characteristic. Stratification may be futile if classes do not differ among themselves. Thus, there should be homogeneity within classes and heterogeneity between classes.

(c) Cluster Sampling or Multistage Sampling

Under this method, the random selection is made of primary, intermediate and final (or the ultimate) units from a given population or stratum. There are several stages in which the sampling process is carried out. At first, the first stage units are sampled by some suitable method, such as simple random sampling. Then, a sample of second stage unit is selected from each of the selected first stage units, again by some suitable method which may be same as or different from the method employed for the first stage units. Further stages may be added as required. The procedure may be illustrated as follows:

Suppose we want to take a sample of 5,000 households from the State of Haryana. At the first stage, the state may be divided into a number of districts and a few districts are selected at random. At the second stage, each district may be sub-divided into a number of villages and a sample of villages may be taken at random. At the third stage, a number of households may be selected from each of the villages selected at second stage. To take another example supposes in a particular survey, we wish to take a

sample of 10,000 students from a University. We may take colleges at the first stage, then draw departments at the second stage, and choose students as the third and last stage.

Merits: Multi-stage sampling introduces flexibility in the sampling method which is lacking in the other methods. It enables existing divisions and sub-divisions of the population to be used as units at various stages, and permits the field work to be concentrated and yet large area to be covered.

Another advantage of this method is that sub-division into second stage units need be carried out for only those first stage units which are included in the sample. It is, therefore, particularly valuable in surveys of under-developed areas where no frame is generally sufficiently detailed and accurate for subdivision of the material into reasonably small sampling units.

Limitations: However, a multi-stage sample is in general less accurate than a sample containing the same number of final stage units which have been selected by some suitable single stage process.

(d) Systematic Sampling

Another sampling form, simple in design and execution, may be employed when the members of population to be sampled are arranged in order, the order corresponding to consecutive numbers. The arrangements of names in a telephone directory or income-tax returns in the income tax department are the illustrations of such orderings. A sample of suitable size is obtained by taking every unit say, seventh unit of the population, one of the first seven units in this ordered arrangement is chosen at random and the sample is completely by selecting every seventh unit from the rest of the list. If the first unit selected is the fifth, the researcher will include in his sample 12th, 19th, 26th, 33rd, etc. We can generalize the approach as follows: if the requirements of the survey call for the inclusion of one unit out of every m units in the population, a unit is chosen at random from the first m units, thereafter, every mth unit in the population when arranged in order, is included in the sample. This mode of selection is called systematic sampling, m is generally referred to as the sampling ratio, i.e., the ratio of the population size to the sample size. Symbolically $m = \frac{N}{n}$.

where N is the population size and n is the sample size. While calculating the value of m, we may get a fractional value. In such cases, it is rounded off to the nearest digit.

Which sampling scheme to select

In sampling, one scheme is said to be more efficient than another when the sample estimates developed by the scheme tend to cluster more closely around the population parameter being estimated. An estimator of the population parameter should possess the following characteristics:

(i) It should be unbiased: An estimator is unbiased when the expected (average) value of the sample statistic is equal to the population parameter being estimated.

(ii) It should be efficient: Efficiency is with respect to sample size and it means that the sample estimates should be clustered as closely possible to the population parameter being estimated for a given sample size. For example, when the population is normally distributed, both the sample mean and the median are unbiased estimators of the population mean. However, for any given sample size, the sample means cluster more closely around the population mean than do the sample medians. Thus, both mean and the median are the unbiased estimators of the population mean. However, the sample mean is the unbiased efficient estimator of the population mean. In stratified random sampling, where stratification is meaningful, a stratified random sample will be more efficient than a simple random sample of the same size. A sampling design is considered efficient with respect to cost if the sample estimates cluster more closely around the population parameter being estimated than they would for any alternative sampling scheme involving equivalent rupee expenditure.

(iii) It should be consistent: An estimator is considered to be consistent if the sample estimates cluster more and more closely around the population parameter being estimated as the sample size increases.

7.3.2 Non-Probability Sampling Methods

There are three methods of sampling in this category. These are explained as follows:

1. Convenience Sampling

In this scheme, a sample is obtained by selecting 'convenient' population elements. For example, a sample selected from the readily available sources or lists such as telephone directory or a register of the small scale industrial units, etc. will give us a convenient sample. In these cases, even if a random approach is used for identifying the units, the scheme will not be considered as simple random

sampling. For example, if one studies the wage structure in a close by textile industry by interviewing a few selected workers, then the scheme adopted here is convenient sampling. The results obtained by convenience sampling method can hardly be said to be representative of the population parameters. Therefore, the results obtained are generally biased and unsatisfactory. However, convenient sampling approach is generally used for making pilot studies, particularly for testing a questionnaire and to obtain preliminary information about the population.

2. Quota Sampling

In this method of sampling, the basic parameters which describe the population are identified first. Then the sample is selected which conform to these parameters. Thus, in a quota sample, quotas are fixed according to these parameters, and each field investigator is assigned with quotas of the number of units to be interviewed. Within the preassigned quotas, the selection of the sample elements depends on the personal judgement. For example, if one is studying the consumer preferences for ice creams among children and college going students and supposes it is fixed to interview 250 individuals from each category. If the city has five colleges, one decides to fix up a quota of 50 students to be interviewed from each college. It entirely depends upon the interviewer who will constitute this sub-sample of 50 students in a college— they may be the first 50 students who visit the ice cream parlour or may be the 50 students who visit the parlour between 4 p.m. and 6 p.m., etc.

Quota sampling method has the advantage that the sample will conform to the selected parameters of the population. The cost and time involved in getting information from the sample will be relatively less for a quota sample but there are many weaknesses too. Some of these are:

(i) It is difficult to validate the information gathered on the elementary units,

(ii) It may be difficult to specify the characteristics of the population and therefore it may be hard to identify it,

(iii) Even when the sample does conform to the characteristics used in the quotas, the sample may be distorted on other factors of importance in the study. For example, interviewing first 50 students or the last 50 students visiting the ice cream parlour can make a lot of difference particularly about their purchasing capacity, tastes, etc. This may completely distort the results.

Quota sampling method is generally used in public opinion studies, election forecast polls, as there is not sufficient time to adopt a probability sampling scheme.

3. Judgement Sampling

Judgement sampling method can also be called as sampling by opinion. In this method, someone who is well acquainted with the population decides which members (elementary units) in his or her judgement would constitute a proper cross-section representing the parameters of relevance to the study. This method of sampling is generally used in studies involving performance of personnel. For example, if one is studying the performance of sales staff in a marketing organisation, the people here are classified into top grade, medium grade and low grade performers. Having specified qualities that are important in the study, the expert (possibly here the Vice-President-sales) indicates the people who, in his or her knowledge, would be representative of each of the three categories mentioned earlier. This, of course, is not a scientific method, but in the absence of better evidence, such a judgement method may have to be used.

7.3.3 Determination of sample size

We prefer samples to complete enumeration because of convenience and reduced cost of data collection. However, in sampling, there is a likelihood of missing some useful information about the population. For a high level of precision, we need to take a larger sample. How large should be the sample and what should be the level of precision? In specifying a sample size, care should be taken such that (i) neither so few are selected so as to render the risk of sampling error intolerably large, nor (ii) too many units are included, which would raise the cost of the study to make it inefficient. It is, therefore, necessary to make a trade-off between (i) increasing sample size, which would reduce the sampling error but increase the cost, and (ii) decreasing the sample size, which might increase the sampling error while decreasing the cost. Therefore, one has to make a compromise between obtaining data with greater precision and with that of lower cost of data collection. Several factors need to be considered before determining the sample size.

The first and the foremost is the size of the error that would be tolerable for the purposes of decision-making. The second consideration would be the degree of confidence with the results of the study, i.e., if one wants to be 100 per cent confident of the results, the entire population must be studied. However, this is generally too impractical and costly. Therefore, one must accept something less than 100 per cent confidence. In practice, the confidence limits most often used are 99 per cent, 95 per cent and 90 per cent. Most commonly used confidence limit is 95 per cent. This means that there is a 5 per cent risk that

the true population statistic is outside the range of possible error specified by the confidence interval. This 5 per cent risk appears to be acceptable in most of the decisions. Thus, for 95 per cent level of confidence, *Z* value is 1.96. The *Z* value can be obtained from normal probability distribution for a specified level of confidence. For determining the sample size, we make use of the following relationship:

$\sigma_{\bar{x}}$ = standard error of the estimate = $\frac{s}{\sqrt{n}}$

$\sigma_{\bar{x}}$ can be calculated if we know the upper and lower confidence limits. Let these limits be Y, then *Z*

$\sigma_{\bar{x}}$ = Y.

Where *Z* is the value of the normal variate for a given confidence level. The procedure has been explained using the illustration given below:

Illustration 7.1. A state cooperative department is performing a survey to determine the annual salary earned by managers numbering 3000 in the cooperative sector within the state. How large a sample size it should take in order to estimate the mean annual earnings within plus and minus 1,000 and at 95 per cent confidence level? The standard deviation of annual earnings of the entire population is known to be Rs. 3,000.

Solution. As the desired upper and lower limit is Rs. 1,000, i.e., we want to estimate the annual earnings within plus and minus Rs. 1,000.

$\therefore \quad z\sigma_{\bar{x}} = 1{,}000$

As the level of confidence is 95 per cent, the Z value is 1.96

$\therefore \quad 1.96\ \sigma_{\bar{x}} = 1{,}000$

$\sigma_{\bar{x}} = \frac{1{,}000}{1.96} = 510.20$

The standard error $\sigma_{\bar{x}}$ is given by $\sigma/\sqrt{n}$ where σ is the population standard deviation

$\therefore \quad \frac{s}{\sqrt{n}} = 510.20$

i.e., $\frac{3000}{\sqrt{n}} = 510.20$

i.e., $\sqrt{n} = \frac{3000}{510.2} = 5.88$

This gives n = 34.57

Therefore, the desired sample size is about 35.

7.3.1. Sample Size for Stratified Sampling

Once the strata have been established, we are interested in the size of the stratified random sample. The size will depend upon whether the proportional or disproportional (optimal) sample is being taken.

A proportional stratified sample is one in which the sample units in a given stratum are allocated in proportion to the relative size of the stratum. The following formula is used for calculation of the proportional sample for each stratum

$$n_i = \frac{N_i}{N} \times n$$

Where n_i = number of sample units from stratum i, N = the total number of units in the population, N_i = the total number of units in the stratum i, n = sample size desired.

The standard error of mean is

$$\sigma_{\bar{x}} = \sqrt{\sum_{i=1}^{k} w_i^2 s_i^2 / n_i}$$

where w_i = the weight of stratum i = N_i/N, σ_i = the standard deviation of the ith stratum, k = the total number of strata. In case of disproportionate stratified sampling, the proportion of units in the sample stratum is not equal to the proportion of the population. The formula for sample allocation in this case is

$$n_i = \frac{w_i s_i n}{\sum_{1}^{k} w_i s_i}$$

Thus, the disproportional stratified sample is more desirable if standard deviation (σ_i) of each stratum is known. The standard error of the mean of a disproportionate stratified sample is

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{k} (w_i s_i)^2}{\frac{1}{\text{å}n_i}}}$$

It may be observed that the standard error for stratified sample is smaller than for *simple* random sample, i.e., much smaller samples may be utilized when the population has been stratified.

Illustration 7.2. In a market area, shops are divided into two categories, viz., those that have daily turnover of more than Rs. 2000 and those that have daily turnover of less than Rs. 2000 for the study of estimating the total sales in the area. The total number of shops in the first stratum are 420 and in the second stratum 180. A sample of 50 was selected, the standard deviation has been found to be 70 for first stratum and 95 for second stratum. What size of stratified random sample should be taken under proportional and disproportional stratified sampling?

Solution. Under the proportional stratified sampling, the sample size is given by

$$n_i = \frac{N_i}{N} \times n$$

and, therefore $n_1 = \frac{420}{600} \times 50 = 35$

and $n_2 = \frac{180}{600} \times 50 = 15$

The standard error $(\sigma_{\bar{x}}) = \sqrt{\sum w_i^2 \frac{s_i^2}{n_i}}$

$$= \sqrt{(0.7)^2 \times \frac{(70)^2}{35} + \frac{(0.3)^2 \times (95)^2}{15}}$$

$$= \sqrt{122.75} = 11.079$$

For disproportionate sampling, the sample size is given by:

$$n_i = \frac{w_i s_i n}{\sum w_i s_i}$$

$$\therefore n_1 = \frac{0.7 \times 70 \times 50}{0.7 \times 70 + 0.3 \times 95} = \frac{2450}{77.5} = 32.0$$

$$\text{and } n_2 = \frac{0.3 \times 95 \times 50}{0.7 \times 70 + 0.3 \times 95} = \frac{1425}{77.5} = 18.0$$

The standard error is given by

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{1}^{k} (w_i s_i)^2}{\sum n_i}} = \sqrt{\frac{(0.7 \times 70 + 0.3 \times 95)^2}{50}}$$

$$= \sqrt{120.125} = 10.96$$

7.3.2. Cost as a Factor in the Determination of the Sample Size

Another consideration in determining the sample size is the cost. Management may reduce the level of confidence in an attempt to reduce the cost of sampling. An illustration will clarify how cost of sampling can be reduced by reducing the sample size.

Illustration 7.3. In a market area there are 600 shops. A researcher wishes to estimate number of customers visiting these shops per day. The researcher wants to estimate the sampling error in the number of customers visiting is no larger than ± 10 with probability of 0.95. The previous studies indicated that the standard deviation is 85 customers. If the cost per interview is Rs. 20 (this includes field work, supervision of interviewers, coding, editing and tabulation of results and report writing, etc.), calculate the total cost involved. Researcher is willing to sacrifice some accuracy in order to reduce cost. If he settles for an estimate with 0.90 probability, how much reduction in cost can be achieved?

Solution. For 95 per cent confidence levels,

$$Z\sigma_{\bar{x}} = Y$$

$$\text{i.e., } 1.96\ \sigma_{\bar{x}} = 10.0$$

$$\therefore\ \sigma_{\bar{x}} = \frac{10}{1.96}$$

Now, $\sigma_{\bar{x}}$ is given by $\sigma/\sqrt{n}$ and therefore, the sample size will be determined by the equation

$$\frac{s}{\sqrt{n}} = \frac{10}{1.96}$$

Since $\sigma = 85$, we have

$$\frac{85}{\sqrt{n}} = \frac{10}{1.96}$$

$$\therefore\ n = 277.6$$

Thus, if the sample is taken as 278, the total cost involved will be 278 × 20 = Rs. 5560. As this cost is considered to be on the higher side by the researcher and in order to reduce the cost, the researcher has

now settled to 90 per cent confidence level. At 90 per cent confidence level, the sample size can be calculated as follows:

$$Z\,\sigma_{\bar{x}} = 10$$

$$1.65\,\sigma_{\bar{x}} = 10$$

$$1.65\,\sigma_{\bar{x}} = 10$$

$$\text{or } \sigma_{\bar{x}} = \frac{10}{1.65}$$

$$\frac{s}{\sqrt{n}} = \frac{10}{1.65}$$

$$\text{i.e., } \frac{85}{\sqrt{n}} = \frac{10}{1.65}$$

$$n = 196.7$$

The cost of survey for this sample size will be 197 × 20 = Rs. 3940. Thus, we have observed that by reducing the confidence level from 95 per cent to 90 per cent, the researcher would reduce the cost from Rs. 5560 to Rs. 3940. The researcher may not like to reduce the confidence level further and so further cost reduction may not be desirable.

7.4 SAMPLING AND NON-SAMPLING ERRORS

The choice of a sample though may be made with utmost care, involves certain errors which may be classified into two types: (a) Sampling errors, and (b) Non-Sampling errors. These errors may occur in the collection, processing and analysis of data.

(a) Sampling Errors

Sampling errors are those which arise due to the method of sampling. Sampling errors arise primarily due to the following reasons:

(1) Faulty selection of the sampling method.

(2) Substituting one sample for the sample due to the difficulties in collecting the sample.

(3) Faulty demarcation of sampling units.

(4) Variability of the population which has different characteristics.

(b) Non-Sampling Errors

Non-Sampling errors are those which creep in due to human factors which always vary from one investigator to another. These errors arise due to any of the following factors:

(1) Faulty planning.

(2) Faulty selection of the sample units.

(3) Lack of trained and experienced staff which collect the data.

(4) Negligence and non-response on the part of the respondent.

(5) Errors in compilation.

(6) Errors due to wrong statistical measures.

(7) Framing of wrong questionnaire.

(8) Incomplete investigation of the sample survey.

7.5 CHECK YOUR PROGRESS

1. In simple random sampling, drawing of elements from the population is ..... and the choice of an element is made in such a way that every element has the same probability of being chosen.

2. In stratified random sampling, the population is sub-divided into ..... before the sample is drawn.

3. In convenience sampling, a sample is obtained by selecting ..... population elements.

4. In a quota sample, quotas are fixed according to these parameters, and each field investigator is assigned with quotas of the ..... to be interviewed.

5. One has to make a compromise between obtaining data with greater ..... and with that of lower cost of data collection.

7.6 SUMMARY

The process of selecting a sample is known as sampling. Thus, the sampling theory is a study of relationship that exists between the population and the samples drawn from the population. The complete enumeration, popularly known as census, may not be feasible either due to non-availability of time or because of high cost involved. A probability sample is one for which the inclusion or exclusion of any individual element of the population depends upon the application of probability methods and not on a personal judgement. It is so designed and drawn that the probability of inclusion of an element is known. The essential feature of drawing such a sample is the randomness. As against the probability sample, we have a variety of other samples, termed as judgement samples, purposive samples, quota

samples, etc. These samples have one common distinguishing feature: personal judgement rather than the random procedure to determine the composition of what is to be taken as a representative sample. The judgement affects the choice of the individual elements. All such samples are non-random, and no objective measure of precision may be attached to the results arrived at.

Non-probability sampling is a procedure of selecting a sample without the use of probability or randomisation. It is based on convenience, judgement, etc. Several factors need to be considered before determining the sample size.The first and the foremost is the size of the error that would be tolerable for the purposes of decision-making. The second consideration would be the degree of confidence with the results of the study

7.7 KEYWORDS

Elementary Units: The attributes that are the object of the study are known as characteristics and the units possessing them are called the elementary units.

Population: The aggregate of elementary units to which the conclusions of the study apply is termed as population/universe.

Sampling Unit: The units that form the basis of the sampling process are called sampling units. The sampling unit may be an elementary unit.

Sample: The sample is defined as an aggregate of sampling units actually chosen in obtaining a representative subset from which inferences about the population are drawn.

Frame: A list or directory, defines all the sampling units in the universe to be covered.

Cluster sampling or multistage sampling: *Under this method, the random selection is made of primary, intermediate and final (or the ultimate) units from a given population or stratum. There are several stages in which the sampling process is carried out.*

Judgement sampling method: In this method, someone who is well acquainted with the population decides which members (elementary units) in his or her judgement would constitute a proper cross-section representing the parameters of relevance to the study.

7.8 SELF-ASSESSMENT TEST

1. Describe the various methods of drawing a sample. Which one would you suggest and why?

2. Describe the importance of sampling. Critically examine the merits of probability sampling and non-probability sampling methods.
3. Specify and explain the factors that make sampling preferable to a complete census in a statistical investigation.
4. How would you determine the sample size for stratified sampling? Explain with the help of a suitable example.
5. To determine the effectiveness of the advertising campaign of a new VCR, management would like to know what percentage of the household are aware of the new brand. The advertising agency thinks that this figure is as high as 70 per cent. The management would like a 95% confidence interval and a margin of error no greater than plus or minus 2 per cent. (a) What sample size should be used for this study? (b) Suppose that management wanted to be 99 per cent confident but could tolerate an error of plus or minus 3 per cent. How would the sample size change?

7.9 ANSWERS TO CHECK YOUR PROGRESS

1. Random
2. Strata
3. Convenient
4. Number of units
5. Precision

7.10 REFERENCES/SUGGESTED READINGS

1. Statistical Methods by S.P. Gupta. Sultan Chand and Sons, New Delhi.
2. Statistics for MBA by T.R. Jain and Dr. S.C. Aggarwal. VK (India) Enterprises, New Delhi. First edition.
3. Business Statistics by Shenoy and Shenoy.
4. Levin, Richard I. and David S. Rubin: Statistics for Management, Prentice Hall, New Delhi.
5. Lawrance B. Moore: Statistics for Business & Economics, Harper Collins, NY.
6. Watsman Terry J. and Keith Parramor: Quantitative Methods in Finance International, Thompson Business Press, London.

Subject: Business Statistics	
Course code: MBA-102	Author: Anil Kumar
Lesson: 08	Vetter: Dr. B. S. Bodla
SAMPLING DISTRIBUTIONS	

STRUCTURE

8.1 Learning Objectives

8.2 Introduction

8.2.1 Sampling Distribution of the Mean

8.2.2 Central Limit Theorem

8.2.3 Sampling Distribution of the Proportion

8.2.4 Sampling Distribution of the Difference of Sample Means

8.2.5 Sampling Distribution of the Difference of Sample Proportions

8.2.6 Small Sampling Distributions

8.2.7 Sampling Distribution of the Variance

8.3 F Distribution and ANOVA

8.4 t-Distributions and Z-Distributions

8.5 Check your Progress

8.6 Summary

8.7 Keywords

8.8 Self-Assessment Questions

8.9 Answers to check your progress

8.10 References/Suggested Readings

8.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- **Understand the concept of sampling distributions.**
- **Meaning and the need of studying sampling distribution of a sample statistic.**

8.2 INTRODUCTION

Having discussed the various methods available for picking up a sample from a population, we would naturally be interested in drawing statistical inferences - making generalizations about the population on the basis of a sample drawn from it. The generalizations to be made about the population are usually either by way of

- Estimating the unknown population parameters, or
- Testing appropriate hypotheses stated in relation to population parameters in the light of sample data

These generalizations, together with the measurement of their reliability, are made in terms of the relationship between the values of any ***sample statistic*** and those of the corresponding ***population parameters***. Population parameter is any number computed (or estimated) for the entire population *viz.* population mean, population median, population proportion, population variance and so on. Population parameter is unknown but fixed, whose value is to be estimated from the sample statistic that is known but random. Sample Statistic is any numbers computed from our sample data viz. sample mean, sample median, sample proportion, sample variance and so on.

It may be appreciated that no single value of the sample statistic is likely to be equal to the corresponding population parameter. This owes to the fact that the sample statistic being random, assumes different values in different samples of the same size drawn from the same population.

Referring to our earlier discussion on the concept of a random variable in the lessons on Probability Distributions, it is not difficult to see that ***any sample statistics is a random variable*** and, therefore, has a probability distribution better known as the Sampling Distribution of the statistic.

The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take when computed from random samples of the same size drawn from a specified population.

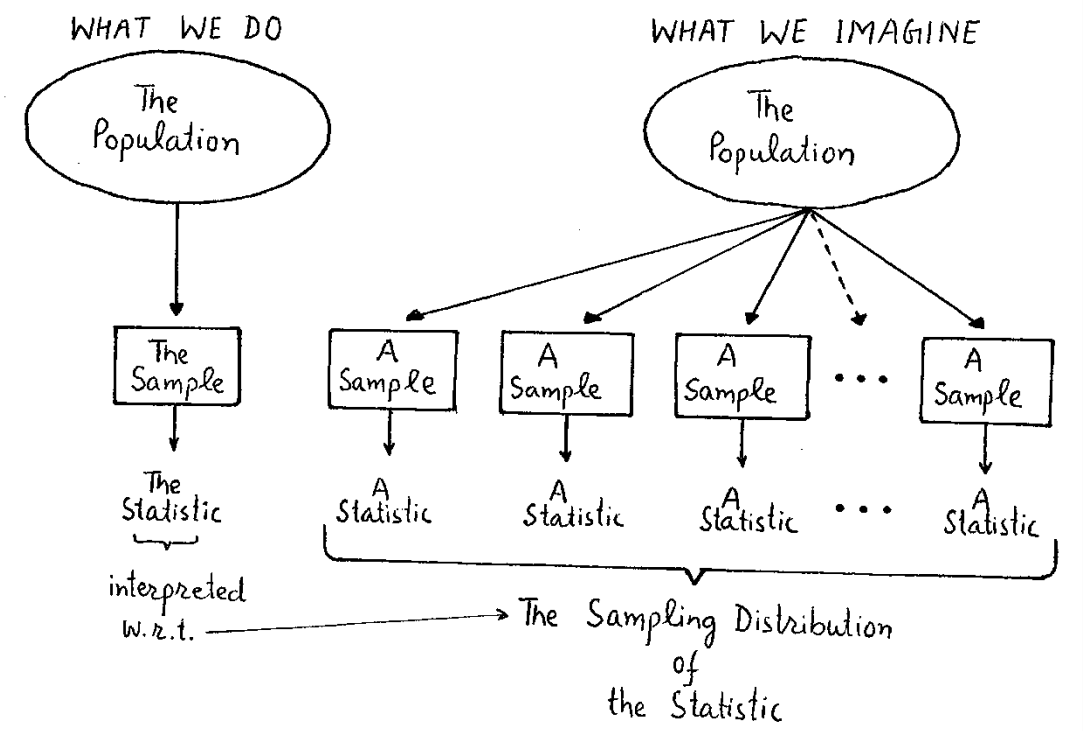


Figure 8-1 Sampling Distribution of a Statistic

In reality, of course we do not have all possible samples and all possible values of the statistic. We have only one sample and one value of the statistic. This value is interpreted with respect to all other outcomes that might have happened, as represented by the sampling distribution of the statistic. In this lesson, we will refer to the sampling distributions of only the commonly used sample statistics like sample mean, sample proportion, sample variance *etc.*, which have a role in making inferences about the population.

Why We Study Sampling Distributions?

Sample statistics form the basis of all inferences drawn about populations. Thus, sampling distributions are of great value in inferential statistics. The sampling distribution of a sample statistic possess well-defined properties which help lay down rules for making generalizations about a population on the basis of a single sample drawn from it. The variations in the value of sample statistic not only determine the shape of its sampling distribution, but also account for the element of error in statistical inference. If we know the probability distribution of the sample statistic, then we can calculate risks (error due to chance) involved in making generalizations about the population. With the help of the properties of

sampling distribution of a sample statistic, we can calculate the probability that the sample statistic assumes a particular value (if it is a discrete random variable) or has a value in a given interval. This ability to calculate the probability that the sample statistic lies in a particular interval is the most important factor in all statistical inferences. We will demonstrate this by an example.

Suppose we know that 40% of the population of all users of hair oil prefers our brand to the next competing brand. A "new improved" version of our brand has been developed and given to a random sample of 100 users for use. If 55 of these prefer our "new improved" version to the next competing brand, what should we conclude? For an answer, we would like to know the probability that the sample proportion in a sample of size 100 is as large as 55% or higher when the true population proportion is only 40%, *i.e.* assuming that the new version is no better than the old. If this probability is quite large, say 0.5, we might conclude that the high sample proportion *viz.* 55% is perhaps because of sampling errors and the new version is not really superior to the old. On the other hand, if this probability works out to a very small figure, say 0.001, then rather than concluding that we have observed a rare event we might conclude that the true population proportion is higher than 40%, *i.e*. the new version is actually superior to the old one as perceived by members of the population. To calculate this probability, we need to know the probability distribution of sample proportion *i.e.* the sampling distribution of the proportion.

8.2.1 Sampling Distribution of the Mean

Suppose we have a simple random sample of size *n*, picked up from a population of size *N*. We take measurements on each sample member in the characteristic of our interest and denote the observation as $x_1, x_2, ...x_n$ respectively. The sample mean for this sample is defined as:

$$\overline{X} = \frac{x_1 + x_2 + ...+ x_n}{n}$$

If we pick up another sample of size *n* from the same population, we might end up with a totally different set of sample values and so a different sample mean. Therefore, there are many (perhaps infinite) possible values of the sample mean and the particular value that we obtain, if we pick up only one sample, is determined only by chance. In other words, the sample mean is a random variable. The possible values of this random variable depends on the possible values of the elements in the random sample from which sample mean is to be computed. The random sample, in turn, depends on the

distribution of the population from which it is drawn. As a random variable, $\overline{X}$ has a *probability distribution.* This probability distribution is the sampling distribution *of* $\overline{X}$.

The sampling distribution of $\overline{X}$ ***is the probability distribution of all possible values the random variable*** $\overline{X}$ ***may take when a sample of size n is taken from a specified population.***

To observe the distribution of $\overline{X}$ empirically, we have to take many samples of size *n* and determine the value of $\overline{X}$ for each sample. Then, looking at the various observed values of $\overline{X}$, it might be possible to get an idea of the nature of the distribution. We will derive the distribution of $\overline{X}$ in three cases:

(a) Sampling from infinite populations

(b) Sampling with replacement from finite populations

(c) Sampling without replacement from finite populations

8.2.1.1 Sampling from Infinite Populations

Let us assume we have a population, with mean μ and variance σ^2, which is infinitely large. If we take a sample of size *n* with individual values $x_1, x_2, \ldots . x_n$, then

$$\text{Sample Mean } \left(\overline{X}\right) = \frac{x_1 + x_2 + \ldots . . + x_n}{n}$$

here x_1 representing the first observed values in the sample, is a random variable since it may take any of the population values. Similarly x_2, representing the second observed value in sample is also a random variable since it may take any of the population values. In other words, we can say that x_i, representing the ith observed value in the sample is a random variable.

Now when the population is infinitely large, whatever is the value of x_1, the distribution of x_2 is not affected by it. This is true for any other pair of random variables as well. In other words; $x_1, x_2, \ldots . x_n$ are independent random variables and all are picked up from the same population.

So $E\left(x_i\right) = \mu$ and $Var\left(x_i\right) = \sigma^2$ for $i = 1, 2, 3, \ldots \ldots n$

Finally, we have

$$\mu_{\bar{x}} = E\left(\overline{X}\right) = E\left(\frac{x_1 + x_2 + \ldots . . + x_n}{n}\right)$$

$$= E\left(\frac{x_1}{n}\right) + E\left(\frac{x_2}{n}\right) + \ldots\ldots + E\left(\frac{x_n}{n}\right) \quad [\text{as } E(\text{A} + \text{B}) = E(\text{A}) + E(\text{B})]$$

$$= \frac{1}{n}E(x_1) + \frac{1}{n}E(x_2) + \ldots\ldots + \frac{1}{n}E(x_2) \quad [\text{as } E(n\text{A}) = n\,E(\text{A})]$$

$$= \frac{1}{n}\,\mu + \frac{1}{n}\,\mu + \ldots\ldots + \frac{1}{n}\,\mu = \mu \text{ and}$$

$$\sigma^2_{\bar{x}} = Var(\bar{X}) = Var\left(\frac{x_1 + x_2 + \ldots\ldots + x_n}{n}\right)$$

$$= Var\left(\frac{x_1}{n}\right) + Var\left(\frac{x_2}{n}\right) + \ldots\ldots + Var\left(\frac{x_n}{n}\right)$$

[as $Var(\text{A} + \text{B}) = Var\,(\text{A}) + Var\,(\text{B})$]

$$= \frac{1}{n^2}Var(x_1) + \frac{1}{n^2}Var(x_2) + \ldots\ldots + \frac{1}{n^2}Var(x_n) \quad [\text{as } Var(n\text{A}) = n^2\,Var(\text{A})]$$

$$= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \ldots\ldots + \frac{1}{n^2}\sigma^2$$

$$= \frac{\sigma^2}{n}$$

So, $\sigma_{\bar{x}} = SD(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$

8.2.1.2 Sampling with Replacement from Finite Populations

The above results have been obtained under the assumption that the random variables $x_1, x_2, \ldots\ldots x_n$ are independent. This assumption is valid when the population is infinitely large. It is also valid when the sampling is done with replacement, so that the population is back to the same form before the next sample member is picked up. Hence, if the sampling is done with replacement, we would again have:

$\mu_{\bar{x}} = E\left(\bar{X}\right) = \mu$ and $\sigma^2_{\bar{x}} = Var(\bar{X}) = \dfrac{\sigma^2}{n}$ or $\sigma_{\bar{x}} = SD(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$

8.2.1.3 Sampling without Replacement from Finite Populations

When sampling without replacement from a finite population, the probability distribution of the second random variable depends on what has been the outcome of the first pick and so on. In other words, the *n*

random variables representing the *n* sample members do not remain independent, the expression for the variance of $\overline{X}$ changes. The results in this case will be:

$$\mu_{\bar{x}} = E\left(\overline{X}\right) = \mu$$

and $\sigma^2_{\bar{x}} = Var(\overline{X}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1}$ or $\sigma_{\bar{x}} = S.D(\overline{X}) = \frac{\sigma}{\sqrt{n}}.\sqrt{\frac{N-n}{N-1}}$

By comparing these expressions with the ones derived above we find that the variance of $\overline{X}$ is the same but further multiplied by a factor $\frac{N-n}{N-1}$. This factor is, therefore, known as the finite population multiplier or the correction factor. In practice, almost all the samples are picked up without replacement. Also, most populations are finite although they may be very large and so the variance of the mean should theoretically be found by using the expression given above. However, if the population size (*N*) is large and consequently the sampling ratio (*n/N*) small, then the finite population multiplier is close to 1 and is not used, thus treating large finite populations as if they were infinitely large. For example, if *N* = 100,000 and *n* = 100, the finite population multiplier will be 0.9995, which is very close to 1 and the variance of the mean would, for all practical purposes, be the same whether the population is treated as finite or infinite. As a rule of that, the finite population multiplier may not be used if the sampling ratio (*n/N*) is smaller than 0.05.

Above discussion on the sampling distribution of mean, presents two very important results, which we shall be using very often in statistical estimation and hypotheses testing. We have seen that the expected value of the sample mean is the same as the population mean. Similarly, that the variance of the sample mean is the variance of the population divided by the sample size (and multiplied by the correction factor when appropriate). The fact that the sampling distribution of $\overline{X}$ has mean μ is very important. It means that, *on the average,* the sample mean is equal to the population mean. The distribution of the statistic is *centered on* the parameter to be estimated, and this makes the statistic $\overline{X}$ a good estimator of μ. This fact will become clearer in the next lesson, where we will discuss estimators and their properties. The fact that the standard deviation *of* $\overline{X}$ is $\sigma/\sqrt{n}$ means that as the sample size *increases*, the standard

deviation of $\overline{X}$ *decreases*, making $\overline{X}$ more likely to be close to μ. This is another desirable property of a good estimator, to be discussed in the next lesson.

If we take a large number of samples of size *n*, then the average value of the sample means tends to be close to the true population mean. On the other hand, if the sample size is increased then the variance of $\overline{X}$ gets reduced and by selecting an appropriately large value of *n*, the variance of $\overline{X}$ can be made as small as desired.

The standard deviation of $\overline{X}$ is also called the *standard error of the mean*. It indicates the extent to which the observed value of sample mean can be away from the true value, due to sampling errors. For example, if the standard error of the mean is small, we may be reasonably confident that whatever sample mean value we have observed cannot be very far away from the true value.

Before discussing the shape of the sampling distribution of mean, let us verify the above results empirically, with the help of a simple example.

Consider a discrete uniform population consisting of the values 1, 2, and 3. If the random variable *X* represents these population values, its mean and variance is

$$\mu = \frac{\sum X_i}{N} = \frac{6}{3} = 2$$

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N} = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}$$

(a) Sampling with Replacement

If random samples of size $n = 2$ are drawn with replacement from this population, we will have $N^n = 3^2 = 9$ possible samples. These are shown in Box 8-1 along with the corresponding sample mean values, which vary from 1 to 3. The resulting distribution of $\overline{X}$ is given below:

$\overline{X}$	:	1	1.5	2	2.5	3
$P(\overline{X})$	:	1/9	2/9	3/9	2/9	1/9

Box 8-1

Sample No. 1 (1,1) $\overline{X}$ = 1	Sample No. 2 (1,2) $\overline{X}$ = 1.5	Sample No. 3 (1,3) $\overline{X}$ = 2
Sample No. 4 (2,1) $\overline{X}$ = 1.5	Sample No. 5 (2,2) $\overline{X}$ = 2	Sample No. 6 (2,3) $\overline{X}$ = 2.5
Sample No. 7 (3,1) $\overline{X}$ = 2	Sample No. 8 (3,2) $\overline{X}$ = 2.5	Sample No. 9 (3,3) $\overline{X}$ = 3

Now we can find out the mean and variance of the sampling distribution, the necessary calculations are given in Table 8-1.

Table 8-1 Calculations for $\mu_{\bar{x}}$ and $\sigma^2_{\bar{x}}$

$\overline{X}$	$P(\overline{X})$	$X.P(\overline{X})$	$P(\overline{X}).[\overline{X}-E(\overline{X})]^2$
1	1/9	1/9	1/9
1.5	2/9	3/9	2/36
2	3/9	6/9	0
2.5	2/9	5/9	2/36
3	1/9	3/9	1/9
	$\sum P(\overline{X})=1$	$\sum X.P(\overline{X})=2$	$\sum P(\overline{X}).[\overline{X}-E(\overline{X})]^2=1/3$

So the mean of the sampling distribution,

$$\mu_{\bar{x}} = E(\overline{X}) = \sum X.P(\overline{X}) = 2 = \mu$$

and the variance of the sampling distribution,

$$\sigma^2_{\bar{x}} = Var(\overline{X}) = \sum P(\overline{X}).[\overline{X}-E(\overline{X})]^2 = 1/3 = \frac{2/3}{2} = \frac{\sigma^2}{n}$$

(b) Sampling Without Replacement

If random samples of size n = 2 are drawn without replacement from this population, we will have ${}^{N}P_n = {}^{3}P_2$ = 6 possible samples. These are shown in Box 8-2 along with the corresponding sample mean values, which vary from 1.5 to 2.5.

Box 8-2

Sample No. 1 (1,2) $\overline{X}$ = 1.5	Sample No. 2 (1,3) $\overline{X}$ = 2	Sample No. 3 (2,1) $\overline{X}$ = 1.5
Sample No. 4 (2,3) $\overline{X}$ = 2.5	Sample No. 5 (3,1) $\overline{X}$ = 2	Sample No. 6 (3,2) $\overline{X}$ = 2.5

The resulting distribution of $\overline{X}$ is given below:

$\overline{X}$	:	1.5	2	2.5
$P(\overline{X})$	:	2/6	2/6	2/6

Now we can find out the mean and variance of the sampling distribution, the necessary calculations are given in Table 8-2.

Table 8-2 Calculations for $\mu_{\bar{x}}$ and $\sigma^2_{\bar{x}}$

$\overline{X}$	$P(\overline{X})$	$X.P(\overline{X})$	$P(\overline{X}).[\overline{X}-E(\overline{X})]^2$
1.5	2/6	3/6	2/24
2	2/6	4/6	0
2.5	2/6	5/6	2/24
	$\sum P(\overline{X})=1$	$\sum X.P(\overline{X})=2$	$\sum P(\overline{X}).[\overline{X}-E(\overline{X})]^2=1/6$

So the mean of the sampling distribution,

$$\mu_{\bar{x}} = E(\overline{X}) = \sum X.P(\overline{X}) = 2 = \mu$$

and the variance of the sampling distribution,

$$\sigma^2_{\bar{x}} = Var(\overline{X}) = \sum P(\overline{X}).[\overline{X}-E(\overline{X})]^2 = 1/6 = \frac{2/3}{2}.\frac{3-2}{3-1} = \frac{\sigma^2}{n}.\frac{N-n}{N-1}$$

Now if we compare the shapes of the parent population and the resulting sampling distribution of mean, we find that although our parent population is uniformly distributed, the sampling distribution of mean is symmetrically distributed as shown in Figure 8-2.

If we increase the sample size *n* we observe an interesting and important fact. As *n* increases.

- the possible values $\overline{X}$ can assume increases, so the number of rectangles increases
- the probability that $\overline{X}$ assumes a particular value decreases *i.e.* the width of rectangles decreases

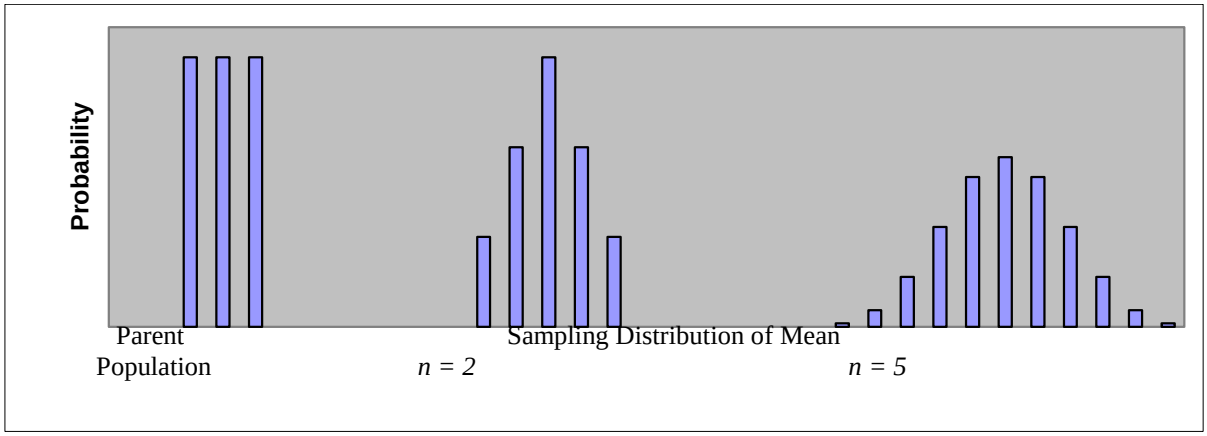
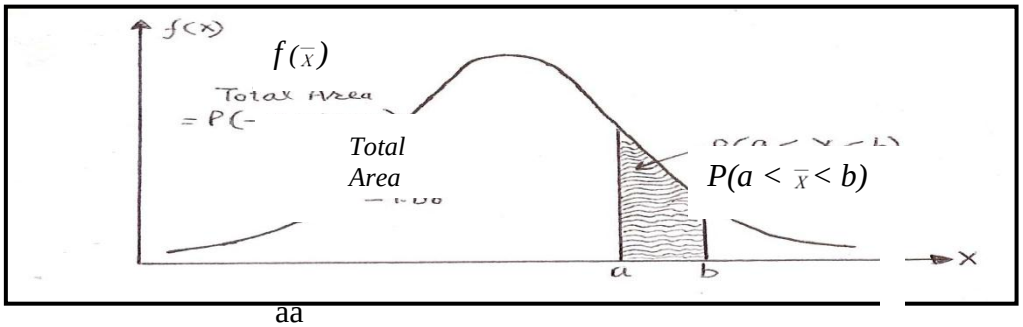


Figure 8-2 Parent Population and Sampling Distribution of Mean for *n = 2* and *n = 5*

In the limiting case when the sample size *n* increases infinitely, the particular values $\overline{X}$ can assume approaches infinity and the probability that $\overline{X}$ assumes a particular value approaches to zero. In other words, the limiting distribution of $\overline{X}$ is normal distribution. Thus as $n \rightarrow \propto$ $\overline{X} \sim N(\mu, \sqrt{\sigma^2/n}^2)$



aa

Figure 8-3 Limiting Distribution of $\overline{X}$

8.2.3 THE CENTRAL LIMIT THEOREM

The result we just stated - *the limiting distribution of* $\overline{X}$ *is the normal distribution* - is one of the most important results in statistics. It is popularly known as the *central limit theorem. When sampling is done from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean* $\overline{X}$ *tends to a normal distribution with mean μ and standard deviation* $\sigma/\sqrt{n}$ *as the sample size n increases.* For "Large Enough" *n:* $\overline{X} \sim N(\mu, \sqrt{\sigma^2/n}^2)$. The central limit theorem is remarkable because it states that the distribution of the sample mean $\overline{X}$ tends to a normal distribution *regardless* of the distribution of the population from which the random sample is drawn. The theorem allows us to make probability statements about the possible range of values the sample mean may take. It allows us

to compute probabilities of how far away $\overline{X}$ may be from the population mean it estimates. We will extensively use the central limit theorem in the next two lessons about statistical estimation and testing of hypotheses.

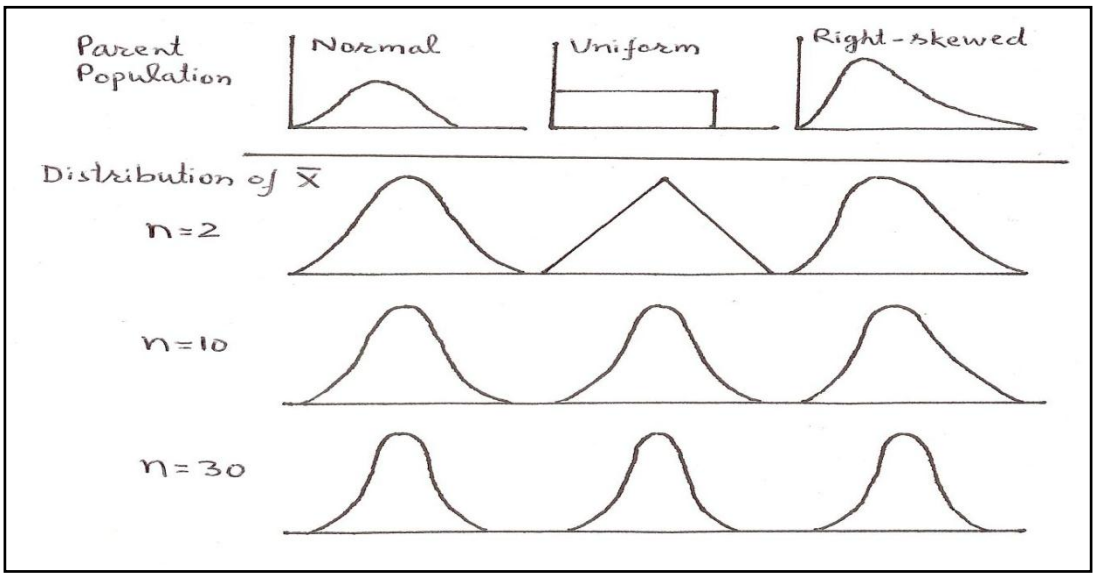


Figure 8-4 Sampling Distributions of $\overline{X}$ for different Sample Sizes

The central limit theorem says that, *in the limit,* as *n goes to* infinity ($n \rightarrow \infty$), the distribution of $\overline{X}$ becomes a normal distribution (regardless of the distribution of the population). The *rate at* which the distribution approaches a normal distribution does depend, however, on the shape of the distribution of the parent population:

- if the population itself is normally distributed, the distribution of $\overline{X}$ is normal for *any* sample size *n*
- if the population distributions are very different from a normal distribution, a relatively large sample size is required to achieve a good normal approximation for the distribution of $\overline{X}$

Figure 8-4 shows several parent population distributions and the resulting sampling distributions of $\overline{X}$ for different sample sizes.

Since we often do not know the shape of the population distribution, it would be useful to have some general rule of thumb telling us when a sample is "Large Enough" that we may apply the central limit theorem:

In general, a sample of 30 or more elements is considered "Large Enough" for the central limit theorem to be applicable.

We emphasize that this is a *general,* and somewhat arbitrary, rule. A larger minimum sample size may be required for a good normal approximation when the population distribution is very different from a normal distribution. By the same reason, a smaller minimum sample size may suffice for a good normal approximation when the population distribution is close to a normal distribution.

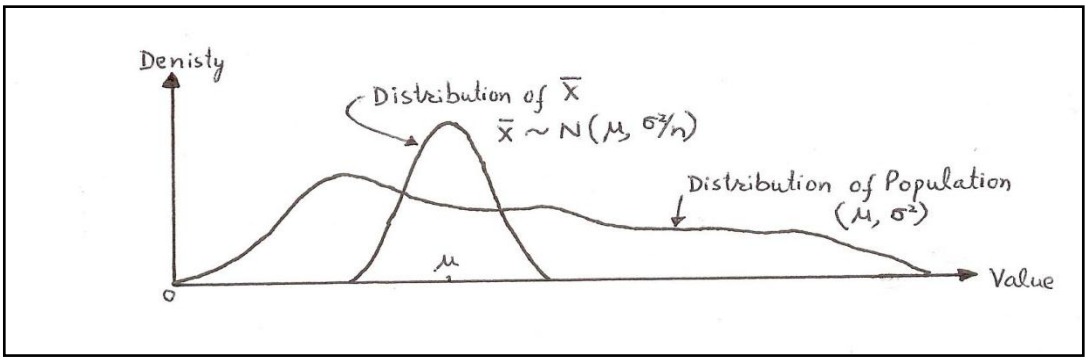


Figure 8-5 Population Distribution and the Sampling Distribution of $\overline{X}$

Figure 8-5 should help clarify the distinction between the population distribution and the sampling distribution of $\overline{X}$. The figure emphasizes the three aspects of the central limit theorem:

1. When the sample size is large enough, the sampling distribution *of* $\overline{X}$ is normal
2. The expected value of $\overline{X}$ is μ
3. The standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$

The last statement is the key to the important fact that as the sample size increases, the variation of $\overline{X}$ about its mean μ decreases. Stated another way, as we buy *more information* (take a larger sample), our *uncertainty* (measured by the standard deviation) about the parameter being estimated *decreases.*

The History of the Central Limit Theorem

What we call the central limit theorem actually comprises several theorems developed over the years. The first such theorem was the discovery of the normal curve by Abraham De Moivre in 1733, when he discovered the normal distribution as the *limit of* the binomial distribution. The fact that the normal distribution appears as a limit of the binomial distribution as *n* increases is a form of the central limit theorem. Around the turn of the twentieth century, Liapunov gave a more general form of the central limit theorem, and in 1922 Lindeberg gave the final form we use in applied statistics. In 1935, W Feller gave the proof of the necessary condition of the theorem.

Let us now look at an example of the use of the central limit theorem.

Example 8-1

ABC Tool Company makes *Laser XR;* a special engine used in speedboats. The company's engineers believe that the engine delivers an average power of 220 horsepower, and that the standard deviation of power delivered is 15 horsepower. A potential buyer intends to sample 100 engines (each engine to be run a single time). What is the probability that the sample mean $\overline{X}$ will be less than 217 horsepower?

Solution: Given that:

Population mean	μ = 220 horsepower
Population standard deviation	σ = 15 horsepower
Sample size	n = 100

Here our random variable $\overline{X}$ is normal (or at least approximately so, by the central limit theorem as our sample size is large).

$$\overline{X} \sim N(\mu, \sqrt{\sigma^2/n}^2)$$

or $$\overline{X} \sim N(220, \sqrt{15^2/100}^2)$$

So we can use the standard normal variable $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ to find the required probability,

$$P(\overline{X} < 217) = P(Z < \frac{217 - 220}{15/\sqrt{100}}) = P(Z < -2) = 0.0228$$

So there is a small probability that the potential buyer's tests will result in a sample mean less than 217 horsepower.

8.2.4 SAMPLING DISTRIBUTION OF THE PROPORTION

Let us assume we have a binomial population, with a proportion p of the population possesses a particular attribute that is of interest to us. This also implies that a proportion q (=1-p) of the population does not possess the attribute of interest. If we pick up a sample of size n with replacement and found x successes in the sample, the sample proportion of success ($\overline{p}$) is given by $\overline{p} = \dfrac{x}{n}$ in which x is a binomial random variable, the possible value of this random variable depends on the composition of the

random sample from which $\bar{p}$ is computed. The probability of *x* successes in the sample of size *n* is given by a binomial probability distribution, *viz*.

$$P(x) = {}^{n}C_x \, p^x \, q^{n-x}$$

Since $\bar{p} = \frac{x}{n}$ and *n* is fixed (determined before the sampling) the distribution of the number of successes (*x*) leads to the distribution of $\bar{p}$.

The sampling distribution of $\bar{p}$ is the probability distribution of all possible values the random variable $\bar{p}$ may take when a sample of size n is taken from a specified population.

The expected value and the variance of *x i.e.* number of successes in a sample of size *n* is known to be:

$$E(x) = n\,p; \qquad Var\,(x) = n\,p\,q$$

Finally we have mean and variance of the sampling distribution of $\bar{p}$

$$\mu_{\bar{p}} = E\left(\bar{p}\right) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x = \frac{1}{n}.n\,p = p \quad \text{and}$$

$$\sigma^2_{\bar{p}} = Var\left(\bar{p}\right) = Var\left(\frac{x}{n}\right) = \frac{1}{n^2}.\,Var(x) = \frac{1}{n^2}.\,n\,p\,q = \frac{pq}{n} \quad \sigma_{\bar{p}} = SD\left(\bar{p}\right) = \sqrt{\frac{pq}{n}}$$

When sampling is without replacement, we can use the finite population correction factor, so sampling distribution of $\bar{p}$ has its

Mean $\mu_{\bar{p}} = p$

Variance $\sigma^2_{\bar{p}} = \frac{pq}{n}.\left(\frac{N-n}{N-1}\right)$

Standard deviation $\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}.\frac{N-n}{N-1}}$

As the sample size *n* increases, the central limit theorem applies here as well. The *rate at* which the distribution approaches a normal distribution does depend, however, on the shape of the distribution of the parent population.

- if the population is symmetrically distributed, the distribution of $\bar{p}$ approaches the normal distribution relatively fast

- if the population distributions are very different from a symmetrical distribution, a relatively large sample size is required to achieve a good normal approximation for the distribution of $\bar{p}$

In order to use the normal approximation for the sampling distribution of $\bar{p}$, the sample size needs to be large. A commonly used rule of thumb says that the normal approximation to the distribution of $\bar{p}$ may be used only if *both n p and n q are greater than 5.* We now state the central limit theorem when sampling for the population proportion $\bar{p}$.

When sampling is done from a population with proportion p, the sampling distribution of the sample proportion $\bar{p}$ ***approaches to a normal distribution with proportion p and standard deviation*** $\sqrt{pq/n}$ ***as the sample size n increases.***

For "Large Enough" *n:* $\bar{p} \sim N(p, \sqrt{pq/n}^2)$

The estimated standard deviation of $\bar{p}$ is also called its *standard error.* We demonstrate the use of the theorem in Example 10-2

Example 8-2

A manufacturer of screws has noticed that on an average 0.02 proportion of screws produced are defective. A random sample of 400 screws is examined for the proportion of defective screws. Find the probability that the proportion of the defective screws ($\bar{p}$) in the sample is between 0.01 and 0.03?

Solution:

Given that:

Population proportion	$p = 0.02$
So	$q = 0.08$ (= 1-0.02)
Sample size	$n = 400$

Since the population is infinite and also the sample size is large, the central limit theorem applies. So

$$\bar{p} \sim N(p, \sqrt{pq/n}^2)$$

$$\bar{p} \sim N(0.02, \sqrt{(0.02)(0.08)/400}^2)$$

We can find the required probability using standard normal variable $Z = \left(\frac{\bar{p} - p}{\sqrt{pq/n}} \right)$

$$P(0.01 < \bar{p} < 0.03) = P\left(\frac{0.01-0.02}{\sqrt{\frac{(0.02)(0.08)}{400}}} < Z < \frac{0.03-0.02}{\sqrt{\frac{(0.02)(0.08)}{400}}}\right)$$

$$= P\left(\frac{-0.01}{0.007} < Z < \frac{0.01}{0.007}\right)$$

$$= P(-1.43 < Z < 1.43)$$

$$= 2\,P(0 < Z < 1.43)$$

$$= 0.8472$$

So there is a very high probability that the sample will result in a proportion between 0.01 and 0.03.

8.2.5 Sampling Distribution of the Difference of Sample Means

In order to bring out the sampling distribution of the difference of sample means, let us assume we have two populations labeled as 1 and 2. So that

μ_1 and μ_2 denote the two population means.

σ_1 and σ_2 denote the two population standard deviations

n_1 and n_2 denote the two sample sizes

$\overline{X_1}$ and $\overline{X_2}$ denote the two sample means

Let us consider independent random sampling from the populations so that the sample sizes need not be same for both populations.

Since $\overline{X_1}$ and $\overline{X_2}$ are random variables so is their difference $\overline{X_1} - \overline{X_2}$. As a random variable, $\overline{X_1} - \overline{X_2}$ has a *probability distribution.* This probability distribution is the sampling distribution *of* $\overline{X_1} - \overline{X_2}$.

The sampling distribution of $\overline{X_1} - \overline{X_2}$ ***is the probability distribution of all possible values the random variable*** $\overline{X_1} - \overline{X_2}$ ***may take when independent samples of size*** n_1 ***and*** n_2 ***are taken from two specified populations.***

Mean and Variance of $\overline{X_1} - \overline{X_2}$

$\mu_{\overline{X_1}-\overline{X_2}} = E\left(\overline{X_1} - \overline{X_2}\right) = E\left(\overline{X_1}\right) - E\left(\overline{X_2}\right) = \mu_1 - \mu_2$ and

$\sigma^2_{\overline{X_1}-\overline{X_2}} = Var\left(\overline{X_1} - \overline{X_2}\right) = Var\left(\overline{X_1}\right) + Var\left(\overline{X_2}\right)$

$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$; when sampling is with replacement

$= \frac{\sigma_1^2}{n_1}.\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{\sigma_2^2}{n_2}.\left(\frac{N_2 - n_2}{N_2 - 1}\right)$; when sampling is without replacement

As the sample sizes n_1 and n_2 increases, the central limit theorem applies here as well. So we state the central limit theorem when sampling for the difference of population means $\overline{X_1} - \overline{X_2}$

When sampling is done from two populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively, the sampling distribution of the difference of sample means $\overline{X_1} - \overline{X_2}$ approaches to a normal distribution with mean $\mu_1 - \mu_2$ and standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ as the sample sizes n_1 and n_2 increases.

For "Large Enough" n_1 and n_2: $\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}^{\,2}\right)$

The estimated standard deviation of $\overline{X_1} - \overline{X_2}$ is also called its *standard error*. We demonstrate the use of the theorem in Example 10-3.

Example 8-3

The makers of Duracell batteries claims that the size AA battery lasts on an average of 45 minutes longer than Duracell's main competitor, the Energizer. Two independent random samples of 100 batteries of each kind are selected. Assuming $\sigma_1 = 84$ minutes and $\sigma_2 = 67$ minutes, find the probability that the difference in the average lives of Duracell and Energizer batteries based on samples does not exceed 54 minutes.

Solution: Given that:

$\mu_1 - \mu_2 = 45$

$\sigma_1 = 84$ and $\sigma_2 = 67$

$n_1 = 100$ and $n_2 = 100$

Let $\overline{X_1}$ and $\overline{X_2}$ denote the two sample average lives of Duracell and Energizer batteries respectively. Since the population is infinite and also the sample sizes are large, the central limit theorem applies.

i.e $\overline{X_1} - \overline{X_2} \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}^2)$

$\overline{X_1} - \overline{X_2} \sim N(45, \sqrt{\frac{84^2}{100} + \frac{67^2}{100}}^2)$

So we can find the required probability using standard normal variable

$$Z = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

So $P(\overline{X_1} - \overline{X_2} < 54) = P(Z < \frac{54-45}{\sqrt{\frac{84^2}{100} + \frac{67^2}{100}}})$

$= P(Z < 0.84) = 1 - 0.20045 = 0.79955$

So there is a very high probability that the difference in the average lives of Duracell and Energizer batteries based on samples does not exceed 54 minutes.

8.2.6 Sampling Distribution of the Difference of Sample Proportions

Let us assume we have two binomial populations labeled as 1 and 2. So that

p_1 and p_2 denote the two population proportions

n_1 and n_2 denote the two sample sizes

$\overline{p_1}$ and $\overline{p_2}$ denote the two sample proportions

Let us consider independent random sampling from the populations so that the sample sizes need not be same for both populations.

Since $\overline{p_1}$ and $\overline{p_2}$ are random variables so is their difference $\overline{p_1} - \overline{p_2}$. As a random variable, $\overline{p_1} - \overline{p_2}$ has a *probability distribution.* This probability distribution is the sampling distribution *of* $\overline{p_1} - \overline{p_2}$.

The sampling distribution of $\overline{p_1} - \overline{p_2}$ ***is the probability distribution of all possible values the random variable*** $\overline{p_1} - \overline{p_2}$ ***may take when independent samples of size*** n_1 ***and*** n_2 ***are taken from two specified binomial populations.***

Mean and Variance of $\overline{p_1} - \overline{p_2}$

$\mu_{\overline{p_1}-\overline{p_2}} = E(\overline{p_1} - \overline{p_2}) = E(\overline{p_1}) - E(\overline{p_2}) = p_1 - p_2$

$\sigma^2_{\overline{p_1}-\overline{p_2}} = Var(\overline{p_1} - \overline{p_2}) = Var(\overline{p_1}) + Var(\overline{p_2})$

$= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$; when sampling is with replacement

$= \frac{p_1 q_1}{n_1}.\left(\frac{N_1 - n_1}{N_1 - 1}\right) + \frac{p_2 q_2}{n_2}.\left(\frac{N_2 - n_2}{N_2 - 1}\right)$; when sampling is without replacement

As the sample sizes n_1 and n_2 increases, the central limit theorem applies here as well. So we state the central limit theorem when sampling for the difference of population proportions $\overline{p_1} - \overline{p_2}$

When sampling is done from two populations with proportions p_1 and p_2 respectively, the sampling distribution of the difference of sample proportions $\overline{p_1} - \overline{p_2}$ approaches to a normal distribution with mean $p_1 - p_2$ and standard deviation $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$ as the sample sizes n_1 and n_2 increases.

For "Large Enough" n_1 and n_2: $\overline{p_1} - \overline{p_2} \sim N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}^2\right)$

The estimated standard deviation of $\overline{p_1} - \overline{p_2}$ is also called its *standard error*. We demonstrate the use of the theorem in Example 10-4.

Example 8-4

It has been experienced that proportions of defaulters (in tax payments) belonging to business class and professional class are 0.20 and 0.15 respectively. The results of a sample survey are:

	Business class	Professional class
Sample size:	$n_1 = 400$	$n_2 = 420$
Proportion of defaulters:	$\overline{p_1} = 0.21$	$\overline{p_2} = 0.14$

Find the probability of drawing two samples with a difference in the two sample proportions larger than what is observed.

Solution: Given that:

$p_1 = 0.20$ $p_2 = 0.15$

$q_1 = 1\text{-}0.20 = 0.80$ $q_2 = 1\text{-}0.15 = 0.85$

$n_1 = 400$ $n_2 = 420$

$\overline{p_1} = 0.21$ $\overline{p_2} = 0.14$

Since the population is infinite and also the sample sizes are large, the central limit theorem applies. *i.e.*

$$\overline{p_1} - \overline{p_2} \sim N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}^2\right)$$

$$\overline{p_1} - \overline{p_2} \sim N\left(0.05, \sqrt{\frac{(0.20)(0.80)}{400} + \frac{(0.15)(0.85)}{420}}^2\right)$$

So we can find the required probability using standard normal variable $Z = \dfrac{(\overline{p_1} - \overline{p_2}) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}}$

$$P(\overline{p_1} - \overline{p_2} > 0.07) = P\left(Z > \frac{0.07 - 0.05}{\sqrt{\dfrac{(0.20)(0.80)}{400} + \dfrac{(0.15)(0.85)}{400}}}\right)$$

$$= P(Z > 0.76)$$

$$= 0.22363$$

So there is a low probability of drawing two samples with a difference in the two sample proportions larger than what is observed.

8.2.7 Small Sampling Distributions

Up to now we were discussing the large sampling distributions in the sense that the various sampling distributions can be well approximated by a normal distribution for "Large Enough" sample sizes. In other words, the *Z*-statistic is used in statistical inference when sample size is large. It may, however, be appreciated that the sample size may be prohibited from being large either due to physical limitations or due to practical difficulties of sampling costs being too high. Consequently, for our statistical inferences, we may often have to contend ourselves with a small sample size and limited information. The consequences of the sample being small; $n < 30$; are that

- the central limit theorem ceases to operate, and
- the sample variance S^2 fails to serve as an unbiased estimator of σ^2

Thus, the basic difference which the sample size makes is that while the sampling distributions based on large samples are approximately normal and sample variance S^2 is an unbiased estimator of σ^2, the same does not occur when the sample is small.

It may be appreciated that the small sampling distributions are also known as exact sampling distributions, as the statistical inferences based on them are not subject to approximation. However, the assumption of population being normal is the basic qualification underlying the application of small sampling distributions.

In the category of small sampling distributions, the Binomial and Poisson distributions were already discussed in lesson 9. Now we will discuss three more important small sampling distributions – the chi-square, the *F* and the student *t*-distribution. The purpose of discussing these distributions at this stage is limited only to understanding the variables, which define them and their essential properties. The application of these distributions will be highlighted in the next two lessons.

The small sampling distributions are defined in terms of the concept of degrees of freedom. We will discuss this before concept proceeding further.

Degrees of Freedom *(df)*

The concept of degrees of freedom *(df)* is important for many statistical calculations and probability distributions. We may define *df* associated with a sample statistic as ***the number of observations contained in a set of sample data which can be freely chosen.*** It refer to ***the number of independent variables which vary freely without being influenced by the restrictions imposed by the sample statistic(s) to be computed.***

Let $x_1, x_2 \ldots . x_n$ be *n* observations comprising a sample whose mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a value known to us.

Obviously, we are free to assign any value to *n*-1 observation out of *n* observations. Once the value are freely assigned to *n*-1observations, freedom to do the same for the nth observation is lost and its value is automatically determined as

n^{th} observation $= n\bar{x}$ - sum of *n*-1 observations = $n\bar{x} - \sum_{i=1}^{n-1} x_i$ As the value of nth observation must satisfy the restriction $\sum_{i=1}^{n} x_i = n\bar{x}$ We say that one degree of freedom, *df* is lost and the sum $n\bar{x}$ of *n* observations has *n*-1 *df* associated with it.

For example, if the sum of four observations is 10, we are free to assign any value to three observations only, say, $x_1 = 2, x_2 = 1 \text{and} x_3 = 4$. Given these values, the value of fourth observation is automatically determined as

$$x_4 = \sum_{i=1}^{4} x_i - (x_1 + x_2 + x_3)$$

$$x_4 = 10 - (2 + 1 + 4)$$

$$x_4 = 3$$

Sampling essentially consists of defining various sample statistics and to make use them in estimating the corresponding population parameters. In this respect, degrees of freedom may be defined as ***the number of n independent observations contained in a sample less the number of parameters m to be estimated on the basis of that sample information, i.e. df = n-m.***

For example, when the population variance σ^2 is not known, it is to be estimated by a particular value of its estimator S^2*;* the sample variance. The number of observations in the sample being *n*, *df = n-m = n-*1 because σ^2 is the only parameter (*i.e. m* =1) to be estimated by the sample variance.

8.2.8 Sampling Distribution of the Variance

We will now discuss the sampling distribution of the variance. We will first introduce the concept of the sample variance as an unbiased estimator of population variance and then present the chi-square distribution, which helps us in working out probabilities for the sample variance.

8.2.8.1 THE SAMPLE VARIANCE

By now it is implicitly clear that we use the sample mean to estimate the population mean and sample proportion to estimate the population proportion, when those parameters are unknown. Similarly, we use a sample statistic called the sample variance to estimate the population variance.

As will see in the next lesson on Statistical Estimation a sample statistic is an unbiased estimator of the population parameter when the expected value of sample statistic is equal to the corresponding population parameter it estimates.

Thus, if we use the sample variance S^2 as an unbiased estimator of population variance σ^2 Then, $E(S^2) = \sigma^2$

However, it can be shown empirically that while calculating S^2 if we divide the sum of square of deviations from mean (SSD) *i.e.* $\sum_{i=1}^{n}\left(x-\bar{x}\right)^2$ by *n*, it will not be an unbiased estimator of σ^2 and $E\left(\frac{\sum_{i=1}^{n}\left(x-\bar{x}\right)^2}{n}\right) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{\sigma^2}{n}$ *i.e.* $\frac{\sum\left(x-\bar{x}\right)^2}{n}$ will underestimate the population variance σ^2 by the factor $\frac{\sigma^2}{n}$. To compensate for this downward bias we divide $\sum_{i=1}^{n}\left(x-\bar{x}\right)^2$ by *n*-1, so that $S^2 = \frac{\sum_{i=1}^{n}\left(x-\bar{x}\right)^2}{n-1}$ is an unbiased estimator of population variance σ^2 and we have:

$$E\left(\frac{\sum_{i=1}^{n}\left(x-\bar{x}\right)^2}{n-1}\right) = \sigma^2$$

In other words *to get the unbiased estimator of population variance* σ^2*, we divide the sum* $\sum_{i=1}^{n}\left(x-\bar{x}\right)^2$ *by the degree of freedom n*-1.

8.2.8.2 The Chi-Square Distribution

Let *X* be a random variable representing *N* values of a population, which is normally distributed with mean μ and variance σ^2, *i. e.*

$$X = \left\{X_1, X_2 \ldots\ldots X_N\right\}$$

We may draw a random sample of size *n* comprising $x_1, x_2 \ldots\ldots x_n$ values from this population. As brought out in section 10.2, each of the *n* sample values $x_1, x_2 \ldots\ldots x_n$ can be treated as an independent normal random variable with mean μ and variance σ^2. In other words,

$$x_i \sim N(\mu, \sigma^2) \qquad \text{where } i = 1, 2, \ldots\ldots n$$

Thus each of these *n* normally distribution random variable may be standardized so that

$$Z_i = \frac{x_i - \mu}{\sigma} \sim N(0, 10) \qquad \text{where } i = 1, 2, \ldots\ldots n$$

A sample statistic *U* may, then, be defined as

$$U = Z_1^2 + Z_2^2 + \ldots\ldots + Z_n^2$$

$$U = \sum_{i=1}^{n} Z_i^2$$

$$U = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

Which will take different values in repeated random sampling. Obviously, *U* is a random variable. It is called chi-square variable, denoted by χ^2. Thus *the chi-square random variable is the sum of several independent, squared standard normal random variables.* The chi-square distribution is the probability distribution of chi-square variable. So, *The chi-square distribution is the probability distribution of the sum of several independent, squared standard normal random variables.* The chi-square distribution is defined as

$$f(\chi^2) = Ce^{-\frac{1}{2}\chi^2} (\chi^2)^{\frac{n}{2}-1} d\chi^2 \qquad \text{for} \qquad \chi^2 \geq 0$$

where e is the base of natural logarithm, *n* denotes the sample size (or the number of independent normal random variables).*C* is a constant to be so determined that the total area under the χ^2 distribution is unity. χ^2 values are determined in terms of degrees of freedom, *df = n*

Properties of χ^2 Distribution

1. A χ^2 distribution is completely defined by the number of degrees of freedom, *df = n*. So there are many χ^2 distributions each with its own *df.*
2. χ^2 is a sample statistic having no corresponding parameter, which makes χ^2distribution a non-parametric distribution.
3. As a sum of squares the χ^2 random variable cannot be negative and is, therefore, bounded on the left by zero.

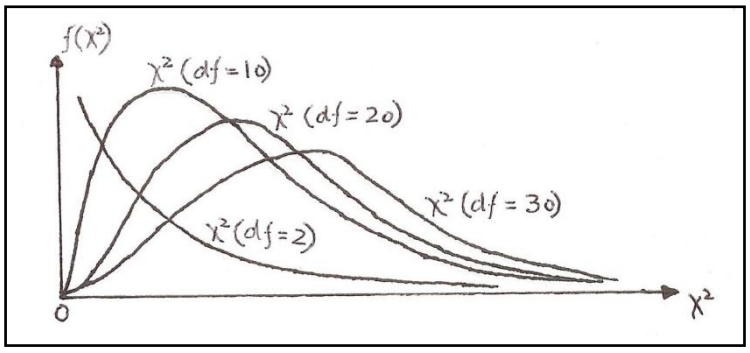


Figure 8-6 χ^2 Distribution with Different Numbers of *df*

4. The mean of a χ^2 distribution is equal to the degrees of freedom *df*. The variance of the distribution is equal to twice the number of degrees of freedom *df* .

$$E(\chi^2) = n; \text{ Var } (\chi^2) = 2n$$

5. Unless the *df* is large, a χ^2 distribution is skewed to the right. As *df* increases, the χ^2 distribution looks more and more like a normal. Thus for large *df*

$$\chi^2 \sim N\,(n, \sqrt{2n}^2)$$

Figure 10-6 shows several χ^2 distributions with different numbers of *df*. In general, for $n \geq 30$, the probability of χ^2 taking a value greater than or less than a particular value can be approximated by using the normal area tables.

6. If $\chi_1^2, \chi_2^2, \chi_3^2, \ldots\ldots\chi_k^2$ are *k* independent χ^2 random variables, with degrees of freedom $n_1, n_2, n_3, \ldots\ldots n_k$. Then their sum $\chi_1^2 + \chi_2^2 + \chi_3^2 + \ldots\ldots + \chi_k^2$ also possesses a χ^2 distribution with *df* $= n_1 + n_2 + n_3 + \ldots\ldots + n_k$.

8.2.8.3 The χ^2 Distribution in terms of Sample Variance S^2

We can write

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(x_i-\bar{x})+(\bar{x}-\mu)\right]^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(x_i-\bar{x})^2+(\bar{x}-\mu)^2+2(x_i-\bar{x})(\bar{x}-\mu)\right]$$

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

$$= \frac{(n-1)S^2}{\sigma^2}+\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2$$

$$\left[\text{since } \sum_{i=1}^{n}(x_i-\bar{x})^2 = (n-1)S^2;\ \sum_{i=1}^{n}(\bar{x}-\mu) = n(\bar{x}-\mu) \text{ and } \sum_{i=1}^{n}(x_i-\bar{x}) = 0\right]$$

Now, we know that the LHS of the above equation is a random variable which has chi-square distribution, with *df* = *n* . We also know that if $\bar{x} \sim N\,(\mu, \sqrt{\sigma^2/n}^2)$ Then $\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2$

will have a chi-square distribution with *df = 1,* Since the two terms on the RHS are independent, $\frac{(n-1)S^2}{\sigma^2}$ will also has a chi-square distribution with *df* = *n*-1. One degree of freedom is lost because all the deviations are measured from $\bar{x}$ and not from μ.

Expected Value and Variance of S^2

In practice, therefore, we work with the distribution of $\frac{(n-1)S^2}{\sigma^2}$ and not with the distribution of S^2 directly.

Since $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with *df* = *n*-1

So $$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1$$

$$\frac{n-1}{\sigma^2}E(S^2) = n-1$$

$$E(S^2) = \sigma^2$$

Also $Var\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1)$

Using the definition of variance, we get

$$E\left[\frac{(n-1)S^2}{\sigma^2} - E\left(\frac{(n-1)S^2}{\sigma^2}\right)\right]^2 = 2(n-1)$$

or $$E\left[\frac{(n-1)S^2}{\sigma^2} - (n-1)\right]^2 = 2(n-1)$$

or $$E\left[\frac{(n-1)^2 S^4}{\sigma^4} + (n-1)^2 - 2(n-1)\frac{(n-1)S^2}{\sigma^2}\right]^2 = 2(n-1)$$

or $$\frac{(n-1)^2}{\sigma^4}E\left[S^4 + \sigma^4 - 2S^2\sigma^2\right]^2 = 2(n-1)$$

or $$\frac{(n-1)^2}{\sigma^4}E(S^2 - \sigma^2)^2 = 2(n-1)$$

or $$E(S^2 - \sigma^2)^2 = \frac{2(n-1)}{(n-1)^2}\sigma^4$$

So $$Var(S^2) = \frac{2\sigma^4}{n-1}$$

It may be noted that the conditions necessary for the central limit theorem to be operative in the case of sample variance S^2 are quite restrictive. For the sampling distribution of S^2 to be approximately normal requires not only that the parent population is normal, but also that the sample is at least as large as 100.

Example 8-5

In an automated process, a machine fills cans of coffee. The variance of the filling process is known to be 30. In order to keep the process in control, from time to time regular checks of the variance of the filling process are made. This is done by randomly sampling filled cans, measuring their amounts and computing the sample variance. A random sample of 101 cans is selected for the purpose. What is the probability that the sample variance is between 21.28 and 38.72?

Solution: We have Population variance $\sigma^2 = 30$, $n = 101$

We can find the required probability by using the chi-square distribution

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

So, $$P(21.28 < S^2 < 38.72) = P\left(\frac{(101-1)21.28}{30} < \chi^2 < \frac{(101-1)38.72}{30}\right)$$

$$= P(70.93 < \chi^2 < 109.06)$$

$$= P(\chi^2 > 70.93) - P(\chi^2 > 109.06)$$

$$\approx 0.990 - 0.025 = 0.965$$

Since our population is normal and also sample size is quite large, we can also estimate the required probability using normal distribution.

We have $$S^2 \sim (\sigma^2, \sqrt{\frac{2\sigma^4}{n-1}}^2)$$

So $P(21.28 < S^2 < 38.72) = P\left(\frac{21.28-\sigma^2}{\sqrt{\frac{2\sigma^4}{n-1}}} < Z < \frac{38.72-\sigma^2}{\sqrt{\frac{2\sigma^4}{n-1}}}\right)$

$$= P\left(\frac{21.28-30}{\sqrt{\frac{2x30x30}{101-1}}} < Z < \frac{38.72-30}{\sqrt{\frac{2x30x30}{101-1}}}\right)$$

$$= P\left(\frac{-8.72}{4.36} < Z < \frac{8.72}{4.36}\right)$$

$$= P(-2 < Z < 2)$$

$$= 2P(0 < Z < 2)$$

$$= 2x0.4772 = 0.9544$$

Which is approximately the same as calculated above using χ^2 distribution

8.3 The *f*-Distribution and Analysis of Variance (ANOVA)

Let us assume two normal population with variances σ_1^2 and σ_2^2 repetitively. For a random sample of size n_1 drawn from the first population, we have the chi-square variable.

$$\chi_1^2 = \frac{(n_1-1)S_1^2}{\sigma_1^2}$$ which process a χ^2 distribution with $\nu_1 = n_1 - 1$ *df*

Similarly, for a random sample of size n_2 drawn from the second population, we have the chi-square variable

$$\chi_2^2 = \frac{(n_2-1)S_2^2}{\sigma_2^2}$$ which process a χ^2 distribution with $\nu_2 = n_2 - 1$ *df*

A new sample statistic defined as $F = \frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2}$

It is a random variable known as ***F statistic***, named in honor of the English statistician Sir Ronald A Fisher.

Being a random variable it has a probability distribution, which is known as ***F* distribution**.

The F distribution is the distribution of the ratio of two chi-square random variables that are independent of each other, each of which is divided by its own degrees of freedom.

Properties of *F*- Distribution

1. The *F* distribution is defined by two kinds of degrees of freedom – the degrees of freedom of the numerator always listed as the first item in the parentheses and the degrees of freedom of the denominator always listed as the second item in the parentheses. So there are a large number of *F* distributions for each pair of v_1 and v_2. Figure 10-7 shows several *F* distributions with different v_1 and v_2.
2. As a ratio of two squared quantities, the *F* random variable cannot be negative and is, therefore, bounded on the left by zero.

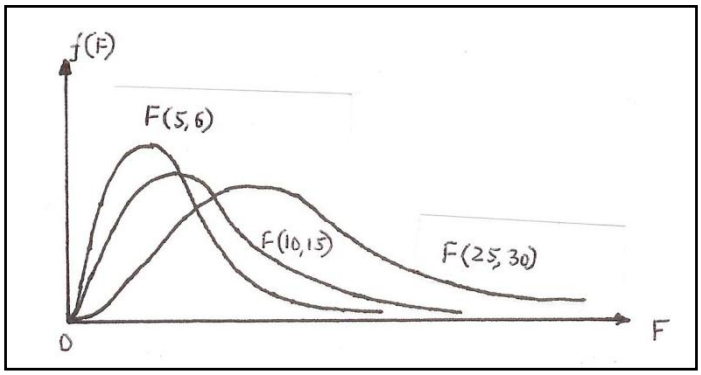


Figure 8-7 *F*- Distribution with different v_1 and v_2

3. The $F_{(v_1,v_2)}$ has no mean for $v_2 \leq 2$ and no variance for $v_2 \leq 4$. However, for v_2 >2, the mean and for $v_2 > 4$, the variance is given as

$$E(F_{(v_1,v_2)}) = \frac{v_2}{v_2 - 2} \qquad Var(F_{(v_1,v_2)}) = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)}$$

4. Unless the v_2 is large, a *F* distribution is skewed to the right. As v_2 increases, the *F* distribution looks more and more like a normal. In general, for $v_2 \geq 30$, the probability of *F* taking a value greater than or less than a particular value can be approximated by using the normal area tables.
5. The *F* distributions defined as $F_{(v_1,v_2)}$ and as $F_{(v_2,v_1)}$ are reciprocal of each other.

$$i.e. \qquad F_{(v_1,v_2)} = \frac{1}{F_{(v_2,v_1)}}$$

Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal. This technique is developed by R.A. Fisher in 1920's. It consists of classifying and cross-classifying statistical results and testing whether the means of a specified classification differ significantly. ANOVA is a method which separate the variations ascribable to one set of causes from the variations ascribable to other set of causes. In other words, analysis of variance is a method of splitting the total variations of a data into constituents parts in which measures different sources of variations. ANOVA enables us to analyze the total variations of data into components which may be attributed to various sources or causes of variation. The total variation is split up into the following two components:

a) Variation within the subgroup of samples.

b) Variations between the subgroups of the samples.

After obtaining the above two variations in viz., a) and b), these two variations are tested for their significance by F-test which is also known as Variance Ratio Test.

Objects of Analysis of Variance

The first objects of analysis of variance is to obtain a measure of the total variation within the series and the second object is to find a measure of variation between or among the components. Then the test of significance of difference between the variations in two series or may be measured. In other words, with ANOVA technique, we can test the hypothesis that the means of all the components constituting a population are equal to the mean of the population or that the samples have come from the same population.

Computation of Test Statistic

The actual analysis of variance is carried out on the basis of ratio between two variances. The variance ratio is obtained by the dividing the variance between the samples by the variance within the samples. This ratio forms the test statistic known as F-Statistic, i.e.,

$$\mathrm{F-Statistic} = \frac{\text{Variance between the samples}}{\text{Variance within the samples}}$$

Assumptions of Analysis of Variance (ANOVA)

The underlying assumptions for the study of analysis of variance are:

1. Each of the samples are a simple random sample.

2. Population from which the samples are selected is normally distributed. If however, the sample sizes are large enough, this assumption of normality is not required.
3. Each one of the samples are independent of other samples.
4. Each one of the populations has the same variance ($\sigma^2_1 = \sigma^2_2 = \ldots\ldots = \sigma^2_n$) and identical means ($\mu_1 = \mu_2 = \mu_3 = \ldots\ldots\ldots = \mu_n$).
5. The effect of various components are additive.

Uses of ANOVA Table

The ANOVA table showing the source of variation, the sum of squares, degree of freedom, mean square (variance) and the formula for the F-ratio is known as ANOVA table. It is used to test whether the means of a number of populations (more than two) are equal. We know that t-statistic is used for testing whether two population means are equal. Thus, the analysis of variance of test may be taken as an extension of t-test for the case of more than two population means.

Classification of Analysis of Variance

The analysis of variance is mainly carried on under the following two classifications:

a) One way Analysis of Variance or One way classification
b) Two way Analysis of Variance or Two way classified Data or Manifold Classification

One way Classification

In one way classification, the data are classified according to only one criterion. In this classification, the influence of only one attributes or factors considered. There are two methods of one way analysis of variance. They are:

a) Direct Method
b) Shortcut Method

Direct Method

The following steps are required under the direct method of one way classification of analysis of variance (ANOVA):

1. **Set Null Hypothesis and Alternate Hypothesis:**

   Null Hypothesis: The means of the populations from which p samples are drawn equal to one another. The notation for Null hypothesis will be as:

$$\mathbf{H_0: \mu_1 = \mu_2 = \ldots\ldots\ldots = \mu_k}$$

Alternate Hypothesis: At least two of the means of the populations are unequal or all the μ_i's are not equal. The notation for Alternate hypothesis will be as:

$$\mathbf{H_1: \ \mu_1 \neq \mu_2 \neq \ldots\ldots\ldots \neq \mu_k}$$

2. **Calculate Variance Between the Samples:** The variance between samples (groups) measures the difference between the sample means of each group and the overall mean weighted by the number of observations in each group. The variance between samples taken into account the random variations from observation to observation. It also measures difference from one group to another. The sum of squares between samples are denoted by SSC. For calculating variance between the samples we take the total of the square of the deviations of the means of various samples from the grand average and divide this total by degree of freedom. Thus the steps in calculating variance between samples will be:
   - Calculate the mean of each sample i.e., $\bar{X}_1$, $\bar{X}_2$, etc.
   - Calculate the grand average $\bar{\bar{X}}$, pronounced " X double bar". Its value obtained as follows:

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \cdots}{N_1 + N_2 + N_3 + \cdots}$$

   - Take the difference between the means of the various samples and the grand average
   - Square these deviations and obtain the total which will give sum of squares between the samples
   - Divide the total obtained in previous step by the degree of freedom. The degree of freedom will be one less than the number of samples, i.e., if there are 4 samples than the degree of freedom will be 4 – 1 = 3 or v = k – 1, where k = number of samples.
3. **Calculate Variance within the Samples:** The variance (or sum of squares) within samples measures those inter-sample differences due to chance only. It is denoted by SSE. The variance within samples (groups) measures variability around the mean of each group. For calculating variance within the samples, we take the total of the sum of squares of the deviation of various items from the mean values of the respective samples and divide this total by the degree of freedom. Thus steps involved in calculating variance within the samples will be:
   - Calculate the mean of each sample i.e., $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, etc.
   - Take the deviations of various items in a sample from the mean values of the respective samples;
   - Square these deviations and obtain the total which will give sum of squares within the samples;

- Divide the total obtained in previous step by the degree of freedom. The degree of freedom is obtained by deduction from the total number of items, the number of samples, i.e., v = N – K, where K refers to the number of samples and N refers to total number of all the observations;

4. Calculate the Ratio F as Follows:

$$\text{F} - \text{Statistic} = \frac{\text{Variance between the samples}}{\text{Variance within the samples}}$$

Symbolically,

$$F = \frac{S_1^2}{S_2^2}$$

The F ratio measures the ratio of the variance between groups to the variance within groups. The variance between the sample means is the numerator and the variance within the sample means is the denominator. If there is no real difference from group to group, any sample difference will be explainable by random variation and the variance between groups should be close to the variance within groups. However, if there is a real difference between the groups, the variance between groups will be significantly larger than the variance within groups.

5. Compare the F value with Table value: After calculation of F value, it is compared with the table value of F for the degrees of freedom at a certain significance (Generally at 5%) level. If calculated value of F is greater than the table value, the difference in sample means is significant. In other words, the samples do not come from the same population. If the calculated value of F is less than the table value, the difference is not significant and variation has arisen due to fluctuations of simple sampling.

The following ANOVA table summarize calculations for sums of squares, together with the r numbers of degrees of freedom and mean squares.

Source of variation	Sum of Squares	Degree of Freedom (v)	Mean Square	F
Between Samples	SSC	v = c - 1	MSC = SSC/(c – 1)	F = MSC/MSE
Within Samples	SSE	v = n - c	MSE = SSE/(n – c)	
Total	SST	n - 1		

SST = Total sum of squares of variations

SSC = Sum of squares between samples (columns)

SSE = Sum of squares within samples (rows)

MSC = Mean sum of squares between samples

MSE = Mean sum of squares within samples

The same procedure for analysis of variance is applicable for both the equal and unequal sample sizes. The following example will illustrate the procedure:

Example 8.6

To assess the significance of possible variation in performance in a certain test between the convert schools of a city, a common test was given to a number of students taken at random from the senior fifth class of each of the four schools concerned. The results are given below. Make an analysis of variance of data.

Schools			
A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Solution:

Computation of Grand Mean

Sample 1 X_1	Sample 2 X_2	Sample 3 X_3	Sample 4 X_4
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15
Total = 45	50	60	65
$\bar{X} = 9$	10	12	13

$$Grand\ Mean\ or\ \bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{N}$$

Where, $\bar{X}_1$, $\bar{X}_2$, etc., represents the mean of each sample and N the number of samples.

$$Grand\ Mean\ or\ \bar{\bar{X}} = \frac{9 + 10 + 12 + 13}{4} = \frac{44}{4} = 11$$

Variation between samples

To calculate variation between samples, calculate the square of the deviation of the various samples from the grand mean or average. The mean of sample 1 is 9 but the grand mean is 11. So, we will take the difference between 9 and 11 and square it. Similarly, other three samples means difference with grand mean calculated and squared. Thus we have the following table:

Computation of variation between samples

Sample 1 $(\bar{X}_1 - \bar{\bar{X}})^2$	Sample 2 $(\bar{X}_2 - \bar{\bar{X}})^2$	Sample 3 $(\bar{X}_3 - \bar{\bar{X}})^2$	Sample 4 $(\bar{X}_4 - \bar{\bar{X}})^2$
4	1	1	4
4	1	1	4
4	1	1	4
4	1	1	4
4	1	1	4
20	**5**	**5**	**20**

Sum of the squares between the samples = 20 + 5 + 5 + 20 = 50

Mean sum of squares between the samples is 50/(4 -1) = 16.7 (because there are four samples and the degrees of freedom are 4 – 1 = 3).

Variance within the Samples

Here, we find the sum of the squares is the deviations of various items in a sample from the mean values of respective samples. Thus, for the first sample, then mean is 9 and so we will take deviations from 10 and so on. The squared deviations are given in the following table:

Computation of Variance within the sample

Sample 1		Sample 2		Sample 3		Sample 4	
X_1	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)^2$	X_3	$(X_3 - \bar{X}_3)^2$	X_4	$(X_4 - \bar{X}_4)^2$
8	1	12	4	18	36	13	0
10	1	11	1	12	0	9	16
12	9	9	1	16	16	12	1
8	1	14	16	6	66	16	9
7	4	4	36	8	16	15	4
	$\sum(X_1 - \bar{X}_1)^2 = 16$		$\sum(X_2 - \bar{X}_2)^2 = 58$		$\sum(X_3 - \bar{X}_3)^2 = 104$		$\sum(X_4 - \bar{X}_4)^2 = 30$

Total sum of squares within the samples = 16 + 58 + 104 + 30 = 208

Mean sum of squares within the samples = 208/20 – 4 = 208/16 = 13

It is advisable to check up the calculations by finding out total variation. Total variation is calculated by taking the squares of the deviation of each item from the general or grand mean or average.

Computation of Total Variation

Sample 1		Sample 2		Sample 3		Sample 4	
X_1	$(X_1 - \bar{\bar{X}})^2$	X_2	$(X_2 - \bar{\bar{X}})^2$	X_3	$(X_3 - \bar{\bar{X}})^2$	X_4	$(X_4 - \bar{\bar{X}})^2$
8	9	12	1	18	49	13	4
10	1	11	0	12	1	9	4
12	1	9	4	16	25	12	1
8	9	14	9	6	25	16	25
7	16	4	49	8	9	15	16
	$\sum(X_1 - \bar{\bar{X}})^2 = 36$		$\sum(X_2 - \bar{\bar{X}})^2 = 63$		$\sum(X_3 - \bar{\bar{X}})^2 = 109$		$\sum(X_4 - \bar{\bar{X}})^2 = 50$

Total sum of squares = 36 + 36 + 109 + 50 = 258

Degree of freedom = 20 – 1= 19

When we add the sum of squares between samples and sum of squares within samples, we get the same total, i.e., 50 + 208 = 258. Hence, our calculations are right.

All the above results can be tabulated as follows:

Summary of above Results

Source of variation	Sum of Squares	Degree of Freedom	Mean Square
Between Samples	50	3	16.7
Within Samples	208	16	13.0
Total	258	19	

$$\text{F Value} = \frac{\text{Variance between the samples}}{\text{Variance within the samples}} = \frac{16.7}{13} = 1.285$$

The table value of F for $v_1 = 3$ and $v_2 = 16$ degree of freedom at 5% significance level = 3.24. The calculated value of F is less than the table value and hence the difference in the mean values of the sample is not significant, i.e., the samples could have come from the same universe.

Short cut Method

The above method of calculating variance between the samples and variance within the samples are cumbersome or difficult. Generally, this method is not in practice because it is time consuming process. An easier method known as short cut method is usually followed which reduces considerably the computational work. The computations by the short cut method shall be as follows:

Computation by Short cut Method

Sample 1		Sample 2		Sample 3		Sample 4	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
8	64	12	144	18	324	13	169
10	100	11	121	12	144	9	81
12	144	9	81	16	256	12	144
8	64	14	196	6	36	16	256
7	49	4	16	8	64	15	225
$\sum X_1$ = 45	$\sum X_1^2$ = 421	$\sum X_2$ = 50	$\sum X_2^2$ = 558	$\sum X_3$ = 60	$\sum X_3^2$ = 824	$\sum X_4$ = 65	$\sum X_4^2$ = 875

The sum of all the items of various samples = $\sum X_1 + \sum X_2 + \sum X_3 + \sum X_4$

= 45 + 50 + 60 + 65 = 220

Correction factor = $T^2 / N = (220)^2 / 20 = 48400/20 = 2420$.

Total sum of squares = $\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2$ – Correction factor

= 421 + 558 + 824 + 875 – 2420

= 2678 – 2420 = 258 (as above)

Sum of squares between the samples

$$= \frac{(\sum X_1)^2}{N} + \frac{(\sum X_2)^2}{N} + \frac{(\sum X_3)^2}{N} + \frac{(\sum X_4)^2}{N} - \frac{T^2}{N}$$

$$= \frac{(45)^2}{5} + \frac{(50)^2}{5} + \frac{(60)^2}{5} + \frac{(65)^2}{5} - 2420$$

$$= \frac{2025}{5} + \frac{2500}{5} + \frac{3600}{5} + \frac{4225}{5} - 2420$$

$$= \frac{12350}{5} - 2420 = 2470 - 2420 = 50 \ (as\ before)$$

Sum of squares within the samples

= Total sum of squares – sum of squares between samples

= 258 – 50 = 208 (as before)

Coding of Data: It refers to the addition, multiplication, subtraction or division of data by a constant. While making analysis of variance, it should be noted that the final quantity tested is a ratio. This means that the original measurement can be coded to simplify calculations without the need for any subsequent adjustments of the results. The previous example 10.6 would illustrate the point as follows:

Let us take 10 as common for the question take in example 10.6. the coded data are given below and the calculations are done therefrom:

Coded Data

A (X_1)	B (X_2)	C (X_3)	D (X_4)
-2	+2	+8	+3
0	+1	+2	-1
+2	-1	+6	+2
-2	+4	-4	+6
-3	-6	-2	+5
$\sum X_1 = 45$	$\sum X_2 = 0$	$\sum X_3 = 10$	$\sum X_4 = 15$
$\bar{X} = 1$	0	2	3

$$Grand\ Mean\ or\ \bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{N} = \frac{-1 + 0 + 2 + 3}{4} = 1$$

Sum of Squares between samples:

Computation of variation between samples

$(\bar{X}_1 - \bar{\bar{X}})^2$	$(\bar{X}_2 - \bar{\bar{X}})^2$	$(\bar{X}_3 - \bar{\bar{X}})^2$	$(\bar{X}_4 - \bar{\bar{X}})^2$
4	1	1	4
4	1	1	4
4	1	1	4
4	1	1	4
4	1	1	4
$\sum(\bar{X}_1 - \bar{\bar{X}})^2 = 20$	$\sum(\bar{X}_2 - \bar{\bar{X}})^2 = 5$	$\sum(\bar{X}_3 - \bar{\bar{X}})^2 = 5$	$\sum(\bar{X}_4 - \bar{\bar{X}})^2 = 20$

Sum of squares between samples = 20 + 5 + 5 + 20 = 50.

Mean squares between samples = 50/ (4 – 1) = 50/3 = 16.7 (as before)

Sum of Square within samples

Computation of Variance within the sample

Sample 1		Sample 2		Sample 3		Sample 4	
X_1	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)^2$	X_3	$(X_3 - \bar{X}_3)^2$	X_4	$(X_4 - \bar{X}_4)^2$
-2	1	+2	4	+8	36	+3	0
0	1	+1	1	+2	0	-1	16
+2	9	-1	1	+6	16	+2	1
-2	1	+4	16	-4	66	+6	9
-3	4	-6	36	-2	16	+5	4
	$\sum(X_1 - \bar{X}_1)^2 = 16$		$\sum(X_2 - \bar{X}_2)^2 = 58$		$\sum(X_3 - \bar{X}_3)^2 = 104$		$\sum(X_4 - \bar{X}_4)^2 = 30$

Total sum of squares within the samples = 16 + 58 + 104 + 30 = 208

Mean squares within the samples = 208/20-4 = 208/16 = 13 (as before)

Two Way Analysis of Variance (ANOVA)

In a manifold classification, we consider two or more characteristics or attributes. When it is believed that two independent factors might have an effect on the response variable of interest, it is possible to design the test so that an analysis of variance can be used to test for the effects of the two factors simultaneously. Such a test is called a two factor analysis of variance. With two factor analysis of variance, we can test two sets of hypothesis with the same data at the same time.

In a two way classification the data are classified according to two different criteria or factors. The procedure for analysis of variance is somewhat different than the one followed while dealing with problems of one way classification. In a two way classification the ANOVA table takes the following form:

Source of variation	Sum of Squares	Degree of Freedom	Mean Square	F Ratio
Between Samples	SSC	c - 1	MSC = SSC/(c – 1)	MSC/MSE
Between Rows	SSR	r - 1	MSR = SSR/(r – 1)	MSR/MSE
Residual or error	SSE	(c – 1) (r – 1)	MSE = SSE/(r – 1) (c – 1)	
Total	SST	n - 1		

SSC = Sum of squares between columns

SSR = Sum of squares between rows

SSE = Sum of squares due to error

SST = Total sum of squares

The sum of squares for the source 'Residual' is obtained by subtracting from the total sum of squares, the sum of squares between columns and rows, i.e., SSE = SST – (SSC + SSR).

The total number of degree of freedom = n – 1 or cr – 1, where c refers to the number of columns and r refers to number of rows. So,

Number of degree of freedom between columns = (c – 1)

Number of degree of freedom between rows = (r – 1)

Number of degree of freedom for residual = (c – 1) (r – 1)

The total sum of squares, sum of squares for between columns and sum of squares for between rows are obtained in the same way as before.

Residual or error sum of square = Total sum of squares – Sum of squares between columns – Sum of squares between rows.

The F value are calculated as follows:

$$F\ value = \frac{MSC}{MSE}$$

In which v_1 = (c – 1) and v_2 = (c – 1) (r – 1)

$$F\ value = \frac{MSR}{MSE}$$

Where, v_1 = (r – 1) and v_2 = (c – 1) (r – 1)

It should be carefully noted that v_1 may not be same in both cases - in one case v_1 = (c – 1) and another case v_2 = (r – 1).

The calculated values of F are compared with the table values. If calculated value of F is greater than the table value at pre-assigned level of significance, the null hypothesis is rejected, otherwise accepted. It would be clear from above that in problem involving two way classification, Residual is the measuring rod for testing significance. It represents the magnitude of variation due to forces called 'chance'. The following examples would illustrate the procedure:

Example 10.7

Perform a two-way ANOVA on the data given below:

Plots of Land	Treatment			
	A	B	C	D
I	38	40	41	39
II	45	42	49	36
III	40	38	42	42

(Use coding method subtracting 40 from the given numbers)

***Solution*:** Let us take the hypothesis that there is no significant difference in the treatment and plots of land. Applying analysis of variance technique.

On subtracting 40 from each value, we get

Plots of Land	Treatment				Total
	A	B	C	D	
I	-2	0	+1	-1	-2
II	+5	+2	+9	-4	+12
III	0	-2	+2	+2	+2
Total	+3	0	+12	-3	+12

$$Correction\ factor = \frac{T^2}{N} = \frac{(12)^2}{12} = \frac{144}{12} = 12$$

Sum of squares between treatments:

$$= \frac{(3)^2}{3} + \frac{(0)^2}{3} + \frac{(12)^2}{3} + \frac{(-3)^2}{3} - \frac{T^2}{N}$$

$$= \frac{9}{3} + \frac{0}{3} + \frac{144}{3} + \frac{9}{3} - 12$$

$$= 3 + 0 + 48 + 3 - 12 = 42$$

Degree of freedom = 4 – 1 = 3

Sum of squares between plot of land:

$$= \frac{(-2)^2}{3} + \frac{(12)^2}{3} + \frac{(2)^2}{3} - \frac{T^2}{N}$$

$$= \frac{4}{4} + \frac{144}{4} + \frac{4}{4} - 12$$

$$= 1 + 36 + 1 - 12 = 26$$

Degree of freedom = (3 – 1) = 2

Total sum of squares

$$= (-2)^2 + (5)^2 + (0)^2 + (0)^2 + (2)^2 + (-2)^2 + (1)^2 + (9)^2 + (2)^2 + (-1)^2 + (-4)^2 + (2)^2 - \frac{T^2}{N}$$

$= 4 + 25 + 0 + 0 + 4 + 4 + 1 + 81 + 4 + 1 + 16 + 4 - 12 = 132$

The following ANOVA table summarize the all calculation related to two way classification analysis of variance table:

Source of variation	Sum of Squares	Degree of Freedom	Mean Square	F Ratio
Between Samples	42	3	MSC = 14	MSC/MSE = 14/10.67 = 1.312
Between Rows	26	2	MSR = 13	MSR/MSE = 13/10.67 = 1.218
Residual or error	64	6	MSE = 10.67	
Total	132	11		

For (3, 6) d. f. $F_{(0.05)} = 4.76$ and for (2, 6) d. f. $F_{(0.05)} = 5.14$. The calculated values of F are less than the table value at 5% level of significance. The hypothesis is accepted. Hence, there is no significant difference between treatments and plots of land.

8.4 THE *t*-DISTRIBUTION AND Z-DISTRIBUTION

t-Distribution

Let us assume a normal population with mean μ and variance σ^2. If x_i represent the *n* values of a sample drawn from this population. Then

$$Z_i = \frac{x_i - \mu}{\sigma} \sim N(0, 10) \text{ where } i = 1, 2, \ldots\ldots n$$

$$\text{And, } U = \sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{\sigma^2} \sim \chi^2 \text{ (n-1 df)} \qquad \text{where } i = 1, 2, \ldots\ldots n$$

A new sample statistic *T* may, then, be defined as

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

$$T = \frac{x_i - \mu}{\sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}}$$

$$T = \frac{x_i - \mu}{S}$$

This statistic - ***the ratio of the standard normal variable Z to the square root of the $\chi 2$ variable divided by its degree of freedom*** - is known as ***'t'*** statistic or ***student 't'*** statistic, named after the pen name of Sir W S Gosset, who discovered the distribution of the quantity.

The random variable $\frac{x_i - \mu}{S}$ follows ***t-distribution*** with *n*-1 degrees of freedom.

$$\frac{x_i - \mu}{S} \sim t \text{ (n-1 df)} \qquad \text{where } i = 1, 2, \ldots\ldots n$$

The *t*-distribution in terms of Sampling Distribution of Sample Mean

We know $\bar{X} \sim N(\mu, \sqrt{\sigma^2/n}^2)$

So $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$

Putting $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ *for* $\frac{x_i - \mu}{\sigma}$ in $T = \frac{\frac{x_i - \mu}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{\sigma^2}}}$, we get

$$T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{\sigma^2}}}$$

or $$T = \frac{\left(\overline{X} - \mu\right)/\sigma}{\frac{1}{\sigma}\sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n(n-1)}}}$$

or $$T = \frac{\overline{X} - \mu}{\sqrt{\frac{1}{n}\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}}$$

or $$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

When defined as above, T again follows ***t-distribution*** with *n*-1 degrees of freedom.

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t \text{ (n-1 df)} \qquad \text{where } i = 1, 2, \ldots\ldots n$$

Properties of *t*- Distribution

1. The *t*-distribution like *Z* distribution, is unimodal, symmetric about mean 0, and the *t*- variable varies from - $\propto$ and $\propto$
2. The t-distribution is defined by the degrees of freedom v = n-1, the df associated with the distribution are the *df* associated with the sample standard deviation.

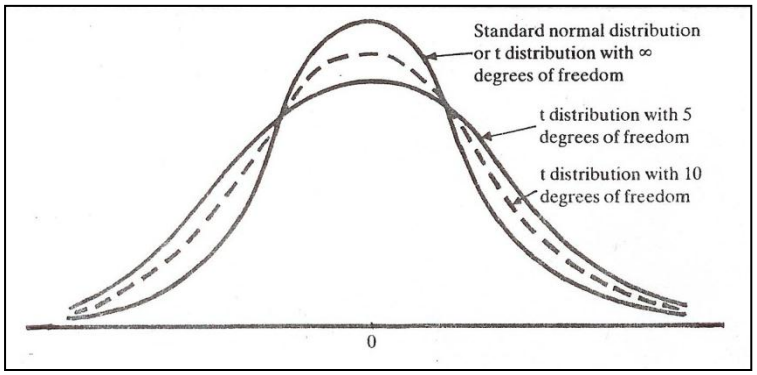


Figure 10-8 *t*-Distribution with different *df*

3. The *t*-distribution has no mean for $n = 2$ *i.e.* for $v = 1$ and no variance for $n \leq 3$ *i.e.* for $v \leq 2$. However, for $v > 1$, the mean and for $v > 2$, the variance is given as $E(T) = 0$; $Var(T) = \frac{v}{v-2}$.
4. The variance $\frac{v}{v-2}$ of the *t*-distribution must always be greater than 1, so it is more variable as against *Z* distribution which has variance 1. This follows from the fact what while *Z* values vary from sample to sample owing to the change in the $\overline{X}$ alone, the variation in *T* values are due to changes in both $\overline{X}$ and *S*.
5. The variance of *t*-distribution approaches 1 as the sample size *n* tends to increase. In general, for $n \geq 30$, the variance of *t*-distribution is approximately the same as that of *Z* distribution. In other words the *t*-distribution is approximately normal for $n \geq 30$.

8.5 CHECK YOUR PROGRESS

1. The variance of *t*-distribution approaches 1 as the sample size *n* tends to……..
2. As a ratio of two squared quantities, the F random variable cannot be negative and is, therefore,……..on the left by zero.
3. The F distributions defined as $F_{(v_1,v_2)}$ and as $F_{(v_2,v_1)}$ are……..of each other.
4. The t-distribution has no mean for n = 2 i.e. for v = 1 and no…….for $n \leq 3$ i.e. for $v \leq 2$.
5. When sampling without replacement from a finite population, the probability distribution of the second random variable depends on what has been the…… of the first pick and so on.

8.6 SUMMARY

Population parameter is any number computed (or estimated) for the entire population *viz.* population mean, population median, population proportion, population variance and so on. Population parameter is unknown but fixed, whose value is to be estimated from the sample statistic that is known but random. Sample Statistic is any numbers computed from our sample data viz. sample mean, sample median, sample proportion, sample variance and so on. The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take when computed from random samples of the same size drawn from a specified population. The sampling distributions of only the commonly used sample statistics like sample mean, sample proportion, sample variance *etc.*, which

have a role in making inferences about the population. The F distribution is the distribution of the ratio of two chi-square random variables that are independent of each other, each of which is divided by its own degrees of freedom. The ratio of the standard normal variable Z to the square root of the $\chi 2$ variable divided by its degree of freedom - is known as‘t’ statistic or student ‘t’ statistic, named after the pen name of Sir W S Gosset, who discovered the distribution of the quantity.

8.7 KEYWORDS

Sampling Distributions: The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take when computed from random samples of the same size drawn from a specified population.

T statistics: The ratio of the standard normal variable Z to the square root of the $\chi 2$ variable divided by its degree of freedom - is known as‘t’ statistic or student ‘t’.

F distribution: It is the distribution of the ratio of two chi-square random variables that are independent of each other, each of which is divided by its own degrees of freedom.

Central Limit Theorem: When sampling is done from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean $\overline{X}$ tends to a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as the sample size n increases.

Sampling distribution of difference of sample mean: The sampling distribution of $\overline{X_1} - \overline{X_2}$ is the probability distribution of all possible values the random variable $\overline{X_1} - \overline{X_2}$ may take when independent samples of size n_1 and n_2 are taken from two specified populations.

Degree of freedom: It refers to the number of independent variables which vary freely without being influenced by the restrictions imposed by the sample statistic(s) to be computed.

8.8 SELF-ASSESSMENT TEST

1. What is a sampling distribution, and what are the uses of sampling distributions?
2. How does the size of population and the kind of random sampling determine the shape of the sampling distributions?
3. (a) A sample of size *n = 5 is* selected from a population. Under what conditions is the sampling distribution of $\overline{X}$ normal?

(b) Suppose the population mean is μ = 105 and the population standard deviation is 20. What are the expected value and the standard deviation *of* $\overline{X}$ *?*

4. What is the most significant aspect of the central limit theorem? Discuss the practical utility of central limit theorem in applied statistics.
5. Under what conditions is the central limit theorem most useful in sampling for making statistical inferences about the population mean?
6. If the population mean is 1,247, the population variance is 10,000, and the sample size is 100, what is the probability that $\overline{X}$ will be less than 1,230?
7. When sampling is from a population with standard deviation σ = 55, using a sample of size n = 150, what is the probability that $\overline{X}$ will be at least 8 units away from the population mean μ?
8. The Colosseum, once the most popular monument in Rome, dates from about AD 70. Since then, earthquakes have caused considerable damage to the huge structure, and engineers are currently trying to make sure the building will survive future shocks. The Colosseum can be divided into several thousand small sections. Suppose that the average section can withstand a quake measuring 3.4 on the Richter scale with a standard deviation of 1.5. A random sample of 100 sections is selected and tested for the maximum earthquake force they can withstand. What is the probability that the average section in the sample can withstand an earthquake measuring at least 3.6 on the Richter scale?
9. On June 10, 1997, the average price per share on the Big Board Composite Index in New York rose 15 cents. Assume the population standard deviation that day was 5 cents. If a random sample of 50 stocks is selected that day, what is the probability that the average price change in this sample was a rise between 14 and 16 cents?
10. An economist wishes to estimate the average family income in a certain population. The population standard deviation is known to be Rs 4,000, and the economist uses a random sample of size n = 225. What is the probability that the sample mean will fall within Rs 750 of the population mean?
11. When sampling is done from a population with population proportion p = 0.2, using a sample size n = 15, what is the sampling distribution of $\overline{p}$ *?* Is it reasonable to use a normal approximation for this sampling distribution? Explain.

12. When sampling is done for the proportion of defective items in a large shipment, where the population proportion is 0.18 and the sample size is 200, what is the probability that the sample proportion will be at least 0.20?
13. A study of the investment industry claims that 55% of all mutual funds outperformed the stock market as a whole last year. An analyst wants to test this claim and obtains a random sample of 280 mutual funds. The analyst finds that only 108 of the funds outperformed the market during the year. Determine the probability that another random sample would lead to a sample proportion as low as or lower than the one obtained by the analyst, assuming the proportion of all mutual funds that outperformed the market is indeed 0.55.
14. In recent years, convertible sport coupes have become very popular in Japan. Toyota is currently shipping Celicas to Los Angeles, where a customizer does a roof lift and ships them back to Japan. Suppose that 25% of all Japanese in a given income and lifestyle category are interested in buying Celica convertibles. A random sample of 100 Japanese consumers in the category of interest is to be selected. What is the probability that at least 20% of those in the sample will express an interest in a Celica convertible?
15. What are the limitations of small samples?
16. What do you understand by small sampling distributions? Why are the small sampling distributions called exact distributions?
17. What do you understand by the concept of degrees of freedom?
18. Define the χ^2 statistic. What are important properties of χ^2 distribution?
19. Define the F statistic. What are important properties of F distribution?
20. Define the t statistic. What are important properties of t-distribution? How does t statistic differ from Z statistic?

8.9 ANSWERS TO CHECK YOUR PROGRESS

1. Increase
2. Bounded
3. Reciprocal
4. Variance
5. Outcome

8.10 REFERENCES/SUGGESTED READINGS

17. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
18. Statistics for Management *by* G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
19. Business Statistics *by* Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
20. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
21. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
22. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
23. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
24. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.

Subject: Business Statistics	
Course code: MBA-102	Author: Dr. B. S. Bodla
Lesson: 09	Vetter: Karam Pal
STATISTICAL ESTIMATION	

STRUCTURE

9.1 Learning Objectives

9.2 Introduction

- 9.2.1 Types of Estimates
- 9.2.2 Criteria of a good estimator
- 9.2.3 Method of Maximum Likelihood
- 9.2.4 Point Estimation
- 9.2.5 Interval Estimation

9.3 Sample size Determination

9.4 Check your progress

9.5 Summary

9.6 Keywords

9.7 Self-Assessment Test

9.8 Answers to check your progress

9.9 References/Suggested Reading

9.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understand the term statistical estimation and types of estimates;
- Construct and interpret confidence interval estimates to know the precision of the estimate of a population mean and proportion; and
- Determine the sample size to ensure that the margin of error will be -within acceptable limits.

9.2 INTRODUCTION

The sampling process is used to draw statistical inference about the characteristics of a population or process of interest. On many occasions we do not have enough information to calculate an exact value of population parameters (such as μ, σ and p) and therefore make the best estimate of this value from the corresponding sample statistics (such as $\bar{x}$, s, and P). The need to use the sample statistic to draw conclusions about the population characteristic is one of the fundamental applications of statistical inference in business and economics. A few applications of statistical estimation are given below:

- A production manager needs to determine the proportion of items being manufactured that do not match with quality standards.
- A mobile phone service company may be interested to know the average length of a long distance telephone call and its standard deviation.
- A bank needs to understand consumer awareness of its services and credit schemes.
- Any service centre needs to determine the average amount of time a customer spends in queue.

In all such cases, a decision-maker needs to examine the following two concepts that are useful for drawing statistical inference about an unknown population or process parameters based upon random samples:

(i) Estimation– a sample statistic to estimate an unknown parameter value

(ii) Hypothesis testing– a claim or belief about an unknown parameter value.

In this lesson we shall discuss methods to estimate unknown population parameters and then to determine the range of values (confidence interval) likely to contain the parameter value.

9.2.1 TYPES OF ESTIMATES

Let us first know the concept of 'estimate' as used in Statistics. According to some dictionaries, an estimate is a valuation based on opinion or roughly made from imperfect or incomplete data. This definition may apply, for example, when an individual who has an opinion about the competence of one of his colleagues. But, in Statistics the term estimate is not used in this sense. In Statistics too the estimates are made when the information available is incomplete or imperfect. However, such estimates are made only when they are based on sound judgement or experience and when the samples are scientifically selected. There are two types of estimates that we can make about a population: *a point estimate* and an *interval estimate.*

A point estimate is a single number, which is used to estimate an unknown population parameter. Although a point estimate may be the most common way of expressing an estimate, it suffers from a major limitation since it fails to indicate how close it is to the quantity it is supposed to estimate. In other words, a point estimate does not give any idea about the reliability of precision of the method of estimation used. For instance, if someone claims that 40 percent of all children in a certain town do not go to the school and are devoid of education, it would not be very helpful if this claim is based on a small number of households, say, 20. However, as the number of households interviewed for this purpose increases from 20 to 100, 500 or even 5,000, the claim that 40 percent of children have no school education would become more and more meaningful and reliable. This makes it clear that a point estimate should always be accompanied by some relevant information so that it is possible to judge how far it is reliable.

The second type of estimate is known as the interval estimate. It is a range of values used to estimate an unknown population parameter. In case of an interval estimate, the error is indicated in two ways: first by the extent of its range; and second, by the probability of the true population parameter lying within that range. Taking our previous example of 40 percent children not having a school education, the statistician may say that actual percentage of such children in that town may lie between 35 percent and 45 percent. Thus, he will have a better idea of the reliability of such an estimate as compared to the point estimate of 40 percent.

Estimator and Estimate

When we make an estimate of a population parameter, we use a sample statistic. This sample statistic is an estimator.

For example, the samples mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

$\bar{x}$ is a point estimator of the population mean μ. The value obtained by the estimator is known as an estimate. Many different Statistics can be used to estimate the same parameter. For example, we may use the sample mean or the sample median or even the range to estimate the population mean. The question here is: how can we evaluate the properties of these estimates, compare then with one another,

and finally, decide which the 'best' is? The answer to this question is possible only when we have certain criteria that a good estimator must satisfy. These criteria are briefly discussed below.

9.2.2 CRITERIA OF A GOOD ESTIMATOR

There are four criteria by which we can evaluate the quality of a statistic as an estimator. These are: unbiasedness, efficiency, consistency and sufficiency.

Unbiasedness

This is a very important property that an estimator should possess. If we take all possible samples of the same size from a population and calculate their means, the mean $\mu_{\bar{x}}$ of all these means will be equal to the mean μ of the population. This means that the sample mean $\bar{x}$ is an unbiased estimator of the population mean μ. When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, the sample statistic is said to be an unbiased estimator.

Suppose we take the smallest sample observation as an estimator of the population mean μ, it can be easily shown that this estimator is biased. Since the smallest observation must be less than the mean, its expected value must be less than μ. Symbolically, $E(Xs) < \mu$, where Xs stands for the smallest item and E stands for the expected value. Thus, this estimator is biased downwards. The extent of bias is the difference between the expected value of the estimator and the value of the parameter. In this case, bias is equal to $E(Xs)- \mu$. In contrast, the biases for the sample mean $\bar{x}$ is zero.

Consistency

Another important characteristic that an estimator should possess is consistency. Let us take the case of the standard deviation of the sampling distribution of $\bar{x}$. The standard deviation of the sampling distribution of sample mean is computed by following formula :

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

The formula states that the standard deviation of the sampling distribution of $\bar{x}$ decreases as the sample size increases and *vice versa*. When the sample size *n* increases, the population standard deviation σ is to be divided by a higher denominator. This results in the reduced value of sample standard deviation σ_{ξ}. Let us take an example.

Illustration 9.1: **A company has 4,000 employees whose average monthly wage comes to Rs.4,800 with a standard deviation of Rs.1,200. Let $\bar{x}$ be the mean monthly wage for a random sample of certain employees selected from this company. Find the mean and standard deviation of ξ for a sample size of (a) 81, (b) 100 and (c) 180.**

Solution

From the given information, for the population of all employees, N = 4,000 μ = Rs.4,800 σ = Rs.1,200.

(a) The mean μ_ξ of the sampling distribution of the ξ is $\mu_\xi = \mu$ = Rs.4,800.

As n = 81 and N = 4,000, which gives n/N = 0.01. At this value is less than 0.05, the standard deviation of ξ is obtained by using the formula. Substituting the values.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ or, } \sigma_{\bar{x}} = \frac{1{,}200}{\sqrt{81}} = \frac{1{,}200}{\sqrt{9}} = \text{Rs.113.33}$$

(b) In this case, n = 100 and n/N = 100/4,000 = 0.025, which is also less than 0.05. The mean and the standard deviation ξ are

$\mu_{\bar{\xi}} = \mu$ = Rs.4,800

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ or, } \sigma_{\bar{x}} = \frac{1{,}200}{\sqrt{100}} = \frac{1{,}200}{10} = \text{Rs.120}$$

(c) In this case, n = 180 and n/N = 180/4,000 = 0.045, which again is less than 0.05. The mean and the standard deviation ξ are

$\mu\bar{x} = \mu$ = Rs.4,800

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ or, } \sigma_{\bar{x}} = \frac{1{,}200}{\sqrt{180}} = \frac{1{,}200}{13.42} = \text{Rs.89.42}$$

From the above three sets of calculation, it becomes clear that the mean of the sampling distribution of $\bar{x}$ is always equal to the mean of the population regardless of the sample size. But, in case of the standard deviation, we find the change. In the given example, we find that standard deviation of $\bar{x}$ decreased from Rs.189.87 to Rs.120 and then to Rs.113.33 as the sample size increased from 40 to 100 and then to 180.

Efficiency

Another desirable property of a good estimator is that it should be efficient. Efficiency is measured in terms of size of the standard error of the statistic. Since an estimator is a random variable, it is necessarily characterised by a certain amount of variability. This means that some estimates may be more variable than others. Just as bias is related to the expected value of the estimator, so efficiency can be defined in terms of the variance. In large samples, for example, the variance of the sample mean is $V(\bar{x})=\sigma^2/n$. As the sample size *n* increases, the variance of the sample mean ($V\bar{x}$) becomes smaller, so the estimator becomes more efficient. This criterion, when applied to large samples, gives better estimates as compared to the small ones.

The efficiency of one estimator in relation to another estimator can be judged by comparing their sampling variances. Thus, efficiency relates to the size of the standard error. Given the same sample size, the statistic that has a smaller standard error is preferable as it is efficient in relation to another statistic that has a larger standard error. The sampling distribution of the mean and the median have the same mean, that is, the population mean. However, the variance of the sampling distribution of the means is smaller than the variance of the sampling distribution of the medians. As such, the sample mean is an efficient estimator of the population mean, while the sample median is an inefficient estimator.

Sufficiency

The fourth property of a good estimator is that it should be sufficient. A sufficient statistic utilises all the information a sample contains about the parameter to be estimated. ξ, for example, is a sufficient estimator of the population mean μ. It implies that no other estimator of μ, such as the sample median, can provide any additional information about the parameter μ. Likewise, we can say that the sample proportion π. Having looked into properties of a good estimator briefly, a pertinent question arises: how can we find estimators with these desirable properties? This brings us to the method of maximum likelihood.

9.2.3 METHOD OF MAXIMUM LIKELIHOOD (ML)

The maximum likelihood method provides estimators with the desirable properties such as efficiency, consistency and sufficiency, which we have just discussed. It usually does not give an unbiased estimate. Let us take an example to explain this method.

Example: *Suppose we want to estimate the average grade of a large number of students. A random sample of size* n = *64 is taken and the sample mean* $\bar{x}$ *is found to be 90 marks. Now, the assumption on which we have to base our reasoning is that the random sample of* n = *64 is representative of the population. We saw how samples that were similar to the population had greater probability of being selected.*

Let us now reverse this reasoning as follows: we have before us a random sample size *n* = 64 and $\bar{x}$ = 90 marks. From which population did it most probably come-a population with μ = 85, 90 or 95? According to our earlier approach, we would think that it most probably came from a population with μ=90 marks. Thus, it can be concluded that the population mean μ, based on our sample, is most likely to be μ=90 marks.

A point worth noting is that the population mean μ is either 90 or not; it has only one value. Hence, we have used the term *likely* instead of probably.

This technique to find the estimators was first used and developed by Sir R.A. Fisher in 1922, who called it the maximum likelihood method.

9.2.4 POINT ESTIMATION

In point estimation, a single sample statistic (such as $\bar{x}$, *s,* and $\bar{p}$) is calculated from the sample to provide a best estimate of the true value of the corresponding population parameter (such as μ, σ and $\bar{p}$). Such a single relevant statistic is termed as *point estimator*, and the value of the statistic is termed as *point estimate.* For example, we may calculate that 10 per cent of the items in a random sample taken from a day's production are defective. The result '10 per cent' is a point estimate of the percentage of items in the whole lot that are defective. Thus, until the next sample of items is not drawn and examined, we may proceed on manufacturing on the assumption that any day's production contains 10 per cent defective items.

9.2.5 INTERVAL ESTIMATION

Generally, a point estimate does not provide information about 'how close is the estimate' to the population parameter unless accompanied by a statement of possible sampling errors involved based on the sampling distribution of the statistic. It is therefore important to know the precision of an estimate before relying on it to make a decision. Thus, decision-makers prefer to use an *interval estimate* that is likely to contain the population parameter value. However, it is also important to state 'how confident' he is that the interval estimate actually contains the parameter value. Hence an interval estimate of a population parameter is therefore a *confidence interval* with a statement of confidence that the interval contains the parameter value.

The confidence interval estimate of a population parameter is obtained by applying the formula:

Point estimate $\pm$ Margin of error

Where Margin of error = $z_c \times$ Standard error of a particular statistic

z_c = critical value of standard normal variable that represents confidence level (probability of being correct) such as 0.90,0.95, and so on.

9.2.5.1 Interval estimation of population mean (σknown)

Suppose the population mean μ is unknown and the true population standard deviation σ is known. Then for a large sample size ($n \geq 30$), the interval estimation of population mean μ is given by

$$\bar{x} \pm z_{\alpha/2}\, \sigma_{\xi} \text{ or, } \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or $$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $z_{a/2}$ is the z-value representing an area a/2 in the right and left tails of the standard normal probability distribution.

In general, a 95 per cent confidence interval estimate implies that if all possible samples of the same size were drawn, then 95 per cent of them would include the true population mean somewhere within the interval around their sample mean and only 5 per cent of them would not. The values for $z_{\alpha/2}$ for the most commonly-used

as well as the other confidence levels can be seen from standard normal probability table as shown in Table 11.1.

Table 9.1 : Values of Standard Normal Probability $z_{\alpha/2}$

Confidence Level, (1-α) (%)	***Acceptable Error Level, α***	***α/2***	$z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

Illustration 9.2: The average monthly electricity consumption for a sample of 100 families is 1250 units. Assuming the standard deviation of electric consumption of all families is 150 units, construct a 95 per cent confidence interval estimate of the actual mean electric consumption.

Solution: The information given is: $\bar{x}$ =12.50, σ=150, n= 100 and confidence level (1-α) = 95 per cent. Using the 'Standard Normal Curve' we find that the half of 0.95 yields a confidence coefficient $z_{\alpha/2} = 1.96$. Thus confidence limits with $_{\alpha/2} = \pm 1.96$ for 95 per cent confidence are given by

$$\xi \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 12.50 \pm 1.96 \frac{150}{\sqrt{100}} = 1250 \pm 29.40 \text{ units}$$

Thus for 95 per cent level of confidence, the population mean μ is likely to fall between 1220.60 units and 1279.40 units, that is, $1220.60 \leq \mu \leq 1279.40$.

Illustrator 9.3: The quality control manager at a factory manufacturing light bulbs is interested to estimate the average life of a large shipment of light bulbs. The standard deviation is known to be 100 hours. A random sample of 50 light bulbs gave a sample average life of 350 hours.

(a) Setup a 95 per cent confidence interval estimate of the true average life of light bulbs in the shipment.

(b) Does the population of light bulb life have to be normally distributed? Explain.

Solution: The following information is given :

$\bar{x}$ = 350, σ = 100, n =50, and confidence level, (1-α) = 95 per cent.

(a) Using the 'Standard Normal Curve', we have $z_{\alpha/2} = \pm 1.96$ for 95 per cent confidence level. Thus confidence limits are given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 350 \pm 1.96 \frac{100}{\sqrt{50}} = 350 \pm 27.72$$

Hence for 95 per cent level of confidence the population mean μ is likely to fall between 322.28 hours to 377.72 hours, that is, $322.28 \leq \mu \leq 377.72$.

(b) No, since σ is known and n = 50, from the central limit theorem we may assume that $\bar{x}$ is normally distributed.

9.2.5.2 Interval Estimation for Difference of Two Means

If all possible samples of large size n_1 and n_2 are drawn from two different populations, then sampling distribution of the difference between two means ξ_1 and ξ_2 is approximately normal with mean $(\mu_1-\mu_2)$ and standard deviation:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For a desired confidence level, the confidence interval limits for the population mean $(\mu_1-\mu_2)$ are given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\, \sigma_{\bar{x}_1-\bar{x}_2}$$

Illustration 9.4: The strength of the wire produced by company A has a mean of 4,500 kg and a standard deviation of 200 kg. Company B has a mean of 4000 kg and a standard deviation of 300 kg. A sample of 50 wires of company A and 100 wires of company B are selected at random for testing the strength. Find 99 per cent confidence limits on the difference in the average strength of the populations of wires produced by the two companies.

Solution: The following information is given:

Company A: $\bar{x}_1 = 4500$, $\sigma_1 = 200$, $n_1 = 50$

Company B: $\bar{x}_2 = 4000$, $\sigma_2 = 300$, $n_2 = 100$

Therefore $\mu_{\bar{x}_1-\bar{x}_2} = 4500 - 4000 = 500$ and $z_{\alpha/2} = 2.576$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} = \sqrt{\frac{40{,}000}{50}+\frac{90{,}000}{100}} = 41.23$$

The required 99 per cent confidence interval limits are given by

$(\bar{x}_1 - \bar{x}_2) \pm z_{a/2}\ \sigma_{\bar{x}_1-\bar{x}_2} = 500 \pm 2.576\ (41.23) = 500 \pm 106.20$

Hence, the 99 percent confidence limits on the difference in the average strength of wires produced by the two companies are likely to fall in the interval $393.80 \le \mu \le 606.20$.

9.2.5.3 Interval estimation of population mean (σknown)

In practice, the standard deviation of a population σ, is not likely to be known. Thus in the large sample case, the sample standard deviation σ, and we use a *z*-table to compute $z_{a/2}$ for providing an area of a/2 in the right tail of the standard normal probability distribution curve. Hence the interval estimate of a population mean for a large sample case (n > 30) with confidence coefficient 1-α is given by

$$\bar{x} \pm z_{\alpha/2}\, s_{\bar{x}} = \bar{x} \pm z_{\alpha/2}\, \frac{s}{\sqrt{n}}$$

When the population standard deviation is not known and the sample size is small, the procedure of interval estimation of population mean is based on a probability distribution known as the *t-distribution*. This distribution is very similar to the normal distribution. However, the *t-distribution* has more area in the tails and less in the center than doe's normal distribution. The t-distribution depends on a parameter known as *degree of freedom*. As the number of degrees of freedom increases, t-distribution gradually approaches the normal distribution, and the sample standard deviation *s* becomes a better estimate of population standard deviation σ.

The interval estimate of a population mean when the sample size is small (n ≤30) with confidence coefficient (1-α), is given by

$$\bar{x} \pm t_{\alpha/2}\, \frac{s}{\sqrt{n}} \text{ or } \bar{x} - t_{\alpha/2}\, \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\, \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the critical value of t-test statistic providing an area α/2 in the right tail of the *t*-distribution with *n*-1 degrees of freedom, and

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

The critical values of *t* for the given degrees of freedom can be obtained from the table of *t*-distribution (see appendix).

The procedure of the confidence interval estimation of population mean μ when population standard deviation is unknown and sample size is large or small, is summarised in Table 9.2.

Table 9.2: Confidence Interval for μ

Sample size	*Interval Estimate of Population Mean μ*
Large • σ assumed known	$\bar{x} \pm z_{\alpha/2} \frac{\sigma_1}{\sqrt{n}}$
• σ estimated by *s*	$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$
Small • σ assumed known	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
• σ estimated by *s*	$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$

Illustration 9.5: A random sample of 64 sales invoices was taken from a large population of sales invoices. The average value was found to be Rs.2000 with a standard deviation of Rs.540. Find a 90 per cent confidence interval for the true mean value of all the sales.

Solution: The information given is: $\bar{x}_1 = 2000, s = 540, n = 64,$ and $\alpha = 10$ per cent. Therefore

$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{540}{\sqrt{64}} = 67.50$ and $z_{\alpha/2}$ = 1.64 (from Normal table)

The required confidence interval of population mean μ is given by

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 2000 + 1.64\ (67.50) = 2000 \pm 110.70$$

Thus the mean of the sales invoices for the whole population is likely to fall between Rs.1889.30 and Rs.2110.70, that is, $1889.30 \leq \mu \leq 2110.70$.

Illustration 9.6: The personnel department of an organization would like to estimate the family dental expenses of its employees to determine the feasibility of providing a dental insurance plan. A random sample of 10 employees reveals the following family dental expenses (in thousand Rs.) in the previous year: 11, 37, 25, 62, 51, 21, 18, 43, 32, 20.

Setup a 99 per cent confidence interval of the average family dental expenses for the employees of this organization.

Solution: The calculations for sample mean $\bar{x}$ and standard deviation are shown in Table 9.3.

Table 9.3 : Calculations for $\bar{x}$ and *s*

Variable, x	$(x-\bar{x})=(x-32)$	$(x-\bar{x})^2$
11	–21	441
37	05	25
25	–07	49
62	30	900
51	19	361
21	–11	121
18	–14	196
43	11	121
32	–	–
20	–12	144
320	0	2358

From the data in Table 11.3, the sample mean $\bar{x} = \frac{\sum x}{n} = \frac{320}{10}$ = Rs.32, and the sample standard deviation $s = \sqrt{\sum(x-\bar{x})^2 / n-1} = \sqrt{\frac{2358}{9}}$ = Rs.5.11. Using this information and $t_{\alpha/2} = 1.833$ at $df = 9$, we have

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 32 \pm 1.833 \frac{5.11}{\sqrt{10}} = 32 \pm 2.962$$

Hence the mean expenses per family are likely to fall between Rs.29.038 and Rs.34.962, that is, 29.038 ≤ μ ≤ 34.962.

9.2.5.4 INTERVAL ESTIMATION FOR POPULATION PROPORTION

(iv) You know that normal distribution as an approximation of the sampling distribution of sample proportion $\bar{p} = x/n$ is based on the large sample conditions: $np \geq 5$ and $nq = n(1-p) \geq 5$, where p is the population proportion as shown in Fig.11.3. The confidence interval estimate for a population proportion at 1-α confidence coefficient is given by

$$\bar{p} \pm z_{\alpha/2}\sigma_{\bar{p}} = \bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \text{ or } \bar{p} - z_{\alpha/2}\sigma_{\bar{p}} \leq p \leq \bar{p} + z_{\alpha/2}\sigma_{\bar{p}}$$

(v) where $z_{\alpha/2}$ is the z-value providing an area of α/2 in the right tail of the standard normal probability distribution and the quantity $z_{\alpha/2}\sigma_{\bar{p}}$ is the margin of error.

(vi) **Illustration 9.7:** Suppose we want to estimate the proportion of families in a town, which have two or more children. A random sample of 144 families shows that 48 families have two or more children. Setup a 95 per cent confidence interval estimate of the population proportion of families having two or more children.

(vii) **Solution:** The sample proportion is : $\bar{p} = \frac{x}{n} = \frac{48}{144} = \frac{1}{3}$

(viii) Using the information, n = 144, $\bar{p} = \frac{1}{3}$ and $z_{\alpha/2}$ = 1.96 at 95 per cent confidence coefficient, we have

(ix) $$\bar{p} \pm z\alpha/2\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \frac{1}{3} \pm 1.96\sqrt{\frac{(\frac{1}{3})(\frac{2}{3})}{144}} = 0.333 \pm 0.077$$

Hence the population proportion of families who have two or more children is likely to be between 25.6 to 41 per cent, that is, $0.256 \leq p \leq 0.410$.

9.3 SAMPLE SIZE DETERMINATION

(x) From previous sections we understand that standard error $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and $\sigma_{\bar{p}} = \sigma/\sqrt{pq/n}$ of sampling distribution of sample statistic $\bar{x}$ and $\bar{p}$ are both inversely proportional to the sample size *n,* which is also related to the width of the confidence intervals $\bar{x} \pm z_{a/2}\sigma_{\bar{x}}$ and $\bar{p} \pm z_{a/2}\sigma_{\bar{p}}$. Obviously, the width or range of the confidence interval can be decreased by increasing the sample size *n.* The decision regarding the appropriate size of the sample, however, depends on (i) deciding in advance how good an estimate is required, and (ii) the availability of funds, time, and ease of sample selection. For example, an insurance company wants to estimate the proportion of claims settled within 2 months of the receipt of claim. For this purpose, the company must decide how much error it is willing to allow in estimating the population proportion of claims settled in a particular financial year. This means, whether accuracy is required to be within ± 80 claims, ± 100 claims, and so on. Also, the company needs to determine in advance the level of confidence for estimating the true population parameter. Hence for determining the sample size for estimating population mean or proportion, such requirements must be kept in mind along with information regarding standard deviation.

9.3.1 Sample Size for Estimating Population Mean

When the distribution of sample mean $\bar{x}$ is normal, the standard normal variable *z* is given as

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \text{ or } \bar{x} - \mu = \frac{\sigma}{\sqrt{n}}$$

The value of *z* can be seen from 'standard normal table' for a specified confidence coefficient 1-α. The value of *z* in the above equation will be positive or negative, depending or whether the sample mean $\bar{x}$ is larger or smaller than, population mean μ as shown in Fig.11.4. This difference between $\bar{x}$ and mean

μ is called the *sampling error* or *margin of error* E. Thus for estimating the population mean μ with a condition that the error in its estimation should not exceed a fixed value, say E, we require that the sample mean $\bar{x}$ should fall within the range, μ ± E with a specified probability. Thus the margin of error *acceptable* (i.e. maximum tolerable difference between unknown population mean μ and the sample estimate at a particular level of confidence) can be written as:

$$\bar{x} - \mu = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ or } E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or $$\sqrt{n} = \frac{z_{\alpha/2}\sigma}{E}, \textbf{ i.e., } n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

This formula for sampling size *n* will provide the tolerable margin of error E, at the chosen confidence level 1-α (which determines the critical value of *z* from the normal table) with known or estimated population standard deviation σ.

Note: If population standard deviation σ is not known, then sample standard deviation *s* can be used to determine the sample size *n*.

Illustration 9.8: Suppose the sample standard deviation of P/E ratios for stocks listed on the Mumbai Stock Exchange (BSE) is *s* = 7.8. Assume that we are interested in estimating the population mean of P/E ratio for all stocks listed on BSE with 95 per cent confidence. How many stocks should be included in the sample if we desire a margin of error of 2?

Solution: The information given is: E=2, *s* =7.8, $z_{\alpha/2}$ = 1.96 at 95 per cent level of confidence.

Using the formula for *n* and substituting the given values, we have

$$n = \frac{(za/2)^2 \sigma 2}{E^2} = \frac{(1.96)^2 (7.8)^2}{(2)^2} = \frac{3.84 \times 60.84}{4} = 59 \text{ approx.}$$

Thus a sample size *n* = 59 should be chosen to estimate the population mean of P/E ratio for all stocks on the BSE.

Note: The general rule used in determining sample size is to always round off to the nearest integer value in order to slightly over-satisfy the desire of estimation.

9.3.2 Sample Size for Estimating Population Proportion

The method for determining a sample size for estimating the population proportion is similar to that used in the previous section. We require that the sample proportion $\bar{p}$ should fall within the range $\bar{p} \pm \mathrm{E}$, with a specified probability

$$\mathrm{E} = z_{\alpha/2}\sigma_{\bar{p}} = z_{\alpha/2}\sqrt{\frac{pq}{n}}; q = 1 - p$$

Or $$\mathrm{E} = (z_{\alpha/2})^2 \frac{pq}{n}, i.e., n = \frac{(z_{\alpha/2})^2 pq}{\mathrm{E}^2}$$

The value of *z* can be calculated from 'Standard normal table' for a specified confidence coefficient. This formula for *n* will provide the desired margin of error E at the chosen confidence level 1-α (which determines the critical value of *z*) with known or estimated population proportion *p*.

(xi) **Illustration 9.9:** A car manufacturing company received a shipment of petrol filters. These filters are to be sampled to estimate the proportion that is unusable. From past experience, the proportion of unusable filter is estimated to be 10 per cent. How large a random sample should be taken to estimate the true proportion of unusable filters to within 0.07 with 99 per cent confidence.

(xii) Solution: The information given is : E=0.07, p = 0.10, and $z_{\alpha/2}$ = 2.576 at 99 per cent confidence level.

(xiii) Using the formula for *n* and substituting the given values, we have

$$n = \frac{(z_{\alpha/2})^2 pq}{\mathrm{E}^2} = \frac{(2.576)^2 (0.10 \times 0.90)}{(0.07)^2} = 121.88$$

Therefore a slightly larger sample size of *n* = 122 filters should be taken

9.3.3 Sample Size Determination for Finite Population

When samples are drawn without replacement from a finite population of size N, the use of finite population correction factor reduces the standard error by a value equal to $\sqrt{(N-n)/(N-1)}$. For

example, for deciding sample size *n* for estimating the population mean μ, the desired margin of error is given by

$$E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-l}}$$

(xiv) Similarly, when estimating the proportion, the desired margin of error is given by

$$\sigma_{\bar{p}} \text{ or } E = z_{\alpha/2}\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$$

Let n_0 be the size for estimating population mean without using correction factor. Then

$$n_0 = \frac{(z_{\alpha/2})^2\sigma^2}{E^2}$$

(xv) The revised sample size, taking into consideration the size of the population, is given by

$$n = \frac{n_0 N}{n_0 + (N-1)}$$

(xvi) **Illustration 9.10:** For a population of 1000, what should be the sampling size necessary to estimate the population mean at 95 per cent confidence with a sampling error of 5 and the standard deviation equal to 20?

Solution: We have E =5, σ=20, $z_{\alpha/2}$ = 1.96 at 95 per cent confidence level, and N=1000. Thus

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.96)^2(20)^2}{(5)^2} = 61.456$$

(xvii) Since the population size is finite, the revised sample size obtained by using the correction factor

$$n = \frac{n_0 N}{n_0 + (N-1)} = \frac{(61.456)(1000)}{61.456 + (1000-1)} = \frac{61456}{1060.456} = 57.952$$

Thus a sample size of *n* = 58 should be taken

9.4 CHECK YOUR PROGRESS

1. If we take all possible samples of the same size from a population and calculate their means, the mean $\mu_{\bar{x}}$ of all these means will be..... to the mean μ of the population.
2. Standard deviation of the sampling distribution of $\bar{x}$ decreases as the sample size ..... and *vice versa*.
3. The efficiency of one estimator in relation to another estimator can be judged by ..... their sampling variances.
4. As the number of degrees of freedom increases, t-distribution gradually approaches the ....., and the sample standard deviation s becomes a better estimate of population standard deviation.
5. For estimating the population mean with a condition that the error in its estimation should not ..... a fixed value, say E, we require that the sample mean $\bar{x}$ should fall within the range, ± E with a specified probability.

9.5 SUMMARY

There are two types of estimates that we can make about a population: *a point estimate* and an *interval estimate.* A point estimate is a single number, which is used to estimate an unknown population parameter. Although a point estimate may be the most common way of expressing an estimate, it suffers from a major limitation since it fails to indicate how close it is to the quantity it is supposed to estimate. The second type of estimate is known as the interval estimate. It is a range of values used to estimate an unknown population parameter. In case of an interval estimate, the error is indicated in two ways: first by the extent of its range; and second, by the probability of the true population parameter lying within that range.

There are four criteria by which we can evaluate the quality of a statistic as an estimator. These are: unbiasedness, efficiency, consistency and sufficiency.

9.6 KEYWORDS

Estimate: An estimate is a valuation based on opinion or roughly made from imperfect or incomplete data.

Point Estimate: A point estimate is a single number, which is used to estimate an unknown population parameter.

Interval Estimate: It is a range of values used to estimate an unknown population parameter.

T-Distribution: When the population standard deviation is not known and the sample size is small, the procedure of interval estimation of population mean is based on a probability distribution known as the *t-distribution*. This distribution is very similar to the normal distribution.

Sampling Error: The difference between $\bar{x}$ and mean μ is called the *sampling error* or *margin of error* E.

9.7 SELF- ASSESSMENT TEST

1. Distinguish between the point estimation and interval estimation. Explain how an interval estimate is better than a point estimate.
2. Explain the concept of 'margin of error' in deciding the size of a sample.
3. Prove that the mean of a simple random sample from a given population is an unbiased estimator of the population mean.
4. Under what circumstances can the normal distribution be used to construct a confidence interval estimate of the population mean?
5. What are the properties of a good estimator? Explain, how these properties are essential for estimating the population characteristic of interest.
6. Distinguish between statistic and parameter and explain the meaning of confidence interval of a population parameter.
7. Explain the following terms with an example
   (a) Point estimate (b) Interval estimate
   (c) Confidence interval (d) Confidence limits
8. Describe the effect of sample size on the margin of sampling error of point estimate of the proportion mean. Doest this error depends on the sample size in the same way?
9. In an effort to estimate the mean amount spend per customer for diner at a city hotel, data were collected for a sample of 49 customers. Assume a population standard deviation of Rs.25.
   At 95 per cent confidence, what is the margin of error? If the sample mean is Rs.124, what is the 95 per cent confidence interval for the population mean?
10. The following data have been collected for a sample from a normal population: 5, 10, 8, 11, 12, 6, 15, 11

What is the point estimate of population mean and standard deviation?

What is the confidence interval for population mean at 95 per cent confidence interval?

11. A machine is producing ball bearing with a diameter of 0.5 inches. It is known that the standard deviation of the ball bearings is 0.005 inch. A sample of 100 ball bearings is selected and their average diameter is found to be 0.48 inch. Determine the 99 per cent confidence interval.
12. Suppose a wholesaler of paints wants to estimate the actual amount of paint contained in 10 kg cans purchased from a paint manufacturing company. It is known from the manufacture's specifications that the standard deviation of the amount of paint is equal to 0.02 kg. A random sample of 50 cans is selected, and the average amount of paint per 10 kg can is 0.995 kg. Setup a 99 per cent confidence interval estimate of the true population average amount of paint included in a 10 kg can. Based on your results, do you think that the wholesaler has a right to complaint to the manufacture? Why?
13. A survey of 672 audited tax returns showed that 448 resulted in additional payments. Construct a 95 per cent confidence interval for the true percentage of all audited tax returns that resulted in additional payments.
14. In a survey carried out in a large city, 170 households out of a random sample of 250 owned at least one pet. Find the 95 per cent confidence interval for the percentage of households in the city who own at least one pet. Does the result support a pet food manufacturer's claim that 75 per cent of all households have at least one pet?
15. A cigarette manufacturer wishes to use random sampling to estimate the average nicotine content. The sampling error should not be more than one milligram above or below the true mean with a 99 per cent confidence coefficient. The population standard deviation is 4 milligrams. What sample size should the company use in order to satisfy these requirements?
16. An agency responsible for electricity distribution would like to estimate the average electric bills for a particular month for single-family homes in a large city. Based on studies conducted in other cities, the standard deviation is assumed to be Rs.40. The agency would like to estimate the average bill for that month to within Rs.10 of the true average. If 95 per cent confidence is desired, then what sample size is necessary?

9.8 ANSWERS TO CHECK YOUR PROGRESS

1. Equal

2. Increases
3. Comparing
4. Normal distribution
5. Exceed

9.9 REFERENCES/SUGGESTED READINGS

1. Spiegel, Murray R.: Theory and Practical of Statistics., London McGraw Hill Book Company.
2. Yamane, T.: Statistics: An Introductory Analysis, New York, Harpered Row Publication.
3. R.P. Hooda: Statistic for Economic and Management McMillan India Ltd.
4. G.C. Beri: Statistics for Mgt., TMA
5. J.K. Sharma: Business Statistics, Pearson Education
6. S.P. Gupta : Statistical Methods, Sultan Chand and Sons.

Subject: Business Statistics	
Course code: MBA-102	Author: Anil Kumar
Lesson: 10	Vetter: Dr. Karam Pal
TESTING OF HYPOTHESES	

STRUCTURE

10.1 Learning Objectives

10.2 Introduction

10.2.1 The Null and the Alternative Hypothesis

10.2.2 Some Basic Concepts

10.2.3 Critical Region in Terms of Test Statistic

10.2.4 General Testing Procedure

10.2.5 Tests of Hypotheses about Population Means

10.2.6 Tests of Hypotheses about Population Proportions

10.2.7 Tests of Hypotheses about Population Variances

10.2.8 The Comparison of Two Populations

10.3 Solved Problems

10.4 Check your Progress

10.5 Summary

10.6 Keywords

10.7 Self-Assessment Test

10.8 Answers to check your progress

10.9 References/Suggested Readings

10.1 LEARNING OBJECTIVES

After going through this lesson, the students will be able to:

- Understand the concept of hypotheses testing
- Specify the most appropriate test of hypothesis in a given situation
- Apply the procedure and make inferences from the results.

10.2 INTRODUCTION

Closely related to Statistical Estimation discussed in the preceding lesson, Testing of Hypotheses is one of the most important aspects of the theory of decision-making. In the present lesson, we will study a class of problems where the decision made by a decision maker depends primarily on the strength of the evidence thrown up by a random sample drawn from a population. We can elaborate this by an example where the operations manager of a cola company has to decide whether the bottling operation is under statistical control or it has gone out of control (and needs some corrective action). Imagine that the company sells cola in bottles labeled *1-liter*, filled by an automatic bottling machine. The implied claim that on the average each bottle contains 1,000 cm^3 of cola may or may not be true.

- If the claim is true, the process is said to be under statistical control. It is in the interest of the company to continue the bottling process
- If the claim is not true *i.e.* the average is either more than or less than 1,000 cm^3, the process is said to be gone out of control. It is in the interest of the company to halt the bottling process and set right the error

Therefore, to decide about the status of the bottling operation, the operations manager needs a tool, which allows him to test such a claim.

Testing of Hypotheses provides such a tool to the decision maker. If the operations manager were to use this tool, he would collect a sample of filled bottles from the on-going bottling process. The sample of bottles will be evaluated and based on the strength of the evidence produced by the sample; the operations manager will accept or reject the implied claim and accordingly make the decision. The implied claim (μ = 1,000 cm^3) is a hypothesis that needs to be tested and the statistical procedure, which allows us to perform such a test, is called *Hypothesis Testing* or *Testing of Hypotheses.*

What is a Hypothesis?

A thesis is some thing that has been proven to be true. A hypothesis is something that has not yet been proven to be true. It is some statement about a population parameter or about a population distribution. Our hypothesis for the example of bottling process could be:

"The average amount of cola in the bottles is equal to 1,000 cm^3"

This statement is tentative as it implies some assumption, which may or may not be found valid on verification. Hypothesis testing is the process of determining whether or not a given hypothesis is true.

If the population is large, there is no way of analyzing the population or of testing the hypothesis directly. Instead, the hypothesis is tested on the basis of the outcome of a random sample.

10.2.1 THE NULL AND THE ALTERNATIVE HYPOTHESIS

As stated earlier, a hypothesis is a statement about a population parameter or about a population distribution. In any testing of hypotheses problem, we are faced with a pair of hypotheses such that one and only one of them is always true. One of this pair is called the null hypothesis and the other one the alternative hypothesis.

A null hypothesis *is an assertion about the value of a population parameter. It is an assertion that we hold as true unless we have sufficient statistical evidence to conclude otherwise.*

For example, a null hypothesis might assert that the population mean is equal to 1,000. Unless we obtain sufficient evidence that it is not 1,000, we will accept it as 1,000. We write the null hypothesis compactly as:

$$H_0: \mu = 1{,}000$$

Where the symbol H_0 denotes the null hypothesis.

*The **alternative hypothesis** is the negation of the null hypothesis.*

For the null hypothesis $H_0: \mu = 1{,}000$, the alternative hypothesis is $\mu \neq 1000$. We will write it as

$$H_1: \mu \neq 1{,}000$$

We use the symbol H_1 *(or* H_a*)* to denote the alternative hypothesis.

The null and alternative hypotheses assert exactly opposite statements. Obviously, both H_0 and H_1 cannot be true and one of them will always be true. Thus, rejecting one is equivalent to accepting the other. At the end of our testing procedure, if we come to the conclusion that H_0 should be rejected, this also amounts to saying that H_1 should be accepted and vice versa.

It is not difficult to identify the pair of hypotheses relevant in any decision situation. Can any one of the two be called the null hypothesis? The answer is a big NO — because the roles of H_0 and H_1 are not symmetrical.

The possible outcomes of a test can be summarized as:

Either: **Accept** $\boldsymbol{H_0}$ -a weak conclusion without any evidence in as a reasonable possibility support of H_0

or: **Reject** $\boldsymbol{H_0}$ and -a strong conclusion with strong evidence **Accept** $\boldsymbol{H_1}$ against H_0

To better understand the role of null and alternative hypotheses, we can compare the process of hypothesis testing with the process by which an accused person is judged to be innocent or guilty. The

person before the bar is assumed to be *"innocent until proven guilty"* So using the language of hypothesis testing, we have:

H_0*: The person is innocent*

H_1*: The person is guilty*

The outcomes of the trial process may result

- Accepting H_0 of innocence: when there was not enough evidence to convict. However, it does not prove that the person is truly innocent
- Rejecting H_0 and accepting H_1 of guilt: when there is enough evidence to rule out innocence as a possibility and to strongly establish guilt

The jury acquitted Michael Jackson, on June 13, of all charges against him in the child molestation case. In other words, using the language of hypothesis testing the jury had to accept the null hypothesis H_0*: Michael Jackson is innocent* because the prosecution could not prove their case against H_0 of innocence.

In a trial case we do not have to rule out guilt in order to find someone innocent, but we do have to rule out innocence in order to find someone guilty. On the similar lines, we do not have to rule out H_1 in order to accept H_0; but we do have to rule out H_0 in order to accept H_1. Thus, it is clear that the two hypotheses - null and alternative - are not interchangeable; each one plays a different, a special role. So it becomes more important to be clear about what the null and alternative hypotheses should be in a given situation, or else the test is meaningless.

One can conceptualize the whole procedure of testing of hypothesis as trying to answer one basic question: ***Is the sample evidence strong enough to enable us to reject H_0?*** This means that H_0 will be rejected only when there is strong sample evidence against it. However, if the sample evidence is not strong enough, we shall conclude that we cannot reject H_0 and so we accept H_0 by default. Thus, H_0 is accepted even without any evidence in support of it whereas it can be rejected only when there is overwhelming evidence against it. In other words, the decision maker is somewhat biased towards the null hypothesis and he does not mind accepting the null hypothesis. However, he would reject the null hypothesis only when the sample evidence against the null hypothesis is too strong to be ignored.

The null hypothesis is called by this name because in many situations, acceptance of this hypothesis would lead to null action. Thus, one way to ensure what the null hypothesis should be is to note that…

…if the null hypothesis is true, then no corrective action would be necessary. If the alternative hypothesis is true, then some corrective action would be necessary.

Recall our example of the cola-company in which an automatic bottling machine fills *1-liter* bottles with cola. Now consider three different situations:

Situation I: The operations manager wants to test the average amount filled, in order to know whether the process is under statistical control.

In this situation, the operations manager will have to take corrective action when the average is either more than or less than 1,000 cm^3. Only when the average equals 1,000 cm^3, no corrective action is necessary. So we have

$$H_0: \quad \mu = 1{,}000\ cm^3$$

$$H_1: \quad \mu \neq 1{,}000\ cm^3$$

Situation II: A consumer advocate suspects that the average amount of cola is less than 1,000 cm^3 and wants to test it.

In this situation, if the average amount of cola is greater than or equal to 1,000 cm^3, no corrective action is needed, but if the average amount is less than 1,000 cm^3, the company has to halt the bottling process and set right the error. So, in this case, we have

$$H_0: \quad \mu \geq 1{,}000\ cm^3$$

$$H_1: \quad \mu < 1{,}000\ cm^3$$

Situation III: The owner of the company suspects that the machine is wasting cola by filling more than 1,000 cm^3 on the average and wants to test it.

From the owner's point of view, no corrective action is necessary if the average is less than or equal to 1,000 cm^3. And, therefore, in this case we have

$$H_0: \quad \mu \leq 1{,}000\ cm^3$$

$$H_1: \quad \mu > 1{,}000\ cm^3$$

As the bottling example indicates, there are three possible cases for the null hypothesis, involving ≥, ≤ and = relationships. The exact null hypothesis should be finalized before any evidence is gathered, or the test will not be valid. **Data snooping -** formulating the null and alternative hypotheses at one's convenience after collecting and looking at the evidence - is unethical.

10.2.2 SOME BASIC CONCEPTS

We will now discuss some concepts, which are essential for setting up a procedure for testing of hypotheses.

10.2.2.1 TYPE I AND TYPE II ERRORS

After the null and alternative hypotheses are spelled out, the next step is to gather evidence from a random sample of the population. An important limitation of making interferences from the sample data is that ***we cannot be 100% confident about it.*** Since variations from one sample to another can never be eliminated until the sample is as large as the population itself, it is possible that the conclusion drawn is incorrect which leads to an error. As shown in Table 12-1 below, there can be two types of errors.

Table 10-1 Type I and Type II Errors of Hypothesis Testing

Decision based on Sample	**States of Population**	
	H_0 True	**H_0 False**
Accept H_0	Correct decision (No Error)	Wrong Decision (Type II Error)
Reject H_0	Wrong Decision (Type I Error)	Correct Decision (No Error)

Type I Error

In the context of statistical testing, the wrong decision of rejecting a true null hypothesis is know as **Type I Error**. If the operations manager reject H_0 and conclude that the process has gone out of control, when in reality it is under control, he would be making a type I error.

Type II Error

The wrong decision of accepting (not rejecting, to be more accurate) a false null hypothesis is known as **Type II Error**. If the operations manager do not reject H_0 and conclude that the process is under control, when in reality it has gone out of control, he would be making a type II error.

Both the type I and type II errors are undesirable and should be reduced to the minimum. Let us analyse how we can minimize the chances of type I and type II errors. It may be easily realized that it is possible, even with imperfect sample evidence, to reduce the probability of type I error all the way down to zero. Just accept the null hypothesis; no matter what the evidence is. Since we will never reject any null hypothesis, we will never reject a true null hypothesis and thus we will never commit a type I error! However, it is obvious that this would be foolish. If we always accept a null hypothesis, then given a false null hypothesis, no matter how wrong it is, we are sure to accept it. In other words, our probability of committing a type II error will be 1. Similarly, we find it foolish to reduce the probability of type II error all the way down to zero by always rejecting a null hypothesis, for we would then reject every true null hypothesis, no matter how right it is. Our probability of type I error will be 1.

Therefore, we cannot and should not try to completely avoid either type of error. We should plan, organize, and settle for some small, optimal probability of each type of error. Before we discuss this issue, we need to learn a few more concepts.

10.2.2.2 TEST STATISTIC AND THE *p*-VALUE

Consider the case of owner's suspicion related to our bottling process example. The null and alternative hypotheses in this case are:

$$H_0: \quad \mu \leq 1{,}000$$

$$H_1: \quad \mu > 1{,}000$$

Suppose the population variance is 25 and a random sample of size 100 yields a sample mean of 1,000.5. Because the sample mean is more than 1,000, the evidence goes against the null hypothesis (H_0). Can we reject H_0 based on this evidence?

- if we reject it, there is some chance that we might be committing a type I error, and
- if we accept it, there is some chance that we might be committing a type II error.

Then what can we do? We should ask a natural question at this situation- *"What is the probability that H_0 can still be true despite the evidence?"* The question asks for the "credibility" of H_0 in light of unfavorable evidence. However, due to mathematical complexities, it is not possible to compute the probability that H_0 is true. We, therefore, settle for a question that comes very close.

"When the actual μ = 1,000, and with sample size 100, what is the probability of getting a sample mean that is more than or equal to 1000.5?"

The answer to this question is then taken, as the "credibility rating" of H_0. Analyzing the question carefully, we note an important aspect:

The condition assumed is $\mu = 1{,}000$; although H_0 states $\mu \leq 1{,}000$. The reason for assuming $\mu = 1{,}000$ is that ***it gives the most benefit of doubt to H_0.*** If we assume $\mu = 999$, for instance, the probability of the sample mean being more than or equal to 1,000.5 will only be smaller, and H_0 will only have less credibility. Thus the assumption $\mu = 1{,}000$ gives the maximum credibility to H_0.

Now using our knowledge of sampling distribution of sample mean, we can easily answer our question.

Since population variance is known and sample size is large enough, the Central Limit Theorem is applicable here *i. e.*

$$\overline{X} \sim N\left(\mu, \left(\sigma/\sqrt{n}\right)^2\right)$$

and the standard normal variable $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is to be used to calculate the required probability

$P(\overline{X} \geq 1{,}000.5)$

So $P(\overline{X} \geq 1{,}000.5) = P\left(Z \geq \dfrac{1{,}000.5 - 1{,}000}{5/\sqrt{100}}\right)$

$= P(Z \geq 1.00)$

$= 0.1587$

≈ 0.16

So the answer to our question is 16%. That is, there is a 16% chance for a sample of size 100 to yield a sample mean more than or equal to 1000.5 when the actual μ = 1,000. Statisticians call this 16% the ***p*-value.** In other words *p*-value-*the probability of observing a sample statistic as extreme as the one observed if the null hypothesis is true*-.

is a kind of "credibility rating" of H_0 in light of the evidence. A *p*-value of zero means H_0 is certainly false and a *p*-value of 1 means that H_0 is certainly true. A *p*-value of 16% means that there is roughly

16% probability that H_0 is true, despite the evidence. Conversely, we can be roughly 84% confident that H_0 is false in light of the evidence. The implication is that if we reject H_0, then there is about an 84% chance that we are doing the right thing, and about a 16% chance that we are committing a type I error. The formal definition of the *p*-value follows:

*Given a null hypothesis and sample evidence with sample size n, the **p-value** is the probability of getting a sample evidence with the same n that is equally or more unfavorable to the null hypothesis while the null hypothesis is actually true. The p-value is calculated giving the null hypothesis the maximum benefit of doubt.*

The random variable, as *Z* in this case, used to calculate the *p*-value is called test statistic. The formal definition of the test statistic follows:

*A **test statistic** is a random variable calculated from the sample evidence, which follows a well-known distribution and thus can be used to calculate the p-value.*

Most of the time, the test statistic we use will be Z, *t,* χ^2, or *F*. The distributions of these random variables are well known and we can calculate the *p*-value.

Up to this point it is very much clear that statistical hypothesis is always stated with reference to a population parameter (mean, proportion or variance). The appropriate random variable calculated from the sample evidence acts as a test statistic and provide the means to decide whether statistical hypothesis is to be rejected or accepted.

10.2.2.3 THE SIGNIFICANCE LEVEL-α

From our discussion on *p*-value, it becomes clear that the *p*-value of a test *i.e.* the credibility of the null hypothesis varies with actual observed value of the sample statistic. This fact necessitates having *a policy* for rejecting H_0 based on *p*-value.

The most common policy in statistical hypothesis testing is to establish a ***significance level,*** denoted by **α,** and to reject H_0 when the *p*-value falls below it. When this policy is followed, one can be sure that the maximum probability of type I error is α.

***Policy:* When the *p*-value is less than α, reject H_0**

In other words, we can say that the rejection region for H_0 is the area under the curve where the *p*-value is less than α. This region is also called critical region

The standard values for α are 10%, 5%, and 1%. Suppose α is set at 5%. In the preceding example, for a sample mean of 1,000.5 the *p*-value was 16%, and H_0 will not be rejected. For a sample mean of 1001 the *p*-value will be 2.28%, which is below α = 5%. Hence H_0 will be rejected.

Let us analyze in some detail the implications of using a significance level α for rejecting a null hypothesis.

- The first thing to note is that *if we do not reject H_0, this does not prove that H_0 is true.* For example, if α = 5% and the *p*-value = 6%, we will not reject H_0. But there is only about 6% chance that H_0 is true, which is hardly proof that H_0 is true. It may be possible that H_0 is false and by not rejecting it, we are committing a type II error. For this reason, we should say *"We cannot reject H_0 at an α of 5%"* rather than *"We accept H_0."*
- The second thing to note is that α is the maximum probability of type I error we set for ourselves. Since α is the maximum *p*-value at which we reject H_0, it is the maximum probability of committing a type I error. In other words, setting α = 5% means that we are willing to put up with up to 5% chance of committing a type I error.
- The third thing to note is that the selected value of α indirectly determines the probability of type II error as well. In general, *other things remaining the same, increasing the value of α will decrease the probability of type II error.* This should be intuitively obvious. For example, increasing α from 5% to 10% means that in those instances with *p*-value in the range 5% to 10% the H_0 that would not have been rejected before would now be rejected. Thus, some cases of false H_0 that escaped rejection before may not escape now. As a result, the probability of type II error will decrease
- The fourth thing to note about α is the meaning of (1 - α). If we set α = 5%, then (1 - α) = 95% is the minimum ***confidence level*** that we set in order to reject H_0. In other words, we want to be *at least* 95% confident that H_0 is false before we reject it.

One-Tailed and Two-Tailed Tests : Consider the null and alternative hypotheses:

$$H_0: \quad \mu \geq 1{,}000$$

$$H_1: \quad \mu < 1{,}000$$

In this case, we will reject H_0 only when X is significantly less than 1,000 or only when Z falls significantly below zero. Thus the rejection occurs only when Z takes a significantly low value in the *left tail* of its distribution.

Such a case where rejection occurs in the *left tail* of the distribution of the test statistic is called a ***left-tailed*** test, as seen in Figure 10-1.

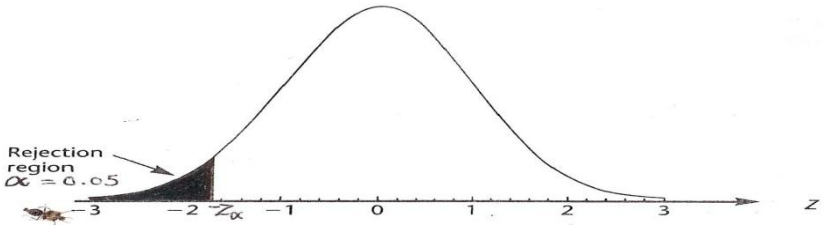


Figure 10-1 A Left-tailed Test

In the case of a left-tailed test, the p-value is the area to the left of the calculated value of the test statistic.

Now consider the case where the null and alternative hypotheses are:

$$H_0: \quad \mu \leq 1{,}000$$

$$H_1: \quad \mu > 1{,}000$$

In this case, we will reject H_0 only when X is significantly more than 1,000 or only when Z is significantly greater than zero. Thus the rejection occurs only when Z takes a significantly high value in the *right tail* of its distribution.

Such a case where rejection occurs in the *right tail* of the distribution of the test statistic is called a ***right-tailed*** test, as seen in Figure 10-2.

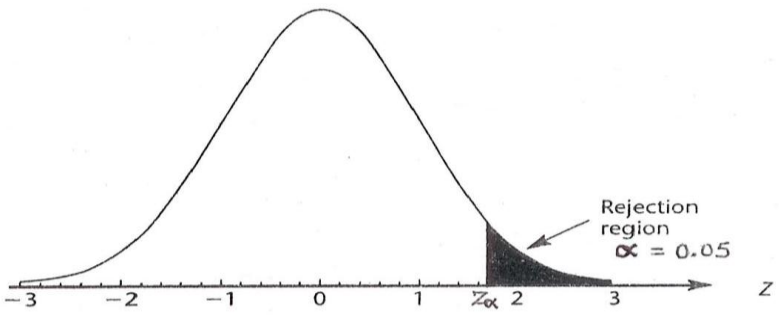


Figure 10-2 A Right-tailed Test

In the case of a right-tailed test, the p-value is the area to the right of the calculated value of the test statistic.

In left-tailed and right-tailed tests, rejection occurs only on one tail. Hence each of them is called a ***one-tailed test.***

Finally, consider the case where the null and alternative hypotheses are:

$$H_0: \quad \mu = 1{,}000$$

$$H_1: \quad \mu \neq 1{,}000$$

In this case, we have to reject H_0 in both cases, that is, whether X is significantly less than or greater than 1,000. Thus, rejection occurs when Z is significantly less than or greater than zero, which is to say that rejection occurs on both tails. Therefore, this case is called a ***two-tailed test.*** See Figure 10-3, where the shaded areas are the rejection regions.

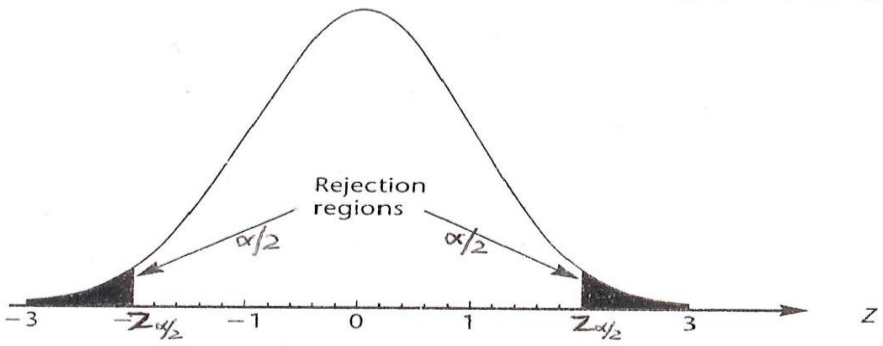


Figure 10-3 A Two-tailed Test

In the case of a two-tailed test, the *p*-value is twice the tail area. If the calculated value of the test statistic falls on the left tail, then we take the area to the left of the calculated value and multiply it by 2. If the calculated value of the test statistic falls on the right tail, then we take the area to the right of the calculated value and multiply it by 2. For example, if the calculated $Z = +1.75$, the area to the right of it is 0.0401. Multiplying that by 2, we get the *p*-value as 0.0802.

Selecting Optimal α

All tests of hypotheses hinge upon this concept of the significance level and it is possible that a null hypothesis can be rejected at $\alpha = 5\%$ whereas the same evidence is not strong enough to reject the null hypothesis at $\alpha = 1\%$. In other words, the inference drawn can be sensitive to the significance level used. We should note that selecting a value for α is a question of compromise between type I and type II error probabilities. In practice, the significance level is supposed to be arrived at after considering the cost consequences of type I error and type II error. However, most of the time the costs are difficult to estimate since they depend, among other things, on the unknown actual value of the parameter being

tested. Thus, arriving at a "calculated" optimal value for α is impractical. Instead, we follow an intuitive approach of assigning one of the three standard values, 1%, 5%, and 10%, to α.

In the intuitive approach, we try to estimate the relative costs of the two types of errors. For example, suppose we are testing the average tensile strength of a large batch of bolts produced by a machine to see if it is above the minimum specified. Here type I error will result in rejecting a good batch of bolts and the cost of the error is roughly equal to the cost of the batch of bolts. Type II error will result in accepting a bad batch of bolts and its cost can be high or low depending on how the bolts are used.

If the bolts are used to hold together a structure, then the cost is high because defective bolts can result in the collapse of the structure, causing great damage. In this case, we should strive to reduce the probability of type II error more than that of type I error. *In such cases where type II error is more costly, we keep a large value for* α*, namely, 10%.*

On the other hand, if the bolts are used to secure the lids on trash cans, then the cost of type II error is not high and we should strive to reduce the probability of type I error more than that of type II error. *In such cases where type I error is more costly, we keep a small value for* α*, namely, 1%.*

Then there are cases where we are not able to determine which type of error is more costly. *If the costs are roughly equal, or if we have not much knowledge about the relative costs of the two types of errors, then we keep* $\alpha = 5\%$.

β and Power of the Test

Denoted by β, Type II error is committed when a wrong decision is taken in accepting a false null hypothesis. It is the probability of accepting H_0 when it should have rejected for being false. It should be noted that β depends on the actual value of the parameter being tested, the sample size, and α. Let us see exactly how it depends.

Consider the null and alternative hypotheses

$$H_0: \quad \mu \leq 1{,}000$$

$$H_1: \quad \mu > 1{,}000$$

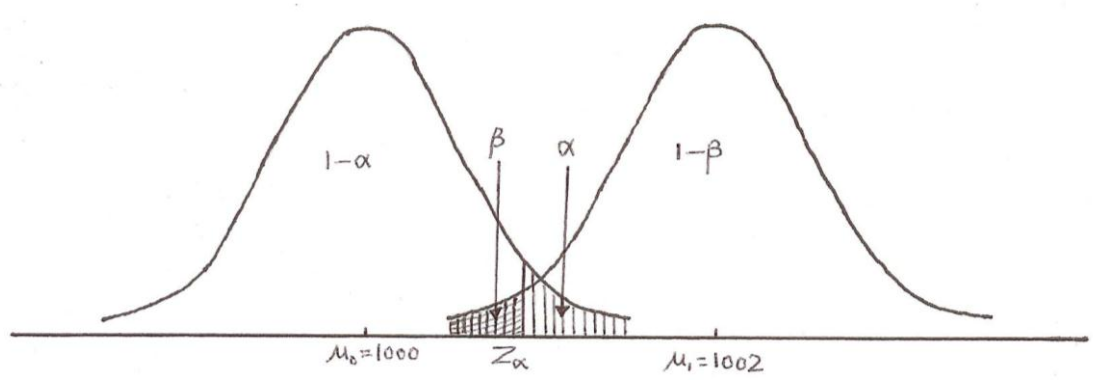


Figure 10-4 Type II Error: H_0: $\mu \leq 1{,}000$ and actual $\mu = 1{,}002$

Suppose the actual value of $\mu = \mu_1$ (say 1,002), such that $\mu_1 > 1{,}000$. Obviously, H_0 is false. The cross-hatched area under the normal curve centered at μ_1 in Figure 12-4 is then the probability of accepting H_0 when it is false. This area - in the acceptance region of the normal curve centered at μ_0 = 1,000; represents the probability that the observed sample mean $\overline{X}$ falls in the acceptance region when $\mu = \mu_1$ (1,002), that is when H_0 is false.

Given the acceptance region (1 - α) for the normal curve centered at $\mu = \mu_0$ = 1,000, a careful analysis of figure reveals the following.

- The value of β decreases as μ_1 move away from μ_0, displaying the entire normal curve centered at μ_1 farther and farther away from the normal curve centered at μ_0.
- The value of β tends to increase as μ_1 moves nearer to μ_0. A limit is reached when μ_1 coincides with μ_0, and the entire acceptance region (1- α) for $\mu = \mu_0$ will represent the value of β. This is important conclusion in the sense that when H_0 is true for $\mu = \mu_0$, the entire acceptance region is Type II error. Hence when H_0 is true, $\beta = 1 - \alpha$ and $\alpha = 1-\beta$.
- The un-shaded area under the normal curve centered at μ_1, which falls outside the acceptance region for $\mu = \mu_0$, represents the probability of rejecting H_0 when it is false for $\mu = \mu_1$. This complement of β; $(1-\beta)$ is known as the *power* of the test. *The **power** of a test is the probability that a false null hypothesis will be detected by the test.*
- A change in the level of significance α means a change in the acceptance region (1- α), which obviously implies a change in the cross hatched area *i.e.* β. In other words, the smaller the α, the

larger the β and vice-versa. Type I and type II errors are, therefore negatively related. Type I error and the power of the test (1-β) are, however, positively related. Thus, the smaller the probability (α) of rejecting H_0 when it is true, the smaller is the probability (1-β) of rejecting H_0 when it is false.

Sample Size

In the discussion above we said that we can keep a α low or aβ low depending on which type of error is more costly. What if both types of error are costly and we want to have low α as well as low β? The only way to do this is to make our evidence more reliable, which can he done only by increasing the sample size. If the sample size increases, then the evidence becomes more reliable and the probability of any error will decrease.

Figure 10-5 shows the relationship between α and β for various values of sample size n. As n increases, the curve shifts downwards reducing both α and β. Thus, when the costs of both types of error are high, the best policy is to have a large sample and a low α, such as 1%.

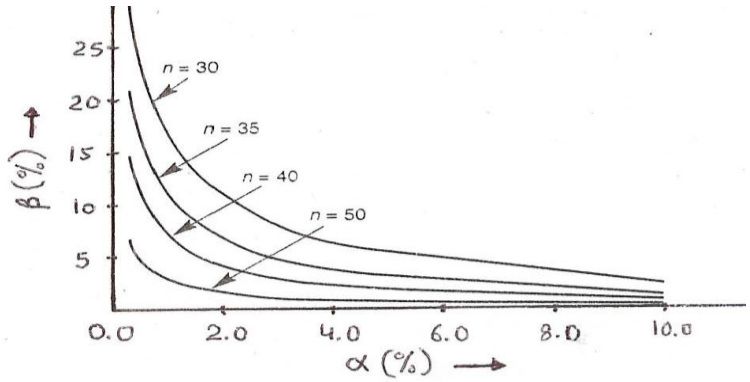


Figure 10-5 β *versus* α for Various Values of *n*

After understanding the basic concepts of testing of hypotheses, we are now, able to develop tests concerning different population parameters. Under different conditions the test procedures have to be developed differently and different test statistics are used for testing. Before proceeding further let us define the critical region in terms of test statistic, which is often more helpful in many situations.

10.2.3 CRITICAL REGION IN TERMS OF TEST STATISTIC

We have seen that the most common policy in statistical hypothesis testing is to establish a ***significance level-*α.** We decide to reject or not to reject the null hypothesis H_0 by comparing the p-value with the significance level. We define the critical or rejection region as:

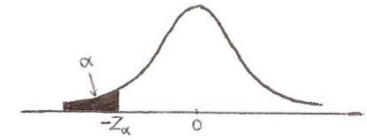
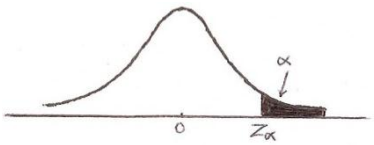
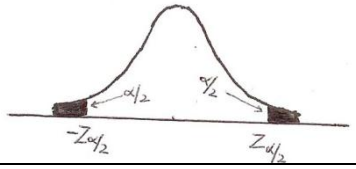
Critical Region: ***p*-value < α**

But in many situations we find it more useful to define the critical region in terms of test statistic. We, then, decide to reject or not to reject the null hypothesis H_0 by comparing the observed value of the test statistic with the cut-off value or the critical value of the test statistic.

Z-test

When in the testing of hypotheses, we use the random variable *Z* for calculating the *p*-value and for defining the critical region of the test; we call the test as *Z*-test. The critical region in terms of *Z* are summarized in Table 10-2

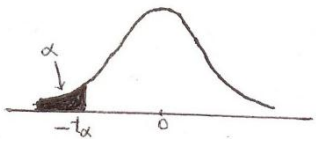
Table 10-2 Critical Region of Z-test

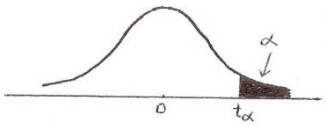
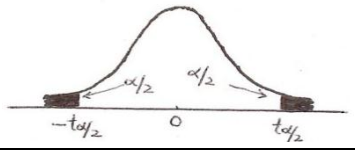
Test		Critical Region
Left-tailed		$Z < -Z_\alpha$
Right-tailed		$Z > Z_\alpha$
Two-tailed		$Z > Z_{\alpha/2}$ and $Z < -Z_{\alpha/2}$

t-test

When in the testing of hypotheses, we use the random variable *t* for calculating the *p*-value and for defining the critical region of the test; we call the test as *t*-test. The critical region in terms of *t* are summarized in Table 10-3

Table 10-3 Critical Region of *t*-test

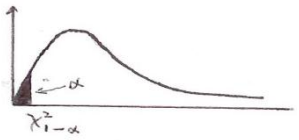
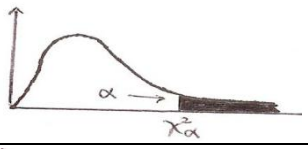
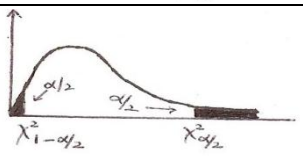
Test		Critical Region
Left-tailed		$t < -t_\alpha$

Right-tailed		$t > t_{\alpha}$
Two-tailed		$t > t_{\alpha/2}$ and $t < -t_{\alpha/2}$

χ^2-test

When in the testing of hypotheses, we use the random variable χ^2 for calculating the *p*-value and for defining the critical region of the test; we call the test as χ^2-test. The critical region in terms of χ^2 are summarized in Table 12-4

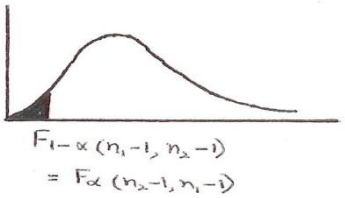
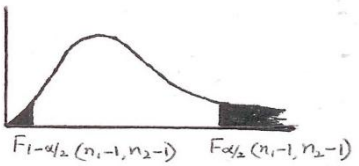
Table 10-4 Critical Region of χ^2-test

Test		Critical Region
Left-tailed		$\chi^2 < \chi^2_{1-\alpha}$
Right-tailed		$\chi^2 > \chi^2_{\alpha}$
Two-tailed		$\chi^2 > \chi^2_{\alpha/2}$ and $\chi^2 < \chi^2_{1-\alpha/2}$

F-test

When in the testing of hypotheses, we use the random variable *F* for calculating the *p*-value and for defining the critical region of the test; we call the test as *F*-test. The critical region in terms of *F* are summarized in Table 10-5

Table 10-5 Critical Region of *F*-test

Test	Critical Region
Left-tailed 	$F < F_{1-\alpha(n_1-1, n_2-1)}$ *i.e.* $F < F_{\alpha(n_2-1, n_1-1)}$
Right-tailed 	$F > F_{\alpha(n_1-1, n_2-1)}$
Two-tailed 	$F > F_{\alpha/2(n_1-1, n_2-1)}$ and $F < F_{1-\alpha/2(n_1-1, n_2-1)}$ *i.e.* $F < F_{\alpha/2(n_2-1, n_1-1)}$

10.2.4 GENERAL TESTING PROCEDURE

We have learnt a number of important concepts about hypothesis testing. We are now in a position to lay down a general testing procedure in a more systematic way. By now it should be clear that there are basically two phases in testing of hypothesis - in the first phase, we design the test and set up the conditions under which we shall reject the null hypothesis. In the second phase, we use the sample evidence and draw our conclusion as to whether the null hypothesis can be rejected. The detailed steps involved are as follows:

Step 1: State the Null and the Alternate Hypotheses. *i.e.* H_0 *and* H_1

Step 2: Specify a level of significance **α**

Step 3: Choose the test statistic and define the critical region in terms of the test statistic

Step 4: Make necessary computations

- calculate the observed value of the test statistic
- find the *p*- value of the test

Step 5: Decide to accept or reject the null hypothesis either

- by comparing the p- value with α or
- by comparing the observed value of the test statistic with the cut-off value or the critical value of the test statistic.

10.2.5 TESTS OF HYPOTHESES ABOUT POPULATION MEANS

When the null hypothesis is about a population mean, the test statistic can be either *Z* or *t*. If we use μ_0 to denote the claimed population mean the null hypothesis can be any of the three usual forms:

H_0: $\mu = \mu_0$ two-tailed test

H_0: $\mu \geq \mu_0$ left-tailed test

H_0: $\mu \leq \mu_0$ right-tailed test

Cases in Which the Test Statistic is *Z*

1. The population standard deviation, σ, is known and the population is normal.
2. The population standard deviation, σ, is known and the sample size, *n*, is at least 30 (The population need not be normal). The formula for calculating the test statistic *Z* in both these cases is

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

3. The population is normal and the population standard deviation, σ, is unknown, but the sample standard deviation, *S*, is known and the sample size, *n*, is large enough. The formula for calculating the test statistic *Z* in this case is

$$Z = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$$

Cases in Which the Test Statistic is *t*

1. The population is normal and the population standard deviation, σ, is unknown, but the sample standard deviation, *S*, is known and the sample size, *n*, is small.
2. The population is not normal and the population standard deviation, σ, is unknown, but the sample standard deviation, *S*, is known and the sample size, *n*, large enough.

The formula for calculating the test statistic *t* in both these cases is

$$t = \frac{\overline{X} - \mu_0}{S / \sqrt{n}}$$

The degrees of freedom for this *t* is *(n-1)*

10.2.6 TESTS OF HYPOTHESES ABOUT POPULATION ROPORTIONS

When the null hypothesis is about a population proportion, the test statistic can be either the Binomial random variable or its Poisson or Normal approximation. If we use p_0 to denote the claimed population proportion the null hypothesis can be any of the three usual forms:

H_0: $p = p_0$ two-tailed test

H_0: $p \geq p_0$ left-tailed test

H_0: $p \leq p_0$ right-tailed test

Cases in which the Test Statistic is Binomial Random Variable *X*

The Binomial distribution can be used whenever we are able to calculate the necessary binomial probabilities. When the Binomial distribution is used, the number of successes *X* serves as the test statistic. It is conveniently applicable to problems where sample size, *n*, is small and p_0 is neither very close to 0 nor to 1.

Cases in which the Test Statistic is Poisson Random Variable *X*

The Poisson approximation of Binomial distribution is conveniently applicable to problems where sample size, *n*, is large and p_0 is either very close to 0 or to 1. When the Poisson distribution is used, the number of successes *X* serves as the test statistic.

Note that the Binomial random variable or its Poisson approximation *X* follows a *discrete* distribution, and recall that the *p*-value is the probability of the test statistic being *equally or more unfavorable to* H_0 *than* the value obtained from the evidence. For example, for a right-tailed test with H_0: $p \leq 0.5$, the *p*-value = $P(X \geq$ observed number of successes).

Cases in Which the Normal Approximation is to be used

The Normal approximation of Binomial distribution is conveniently applicable to problems where sample size, *n*, is large and p_0 is neither very close to 0 nor to 1. When the normal distribution is used, the test statistic *Z* is calculated as:

$$Z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

10.2.7 TESTS OF HYPOTHESES ABOUT POPULATION VARIANCES

When the null hypothesis is about a population variance, the test statistic is χ^2. If we use σ_0 to denote the claimed population proportion the null hypothesis can be any of the three usual forms:

H_0:	$\sigma = \sigma_0$	two-tailed test
H_0:	$\sigma \geq \sigma_0$	left-tailed test
H_0:	$\sigma \leq \sigma_0$	right-tailed test

The formula for calculating the test statistic χ^2 is:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

The degree of freedom for this $\chi 2$ is *(n - 1).*

10.2.8 THE COMPARISON OF TWO POPULATIONS

Almost daily we compare products, services, investment opportunities, management styles and so on. In all such situations we are interested in the comparisons of two populations with respect to some population parameter - the population mean, the population proportion, or the population variance. Now we will learn how to conduct such comparisons in an objective and meaningful way.

10.2.8.1 TESTING FOR DIFFERENCE BETWEEN MEANS

When we want to arrive at same conclusion about the difference between two population means, we draw one sample from each of the population. The samples drawn may be dependent on each other or these may be independent of each other.

Dependent Samples- Paired Observations

In many situations, we can design our test in such a way that the samples drawn are dependent on each other and our observations come from two populations and are paired in some way. In general, when possible, it is often advisable to pair the observations, as this makes the experiment more precise. We can see the advantage of pairing observations with the helps of an example.

Consider a sales manager who wants to know if display at point of purchase helps in increasing the sales of his product. He may design the experiment in two ways:

Design I: He picks up a sample of, say 12, retail shops with no display at point of purchase. Similarly he picks up a sample of, say 10, retail shops with display at point of purchase. He will note his observations from both samples independently of each other.

Design II: He picks-up a random sample, of say 11, retail shops and note down the observations about weekly sale in each of these shops. Next he introduces display at point of purchase at each of these shops and again observes the weekly sales in them.

Obviously design II much better, as this tends to remove much of the extraneous variations in sales – the variation in the location of the soap, experimental conditions and other extraneous factors. Now after eliminating the effect of all other major factors, we can attribute the difference only to the ***'treatment***' we are studying -the display at point of purchase.

Let us label the two populations as 1 and 2. Under the situation of paired observations, it is easy to see that the variable in which we are interested is the differences between the two observations *i.e.* $d = x_1 - x_2$. In other words our two-population comparison test is reduced to a hypothesis test about one parameter - the difference between the means of two populations' *i.e.* $\mu_d = \mu_1 - \mu_2$

Thus the null hypothesis can be any of the three usual forms:

H_0:	$\mu_1 - \mu_2 = \mu_{d0}$	or	$\mu_d = \mu_{d0}$	two-tailed test
H_0:	$\mu_1 - \mu_2 \geq \mu_{d0}$	or	$\mu_d \geq \mu_{d0}$	left-tailed test
H_0:	$\mu_1 - \mu_2 \leq \mu_{d0}$	or	$\mu_d \leq \mu_{d0}$	right-tailed test

The test statistic can be either *t* or *Z*.

Cases in Which the Test Statistic is *t*

The population standard deviation of the difference, $\sigma_{d,}$ is not known and the sample size, *n*, is small. The formula for calculating the test statistic *t* is

$$t = \frac{\bar{d} - \mu_{d_0}}{S_d / \sqrt{n}}$$

The degrees of freedom for this *t* is *(n-1)*

Cases in Which the Test Statistic is *Z*

The sample size, *n*, is large and/or we happen to know the population standard deviation of the difference, σ_d. The formula for calculating the test statistic *t* is

$$Z = \frac{\bar{d} - \mu_{d_0}}{S_d / \sqrt{n}}$$

or

$$Z = \frac{\bar{d} - \mu_{d_0}}{\sigma_d / \sqrt{n}}$$

10.2.8.2 Independent Samples

When independent random sample are taken, the sample size need not be same for both populations. Let us label the two populations as 1 and 2. So that

μ_1 and μ_2 denote the two population means.

σ_1 and σ_2 denote the two population standard deviations

n_1 and n_2 denote the two sample sizes

$\overline{X_1}$ and $\overline{X_2}$ denote the two sample means

S_1 and S_2 denote the two sample standard deviations

If we use $(\mu_1 - \mu_2)_0$ to denote the claimed difference between the two population means, then the null hypothesis can be any of the three usual forms:

H_0:	$\mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$	two-tailed test
H_0:	$\mu_1 - \mu_2 \geq (\mu_1 - \mu_2)_0$	left-tailed test
H_0:	$\mu_1 - \mu_2 \leq (\mu_1 - \mu_2)_0$	right-tailed test

The test statistic can be either *Z* or *t*.

Cases in Which the Test Statistic is *Z*

1. The population standard deviations; σ_1 and σ_2; are known and both the populations are normal.
2. The population standard deviations; σ_1 and σ_2; are known and the sample sizes; n_1 and n_2; are both at least 30 (The population need not be normal).

The formula for calculating the test statistic *Z* in both these cases is

$$Z = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Cases in Which the Test Statistic is *t*

The populations are normal; the population standard deviations; σ_1 and σ_2; are unknown, but the sample standard deviations; S_1 and S_2; are known. The formula for calculating the test statistic *t* depends on two sub cases:

Subcase I: σ_1 and σ_2 are believed to be equal (although unknown)

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)_0}{\sqrt{S_p^2\left(1/n_1 + 1/n_2\right)}}$$

Where S_p^2 is the pooled variance of the two samples, which serves as the estimator of the common population variance.

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

The degrees of freedom for this *t* is $(n_1 + n_2 - 2)$.

Subcase II: σ_1 and σ_2 are believed to be unequal (although unknown)

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)_0}{\sqrt{\left(S_1^2/n_1 + S_2^2/n_2\right)}}$$

The degrees of freedom for this *t* is given by:

$$df = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1}\right)\left(\frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}\right)}$$

TESTING FOR DIFFERENCE BETWEEN POPULATION PROPORTIONS

We will consider the large-sample tests for the difference between population proportions. For ***'large enough'*** sample sizes the distribution of the two sample proportions and also the distribution of the difference between the two sample proportions is approximated well by a normal distribution. This gives rise to *Z*-test for comparing the two population proportions. Let us assume independent random sampling from the two populations, labeled as 1 and 2, so that

p_1 and p_2 denote the two population proportions

n_1 and n_2 denote the two sample sizes

$\bar{p}_1$ and $\bar{p}_2$ denote the two sample proportions

We will use $(p_1 - p_2)_0$ to denote the claimed difference between the two population proportions. Then the null hypothesis can be any of the three usual forms:

H_0:	$p_1 - p_2 = (p_1 - p_2)_0$	two-tailed test
H_0:	$p_1 - p_2 \geq (p_1 - p_2)_0$	left-tailed test
H_0:	$p_1 - p_2 \leq (p_1 - p_2)_0$	right-tailed test

The formula for calculating the test statistic *Z* depends on two cases.

Case I: When $(p_1 - p_2)_0 = 0$ *i.e.* the claimed difference between the two population proportions is zero

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - 0}{\sqrt{\bar{p}(1-\bar{p})\left(1/n_1 + 1/n_2\right)}}$$

Where $\bar{p}$ is the combined sample proportion in both the samples

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Case II: When $(p_1 - p_2)_0 \neq 0$ *i.e.* the claimed difference between the two population proportions is some number other than zero

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)_0}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}}$$

TESTING FOR EQUALITY OF TWO POPULATION VARIANCES

Many a times, we may be interested in comparing the degree of variability or dispersion of two different populations. Here the problem essentially involves testing the equality of two population variances. Let us assume independent random sampling from the two populations, labeled as 1 and 2, so that

σ_1^2 and σ_2^2 denote the two population variances

n_1 and n_2 denote the two sample sizes

S_1^2 and S_2^2 denote the two sample variances

Then the null hypothesis can be any of the three usual forms:

H_0: $\sigma_1^2 = \sigma_2^2$ two-tailed test

H_0: $\sigma_1^2 \geq \sigma_2^2$ left-tailed test

H_0: $\sigma_1^2 \leq \sigma_2^2$ right-tailed test

The formula for calculating the test statistic F is:

$$F_{(n_1-1,\, n_2-1)} = \frac{S_1^2}{S_2^2}$$

The degrees of freedom for this F is $(n_1\text{-}1,\, n_2\text{-}1)$

10.3 SOLVED PROBLEMS

Now we will solve some problems relating to testing the hypotheses stated about different population parameters, under different conditions.

Example 10-1

An automatic bottling machine fills oil into *2-liter* (2,000 cm^3) bottles. A consumer advocate wants to test the null hypothesis that the average amount filled by the machine into a bottle is at least 2,000 cm^3. A random sample of 40 bottles coming out of the machine was selected and the exact contents of the selected bottles are recorded. The sample mean was 1,999.6 cm^3. The population standard deviation is known from past experience to be 1.30 cm^3.

(a) Test the null hypothesis at an **α** of 5%.

(b) Assume that the population is normally distributed with the same standard deviation of 1.30 cm^3. Assume that the sample size is only 20 but the sample mean is the same 1,999.6 cm^3. Conduct the test once again at an α of 5%.

If there is a difference in the two test results, explain the reason for the difference.

Solution: (a)

1. **The null and alternative hypotheses:**

$$H_0: \quad \mu \geq 2{,}000$$

$$H_1: \quad \mu < 2{,}000$$

The test is a *left-tailed* test

2. **Level of significance:** α = 5% or 0.05
3. **Test statistic: Z;** as the population standard deviation **is** known and sample size is greater than 30
4. **Critical region:** $Z < -Z_{0.05}$ Where $Z_{0.05}=1.645$
5. **Computations:** $\overline{X}$ =1,999.6; σ = 1.30; n = 40

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}; \quad Z = \frac{1{,}999.6 - 2{,}000}{1.30/\sqrt{40}}; \quad Z = -1.95$$

6. **Conclusion:** We reject the null hypothesis at α =0.05 since $Z = -1.95 < -Z_{0.05} = -1.645$.

(b) Since the population is normally distributed, the test statistic is once again Z

Computations: $\overline{X}$ =1,999.6; σ = 1.30; n = 20

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}; \; Z = \frac{1{,}999.6 - 2{,}000}{1.30/\sqrt{20}} \; ; \; Z = -1.38$$

Conclusion: We do not reject the null hypothesis at α=0.05 since $Z = -1.38 > -Z_{0.05} = -1.645$

(c) In the first case we could reject the null hypothesis but in the second we could not, although in both cases the sample mean was the same. The reason is that in the first case the sample size was larger and therefore the evidence against the null hypothesis was more reliable. This produced a smaller *p*-value in the first case.

Example 10-2

An automobile manufacturer substitutes a different engine in cars that were known to have an average miles-per-gallon rating of 31.5 on the highway. The manufacturer wants to test whether the new engine changes the miles-per-gallon rating of the automobile model. A random sample of 100 trial runs gives $\overline{X}$ = 29.8 miles per gallon and S = 6.6 miles per gallon. Using the 0.05 level of significance, is the

average miles-per-gallon rating on the highway for cars using the new engine different from the rating for cars using the old engine?

Solution:

1. **The null and alternative hypotheses:**

$$H_0: \quad \mu = 31.5$$

$$H_1: \quad \mu \neq 31.5$$

The test is a *two-tailed* test

2. **Level of significance:** $\alpha = 5\%$ or 0.05
3. **Test statistic:** *Z;* as the sample standard deviation is known and sample size is greater than 30
4. **Critical region:** $Z_{0.025} < Z < -Z_{0.025}$ Where $Z_{0.025} = 1.96$
5. **Computations:** $\overline{X} = 29.8$, $S = 6.6$, $n = 100$

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}; \quad Z = \frac{29.8 - 31.5}{6.6/\sqrt{100}}; \quad Z = -2.57$$

6. **Conclusion:** We reject the null hypothesis at $\alpha = 0.05$ since $Z = -2.57 < -Z_{0.025} = -1.96$. So we conclude that the average miles-per-gallon rating on the highway for cars using the new engine is different from the rating for cars using the old engine.

Example 10-3

Sixteen oil tins are taken at random from an automatic filling machine. The mean weight of the tins is 12.2 kg, with a standard deviation of 0.40 kg. Can we conclude that the filling machine is wasting oil by filling more than the intended weight of 12 kg, at a significance level of 5%?

Solution:

1. **The null and alternative hypotheses:**

$$H_0: \quad \mu \leq 12.2$$

$$H_1: \quad \mu > 12.2$$

The test is a *right-tailed* test

2. **Level of significance:** $\alpha = 5\%$ or 0.05
3. **Test statistic:** *t;* as the sample standard deviation is known and sample size is small.
4. **Critical region:** $t > t_{0.05}$ Where $t_{0.05}$ for *15 df* $= 1.7530$

5. **Computations:** $\overline{X}$ =12.2; S = 0.40; n = 16

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}};\ t = \frac{14.2 - 14}{0.40/\sqrt{16}};\ t = 2$$

6. **Conclusion:** We reject the null hypothesis at α = 0.05 since $t = 2 > t_{0.05} = 1.7530$. So we conclude that the filling machine is wasting oil by filling more than the intended weight of 12 kg.

Example 10-4

A coin is to be tested for fairness. It is tossed 15 times and only 8 heads are observed. Test if the coin is fair at α = 5%.

Solution:

1. **The null and alternative hypotheses:**

   H0: p = 0.5

   H1: p ≠ 0.5

   The test is a two-tailed test
2. **Level of significance:** α = 5% or 0.05
3. **Test statistic:** Binomial random variable X
4. **Critical region: p-value < α**
5. **Computations:** p_0 = 0.5; n = 15

$$p\text{-value} = 2*P(X \le 8) = 2*\left(\sum_{X=0}^{8} {}^{n}C_X\, p^X (1-p)^{n-X}\right) = 2*\left(\sum_{X=0}^{8} {}^{15}C_X\, 0.5^X (1-0.5)^{n-X}\right)$$

$$= 0.5034$$

6. **Conclusion:** We cannot reject the null hypothesis at α = 0.05 since p-value > α. So we accept that the coin is fair.

Example 10-5

A wholesaler received a shipment of goods, which is reported to be containing at most 2% defective items. He will accept the shipment if the claim is found true and reject if the percentage of defective items is more. To verity this claim, he draws a sample of 200 items and finds that 10 items are defective. What should be his decision at 5% level of significance?

Solution:

1. **The null and alternative hypotheses:**

$$H_0: \quad p \leq 0.02$$

$$H_1: \quad p > 0.02$$

The test is a *right-tailed* test

2. **Level of significance:** α = 5% or 0.05
3. **Test statistic:** Poisson random variable *X* since p_0 is very small and the sample size is large enough to use poisson approximation of binomial distribution.
4. **Critical region:** *p*-value < α
5. **Computations:** p_0 = 0.02; *n* = 200; μ = 4

$$p\text{-value} = P(X \geq 10) = 1 - P(X \leq 9) = 1 - \sum_{X=0}^{9}\left(\frac{e^{-\mu}\mu^X}{X!}\right) = 1 - 0.9919 = 0.0081$$

6. **Conclusion:** We reject the null hypothesis at α =0.05 since *p*-value < α. So the wholesaler will not accept the shipment.

Example 10-6

SBI claims that more than 55% of the saving accounts in Haryana are at SBI. A sample survey of 400 account holders revealed that only 180 account holders have account at SBI. Verify, using 5% level of significance, if the sample results underestimate the claim of SBI.

Solution:

1. **The null and alternative hypotheses:**

$$H_0: \quad p \geq 0.55$$

$$H_1: \quad p < 0.55$$

The test is a *left-tailed* test

2. **Level of significance:** α = 5% or 0.05
3. **Test statistic:** *Z;* since p_0 is not too close to 0 or 1 and the sample size is large enough to use normal approximation of binomial distribution.
4. **Critical region:** $Z < -Z_{0.05}$ Where $Z_{0.05} = 1.645$
5. **Computations:** p_0 = 0.55, $\bar{p}$ = 180/400 = 0.45, *n* = 400

$$Z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}; \quad Z = \frac{0.45 - 0.55}{\sqrt{0.55(1-0.55)/400}}; \quad Z = \frac{-20000}{4975}; \quad Z = -4.02$$

6. **Conclusion:** We reject the null hypothesis at α = 0.05 since $Z = -4.02 < -Z_{0.05} = -1.645$. So the sample results underestimate the claim of SBI.

Example 10-7

A manufacturer of golf balls claims that the company controls the weights of the golf balls accurately so that the variance of the weights is not more than 1 mg^2. A random sample of 31 golf balls yields a sample variance of 1.62 mg^2.Is that sufficient evidence to reject the claim at an α of 5%?

Solution:

1. **The null and alternative hypotheses:**

$$H_0: \quad \sigma^2 \leq 1$$

$$H_1: \quad \sigma^2 > 1$$

The test is a *right-tailed* test

2. **Level of significance:** α = 5% or 0.05
3. **Test statistic:** χ^2
4. **Critical region:** $\chi^2 > \chi^2_{0.05}$ Where $\chi^2_{0.05}$ for $30\ df = 43.7729$
5. **Computations:** $\sigma_0^2 = 1$; $S^2 = 1.62$; $n = 31$

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{30x1.62}{1} = 48.6$$

6. **Conclusion:** We reject the null hypothesis at α =0.05 since $\chi^2 = 48.6 > \chi^2_{0.05} = 43.7729$. So we conclude that there is sufficient evidence to reject the claim of the company.

Example 10-8

A sales manager wants to know if display at point of purchase helps in increasing the sales of his product. He notes the following observations:

Shop No.	1	2	3	4	5	6	7	8	9	10	11
Sales before display	4500	5275	7235	6844	5991	6672	4943	7615	6128	5623	5154

Sales after display	4834	5010	7562	6957	6401	6423	5334	8004	6729	6277	5769
Difference*(d)*	-334	265	-327	-113	-410	249	-391	-389	-581	-654	-615

$\bar{d}$ = -300 S_d = 312.53

Is there sufficient evidence to conclude that display at point of purchase helps in increasing the sales of his product?

Solution:

1. **The null and alternative hypotheses:**

H_0: $\mu_d \geq 0$

H_1: $\mu < 0$

The test is a *left-tailed* test

2. **Level of significance:** α = 5% or 0.05
3. **Test statistic:** *t;* as the population standard deviation of the difference, σ_d, *is* not known and the sample size, *n*, is small.
4. **Critical region:** $t < -t_{0.05}$ Where $t_{0.05}$ for *10 df =1.812*
5. **Computations:** $\bar{d}$ =-300; *S* = 312.53; *n* = 11

$$t = \frac{\bar{d} - \mu_{d_0}}{S_d / \sqrt{n}}; \quad t = \frac{-300 - 0}{314.53 / \sqrt{11}}; \quad t = -3.16$$

6. **Conclusion:** We reject the null hypothesis at α =0.05 since $t = -3.16 < t_{0.05} = -1.812$. So the sales manager has sufficient evidence to conclude that display at point of purchase helps in increasing the sales.

Example 10-9

The makers of Duracell batteries want to demonstrate that their size AA battery lasts on an average of at least 45 minutes longer than Duracell's main competitor, the Energizer. Two independent random samples of 100 batteries of each kind are selected. The sample average lives for Duracell and Energizer batteries are found to be $\bar{X}_1$ = 308 minutes and $\bar{X}_2$ = 254 minutes respectively. Assume $\sigma_1 = 84$ minutes and $\sigma_2 = 67$ minutes. Is there evidence to substantiate Duracell's claim that its batteries last, on an average, at least 45 minutes longer than Energizer of the same size?

Solution:

1. **The null and alternative hypotheses:**

$$H_0: \quad \mu_1 - \mu_2 \leq 45$$

$$H_1: \quad \mu_1 - \mu_2 > 45$$

The test is a *right-tailed* test

2. **Level of significance:** α = 5% or 0.05
3. **Test statistic:** *Z*
4. **Critical region:** $Z > Z_{0.05}$ Where $Z_{0.05} = 1.645$
5. **Computations:** $\overline{X}_1 = 308$; $\overline{X}_2 = 254$; $\sigma_1 = 84$; $\sigma_2 = 67$; $n_1 = n_2 = 100$

$$Z = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}; \; Z = \frac{308 - 254 - 45}{\sqrt{84^2/100 + 67^2/100}}; \; Z = 0.838$$

6. **Conclusion:** We cannot reject the null hypothesis at α = 0.05 since $Z = 0.838 < Z_{0.05} = 1.645$. In fact the observed value of the test statistic falls in the non-rejection region of our *right-tailed* test at any conventional level of significance. So we must conclude that there is insufficient evidence to support Duracell's claim.

Example 10-10

The following information relate to the prices (in Rs) of a product in two cities A and B.

	City A	City B
Mean price	22	17
Standard deviation	5	6

The observations related to prices are made for 9 months in city A and for 11 months in city B. Test at 0.01 level whether there is any significant difference between prices in two cities, assuming (a) $\sigma_1^2 = \sigma_2^2$ (b) $\sigma_1^2 \neq \sigma_2^2$

Solution:

1. **The null and alternative hypotheses:**

$$H_0: \quad \mu_1 - \mu_2 = 0$$

$$H_1: \quad \mu_1 - \mu_2 \neq 0$$

The test is a *two-tailed* test

2. **Level of significance:** α =1% or 0.01
3. **Test statistic:** *t;* since the population standard deviations, σ_1 and σ_2, are unknown, but the sample standard deviations, S_1 and S_2, are known and sample sizes are small.
4. **Critical region:** $t_{0.005} < t < -t_{0.005}$
5. **Computations:** $\overline{X}_1 = 22$; $\overline{X}_2 = 17$; $S_1 = 5$; $S_2 = 6$; $n_1 = 9$; $n_2 = 11$

(a) $\sigma_1^2 = \sigma_2^2$

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)_0}{\sqrt{S_P^2\left(1/n_1 + 1/n_2\right)}} \quad t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)_0}{\sqrt{\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\right)\left(1/n_1 + 1/n_2\right)}}$$

$$t = \frac{22-17}{\sqrt{\left(\frac{8x25+10x36}{18}\right)\left(1/9 + 1/11\right)}}; \; t = \frac{5}{2.51}; \; t = 1.99$$

The degrees of freedom for this *t* are $n_1 + n_2 - 2$ *i.e.* 9 +11-2 =18

For 18 *df* , $t_{0.005} = 2.88$

(b) $\sigma_1^2 \neq \sigma_2^2$

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)_0}{\sqrt{\left(S_1^2/n_1 + S_2^2/n_2\right)}}; t = \frac{22-17-0}{\sqrt{\left(25/9 + 36/11\right)}}; \; t = \frac{5}{2.46}; \; t = 2.03$$

The degrees of freedom for this *t* are given by

$$df = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1}\right)\left(\frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}\right)}, \; df = \frac{\left(25/9 + 36/11\right)^2}{\left(\frac{\left(25/9\right)^2}{8}\right)\left(\frac{\left(36/11\right)^2}{10}\right)} = 18$$

Against which, $t_{0.005} = 2.88$

6. **Conclusion: (a)** We cannot reject the null hypothesis at α = 0.01, when $\sigma_1^2 = \sigma_2^2$ since $t=1.99 < t_{0.005} = 2.88$. **(b)** We cannot reject the null hypothesis at α = 0.01, when $\sigma_1^2 \neq \sigma_2^2$ since $t=2.03 < t_{0.005} = 2.88$.

Example 10-11

A sample survey of tax-payers belonging to business class and professional class yielded the following results:

	Business Class	Professional Class
Sample size	$n_1 = 400$	$n_2 = 420$
Defaulters in tax payment	$x_1 = 80$	$x_2 = 65$

Given these sample data, test the hypothesis at α = 5% that

(a) the defaulters rate is the same for the two classes of tax-payers

(b) the defaulters rate in the case of business class is more than that in the case of professional class by 0.07.

Solution: (a)

1. **The null and alternative hypotheses:**

H_0: $p_1 - p_2 = 0$

H_1: $p_1 - p_2 \neq 0$

The test is a *two-tailed* test

2. **Level of significance:** α =1% or 0.01

3. **Test statistic:** *Z;* since the sample sizes are large enough.

4. **Critical region:** $Z_{0.005} < Z < -Z_{0.005}$ Where $Z_{0.005} = 2.58$

5. **Computations:**

$$\overline{p_1} = \frac{x_1}{n_1} = \frac{80}{400} = 0.20; \quad \overline{p_2} = \frac{x_2}{n_2} = \frac{65}{420} = 0.15 \quad \overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{80+65}{400+420} = 0.177$$

$$Z = \frac{(\overline{p}_1 - \overline{p}_2) - 0}{\sqrt{\overline{p}(1-\overline{p})\left(1/n_1 + 1/n_2\right)}}; Z = \frac{0.20 - 0.15}{\sqrt{(0.177x0.823)\left(1/400 + 1/420\right)}} \; Z = 1.87$$

6. **Conclusion:** We cannot reject the null hypothesis at α = 0.05 since $Z = 1.87 < Z_{0.005} = 2.58$

(b)

1. **The null and alternative hypotheses:**

H_0: $p_1 - p_2 = 0.07$

H_1: $p_1 - p_2 \neq 0.07$

The test is a *two-tailed* test

2. **Level of significance:** α =1% or 0.01
3. **Test statistic:** *Z;* since the sample sizes are large enough.
4. **Critical region:** $Z_{0.005} < Z < -Z_{0.005}$ Where $Z_{0.005} = 2.58$
5. **Computations:**

$$\overline{p_1} = \frac{x_1}{n_1} = \frac{80}{400} = 0.20 \quad \overline{p_2} = \frac{x_2}{n_2} = \frac{65}{420} = 0.15 \quad Z = \frac{(\overline{p}_1 - \overline{p}_2) - (p_1 - p_2)_0}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}}$$

$$Z = \frac{(0.20 - 0.15) - (0.07)}{\sqrt{\frac{0.20x0.80}{400} + \frac{0.15x0.85}{420}}} \qquad Z = -0.76$$

6. Conclusion: We cannot reject the null hypothesis at α = 0.05 since $Z = -0.76 > -Z_{0.01} = -2.58$.

Example 10-12

Use the data of Problem 12-10: n_1 = 9, n_2 =11 and $S_1 = 5$, $S_2 = 6$ to test the assumption of equal population variances.

Solution:

1. **The null and alternative hypotheses:**

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

The test is a *two-tailed* test

2. **Level of significance:** α = 5% or 0.05
3. **Test statistic:** *F*
4. **Critical region:** $F_{(n_1-1,n_2-1)} > F_{\alpha/2(n_1-1,\,n_2-1)}$ and $F_{(n_1-1,n_2-1)} < F_{1-\alpha/2(n_1-1,\,n_2-1)}$ *i.e.* $F_{(8,10)} > F_{0.025(8,10)} = 3.85$ and $F_{(8,10)} < F_{0.095(8,10)} = 0.23$

5. Computations: $S_1 = 5;\ S_2 = 6$; $n_1 = 9$; $n_2 = 11$

$$F_{(n_1-1,n_2-1)} = \frac{S_1^2}{S_2^2} \qquad F_{(8,10)} = \frac{25}{36} = 0.694$$

6. Conclusion: We cannot reject the null hypothesis at α = 0.05 since $F_{(8,10)} < F_{0.025(8,10)} = 3.85$ and $F_{(8,10)} > F_{0.095(8,10)} = 0.23$. So the sample evidence supports the view that the two populations do not have different variances.

10.4 CHECK YOUR PROGRESS

1. We cannot …… H_0 at an α of 5%" rather than "We accept H_0."

2. In general, other things remaining the same, ……the value of α will decrease the probability of type II error.

3. In the case of a left-tailed test, the p-value is the area to the …… of the calculated value of the test statistic.

4. The value of β tends to …….. as μ_1 moves nearer to μ_0.

5. When the null hypothesis is about a population proportion, the test statistic can be either the ……. or its Poisson or Normal approximation.

10.5 SUMMARY

A hypothesis is something that has not yet been proven to be true. It is some statement about a population parameter or about a population distribution. This statement is tentative as it implies some assumption, which may or may not be found valid on verification. Hypothesis testing is the process of determining whether or not a given hypothesis is true. If the population is large, there is no way of analyzing the population or of testing the hypothesis directly. Instead, the hypothesis is tested on the basis of the outcome of a random sample. In any testing of hypotheses problem, we are faced with a pair of hypotheses such that one and only one of them is always true. One of this pair is called the null hypothesis and the other one the alternative hypothesis. Almost daily we compare products, services, investment opportunities, management styles and so on. In all such situations we are interested in the comparisons of two populations with respect to some population parameter - the population mean, the population proportion, or the population variance.

10.6 KEYWORDS

Null Hypothesis: A null hypothesis is an assertion about the value of a population parameter. It is an assertion that we hold as true unless we have sufficient statistical evidence to conclude otherwise.

Alternative Hypothesis: The alternative hypothesis is the negation of the null hypothesis.

Type I Error: In the context of statistical testing, the wrong decision of rejecting a true null hypothesis is known as Type I Error.

Type II Error: The wrong decision of accepting (not rejecting, to be more accurate) a false null hypothesis is known as Type II Error.

P-value: The probability of observing a sample statistic as extreme as the one observed if the null hypothesis is true.

10.7 SELF-ASSESSMENT TEST

1. What is a Hypothesis? Explain how Hypothesis Testing is useful to management?
2. What are Null and Alternative hypotheses? How you will set up null and alternative hypotheses under following conditions:
   (a) A pharmaceutical company claims that four out of five doctors prescribe the pain medicine it produces. You wish to test this claim.
   (b) A manufacturer of golf balls claims that the variance of the weights of the company's golf balls is controlled within 0.0028 oz^2. You wish to test this claim.
   (c) A medicine is effective only if the concentration of a certain chemical in it is at least 200 parts per million (ppm). At the same time the medicine would produce an undesirable side effect if the concentration of the same chemical exceeds 200 parts per million (ppm). You wish to test the concentration of the chemical in the medicine.
3. What are Type I and Type II Errors in hypothesis testing? Explain the relationship between the two types of errors.
4. What is a Test Statistic? Why do we have to know the distribution of the test statistic? What are the commonly used test statistics in hypotheses testing?
5. Distinguish between a One-tailed and Two-tailed test, give a diagram and an example in each case.
6. What is the *p*-value of a test? How it is calculated? Find the *p*-value of a (a) left-tailed, (b) right-tailed, and (c) two-tailed test if

(i) In the test, the test statistic Z = -1.86. In which of these three cases will H_0 be rejected at an α of 5%?

(ii) In the test, the test statistic Z = 1.75. In which of these three cases will H_0 be rejected at an α of 5%?

7. What do you mean by Level of Significance of a test? "Level of significance should be specified after due consideration to the costs associated with Type I and Type II errors". Explain this statement.
8. What do you mean by Critical Regain and Acceptance Region of a test?
9. What is the Power of a hypothesis test? Why is it important? How is the power of a hypothesis test related to

   (a) the significance level?

   (b) the sample size?

   (c) the actual value of the parameter?
10. Consider the use of metal detectors in airports to test people for concealed weapons. In essence, this is a form of hypothesis testing.

    (a) What are the null and alternative hypotheses?

    (b) What are type I and type II errors in this case?

    (c) Which type of error is more costly?

    (d) Based on your answer to part (c), what value of α would you recommend for this test?

    (e) If the sensitivity of the metal detector is increased, how would the probabilities of type I and type II errors be affected?

    (f) If α is to be increased, should the sensitivity of the metal detector be increased or decreased?
11. When planning a hypothesis test, what should be done if the probabilities of both type I and type II errors are to be small?
12. "Not – rejecting a Null hypothesis" is a more precise term rather than "Accepting a Null hypothesis". Do you agree with this statement? Explain.
13. What step are involved in statistical testing of a hypothesis?
14. A company is engaged in the packaging of a superior quality tea in jars of 500gm each. The company is of the view that as long as the jars contains 500gm of tea, the process is under control.

The standard deviation of the process is 50gm. A sample of 225 jars is taken at random and the sample average is found to be 510 gm. Has the process gone out of control?

15. A sample of size 400 was drawn and the sample mean found to be 99. Test, at 5% level of significance, whether this sample could have come form normal population with mean 100 and variance 64.
16. A manufacturer of a new motorcycle claims for it an average mileage of 60 km/liter under city conditions. However, the average mileage in 16 trials is found to be 57 km, with a standard deviation of 2 km. Is the manufacturer's claim justified?
17. In a big city, 450 men out of a sample of 850 men were found to be smokers. Does this information, at 5% level of significance, supports the view that the majority of men in this city are smokers?
18. A stock-broker claims that she can predict with 85% accuracy whether a stock's market value will rise or fall during the coming month. Test the stock-broker's claim at 5% level of significance if, as a test, she predict the outcome of 6 stocks and is correct in 5 of the predictions.
19. A company engaged in manufacturing of radio tubes, finds that the life of its tubes has a variance of 0.7 years. As a result of some qualitative improvement brought about in the product, the company claims that the variance of the life of its tubes has reduced. If the sample variance, S2, on observation of 9 tubes is observed 0.55 years at test the claim of the company (a) 5% level of significance (b) 1% level of significance.
20. Seven persons were appointed in officer cadre in an organisation. Their performance was evaluated by giving a test and the marks were recorded out of 100. They were given two-month training and another test was held and marks were recorded out of 100.

Officer:	a	b	c	d	e	f	g
Score Before Training:	80	76	92	60	70	56	74
Score After Training:	84	70	96	80	70	52	84

Can it be concluded that the training has benefited the employees? Use 5% level of significance.

21. The makers of Philips bulb want to demonstrate that their bulb lasts on an average of at least 100 hours longer than Philips' main competitor, Surya. Two independent random samples of 100 bulbs of each kind are selected. The sample average lives for Philips and Surya bulbs are found to be $\overline{X}_1$ = 1232 hours and $\overline{X}_2$ = 1016 hours respectively. Assume $\sigma_1 = 84$ hours and $\sigma_2 = 67$hours. Is

there evidence to substantiate Philips' claim that its bulbs last, on an average, at least 180 hours longer than Surya bulb of the same size?

22. Consider the following data:

	Sample A	Sample B
Sample Mean	100	105
Standard Deviation	16	24
Sample Size	800	1600

Test, at 5% level of significance, the difference between means of two populations from which samples are taken.

23. The following information relate to the wages (in Rs) of mill workers in two cities A and B.

	City A	City B
Mean wage	40	34
Standard deviation	5	6

The observations related to wages are for 8 workers in city A and for 10 workers in city B. Test at 0.01 level whether there is any significant difference between wages in two cities, assuming (a) $\sigma_1^2 = \sigma_2^2$ (b) $\sigma_1^2 \neq \sigma_2^2$

24. Test market result of two advertisements A and B, yielded the following results:

	A	B
Who saw the Advertisements	$n_1 = 200$	$n_2 = 220$
Who tried the Product	$x_1 = 40$	$x_2 = 35$

Given the data, test the hypotheses at $\alpha = 5\%$ that

(a) both the advertisements are equally effective

(b) advertisement A is more effective than advertisement B by more than 0.05

Effectiveness of the advertisements are measured as proportion of viewers who tried the product.

25. Use the data of Problem 22: $n_1 = 8$, $n_2 = 10$ and $S_1 = 5$, $S_2 = 6$ to test the assumption of equal population variances.

10.8 ANSWERS TO CHECK YOUR PROGRESS

1. Reject

2. Increasing
3. Left
4. Increase
5. Binomial random variable

10.9 REFERNCES/SUGGESTED READINGS

25. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
26. Statistics for Management *by* G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
27. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
28. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
29. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons. New Delhi.
30. Statistical Method by S.P. Gupta. Sultan Chand and Sons. New Delhi.
31. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
32. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.

Course: Business Statistics	
Course Code: MBA-102	Author: Dr. B.S. Bodla
Lesson: 11	Vetter: Karam Pal
NON-PARAMETRIC TESTS	

STRUCTURE

11.1 Learning Objectives

11.2 Introduction

- 11.2.1 Sign tests
- 11.2.2 The two-sample and K-sample Median Tests
- 11.2.3 Wilcoxon matched-pairs test (or Signed Rank Test)
- 11.2.4 The Mann-Whitney U Test
- 11.2.5 The Kruskal-Wallis Test
- 11.2.6 The spearman's rank correlation test
- 11.2.7 Tests of Randomness: Runs Above and Below the Median
- 11.2.8 Kolmogorov-Smirnov One-sample Test

11.3 Check your Progress

11.4 Summary

11.5 Keywords

11.6 Self- Assessment Test

11.7 Answers to check your progress

11.8 References/Suggested readings

11.1 LEARNING OBJECTIVES

After going through this lesson, the students will be able to:

- Differentiate between parametric and nonparametric tests

- Understand the relevance of non-parametric test in data analysis
- Understand the procedure involved in carrying out non-parametric tests
- Design and conduct some selected non-parametric tests

11.2 INTRODUCTION

In contrast to parametric tests, non-parametric tests do not require any assumptions about the parameters or about the nature of population. It is because of this that these methods are sometimes referred to as the distribution free methods. Most of these methods, however, are based upon the weaker assumptions that observations are independent and that the variable under study is continuous with approximately symmetrical distribution. In addition to this, these methods do not require measurements as strong as that required by parametric methods. Most of the non-parametric tests are applicable to data measured in an ordinal or nominal scale. As opposed to this, the parametric tests are based on data measured at least in an interval scale. The measurements obtained on interval and ratio scale are also known as high level measurements.

Level of measurement

1. ***Nominal scale:*** This scale uses numbers or other symbols to identify the groups or classes to which various objects belong. These numbers or symbols constitute a nominal or classifying scale. For example, classification of individuals on the basis of sex (male, female) or on the basis of level of education (matric, senior secondary, graduate, post graduate), etc. This scale is the weakest of all the measurements.
2. ***Ordinal scale:*** This scale uses numbers to represent some kind of ordering or ranking of objects. However, the differences of numbers, used for ranking, don't have any meaning. For example, the top 4 students of class can be ranked as 1, 2, 3, 4, according to their marks in an examination.
3. ***Interval scale:*** This scale also uses numbers such that these can be ordered and their differences have a meaningful interpretation.
4. ***Ratio scale:*** A scale possessing all the properties of an interval scale along with a *true zero point* is called a ratio scale. It may be pointed out that a zero point in an interval scale is arbitrary. For example, freezing point of water is defined at 0° Celsius or 32° Fahrenheit, implying thereby that

the zero on either scale is arbitrary and doesn't represent total absence of heat. In contrast to this, the measurement of distance, say in metres, is done on a ratio scale. The term ratio is used here because ratio comparisons are meaningful. For example, 100 kms of distance is four times larger than a distance of 25 kms while 100°F may not mean that it is twice as hot as 50°F.

It should be noted here that a test that can be performed on high level measurements can always be performed on ordinal or nominal measurements but not vice-versa. However, if along with the high level measurements the conditions of a parametric test are also met, the parametric test should invariably be used because this test is most powerful in the given circumstances.

From the above, we conclude that a non-parametric test should be used when either the conditions about the parent population are not met or the level of measurements is inadequate for a parametric test.

Advantages

The non-parametric tests have gained popularity in recent years because of their usefulness in certain circumstances. Some advantages of non-parametric tests are mentioned below:

1. Non-parametric tests require less restrictive assumptions vis-à-vis a comparable parametric test.
2. These tests often require very few arithmetic computations.
3. There is no alternative to using a non-parametric test if the data are available in ordinal or nominal scale.
4. None of the parametric tests can handle data made up of samples from several populations without making unrealistic assumptions. However, there are suitable non-parametric tests available to handle such data.

Disadvantages

1. It is often said that non-parametric tests are less efficient than the parametric tests because they tend to ignore a greater part of the information contained in the sample. In spite of this, it is argued that although the non-parametric tests are less efficient, a researcher using them has more confidence in using his methodology than he does if he must adhere to the unsubstantial assumptions inherent in parametric tests.

2. The non-parametric tests and their accompanying tables of significant values are widely scattered in various publications. As a result of this, the choice of most suitable method, in a given situation, may become a difficult task.

11.2.1 Sign tests

One of the easiest non-parametric tests is the sign test. The test is known as the sign test as it is based on the direction of the plus or minus signs of observations in a sample instead of their numerical values. There are two types of sign tests: (a) *One-sample sign test,* and (b) *Two-sample sign test*.

One-sample sign test

The one-sample sign test is a very simple non-parametric test applicable on the assumption that we are dealing with a population having a continuous symmetrical distribution. As such, the probability of getting a value less than the mean is 0.5. Likewise, the probability of getting a value greater than the mean is also 0.5. To test the null hypothesis $\mu = \mu_0$ against an appropriate alternative, each sample value greater than μ_0 is replaced by plus (+) sign and each sample value less than μ_0 with a minus (-) sign. Having done this, we can test the null hypothesis that the probabilities of getting both plus and minus signs are 0.5. It may be noted that if a sample value happens to be equal to μ_0, it is simply discarded.

To perform the actual test, we use either of the two methods. When the sample is small, the test is performed by computing the binomial probabilities or by referring to the binomial probabilities table. When the sample is large, the normal distribution is used as an approximation of the binomial distribution. Let us take an example to show how the one-sample sign test is applied.

Example 1: We are required to test the hypothesis that the mean value μ of a continuous distribution is 20 against the alternative hypothesis $\mu \neq 20$. Fifteen observations were taken and the following results were obtained:

18, 19, 25, 21, 16, 13, 19, 22, 24, 21, 18, 17, 13, 26 and 24.

We may use $\alpha = 0.05$ level of significance.

Solution: Replacing each value greater than 20 with a plus (+) sign and each value less than 20 with a minus (-) sign, we get

- - + + - - - + + + - - - + +

Now, the question before us is whether 7 plus signs observed in 13 trials support the null hypothesis p = 0.5 or the alternative hypothesis p ≠ 0.5. Using the binomial probability tables or binomial probabilities, we find that the probability of 7 or more successes is 0.196 + 0.196 + 0.133 + 0.092 + 0.042 + 0.014 + 0.003 = 0.696* and p = 0.5 and since this value is greater than α/2 = 0.025, we find that the null hypothesis will have to be accepted. We can also use normal approximation to the binomial distribution when np ≥ 5. As here p = ½, the condition for the normal approximation to the binomial distribution is satisfied as n > 10. As such, we can use the Z statistic for which the following formula is to be used.

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{X - (np)}{\sqrt{\frac{n}{4}}}$$

$$= \frac{7 - (15/2)}{\sqrt{\frac{15}{4}}} = \frac{\frac{14-15}{2}}{1.9365} = \frac{-0.5}{1.9365} = -0.26$$

Since calculated Z = -0.26 lies between Z = - 1.96 and Z = 1.96 (the critical value of Z at 0.05 level of significance), the null hypothesis is accepted.

The two-sample sign test

The sign test can be applied to problems that deal with paired data. In such problems, each pair can be replaced with a plus sign if the first value is greater than the second or a minus sign if the first value is smaller than the second. In case the two values in the pair turn out to be equal, these are discarded. These are essentially two kinds of situations: (a) the data are actually given as pairs and (b) the data comprise two independent samples that are randomly paired.

Example 2: *Suppose we have the following table indicating the ratings assigned to two brands of cold drink X and Y by 12 consumers. Each respondent was asked to taste the two brands of cold drink and then rate them.*

Table 11.1. Ratings of brands X and Y cold drinks

Brand X	26	30	44	23	18	50	34	16	25	49	37	20
Brand Y	22	27	39	7	11	56	30	14	18	51	33	16
Sign	+	+	+	+	+	-	+	+	+	-	+	+

We have to apply the two-sample sign test. H_0 being both brands enjoy equal preference.

Solution: Row three of Table 13.1 shows + and – signs. When X's rating is higher than that of Y, then the third row shows the '+' sign. As against this, when X's rating is lower than that of Y, then it shows the '-' sign. The table shows 10 plus signs and 2 minus signs. Now, we have to examine whether '10 successes in 12 trials' supports the null hypothesis $p = ½$ or the alternative hypothesis $p > ½$. The null hypothesis implies that both the brands enjoy equal preferences and none is better than the other. The alternative hypothesis is that the brand X is better than brand Y. Referring to the binomial probabilities table, we find that for $n = 12$ and $p = ½$ the probability of '10 or more successes' is $0.016 + 0.003 = 0.019$. It follows that the null hypothesis can be rejected at $\alpha = 0.05$ level of significance. We can, therefore, conclude that brand X is a preferred brand as compared to brand Y.

Example 3: To illustrate the second case, which relates to two independent samples, let us consider the following data pertaining to the downtimes (periods in which computers were inoperative on account of failures, in minutes of two different computers. We have to apply the two-sample sign test.

Computer A	58	60	42	62	65	59	60	52	50	75	59
	52	57	30	46	66	40	78	55	52	58	44
Computer B	32	48	50	41	45	40	43	43	70	60	80
	45	36	56	40	70	50	53	50	30	42	45

Solution: These data are shown in Table 13.2 along with + or – sign as may be applicable in case of each pair of values. A plus sign is assigned when the downtime for computer A is greater than that for computer B and a minus sign is given when the downtime for computer B is greater than that for computer A.

Table 11.2: Downtime of computers A and B (Minutes)

Computer A	58	60	42	62	65	59	60	52	50	75	59
Computer B	32	48	50	41	45	40	43	43	70	60	80
Sign	+	+	-	+	+	+	+	+	-	+	-
Computer A	52	57	30	46	66	40	78	55	52	58	44
Computer B	45	36	56	40	70	50	53	50	30	42	45
Sign	+	+	-	+	_	_	+	+	+	+	-

It will be seen that there are 13 plus signs and 7 minus signs. Thus, we have to ascertain whether '13 successes in 22 trials' support the null hypothesis p = ½. The null hypothesis implies that the true average downtime is the same for both the computers A and B. The alternative hypothesis is p ≠ ½. The null hypothesis implies that the true average downtime is the same for both the computers A and B. The alternative hypothesis is p = ½.

Let us use in this case the normal approximation of the binomial distribution. This can be done since np and n (1 – p) are both equal to 11 in this example. Substituting n = 22 and p = ½ into the formulas for the mean and the standard deviation of the binomial distribution, we get μ = np = 22 (½) = 11 and

$\sigma = \sqrt{np\,(1 - p)} = \sqrt{22.½.½} = 2.345$

Hence, $Z = (X – \mu)/\sigma = (13 – 11)/2.345 = 1.71$

Since this value of 1.71 falls between $– Z_{0.025} = - 1.96$ and $Z_{0.025} = 1.96$, we find that the null hypothesis cannot be rejected. This means that the downtime in the two computers is the same.

This seems to be surprising as we find that there are substantial differences. The two sample means, for example, are 55.5 for A and 48.6 for B. This example illustrates the point that at times the sign test can be quite a waste of information. It may also be noted that had the continuity correction been used, we would have obtained:

Z = 3.5/2.345 = 1.49

This would not have changed our earlier conclusion.

11.2.2 Median test for two independent samples

In order to perform this test, let us use our previous example, which pertains to the downtimes of the two computers. The median of the combined data is 52, which can easily be checked. There are 5 values below 52 and 13 values above it, in case of computer A. As regards computer B, the corresponding figures are 16 and 6. All this information is summarised in Table 13.3, which also indicates the totals of the rows and columns.

Table 11.3. Classification of downtime for computers A and B

	Below median	Above median	Total
Computer A	5	13	20
Computer B	16	6	22
Total	**21**	**21**	**42**

Our null hypothesis H_0 is that there is no difference in the median downtime for the two computers. The alternative hypothesis H_1 is that there is difference in the downtime of the two computers.

We now calculate the expected frequencies by the formula ($Row_i \times Column_i$)/Grand total. Thus, Table 11.4 shows both the observed and the expected frequencies. Of course, we could have obtained these results by arguing that half the values in each sample can be expected to fall above the median and the other half below it.

Table 11.4. Calculation of chi-square

Observed frequencies (O)	Expected frequencies (E)	O – E	$(O – E)^2$	$(O – E)^2/E$
5	10	-5	25	2.50
13	10	5	25	2.50
16	11	5	25	2.27
6	11	-5	25	2.27
			Total	**9.54**

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 9.54$$

The critical value of χ^2 at 0.05 level of significance for (2 – 1) (2 – 1) = 1 degree of freedom is 3.841 (χ^2-Table). Since the calculated value of χ^2 exceeds the critical value, the null hypothesis has to be rejected. In other words, there is no evidence to suggest that the downtime is the same in case of the two computers.

It may be recalled that in the previous example having the same data, the null hypothesis could not be rejected. In contrast, we find here that the two-sample median test has led to the rejection of the null hypothesis. This may be construed as evidence that the median test is not quite as wasteful of the information as the sign test. However, in general, it is very difficult to make a meaningful comparison of the merits of two or more non-parametric tests, which can be used for the same purpose.

11.3.1. The K-sample Median Test

The median test can easily be generalised so that it can be applied to K-samples. In accordance with the earlier procedure, first find the median of the combined data. We then determine how many of the

values in each sample fall above or below the median. Finally, we analyse the resulting contingency table by the method of chi-square. Let us take an example.

Example 4: Suppose that we are given the following data relating to marks obtained by students in Statistics in the three different sections of a MBA class in G.J.U. Hisar. The maximum marks were 100.

Section A	46	60	58	80	66	39	56	61	81	70
	75	48	43	64	57	59	87	50	73	62
Section B	60	55	82	70	46	63	88	69	61	43
	76	54	58	65	73	52				
Section C	74	67	37	80	72	92	19	52	70	40
	83	76	68	21	90	74	49	70	65	58

Test whether the differences among the three sample means are significant.

Solution: In case of such problems, analysis of variance is ordinarily performed. However, here we find that the data for Section C have much more variability as compared to the data for the other two sections. In view of this, it would be wrong to assume that the three population standard deviations are the same. This means that the method of one-way analysis of variance cannot be used.

In order to perform a median test, we should first determine the median of the combined data. This comes out to 63.5, as can easily be checked. Then we count how many of the marks in each sample fall below or above the median. Thus, the results obtained are shown in Table 11.5.

Table 11.5. Worksheet for calculating chi-square

	Below median	Above median
Section A	12	8
Section B	9	7
Section C	7	13

Since the corresponding expected frequencies for Section A are 10 and 10, for Section B are 8 and 8, and for Section C 10 and 10, we can obtain the value of chi-square. These calculations are shown below:

$$\chi^2 = \frac{(12-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-8)^2}{8} + \frac{(7-8)^2}{8} + \frac{(7-10)^2}{10} + \frac{(13-10)^2}{10}$$

$$= 0.4 + 0.4 + 0.125 + 0.125 + 0.9 + 0.9 = 2.85$$

Now, we have to compare this value with the critical value of χ^2 at 5 per cent level of significance. This value is 5.991 for 2 (K – 1 = 3 – 1) degrees of freedom (Chi-square Table). As the calculated value of χ^2 is less than the critical value, the null hypothesis cannot be rejected. In other words, we cannot conclude that there is a difference in the true average (median) marks obtained by the students in Statistics test from the three sections.

11.2.3 Wilcoxon

Wilcoxon matched-pairs test is an important non-parametric test, which can be used in various situations in the context of two related samples such as a study where husband and wife are matched or when the output of two similar machines are compared. In such cases we can determine both direction and magnitude of difference between matched values, using Wilcoxon matched-pairs test.

The procedure involved in using this test is simple. To begin with, the difference (d) between each pair of values is obtained. These differences are assigned ranks from the smallest to the largest, ignoring signs. The actual signs of differences are then put to corresponding ranks and the test statistic T is calculated, which happens to be the smaller of the two sums, namely, the sum of the negative ranks and the sum of the positive ranks.

There may arise two types of situations while using this test. One situation may arise when the two values of some matched-pair(s) is/are equal as a result the difference (d) between the values is zero. In such a case, we do not consider the pair(s) in the calculations. The other situation may arise when we get the same difference (d) in two or more pairs. In such a case, ranks are assigned to such pairs by averaging their rank positions. For instance, if two pairs have rank score of 8, then each pair is assigned 8.5 rank [(8 + 9)/2 = 8.5] and the next largest pair is assigned the rank 10.

After omitting the number of tied pairs, if the given number or matched pairs is equal to or less than 25, then the table of critical value T is used for testing the null hypothesis. When the calculated value of T is equal to or smaller than the table (i.e. critical) value at a desired level of significance, then the null hypothesis is rejected. In case the number exceeds 25, the sampling distribution of T is taken as approximately normal with mean $\mu_T = n (n + 1)/\mu$ and standard deviation

$$\sigma_T = \sqrt{n (n + 1) (2n + 1)/24}$$

where n is taken as the number of given matched pairs- number of tied pairs omitted, if any. In such a situation, the test Z statistic is worked out as follows:

$Z = (T - \mu_r)/\sigma_r$

Let us now take an example to illustrate the application of Wilcoxon matched-pairs test.

Example 5: The management of the Punjab National Bank wants to test the effectiveness of an advertising company that is intending to enhance the awareness of the bank's service features. It administered a questionnaire before the advertising campaign, designed to measure the awareness of services offered. After the advertising campaign, the bank administered the same questionnaire to the same group of people. Both the before and after advertising campaign scores are given in the following table.

Consumer awareness of bank services offered

Consumer	1	2	3	4	5	6	7	8	9	10
Before ad campaign	82	81	89	74	68	80	77	66	77	75
After ad campaign	87	84	84	76	78	81	79	81	81	83

Using Wilcoxon matched-pairs test, test the hypothesis that there is no difference in awareness of services offered after the advertising campaign.

Solution:

Table 11.6. Application of Wilcoxon matched-pairs test

Consumer	After Ad campn.	Before Ad campn.	Diff. d_i	Rank of d_i	Rank (-) sign	Rank (+) sign
1	87	82	5	6.5		6.5
2	84	81	3	4		4
3	84	89	-5	6.5	-6.5	
4	76	74	2	2.5		2.5
5	78	68	10	9		9
6	81	80	1	1		1
7	79	77	2	2.5		2.5

8	81	66	13	10		10
9	81	77	4	5		5
10	83	75	8	8		8
				Total	**-6.5**	**+48.5**

Null hypothesis H_0: There is no difference in the awareness of bank services after the ad campaign. Alternative hypothesis H_1: There is a difference in the awareness of bank services after the ad campaign.

Computed 'T' value is 6.5. The critical value of T for n = 10 at 5 per cent level of significance is 8 (Area Table). Since the computed T value is less than the critical T value, the null hypothesis is rejected. We can conclude that after the ad campaign there is difference in the consumer awareness of the bank's services needs some explanation. Had there been no difference in the awareness before and after the ad campaigns, the sum of positive and negative ranks would have been almost equal. However, if the difference between the two series being compared is larger, then the value of T will tend to be smaller as it is defined as smaller of ranks. This is the case we find in this problem. It may be noted that with this test the calculated value of T must be smaller than the critical value in order to reject the null hypothesis.

11.2.4 The Mann-Whitney U Test

One of the most common and best known distribution-free tests is the Mann-Whitney test for two independent samples. The logical basis of this test is particularly easy to understand. Suppose we have two independent treatment groups, with n_1 observations in Group 1 and n_2 observations in Group 2. Now, we assume that the population from which Group 1 scores have been sampled contained generally lower values than the population from which Group 2 scores were drawn. If we were to rank these scores disregarding the group to which they belong then the lower ranks would generally fall to Group 1 scores and the higher ranks would generally fall to Group 2 scores. Proceeding one step further, if we were to add together the ranks assigned to each group, the sum of the ranks in Group 1 would be expected to be considerably smaller than the sum of the ranks in Group 2. This would result in the rejection of the null hypothesis.

Let us now take another situation where the null hypothesis is true and the scores for the two groups are sampled from identical populations. If we were to rank all N scores regardless of the group, we would expect a mix of low and high ranks in each group. Thus, the sum of the ranks assigned to Group 1 would be broadly equal to the sum of the ranks assigned to Group 2.

The Mann-Whitney test is based on the logic just described, using the sum of the ranks in one of the groups as the test statistic. In case that sum turns out to be too small as compared to the other sum, the null hypothesis is rejected. The common practice is to take the sum of the ranks assigned to the smaller group, or if $n_1 = n_2$, the smaller of the two sums as the test statistic. This value is then compared with the critical value that can be obtained from the table of the Mann-Whitney statistic (W_s) to test the null hypothesis.

Let us take an example to illustrate the application of this test.

Example 6: The following data indicate the lifetime (in hours) of samples of two kinds of light bulbs in continuous use:

Brand A	603	625	641	622	585	593	660	600	633	580	613	648
Brand B	620	640	646	620	652	639	590	646	631	669	610	619

We are required to use the Mann-Whitney test to compare the lifetimes of brands A and B light bulbs.

Solution: The first step for performing the Mann-Whitney test is to rank the given data *jointly* (as if they were one sample) in an increasing or decreasing order of magnitude. For our data, we thus obtain the following array where we use the letters A and B to denote whether the light bulb was from brand A or brand B.

Table 11.7. Ranking of light bulbs of brands A and B

Sample score	Group	Rank	Sample score	Group	Rank
580	A	1	625	A	13
585	A	2	631	B	14
590	B	3	633	A	13
593	A	4	639	B	16
600	A	5	640	B	17
603	A	6	641	A	18

610	B	7	646	B	19.5
613	A	8	646	B	19.5
619	B	9	648	A	21
620	B	10.5	652	B	22
620	B	10.5	660	A	23
622	A	12	669	B	24

As both the samples come from identical populations, it is reasonable to assume that the means of the ranks assigned to the values of the two samples are more or less the same. As such, our null hypothesis is:

H_0: Means of ranks assigned to the values in the two groups are the same.

H_1: Means are not the same.

However, instead of using the means of the ranks, we shall use *rank sums* for which the following formula will be used.

$U = n_1n_2 + [n_1(n_1 + 1)]/2 – R_1$

Where n_1 and n_2 are the sample sizes of Group 1 and Group 2, respectively, and R_1 is the sum of the ranks assigned to the values of the first sample. In our example, we have $n_1 = 12$, $n_2 = 12$ and $R_1 = 1 + 2 + 4 + 5 + 6 + 8 + 12 + 13 + 13 + 18 + 21 + 23 = 128$. Substituting these values in the above formula,

$U = (12)(12) + [12(12 + 1)]/2 – 128$

$= 144 + 78 – 128$

$= 94$

From Appendix Table 9 for n_1 and n_2, each equal to 12, and for 0.05 level of significance is 37. Since the critical value is smaller than the calculated value of 94, we accept the null hypothesis and conclude that there is no difference in the average lifetimes of the two brands of light bulbs.

The test statistic we have just applied is suitable when n_1 and n_2 are less than or equal to 25. For larger values of n_1 and/or n_2, we can make use of the fact that the distribution of W_s approaches a normal distribution as sample sizes increase. We can then use the Z test to test the hypothesis.

The Normal Approximation

Although our observations are limited, we may apply the normal approximation to this problem. For this, we have to use the Z statistic.

1. Mean = μu = $[(N_1N_2/2] = [(12 \times 12)/2] = 72$

2. Standard error = $\sqrt{\frac{n_1n_2\,(n_1 + n_2 + 1)}{12}}$

$= \sqrt{\frac{12 \times 12\,(12 + 12 + 1)}{12}}$

$= \sqrt{300} = 17.3$

3. (Statistic – Mean)/Standard deviation

= (94 – 72)/17.3 = 1.27

The critical value of Z at 0.05 level of significance is 1.64. Since the calculated value of Z = 1.27 is smaller than 1.64, the null hypothesis is accepted. This shows that there is no difference in average lifetimes of brands A and B bulbs. The Z test is more dependable as compared to the earlier one. It may be noted that Mann-Whitney test required fewer assumptions than the corresponding standard test. In fact, the only assumption required is that the populations from which samples have been drawn are continuous. In actual practice, even when this assumption turns out to be wrong, this is not regarded serious.

11.2.5 The Kruskal-Wallis test

This test is used to determine whether k independent samples can be regarded to have been obtained from identical populations with respect to their means. The Kruskal-Wallis Test is the non-parametric counter part of the one-way analysis of variance. The assumption of the F-test, used in analysis of variance, was that each of the k populations should be normal with equal variance. In contrast to this, the Kruskal-Wallis test only assumes that the k populations are continuous and have the same pattern (symmetrical or skewed) of distribution. The null and the alternative hypotheses of the Kruskal-Wallis test are:

H_0: $m_1 = m_2 = \ldots = m_k$ (i.e., means of the k populations are equal)

H_a: Not all m_i's are equal.

The Test Statistic: The computation of the test statistic follows a procedure that is very similar to the Mann-Whitney Wilcoxon test.

(i) Rank all the $n_1 + n_2 + \ldots + n_k = n$ observations, arrayed in ascending order.

(ii) Find R_1, $R_{2, \ldots} R_k$, where R_i is the sum of ranks of the ith sample.

The test statistic, denoted by H, is given by

$$H = \frac{12}{n(n+1)} \left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \ldots + \frac{R_k^2}{n_k} \right) - 3(n+1).$$

It can be shown that the distribution of H is c^2 with k – 1 d.f., when size of each sample is at least 5. Thus, if $H > \chi^2_{k-1}$, H_0 is rejected.

Example 7: To compare the effectiveness of three types of weight-reducing diets, a homogeneous groups of 22 women was divided into three sub-groups and each sub-group followed one of these diet plans for a period of two months. The weight reductions, in kgs, were noted as given below:

	I	4.3	3.2	2.7	6.2	5.0	3.9			
Diet Plans	II	5.3	7.4	8.3	5.5	6.7	7.2	8.5		
	III	1.4	2.1	2.7	3.1	1.5	0.7	4.3	3.5	0.3

Use the Kruskal-Wallis test to test the hypothesis that the effectiveness of the three weight reducing diet plans is same at 5% level of significance.

Solution:

It is given that $n_1 = 6$, $n_2 = 7$ and $n_3 = 9$.

The total number of observations is 6 + 7 + 9 = 22. These are ranked in their ascending order as given below:

Diet Plans	I	12.5	9	6.5	17	14	11		70
	II	13	20	21	16	18	19	22	131
	III	3	5 1	6.5	8	4	2	12.5 10	52

From the above table, we get $R_1 = 70$, $R_2 = 131$ and $R_3 = 52$.

$$\therefore H = \frac{12}{22 \times 23} \left(\frac{70^2}{6} + \frac{131^2}{7} + \frac{52^2}{9} \right) - 3 \times 23 = 13.63$$

The tabulated value of χ^2 at 2 d.f. and 5% level of significance is 5.99. Since H is greater than this value, H_0 is rejected at 5% level of significance.

11.2.6 The spearman's rank correlation test

The Spearman's Rank Correlation $r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$, can be used to test the significance of correlation in population. We can write H_0: $r_s = 0$, where r_s is the coefficient of correlation in population.

The test statistic: It can be shown that for $n \geq 10$, the distribution of r_s, under H_0, is approximately normal with mean 0 and standard error $\frac{1}{\sqrt{n-1}}$. Thus, $z = r_s\sqrt{n-1}$ is a standard normal variate.

Example 8: Twelve entries in a painting competition were ranked by two judges, as shown below:

Entry:	A	B	C	D	E	F	G	H	I	J	K	L
Judge I:	5	2	3	5	1	6	8	7	10	9	12	11
Judge II:	4	5	2	1	6	7	10	9	11	12	3	8

Test the hypothesis that coefficient of rank correlation in population is positive.

Solution: We have to test H_0:$\sigma_s \leq 0$ against H_a:$\sigma_s > 0$.

From the given data, we can find $d_1 = R_u - R_{2i}$ and then $\sum d_i^2 = 134$.

$\therefore\ r_s = 1 - \frac{6 \times 154}{12 \times 143} = 0.46$ and $z = 0.46\sqrt{11} = 1.53$.

Since the value of z is less than 1.645, there is no evidence against H_0 at 5% level of significance. Hence, the correlation in population cannot be regarded as positive.

11.2.7 The Median Test for Randomness

Any sample comprising numerical observations can be treated in the same manner by using the letters a and b to denote, respectively, values above the median and values below the median of the sample. In case an observation is equal to the median, it is omitted. The resulting series of as and bs (representing the data in their original order) can be tested for randomness on the basis of the total number of runs above and below the median, respectively. Let us take an example.

Example 9: Suppose we have the following series of 29 college students. After performing a set of study exercises, increases in their pulse rate were recorded as follows:

22, 23, 21, 25, 33, 32, 25, 30, 17, 20, 26, 12, 21, 20, 27, 24, 28, 14, 29, 23, 22, 36, 25, 21, 23, 19, 17, 26 and 26.

We have to test the randomness of these data.

Solution: First, we have to calculate the median of this series. If we arrange these values in an ascending order, we find that the size of (n +1)/2th item, that is, 13th item is 24. Thus, the median is 24. As there is one value, which is 24 we omit it and get the following arrangement of as and bs where a stands for an item greater than (or above) the median and b stands for an item lower than (or below) the median:

bbb aaaaa bb a bbb aa b a bb aa bbbb aa

On the basis of this arrangement, we find that n_1, (i.e. a) = 13, n_2, (i.e. bs) = 13, and u = 12, we get

$\mu_r = [(2n_1n_2)/(n_1 + n_2)] + 1$

$= [(2 \times 13 \times 13)/(13 + 13)] + 1 = (390/28) + 1 = 14.93$

$$\sigma_u = \sqrt{\frac{2n_1n_2\,(2n_1n_2 - n_1 - n_2)}{(n_1n_2)^2\,(n_1 + n_2 - 1)}}$$

$$\sigma_u = \sqrt{\frac{(2 \times 13 \times 15)\,(2 \times 13 \times 15 - 13 - 15)}{(13 + 15)^2\,(13 + 15 - 1)}}$$

$$= \sqrt{\frac{390 \times 362}{(28)^2\,(27)}}$$

$$= \sqrt{\frac{141180}{21168}} = \sqrt{6.6695} = 2.58$$

$Z = (u - \mu_r)/\sigma_u = (12 - 14.93)/2.58 = -2.93/2.58 = -1.14$

Since Z = -1.14 falls between $-Z_{0.025}$ = -1.96 and $Z_{0.025}$ = 1.96, the null hypothesis cannot be rejected at the level of significance α = 0.05. We can, therefore, conclude that the randomness of the original sample cannot be questioned.

It may be noted that this test is particularly useful in detecting trends and cyclic patterns in a series. If there is a trend, there would be first mostly as and later mostly bs or vice versa. In case of a repeated cyclic pattern, there would be a systematic alternation of as and bs and probably, too many runs.

11.2.8 Kolmogorov-Smirnov one-sample test

This test is concerned with the degrees of agreement between a set of observed values and the values specified by the null hypothesis. It is similar to the chi-square test of goodness-of-fit. It is used when one is interested in comparing a set of values on an ordinal scale. Let us take an example.

Example 10: Suppose that a company has conducted a field survey covering 200 respondents. Apart from other questions, it asked the respondents to indicate on a 5-point scale how much the durability of a particular product is important to them. The respondents indicated as follows:

Very important	50
Somewhat important	60
Neither important nor unimportant	20
Somewhat unimportant	40
Very unimportant	30
Total respondents	**200**

We have been asked to use the Kolmogorov-Smirnov test to test the hypothesis that there is no difference in importance ratings for durability among the respondents.

Solution: In order to apply the Kolmogorov-Smirnov test to the above data, first of all we should have the cumulative frequency distribution from the sample. Second, we have to establish the cumulative frequency distribution, which would be expected on the basis of the null hypothesis. Third, we have to determine the largest absolute deviation between the two distributions mentioned above. Finally, this value is to be compared with the critical value to ascertain its significance.

Table 11.8 shows the calculations.

Table 11.8: Worksheet for the Kolmogorov-Smirnov D

Importance of durability	Observed number	Observed proportion	Observed cumulative proportion	Null proportion	Null cumulative proportion	Absolute difference observed and Null
Very important	50	0.25	0.25	0.2	0.2	0.05
Somewhat important	60	0.30	0.55	0.2	0.4	0.13

Neither important nor unimportant	20	0.10	0.65	0.2	0.6	0.05
Somewhat unimportant	40	0.20	0.85	0.2	0.8	0.05
Very unimportant	30	0.13	1.00	0.2	1.0	0.00

From Table 11.8, we find that the largest absolute difference is 0.13, which is known as the Kolmogorov-Smirnov D value. For a sample size of more than 35, the critical value of D at an $\alpha = 0.05$ is $1.36/\sqrt{n}$. As sample size in this example is 200, $D = 1.36/\sqrt{200} = 0.096$. As the calculated D exceeds the critical value of 0.096, the null hypothesis that there is no difference in importance ratings for durability among the respondents is rejected.

Although there are a number of non-parametric tests, we have presented some of the more frequently used tests in this chapter. While using these tests, we must know that the advantages we derive by limiting our assumptions may be offset by the loss in the power of such tests. However, when basic assumptions as required for parametric tests are valid, the use of non-parametric tests may lead to a false hypothesis and thus we may commit a Type II error. We have to consider this aspect very carefully before deciding in favour of non-parametric tests. It may be reiterated that such tests are more suitable in case of ranked, scaled or rated data.

11.3 CHECK YOUR PROGRESS

1. Sign test is based on the ..... of the plus or minus signs of observations in a sample instead of their numerical values.
2. We can determine both..... of difference between matched values, using Wilcoxon matched-pairs test.
3. In Mann-Whitney test, the only assumption required is that the populations from which samples have been drawn are ..... .
4. The Kruskal-Wallis Test is the ..... counter part of the one-way analysis of variance.
5. Kolmogorov-Smirnov test is used when one is interested in comparing a set of values on a ..... .

11.4 SUMMARY

In contrast to parametric tests, non-parametric tests do not require any assumptions about the parameters or about the nature of population. It is because of this that these methods are sometimes referred to as the distribution free methods. Most of these methods, however, are based upon the weaker assumptions that observations are independent and that the variable under study is continuous with approximately symmetrical distribution. In addition to this, these methods do not require measurements as strong as that required by parametric methods. Most of the non-parametric tests are applicable to data measured in an ordinal or nominal scale. As opposed to this, the parametric tests are based on data measured at least in an interval scale. The measurements obtained on interval and ratio scale are also known as high level measurements. It should be noted here that a test that can be performed on high level measurements can always be performed on ordinal or nominal measurements but not vice-versa. However, if along with the high level measurements the conditions of a parametric test are also met, the parametric test should invariably be used because this test is most powerful in the given circumstances. From the above, we conclude that a non-parametric test should be used when either the conditions about the parent population are not met or the level of measurements is inadequate for a parametric test. There are different advantages and disadvantages of it. There are different test for it like sign test, Wilcoxon test, Kolmogorov-Smirnov D etc.

11.5 KEYWORDS

Non-parametric tests: Tests that rely less on parameter estimation and/or assumptions about the shape of a population distribution.

One-Sample Runs test: A non-parametric test used for determining whether the items in a sample have been selected randomly.

Run: A sequence of identical occurrences that may be preceded and followed by different occurrences. At times, they may not be preceded or followed by any occurrences.

Sign test: A non-parametric test that takes into account the difference between paired observations where plus (+) and minus (-) signs are substituted for quantitative values.

Theory of runs: A theory concerned with the testing of samples for the randomness of the order in which they have been selected.

Wilcoxon Matched-pairs Test (or Signed Rank Test): A non-parametric test that can be used in various situations in the context of two related samples.

Kolmogorov-Smirnov test: A non-parametric test that is concerned with the degrees of agreement between a set of observed ranks (sample values) and a theoretical frequency distribution.

Kurskal-Wallis test: A non-parametric method for testing the null hypothesis that K independent random samples come from identical populations. It is a direct generalisation of the Mann-Whitney test.

Mann-Whitney U test: A non-parametric test that is used to determine whether two different samples come from identical populations or whether these populations have different means.

11.6 SELF-ASSESSMENT TEST

1. What do you understand by non-parametric or distribution free methods?
2. What are the major advantages of non-parametric methods over parametric methods?
3. What are the main limitations of non-parametric tests?
4. Enumerate the different non-parametric tests and explain any two of them.
5. The sequence of occurrence of 'zeros' and 'ones' in a message sent in a digital code is shown below. Test at 5 per cent whether the sequence of '0' and '1' is random 00110 11011 00001 11100 00110 11001 11110 00011 00100 11000 11100 00011 00111 11100 00000 11111 10001 11000 10001 01110.
6. The proprietor of a small business computed his average earnings per day over a period of 12 days. For each day, an L was recorded if the earnings were less than the average, otherwise an M was recorded. These data are given below:

   L L L L M M L L L L M M

7. In a metropolitan city, a city bus service was scheduled to reach a major bus stop at 11 a.m. each day. If the bus reached that stop within 5 minutes of 11 a.m. it was considered to be on time. Over a 13-day period, an A was recorded if the bus was on time, otherwise a B was recorded. The picture that emerged after ten days was as follows:

   A A B A B B A B A A B B B A A

8. The following data show employees' rate of substandard performance before and after a new incentive scheme. Determine whether the introduction of the new incentive scheme has reduced the substandard performance at 0.05 level of significance.

Before	7	8	5	9	10	6	5	9	6	8
After	5	6	7	6	8	7	6	6	5	7

9. A company manufacturing electronic toys toys has recently been taken over by another company. Prior to the takeover of the company, certain workers were approached to ascertain their satisfaction levels. The same workers were again approached to know their satisfaction level after the takeover of the company. The two sets of data are given below.

Before	69	73	58	76	82	65	75	64	87	70
After	65	75	63	75	82	68	71	65	85	68

Using an appropriate test, find out whether there has been an improvement in the satisfaction level of workers after the takeover of their company by a new company

10. The following data relate to the costs of building comparable lots in the two Resons A and B (in million rupees):

Resort A	30.9	32.5	44.3	39.5	35.0	48.9
Resort B	53.9	61.0	36.0	42.5	40.9	47.9

The company owning the resort area A claimed that the median price of building lots was less in area A as compared to resort area B. You are asked to test this claim, using a nonparametric test with a 1 per cent level of significance.

11. On 13 different days, A had to wait for the city bus to reach his office as shown below:

17, 12, 18, 20, 25, 30, 10, 13, 7, 10, 9, 11, 5, 11 and 20 minutes.

Use the sign test at 5 per cent level of significance to test the bus company's claim that on an average A should not have to wait for more than 13 minutes.

12. A company used three different methods of advertising its product in three cities. It later found the increased sales (in thousand rupees) in identical retail outlets in the three cities as follows:

City A	70	58	60	45	55	62	80	72	
City B	65	57	48	55	75	68	45	52	63
City C	53	59	71	70	63	60	58	75	

Use Kruskal-Wallis method to test the hypothesis that the mean increase in sales on account of three different methods of advertising was the same in the retail outlets in A, B and C cities. Use 5 per cent level of significance.

11.7 ANSWERS TO CHECK YOUR PROGRESS

1. Direction
2. Direction and magnitude
3. Continuous
4. Non-parametric
5. Ordinal scale.

11.8 REFERENCES/SUGGESTED READINGS

1. Spiegel, Murray R.: Theory and Practical of Statistics, London McGraw Hill Book Company.
2. Yamane, T.: Statistics: An Introductory Analysis, New York, Harpered Row Publication
3. R.P. Hooda: Statistic for Economic and Management McMillan India Ltd.
4. G.C. Beri: Statistics for Mgt., TMA
5. J.K. Sharma: Business Statistics, Pearson Education
6. S.P. Gupta : Statistical Methods, Sultan Chand and Sons.

Subject: Business Statistics	
Course code: MBA-102	Author: Anil Kumar
Lesson: 12	Vetter : Prof. Harbhajan Bansal
CORRELATION ANALYSIS	

STRUCTURE

12.1 Learning Objectives

12.2 Introduction

12.2.1 What is Correlation?

12.3 Correlation Analysis

12.3.1 Scatter Diagram

12.3.2 Correlation Graph

12.3.3 Pearson's Coefficient of Correlation

12.3.4 Spearman's Rank Correlation

12.3.5 Concurrent Deviation Method

12.4 Limitations of Correlation Analysis

12.5 Partial Correlation

12.5.1 Zero Order Correlation

12.5.2 First Order Correlation

12.5.3 Second Order Correlation

12.5.4 Limitations of Partial Correlation

12.6 Multiple Correlation

12.6.1 Calculation of Multiple Correlation Coefficient

12.6.2 Limitations of Multiple Correlation

12.7 Check your Progress

12.8 Summary

12.9 Keywords

12.10 Self-Assessment Test

12.11 Answers to check your progress

12.12 References/Suggested Readings

12.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understand the bivariate linear correlation
- Understand the importance of correlation analysis
- Understand the limitations of correlation analysis

12.2 INTRODUCTION

...if we have information on more than one variables, we might be interested in seeing if there is any connection - any association - between them.

Statistical methods of measures of central tendency, dispersion, skewness and kurtosis are helpful for the purpose of comparison and analysis of distributions involving only one variable *i.e.* univariate distributions. However, describing the relationship between two or more variables, is another important part of statistics.

In many business research situations, the key to decision making lies in understanding the relationships between two or more variables. *For example,* in an effort to predict the behavior of the bond market, a broker might find it useful to know whether the interest rate of bonds is related to the prime interest rate. While studying the effect of advertising on sales, an account executive may find it useful to know whether there is a strong relationship between advertising dollars and sales dollars for a company.

The statistical methods of *Correlation* (discussed in the present lesson) and *Regression* (to be discussed in the next lesson) are helpful in knowing the relationship between two or more variables which may be related in same way, *like* interest rate of bonds and prime interest rate; advertising expenditure and sales; income and consumption; crop-yield and fertilizer used; height and weights and so on.

In all these cases involving two or more variables, we may be interested in seeing:

- if there is any association between the variables;
- if there is an association, is it strong enough to be useful;
- if so, what form the relationship between the two variables takes;
- how we can make use of that relationship for predictive purposes, that is, forecasting; and
- how good such predictions will be.

Since these issues are inter related, correlation and regression analysis, as two sides of a single process, consists of methods of examining the relationship between two or more variables. If two (or more) variables are correlated, we can use information about one (or more) variable(s) to predict the value of the other variable(s), and can measure the error of estimations - *a job of regression analysis.*

12.2.1 WHAT IS CORRELATION?

Correlation is a measure of association between two or more variables. When two or more variables very in sympathy so that movement in one tends to be accompanied by corresponding movements in the other variable(s), they are said to be correlated.

"The correlation between variables is a measure of the nature and degree of association between the variables".

As *a measure of the degree of relatedness of two variables,* correlation is widely used in exploratory research when the objective is to locate variables that might be related in some way to the variable of interest.

TYPES OF CORRELATION

Correlation can be classified in several ways. The important ways of classifying correlation are:

(i) Positive and negative,

(ii) Linear and non-linear (curvilinear) and

(iii) Simple, partial and multiple.

Positive and Negative Correlation

If both the variables move in the same direction, we say that there is a positive correlation, *i.e.*, if one variable increases, the other variable also increases on an average or if one variable decreases, the other variable also decreases on an average.

On the other hand, if the variables are varying in opposite direction, we say that it is a case of negative correlation; *e.g.,* movements of demand and supply.

Linear and Non-linear (Curvilinear) Correlation

If the change in one variable is accompanied by change in another variable in a constant ratio, it is a case of linear correlation. Observe the following data:

X	:	10	20	30	40	50
Y	:	25	50	75	100	125

The ratio of change in the above example is the same. It is, thus, a case of linear correlation. If we plot these variables on graph paper, all the points will fall on the same straight line.

On the other hand, if the amount of change in one variable does not follow a constant ratio with the change in another variable, it is a case of non-linear or curvilinear correlation. If a couple of figures in either series *X* or series *Y* are changed, it would give a non-linear correlation.

Simple, Partial and Multiple Correlation

The distinction amongst these three types of correlation depends upon the number of variables involved in a study. If only two variables are involved in a study, then the correlation is said to be simple correlation. When three or more variables are involved in a study, then it is a problem of either partial or multiple correlation. In multiple correlation, three or more variables are studied simultaneously. But in partial correlation we consider only two variables influencing each other while the effect of other variable(s) is held constant.

Suppose we have a problem comprising three variables *X, Y* and *Z. X* is the number of hours studied, *Y* is I.Q. and *Z* is the number of marks obtained in the examination. In a multiple correlation, we will study the relationship between the marks obtained *(Z)* and the two variables, number of hours studied *(X)* and I.Q. (*Y*). In contrast, when we study the relationship between *X* and *Z,* keeping an average I.Q. (*Y*) as constant, it is said to be a study involving partial correlation. In this lesson, we will study linear correlation between two variables.

CORRELATION DOES NOT NECESSARILY MEAN CAUSATION

The correlation analysis, in discovering the nature and degree of relationship between variables, does not necessarily imply any cause and effect relationship between the variables. Two variables may be related to each other but this does not mean that one variable causes the other. *For example*, we may find that logical reasoning and creativity are correlated, but that does not mean if we could increase peoples' logical reasoning ability, we would produce greater creativity. We need to conduct an actual experiment to unequivocally demonstrate a causal relationship. But if it is true that influencing someones' logical reasoning ability does influence their creativity, then the two variables must be correlated with each other. In other words, *causation always implies correlation, however converse is not true.* Let us see some situations:

1. The correlation may be due to chance particularly when the data pertain to a small sample. A small sample bivariate series may show the relationship but such a relationship may not exist in the universe.
2. It is possible that both the variables are influenced by one or more other variables. For example, expenditure on food and entertainment for a given number of households show a positive relationship because both have increased over time. But, this is due to rise in family incomes over the same period. In other words, the two variables have been influenced by another variable - increase in family incomes.
3. There may be another situation where both the variables may be influencing each other so that we cannot say which is the cause and which is the effect. *For example,* take the case of price and demand. The rise in price of a commodity may lead to a decline in the demand for it. Here, price is the cause and the demand is the effect. In yet another situation, an increase in demand may lead to a

rise in price. Here, the demand is the cause while price is the effect, which is just the reverse of the earlier situation. In such situations, it is difficult to identify which variable is causing the effect on which variable, as both are influencing each other.

The foregoing discussion clearly shows that correlation does not indicate any causation or functional relationship. Correlation coefficient is merely a mathematical relationship and this has nothing to do with cause and effect relation. It only reveals co-variation between two variables. Even when there is no cause-and-effect relationship in bivariate series and one interprets the relationship as causal, such a correlation is called spurious or non-sense correlation. Obviously, this will be misleading. As such, one has to be very careful in correlation exercises and look into other relevant factors before concluding a cause-and-effect relationship.

12.3 CORRELATION ANALYSIS

Correlation Analysis is a statistical technique used to indicate the nature and degree of relationship existing between one variable and the other(s). It is also used along with regression analysis to measure how well the regression line explains the variations of the dependent variable with the independent variable.

The commonly used methods for studying linear relationship between two variables involve both graphic and algebraic methods. Some of the widely used methods include:

1. Scatter Diagram
2. Correlation Graph
3. Pearson's Coefficient of Correlation
4. Spearman's Rank Correlation
5. Concurrent Deviation Method

12.3.1 SCATTER DIAGRAM

This method is also known as Dotogram or Dot diagram. Scatter diagram is one of the simplest methods of diagrammatic representation of a bivariate distribution. Under this method, both the variables are plotted on the graph paper by putting dots. The diagram so obtained is called "Scatter Diagram". By

studying diagram, we can have rough idea about the nature and degree of relationship between two variables. The term scatter refers to the spreading of dots on the graph. We should keep the following points in mind while interpreting correlation:

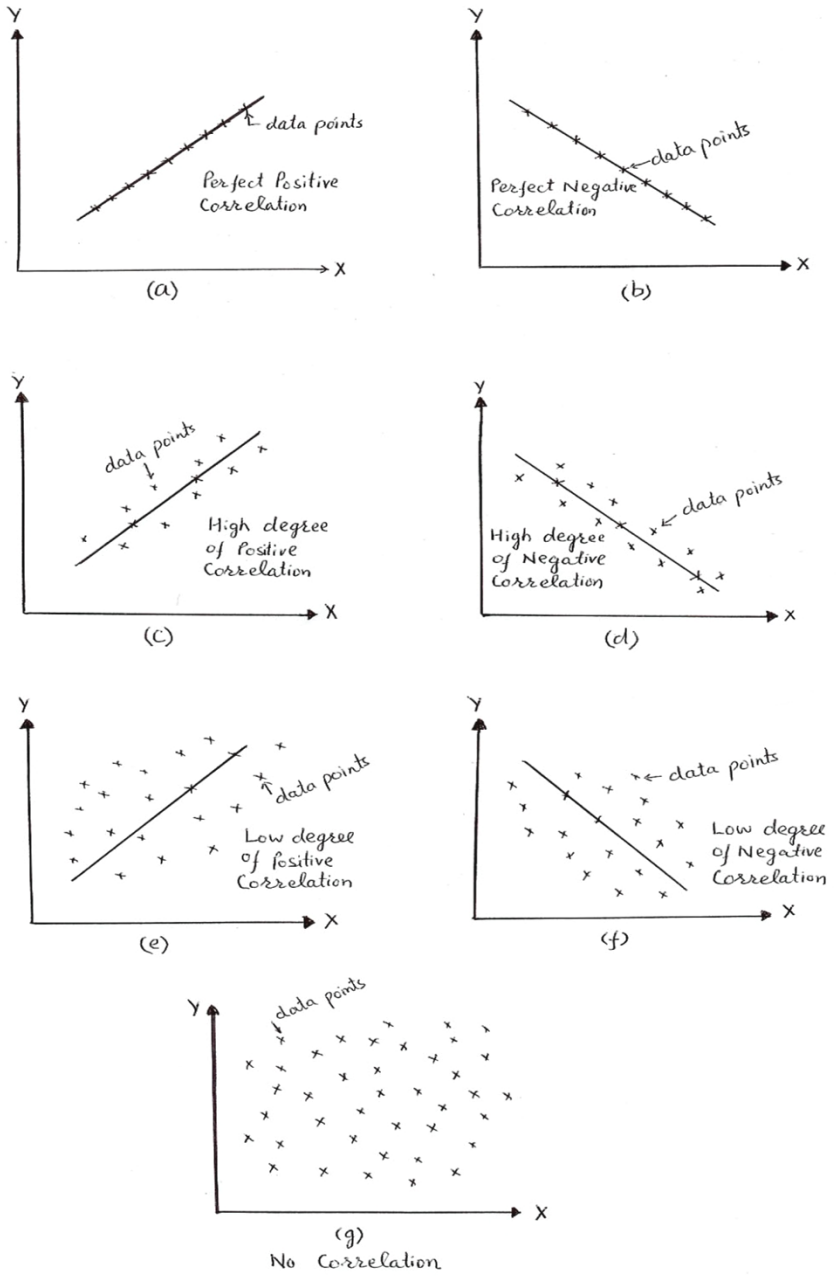


Figure 12.1 Scatter Diagrams

- if the plotted points are very close to each other, it indicates high degree of correlation. If the plotted points are away from each other, it indicates low degree of correlation.
- if the points on the diagram reveal any trend (either upward or downward), the variables are said to be correlated and if no trend is revealed, the variables are uncorrelated.
- if there is an upward trend rising from lower left hand corner and going upward to the upper right hand corner, the correlation is positive since this reveals that the values of the two variables move in the same direction. If, on the other hand, the points depict a downward trend from the upper left hand corner to the lower right hand corner, the correlation is negative since in this case the values of the two variables move in the opposite directions.
- in particular, if all the points lie on a straight line starting from the left bottom and going up towards the right top, the correlation is perfect and positive, and if all the points like on a straight line starting from left top and coming down to right bottom, the correlation is perfect and negative.

The various diagrams of the scattered data in Figure 12.1 depict different forms of correlation.

Example 12-1

Given the following data on sales (in thousand units) and expenses (in thousand rupees) of a firm for 10 month:

Month :	J	F	M	A	M	J	J	A	S	O
Sales:	50	50	55	60	62	65	68	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

a) Make a Scatter Diagram

b) Do you think that there is a correlation between sales and expenses of the firm? Is it positive or negative? Is it high or low?

Solution:(a) The Scatter Diagram of the given data is shown in Figure 4-2

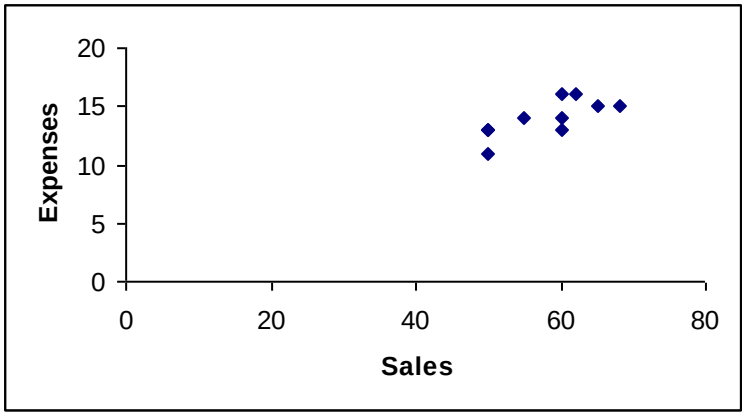


Figure 12.2 Scatter Diagram

(b) Figure 12.2 shows that the plotted points are close to each other and reveal an upward trend. So there is a high degree of positive correlation between sales and expenses of the firm.

12.3.2 CORRELATION GRAPH

This method, also known as Correlogram is very simple. The data pertaining to two series are plotted on a graph sheet. We can find out the correlation by examining the direction and closeness of two curves. If both the curves drawn on the graph are moving in the same direction, it is a case of positive correlation. On the other hand, if both the curves are moving in opposite direction, correlation is said to be negative. If the graph does not show any definite pattern on account of erratic fluctuations in the curves, then it shows an absence of correlation.

Example 12.2

Find out graphically, if there is any correlation between price yield per plot (qtls); denoted by *Y* and quantity of fertilizer used (kg); denote by *X*.

Plot No.:	1	2	3	4	5	6	7	8	9	10
Y:	3.5	4.3	5.2	5.8	6.4	7.3	7.2	7.5	7.8	8.3
X:	6	8	9	12	10	15	17	20	18	24

Solution: The Correlogram of the given data is shown in Figure 4-3

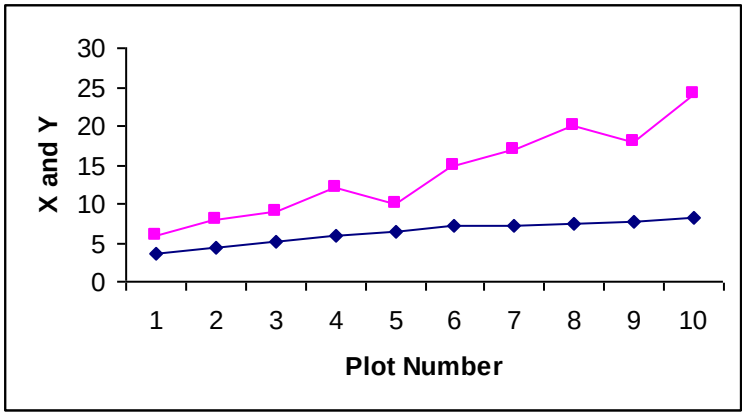


Figure 12.3 Correlation Graph

Figure 12.3 shows that the two curves move in the same direction and, moreover, they are very close to each other, suggesting a close relationship between price yield per plot (qtls) and quantity of fertilizer used (kg)

Remark: Both the Graphic methods - scatter diagram and correlation graph provide a *'feel for'* of the data – by providing visual representation of the association between the variables. These are readily comprehensible and enable us to form a fairly good, though rough idea of the nature and degree of the relationship between the two variables. However, these methods are unable to quantify the relationship between them. To quantify the extent of correlation, we make use of algebraic methods - which calculate correlation coefficient.

12.3.3 PEARSON'S COEFFICIENT OF CORRELATION

A mathematical method for measuring the intensity or the magnitude of *linear relationship* between two variables was suggested by Karl Pearson (1867-1936), a great British Biometrician and Statistician and, it is by far the most widely used method in practice.

Karl Pearson's measure, known as Pearsonian correlation coefficient between two variables *X* and *Y*, usually denoted by *r(X,Y)* or r_{xy} or simply *r* is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between *X* and *Y*, to the product of the standard deviations of *X* and *Y*.

Symbolically

$$r_{xy} = \frac{Cov(X,Y)}{S_x . S_y} \qquad \ldots\ldots\ldots\ldots (4.1)$$

when, $(X_1,Y_1);(X_2,Y_2);\ldots\ldots\ldots\ldots(X_n,Y_n)$ are *N* pairs of observations of the variables *X* and *Y* in a bivariate distribution,

$$Cov(X,Y)=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{N} \qquad \ldots\ldots\ldots\ldots(4.2a)$$

$$S_x=\sqrt{\frac{\sum(X-\overline{X})^2}{N}} \qquad \ldots\ldots\ldots\ldots(4.2b)$$

$$\text{and } S_y=\sqrt{\frac{\sum(Y-\overline{Y})^2}{N}} \qquad \ldots\ldots\ldots\ldots(4.2c)$$

Thus by substituting *Eqs. (4.2)* in *Eq. (4.1),* we can write the Pearsonian correlation coefficient as

$$r_{xy}=\frac{\frac{1}{N}\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\frac{1}{N}\sum(X-\overline{X})^2}\sqrt{\frac{1}{N}\sum(Y-\overline{Y})^2}}$$

$$r_{xy}=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{\sum(Y-\overline{Y})^2}} \qquad \ldots\ldots\ldots\ldots(4.3)$$

If we denote, $d_x=X-\overline{X}$ and $d_y=Y-\overline{Y}$

Then $r_{xy}=\frac{\sum d_x d_y}{\sqrt{\sum d_x{}^2 d_y{}^2}}$ $\ldots\ldots\ldots\ldots(4.3a)$

We can further simply the calculations of *Eqs. (4.2)*

We have

$$Cov(X,Y)=\frac{1}{N}\sum(X-\overline{X})(Y-\overline{Y})$$

$$=\frac{1}{N}\sum XY-\overline{X}\overline{Y}$$

$$=\frac{1}{N}\sum XY-\frac{\sum X}{N}\frac{\sum Y}{N}$$

$$=\frac{1}{N^2}\left[N\sum XY-\sum X\sum Y\right] \qquad \ldots\ldots\ldots\ldots(4.4)$$

and $S_x^2 = \frac{1}{N}\sum(X-\overline{X})^2$

$$= \frac{1}{N}\sum X^2 - (\overline{X})^2$$

$$= \frac{1}{N}\sum X^2 - \left(\frac{\sum X}{N}\right)^2$$

$$= \frac{1}{N^2}\left[N\sum X^2 - \left(\sum X\right)^2\right] \qquad \ldots\ldots\ldots\ldots (4.5a)$$

Similarly, we have

$$S_y^2 = \frac{1}{N^2}\left[N\sum Y^2 - \left(\sum Y\right)^2\right] \qquad \ldots\ldots\ldots\ldots (4.5b)$$

So Pearsonian correlation coefficient may be found as

$$r_{xy} = \frac{\frac{1}{N^2}\left[N\sum XY - \sum X\sum Y\right]}{\sqrt{\frac{1}{N^2}\left[N\sum X^2 - \left(\sum X\right)^2\right]}\sqrt{\frac{1}{N^2}\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

or

$$r_{xy} = \frac{N\sum XY - \sum X\sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}} \qquad \ldots\ldots\ldots\ldots (4.6)$$

Remark: *Eq. (4.3)* or *Eq. (4.3a)* is quite convenient to apply if the means $\overline{X}$ and $\overline{Y}$ come out to be integers. If $\overline{X}$ or/and $\overline{Y}$ is (are) fractional then the *Eq. (4.3)* or *Eq. (4.3a)* is quite cumbersome to apply, since the computations of $\sum(X-\overline{X})^2$, $\sum(Y-\overline{Y})^2$ and $\sum(X-\overline{X})(Y-\overline{Y})$ are quite time consuming and tedious. In such a case *Eq. (4.6)* may be used provided the values of *X* or/ and *Y* are small. But if *X* and *Y* assume large values, the calculation of $\sum X^2$, $\sum Y^2$ and $\sum XY$ is again quite time consuming.

Thus if *(i)* $\overline{X}$ and $\overline{Y}$ are fractional and *(ii) X* and *Y* assume large values, the *Eq. (4.3)* and *Eq. (4.6)* are not generally used for numerical problems. In such cases, the step deviation method where we take the deviations of the variables *X* and *Y* from any arbitrary points is used. We will discuss this method in the properties of correlation coefficient.

Properties of Pearson Correlation Coefficient

The following are important properties of Pearson correlation coefficient:

1. *Pearson correlation coefficient cannot exceed 1 numerically*. In other words it lies between –1 and +1. Symbolically, $-1 \leq r \leq 1$

Remarks:

(i) This property provides us a check on our calculations. If in any problem, the obtained value of *r* lies outside the limits $\pm$ 1, this implies that there is some mistake in our calculations.

(ii) The sign of *r* indicate the nature of the correlation. Positive value of *r* indicates positive correlation, whereas negative value indicates negative correlation. *r = 0* indicate absence of correlation.

(iii) The following table sums up the degrees of correlation corresponding to various values of *r:*

Value of *r*	**Degree of correlation**
±1	Perfect correlation
±0.90 or more	Very high degree of correlation
±0.75 to ±0.90	Sufficiently high degree of correlation
±0.60 to ±0.75	Moderate degree of correlation
±0.30 to ±0.60	Only the possibility of a correlation
less than ±0.30	Possibly no correlation
0	Absence of correlation

2. *Pearsonian Correlation coefficient is independent of the change of origin and scale*. Mathematically, if given variables *X* and *Y* are transformed to new variables *U* and *V* by change of origin and scale, *i. e.*

$$U = \frac{X - A}{h} \quad \text{and} \quad V = \frac{Y - B}{k}$$

Where *A, B, h* and *k* are constants and *h > 0, k > 0;* then the correlation coefficient between *X* and *Y* is same as the correlation coefficient between *U* and *V i.e.*,

$$r(X,Y) = r(U, V) \Rightarrow r_{xy} = r_{uv}$$

Remark: This is one of the very important properties of the correlation coefficient and is extremely helpful in numerical computation of *r*. We had already stated that *Eq. (4.3)* and *Eq.(4.6)* become quite

tedious to use in numerical problems if X and/or Y are in fractions or if X and Y are large. In such cases we can conveniently change the origin and scale (if possible) in X or/and Y to get new variables U and V and compute the correlation between U and V by the *Eq. (4.7)*

$$r_{xy} = r_{uv} = \frac{N\sum UV - \sum U \sum V}{\sqrt{N\sum U^2 - \left(\sum U\right)^2}\sqrt{N\sum V^2 - \left(V\right)^2}} \quad \ldots\ldots\ldots\ldots (4.7)$$

3. Two independent variables are uncorrelated but the converse is not true

If X and Y are independent variables then

$$r_{xy} = 0$$

However, the converse of the theorem is not true *i.e.*, uncorrelated variables need not necessarily be independent. As an illustration consider the following bivariate distribution.

X	:	1	2	3	-3	-2	-1
Y	:	1	4	9	9	4	1

For this distribution, value of r will be 0.

Hence in the above example the variable X and Y are uncorrelated. But if we examine the data carefully we find that X and Y are not independent but are connected by the relation $Y = X^2$. The above example illustrates that uncorrelated variables need not be independent.

Remarks: One should not be confused with the words uncorrelation and independence. $r_{xy} = 0$ *i.e.*, uncorrelation between the variables X and Y simply implies the absence of any linear (straight line) relationship between them. They may, however, be related in some other form other than straight line *e.g.*, quadratic (as we have seen in the above example), logarithmic or trigonometric form.

4. Pearson coefficient of correlation is the geometric mean of the two regression coefficients, i.e.

$$r_{xy} = \pm\sqrt{b_{xy}.b_{yx}}$$

The signs of both the regression coefficients are the same, and so the value of r will also have the same sign. This property will be dealt with in detail in the next lesson on Regression Analysis.

5. *The square of Pearsonian correlation coefficient is known as the coefficient of determination*.

Coefficient of determination, which measures the percentage variation in the dependent variable that is accounted for by the independent variable, is a much better and useful measure for interpreting the value of r. This property will also be dealt with in detail in the next lesson.

Probable Error of Correlation Coefficient

The correlation coefficient establishes the relationship of the two variables. After ascertaining this level of relationship, we may be interested to find the extent upto which this coefficient is dependable. Probable error of the correlation coefficient is such a measure of testing the reliability of the observed value of the correlation coefficient, when we consider it as satisfying the conditions of the random sampling.

If *r* is the observed value of the correlation coefficient in a sample of *N* pairs of observations for the two variables under consideration, then the Probable Error, denoted by *PE* (*r*) is expressed as

$$PE(r) = 0.6745\ SE(r)$$

or

$$PE(r) = 0.6745\frac{1-r^2}{\sqrt{N}}$$

There are two main functions of probable error:

1. ***Determination of limits***: The limits of population correlation coefficient are $r \pm PE(r)$, implying that if we take another random sample of the size *N* from the same population, then the observed value of the correlation coefficient in the second sample can be expected to lie within the limits given above, with 0.5 probability. When sample size *N* is small, the concept or value of *PE* may lead to wrong conclusions. Hence to use the concept of *PE* effectively, sample size *N* it should be fairly large.
2. ***Interpretation of 'r':*** The interpretation of *'r'* based on *PE* is as under:
   - If $r < PE(r)$, there is no evidence of correlation, *i.e.* a case of insignificant correlation.
   - If $r > 6\ PE(r)$, correlation is significant. *If* $r < 6\ PE(r)$, it is insignificant.
   - If the probable error is small, correlation exist where $r > 0.5$

Example 12-3

Find the Pearsonian correlation coefficient between sales (in thousand units) and expenses (in thousand rupees) of the following 10 firms:

Firm:	1	2	3	4	5	6	7	8	9	10
Sales:	50	50	55	60	65	65	65	60	60	50
Expenses:	11	13	14	16	16	15	15	14	13	13

Solution: Let sales of a firm be denoted by *X* and expenses be denoted by *Y*

Calculations for Coefficient of Correlation

{Using *Eq. (4.3)* or *(4.3a)*}

Firm	X	Y	$d_x = X - \overline{X}$	$d_y = Y - \overline{Y}$	d_x^2	d_y^2	$d_x.d_y$
1	50	11	-8	-3	64	9	24
2	50	13	-8	-1	64	1	8
3	55	14	-3	0	9	0	0
4	60	16	2	2	4	4	4
5	65	16	7	2	49	4	14
6	65	15	7	1	49	1	7
7	65	15	7	1	49	1	7
8	60	14	2	0	4	0	0
9	60	13	2	-1	4	1	-2
10	50	13	-8	-1	64	1	8
	$\sum X$ = 580	$\sum Y$ = 140			$\sum d_x^2$ =360	$\sum d_y^2$ =22	$\sum d_x d_y$ =70

$$\overline{X} = \frac{\sum X}{N} = \frac{580}{10} = 58 \qquad \text{and} \qquad \overline{Y} = \frac{\sum Y}{N} = \frac{140}{10} = 14$$

Applying the *Eq. (4.3a), w*e have, Pearsonian coefficient of correlation

$$r_{xy} = \frac{\sum d_x d_y}{\sqrt{\sum d_x{}^2 d_y{}^2}}$$

$$r_{xy} = \frac{70}{\sqrt{360x22}}$$

$$r_{xy} = \frac{70}{\sqrt{7920}} = 0.78$$

The value of $r_{xy} = 0.78$, indicate a high degree of positive correlation between sales and expenses.

Example 12-4

The data on price and quantity purchased relating to a commodity for 5 months is given below:

Month :	January	February	March	April	May
Prices(Rs):	10	10	11	12	12
Quantity(Kg):	5	6	4	3	3

Find the Pearsonian correlation coefficient between prices and quantity and comment on its sign and magnitude.

Solution: Let price of the commodity be denoted by *X* and quantity be denoted by *Y*

Calculations for Coefficient of Correlation

{Using *Eq. (4.6)*}

Month	X	Y	X^2	Y^2	XY
1	10	5	100	25	50
2	10	6	100	36	60
3	11	4	121	16	44
4	12	3	144	9	36
5	12	3	144	9	36
	$\sum X = 55$	$\sum Y = 21$	$\sum X^2 = 609$	$\sum Y^2 = 95$	$\sum XY = 226$

Applying the *Eq. (4.6), w*e have, Pearsonian coefficient of correlation

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r_{xy} = \frac{5x226 - 55x21}{\sqrt{(5x609 - 55x55)(5x95 - 21x21)}}$$

$$r_{xy} = \frac{1130 - 1155}{\sqrt{20x34}} =$$

$$r_{xy} = \frac{-25}{\sqrt{680}}$$

$$r_{xy} = -0.98$$

The negative sign of *r* indicate negative correlation and its large magnitude indicate a very high degree of correlation. So there is a high degree of negative correlation between prices and quantity demanded.

Example 12-5

Find the Pearsonian correlation coefficient from the following series of marks obtained by 10 students in a class test in mathematics (*X)* and in Statistics (*Y)*:

X :	45	70	65	30	90	40	50	75	85	60
Y :	35	90	70	40	95	40	60	80	80	50

Also calculate the Probable Error.

Solution:

Calculations for Coefficient of Correlation

{Using *Eq. (4.7)*}

X	*Y*	*U*	*V*	U^2	V^2	*UV*
45	35	-3	-6	9	36	18
70	90	2	5	4	25	10
65	70	1	1	1	1	1
30	40	-6	-5	36	25	30
90	95	6	6	36	36	36
40	40	-4	-5	16	25	20

50	60	-2	-1	4	1	2
75	80	3	3	9	9	9
85	80	5	3	25	9	15
60	50	0	-3	0	9	0
		$\sum U = 2$	$\sum V = -2$	$\sum U^2 = 140$	$\sum V^2 = 176$	$\sum UV = 141$

We have, defined variables *U* and *V* as

$$U = \frac{X - 60}{5} \qquad \text{and} \qquad V = \frac{Y - 65}{5}$$

Applying the *Eq. (4.7)*

$$r_{xy} = r_{uv} = \frac{N\sum UV - \left(\sum U \sum V\right)}{\sqrt{N\sum U^2 - \left(\sum U\right)^2}\sqrt{N\sum V^2 - \left(\sum V\right)^2}}$$

$$= \frac{10x141 - 2x(-2)}{\sqrt{10x140 - 2x2}\sqrt{10x176 - (-2)x(-2)}}$$

$$= \frac{1410 + 4}{\sqrt{1400 - 4}\sqrt{1760 - 4}}$$

$$= \frac{1414}{\sqrt{2451376}} = 0.9$$

So, there is a high degree of positive correlation between marks obtained in Mathematics and in Statistics.

Probable Error, denoted by *PE* (*r*) is given as

$$PE(r) = 0.6745\frac{1 - r^2}{\sqrt{N}}$$

$$PE(r) = 0.6745\frac{1 - (0.9)^2}{\sqrt{10}}$$

$$PE(r) = 0.0405$$

So the value of *r* is highly significant.

12.3.4 SPEARMAN'S RANK CORRELATION

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as honesty, beauty, character, morality, *etc.*, which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearson's coefficient of correlation cannot be used as such. Charles Edward Spearman, a British Psychologist, developed a formula in 1904, which consists in obtaining the correlation coefficient between the ranks of *N* individuals in the two attributes under study.

Suppose we want to find if two characteristics *A*, say, intelligence and *B*, say, beauty are related or not. Both the characteristics are incapable of quantitative measurements but we can arrange a group of *N* individuals in order of merit (ranks) *w.r.t.* proficiency in the two characteristics. Let the random variables *X* and *Y* denote the ranks of the individuals in the characteristics *A* and *B* respectively. If we assume that there is no tie, *i.e.*, if no two individuals get the same rank in a characteristic then, obviously, *X* and *Y* assume numerical values ranging from *1* to *N*.

The Pearsonian correlation coefficient between the ranks X and Y is called the rank correlation coefficient between the characteristics A and B for the group of individuals.

Spearman's rank correlation coefficient, usually denoted by ρ(Rho) is given by the equation

$$\rho = 1 - \frac{6\sum d^2}{N(N^2-1)} \qquad \ldots\ldots\ldots\ldots (4.8)$$

Where *d* is the difference between the pair of ranks of the same individual in the two characteristics and *N* is the number of pairs.

Example 12-6

Ten entries are submitted for a competition. Three judges study each entry and list the ten in rank order. Their rankings are as follows:

Entry:	A	B	C	D	E	F	G	H	I	J
Judge J_1:	9	3	7	5	1	6	2	4	10	8
Judge J_2:	9	1	10	4	3	8	5	2	7	6
Judge J_3:	6	3	8	7	2	4	1	5	9	10

Calculate the appropriate rank correlation to help you answer the following questions:

(i) Which pair of judges agrees the most?

(ii) Which pair of judges disagrees the most?

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq.(4.8)*}

Entry	Rank by Judges			Difference in Ranks					
	J_1	J_2	J_3	$d(J_1 \& J_2)$	d^2	$d(J_1 \& J_3)$	d^2	$d(J_2 \& J_3)$	d^2
A	9	9	6	0	0	+3	9	+3	9
B	3	1	3	+2	4	0	0	-2	4
C	7	10	8	-3	9	-1	1	+2	4
D	5	4	7	+1	1	-2	4	-3	9
E	1	3	2	-2	4	-1	1	+1	1
F	6	8	4	-2	4	+2	4	+4	16
G	2	5	1	-3	9	+1	1	+4	16
H	4	2	5	+2	4	-1	1	-3	9
I	10	7	9	+3	9	+1	1	-2	4
J	8	6	10	+2	4	-2	4	-4	16
					$\sum d^2 = 48$		$\sum d^2 = 26$		$\sum d^2 = 88$

$$\rho(J_1 \& J_2) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 48}{10(10^2-1)} = 1 - \frac{288}{990} = 1 - 0.29 = +0.71$$

$$\rho(J_1 \& J_3) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 26}{10(10^2-1)} = 1 - \frac{156}{990} = 1 - 0.1575 = +0.8425$$

$$\rho(J_2 \& J_3) = 1 - \frac{6\sum d^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 88}{10(10^2-1)} = 1 - \frac{528}{990} = 1 - 0.53 = +0.47$$

So *(i)* Judges J_1 and J_3 agree the most

(ii) Judges J_2 and J_3 disagree the most

Spearman's rank correlation *Eq.(4.8)* can also be used even if we are dealing with variables, which are

measured quantitatively, *i.e.* when the actual data but not the ranks relating to two variables are given. In such a case we shall have to convert the data into ranks. The highest (or the smallest) observation is given the rank 1. The next highest (or the next lowest) observation is given rank 2 and so on. It is immaterial in which way (descending or ascending) the ranks are assigned. However, the same approach should be followed for all the variables under consideration.

Example 12-7

Calculate the rank coefficient of correlation from the following data:

X:	75	88	95	70	60	80	81	50
Y:	120	134	150	115	110	140	142	100

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq.(4.8)*}

X	Ranks R_X	Y	Ranks R_Y	$d = R_X - R_Y$	d^2
75	5	120	5	0	0
88	2	134	4	-2	4
95	1	150	1	0	0
70	6	115	6	0	0
60	7	110	7	0	0
80	4	140	3	+1	1
81	3	142	2	+1	1
50	8	100	8	0	0

$\sum d^2 = 6$

$$\rho = 1 - \frac{6\sum d^2}{N(N^2-1)} = 1 - \frac{6 \times 6}{8(8^2-1)} = 1 - \frac{36}{504} = 1 - 0.07 = +0.93$$

Hence, there is a high degree of positive correlation between *X* and *Y*

Repeated Ranks

In case of attributes if there is a tie *i.e.,* if any two or more individuals are placed together in any classification *w.r.t.* an attribute or if in case of variable data there is more than one item with the same

value in either or both the series then Spearman's *Eq.(4.8)* for calculating the rank correlation coefficient breaks down, since in this case the variables *X* [the ranks of individuals in characteristic A (1st series)] and *Y* [the ranks of individuals in characteristic B (2nd series)] do not take the values from *1* to *N*.

In this case common ranks are assigned to the repeated items. These common ranks are the arithmetic mean of the ranks, which these items would have got if they were different from each other and the next item will get the rank next to the rank used in computing the common rank. For example, suppose an item is repeated at rank 4. Then the common rank to be assigned to each item is (4+5)/2, *i.e.*, 4.5 which is the average of 4 and 5, the ranks which these observations would have assumed if they were different. The next item will be assigned the rank 6. If an item is repeated thrice at rank 7, then the common rank to be assigned to each value will be (7+8+9)/3, *i.e.*, 8 which is the arithmetic mean of 7,8 and 9 *viz.*, the ranks these observations would have got if they were different from each other. The next rank to be assigned will be 10.

If only a small proportion of the ranks are tied, this technique may be applied together with *Eq.(4.8)*. If a large proportion of ranks are tied, it is advisable to apply an adjustment or a correction factor to *Eq.(4.8)*as explained below:

"In the Eq.(4.8) add the factor

$$\frac{m(m^2-1)}{12} \qquad \text{……(4.8a)}$$

to $\sum d^2$ *; where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the series".*

Example 12-8

For a certain joint stock company, the prices of preference shares (*X*) and debentures (*Y*) are given below:

X:	73.2	85.8	78.9	75.8	77.2	81.2	83.8
Y:	97.8	99.2	98.8	98.3	98.3	96.7	97.1

Use the method of rank correlation to determine the relationship between preference prices and debentures prices.

Solution:

Calculations for Coefficient of Rank Correlation

{Using *Eq. (4.8)* and *(4.8a)*}

X	Y	Rank of X (X_R)	Rank of Y (Y_R)	$d = X_R - Y_R$	d^2
73.2	97.8	7	5	2	4
85.8	99.2	1	1	0	0
78.9	98.8	4	2	2	4
75.8	98.3	6	3.5	2.5	6.25
77.2	98.3	5	3.5	1.5	2.25
81.2	96.7	3	7	-4	16
83.8	97.1	2	6	-4	16
				$\sum d = 0$	$\sum d^2 = 48.50$

In this case, due to repeated values of Y, we have to apply ranking as average of 2 ranks, which could have been allotted, if they were different values. Thus ranks 3 and 4 have been allotted as 3.5 to both the values of Y = 98.3. Now we also have to apply correction factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, where m in the number of times the value is repeated, here $m = 2$.

$$\rho = \frac{6\left[\sum d^2 + \frac{m(m^2-1)}{2}\right]}{N(N^2-1)} = \frac{6\left[48.5 + \frac{2(4-1)}{12}\right]}{7(7^2-1)} = 1 - \frac{6\,x\,49}{7\,x\,48} = 0.125$$

Hence, there is a very low degree of positive correlation, probably no correlation, between preference share prices and debenture prices.

Remarks on Spearman's Rank Correlation Coefficient

1. We always have $\sum d = 0$, which provides a check for numerical calculations.
2. Since Spearman's rank correlation coefficient, ρ, is nothing but Karl Pearson's correlation coefficient, r, between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.
3. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure, which is distribution free (or non-parametric). Spearman's ρ is such a distribution free measure, since no strict assumption are made about the from of the population from which sample observations are drawn.
4. Spearman's formula is easy to understand and apply as compared to Karl Pearson's formula. The values obtained by the two formulae, *viz* Pearsonian r and Spearman's ρ are generally different. The

difference arises due to the fact that when ranking is used instead of full set of observations, there is always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics, which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.
6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution. For N >30, this formula should not be used unless the ranks are given.

12.3.5 CONCURRENT DEVIATION METHOD

This is a casual method of determining the correlation between two series when we are not very serious about its precision. This is based on the signs of the deviations (*i.e.* the direction of the change) of the values of the variable from its preceding value and does not take into account the exact magnitude of the values of the variables. Thus we put a plus (+) sign, minus (-) sign or equality (=) sign for the deviation if the value of the variable is greater than, less than or equal to the preceding value respectively. The deviations in the values of two variables are said to be concurrent if they have the same sign (either both deviations are positive or both are negative or both are equal). The formula used for computing correlation coefficient r_c by this method is given by

$$r_c = \pm\sqrt{\pm\left(\frac{2c-N}{N}\right)} \quad \ldots\ldots\ldots\ldots (4.9)$$

Where c is the number of pairs of concurrent deviations and N is the number of pairs of deviations. If ($2c$-N) is positive, we take positive sign in and outside the square root in *Eq. (4.9)* and if ($2c$-N) is negative, we take negative sign in and outside the square root in *Eq. (4.9).*

Remarks: (i) It should be clearly noted that here N is not the number of pairs of observations but it is the number of pairs of deviations and as such it is one less than the number of pairs of observations.

(ii) Coefficient of concurrent deviations is primarily based on the following principle:

> *"If the short time fluctuations of the time series are positively correlated or in other words, if their deviations are concurrent, their curves would move in the same direction and would indicate positive correlation between them"*

Example 12-9

Calculate coefficient of correlation by the concurrent deviation method

Supply:	112	125	126	118	118	121	125	125	131	135
Price:	106	102	102	104	98	96	97	97	95	90

Solution:

Calculations for Coefficient of Concurrent Deviations

{Using *Eq. (4.9)*}

Supply (X)	Sign of deviation from preceding value (X)	Price (Y)	Sign of deviation preceding value (Y)	Concurrent deviations
112		106		
125	+	102	-	
126	+	102	=	
118	-	104	+	
118	=	98	-	
121	+	96	-	
125	+	97	+	+(c)
125	=	97	=	= (c)
131	+	95	-	
135	+	90	-	

We have

Number of pairs of deviations, N =10 – 1 = 9

c = Number of concurrent deviations

= Number of deviations having like signs

= 2

Coefficient of correlation by the method of concurrent deviations is given by:

$$r_c = \pm\sqrt{\pm\left(\frac{2c-N}{N}\right)}$$

$$r_c = \pm\sqrt{\pm\left(\frac{2x2-9}{9}\right)}$$

$$r_c = \pm\sqrt{\pm(-0.5556)}$$

Since $2c - N$ = -5 (negative), we take negative sign inside and outside the square root

$$r_c = -\sqrt{-(-0.5556)}$$

$$r_c = -\sqrt{0.5556}$$

$$r_c = -0.7$$

Hence there is a fairly good degree of negative correlation between supply and price.

12.4 LIMITATIONS OF CORRELATION ANALYSIS

As mentioned earlier, correlation analysis is a statistical tool, which should be properly used so that correct results can be obtained. Sometimes, it is indiscriminately used by management, resulting in misleading conclusions. We give below some *errors* frequently made in the use of correlation analysis:

1. Correlation analysis cannot determine cause-and-effect relationship. One should not assume that a change in *Y* variable is caused by a change in *X* variable unless one is reasonably sure that one variable is the cause while the other is the effect. Let us take an example.

   Suppose that we study the performance of students in their graduate examination and their earnings after, say, three years of their graduation. We may find that these two variables are highly and positively related. At the same time, we must not forget that both the variables might have been influenced by some other factors such as quality of teachers, economic and social status of parents, effectiveness of the interviewing process and so forth. If the data on these factors are available, then it is worthwhile to use multiple correlation analysis instead of bivariate one.
2. Another mistake that occurs frequently is on account of misinterpretation of the coefficient of correlation. Suppose in one case $r = 0.7$, it will be wrong to interpret that correlation explains 70 percent of the total variation in *Y*. The error can be seen easily when we calculate the coefficient of determination. Here, the coefficient of determination r^2 will be 0.49. This means that only 49 percent of the total variation in *Y* is explained.

   Similarly, the coefficient of determination is misinterpreted if it is also used to indicate causal relationship, that is, the percentage of the change in one variable is due to the change in another variable.
3. Another mistake in the interpretation of the coefficient of correlation occurs when one concludes a positive or negative relationship even though the two variables are actually unrelated. For example, the age of students and their score in the examination have no relation with each other. The two variables may show similar movements but there does not seem to be a common link between them.

To sum up, one has to be extremely careful while interpreting coefficient of correlation. Before one concludes a causal relationship, one has to consider other relevant factors that might have any influence

on the dependent variable or on both the variables. Such an approach will avoid many of the pitfalls in the interpretation of the coefficient of correlation. It has been rightly said that the coefficient of correlation is not only one of the most widely used, but also one of the widely abused statistical measures.

12.5 PARTIAL CORRELATION

It is a statistical technique to analyze the association between dependent variables and one of the independent variables by eliminating the effect of other variables. Partial correlation is also known as Net Correlation. It is a statistical technique to study the association between one dependent variable and one independent variable by keeping other independent variables constant. In simple correlation, the effect of other independent variables was ignored. It is mainly divided into three categories:

- Zero order coefficient
- First order coefficient
- Second order coefficient

12.5.1 Zero Order Coefficient

It is simple correlation between two variables whereby no other variables are held constant. It can be understood with the help of following example:

It is assumed that there are three variables named as x_1, x_2 and x_3. In this case, simple correlation will be calculated between any of two variables by ignoring third variable completely in the calculation. Then,

- r_{12} = Simple correlation between x_1 and x_2 variables by ignoring x_3 variable.
- r_{23} = Simple correlation between x_2 and x_3 variables by ignoring x_1 variable.
- r_{31} = Simple correlation between x_3 and x_1 variables by ignoring x_2 variable.

The formulas for calculating simple correlation between two variables has already been discussed in section 12.3 of this lesson. Therefore, you can refer to the section 12.3 to understand the simple correlation calculation techniques.

12.5.2 First Order Coefficient

It is a partial correlation between two variables keeping the third variable constant. You can keep any one variable constant out of three variables in final calculation of partial correlation. Hence, there can be three cases of first order coefficient:

- $r_{12.3}$ = Simple correlation between x_1 and x_2 variables keeping x_3 variable constant.

- $r_{23.1}$ = Simple correlation between x_2 and x_3 variables keeping x_1 variable constant.
- $r_{13.2}$ = Simple correlation between x_1 and x_3 variables keeping x_2 variable constant.

Following are the formulas for calculating first order partial correlation:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)}\ \sqrt{(1 - r_{23}^2)}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{(1 - r_{12}^2)}\ \sqrt{(1 - r_{13}^2)}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{\sqrt{(1 - r_{12}^2)}\ \sqrt{(1 - r_{23}^2)}}$$

Example 12.10 The simple coefficients of correlation between two variables out of three are: $r_{12} = 0.8$, $r_{13} = 0.7$ and $r_{23} = 0.6$ Find the partial coefficients of correlation, $r_{12.3}$, $r_{23.1}$ and $r_{13.2}$

Solution:

$$r_{12.3} = \frac{0.8 - (0.7 \times 0.6)}{\sqrt{(1 - 0.7^2)}\ \sqrt{(1 - 0.6^2)}} = \frac{0.8 - 0.42}{0.714 \times 0.8} = \frac{0.38}{0.5712} = 0.665$$

$$r_{23.1} = \frac{0.6 - (0.8 \times 0.7)}{\sqrt{(1 - 0.8^2)}\ \sqrt{(1 - 0.7^2)}} = \frac{0.6 - 0.56}{0.6 \times 0.714} = \frac{0.04}{0.4284} = 0.093$$

$$r_{13.2} = \frac{0.7 - (0.8 \times 0.6)}{\sqrt{(1 - 0.8^2)}\ \sqrt{(1 - 0.6^2)}} = \frac{0.7 - 0.48}{0.6 \times 0.8} = \frac{0.22}{0.48} = 0.458$$

12.5.3 Second Order Coefficient

It is a partial correlation between two variables keeping other two variables constant. When four variables are included in a problem than any two of them are kept constant and the correlation is calculated between remaining two variables. Let's assume that there are four variables including x_1, x_2, x_3 and x_4. Thus, if correlation is calculated between x_1 and x_2 then two variables x_3 and x_4 will be kept constant. It can be symbolically represented by $r_{12.34}$. In this way, there can be six cases of partial correlation of second order including:

- $r_{12.34}$ = correlation between x_1 and x_2 variables keeping x_3 and x_4 constant.
- $r_{13.24}$ = correlation between x_1 and x_3 variables keeping x_2 and x_4 constant.
- $r_{14.23}$ = correlation between x_1 and x_4 variables keeping x_2 and x_3 constant.
- $r_{23.14}$ = correlation between x_2 and x_3 variables keeping x_1 and x_4 constant.
- $r_{24.13}$ = correlation between x_2 and x_4 variables keeping x_1 and x_3 constant.
- $r_{34.12}$ = correlation between x_3 and x_4 variables keeping x_1 and x_2 constant.

Following are the formulas for calculating second order partial correlation:

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}\, r_{23.4}}{\sqrt{(1 - r^2_{13.4})}\ \sqrt{(1 - r^2_{23.4})}}$$

$$r_{13.24} = \frac{r_{13.4} - r_{12.4}\, r_{23.4}}{\sqrt{(1 - r^2_{12.4})}\ \sqrt{(1 - r^2_{23.4})}}$$

$$r_{14.23} = \frac{r_{14.3} - r_{12.3}\, r_{24.3}}{\sqrt{(1 - r^2_{12.3})}\ \sqrt{(1 - r^2_{24.3})}}$$

$$r_{23.14} = \frac{r_{23.4} - r_{12.4}\, r_{13.4}}{\sqrt{(1 - r^2_{12.4})}\ \sqrt{(1 - r^2_{13.4})}}$$

$$r_{24.13} = \frac{r_{24.3} - r_{12.3}\, r_{14.3}}{\sqrt{(1 - r^2_{12.3})}\ \sqrt{(1 - r^2_{14.3})}}$$

$$r_{34.12} = \frac{r_{34.2} - r_{13.2}\, r_{14.2}}{\sqrt{(1 - r^2_{13.2})}\ \sqrt{(1 - r^2_{14.2})}}$$

Remarks:

1. The partial correlation coefficient lies between -1 and +1.
2. These correlation coefficient are calculated on the basis of zero order coefficient or simple coefficients of correlation where no variable is kept constant. Thus r_{12} is a simple correlation coefficient between x_1 and x_2.

Example: 12.11 Given: $r_{12.4} = 0.6$, $r_{13.4} = 0.5$ and $r_{23.4} = 0.7$. Find $r_{12.34}$ and $r_{13.24}$.

Solution:

$$r_{12.34} = \frac{0.6 - (0.5 \times 0.7)}{\sqrt{(1 - 0.5^2)}\ \sqrt{(1 - 0.7^2)}}$$

$$= \frac{0.6-0.35}{\sqrt{0.75}\ \sqrt{0.51}} = \frac{0.25}{0.867 \times 0.714} = \frac{0.25}{0.619} = 0.404$$

$$r_{13.24} = \frac{0.5 - (0.6 \times 0.7)}{\sqrt{(1-0.6^2)}\ \sqrt{(1-0.7^2)}}$$

$$= \frac{0.5-0.42}{\sqrt{0.64}\ \sqrt{0.51}} = \frac{0.8}{0.8 \times 0.714} = \frac{0.8}{0.5712} = 0.14$$

Example 12.12 It is observed that weight of fish depends on food consumption, type of water and temperature. Find the partial correlation between weight of fish (x_1) and temperature (x_4) keeping food consumption (x_2) and type of water (x_3) constant. Given that $r_{14.3} = 0.6$, $r_{12.3} = 0.5$, $r_{24.3}$ 0.7. Find $r_{14.23}$.

Solution:

$$r_{14.23} = \frac{0.6 - (0.5 \times 0.7)}{\sqrt{(1-0.5^2)}\ \sqrt{(1-0.7^2)}}$$

$$= \frac{0.6-0.35}{\sqrt{0.75}\ \sqrt{0.51}} = \frac{0.25}{0.867 \times 0.714} = \frac{0.25}{0.619} = 0.404$$

This gives the partial correlation between weight of fish and temperature keeping food consumption and type of water constant.

12.5.4 Limitations of Partial Correlation

Followings are the main limitations of partial correlation:

1. In the calculation of partial correlation coefficient, it is presumed that there should exist a linear relation between variables. In real situation, this condition lacks in some cases.
2. The independent variable in the study of partial correlation should be linearly independent.
3. The reliability of partial correlation coefficient decreases as their order goes up. This means that the second order partial coefficients are not as dependable as the first order ones are. Therefore, it is necessary that the size of the items in the gross correlation should be large.
4. It involves a lot of calculation work and its analysis is not easy.

12.6 MULTIPLE CORRELATION:

Multiple correlation is a technique to study the relationship between three or four variables simultaneously. In this, effect of all independent variables on a dependent variable is analyze. The coefficient of multiple linear correlation is represented by R. It can be understood as follow:

Assuming three variable x_1, x_2 and x_3 where as x_1 is the dependent variable and other two are independent variables. Here, the multiple correlation coefficient can be defined as follow:

- $R_{1.23}$ = Multiple correlation coefficient with x_1 as dependent variable and x_2 and x_3 as independent variables.
- $R_{2.13}$ = Multiple correlation coefficient with x_2 as dependent variable and x_1 and x_3 as independent variables.
- $R_{3.12}$ = Multiple correlation coefficient with x_3 as dependent variable and x_1 and x_2 as independent variables.

12.6.1 Calculation of Multiple Correlation Coefficient

Following are the formulas for calculating multiple correlation coefficient:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13} r_{23}}{1 - r_{23}^2}}$$

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}\, r_{13} r_{23}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}\, r_{13} r_{23}}{1 - r_{12}^2}}$$

If there are three independent variables and one dependent variable the formula for finding out the multiple correlation is

$$R_{1.234} = \sqrt{1 - (1 - r^2{}_{14})\,(1 - r^2{}_{12.3})\,(1 - r^2{}_{12.34})}$$

Remarks

1. Multiple correlation coefficient is a positive coefficient always. Its value ranges between 0 and 1. It cannot be a negative value.
2. If $R_{1.23} = 0$, then $r_{12} = 0$ and $r_{13} = 0$.
3. $R_{1.23} \geq r_{12}$ and $R_{1.23} \geq r_{13}$
4. $R_{1.23}$ is the same thing as $R_{1.32}$. The position of the subscript to the right of the dot does not make a difference.

5. By squaring $R_{1.23}$, we get the coefficient of multiple determination.

Example 12.13 The following zero order correlation coefficient are given: $r_{12} = 0.5$, $r_{13} = 0.6$, $r_{23} = 0.7$. Calculate the multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

Solution:

$$R_{1.23} = \sqrt{\frac{(0.5)^2 + (0.6)^2 - 2 \times 0.5 \times 0.6 \times 0.7}{1 - (0.7)^2}}$$

$$= \sqrt{\frac{0.25 + 0.36 - 0.42}{0.51}} = \sqrt{\frac{0.19}{0.51}} = \sqrt{0.3725} = 0.610$$

$$R_{2.13} = \sqrt{\frac{(0.5)^2 + (0.7)^2 - 2 \times 0.5 \times 0.6 \times 0.7}{1 - (0.6)^2}}$$

$$= \sqrt{\frac{0.25 + 0.49 - 0.42}{0.64}} = \sqrt{\frac{0.32}{0.64}} = \sqrt{0.5} = 0.707$$

$$R_{3.12} = \sqrt{\frac{(0.6)^2 + (0.7)^2 - 2 \times 0.5 \times 0.6 \times 0.7}{1 - (0.5)^2}}$$

$$= \sqrt{\frac{0.36 + 0.49 - 0.42}{0.75}} = \sqrt{\frac{0.43}{0.75}} = \sqrt{0.5733} = 0.757$$

12.6.2 Limitations of Multiple Correlation

Followings are the main limitations of multiple correlation coefficients:

1. It also assume a linear relationship between the zero order coefficients like partial correlation coefficients.
2. It assume that the independent variables affect the dependent variable in an independent manner and have an additive property.
3. If, however, there is interrelation between independent variables, their effect cannot be distinct nor can they be additive.
4. The calculation of multiple correlation is cumbersome and difficult.

12.7 CHECK YOUR PROGRESS

1. If both the variables move in the same direction, we say that there is a …… correlation.
2. If the change in one variable is accompanied by change in another variable in a constant ratio, it is a case of …….. correlation.
3. When three or more variables are involved in a study, then it is a problem of either ……correlation.
4. The Pearsonian correlation coefficient between the ranks X and Y is called the…… between the characteristics A and B for the group of individuals.
5. Correlation analysis cannot determine …… relationship.

12.8 SUMMARY

The statistical methods of Correlation (discussed in the present lesson) and Regression (to be discussed in the next lesson) are helpful in knowing the relationship between two or more variables which may be related in same way, *like* interest rate of bonds and prime interest rate. Correlation can be classified in several ways. The important ways of classifying correlation are: Positive and negative, Linear and non-linear (curvilinear) and Simple, partial and multiple. The commonly used methods for studying linear relationship between two variables involve both graphic and algebraic methods. Some of the widely used methods include: Scatter Diagram, Correlation Graph, Pearson's Coefficient of Correlation, Spearman's Rank Correlation and Concurrent Deviation Method. There are measure uses of correlation analysis, but some limitations too. It has been rightly said that the coefficient of correlation is not only one of the most widely used, but also one of the widely abused statistical measures.

12.9 KEYWORDS

Correlation: When two or more variables very in sympathy so that movement in one tends to be accompanied by corresponding movements in the other variable(s), they are said to be correlated.

Scatter Diagram: Under this method, both the variables are plotted on the graph paper by putting dots. By studying diagram, we can have rough idea about the nature and degree of relationship between two variables.

Correlation Graph: The data pertaining to two series are plotted on a graph sheet and then find out the correlation by examining the direction and closeness of two curves.

Pearson's correlation coefficient: It is denoted by $r(X,Y)$ or r_{xy} or simply r is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between X and Y, to the

product of the standard deviations of X and Y.

Concurrent Deviation Method: The deviations in the values of two variables are said to be concurrent if they have the same sign.

12.10 SELF-ASSESSMENT TEST

1. "Correlation and Regression are two sides of the same coin". Explain.
2. Explain the meaning and significance of the concept of correlation. Does correlation always signify casual relationships between two variables? Explain with illustration on what basis can the following correlation be criticized?
   (a) Over a period of time there has been an increased financial aid to under developed countries and also an increase in comedy act television shows. The correlation is almost perfect.
   (b) The correlation between salaries of school teachers and amount of liquor sold during the period 1940 – 1980 was found to be 0.96
3. Write short not on the following
   (a) Spurious correlation
   (b) Positive and negative correlation
   (c) Linear and non-linear correlation
   (d) Simple, multiple and partial correlation
4. What is a scatter diagram? How does it help in studying correlation between two variables, in respect of both its nature and extent?
5. Write short note on the following
   (a) Karl Pearson's coefficient of correlation
   (b) Probable Error
   (c) Spearman's Rank Correlation Coefficient
   (d) Coefficient of Concurrent Deviation
6. Draw a scatter diagram from the data given below and interpret it.

X:	10	20	30	40	50	60	70	80
Y:	32	20	24	36	40	28	38	44

7. Calculate Karl Pearson's coefficient of correlation between expenditure on advertising (X) and sales (Y) from the data given below:

X:	39	65	62	90	82	75	25	98	36	78
Y:	47	53	58	86	62	68	60	91	51	84

8. To study the effectiveness of an advertisement a survey is conducted by calling people at random by asking the number of advertisements read or seen in a week (*X*) and the number of items purchased (*Y*) in that week.

X:	5	10	4	0	2	7	3	6
Y:	10	12	5	2	1	3	4	8

Calculate the correlation coefficient and comment on the result.

9. Calculate coefficient of correlation between *X* and *Y* series from the following data and calculate its probable error also.

X:	78	89	96	69	59	79	68	61
Y:	125	137	156	112	107	136	123	108

10. In two set of variables *X* and *Y*, with 50 observations each, the following data are observed:

$\overline{X}$ = 10, SD of *X* = 3

$\overline{Y}$ = 6, SD of *Y* = 2 $r_{xy} = 0.3$

However, on subsequent verification, it was found that one value of *X* (=10) and one value of *Y* (= 6) were inaccurate and hence weeded out with the remaining 49 pairs of values. How the original value of is $r_{xy} = 0.3$ affected?

11. Calculate coefficient of correlation *r* between the marks in statistics (*X*) and Accountancy (*Y*) of 10 students from the following:

X:	52	74	93	55	41	23	92	64	40	71
Y:	45	80	63	60	35	40	70	58	43	64

Also determine the probable error or *r*.

12. The coefficient of correlation between two variables *X* and *Y* is 0.48. The covariance is 36. The variance of *X* is 16. Find the standard deviation of *Y*.

13. Twelve entries in painting competition were ranked by two judges as shown below:

Entry:	A	B	C	D	E	F	G	H	I	J
Judge I:	5	2	3	4	1	6	8	7	10	9
Judge II:	4	5	2	1	6	7	10	9	3	8

Find the coefficient of rank correlation.

14. Calculate Spearman's rank correlation coefficient between advertisement cost (*X*) and sales (*Y*) from the following data:

X:	39	65	62	90	82	75	25	98	36	78
Y:	47	53	58	86	62	68	60	91	51	84

15. An examination of eight applicants for a clerical post was taken by a firm. From the marks obtained by the applicants in the Accountancy (*X*) and Statistics (*Y*) paper, compute rank coefficient of correlation.

Applicant:	A	B	C	D	E	F	G	H
X:	15	20	28	12	40	60	20	80
Y:	40	30	50	30	20	10	30	60

16. Calculate the coefficient of concurrent deviation from the following data:

Year:	1993	1994	1995	1996	1997	1998	1999	2000
Supply:	160	164	172	182	166	170	178	192
Price:	222	280	260	224	266	254	230	190

17. Obtain a suitable measure of correlation from the following data regarding changes in price index of the shares *A* and *B* during nine months of a year:

Month:	A	M	J	J	A	S	O	N	D
A:	+4	+3	+2	-1	-3	+4	-5	+1	+2
B:	-2	+5	+3	-2	-1	-3	+4	-1	-3

18. The cross-classification table shows the marks obtained by 105 students in the subjects of Statistics and Finance:

		Marks in Statistics				
		50-54	55-59	60-64	65-74	Total
Marks in Finance	50-59	4	6	8	7	25
	60-69	-	10	12	13	35
	70-79	16	9	20	-	45
	80-89	-	-	-	-	-
	Total	20	25	40	20	105

Find the coefficient of correlation between marks obtained in two subjects.

12.11 ANSWERS TO CHECK YOUR PROGRESS

1. Positive
2. Linear
3. Partial or multiple

4. Rank correlation coefficient
5. Cause-and-effect

12.12 REFERNCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics by Kohlar Heinz. Harper Collins., New York.

Subject: Business Statistics	
Course Code: MBA-102	Author: Anil Kumar
Lesson: 13	Vetter: Prof. Harbhajan Bansal
REGRESSION ANALYSIS	

STRUCTURE

13.1 Learning Objectives

13.2 What is Regression?

13.2.1 Linear Regression

13.2.2 Regression Line of Y on X

13.2.2.1 Scatter Diagram

13.2.2.2 Fitting a Straight Line

13.2.2.3 Predicting an Estimate and its Preciseness

13.2.2.4 Error of Estimate

13.2.3 Regression Line of X on Y

13.3 Properties of Regression Coefficients

13.3.1 Regression Lines and Coefficient of Correlation

13.3.2 Coefficient of Determination

13.3.3 Correlation Analysis Versus Regression Analysis

13.4 Solved Problems

13.5 Multiple Regression Analysis

13.5.1 Regression Equations

13.5.2 Assumptions of Multiple Linear Regression Analysis

13.5.3 Methods of Multiple Regression Analysis

13.5.4 Standard Error of the Estimates

13.6 Solved Problems of Multiple Regression Analysis

13.7 Check your progress
13.8 Summary
13.9 Keywords
13.10 Self-Assessment Test
13.11 Answers to check your progress
13.12 References/Suggested Readings

13.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- **Understand the concept of linear regression**
- **Understand the importance of regression analysis**
- **Understand the limitations of regression analysis.**

13.2 WHAT IS REGRESSION?

"if we find any association between two or more variables, we might be interested in estimating the value of one variable for known value(s) of another variable(s)"

In business, several times it becomes necessary to have some forecast so that the management can take a decision regarding a product or a particular course of action. In order to make a forecast, one has to ascertain some relationship between two or more variables relevant to a particular situation. For example, a company is interested to know how far the demand for television sets will increase in the next five years, keeping in mind the growth of population in a certain town. Here, it clearly assumes that the increase in population will lead to an increased demand for television sets. Thus, to determine the nature and extent of relationship between these two variables becomes important for the company.

In the preceding lesson, we studied in some depth linear correlation between two variables. Here we have a similar concern, the association between variables, except that we develop it further in two respects. *First,* we learn how to build statistical models of relationships between the variables to have a better understanding of their features. *Second,* we extend the models to consider their use in forecasting. For this purpose, we have to use the technique - regression analysis - which forms the subject-matter of this lesson.

In 1889, Sir Francis Galton, a cousin of Charles Darwin published a paper on heredity, *"Natural Inheritance".* He reported his discovery that sizes of seeds of sweet pea plants appeared to "revert" or "regress", to the mean size in successive generations. He also reported results of a study of the relationship between heights of fathers and heights of their sons. A straight line was fit to the data pairs: height of father versus height of son. Here, too, he found a "regression to mediocrity" The heights of the sons represented a movement away from their fathers, towards the average height. We credit Sir Galton with the idea of statistical regression.

While most applications of regression analysis may have little to do with the "regression to the mean" discovered by Galton, the term "regression" remains. It now refers to the statistical technique of modeling the relationship between two or more variables. In general sense, regression analysis means the estimation or prediction of the unknown value of one variable from the known value(s) of the other variable(s). It is one of the most important and widely used statistical techniques in almost all sciences - natural, social or physical.

In this lesson we will focus only on simple regression –linear regression involving only two variables: a dependent variable and an independent variable. Regression analysis for studying more than two variables at a time is known as multiple regressions.

INDEPENDENT AND DEPENDENT VARIABLES

Simple regression involves only two variables; one variable is predicted by another variable. *The variable to be predicted* is called the dependent variable. *The predictor* is called the independent variable, or *explanatory variable.* For example, when we are trying to predict the demand for television sets on the basis of population growth, we are using the demand for television sets as the dependent variable and the population growth as the independent or predictor variable.

The decision, as to which variable is which sometimes, causes problems. Often the choice is obvious, as in case of demand for television sets and population growth because it would make no sense to suggest that population growth could be dependent on TV demand! The population growth has to be the independent variable and the TV demands the dependent variable.

If we are unsure, here are some points that might be of use:

- If we have control over one of the variables then that is the independent. For example, a manufacturer can decide how much to spend on advertising and expect his sales to be dependent upon how much he spends
- It there is any lapse of time between the two variables being measured, then the latter must depend upon the former, it cannot be the other way round

- If we want to predict the values of one variable from your knowledge of the other variable, the variable to be predicted must be dependent on the known one

13.2.1 LINEAR REGRESSION

The task of bringing out linear relationship consists of developing methods of fitting a straight line, or a regression line as is often called, to the data on two variables.

The line of Regression is the graphical or relationship representation of the best estimate of one variable for any given value of the other variable. The nomenclature of the line depends on the independent and dependent variables. If *X* and *Y* are two variables of which relationship is to be indicated, a line that gives best estimate of *Y* for any value of *X*, it is called *Regression line of Y on X*. If the dependent variable changes to *X*, then best estimate of *X* by any value of *Y* is called Regression line of X on Y.

13.2.2 REGRESSION LINE OF *Y* ON *X*

For purposes of illustration as to how a straight line relationship is obtained, consider the sample paired data on sales of each of the $N = 5$ months of a year and the marketing expenditure incurred in each month, as shown in Table 5-1

Table 13-1

Month	Sales (Rs lac) *Y*	Marketing Expenditure (Rs thousands) *X*
April	14	10
May	17	12
June	23	15
July	21	20
August	25	23

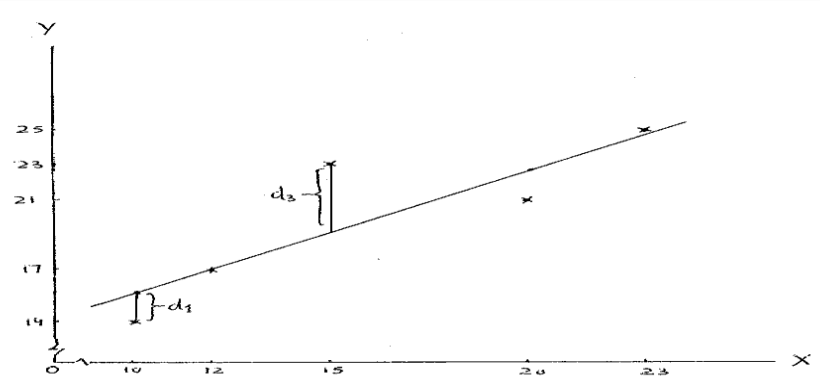
Let *Y*, the sales, be the dependent variable and *X*, the marketing expenditure, the independent variable. We note that for each value of independent variable *X*, there is a specific value of the dependent variable *Y*, so that each value of *X* and *Y* can be seen as paired observations.

13.2.2.1 Scatter Diagram

Before obtaining a straight-line relationship, it is necessary to discover whether the relationship between the two variables is linear, that is, the one which is best explained by a straight line. A good way of doing this is to plot the data on *X* and *Y* on a graph so as to yield a scatter diagram, as may be seen in Figure 13-1. A careful reading of the scatter diagram reveals that:

- The overall tendency of the points is to move upward, so the relationship is positive
- The general course of movement of the various points on the diagram can be best explained by a straight line
- There is a high degree of correlation between the variables, as the points are very close to each other

Figure 13-1 Scatter Diagram with Line of Best Fit



13.2.2.2 Fitting a Straight Line on the Scatter Diagram

If the movement of various points on the scatter diagram is best described by a straight line, the next step is to fit a straight line on the scatter diagram. It has to be so fitted that on the whole it lies as close as possible to every point on the scatter diagram. The necessary requirement for meeting this condition being that the sum of the squares of the vertical deviations of the observed Y values from the straight line is minimum.

As shown in Figure 5-1, if d_l, d_2,..., d_N are the vertical deviations' of observed Y values from the straight line, fitting a straight line requires that is the minimum.

$$d_1^2 + d_2^2 + \ldots\ldots\ldots\ldots\ldots + d_N^2 = \sum_{j=1}^{N} d_j^2$$

The deviations d_j have to be squared to avoid negative deviations canceling out the positive deviations. Since a straight line so fitted best approximates all the points on the scatter diagram, it is better known as the best approximating line or the line of best fit. A line of best fit can be fitted by means of:

1. Free hand drawing method, and
2. Least square method

Free Hand Drawing:

Free hand drawing is the simplest method of fitting a straight line. After a careful inspection of the movement and spread of various points on the scatter diagram, a straight line is drawn through these points by using a transparent ruler such that on the whole it is closest to every point. A straight line so drawn is particularly useful when future approximations of the dependent variable are promptly required.

Whereas the use of free hand drawing may yield a line nearest to the line of best fit, the major drawback is that the slope of the line so drawn varies from person to person because of the influence of subjectivity. Consequently, the values of the dependent variable estimated on the basis of such a line may not be as accurate and precise as those based on the line of best fit.

Least Square Method:

The least square method of fitting a line of best fit requires minimizing the sum of the squares of vertical deviations of each observed *Y* value from the fitted line. These deviations, such as d_1 and d_3, are shown in Figure 5-1 and are given by $Y - Y_c$, where *Y* is the observed value and Y_c the corresponding computed value given by the fitted line for the ith value of *X*.

$$Y_c = a + bX_i \quad \ldots\ldots\ldots (5.1)$$

The straight line relationship in *Eq.(5.1)*, is stated in terms of two constants *a* and *b*

- The constant *a* is the *Y*-intercept; it indicates the height on the vertical axis from where the straight line originates, representing the value of *Y* when *X* is zero.
- Constant *b* is a measure of the slope of the straight line; it shows the absolute change in *Y* for a unit change in *X*. As the slope may be positive or negative, it indicates the nature of relationship between *Y* and *X*. Accordingly, *b* is also known as the regression coefficient of *Y* on *X*.

Since a straight line is completely defined by its intercept *a* and slope *b*, the task of fitting the same reduces only to the computation of the values of these two constants. Once these two values are known, the computed Y_c values against each value of *X* can be easily obtained by substituting *X* values in the linear equation.

In the method of least squares the values of *a* and *b* are obtained by solving simultaneously the following pair of normal equations

$$\sum Y = aN + b\sum X \quad \ldots\ldots\ldots (5.2)$$

$$\sum XY = a\sum X + b\sum X^2 \quad \ldots\ldots (5.2)$$

The value of the expressions - $\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$ can be obtained from the given observations and then can be substituted in the above equations to obtain the value of *a* and *b.*

Since simultaneous solving the two normal equations for *a* and *b* may quite often be cumbersome and time consuming, the two values can be directly obtained as

$$a = \bar{Y} - b\bar{X} \quad \ldots\ldots\ldots(5.3)$$

and

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2} \ldots\ldots\ldots(5.4)$$

Note: *Eq. (5.3) is obtained simply by dividing both sides of the first of Eqs. (5.2) by N and Eq.(5.4) is obtained by substituting* ($\bar{Y} - b\bar{X}$) *in place of a in the second of Eqs. (5.2)*

Instead of directly computing *b*, we may first compute value of *a* as

$$a = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N\sum X^2 - \left(\sum X\right)^2} \ldots\ldots\ldots(5.5)$$

and

$$b = \frac{\bar{Y} - a}{\bar{X}} \ldots\ldots\ldots(5.6)$$

Note: *Eq. (5.5) is obtained by substituting* $\frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2}$ *for b in Eq. (5.3) and Eq. (5.6) is obtained simply by rearranging Eq. (5.3)*

Table 13-2

Computation of a *and* b

Y	X	XY	X^2	Y^2
14	10	140	100	196
17	12	204	144	289
23	15	345	225	529
21	20	420	400	441
25	23	575	529	625
$\sum Y = 100$	$\sum X = 80$	$\sum XY = 1684$	$\sum X^2 = 1398$	$\sum Y^2 = 2080$

So using *Eqs. (5.5)* and *(5.4)*

$$a = \frac{100x1398 - 80x1684}{5x1398 - (80)^2} = \frac{139800 - 134720}{6990 - 6400} = \frac{5080}{590} = 8.6101695$$

$$b = \frac{5x1684 - 80x100}{5x1398 - (80)^2} = \frac{8420 - 8000}{6990 - 6400} = \frac{420}{590} = 0.7118644$$

Now given $a = 8.61$ and $b = 0.71$

The regression *Eq.(5.1)* takes the form

$$Y_c = 8.61 + 0.71X \ldots\ldots\ldots\ldots (5.1a)$$

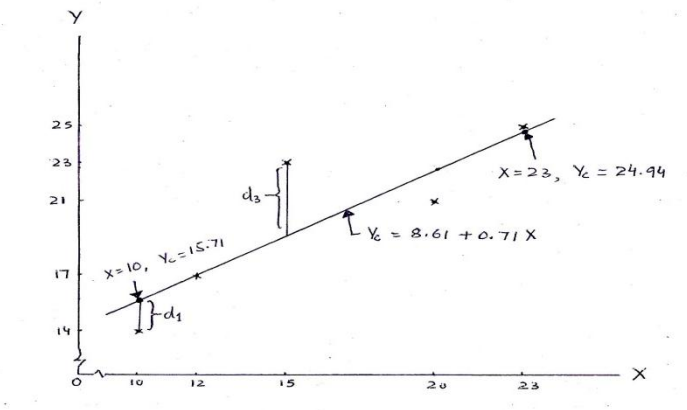


Figure 13-2 Regression Line of *Y* on *X*

Then, to fit the line of best fit on the scatter diagram, only two computed Y_c values are needed. These can be easily obtained by substituting any two values of *X* in *Eq. (5.1a).* When these are plotted on the diagram against their corresponding values of *X*, we get two points, by joining which (by means of a straight line) gives us the required line of best fit, as shown in Figure 5-2

Some Important Relationships

We can have some important relationships for data analysis, involving other measures such as $\overline{X}$, $\overline{Y}$, S_x, S_y and the correlation coefficient r_{xy}.

Substituting $\overline{Y} - b\overline{X}$ [from *Eq.(5.3)*] for *a* in *Eq.(5.1)*

$$Y_c = (\overline{Y} - b\overline{X}) + bX$$

or

$$Y_c - \overline{Y} = b(X - \overline{X}) \ldots\ldots\ldots\ldots (5.7)$$

Dividing the numerator and denominator of *Eq.(5.4)* by N^2, we get

$$b = \frac{\frac{\sum XY}{N} - \left(\frac{\sum X}{N}\right)\left(\frac{\sum Y}{N}\right)}{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2}$$

$$\text{or } b = \frac{\frac{\sum XY}{N} - \overline{X}\,\overline{Y}}{S_x^2}$$

$$\text{or } b = \frac{Cov(X,Y)}{S_x^2} \qquad \ldots\ldots\ldots\ldots (5.8)$$

We know, coefficient of correlation, r_{xy} is given by

$$r_{xy} = \frac{Cov(X,Y)}{S_x\, S_y}$$

$$\text{or } Cov(X,Y) = r_{xy} S_x\, S_y$$

So *Eq. (5.8)* becomes

$$b = r_{xy} \frac{S_x S_y}{S_x^2}$$

$$b = r_{xy} \frac{S_y}{S_x} \ldots\ldots\ldots\ldots (5.9)$$

Substituting $r_{xy} \frac{S_y}{S_x}$ for b in *Eq.(5.7),* we get

$$Y_c - \overline{Y} = r_{xy} \frac{S_y}{S_x} (X - \overline{X}) \ldots\ldots\ldots (5.10)$$

These are important relationships for data analysis.

13.2.2.3 Predicting an Estimate and its Preciseness

The main objective of regression analysis is to know the nature of relationship between two variables and to use it for predicting the most likely value of the dependent variable corresponding to a given, known value of the independent variable. This can be done by substituting in *Eq.(5.1a)* any known value of *X* corresponding to which the most likely estimate of *Y* is to be found.

For example, the estimate of Y (*i.e.* Y_c), corresponding to $X = 15$ is

$$Y_c = 8.61 + 0.71(15)$$

$$= 8.61 + 10.65$$

$$= 19.26$$

It may be appreciated that an estimate of Y derived from a regression equation will not be exactly the same as the Y value which may actually be observed. The difference between estimated Y_c values and the corresponding observed Y values will depend on the extent of scatter of various points around the line of best fit.

The closer the various paired sample points *(Y, X)* clustered around the line of best fit, the smaller the difference between the estimated Y_c and observed Y values, and vice-versa. On the whole, the lesser the scatter of the various points around, and the lesser the vertical distance by which these deviate from the line of best fit, the more likely it is that an estimated Y_c value is close to the corresponding observed Y value.

The estimated Y_c values will coincide the observed Y values only when all the points on the scatter diagram fall in a straight line. If this were to be so, the sales for a given marketing expenditure could have been estimated with l00 percent accuracy. But such a situation is too rare to obtain. Since some of the points must lie above and some below the straight line, perfect prediction is practically non-existent in the case of most business and economic situations.

This means that the estimated values of one variable based on the known values of the other variable are always bound to differ. The smaller the difference, the greater the precision of the estimate, and vice-versa. Accordingly, the preciseness of an estimate can be obtained only through a measure of the magnitude of error in the estimates, called the error of estimate.

13.2.2.4 Error of Estimate

A measure of the error of estimate is given by the standard error of estimate of *Y on X*, denoted as S_{yx} and defined as

$$S_{yx} = \sqrt{\frac{\sum (Y - Y_c)^2}{N}} \quad \text{……..(5.11)}$$

S_{yx} measures the average absolute amount by which observed Y values depart from the corresponding computed Y_c values. Computation of S_{yx} becomes little cumbersome where the number of observations N is large. In such cases S_{yx} may be computed directly by using the equation:

$$S_{yx} = \sqrt{\frac{\sum Y^2 - a\left(\sum Y\right) - b\sum XY}{N}} \quad \ldots\ldots\ldots(5.12)$$

By substituting the values of $\sum Y^2$, $\sum Y$, and $\sum XY$ from the Table 5-2, and the calculated values of *a* and *b*

We have

$$S_{yx} = \sqrt{\frac{2080 - 8.61x100 - 0.71x1684}{5}}$$

$$= \sqrt{\frac{2080 - 861 - 1195.64}{5}} = \sqrt{\frac{23.36}{5}}$$

$$= \sqrt{4.67} = 2.16$$

Interpretations of S_{yx}

A careful observation of how the standard error of estimate is computed reveals the following:

1. S_{yx} is a concept statistically parallel to the standard deviation S_y . The only difference between the two being that the standard deviation measures the dispersion around the mean; the standard error of estimate measures the dispersion around the regression line. Similar to the property of arithmetic mean, the sum of the deviations of different *Y* values from their corresponding estimated Y_c values is equal to zero. That is

   $$\sum(Y_i - \bar{Y}) = \sum (Y_i - Y_c) = 0 \text{ where } i = 1, 2, ..., N.$$

2. S_{yx} tells us the amount by which the estimated Y_c values will, on an average, deviate from the observed *Y* values. Hence it is an estimate of the average amount of error in the estimated Y_c values. The actual error (the residual of *Y* and Y_c) may, however, be smaller or larger than the average error. Theoretically, these errors follow a normal distribution. Thus, assuming that $n \geq 30$, $Y_c \pm 1.S_{yx}$ means that 68.27% of the estimates based on the regression equation will be within $1.S_{yx}$ Similarly, $Y_c \pm 2.S_{yx}$ means that 95.45% of the estimates will fall within $2.S_{yx}$

   Further, for the estimated value of sales against marketing expenditure of Rs 15 thousand being Rs 19.26 lac, one may like to know how good this estimate is. Since S_{yx} is estimated to be Rs 2.16 lac, it means there are about 68 chances (68.27) out of 100 that this estimate is in error by not more than Rs 2.16 lac above or below Rs 19.26 lac. That is, there are 68% chances that actual sales would fall between (19.26 - 2.16) = Rs 17.10 lac and (19.26 + 2.16) = Rs 21.42 lac.

3. Since S_{yx} measures the closeness of the observed Y values and the estimated Y_c values, it also serves as a measure of the reliability of the estimate. Greater the closeness between the observed and estimated values of Y, the lesser the error and, consequently, the more reliable the estimate. And vice-versa.
4. Standard error of estimate S_{yx} can also be seen as a measure of correlation insofar as it expresses the degree of closeness of scatter of observed Y values about the regression line. The closer the observed Y values scattered around the regression line, the higher the correlation between the two variables.

A major difficulty in using S_{yx} as a measure of correlation is that it is expressed in the same units of measurement as the data on the dependent variable. This creates problems in situations requiring comparison of two or more sets of data in terms of correlation. It is mainly due to this limitation that the standard error of estimate is not generally used as a measure of correlation. However, it does serve as the basis of evolving the coefficient of determination, denoted as r^2, which provides an alternate method of obtaining a measure of correlation.

13.2.3 REGRESSION LINE OF *X* ON *Y*

So far we have considered the regression of Y on X, in the sense that Y was in the role of dependent and X in the role of an independent variable. In their reverse position, such that X is now the dependent and Y the independent variable, we fit a line of regression of X on Y. The regression equation in this case will be

$$X_c = a' + b'Y \ldots\ldots\ldots\ldots (5.13)$$

Where X_c denotes the computed values of X against the corresponding values of Y. a' is the X-intercept and b' is the slope of the straight line.

Two normal equations to solve $a'$and b' are

$$\sum X = a'N + b'\sum Y \ldots\ldots (5.14)$$

$$\sum XY = a'\sum Y + b'\sum Y^2 \ldots\ldots (5.14)$$

The value of a' and b' can also be obtained directly

$$a' = \overline{X} - b'\overline{Y} \ldots\ldots\ldots\ldots (5.15)$$

and

$$b' = \frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - \left(\sum Y\right)^2} \ldots\ldots\ldots\ldots (5.16)$$

or

$$a' = \frac{\sum X \sum Y^2 - \sum Y \sum XY}{N\sum Y^2 - \left(\sum Y\right)^2} \quad \ldots\ldots\ldots\ldots (5.17)$$

and

$$b' = \frac{\overline{X} - a'}{\overline{Y}} \quad \ldots\ldots\ldots\ldots (5.18)$$

$$b' = \frac{Cov(Y, X)}{S_y^2} \quad \ldots\ldots\ldots\ldots (5.19)$$

$$b' = r_{yx} \frac{S_x}{S_y} \quad \ldots\ldots\ldots\ldots (5.20)$$

So, Regression equation of X on Y may also be written as

$$X_c - \overline{X} = b' (Y - \overline{Y}) \ldots\ldots\ldots\ldots (5.21)$$

$$X_c - \overline{X} = r_{yx} \frac{S_x}{S_y} (Y - \overline{Y}) \ldots\ldots\ldots\ldots (5.22)$$

As before, once the values of a' and b' have been found, their substitution in *Eq.(5.13)* will enable us to get an estimate of X corresponding to a known value of Y

Standard Error of estimate of X on Y *i.e.* S_{xy} will be

$$S_{xy} = \sqrt{\frac{(X - X_c)^2}{N}} \quad \ldots\ldots\ldots\ldots (5.23)$$

or

$$S_{xy} = \sqrt{\frac{\sum X^2 - a'\sum X - b'\sum XY}{N}} \quad \ldots\ldots\ldots\ldots (5.24)$$

For example, if we want to estimate the marketing expenditure to achieve a sale target of Rs 40 lac, we have to obtain regression line of X on Y *i. e.*

$$X_c = a' + b'Y$$

So using *Eqs. (5.17)* and *(5.16),* and substituting the values of $\sum X, \sum Y^2, \sum Y$ and $\sum XY$ from Table 5-2, we have

$$a' = \frac{80x2080 - 100x1684}{5x2080 - (100)^2} = \frac{166400 - 168400}{10400 - 10000} = \frac{-2000}{400} = -5.00$$

$$b' = \frac{5x1684 - 80x100}{5x2080 - (100)^2} = \frac{8420 - 8000}{10400 - 10000} = \frac{420}{400} = 1.05$$

Now given that a'= -5.00 and b'=1.05, Regression equation *(5.13)* takes the form

$$X_c = -5.00 + 1.05Y$$

So when Y = 40(Rs lac), the corresponding X value is

$$X_c = -5.00 + 1.05x40$$

$$= -5 + 42$$

$$= 37$$

That is to achieve a sale target of Rs 40 lac, there is a need to spend Rs 37 thousand on marketing.

13.3 PROPERTIES OF REGRESSION COEFFICIENTS

As explained earlier, the slope of regression line is called the regression coefficient. It tells the effect on dependent variable if there is a unit change in the independent variable. Since for a paired data on X and Y variables, there are two regression lines: regression line of Y on X and regression line of X on Y, so we have two regression coefficients:

(i) Regression coefficient of Y on X, denoted by b_{yx} [b in *Eq.(5.1)*]

(ii) Regression coefficient of X on Y, denoted by b_{xy} [b' in *Eq.(5.13)*]

The following are the important properties of regression coefficients that are helpful in data analysis

1. The value of both the regression coefficients cannot be greater than 1. However, value of both the coefficients can be below 1 or at least one of them must be below 1, so that the square root of the product of two regression coefficients must lie in the limit ±1.
2. Coefficient of correlation is the geometric mean of the regression coefficients, *i.e*.

   $$r = \pm\sqrt{b.b'} \quad \ldots\ldots\ldots\ldots (5.25)$$

   The signs of both the regression coefficients are the same, and so the value of r will also have the same sign.
3. The mean of both the regression coefficients is either equal to or greater than the coefficient of correlation, *i.e.*

$$\frac{b+b'}{2} \geq r$$

4. Regression coefficients are independent of change of origin but not of change of scale. Mathematically, if given variables *X* and *Y* are transformed to new variables *U* and *V* by change of origin and scale, *i. e.*

$$U = \frac{X-A}{h} \quad \text{and} \quad V = \frac{Y-B}{k}$$

Where *A, B, h* and *k* are constants, *h > 0, k > 0* then

Regression coefficient of *Y* on *X* = *k/h* (Regression coefficient of *V* on *U*)

$$b_{yx} = \frac{k}{h} b_{vu} \quad \text{and}$$

Regression coefficient of *X* on *Y* = *h/k* (Regression coefficient of *U* on *V*)

$$b_{xy} = \frac{h}{k} b_{uv}$$

5. Coefficient of determination is the product of both the regression coefficients *i.e.*

$$r^2 = b.b'$$

13.3.1 REGRESSION LINES AND COEFFICIENT OF CORRELATION

The two regression lines indicate the nature and extent of correlation between the variables.

The two regression lines can be represented as

$$Y - \bar{Y} = r\frac{S_y}{S_x}(X - \bar{X}) \quad \text{and} \quad X - \bar{X} = r\frac{S_x}{S_y}(Y - \bar{Y})$$

We can write the slope of these lines, as

$$b = r\frac{S_y}{S_x} \quad \text{and} \quad b' = r\frac{S_x}{S_y}$$

If θ is the angle between these lines, then

$$\tan\theta = \frac{b-b'}{1+bb'} = \frac{S_x S_y}{S_x^2 + S_y^2}\left(\frac{r^2-1}{r}\right)$$

$$\text{or } \theta = \tan^{-1}\left[\frac{S_x S_y}{S_x^2 + S_y^2}\left(\frac{r^2-1}{r}\right)\right] \ldots\ldots\ldots\ldots (5.26)$$

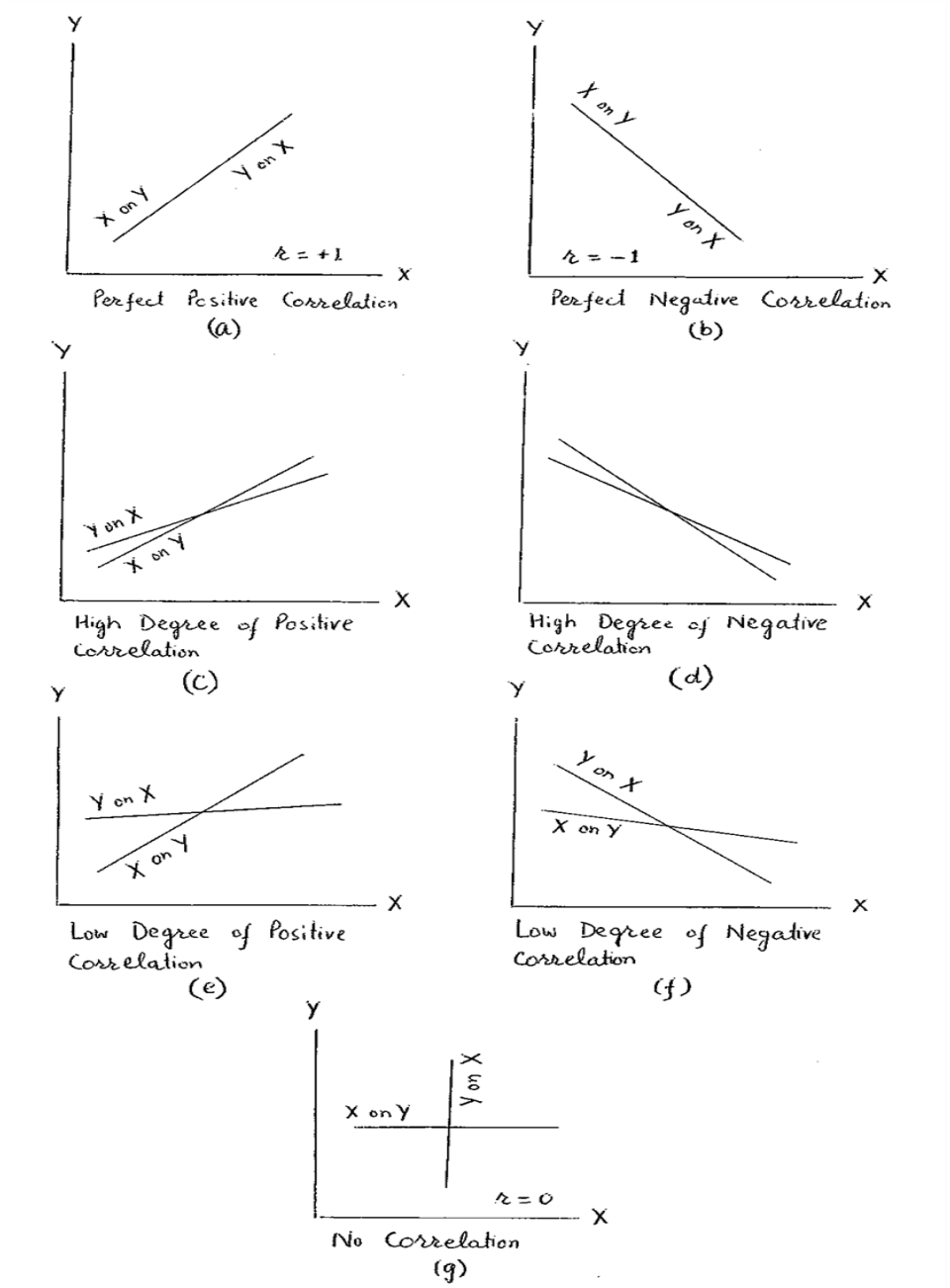


Figure 13-3 Regression Lines and Coefficient of Correlation

Eq. (5.26) reveals the following:

- In case of perfect positive correlation (r = +1) and in case of perfect negative correlation (r = -1), $\theta = 0$, so the two regression lines will coincide, *i.e.* we have only one line, see (a) and (b) in Figure 5-3.

The farther the two regression lines from each other, lesser will be the degree of correlation and nearer the two regression lines, more will be the degree of correlation, see (c) and (d) in Figure 5-3.

- If the variables are independent *i.e.* $r = 0$, the lines of regression will cut each other at right angle. See (g) in Figure 5-3.

Note: *Both the regression lines cut each other at mean value of X and mean value of Y i.e. at* $\overline{X}$ *and* $\overline{Y}$.

13.3.2 COEFFICIENT OF DETERMINATION

Coefficient of determination gives the percentage variation in the dependent variable that is accounted for by the independent variable. In other words, the coefficient of determination gives the ratio of the explained variance to the total variance. The coefficient of determination is given by the square of the correlation coefficient, *i.e.* r^2. Thus, Coefficient of determination

$$r^2 = \frac{Explained Variance}{Total Variance}$$

$$r^2 = \frac{\sum\left(Y_c - \overline{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2} \quad \ldots\ldots\ldots\ldots (5.27)$$

We can calculate another coefficient K^2, known as coefficient of Non-Determination, which is the ratio of unexplained variance to the total variance.

$$K^2 = \frac{Unexplained Variance}{Total Variance}$$

$$K^2 = \frac{\sum\left(Y - Y_c\right)^2}{\sum\left(Y - \overline{Y}\right)^2} \quad \ldots\ldots (5.28)$$

$$K^2 = 1 - \frac{Explained Variance}{Total Variance}$$

$$= 1 - r^2 \ldots\ldots (5.29)$$

The square root of the coefficient of non-determination, *i.e. K* gives the coefficient of alienation

$$K = \pm\sqrt{1 - r^2} \quad \ldots\ldots\ldots (5.30)$$

Relation Between S_{yx} and *r*:

A simple algebraic operation on *Eq. (5.30)* brings out some interesting points about the relation between S_{yx} and *r*. Thus, since

$$\sum (Y - Y_c)^2 = N S_{yx}^2 \quad \text{and} \quad \sum (Y - \bar{Y})^2 = N S_y^2$$

So we have coefficient of Non-determination

$$K^2 = \frac{\sum (Y - Y_c)^2}{\sum (Y - \bar{Y})^2}$$

$$K^2 = \frac{N S_{yx}^2}{N S_y^2} = \frac{S_{yx}^2}{S_y^2}$$

So $$1 - r^2 = \frac{S_{yx}^2}{S_y^2}$$

or $$\frac{S_{yx}}{S_y} = \sqrt{1 - r^2} \quad \ldots\ldots\ldots\ldots (5.31)$$

If coefficient of correlation, *r*, is defined as the under root of the coefficient of determination

$$r = \sqrt{r^2}$$

$$r^2 = 1 - \frac{S_{yx}^2}{S_y^2}$$

$$r = \sqrt{1 - \frac{S_{yx}}{S_y^2}} \quad \ldots\ldots\ldots\ldots (5.32)$$

On carefully observing *Eq. (5.32)*, it will be noticed that the ratio S_{yx}/S_y will be large if the coefficient of determination is small, and it will be small when the coefficient of determination is large. Thus

- ✓ if $r^2 = r = 0$, $S_{yx}/S_y = 1$, which means that $S_{yx} = S_y$.
- ✓ if $r^2 = r = 1$, $S_{yx}/S_y = 0$, which means that $S_{yx} = 0$.
- ✓ when $r = 0.865$, $S_{yx} = 0.427\ S_y$ means that S_{yx} is 42.7% of S_y.

Eq. (5.32) also implies that S_{yx} is generally less than S_y. The two can at the most be equal, but only in the extreme situation when $r = 0$.

Interpretations of r^2:

1. Even though the coefficient of determination, whose under root measures the degree of correlation, is based on S_{yx},; it is expressed as 1 - (S_{yx}/S_y). As it is a dimensionless pure number, the unit in which S_{yx} is measured becomes irrelevant. This facilitates comparison between the two sets of data

in terms of their coefficient of determination r^2 (or the coefficient of correlation *r).* This was not possible in terms of S_{yx} as the units of measurement could be different.

2. The value of r^2 can range between 0 and 1. When $r^2 = 1$, all the points on the scatter diagram fall on the regression line and the entire variations are explained by the straight line. On the other hand, when $r^2 = 0$, none of the points on the scatter diagram falls on the regression line, meaning thereby that there is no relationship between the two variables. However, being always non-negative coefficient of determination does not tell us about the direction of the relationship (whether it is positive or negative) between the two variables.
3. When $r^2 = 0.7455$ (or any other value), 74.55% of the total variations in sales are explained by the marketing expenditure used. What remains is the coefficient of non-determination $K^2 (= 1 - r^2) = 0.2545$. It means 25.45% of the total variations remain unexplained, which are due to factors other than the changes in the marketing expenditure.
4. r^2 provides the necessary link between regression and correlation which are the two related aspects of a single problem of the analysis of relationship between two variables. Unlike regression, correlation quantifies the degrees of relationship between the variables under study, without making a distinction between the dependent and independent ones. Nor does it, therefore, help in predicting the value of one variable for a given value of the other.
5. The coefficient of correlation overstates the degree of relationship and it's meaning is not as explicit as that of the coefficient of determination. The coefficient of correlation $r = 0.865$, as compared to $r^2 = 0.7455$, indicates a higher degree of correlation between sales and marketing expenditure. Therefore, the coefficient of' determination is a more objective measure of the degree of relationship.
6. The sum of r and K never adds to one, unless one of the two is zero. That is, $r + K$ can be unity either when there is no correlation or when there is perfect correlation. Except in these two extreme situations, $(r + K) > 1$.

13.3.3 CORRELATION ANALYSIS VERSUS REGRESSION ANALYSIS

Correlation and Regression are the two related aspects of a single problem of the analysis of the relationship between the variables. If we have information on more than one variable, we might be interested in seeing if there is any connection - any association - between them. If we found such a

association, we might again be interested in predicting the value of one variable for the given and known values of other variable(s).

1. Correlation literally means the relationship between two or more variables that vary in sympathy so that the movements in one tend to be accompanied by the corresponding movements in the other(s). On the other hand, regression means stepping back or returning to the average value and is a mathematical measure expressing the average relationship between the two variables.
2. Correlation coefficient r_{xy} between two variables X and Y is a measure of the direction and degree of the linear relationship between two variables that is mutual. It is symmetric, *i.e.*, $r_{yx} = r_{xy}$ and it is immaterial which of X and Y is dependent variable and which is independent variable. Regression analysis aims at establishing the functional relationship between the two(or more) variables under study and then using this relationship to predict or estimate the value of the dependent variable for any given value of the independent variable(s). It also reflects upon the nature of the variable, *i.e.*, which is dependent variable and which is independent variable. Regression coefficient are not symmetric in X and Y, *i.e.*, $b_{yx} \neq b_{xy}$.
3. Correlation need not imply cause and effect relationship between the variable under study. However, regression analysis clearly indicates the cause and effect relationship between the variables. The variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.
4. Correlation coefficient r_{xy} is a relative measure of the linear relationship between X and Y and is independent of the units of measurement. It is a pure number lying between ±1. On the other hand, the regression coefficients, b_{yx} and b_{xy} are absolute measures representing the change in the value of the variable Y (or X), for a unit change in the value of the variable X (or Y). Once the functional form of regression curve is known; by substituting the value of the independent variable we can obtain the value of the dependent variable and this value will be in the units of measurement of the dependent variable.
5. There may be non-sense correlation between two variables that is due to pure chance and has no practical relevance, *e.g.*, the correlation, between the size of shoe and the intelligence of a group of individuals. There is no such thing like non-sense regression.

13.4 SOLVED PROBLEMS

Example 13-1

The following table shows the number of motor registrations in a certain territory for a term of 5 years and the sale of motor tyres by a firm in that territory for the same period.

Year	Motor Registrations	No. of Tyres Sold
1	600	1,250
2	630	1,100
3	720	1,300
4	750	1,350
5	800	1,500

Find the regression equation to estimate the sale of tyres when the motor registration is known. Estimate sale of tyres when registration is 850.

Solution: Here the dependent variable is number of tyres; dependent on motor registrations. Hence we put motor registrations as X and sales of tyres as Y and we have to establish the regression line of Y on X.

Calculations of values for the regression equation are given below:

X	Y	$d_x = X - \overline{X}$	$d_y = Y - \overline{Y}$	d_x^2	$d_x d_y$
600	1,250	-100	-50	10,000	5,000
630	1,100	-70	-200	4,900	14,000
720	1,300	20	0	400	0
750	1,350	50	50	2,500	2,500
800	1,500	100	200	10,000	20,000
$\sum X =$ 3,500	$\sum Y = 6,500$	$\sum d_x = 0$	$\sum d_y = 0$	$\sum d_x^2 =$ 27,800	$\sum d_x d_y =$ 41,500

$$\overline{X} = \frac{\sum X}{N} = \frac{3,500}{5} = 700 \quad \text{and} \quad \overline{Y} = \frac{\sum Y}{N} = \frac{6,500}{5} = 1,300$$

b_{yx} = Regression coefficient of Y on X

$$b_{yx} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{\sum d_x d_y}{\sum d_x^2} = \frac{4{,}1500}{2{,}7800} = 1.4928$$

Now we can use these values for the regression line

$$Y-\bar{Y} = b_{yx}\ (X-\bar{X})$$

$$\text{or } Y - 1300 = 1.4928\ (X - 700)$$

$$Y = 1.4928\ X + 255.04$$

When X = 850, the value of Y can be calculated from the above equation, by putting X = 850 in the equation.

$$Y = 1.4928 \times 850 + 255.04$$

$$= 1523.92$$

$$= 1{,}524 \text{ Tyres}$$

Example 13-2

A panel of Judges A and B graded seven debators and independently awarded the following marks:

Debator	Marks by A	Marks by B
1	40	32
2	34	39
3	28	26
4	30	30
5	44	38
6	38	34
7	31	28

An eighth debator was awarded 36 marks by judge A, while Judge B was not present. If Judge B were also present, how many marks would you expect him to award to the eighth debator, assuming that the same degree of relationship exists in their judgement?

Solution:

Let us use marks from Judge A as *X* and those from Judge B as *Y*. Now we have to work out the regression line of *Y* on *X* from the calculation below:

Debtor	*X*	*Y*	*U = X-35*	*V = Y-30*	U^2	V^2	*UV*
1	40	32	5	2	25	4	10
2	34	39	-1	9	1	81	-9
3	28	26	-7	-4	49	16	28
4	30	30	-5	0	25	0	0
5	44	38	9	8	81	64	72
6	38	34	3	4	9	16	12
7	31	28	-4	-2	16	4	8
N = 7			$\sum U = 0$	$\sum V = 17$	$\sum U^2 = 206$	$\sum V^2 = 185$	$\sum UV = 121$

$$\bar{X} = A + \frac{\sum U}{N} = 35 + \frac{0}{7} = 35 \qquad \text{and} \qquad \bar{Y} = A + \frac{\sum V}{N} = 30 + \frac{17}{7} = 32.43$$

$$b_{yx} = b_{vu} = \frac{N\sum UV - \left(\sum U \sum V\right)}{N\sum U^2 - \left(\sum U\right)^2} = \frac{7x121 - 0x17}{7x206 - 0} = 0.587$$

Hence regression equation can be written as

$Y - \bar{Y} = b_{yx}\ (X - \bar{X})$

$Y - 32.43 = 0.587\ (X-35)$

or $Y = 0.587X + 11.87$

When $X = 36$ (awarded by Judge A)

$Y = 0.587 \times 36 + 11.87$

$= 33$

Thus if Judge B were present, he would have awarded 33 marks to the eighth debator.

Example 13-3

For some bivariate data, the following results were obtained.

Mean value of variable X = 53.2

Mean value of variable Y = 27.9

Regression coefficient of Y on X = - 1.5

Regression coefficient of X on Y = - 0.2

What is the most likely value of Y, when $X = 60$?

What is the coefficient of correlation between X and Y?

Solution: Given data indicate

$\overline{X}$ = 53.2 $\overline{Y}$ = 27.9

b_{yx} = -1.5 b_{xy} = -0.2

To obtain value of Y for $X = 60$, we establish the regression line of Y on X,

$$Y - \overline{Y} = b_{yx}\ (X - \overline{X})$$

$$Y - 27.9 = -1.5\ (X - 53.2)$$

or $$Y = -1.5X + 107.7$$

Putting value of $X = 60$, we obtain

$$Y = -1.5 \times 60 + 107.7$$

$$= 17.7$$

Coefficient of correlation between X and Y is given by G.M. of b_{yx} and b_{xy}

$$r^2 = b_{yx}\, b_{xy}$$

$$= (-1.5) \times (-0.2)$$

$$= 0.3$$

So $$r = \pm\sqrt{0.3} = \pm 0.5477$$

Since both the regression coefficients are negative, we assign negative value to the correlation coefficient

$$r = - 0.5477$$

Example 13-4

Write regression equations of *X* on *Y* and of *Y* on *X* for the following data

X:	45	48	50	55	65	70	75	72	80	85
Y:	25	30	35	30	40	50	45	55	60	65

Solution:

We prepare the table for working out the values for the regression lines.

X	*Y*	*U = X-65*	*V = Y-45*	U^2	*UV*	V^2
45	25	-20	-20	400	400	400
48	30	-17	-15	289	255	225
50	35	-15	-10	225	150	100
55	30	-10	-15	100	150	225
65	40	0	-5	0	0	25
70	50	5	5	25	25	25
75	45	10	0	100	0	0
72	55	7	5	49	35	25
80	60	15	15	225	225	225
85	65	20	20	400	400	400
$\sum X =$ 645	$\sum Y = 435$	$\sum U = 5$	$\sum V = -20$	$\sum U^2 = 1813$	$\sum V^2 = 1415$	$\sum UV = 1675$

We have, $\bar{X} = \frac{\sum X}{N} = \frac{645}{10} = 64.5$ and $\bar{Y} = \frac{\sum Y}{N} = \frac{435}{10} = 43.5$

$$b_{yx} = \frac{N\sum UV - \left(\sum U \sum V\right)}{N\sum U^2 - \left(\sum U\right)^2} = \frac{(10)x1415 - (5)x(-20)}{(10)x1813 - (5)^2}$$

$$= \frac{14150+100}{18130-25} = \frac{14250}{18105} = 0.787$$

Regression equation of *Y* on *X* is

$$Y-\bar{Y} = b_{yx}\ (X-\bar{X})$$

$$Y - 43.5 = 0.787\ (X-64.5)$$

or $$Y = 0.787X + 7.26$$

Similarly b_{xy} can be calculated as

$$b_{xy} = \frac{N\sum UV - \left(\sum U \sum V\right)}{N\sum V^2 - \left(\sum V\right)^2} = \frac{(10)x1415-(5)x(-20)}{(10)x1675-(-20)^2}$$

$$= \frac{14150+100}{16750-400} = \frac{14250}{16350} = 0.87$$

Regression equation of X on *Y* will be

$$X - \bar{X} = b_{xy}\ (Y-\bar{Y})$$

$$X - 64.5 = 0.87\ (Y-43.5)$$

or $$X = 0.87Y + 26.65$$

Example 13-5

The lines of regression of a bivariate population are

$$8X - 10Y + 66 = 0$$

$$40X - 18Y = 214$$

The variance of *X* is 9. Find

(i) The mean value of *X* and *Y*

(ii) Correlation coefficient between X and Y

(iii) Standard deviation of Y

Solution: The regression lines given are

$$8X - 10Y + 66 = 0$$

$$40X - 18Y = 214$$

Since both the lines of regression pass through the mean values, the point $(\bar{X}, \bar{Y})$ will satisfy both the equations.

Hence these equations can be written as

$$8\ \bar{X} - 10\ \bar{Y} + 66 = 0$$

$$40\ \bar{X} - 18\bar{Y} - 214 = 0$$

Solving these two equations for $\overline{X}$ and $\overline{Y}$, we obtain

$\overline{X} = 13$ and $\overline{Y} = 17$

(ii) For correlation coefficient between *X* and *Y*, we have to calculate the values of b_{yx} and b_{xy}

Rewriting the equations

$10Y = 8X + 66$

$b_{yx} = + 8/10 = + 4/5$

Similarly, $40X = 18Y + 214$

$b_{xy} = 18/40 = 9/20$

By these values, we can now work out the correlation coefficient.

$r^2 = b_{yx} \cdot b_{xy}$

$= 4/5 \times 9/20 = 9/25$

So $r = \pm\sqrt{9/25}$

$= \pm 0.6$

Both the values of the regression coefficients being positive, we have to consider only the positive value of the correlation coefficient. Hence $r = 0.6$

(iii) We have been given variance of *X i.e* $S_x^2 = 9$

$S_x = \pm 3$

We consider $S_x = 3$ as SD is always positive

Since $b_{yx} = r\, S_y / S_x$

Substituting the values of b_{yx}, r and S_x we obtain,

$S_y = 4/5 \times 3/0.6$

$= 4$

Example 13-6

The height of a child increases at a rate given in the table below. Fit the straight line using the method of least-square and calculate the average increase and the standard error of estimate.

Month:	1	2	3	4	5	6	7	8	9	10
Height:	52.5	58.7	65	70.2	75.4	81.1	87.2	95.5	102.2	108.4

Solution:

For Regression calculations, we draw the following table

Month (X)	Height (Y)	X^2	XY
1	52.5	1	52.5
2	58.7	4	117.4
3	65.0	9	195.0
4	70.2	16	280.8
5	75.4	25	377.0
6	81.1	36	486.6
7	87.2	49	610.4
8	95.5	64	764.0
9	102.2	81	919.8
10	108.4	100	1084.0
$\sum X$ =55	$\sum Y$ =796.2	$\sum X^2 = 385$	$\sum XY = 4887.5$

Considering the regression line as $Y = a + bX$, we can obtain the values of a and b from the above values.

$$a = \frac{\sum Y \sum X^2 - \sum X \sum XY}{N\sum X^2 - \left(\sum X\right)^2} = \frac{796.2 x 385 \ - 55 \ x 4887.5}{10 x 385 - 55 x 55} = 45.73$$

$$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2} = \frac{10 x 4887.5 - 55 \ x 796.2}{10 x 385 - 55 x 55} = 6.16$$

Hence the regression line can be written as

$$Y = 45.73 + 6.16X$$

For standard error of estimation, we note the calculated values of the variable against the observed values,

When $X = 1$, $Y_1 = 45.73 + 6.16 = 51.89$

for $X = 2$, $Y_2 = 45.73 + 616 \ x \ 2 = 58.05$

Other values for $X = 3$ to $X = 10$ are calculated and are tabulated as follows:

Month (X)	Height (Y)	Y_i	$Y-Y_i$	$(Y-Y_i)^2$
1	52.5	51.89	0.61	0.372
2	58.7	58.05	0.65	0.423
3	65.0	64.21	0.79	0.624
4	70.2	70.37	-0.17	0.029
5	75.4	76.53	-1.13	1.277
6	81.1	82.69	-1.59	2.528
7	87.2	88.85	-1.65	2.723
8	95.5	95.01	0.49	0.240
9	102.2	101.17	1.03	1.061
10	108.4	107.33	1.07	1.145
				$\sum(Y-Y_i)^2 = 10.421$

Standard error of estimation

$$S_{yx} = \sqrt{\frac{1}{N}\sum(Y-Y_i)^2}$$

$$= \sqrt{\frac{10.421}{10}}$$

$$= 1.02$$

Example 13-7

Given $X = 4Y+5$ and $Y = kX + 4$ are the lines of regression of X on Y and of Y on X respectively. If k is positive, prove that it cannot exceed ¼.

If k = 1/16, find the means of the two variables and coefficient of correlation between them.

Solution: Line $X = 4Y + 5$ is regression line of X on Y

So $b_{xy} = 4$

Similarly from regression line of Y on X , $Y = kX + 4$,

We get $b_{yx} = k$

Now

$$r^2 = b_{xy}.\ b_{yx}$$

$$= 4k$$

Since $0 \le r^2 \le 1$, we obtain $0 \le 4k \le 1$,

Or $$0 \le k \le \frac{1}{4},$$

Now for k = $\frac{1}{16}$,

$$r^2 = 4x\frac{1}{16} = \frac{1}{4}$$

$$r = + ½$$

= ½ since b_{yx} and b_{yx} are positive

When k = $\frac{1}{16}$, the regression line of Y on X becomes

$$Y = \frac{1}{16}X + 4$$

Or $$X - 16Y + 64 = 0$$

Since line of regression pass through the mean values of the variables, we obtain revised equations as

$$\overline{X} - 4\overline{Y} - 5 = 0$$

$$\overline{X} - 16\overline{Y} + 64 = 0$$

Solving these two equations, we get

$\overline{X}$ = 28 and $\overline{Y}$ = 5.75

Example 13-8

A firm knows from its past experience that its monthly average expenses (X) on advertisement are Rs 25,000 with standard deviation of Rs 25.25. Similarly, its average monthly product sales (Y) have been Rs 45,000 with standard deviation of Rs 50.50. Given this information and also the coefficient of correlation between sales and advertisement expenditure as 0.75, estimate

(i) the most appropriate value of sales against an advertisement expenditure of Rs 50,000.

(ii) the most appropriate advertisement expenditure for achieving a sales target of Rs 80,000

Solution: Given the following

$\overline{X}$ = Rs 25,000 S_x = Rs 25.25

$\overline{Y}$ = Rs 45,000 S_y = Rs 50.50

r = 0.75

(i) Using equation $Y_c - \overline{Y} = r \frac{S_y}{S_x} (X - \overline{X})$, the most appropriate value of sales Y_c for an advertisement expenditure X = Rs 50,000 is

$$Y_c - 45{,}000 = 0.75 \frac{50.50}{25.25}(50{,}000 - 25{,}000)$$

$$Y_c = 45{,}000 + 37{,}500$$

$$= \text{Rs } 82{,}500$$

(ii) Using equation $X_c - \overline{X} = r \frac{S_x}{S_y}(Y - \overline{Y})$, the most appropriate value of advertisement expenditure X_c for achieving a sales target Y= Rs 80,000 is

$$X_c - 25{,}000 = 0.75 \frac{25.25}{50.50}(80{,}000 - 45{,}000)$$

$$X_c = 13{,}125 + 25{,}000$$

$$= \text{Rs } 38{,}125$$

13.5 MULTIPLE REGRESSION ANALYSIS

In multiple regression analysis, the effect of two or more independent variables on one dependent variable is studied. It uses three or more variables to estimate the value of dependent variable. Let's take three variables say X_1, X_2 and X_3. Now, take X_1 as the dependent variable and try to find out its relative movement for movements in both X_2 and X_3 which are independent variables. The prime objectives of multiple regression analysis are:

- To estimate an equation which provides estimates of the dependent variable from the values of the two or more independent variables.
- To obtain a measure of error involved in using regression equation as a basis for estimation.
- To obtain a measure of the proportion of variance in the dependent variable explained by the independent variables.

The first objective is accomplished by estimating an appropriate regression equation by the method of least squares. The second objective is achieved through the calculation of a standard error of estimate. The third objective is attained by computing the multiple coefficient of determination.

13.5.1 Regression Equations

A regression equation is an equation for estimating a dependent variable say X_1 from the independent variables X_2, X_3 and is called a regression equation of X_1 on X_2 and X_3. The procedure of estimating multiple regression is similar to the simple regression with the difference that the other variables are added in the regression equation. If there are three variables X_1 X_2 and X_3, the multiple regression has the following form:

$$\boldsymbol{X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3}$$

There are three constants $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ in the above equation. The subscript after the dot indicates the variables which are held constant.

Interpretation of Constants: In the above equation:

$a_{1.23}$ = The constant **$a_{1.23}$** is the intercept made by the regression plane. It gives the value of dependent variable when X_2 and X_3 independent variables are 0.

$b_{12.3}$ = It indicates the slope of the regression line of X_1 on X_2 when X_3 is held constant. It measures the amount by which a unit change in X_2 is expected to affect X_1 when X_3 is held constant.

$b_{13.2}$ = It indicates the slope of the regression line of X_1 on X_3 when X_2 is held constant. It measures the amount by which a unit change in X_3 is expected to affect X_1 when X_2 is held constant.

In this way, three different regression equations can be formed using three variables X_1, X_2 and X_3 which are explained below:

a. The regression plane of X_1 on X_2 and X_3
b. The regression plane of X_2 on X_1 and X_3
c. The regression plane of X_3 on X_1 and X_2

The regression plane of X_1 on X_2 and X_3

$$\boldsymbol{X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3}$$

In the above equation, the value of the value of $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneously the following three normal equations:

$$\boldsymbol{\sum X_1 = Na_{1.23} + b_{12.3}\sum X_2 + b_{13.2}\sum X_3}$$

$$\sum X_1X_2 = a_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3$$

$$\sum X_1X_3 = a_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2$$

The last two of these three equations can be obtained if we multiply first equation by X_2 and X_3 on both sides.

The regression plane of X_2 on X_1 and X_3

$$X_2 = a_{2.13} + b_{21.3}X_1 + b_{23.1}X_3$$

In the above equation, the value of the value of $b_{21.3}$ and $b_{23.1}$ are determined by solving simultaneously the following three normal equations:

$$\sum X_2 = Na_{2.13} + b_{21.3}\sum X_1 + b_{23.1}\sum X_3$$

$$\sum X_1X_2 = a_{2.13}\sum X_1 + b_{21.3}\sum X_1^2 + b_{23.1}\sum X_1X_3$$

$$\sum X_2X_3 = a_{2.13}\sum X_3 + b_{21.3}\sum X_1X_3 + b_{23.1}\sum X_3^2$$

The last two of these three equations can be obtained if we multiply first equation by X_1 and X_3 on both sides.

The regression plane of X_3 on X_1 and X_2

$$X_3 = a_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$$

In the above equation, the value of the value of $b_{31.2}$ and $b_{32.1}$ are determined by solving simultaneously the following three normal equations:

$$\sum X_3 = Na_{3.12} + b_{31.2}\sum X_1 + b_{32.1}\sum X_2$$

$$\sum X_1X_3 = a_{3.12}\sum X_1 + b_{31.2}\sum X_1^2 + b_{32.1}\sum X_1X_2$$

$$\sum X_2X_3 = a_{3.12}\sum X_2 + b_{31.2}\sum X_1X_2 + b_{32.1}\sum X_2^2$$

The last two of these three equations can be obtained if we multiply first equation by X_1 and X_2 on both sides.

13.5.2 Assumptions of Linear Multiple Regression Analysis

Followings are the main assumption of linear multiple regression for point estimation:

1. The dependent variable is a random variable whereas the independent variables need not to be random variables.
2. There must be a linear relationship between the several independent variables and the one dependent variable.
3. The variance of the conditional distributions of the dependent variable, given various combinations of values of the independent variables, are all equal.
4. For internal estimation, an additional assumption is that the conditional distributions for the dependent variable follow the normal probability distribution.

13.5.3 Methods of Multiple Regression Analysis

There are three methods of estimating multiple regression equation when deviations are taken from actual means:

First Method

Let $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ be the actual means of the three variable X_1, X_2 and X_3 respectively. The calculation involved in solving these regression equations are considerably reduced if the deviations of the various variables are taken from their respective means. If $x_1 = (X_1 - \bar{X}_1)$, $x_2 = (X_2 - \bar{X}_2)$ and $x_3 = (X_3 - \bar{X}_3)$. The multiple regression equation takes the following shape:

$$\boldsymbol{x_1 = b_{12.3}x_2 + b_{13.2}x_3}$$

We can find the values of $b_{12.3}$ and $b_{13.2}$ by simultaneously solving the following two normal equations:

$$\boldsymbol{\sum x_1x_2 = b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3}$$

$$\boldsymbol{\sum x_1x_3 = b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2}$$

The values of $b_{12.3}$ and $b_{13.2}$ can also be obtained as follows:

$$\boldsymbol{b_{12.3} = r_{12.3} \times \frac{\sigma_{1.23}}{\sigma_{2.13}}}$$

$$\boldsymbol{b_{13.2} = r_{13.2} \times \frac{\sigma_{1.32}}{\sigma_{3.12}}}$$

Similar, calculations are possible for the regression equations of x_2 on x_1 and x_3; and for regression equations of x_3 on x_1 and x_2.

Second Method

Multiple Regression Equation of X_1 on X_2 and X_3: It can be expressed as follows:

$$X_1 - \bar{X}_1 = \frac{\sigma_1}{\sigma_2}\left[\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right](X_2 - \bar{X}_2) + \frac{\sigma_1}{\sigma_3}\left[\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right](X_2 - \bar{X}_3)$$

Multiple Regression Equation of X_2 on X_3 and X_1: It can be expressed as follows:

$$X_2 - \bar{X}_2 = \frac{\sigma_2}{\sigma_3}\left[\frac{r_{23} - r_{12}r_{13}}{1 - r_{13}^2}\right](X_3 - \bar{X}_3) + \frac{\sigma_2}{\sigma_1}\left[\frac{r_{12} - r_{23}r_{13}}{1 - r_{13}^2}\right](X_1 - \bar{X}_1)$$

Multiple Regression Equation of X_3 on X_1 and X_2: It can be expressed as follows:

$$X_3 - \bar{X}_3 = \frac{\sigma_3}{\sigma_2}\left[\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}\right](X_2 - \bar{X}_2) + \frac{\sigma_3}{\sigma_1}\left[\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right](X_1 - \bar{X}_1)$$

Third Method

If σ_1, σ_2, σ_3 are not given, we can calculate S_1, S_2 and S_3 from the given data. Then the regression equations are:

Regression equation of X_1 on X_2 and X_3

$$X_1 - \bar{X}_1 = \frac{S_1}{S_2}\left[\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right](X_2 - \bar{X}_2) + \frac{S_1}{S_3}\left[\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right](X_2 - \bar{X}_3)$$

Regression equation of X_2 on X_3 and X_1

$$X_2 - \bar{X}_2 = \frac{S_2}{S_3}\left[\frac{r_{23} - r_{12}r_{13}}{1 - r_{13}^2}\right](X_3 - \bar{X}_3) + \frac{S_2}{S_1}\left[\frac{r_{12} - r_{23}r_{13}}{1 - r_{13}^2}\right](X_1 - \bar{X}_1)$$

Regression equation of X_3 on X_2 and X_1

$$X_3 - \bar{X}_3 = \frac{S_3}{S_2}\left[\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}\right](X_2 - \bar{X}_2) + \frac{S_3}{S_1}\left[\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right](X_1 - \bar{X}_1)$$

13.5.4 Standard Error of Estimates

The reliability of the estimates obtained through multiple regression equations is studied through the calculations of standard error of the estimate. It is an estimate of unexplained variations in the values of

dependent variable X_1. If the total variations of X_1 is divided into two parts, the standard error would represent the unexplained variations. The explained variations would be due to independent variables. The standard error of the estimate of multiple regression equation X_1 on X_2 and X_3 is calculated by the following formula:

$$S_{1.23} = \sqrt{\frac{\sum(X_1 - Y_1)^2}{N-3}}$$

Where $S_{1.23}$ is the standard error of the estimate of X_1 on X_2 and X_3. X_1 is the original value of X and Y_1 is the estimated value on the basis of the regression equations.

The standard error of the estimate in terms of the correlation coefficients r_{12}, r_{13} and r_{23} is:

$$S_{1.23} = \sigma_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2\, r_{12} r_{13} r_{23}}{1 - r_{23}^2}}$$

If it is proposed not to calculate the estimated values of X for all data points and also not to calculate the partial correlation coefficients, the standard error of the regression estimate is calculate by the following formula:

$$S_{1.23} = \sigma_1 \sqrt{\frac{\sum X_1^2 - a \sum X_1 - b_1 \sum X_2 X_1 - b_2 \sum X_2 X_1}{N - 3}}$$

13.6 SOLVED EXAMPLES OF MULTIPLE REGRESSION ANALYSIS

Example 13.9 Find the multiple linear regression equation of X_1 on X_2 and X_3 by using normal equations from the data given below:

X_1	2	4	6	8
X_1	3	5	7	9
X_1	4	6	8	10

Solution: **The regression equation of X_1 on X_2 and X_3 is:**

$$\boldsymbol{X_1 = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3} \ldots\ldots \text{(A)}$$

Where the value of three constants $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are obtained by solving the following three normal equations:

$$\sum \boldsymbol{X_1} = \boldsymbol{N a_{1.23}} + \boldsymbol{b_{12.3}} \sum \boldsymbol{X_2} + \boldsymbol{b_{13.2}} \sum \boldsymbol{X_3}$$

$$\sum X_1X_2 = a_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3$$

$$\sum X_1X_3 = a_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2$$

Table: Calculation of Required Values

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_1^2	X_2^2	X_3^2
2	3	4	6	8	12	4	9	16
4	5	6	20	24	30	16	25	36
6	7	8	42	48	56	36	49	64
8	9	10	72	80	90	64	81	100
$\sum X_1$ =20	$\sum X_2$ =24	$\sum X_3$ =28	$\sum X_1X_2$ =140	$\sum X_1X_3$ =160	$\sum X_2X_3$ =188	$\sum X_1^2$ =120	$\sum X_2^2$ =164	$\sum X_3^2$ =216

Substituting the values in the normal equations, we get

$6a_{1.23} + 24b_{12.3} + 28b_{13.2} = 20$ …….. (i)

$24a_{1.23} + 164b_{12.3} + 188b_{13.2} = 140$ …… (ii)

$28a_{1.23} + 188b_{12.3} + 216b_{13.2} = 160$ …… (iii)

Multiplying equation (i) by 4 and Subtracting it from the equation (ii), we get:

$68b_{12.3} + 76b_{13.2} = 60$ ……. (iv)

Multiplying equation (ii) by 7 and equation (iii) by 6, we get:

$168a_{1.23} + 114b_{12.3} + 1316b_{13.2} = 980$ ……(v)

$168a_{1.23} + 1128b_{12.3} + 1296b_{13.2} = 960$ ……(vi)

Subtracting (vi) from (v), we obtain:

$20b_{12.3} + 20b_{13.2} = 20$ ……. (vii)

Multiply equation (iv) by 5 and equation (vii) by 7 to get:

$340b_{12.3} + 380b_{13.2} = 300$……(viii)

$340b_{12.3} + 340b_{13.2} = 340$……(ix)

Subtracting (ix) from (viii), we get

$40b_{13.2} = -40 = b_{13.2} = -1$…….(x)

Substitute the value of $b_{13.2}$ in equation (vii), we have:

$$20b_{12.3} + 20(-1) = 20$$

$$20b_{12.3} - 20 = 20$$

$$20b_{12.3} = 40$$

$$b_{12.3} = 40/20 = 2 \quad \ldots\ldots\ldots\ldots\ldots\ldots \text{(xi)}$$

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in equation (i), we get:

$$6a_{1.23} + 24(2) + 28(-1) = 20$$

$$6a_{1.23} + 48 - 28 = 20$$

$$6a_{1.23} + 20 = 20$$

$$6a_{1.23} = 20 - 20$$

$$a_{1.23} = 0/6 = 0$$

Putting the value of $a_{1.23} = 0$, $b_{12.3} = 2$ and $b_{13.2} = -1$, in equation (A) we get the required equation of X_1 on X_2 and X_3 as:

$$X_1 = 0 + 2X_2 - X_3$$

$$X_1 = 2X_2 - X_3$$

It should be noted that in the above problem X_1, X_2, X_3 are linear and as such the estimated values of X_1 for given value of X_2 and X_3 would remain unchanged.

Example 13.10

In a trivariate distribution: $\sigma_1 = 3$; $\sigma_2 = 4$; $\sigma_3 = 5$; $r_{23} = 0.4$; $r_{31} = 0.6$; $r_{12} = 0.7$.

Determine the regression of X_1 on X_2 and X_3, when variates are measured from then means.

Solution: The regression equation of X_1 on X_2 and X_3, when variates are measured from then mean is:

$$\boldsymbol{x_1 = b_{12.3}x_2 + b_{13.2}x_3} \qquad \text{(i)}$$

$$b_{12.3} = \frac{\sigma_1}{\sigma_2}\left[\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right] = \frac{3}{4} \times \left[\frac{0.7 - (0.6 \times 0.4}{1 - (0.4^2)}\right] = \frac{0.75 \times 0.46}{0.84} = \frac{0.345}{0.84} = 0.41$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3}\left[\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right] = \frac{3}{5} \times \left[\frac{0.6 - (0.7 \times 0.4}{1 - (0.4^2)}\right] = \frac{0.6 \times 0.32}{0.84} = \frac{0.192}{0.84} = 0.229$$

Substituting $b_{12.3} = 0.41$ and $b_{13.2} = 0.229$ in the equation (i), we get the required regression equation of X_1 on X_2 and X_3 as:

$$\boldsymbol{x_1 = 0.41\, x_2 + 0.229\, x_3}$$

Example 13.11

The given data are: $\bar{X}_1 = 6$, $\bar{X}_2 = 7$, $\bar{X}_3 = 8$; $\sigma_1 = 1$, $\sigma_2 = 2$, $\sigma_3 = 3$; $r_{12} = 0.6$, $r_{13} = 0.7$, $r_{23} = 0.8$. Find the regression equation of X_3 on X_1 and X_2. Estimate the value when $X_1 = 4$, and $X_2 = 5$.

Solution: We shall solve this problem by the third method discussed earlier. According to this method, the regression equation of X_3 on X_1 and X_2 is:

$$\boldsymbol{X_3 - \bar{X}_3 = \frac{\sigma_3}{\sigma_2}\left[\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}\right](X_2 - \bar{X}_2) + \frac{\sigma_3}{\sigma_1}\left[\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right](X_1 - \bar{X}_1)}$$

Substituting the given values in (i) we have:

$$\Rightarrow \quad \boldsymbol{X_3 - 8 = \frac{3}{2}\left[\frac{0.8 - (0.6 \times 0.7)}{1 - (0.6)^2}\right](X_2 - 7) + \frac{3}{1}\left[\frac{0.7 - (0.8 \times 0.6)}{1 - (0.6)^2}\right](X_1 - 6)}$$

$$\Rightarrow \quad \boldsymbol{X_3 - 8 = 1.5\left[\frac{0.8 - 0.42}{0.64}\right](X_2 - 7) + 3\left[\frac{0.7 - 0.48}{0.64}\right](X_1 - 6)}$$

$$\Rightarrow \quad \boldsymbol{X_3 - 8 = 1.5\left[\frac{0.38}{0.64}\right](X_2 - 7) + 3\left[\frac{0.22}{0.64}\right](X_1 - 6)}$$

$$\Rightarrow \quad \boldsymbol{X_3 - 8 = 1.5 \times 0.59375\,(X_2 - 7) + 3 \times 0.34375\,(X_1 - 6)}$$

$$\Rightarrow \quad \boldsymbol{X_3 - 8 = 0.89\,(X_2 - 7) + 1.03\,(X_1 - 6)}$$

$$\Rightarrow \quad \boldsymbol{X_3 - 8 = 0.89\,X_2 - 6.23 + 1.03\,X_1 - 6.18}$$

$$\Rightarrow \quad \boldsymbol{X_3 = 0.89\,X_2 + 1.03\,X_1 - 6.23 - 6.18 + 8}$$

$$\Rightarrow \quad \boldsymbol{X_3 = 0.89\,X_2 + 1.03\,X_1 - 4.41}$$

It is the regression equation of X_3 on X_1 and X_2 for estimating the value of X_3.

Estimation of X_3 by putting $X_1 = 4$ and $X_2 = 5$ in the equation we get:

$$\Rightarrow \quad \boldsymbol{X_3 = 0.89 \times 5 + 1.03 \times 4 - 4.41}$$

$$\Rightarrow \quad \boldsymbol{X_3 = 4.45 + 4.08 - 4.41 = 4.12}.$$

Hence $X_3 = 4.12$, When $X_1 = 4$ and $X_2 = 5$.

Example 13.12 The following table shows the corresponding shows the corresponding values of three variables X_1 X_2 and X_3. Find the least square regression equation of X_3 on X_1 and X_2. Estimate X_3 when $X_1 = 10$ and $X_2 = 6$.

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

Solution: The regression equation of X_3 on X_1 and X_2 can be written as follows:

$$\boldsymbol{X_3 - \bar{X}_3 = \frac{S_3}{S_2}\left[\frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}\right](X_2 - \bar{X}_2) + \frac{S_3}{S_1}\left[\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right](X_1 - \bar{X}_1)} \ldots(i)$$

Table: Computation of $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, S_1, S_2, S_3, r_{12}, r_{13}, r_{23}

$(X_1 - \bar{X}_1) = x_1$			$(X_2 - \bar{X}_2) = x_1$			$(X_3 - \bar{X}_3) = x_3$					
X_1	x_1	x_1^2	X_2	x_2	x_2^2	X_2	x_3	x_3^2	$x_1 x_2$	$x_1 x_3$	$x_2 x_3$
3	-5	25	16	+9	81	90	+40	1600	-45	-200	+360
5	-3	9	10	+3	9	72	+22	484	-9	-66	+66
6	-2	4	7	0	0	54	+4	16	0	-8	0
8	0	0	4	-3	9	42	-8	64	0	0	+24
12	+4	16	3	-4	16	30	-20	400	-16	-80	+80
14	+6	36	2	-5	25	12	-38	1444	-30	-228	+190
$\sum X_1 = 48$	$\sum x_1 = 0$	$\sum x_1^2 = 90$	$\sum X_2 = 42$	$\sum x_2 = 0$	$\sum x_2^2 = 140$	$\sum X_3 = 300$	$\sum x_3 = 0$	$\sum x_3 = 4008$	$\sum x_1x_2 = -100$	$\sum x_1x_3 = -582$	$\sum x_2x_3 = 720$

$$\bar{X}_1 = \frac{48}{6} = 8; \qquad \bar{X}_2 = \frac{42}{6} = 7; \qquad \bar{X}_3 = \frac{300}{6} = 50$$

$$S_1 = \sqrt{\frac{\sum(X_1 - \bar{X}_1)^2}{N}} = \sqrt{\frac{90}{6}} = \sqrt{15} = 3.87$$

$$S_2 = \sqrt{\frac{\sum(X_2 - \bar{X}_2)^2}{N}} = \sqrt{\frac{140}{6}} = \sqrt{23.33} = 4.83$$

$$S_3 = \sqrt{\frac{\sum(X_3 - \bar{X}_3)^2}{N}} = \sqrt{\frac{4008}{6}} = \sqrt{668} = 25.85$$

$$r_{12} = \frac{\sum x_1x_2}{\sqrt{\sum x_1^2 \times x_2^2}} = \frac{-100}{\sqrt{90 \times 140}} = \frac{-100}{112.25} = -0.891$$

$$r_{13} = \frac{\sum x_1x_3}{\sqrt{\sum x_1^2 \times x_3^2}} = \frac{-582}{\sqrt{90 \times 4008}} = \frac{-582}{600.599} = -0.969$$

$$r_{23} = \frac{\sum x_2x_3}{\sqrt{\sum x_2^2 \times x_3^2}} = \frac{720}{\sqrt{140 \times 4008}} = \frac{720}{749.08} = 0.961$$

Substituting these values in the equation (i), we get:

$$\boldsymbol{X_3 - 50 = \frac{25.85}{4.83}\left[\frac{0.961 - (-0.969 \times -0.891)}{1 - (-0.9)^2}\right](X_2 - 7) + \frac{25.85}{3.87}\left[\frac{-0.969 - (0.961 \times -0.891)}{1 - (-0.9)^2}\right](X_1 - 8)}$$

$$\Rightarrow \quad X_3 - 50 = 2.546\,(X_2 - 7) - 3.664\,(X_1 - 8)$$
$$\Rightarrow \quad X_3 - 50 = 2.546\,X_2 - 17.822 - 3.664\,X_1 + 29.312$$
$$\Rightarrow \quad X_3 = 2.546\,X_2 - 3.664\,X_1 + 61.49$$

It is required regression equation of X_3 on X_1 and X_2.

Estimation of value X_3 putting value of $X_1 = 10$ and $X_2 = 6$ in the equation, we get:

$$\Rightarrow \quad X_3 = 2.546 \times 6 - 3.664 \times 10 + 61.49$$
$$\Rightarrow \quad X_3 = 15.276 - 36.64 + 61.49 = 40.126$$

Hence, When $X_1 = 10$, $X_2 = 6$, then $X_3 = 40.126$.

13.7 CHECK YOUR PROGRESS

1. The sum of the squares of the vertical deviations of the observed Y values from the straight line is …… .

2. The value of both the regression coefficients cannot be……than 1.

3. The mean of both the regression coefficients is either equal to or greater than the …….

4. Regression coefficients are independent of change of …… but not of change of scale.

5. r^2 provides the necessary link between regression and correlation which are the two related aspects of a single problem of the analysis of …… between two variables.

13.8 SUMMARY

Regression analysis means the estimation or prediction of the unknown value of one variable from the known value(s) of the other variable(s). It is one of the most important and widely used statistical techniques in almost all sciences - natural, social or physical. Regression analysis for studying more than two variables at a time is known as multiple regressions. If *X* and *Y* are two variables of which relationship is to be indicated, a line that gives best estimate of *Y* for any value of *X*, it is called Regression line of Y on X. If the dependent variable changes to *X*, then best estimate of *X* by any value of *Y* is called Regression line of X on Y. A line of best fit can be fitted by means of: Free hand drawing method, and least square method. The two regression lines indicate the nature and extent of correlation between the variables. Correlation and Regression are the two related aspects of a single problem of the analysis of the relationship between the variables. If we have information on more than one variable, we might be interested in seeing if there is any connection - any association - between them. If we found

such a association, we might again be interested in predicting the value of one variable for the given and known values of other variable(s).

13.9 KEYWORDS

Dependent variable: The variable to be predicted is called the dependent variable. **Independent variable:** The predictor is called the independent variable, or explanatory variable.

The line of Regression: It is the graphical or relationship representation of the best estimate of one variable for any given value of the other variable.

Error of estimate: The preciseness of an estimate can be obtained only through a measure of the magnitude of error in the estimates, called the error of estimate.

Coefficient of determination: It gives the percentage variation in the dependent variable that is accounted for by the independent variable.

13.10 SELF-ASSESSMENT TEST

1. Explain clearly the concept of Regression. Explain with suitable examples its role in dealing with business problems.
2. What do you understand by linear regression?
3. What is meant by 'regression'? Why should there be in general, two lines of regression for each bivariate distribution? How the two regression lines are useful in studying correlation between two variables?
4. Why is the regression line known as line of best fit?
5. Write short note on
   (i) Regression Coefficients
   (ii) Regression Equations
   (iii) Standard Error of Estimate
   (iv) Coefficient of Determination
   (v) Coefficient of Non-determination
6. Distinguish clearly between correlation and regression as concept used in statistical analysis.
7. Fit a least-square line to the following data:
   (i) Using *X* as independent variable
   (ii) Using *X* as dependent variable

X	:	1	3	4	8	9	11	14
Y	:	1	2	4	5	7	8	9

Hence obtain

c) The regression coefficients of Y on X and of X on Y

d) $\overline{X}$ and $\overline{Y}$

e) Coefficient of correlation between and X and Y

f) What is the estimated value of Y when X = 10 and of X when Y = 5?

8. What are regression coefficients? Show that $r^2 = b_{yx.}\ b_{xy}$ where the symbols have their usual meanings. What can you say about the angle between the regression lines when *(i)* $r = 0$, *(ii)* $r = 1$ *(iii)* r increases from 0 to 1?

9. Obtain the equations of the lines of regression of Y on X from the following data.

X	:	12	18	24	30	36	42	48
Y	:	5.27	5.68	6.25	7.21	8.02	8.71	8.42

Estimate the most probable value of Y, when $X = 40$.

10. The following table gives the ages and blood pressure of 9 women.

Age (X) :	56	42	36	47	49	42	60	72	63
Blood Pressure(Y)	147	125	118	128	145	140	155	160	149

Find the correlation coefficient between X and Y.

(i) Determine the least square regression equation of Y on X.

(ii) Estimate the blood pressure of a woman whose age is 45 years.

11. Given the following results for the height (X) and weight (Y) in appropriate units of 1,000 students:

$\overline{X} = 68$, $\overline{Y} = 150$, $S_x = 2.5$, $S_y = 20$ and $r = 0.6$.

Obtain the equations of the two lines of regression. Estimate the height of a student A who weighs 200 units and also estimate the weight of the student B whose height is 60 units.

12. From the following data, find out the probable yield when the rainfall is 29".

	Rainfall	Yield
Mean	25"	40 units per hectare
Standard Deviation	3"	6 units per hectare

Correlation coefficient between rainfall and production = 0.8.

13. A study of wheat prices at two cities yielded the following data:

	City A	City B
Average Price	Rs 2,463	Rs 2,797
Standard Deviation	Rs 0.326	Rs 0.207

Coefficient of correlation r is 0.774. Estimate from the above data the most likely price of wheat

(i) at City A corresponding to the price of Rs 2,334 at City B

(ii) at city B corresponding to the price of Rs 3.052 at City A

14. Find out the regression equation showing the regression of capacity utilization on production from the following data:

	Average	Standard Deviation
Production (in lakh units)	35.6	10.5
Capacity Utilization (in percentage)	84.8	8.5

$r = 0.62$. Estimate the production, when capacity utilisation is 70%.

15. The following table shows the mean and standard deviation of the prices of two shares in a stock exchange.

Share	Mean (in Rs)	Standard Deviation (in Rs)
A Ltd.	39.5	10.8
B Ltd.	47.5	16.0

If the coefficient of correlation between the prices of two shares is 0.42, find the most likely price of share A corresponding to a price of Rs 55, observed in the case of share B.

16. Find out the regression coefficients of Y on X and of X on Y on the basis of following data:

$\sum X = 50, \quad \overline{X} = 5, \quad \sum Y = 60, \quad \overline{Y} = 6, \quad \sum XY = 350$

Variance of $X = 4$, Variance of $Y = 9$

17. Find the regression equation of X and Y and the coefficient of correlation from the following data:

$\sum X = 60, \sum Y = 40, \sum XY = 1150, \sum X^2 = 4160, \sum Y^2 = 1720$ and $N = 10$.

18. By using the following data, find out the two lines of regression and from them compute the Karl Pearson's coefficient of correlation.

$\sum X = 250, \sum Y = 300, \sum XY = 7900, \sum X^2 = 6500, \sum Y^2 = 10000, N = 10$

19. The equations of two regression lines between two variables are expressed as

$$2X - 3Y = 0 \text{ and } 4Y - 5X\text{-}8 = 0.$$

(i) Identify which of the two can be called regression line of Y on X and of X on Y.

*(ii)*Find $\overline{X}$ and $\overline{Y}$ and correlation coefficient r from the equations

20. If the two lines of regression are $4X - 5Y + 30 = 0$ and $20X - 9Y - 107 = 0$

Which of these is the lines of regression of X and Y. Find r_{xy} and S_y when $S_x = 3$

21. The regression equation of profits (X) on sales (Y) of a certain firm is $3Y - 5X + 108 = 0$. The average sales of the firm were Rs 44,000 and the variance of profits is $9/16^{th}$ of the variance of sales. Find the average profits and the coefficient of correlation between the sales and profits.
22. Explain the concept of multiple regression, and try to find out an example in practical field where the multiple regression analysis is likely to be helpful.
23. Explain multiple regression. How does it different from partial regression? Explain with the help of example.
24. Given $r_{12} = 0.8$, $r_{13} = 0.7$, $r_{23} = 0.6$, $\sigma_1 = 10$, $\sigma_2 = 8$, $\sigma_3 = 5$. Determine the regression equation of x_2 on x_1 and x_3. (Hint: Here $b_{21.3} = 0.596$, $b_{23.1} = 0.125$)

13.11 ANSWERS TO CHECK YOUR PROGRESS

1. Minimum
2. Greater
3. Coefficient of correlation
4. Origin
5. Relationship

13.12 REFERNCES/SUGGESTED READINGS

33. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
34. Statistics for Management *by* G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
35. Business Statistics *by* Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.

36. Statistics for Business and Economics *by* R.P. Hooda. MacMillan India Ltd., New Delhi.

37. Business Statistics *by* S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.

38. Statistical Method *by* S.P. Gupta. Sultan Chand and Sons., New Delhi.

39. Statistics for Management *by* Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.

40. Statistics for Business and Economics *by* Kohlar Heinz. Harper Collins., New York.

Subject: Business Statistics	
Course Code: MBA-102	Author: Dr. B. S. Bodla
Lesson No. 14	Vetter: Karam Pal
ANALYSIS OF TIME SERIES	

STRUCTURE

14.1. Learning Objectives

14.2 Introduction

14.2.1 Objectives of time series analysis

14.2.2 Components of time series

14.2.3 Time series decomposition models

14.2.4 Measurement of secular trend

14.3 Seasonal variations

14.4 Measurement of cyclical and irregular variations

14.5 Check your progress

14.6 Summary

14.7 Keywords

14.8 Self-Assessment Test

14.9 Answers to check your progress

14.10 References/Suggested readings

14.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understand the meaning and importance of time series

- Explain the models and components of time series
- Explain the details of methods of measuring trends.

14.2 INTRODUCTION

A series of observations, on a variable, recorded after successive intervals of time is called a time series. The successive intervals are usually equal time intervals, e.g., it can be 10 years, a year, a quarter, a month, a week, a day, and an hour, etc. The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.

14.2.1 OBJECTIVES OF TIME SERIES ANALYSIS

The analysis of time series implies its decomposition into various factors that affect the value of its variable in a given period. It is a quantitative and objective evaluation of the effects of various factors on the activity under consideration.

There are two main objectives of the analysis of any time series data:

(i) To study the past behaviour of data.

(ii) To make forecasts for future.

The study of past behaviour is essential because it provides us the knowledge of the effects of various forces. This can facilitate the process of anticipation of future course of events, and, thus, forecasting the value of the variable as well as planning for future.

14.2.2 Components of a Time Series

In the typical time-series there are three main components which seem to be independent of the and seems to be influencing time-series data.

Trend- It is the broad long-term tendency of either upward or downward movement in the average (or mean) value of the forecast variable *y* over time. The rate of trend growth usually varies over time, as shown in fig 16.1(a) and (b).

Cycles- An upward and downward oscillation of uncertain duration and magnitude about the trend line due to seasonal effect with fairly regular period or long period with irregular swings is called a *cycle.* A business cycle may vary in length, usually greater than one year but less than 5 to 16 years. The movement is through four phases: from *peak* (prosperity) to *contradiction* (recession) to *trough* (depression) to *expansion* (recovery or growth) as shown in Fig. 16.1 (b) and (c).

Seasonal- It is a special case of a cycle component of time series in which the magnitude and duration of the cycle do not vary but happen at a regular interval each year. For example, average sales for a retail store may increase greatly during festival seasons.

Irregular- An irregular or erratic (or residual) movement in a time series is caused by short-term unanticipated and non-recurring factors. These follow no specific pattern.

14.2.3 TIME SERIES DECOMPOSITION MODELS

The analysis of time series consists of two major steps:

1. Identifying the various forces (influences) or factors which produce the variations in the time series, and
2. Isolating, analysing and measuring the effect of these factors separately and independently, by holding other things constant.

The purpose of decomposition models is to break a time series into its components: Trend (T), Cyclical (*C*), Seasonality (*S*), and Irregularity (*I*). Decomposition of time series provides a basis for forecasting. There are many models by which a time series can be analysed; two models commonly used for decomposition of a time series are discussed below.

14.2.3.1. Multiplicative Model

This is a most widely used model which assumes that forecast (*Y*) is the product of the four components at a particular time period. That is, the effect of four components on the time series is interdependent.

$$Y = T \times C \times S \times I \leftarrow \text{Multiplicative model}$$

The multiplicative model is appropriate in situations where the effect of *S, C,* and *I* is measured in relative sense and is not in absolute sense. The geometric mean of *S, C,* and *I* is assumed to be less than one. For example, let the actual sales for period 20 be Y_{20} = 423.36. Further let, this value be broken down into its components as: let trend component (mean sales) be 400; effect of current cycle (0.90) is to depress sales by 10 per cent; seasonality of the series (1.20) boosts sales by 20 per cent. Thus besides the random fluctuation, the expected value of sales for the period is 400 × 0.90 × 1.20 = 432. If the random factor depresses sales by 2 per cent in this period, then the actual sales value will be 432 × 0.98 = 423.36.

14.2.3.2. Additive Model

In this model, it is assumed that the effect of various components can be estimated by adding the various components of a time-series. It is stated as:

$$Y = T + C + S + I \quad \leftarrow \text{Additive model}$$

Here *S, C,* and *I* are absolute quantities and can have positive or negative values. It is assumed that these four components are independent of each other. However, in real-life time series data this assumption does not hold good.

14.2.4 MEASUREMENT OF SECULAR TREND

The principal methods of measuring trend fall into following categories:

1. Free Hand Curve methods
2. Method of Averages
3. Method of least squares

The *time series methods* are concerned with taking some observed historical pattern for some variable and projecting this pattern into the future using a mathematical formula. These methods do not attempt to suggest why the variable under study will take some future value. This limitation of the time series approach is taken care by the application of a causal method. The causal method tries to identify factors which influence the variable is some way or cause it to vary in some predictable manner.

The two causal methods, regression analysis and correlation analysis, have already been discussed previously.

A few time series methods such as *freehand curves* and *moving averages* simply describe the given data values, while other methods such as *semi-average* and *least squares* help to identify a trend equation to describe the given data values.

14.2.4.1. Freehand Method

A freehand curve drawn smoothly through the data values is often an easy and, perhaps, adequate representation of the data. The forecast can be obtained simply by extending the trend line. A trend line fitted by the freehand method should conform to the following conditions:

(i) The trend line should be smooth- a straight line or mix of long gradual curves.

(ii) The sum of the vertical deviations of the observations above the trend line should equal the sum of the vertical deviations of the observations below the trend line.

(iii) The sum of squares of the vertical deviations of the observations from the trend line should be as small as possible.

(iv) The trend line should bisect the cycles so that area above the trend line should be equal to the area below the trend line, not only for the entire series but as much as possible for each full cycle.

Example 14.1: Fit a trend line to the following data by using the freehand method.

Year	1991	1992	1993	1994	1995	1996	19916	1998
Sales turnover: (Rs. in lakh)	80	90	92	83	94	99	92	104

Solution: Figure 16.2 presents the freehand graph of sales turnover (Rs. in lakh) from 1991 to 1998. Forecast can be obtained simply by extending the trend line.

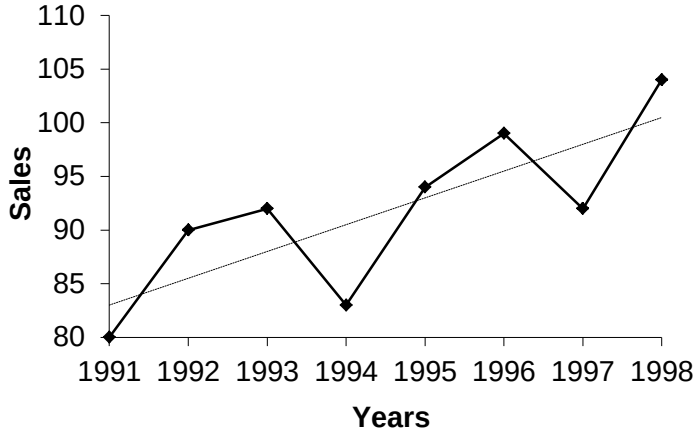


Fig. 16.2:

Graph of Sales Turnover

Limitations of freehand method

(i) This method is highly subjective because the trend line depends on personal judgement and therefore what happens to be a good-fit for one individual may not be so for another.

(ii) The trend line drawn cannot have much value if it is used as a basis for predictions.

(iii) It is very time-consuming to construct a freehand trend if a careful and conscientious job is to be done.

14.2.4.2. Method of Averages

The objective of smoothing methods into smoothen out the random variations due to irregular components of the time series and thereby provide us with an overall impression of the pattern of movement in the data over time. In this section, we shall discuss three smoothing methods.

(i) Moving averages

(ii) Weighted moving averages

(iii) Semi-averages

The data requirements for the techniques to be discussed in this section are minimal and these techniques are easy to use and understand.

Moving Averages

If we are observing the movement of some variable values over a period of time and trying to project this movement into the future, then it is essential to smooth out first the irregular pattern in the historical values of the variable, and later use this as the basis for a future projection. This can be done by calculating a series of moving averages.

This method is a subjective method and depends on the length of the period chosen for calculating moving averages. To remove the effect of cyclical variations, the period chosen should be an integer value that corresponds to or is a multiple of the estimated average length of a cycle in the series.

The moving averages which serve as an estimate of the next period's value of a variable given a period of length *n* is expressed as:

$$\text{Moving average, } Ma_{t+1} = \frac{\sum\{D_t + D_{t-1} + D_{t-2}. + .. + D_{t-n+1}\}}{n}$$

where t = current time period

D = actual data which is exchanged each period

n = length of time period

In this method, the term 'moving' is used because it is obtained by summing and averaging the values from a given number of periods, each time deleting the oldest value and adding a new value.

The limitation of this method is that it is highly subjective and dependent on the length of period chosen for constructing the averages. Moving averages have the following three limitations:

(i) As the size of n (the number of periods averaged) increases, it smoothens the variations better, but it also makes the method less sensitive to real changes in the data.

(ii) Moving averages cannot pick-up trends very well. Since these are averages, it will always stay within past levels and will not predict a change to either a higher or lower level.

(iii) Moving average requires extensive records of past data.

Example 14.2: Using three-yearly moving averages, determine the trend and short-term-error.

Year	Production (in '000 tonnes)	Year	Production (in '000 tonnes)
19816	21	1992	22
1988	22	1993	25
1989	23	1994	26
1990	25	1995	216
1991	24	1996	26

Solution: The moving average calculation for the first 3 years is:

$$\text{Moving average (year 1-3)} = \frac{21 + 22 + 23}{3} = 22$$

Similarly, the moving average calculation for the next 3 years is:

$$\text{Moving average (year 2-4)} = \frac{22 + 23 + 25}{3} = 22.33$$

A complete summary of 3-year moving average calculations is given in Table 14.1

Table 14.1: Calculation of Trend and Short-term Fluctuations

Year	Production Y	3-Year Moving Total	3-yearly Moving Average (Trend values $\hat{y}$	Forecast Error (y- $\hat{y}$)
19816	21	-	-	-
1988	22 →	66	22.00	0
1989	23 →	160	23.33	-0.33
1990	25 →	162	24.00	1.00
1991	24	161	23.616	0.33
1992	22	161	23.616	-1.616
1993	25	163	24.33	0.616
1994	26	168	26.00	0
1995	216	169	26.33	0.616
1996	26	-	-	-

Odd and Even Number of Years

When the chosen period of length n is an odd number, the moving average at year i is centred on i, the middle year in the consecutive sequence of n yearly values used to compute i. For instance with n =5, $MA_3(5)$ is centred on the third year, $MA_4(5)$ is centred on the fourth year…, and $MA_9(5)$ is centred on the ninth year.

No moving average can be obtained for the first $(n-1)/2$ years or the last $(n-1/2)$ year of the series. Thus for a 5-year moving average, we cannot make computations for the just two years or the last two years of the series.

When the chosen period of length n is an even numbers, equal parts can easily be formed and an average of each part is obtained. For example, if n = 4, then the first moving average M_3 (placed at period 3) is an average of the first four data values, and the second moving average M_4 (placed at period 4) is the average of data values 2 through 5). The average of M_3 and M_4 is placed at period 3 because it is an average of data values for period 1 through 5.

Example 14.3: Assume a four-yearly cycle and calculate the trend by the method of moving average from the following data relating to the production of tea in India.

Year	*Production (million lbs)*	*Year*	*Production (million lbs)*
19816	464	1992	540
1988	515	1993	5516
1989	518	1994	5161
1990	4616	1995	586
1991	502	1996	612

Solution: The first 4-year moving average is:

$$MA_3(4) = \frac{464 + 515 + 518 + 4616}{4} = \frac{1964}{4} = 491.00$$

This moving average is centred on the middle value, that is, the third year of the series. Similarly,

$$MA_4(4) = \frac{515 + 518 + 4616 + 502}{4} = \frac{2002}{4} = 500.50$$

This moving average is centred on the fourth year of the series.

Table 14.2. presents the data along with the computations of 4-year moving averages.

Table 14.2: Calculation of Trend and Short-term Fluctuations

Year	Production (mm lbs)	4-yearly Moving Totals	4-Yearly Moving Average	4-Yearly Moving Average Centred
19816	464	-	-	-
1988	515	-	-	-
		1964	491.00	
1989	518			495.165
		2002	500.50	
1990	4616			503.62

		20216	506.165	
1991	502			511.62
		2066	516.50	
1992	540			529.50
		21160	542.50	
1993	5516			553.00
		2254	563.50	
1994	5161			5162.00
		2326	581.50	-
1995	586	-	-	-
1996	612	-	-	-

Weighted Moving Averages

In moving averages, each observation is given equal importance (weight). However, different values may be assigned to calculate a weighted average of the most recent *n* values. Choice of weights is somewhat arbitrary because there is no set formula to determine them. In most cases, the most recent observation receives the most weightage, and the weight decreases for older data values.

A weighted moving average may be expressed mathematically as

$$\text{Weighted moving average} = \frac{\Sigma(\text{Weight for period } n)(\text{Data value in period } n)}{\Sigma\text{Weights}}$$

Example 14.4: Vaccum cleaner sales for 12 months is given below. The owner of the supermarket decides to forecast sales by weighting the past three months as follows:

Weight Applied	**Month**
3	Last month
2	Two months ago
1	Three months ago
6	

Month	:	1	2	3	4	5	6	16	8	9	10	11	12
Actual sales (in units)	:	10	12	13	16	19	23	26	30	28	18	16	14

Solution:

The results of 3-month weighted average are shown in Table 14.3.

$$\text{Forecast for the Current month} = \frac{3 \times \text{Sales last month} + 2 \times \text{Sales two months ago} + 1 \times \text{Sales three months ago}}{6}$$

Table 14.3: Weighted Moving Average

Month	Actual Sales	Three-month Weighted Moving Average
1	10	-
2	12	-
3	13	-
4	16	$\frac{1}{6}[3\times13)+(2\times12)+1\times10]=\frac{121}{6}$
5	19	$\frac{1}{6}[3\times16)+(2\times13)+1\times12]=\frac{141}{3}$
6	23	$\frac{1}{6}[3\times19)+(2\times16)+1\times13]=17$
16	26	$\frac{1}{6}[3\times23)+(2\times19)+1\times16]=\frac{201}{2}$
8	30	$\frac{1}{6}[3\times26)+(2\times23)+1\times19]=\frac{235}{6}$
9	28	$\frac{1}{6}[3\times30)+(2\times26)+1\times23]=\frac{271}{2}$
10	18	$\frac{1}{6}[3\times28)+(2\times30)+1\times26]=\frac{289}{3}$
11	16	$\frac{1}{6}[3\times18)+(2\times28)+1\times30]=\frac{231}{3}$
12	14	$\frac{1}{6}[3\times16)+(2\times18)+1\times28]=\frac{182}{3}$

Example 14.5: A food processor uses a moving average to forecast next month's demand. Past actual demand (in units) is shown below:

Month	:	43	44	45	46	416	48	49	50	51
Actual demand (in units)	:	105	106	110	110	114	121	130	128	1316

(a) Compute a simple five-month moving average to forecast demand for month 52.

(b) Compute a weighted three-month moving average where the weights are highest for the latest months and descend in order of 3, 2, 1.

Solution: Calculation for five-month moving average are shown in Table 14.4.

Month	Actual Demand	5-month Moving Total	5-month Moving Average
43	105	-	-
44	106	-	-
45	110	545	109.50
46	110	561	112.2
416	114	585	1116.0
48	121	603	120.6
49	130	630	126.0
50	128	-	-
51	1316	-	-

(a) Five-month average demand for month 52 is

$$\frac{\Sigma x}{\text{Number of periods}} = \frac{114 + 121 + 130 + 128 + 1316}{5} = 126 \text{ units}$$

(b) Weighted three-month average as per weights is as follows:

$$MA_{Wt} = \frac{\Sigma \text{ Weight} \times \text{Data value}}{\Sigma \text{ weight}}$$

Where Month Weight × Value = Total

51	3 × 1316	=	141
50	2 × 128	=	256
49	1 × 130	=	130
	6		16916

$MA_{WT} = \frac{797}{6}$ = 133 units

Semi-Average Method

The semi-average method permits us to estimate the slope and intercept of the trend the quite easily if a linear function will adequately described the data. The procedure is simply to divide the data into two parts and compute their respective arithmetic means. These two points are plotted corresponding to their midpoint of the class interval covered by the respective part and then these points are joined by a straight line, which is the required trend line. The arithmetic mean of the first part is the intercept value, and the slope is determined by the ratio of the difference in the arithmetic mean of the number of years between them, that is, the change per unit time. The resultant is a time series of the form : $\hat{y} = a + bx$. The $\hat{y}$ is the calculated trend value and a and b are the intercept and slope values respectively. The equation should always be stated completely with reference to the year where x =0 and a description of the units of x and y.

The semi-average method of developing a trend equation is relatively easy to commute and may be satisfactory if the trend is linear. If the data deviate much from linearity, the forecast will be biased and less reliable.

Example 14.6: Fit a trend line to the following data by the method of semi-average and forecast the sales for the year 2002.

Year	*Sales of Firm (thousand units)*	*Year*	*Sales of Firm (thousand units)*
1993	102	19916	108
1994	105	1998	116

1995	114	1999	112
1996	110		

Solution: Since number of years are odd in number, therefore divide the data into equal parts (A and B) of 3 years ignoring the middle year (1996). The average of part A and B is

$$\bar{y}_A = \frac{102 + 105 + 114}{3} = \frac{321}{3} = 1016 \text{ units}$$

$$\bar{y}_B = \frac{108 + 116 + 112}{3} = \frac{336}{2} = 112 \text{ units}$$

Part A is centred upon 1994 and part B on 1998. Plot points 1016 and 112 against their middle years, 1994 and 1998. By joining these points, we obtain the required trend line as shown Fig. 14.3. The line can be extended and be used for prediction.

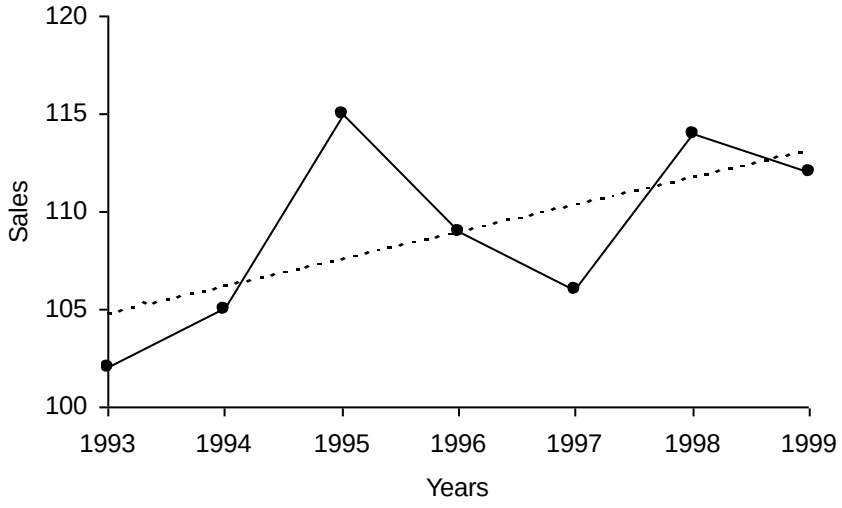


Fig. 14.3: Trend Line by the Method of Semi-Average

To calculate the time-series $\hat{y} = a + bx$, we need

$$\text{Slope } b = \frac{\Delta y}{\Delta x} = \frac{\text{Change in sales}}{\text{Change in year}}$$

$$= \frac{112 - 1016}{1998 - 1994} = \frac{5}{4} = 1.25$$

Intercept = a = 1016 units at 1994

Thus, the trend line is: $\hat{y} = 1016 + 1.25x$

Since 2002 is 8 year distant from the origin (1994), therefore we have

$$\hat{y} = 1016 + 1.25(8) = 1116$$

Exponential Smoothing Methods

Exponential smoothing is a type of moving-average forecasting technique which weighs past data in an exponential manner so that the most recent data carries more weight in the moving average. Simple exponential smoothing makes no explicit adjustment for trend effects whereas adjusted exponential smoothing does take trend effect into account (see next section for details).

Simple Exponential Smoothing

With simple exponential smoothing, the forecast is made up of the last period forecast plus a portion of the difference between the last period's actual demand and the last period's forecast.

$$F_t = F_{t-1} + \alpha (D_{t-1} - F_{t-1}) = (1-\alpha)F_{t-1} + \alpha D_{t-1} \quad \ldots(16.1)$$

Where F_t = current period forecast

F_{t-1} = last period forecast

α = a weight called smoothing constant ($0 \leq \alpha \leq 1$)

D_{t-1} = last period actual demand

From Eqn. (16.1), we may notice that each forecast is simply the previous forecast plus some correction for demand in the last period. If demand was above the last period forecast the correction will be positive, and if below it will negative.

When *smoothing constant* α is low, more weight is given to past data, and when it is high, more weight is given to recent data. When α is equal to 0.9, then 99.99 per cent of the forecast value is determined by the four most recent demands. When α is as low as 0.1, only 34.39 per cent of the average is due to these last 4 periods and the smoothing effect is equivalent to a 19-period arithmetic moving average.

If α were assigned a value as high as 1, each forecast would reflect total adjustment to the recent demand and the forecast would simply be last period's actual demand, that is, $F_t = 1.0D_{t-1}$. Since demand fluctuations are typically random, the value of α is generally kept in the range of 0.005 to 0.30 in order to 'smooth' the forecast. The exact value depends upon the response to demand that is best for the individual firm.

The following table helps illustrate this concept. For example, when $\alpha = 0.5$, we can see that the new forecast is based on demand in the last three or four periods. When $\alpha = 0.1$, the forecast places little weight on recent demand and takes a 19-period arithmetic moving average.

	Weight Assigned to				
Smoothing Constant	*Most Recent Period* (α)	*2nd Most Recent Period* $\alpha(1-\alpha)$	*3rd Most Recent Period* $\alpha(1-\alpha)^2$	*4th Most Recent Period* $\alpha(1-\alpha)^3$	*5th Most Recent Period* $\alpha(1-\alpha)^4$
$\alpha = 0.1$	0.1	0.09	0.081	0.0163	0.066
$\alpha = 0.5$	0.5	0.25	0.125	0.063	0.031

Selecting the smoothing constant

The exponential smoothing approach is easy to use and it has been successfully applied by banks, manufacturing companies, wholesalers, and other organizations. The appropriate value of the smoothing constant, α, however, can make the difference between an accurate and an inaccurate forecast. In picking a value for the smoothing constant, the objective is to obtain the most accurate forecast.

The correct α-value facilitates scheduling by providing a reasonable reaction to demand without incorporating too much random variation. An approximate value of α which is equivalent to an arithmetic moving average, in terms of degree of smoothing, can be estimated as: $\alpha = 2/(n+1)$. The accuracy of a forecasting model can be determined by comparing the forecasting values with the actual or observed values.

The forecast error is defined as:

Forecast error = Actual values – Forecasted values

One measure of the overall forecast error for a model is the *mean absolute deviation (MAD)*. This is computed by taking the sum of the absolute values of the individual forecast errors and dividing by the number of periods n of data.

$$MAD = \frac{\Sigma \left| \text{Forecast errors} \right|}{n}$$

where Standard deviation σ = 1.25 MAD

The exponential smoothing method also facilities continuous updating of the estimate of MAD. The current MAD_t is given by

$MAD_t = \alpha \left| \text{Actual values- Forecasted values} \right| + (1-\alpha)\, MAD_{t-1}$

Higher values of smoothing constant α make the current MAD more responsive to current forecast errors.

Example 14.16: A firm uses simple exponential smoothing with α =0.1 to forecast demand. The forecast for the week of February 1 was 500 units whereas actual demand turned out to be 450 units.

(a) Forecast the demand for the week of February 8.

(b) Assume the actual demand during the week of February 8 turned out to be 505 units. Forecast the demand for the week of February 15. Continue forecasting through March 15, assuming that subsequent demands were actually 516, 488, 4616, 554 and 510 units.

Solution: Given $F_{t-1} = 500$, $D_{t-1} = 450$, and $\alpha = 0.1$

(a) $F_t = F_{t-1} - \alpha(D_{t-1} - F_{t-1}) = 500 + 0.1(450-500) = 495$ units

(b) Forecast of demand for the week of February 15 is shown in Table 14.5

Table 14.5: Forecast of Demand

Week	Demand D_{t-1}	Old Forecast F_{t-1}	Forecast Error $(D_{t-1} - F_{t-1})$	Correction $\alpha(D_{t-1} - F_{t-1})$	New Forecast (F_t) $F_{t-1} + \alpha(D_{t-1} - F_{t-1})$
Feb. 1	450	500	-50	-5	495
Feb. 8	505	495	10	1	496
Feb. 15	516	496	20	2	498
Feb. 22	488	498	-10	-1	4916
Mar. 1	4616	4916	-30	-3	494
Mar. 8	554	494	60	6	500
Mar. 15	510	500	10	1	501

If no previous forecast value is known, the old forecast starting point may be estimated or taken to be an average of some preceding periods.

Example 14.8: A hospital has used a 9 month moving average forecasting method to predict drug and surgical inventory requirements. The actual demand for one item is shown in the table below. Using the previous moving average data, convert to an exponential smoothing forecast for month 33.

Month	:	24	25	26	216	28	29	30	31	32
Demand (in units)	:	168	65	90	161	80	101	84	60	163

Solution: The moving average of a 9-month period is given by

$$MA = \frac{\sum \text{Demand (x)}}{\text{Number of periods}} = \frac{168 + 65 \ldots + 163}{9} = 168$$

Assume $F_{t-1} = 168$. Therefore, estimated $\alpha = \frac{2}{n+1} = \frac{2}{9+1} = 0.2$

Thus, $F_t = F_{t-1} + \alpha(D_{t-1} - F_{t-1}) = 168 + 0.2\ (163 - 168) = 1616$ units

Methods of least square

The trend project method fits a trend line to a series of historical data points and then projects the line into the future for medium-to-long range forecasts. Several mathematical trend equations can be developed (such as exponential and quadratic), depending upon movement of time-series data.

Reasons to study trend: A few reasons to study trends are as follows:

1. The study of trend allows us to describe a historical pattern so that we may evaluate the success of previous policy.
2. The study allows us to use trends as an aid in making intermediate and long-range forecasting projections in the future.
3. The study of trends helps us to isolate and then eliminate its influencing effects on the time-series model as a guide to short-run (one year or less) forecasting of general business cycle conditions.

Linear Trend Model

If we decide to develop a linear trend line by a precise statistical method, we can apply the *least squares method.* A least squares line is described in terms of its *y*-intercept (the height at which it intercepts the *y*-axis) and its slope (the angle of the line). If we can compute the *y*-intercept and slope, we can express the line with the following equation

$$\hat{y} = a + bx$$

where $\hat{y}$ = predicted value of the dependent variable

a = y-axis intercept

b = slope of the regression line (or the rate of change in y for a given change in x)

x = independent variable (which is *time* in this case)

Least squares is one of the most widely used methods of fitting trends to data because it yields what is mathematically described as a 'line of best fit'. This trend line has the properties that (i) the summation of all vertical deviations about it is zero, that is, $\Sigma(y-\hat{y}) = 0$, (ii) the summation f all vertical deviations squared is a

minimum, that is, $\Sigma(y-\hat{y})$ is least, and (iii) the line goes through the mean values of variables x and y. For linear equations, it is found by the simultaneous solution for a and b of the two normal equations:

$$\Sigma y = na + b\Sigma x \text{ and } \Sigma xy = a\Sigma x + b\Sigma x^2$$

Where the data can be coded so that $\sum x = 0$, two terms in three equations drop out and we have $\Sigma y = na$ and $\Sigma xy = b\Sigma x^2$

Coding is easily done with time-series data. For coding the data, we choose the centre of the time period as $x = 0$ and have an equal number of plus and minus periods on each side of the trend line which sum to zero.

Alternately, we can also find the values of constants a and b for any regression line as:

$$b= \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} \text{ and } a = \bar{y} - b\bar{x}$$

Example 14.9: Below are given the figures of production (in thousand quintals) of a sugar factory:

Year	:	1992	1993	1994	1995	1996	19916	1998
Production	:	80	90	92	83	94	99	92

(a) Fit a straight line trend to these figures.

(b) Plot these figures on a graph and show the trend line.

(c) Estimate the production in 2001.

Solution: (a) Using normal equations and the sugar production data we can compute constants a and b as shown in Table 14.6:

Table 14.6: Calculations for Least Squares Equation

Year	*Time*	*Production*	x^2	*xy*	*Trend*

	Period (x)	*(x)*			*Values* $\bar{y}$
1992	1	80	1	80	84
1993	2	90	4	180	86
1994	3	92	9	2166	88
1995	4	83	16	332	90
1996	5	94	25	4160	92
19916	6	99	36	594	94
1998	16	92	49	644	96
Total	**28**	**630**	**140**	**25166**	

$$\bar{x} = \frac{\sum x}{n} = \frac{28}{7} = 4, \quad \bar{y} = \frac{\sum y}{n} = \frac{630}{7} = 90$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = \frac{2576 - 7(4)(90)}{140 - 7(4)^2} = \frac{56}{28} = 2$$

$$a = \bar{y} - b\bar{x} = 90 - 2(4) = 82$$

Therefore, linear trend component for the production of sugar is:

$$\hat{y} = a + bx = 82 + 2x$$

The slope $b = 2$ indicates that over the past 16 years, the production of sugar had an average growth of about 2 thousand quintals per year.

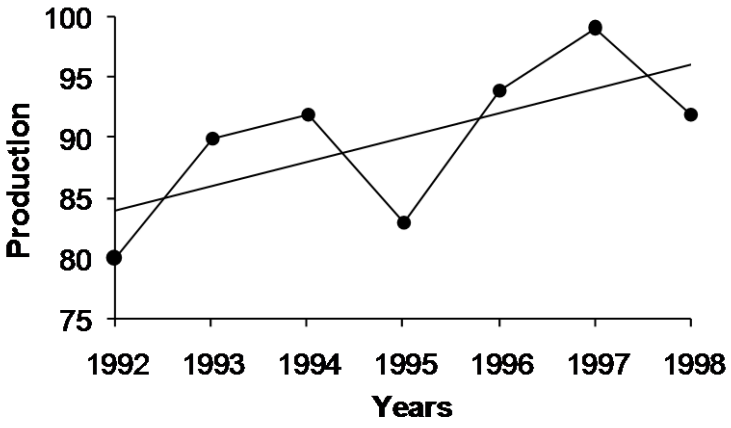


Fig.14.4: Linear Trend for Production of Sugar

(b) Plotting points on the graph paper, we get an actual graph representing production of sugar over the past 16 years. Join the point $a = 82$ and $b = 2$ (corresponds to 1993) on the graph we get a trend line as shown in Fig. 14.4.

(c) The production of sugar for year 2001 will be $\hat{y} = 82 + 2\ (10) = 102$ thousand quintals.

Parabolic Trend Model

The curvilinear relationship for estimating the value of a dependent variable y from an independent variable x might take the form

$$\hat{y} = a + bx + cx^2$$

This trend line is called the *parabola*.

For a non-linear equation $\hat{y} = a + bx - cx^2$, the values of constants a, b, and c can be determined by solving three normal equations.

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

When the data can be coded so that $\Sigma x = 0$ and $\Sigma x^3 = 0$, two term in the above expressions drop out and we have

$$\Sigma y = na + c\Sigma x^2$$

$$\Sigma xy = b\Sigma x^2$$

$$\Sigma x^2 y = a\Sigma x^2 + c\Sigma x^4$$

To find the exact estimated value of the variable y, the values of constants a, b, and c need to be calculated. The values of these constants can be calculated by using the following shortest method:

$$a = \frac{\sum y - c\sum x^2}{n}; b = \frac{\sum xy}{\sum x^2} \text{ and } c = \frac{n\sum x^2 y - \sum x^2 \sum y}{n\sum x^4 - (\sum x^2)^2}$$

Example 14.10: The prices of a commodity during 1999-2004 are given below. Fit a parabola to these data. Estimate the price of the commodity for the year 2005.

Year	Price	Year	Price
1999	100	2002	140
2000	1016	2003	181
2001	128	2004	192

Also plot the actual and trend values on a graph.

Solution: To fit a parabola $\hat{y} = a + bx + cx^2$, the calculations to determine the values of constants a, b, and c are shown in Table 14.16.

Table 14.16: Calculations for Parabola Trend Line

Year	*Time Scale (x)*	*Price (y)*	x^2	x^3	x^4	xy	x^2y	*Trend Values* ($\hat{y}$)
1999	-2	100	4	-8	16	-200	400	916.162
2000	-1	1016	1	-1	1	-1016	1016	110.34
2001	0	128	0	0	0	0	0	126.68
2002	1	140	1	1	1	140	140	146.50
2003	2	181	4	8	16	362	1624	169.88
2004	3	192	9	216	81	5166	11628	196.82
	3	**848**	**19**	**216**	**115**	**16161**	**3099**	**8416.94**

(i) $\Sigma y = na + b\Sigma x + c\Sigma x^2$

$848 = 6a + 3b + 19c$

(ii) $\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$

$16161 = 3a + 19b + 216c$

(iii) $\Sigma x^2y = a\Sigma x^2 + b\Sigma x^2 + c\Sigma x^4$

$3099 = 19a + 216b + 115c$

Eliminating a from eqns. (i) and (ii), we get

(iv) $694 = 35b + 35c$

Eliminating a from eqns. (ii) and (iii), we get

(v) $5352 = 280b + 168c$

Solving eqns. (iv) and (v) for b and c we get b =18.04 and c = 1.168. Substituting values of b and c in eqn. (i), we get a = 126.68.

Hence, the required non-linear trend line becomes

$$y = 126.68 + 18.04x + 1.168x^2$$

Several trend values as shown in Table 14.16 can be obtained by putting x = -2, -1, 0, 1, 2 and 3 in the trend line. The trend values are plotted on a graph paper. The graph is shown in Fig. 14.5.

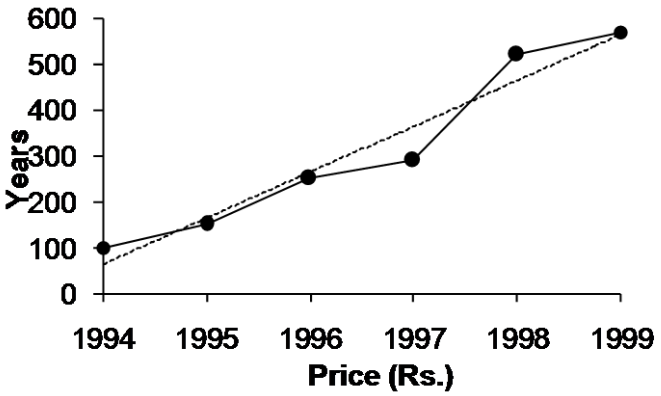


Fig. 14.5

Exponential Trend Model

When the given values of dependent variable y from approximately a geometric progression while the corresponding independent variable x values form an arithmetic progression, the relationship between variables x and y is given by an exponential function, and the best fitting curve is said to describe the *exponential trend.* Data from the fields of biology, banking, and economics frequently exhibit such a trend. For example, growth of bacteria, money accumulating at compound interest, sales or earnings over a short period, and so on, follow exponential growth.

The characteristics property of this law is that the rate of growth, that is, the rate of change of y with respect to x is proportional to the values of the function. The following function has this property.

$$y = ab^{cx},\ a > 0$$

The letter b is a fixed constant, usually either 10 or e, where a is a constant to be determined from the data.

To assume that the law of growth will continue is usually unwarranted, so only short range predictions can be made with any considerable degree or reliability.

If we take logarithms (with base 10) of both sides of the above equation, we obtain

$$\text{Log } y = \log a + (c \log b)\, x \qquad (16.2)$$

For b =10, log b =1, but for b=e, log b =0.4343 (approx.). In either case, this equation is of the form $y' = c + dx$

Where $y' = \log y$, $c = \log a$, and $d = c \log b$.

Equation (16.2) represents a straight line. A method of fitting an exponential trend line to a set of observed values of y is to fit a straight trend line to the logarithms of the y-values.

In order to find out the values of constants a and b in the exponential function, the two normal equations to be solved are

$\Sigma \log y = n \log a + \log b \Sigma x$

$\Sigma x \log y = \log a \Sigma x + \log b \Sigma x^2$

When the data is coded so that $\Sigma x = 0$, the two normal equations become

$$\Sigma \log y = n \log a \text{ or } \log a = \frac{1}{n} \Sigma \log y$$

and $\Sigma x \log y = \log b\, \Sigma x^2$ or $\log b = \dfrac{\sum x \log y}{\sum x^2}$

Coding is easily done with time-series data by simply designating the center of the time period as x =0, and have equal number of plus and minus period on each side which sum to zero.

Example 14.11: The sales (Rs. In million) of a company for the years 1995 to 1999 are:

Year :	1995	1996	19916	1998	1999
Sales :	1.6	4.5	13.8	40.2	125.0

Find the exponential trend for the given data and estimate the sales for 2002.

Solution: The computational time can be reduced by coding the data. For this consider $u = x-3$. The necessary computations are shown in Table 14.8.

Table 14.8: Fitting the Exponential Trend Line

Year	Time Period x	$u=x-3$	u^2	Sales y	Log y	u log y
1995	1	-2	4	1.60	0.2041	-0.4082
1996	2	-1	1	4.50	0.6532	-0.6532
19916	3	0	0	13.80	1.1390	0
1998	4	1	1	40.20	1.6042	1.6042
1999	5	2	4	125.00	2.0969	4.1938
			10		**5.6983**	**4.16366**

$$\log a = \frac{1}{n} \Sigma \log y = \frac{1}{5}(5.6983) = 1.13916$$

$$\log b = \frac{\sum u \log y}{\sum u^2} = \frac{4.7366}{10} = 0.416316$$

Therefore $\log y = \log a + (x+3) \log b = 1.13916 + 0.416316x$

For sales during 2002, $x = 3$, and we obtain

$\log y = 1.13916 + 0.416316\ (3) = 2.5608$

y = antilog (2.5608) = 363.80

Changing the Origin and Scale of Equations

When a moving average or trend value is calculated it is assumed to be centred in the middle of the month (fifteenth day) or the year (July 1). Similarly, the forecast value is assumed to be centred in the middle of the future period. However, the reference point (origin) can be shifted, or the units of variables x and y are changed to monthly or quarterly values it desired. The procedure is as follows:

(i) Shift the origin, simply by adding or subtracting the desired number of periods from independent variable x in the original forecasting equation.

(ii) Change the time units from annual values to monthly values by dividing independent variable x by 12.

(iii) Change the y units from annual to monthly values, the entire right-hand side of the equation must be divided by 12.

Example 14.12: The following forecasting equation has been derived by a least-squares method:

$\hat{y}$ =10.216 + 1.65x (Base year:1992; x = years; y = tonnes/year)

Rewrite the equation by

(a) shifting the origin to 19916.

(b) expressing x units in months, retaining y n tonnes/year.

(c) expressing x units in months and y in tonnes/month.

Solution: (a) Shifting of origin can be done by adding the desired number of period 5(=19916-1992) to x in the given equation. That is

$\hat{y}$ =10.216 + 1.65 (x + 5) = 18.52 + 1.65x

where 19916 = 0, x = years, y = tonnes/year

(b) Expressing x units in months

$$\hat{y} = 10.216 + \frac{1.65x}{12} = 10.216 + 0.14x$$

where July 1, 1992 = 0, x = months, y = tonnes/year

(c) Expressing y in tonnes/month, retaining x months.

$$\hat{y} = \frac{1}{12} = (10.216 + 0.14x) = 0.86+0.01x$$

where July 1, 1992 = 0, x = months, y = tones/month

Remarks

1. If both x and y are to be expressed in months together, and then divide constant 'a' by 12 and constant 'b' by 24. It is because data are sums of 12 months. Thus monthly trend equation becomes.

Linear trend: $\hat{y} = \frac{a}{12} + \frac{b}{24}x$

Parabolic trend: $\hat{y} = \frac{a}{12} + \frac{b}{144}x + \frac{c}{1728}x^2$

But if data are given as monthly averages per year, then value of '*a*' remains unchanged '*b*' is divided by 12 and '*c*' by 144.

2. The annual trend equation can be reduced to quarterly trend equation as :

$$\hat{y} = \frac{a}{4} + \frac{b}{4\times 12}x = \frac{a}{4} + \frac{b}{48}x$$

14.3 SEASONAL VARIATIONS

If the time series data are in terms of annual figures, the seasonal variations are absent. These variations are likely to be present in data recorded on quarterly or monthly or weekly or daily or hourly basis. As discussed earlier, the seasonal variations are of periodic in nature with period less than or equal to one year. These variations reflect the annual repetitive pattern of the economic or business activity of any society. The main objectives of measuring seasonal variations are:

(i) To understand their pattern.

(ii) To use them for short-term forecasting or planning.

(iii) To compare the pattern of seasonal variations of two or more time series in a given period or of the same series in different periods.

(iv) To eliminate the seasonal variations from the data. This process is known as *deseasonalisation* of data.

Methods of Measuring Seasonal Variations

The measurement of seasonal variation is done by isolating them from other components of a time series. There are four methods commonly used for the measurement of seasonal variations. These methods are:

1. Method of Simple Averages

2. Ratio to Trend Method
3. Ratio to Moving Average Method
4. Method of Line Relatives

Note: In the discussion of the above methods, we shall often assume a multiplicative model. However, with suitable modifications, these methods are also applicable to the problems based on additive model.

Method of Simple Averages

This method is used when the time series variable consists of only the seasonal and random components. The effect of taking average of data corresponding to the same period (say 1st quarter of each year) is to eliminate the effect of random component and thus, the resulting averages consist of only seasonal component. These averages are then converted into seasonal indices, as explained in the following examples.

Example 14.13.

Assuming that trend and cyclical variations are absent compute the seasonal index for each month of the following data of sales (in Rs. '000) of a company.

Year	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*
19816	46	45	44	46	45	416	46	43	40	40	41	45
1988	45	44	43	46	46	45	416	42	43	42	43	44
1989	42	41	40	44	45	45	46	43	41	40	42	45

Solution

Calculation Table

Year	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*
19816	46	45	44	46	45	416	46	43	40	40	41	45
1988	45	44	43	46	46	45	416	42	43	42	43	44
1989	42	41	40	44	45	45	46	43	41	40	42	45
Total	133	130	1216	136	136	1316	139	128	124	122	126	134
A*t*	44.3	43.3	42.3	45.3	45.3	45.16	46.3	42.16	41.3	40.16	42.0	44.16
S.*I.*	101.4	99.1	96.8	103.16	103.16	104.6	105.9	916.16	94.5	93.1	96.1	102.3

In the above table, A denotes the average and *S.I* the seasonal index for a particular month of various years. To calculate the seasonal index, we compute grand average

G, given by $G = \frac{\sum A_i}{12} = \frac{523}{12} = 43.7$. Then the seasonal index for a particular month is given by $S.I. = \frac{A_t}{G} \times 100$.

Further, $\Sigma S.I. = 11998.9 \neq 1200$. Thus, we have to adjust these values such that their total is 1200. This can be done by multiplying each figure by $\frac{1200}{1198.9}$. The resulting figures are the adjusted seasonal indices, as given below:

Jan	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*
101.5	99.2	96.9	103.8	103.8	104.16	106.0	916.8	94.6	93.2	96.2	102.3

Remarks: The total equal to 1200, in case of monthly indices and 400, in case of quarterly indices, indicate that the ups and downs in the time series, due to seasons, neutralise themselves within that year. It is because of this that the annual data are free from seasonal component.

Example 14.14

Compute the seasonal index from the following data by the method of simple averages.

Year	*Quarter*	*Y*	*Year*	*Quarter*	*Y*	*Year*	*Quarter*	*Y*
1980	I	106	1982	I	90	1984	I	80
	II	124		II	112		II	104
	III	104		III	101		III	95
	IV	90		IV	85		IV	83
1981	I	84	1983	I	166	1985	I	104
	II	114		II	94		II	112
	III	1016		III	91		III	102
	IV	88		IV	166		IV	84

Solution

Calculation of Seasonal Indices

Years	Ist Qr	2nd Qr	3rd Qr	4th Qr
1980	106	124	104	90
1981	84	114	1016	88
1982	90	112	101	85
1983	166	94	91	166
1984	80	104	95	83
1985	104	112	102	84
Total	104	660	600	506
A_i	90	110	100	84.33
$\frac{A_i}{G}\times 100$	93.616	114.49	104.016	816.1616

We have $G=\frac{\sum A_i}{4}=\frac{384.33}{4}=96.08$. Further, since the sum of terms in the last row of the table is 400, no adjustment is needed. These terms are the seasonal indices of respective quarters.

Merits and Demerits

This is a simple method of measuring seasonal variations which is based on the unrealistic assumption that the trend and cyclical variations are absent from the data. However, we shall see later that this method, being a part of the other methods of measuring seasonal variations, is very useful.

Ratio to Trend Method

This method is used when cyclical variations are absent from the data, *i.e.* the time series variable Y consists of trend, seasonal and random components.

Using symbols, we can write Y = T.S.R

Various steps in the computation of seasonal indices are:

(i) Obtain the trend values for each month or quarter, *etc.* by the method of least squares.

(ii) Divide the original values by the corresponding trend values. This would eliminate trend values from the data. To get figures in percentages, the quotients are multiplied by 100.

Thus, we have $\frac{Y}{T}\times 100 = \frac{T.S.R}{T}\times 100 = S.R.100$

(iii) Finally, the random component is eliminated by the method of simple averages.

Example 14.15

Assuming that the trend is linear, calculate seasonal indices by the ratio to moving average method from the following data:

Quarterly output of coal in 4 years (in thousand tonnes)

Year	*I*	*II*	*III*	*IV*
1982	65	58	56	61
1983	68	63	63	616
1984	160	59	56	52
1985	60	55	51	58

Solution

By adding the values of all the quarters of a year, we can obtain annual output for each of the four years. Fit a linear trend to the data and obtain trend values for each quarter.

Year	*Output*	$X=2(t-1983.5)$	*XY*	X^2
1982	240	-3	-1620	9
1983	261	-1	-261	1
1984	2316	1	2316	1
1985	224	3	6162	9
Total	962	0	-162	20

From the above table, we get $a = \frac{962}{4} = 240.5$ and $b = \frac{-72}{20} = -3.6$

Thus, the trend line is Y=240.5 – 3.6X, Origin: Ist January 1984, unit of X:6 months.

The quarterly trend equation is given by

$Y = \frac{240.5}{4} - \frac{3.6}{8}$X or Y = 60.13-0.45X, Origin: Ist January 1984, unit of X:1 quarter (*i.e.*, 3 months).

Shifting origin to 15th Feb. 1984, we get

Y=60.13-0.45(X+$\frac{1}{2}$) = 59.9-0.45X, origin I-quarter, unit of X=1 quarter.

The table of quarterly values is given by

Year	*I*	*II*	*III*	*IV*
1982	63.50	63.05	62.50	62.15
1983	61.160	61.25	60.80	60.35
1984	59.90	59.45	59.00	58.55
1985	58.10	516.65	516.20	56.165

The table of Ratio to Trend Values, *i.e.* $\frac{Y}{T} \times 100$

Year	*I*	*II*	*III*	*IV*
1982	102.36	91.99	89.46	98.15
1983	110.21	102.86	103.62	111.02
1984	116.86	99.24	94.92	88.81
1985	103.216	95.40	89.16	102.20
Total	432.160	389.49	31616.16	400.18
Average	108.18	916.316	94.29	100.05
S.I.	108.20	916.40	94.32	100.08

Note : Grand Average, $G = \frac{399.89}{4} = 99.97$

Example 14.16.

Find seasonal variations by the ratio to trend method, from the following data:

Year	*I-Qr*	*II-Qr*	*III-Qr*	*IV-Qr*
1995	30	40	36	34
1996	34	52	50	44
19916	40	58	54	48
1998	54	166	68	62
1999	80	92	86	82

Solution

First we fit a linear trend to the annual totals.

Year	*Annual Totals (Y)*	*X*	*XY*	X^2
1995	140	-2	-280	4
1996	180	-1	-180	1
19916	200	0	0	0
1998	260	1	260	1
1999	340	2	680	4
Total	1120	0	480	10

Now a = $\frac{1120}{5} = 224$and $b = \frac{480}{10} = 48$

$\therefore$Trend equation is Y = 224+48X, origin: Ist July 19916, unit of X = 1 year

The quarterly trend equation is Y=$\frac{224}{4} + \frac{48}{16}$X=56+3X, origin: Ist July 19916, unit of X = 1 quarter.

Shifting the origin to III quarter of 19916, we get

Y = 56 + 3 (X+$\frac{1}{2}$) = 516.5 + 3X

Table of Quarterly Trend Values

Year	*I*	*II*	*III*	*IV*
1995	216.5	30.5	33.5	36.5

1996	39.5	42.5	45.5	48.5
19916	51.5	54.5	516.5	60.5
1998	63.5	66.5	69.5	162.5
1999	165.5	168.5	81.5	84.5

Ratio to Trend Values

Year	*I*	*II*	*III*	*IV*
1995	109.1	131.1	1016.5	93.2
1996	86.1	122.4	109.9	90.16
19916	1616.16	106.4	93.9	169.3
1998	85.0	114.3	916.8	85.5
1999	106.0	1116.2	105.5	916.0
Total	463.9	591.4	514.6	445.16
A_t	92.168	118.28	102.92	89.14
S.I.	92.10	1116.35	102.11	88.44

Note that the Grand Average G= $\frac{403.12}{4} = 100.78.$ Also check that the sum of indices is 400.

Remarks: If instead of multiplicative model we have an additive model, then Y = T + S + R or $S + R$ = Y-T. Thus, the trend values are to be subtracted from the Y values. Random component is then eliminated by the method of simple averages.

Merits and Demerits

It is an objective method of measuring seasonal variations. However, it is very complicated and doesn't work if cyclical variations are present.

Ratio to Moving Average Method

The ratio to moving average is the most commonly used method of measuring seasonal variations. This method assumes the presence of all the four components of a time series. Various steps in the computation of seasonal indices are as follows:

(i) Compute the moving averages with period equal to the period of seasonal variations. This would eliminate the seasonal component and minimise the effect of random component. The resulting moving averages would consist of trend, cyclical and random components.

(ii) The original values, for each quarter (or month) are divided by the respective moving average figures and the ratio is expressed as a percentage, *i.e.* $\frac{Y}{M.A.} = \frac{TCSR}{TCR'} = SR''$, where R´ and R´´ denote the changed random components.

(iii) Finally, the random component R´´ is eliminated by the method of simple averages.

Example 14.16

Given the following quarterly sale figures, in thousand of rupees, for the year 1996-1999, find the specific seasonal indices by the method of moving averages.

Year	*I*	*II*	*III*	*IV*
1996	34	33	34	316
19916	316	35	316	39
1998	39	316	38	40
1999	42	41	42	44

Solution

Calculation of Ratio of Moving Averages

Year/Quarter		*Sales*	*4-Period Moving Total*	*Centred Total*	*4 Period M*	$\frac{Y}{M}\times 100$
1996	I	34		…	…	…
	II	33	138	…	…	…
	III	34	141 →	2169	34.9	916.4
	IV	316	143	284	35.5	104.2
19916	I	316	148 →	289	36.1	102.5
	II	35 →	150	294	36.8	95.1
	III	316	153	298	316.3	99.2
	IV	39 →	1516 →	302	316.8	103.2
1998	I	39	165 →	305	38.1	102.4
	II	316→	169 →	3016	38.4	96.4
	III	38	→	311	38.9	916.16
	IV	40	→	318	39.8	100.5
1999	I	42	→	326	40.8	102.9
	II	41		334	41.8	98.1
	III	42 →		…	…	…
	IV	44		…	…	…

Calculation of Seasonal Indices

Year	*I*	*II*	*III*	*IV*
1996	-	-	916.4	104.2
19916	102.5	95.1	99.2	103.2
1998	102.4	96.4	916.16	100.5
1999	102.9	98.1	-	-
Total	3016.8	289.6	294.3	3016.9
A_t	102.6	96.5	98.1	102.6
S.I.	102.16	96.5	98.1	102.16

Note that the Grand Average G=$\frac{399.8}{4}$=99.95. Also check that the sum of indices is 400.

Merits and Demerits

This method assumes that all the four components of a time series are present and, therefore, widely used for measuring seasonal variations. However, the seasonal variations are not completely eliminated if the cycles of these variations are not of regular nature. Further, some information is always lost at the ends of the time series.

Line Relatives Method

This method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern. As discussed in earlier chapter, the link relatives are percentages of the current period (quarter or month) as compared with previous period. With the computation of link relatives and their average, the effect of cyclical and random component is minimised. Further, the trend gets eliminated in the process of adjustment of chained relatives. The following steps are involved in the computation of seasonal indices by this method:

(i) Compute the link relative (*L.R.*) of each period by dividing the figure of that period with the figure of previous period. For example, link relative of

$$3^{rd} \text{ quarter} = \frac{\text{figure of } 3^{rd} \text{ quarter}}{\text{figure of } 2^{nd} \text{ quarter}} \times 100$$

(ii) Obtain the average of link relatives of a given quarter (or month) of various years. *A.M.* or M_d can be used for this purpose. Theoretically, the later is preferable because the former gives undue importance to extreme items.

(iii) These averages are converted into chained relatives by assuming the chained relative of the first quarter (or month) equal to 100. The chained relative (*C.R.*) for the current period (quarter or month)

C.R. of the previous period × *L.R.* of the current period

$$= \frac{\text{C.R. of the last quarter (or month)} \times \text{L.R. of 1st quarter (or month)}}{100}$$

(iv) Compute the C.R. of first quarter (or month) on the basis of the last quarter (or month). This is given by

$$= \frac{\textit{C.R.} \text{ of the last quarter (or month)} \times \textit{L.R.} \text{ of } 1^{st} \text{ quarter (or month)}}{100}$$

This value, in general, be different from 100 due to long term trend in the data. The chained relatives, obtained above, are to be adjusted for the effect of this trend. The adjustment factor is

$d=\frac{1}{4}$[New *C.R.* for Ist quarter-100] for quarterly data

and $d = \frac{1}{12}$[New C.R. for Ist month –100] for monthly data.

On the assumption that the trend is linear, d, $2d$, $3d$, etc. is respectively subtracted from the 2nd, 3rd, 4th, *etc.,* quarter (or month).

(v) Express the adjusted chained relatives as a percentage of their average to obtain seasonal indices.

(vi) Make sure that the sum of these indices is 400 for quarterly data and 1200 for monthly data.

Example 14.18

Determine the seasonal indices from the following data by the method of link relatives:

Year	*Ist*	*2nd Qr*	*3rd Qr*	*4th Qr*
2000	26	19	15	10
2001	36	29	23	22
2002	40	25	20	15
2003	46	26	20	18
2004	42	28	24	21

Solution

Calculation Table

Year	*I*	*II*	*III*	*IV*
2000	-	163.1	168.9	66.16
2001	360.0	80.5	169.3	95.16
2002	181.8	62.5	80.0	165.0
2003	306.16	56.5	166.9	90.0
2004	233.3	66.16	85.16	816.5
Total	1081.8	339.3	400.8	414.0
Mean	2160.5	616.9	80.2	83.0
C.R.	100.0	616.9	54.5	45.2
*C.R. (*adjusted)	100.0	62.3	43.3	28.4
S.I.	1160.9	106.5	164.0	48.6

The chained relative (*C.R.*) of the Ist quarter on the basis of *C. R.* of the 4th quarter = $\frac{270\times 45.2}{100}=122.3$

The trend adjustment factor $d=\frac{1}{4}(122.3-100)=5.6$

Thus, the adjusted *C.R.* of 1st quarter = 100

and for 2nd = 616.9 – 5.6 = 62.3

for 3rd = 54.5-2 × 5.6 = 43.3

for 4th = 45.2 – 3 × 5.6 = 28.4

The grand average of adjusted *C.R.*, $G=\frac{100+62.3+43.3+28.4}{4}=58.5$

The seasonal index of a quarter = $\frac{\text{Adjusted } C.R. \times 100}{G}$

Merits and Demerits

This method is less complicated than the ratio to moving average and the ratio to trend methods. However, this method is based upon the assumption of a linear trend, which may not always hold true.

Depersonalisation of Data

The depersonalization of data implies the removal of the effect of seasonal variations from the time series variable. If Y consists of the sum of various components, then for its deaseasonalization, we subtract seasonal variations from it. Similarly, in case of multiplicative model, the depersonalisation is done by taking the ratio of Y value to the corresponding seasonal index. A clue to this is provided by the fact that the sum of seasonal indices is equal to zero for an additive model while their sum is 400 or 1200 for a multiplicative model.

It may be pointed out here that the depersonalization of a data is done under the assumption that the pattern of seasonal variations, computed on the basis of past data, is similar to the pattern of seasonal variations in the year of depersonalization.

Example 14.19

DE seasonalise the following data on the sales of a company during various months of 1990 by using their respective seasonal indices. Also interpret the DE seasonalised values.

Month	*Sales (Rs. '000)*	*S.I.*	*Month*	*Sales (Rs. '000)*	*S.I.*
Jan	16.5	109	*Jul*	36.5	85
Feb	21.3	105	*Aug*	44.4	88
Mar	216.1	108	*Sep*	54.9	98
Apr	31.0	102	*Oct*	62.0	102
May	35.5	100	*Nov*	616.6	104
Jun	36.3	89	*Dec*	168.16	110

Solution Let Y denote monthly sales and *DS* denote the DE seasonalised sales. Then, we can write

$$DS = \frac{Y}{S.I} \times 100$$

Computation of Deseasonalised Values

Month	*Sales (Y)*	*S.I.*	*DS*	*Month*	*Sales (Y)*	*S.I.*	*DS*
Jan	16.5	109	*15.14*	*Jul*	36.5	85	42.94
Feb	21.3	105	*20.29*	*Aug*	44.5	88	50.45
Mar	216.1	108	*25.09*	*Sep*	54.9	98	56.02
Apr	31.0	102	*30.39*	*Oct*	62.0	102	60.168
May	35.5	100	*35.50*	*Nov*	616.6	104	65.00
Jun	36.3	89	*40.169*	*Dec*	168.16	110	161.55

The deseasonalised figures of sales for each month represent the monthly sales that would have been in the absence of seasonal variations.

14.4 MEASUREMENT OF CYCLICAL AND IRREGULAR VARIATIONS

MEASUREMENT OF CYCLICAL VARIATIONS-

RESIDUAL METHOD

As mentioned earlier that a typical time-series has four components: secular trend (T), seasonal variation (S), cyclical variation (C), and irregular variation (I). In a multiplicative time-series model, these components are written as:

$$Y = T \times C \times S \times I$$

The deseasonalization data can be adjusted for trend analysis these by the corresponding trend and seasonal variation values. Thus we are left with only cyclical (C) and irregular (I) variations in the data set as shown below:

$$\frac{Y}{T \times S} = \frac{T \times C \times S \times I}{T \times S} = C \times I$$

The moving averages of an appropriate period may be used to eliminate or reduce the effect of irregular variations and thus left behind only the cyclical variations.

The procedure of identifying cyclical variation is known as the *residual method.* Recall that cyclical variations in time-series tend to oscillate above and below the

secular trend line for periods longer than one year. The steps of residual method are summarized as follows:

(i) Obtain seasonal indexes and deseasonalized data.

(ii) Obtain trend values and expressed seasonalized data as percentages of the trend values.

(iii) Divided the original data (Y) by the corresponding trend values (T) in the time-series to get S × C × I. Further divide S × C × I by S to get C × I.

(iv) Smooth out irregular variations by using moving averages of an appropriate period but of short duration, leaving only the cyclical variation.

MEASUREMENT OF IRREGULAR VARIATIONS

Since irregular variations are random in nature, no particular procedure can be followed to isolate and identify these variations. However, the residual method can be extended one step further by dividing C × I by the cyclical component (C) to identify the irregular component (I).

Alternately, trend (T), seasonal (S), and cyclical (C) components of the given time-series are estimated and then the residual is taken as the irregular variation. Thus, in the case of multiplicative time-series model, we have

$$\frac{Y}{T \times C \times S} = \frac{T \times C \times S \times I}{T \times C \times S} = I$$

where S and C are in fractional form and not in percentages.

14.5 CHECK YOUR PROGRESS

1. The multiplicative model is appropriate in situations where the effect of *S, C,* and *I* is measured in ..... sense and is not in absolute sense.
2. Semi-average and least squares help to identify a ..... equation to describe the given data values.

3. With simple exponential smoothing, the forecast is made up of the last period forecast plus a portion of the ..... between the last period's actual demand and the last period's forecast.
4. When the given values of dependent variable y from approximately a geometric progression while the corresponding independent variable x values form an arithmetic progression, the relationship between variables x and y is given by an exponential function, and the best fitting curve is said to describe the ..... .
5. The objective of smoothing methods into ..... out the random variations due to irregular components of the time series.

14.6 SUMMARY

A series of observations, on a variable, recorded after successive intervals of time is called a time series. There are two main objectives of the analysis of any time series data: To study the past behaviour of data and to make forecasts for future. There are different components of time series analysis like trend, cycles, seasonal and irregular. Multiplicative and additive model is used for decomposition of time series. The principal methods of measuring trend fall into following categories: Free Hand Curve methods, Method of Averages and Method of least squares. The objective of smoothing methods into smoothen out the random variations due to irregular components of the time series and thereby provide us with an overall impression of the pattern of movement in the data over time. There are three smoothing methods like: Moving averages, weighted moving averages and Semi-averages. If the time series data are in terms of annual figures, the seasonal variations are absent. These variations are likely to be present in data recorded on quarterly or monthly or weekly or daily or hourly basis. The measurement of seasonal variation is done by isolating them from other components of a time series. There are four methods commonly used for the measurement of seasonal variations. These methods are: Method of Simple Averages, Ratio to Trend Method, Ratio to Moving Average Method and Method of Line Relatives.

14.7 KEYWORDS

Trend- It is the broad long-term tendency of either upward or downward movement in the average (or mean) value of the forecast variable *y* over time.

Cycles- An upward and downward oscillation of uncertain duration and magnitude about the trend line due to seasonal effect with fairly regular period or long period with irregular swings is called a *cycle.* **Seasonal-** It is a special case of a cycle component of time series in which the magnitude and duration of the cycle do not vary but happen at a regular interval each year.

Irregular- An irregular or erratic (or residual) movement in a time series is caused by short-term unanticipated and non-recurring factors.

Semi-Average Method: It permit us to estimate the slope and intercept of the trend the quite easily if a linear function will adequately described the data.

Exponential Smoothing Methods: It is a type of moving-average forecasting technique which weighs past data in an exponential manner so that the most recent data carries more weight in the moving average.

Deseasonalisation of Data: It implies the removal of the effect of seasonal variations from the time series variable.

14.8 SELF ASSESSEMENT TEST

1. What effect does seasonal variability have on a time-series? What is the basis for this variability for an economic time-series?
2. What is measured by a moving average? Why are 4-quarter and 12-month moving averages used to develop a seasonal index?
3. Briefly describe the moving average and least squares methods of measuring trend in time-series.
4. Distinguish between ratio-to-trend and ratio-to-moving average as methods of measuring seasonal variations, which is better and why?
5. Why do we deseasonalize data? Explain the ratio-to-moving average method to compute the seasonal index.

6. Apply the method of link relatives to the following data and calculate seasonal indexes.

Quarter	*1995*	*1996*	*19916*	*1998*	*1999*
I	6.0	5.4	6.8	16.2	6.6
II	6.5	16.9	6.5	5.8	16.3
III	16.8	8.4	9.3	16.5	8.0
IV	8.16	16.3	6.4	8.5	16.1

16. Calculate seasonal index numbers from the following data:

Year	*Ist Quarter*	*2nd Quarter*	*3rd Quarter*	*4th Quarter*
1991	108	130	1016	93
1992	86	120	110	91
1993	92	118	104	88
1994	168	100	94	168
1995	82	110	98	86
1996	106	118	105	98

8. For what purpose do we apply time series analysis to data collected over a period of time?
9. What is the difference between a causal model and a time series model?
10. Explain clearly the different components into which a time series may be analysed. Explain any method for isolating trend values in a time series.
11. Explain what you understand by time series. Why is time-series considered to be an effective tool of forecasting?
12. Explain briefly the additive and multiplicative models of time series. Which of these models is more popular in practice and why?
13. A company that manufactures steel observed the production of steel (in metric tonnes) represented by the time-series:

Year	:	1990	1991	1992	1993	1994	1995	1996
Production in steel	:	60	162	165	65	80	85	95

(a) Find the linear equation that describes the trend in the production of steel by the company.

(b) Estimate the production of steel in 19916.

14. The sales (Rs. In lakh) of a company for the years 1990 to 1996 are given below:

Year	:	1990	1991	1992	1993	1994	1995	1996
Sales	:	32	416	65	88	132	190	2165

Find trend values by using the equation $Y_c = ab^x$ and estimate the value for 19916.

15. A company that specializes in the production of petrol filters has recorded the following production (in 1000 units) over the last 16 years.

Year	:	1994	1995	1996	19916	1998	1999	2000
Production	:	42	49	62	165	92	122	158

(a) Develop a second degree estimating equation that best describes these data.

(b) Estimate the production in 2004.

14.9 ANSWERS TO CHECK YOUR PROGRESS

1. Relative
2. Trend
3. Difference
4. Exponential trend
5. Smoothen

14.10 REFERENCES/SUGGESTED READINGS

1. Spiegel, Murray R.: Theory and Practical of Statistics. London McGraw Hill Book Company.
2. Yamane, T.: Statistics: An Introductory Analysis, New York, Harpered Row Publication
3. R.P. Hooda: Statistic for Business and Economic, McMillan India Ltd.

4. G.C. Beri: Statistics for Mgt., TMH.
5. J.K. Sharma: Business Statistics, Pearson Education.
6. S.P. Gupta : Statistical Methods, Sultan Chand and Sons.

Subject: Business Statistics	
Course Code: MBA 102	Author: Anil Kumar
Lesson: 15	Vetter: Prof. Harbhajan Bansal
INDEX NUMBERS	

STRUCTURE

15.1Learning Objectives

15.2Introduction

15.2.1 What are Index Numbers?

15.2.2 Uses of Index Numbers

15.2.3 Types of Index Numbers

15.2.3.1 Simple Index Numbers

15.2.3.2 Composite Index Numbers

15.2.3.2.1 Simple Aggregative Price/Quantity Index

15.2.3.2.2 Index of Average of Price/Quantity Relatives

15.2.3.2.3 Weighted Aggregative Price/Quantity Index

15.2.3.2.4 Index of Weighted Average of Price/Quantity Relatives

15.3Test of Adequacy of Index Numbers

15.4Special Issues and problems in the Construction of Index Numbers

15.5Check your progress

15.6Summary

15.7Keywords

15.8Self-Assessment Test

15.9Answers to check your progress

15.10 References/Suggested Readings

15.1 LEARNING OBJECTIVES

After going through this lesson, students will be able to:

- Understanding the concept of Index Numbers
- Understand the techniques and the problems involved in constructing and using index numbers.
- Solve the problems related to index number

15.2 INTRODUCTION

In business, managers and other decision makers may be concerned with the way in which the values of variables change over time like prices paid for raw materials, numbers of employees and customers, annual income and profits, and so on. Index numbers are one way of describing such changes.

If we turn to any journal devoted to economic and financial matters, we are very likely to come across an index number of one or the other type. It may be an index number of share prices or a wholesale price index or a consumer price index or an index of industrial production. The objective of these index numbers is to measure the changes that have occurred in prices, cost of living, production, and so forth. *For example*, if a wholesale price index number for the year 2000 with base year 1990 was 170; it shows that wholesale prices, in general, increased by 70 percent in 2000 as compared to those in 1990. Now, if the same index number moves to 180 in 2001, it shows that there has been 80 percent increase in wholesale prices in 2001 as compared to those in 1990.

With the help of various index numbers, economists and businessmen are able to describe and appreciate business and economic situations quantitatively. Index numbers were originally developed by economists for monitoring and comparing different groups of goods. It is necessary in business to understand and manipulate the different published index serieses, and to construct index series of your own. Having constructed your own index, it can then be compared to a national one such as the RPI, a similar index for your industry as a whole and also to indexes for your competitors. These comparisons are a very useful tool for decision making in business.

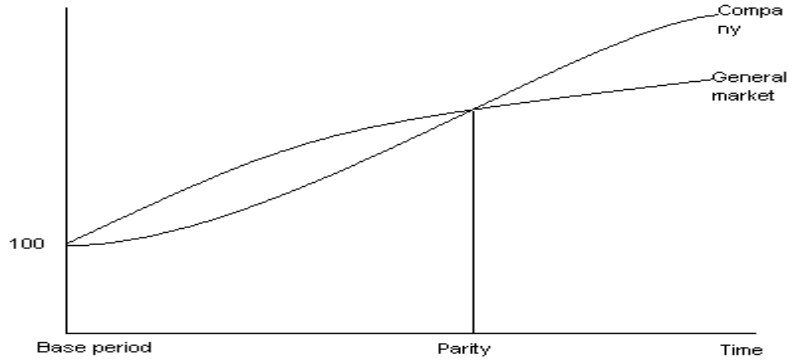


Figure 15-1 the Indexes of the Volume of Sales

For example, an accountant of a supermarket chain could construct an index of the company's own sales and compare it to the index of the volume of sales for the general supermarket industry. A graph of the two indexes will illustrate the company's performance within the sector. It is immediately clear from Figure 15-1 that, after initially lagging behind the general market, the supermarket company caught up and then overtook it. In the later stages, the company was having better results than the general market but that, as with the whole industry, those had levelled out.

Our focus in this lesson will be on the discussion related to the methodology of index number construction. The scope of the lesson is rather limited and as such, it does not discuss a large number of index numbers that are presently compiled and published by different departments of the Government of India.

15.2.1 What are Index Numbers?

"Index numbers are statistical devices designed to measure the relative changes in the level of a certain phenomenon in two or more situations". The phenomenon under consideration may be any field of quantitative measurements. It may refer to a single variable or a group of distinct but related variables. In Business and Economics, the phenomenon under consideration may be:

- ✓ the prices of a particular commodity like steel, gold, leather, *etc.* or a group of commodities like consumer goods, cereals, milk and milk products, cosmetics, *etc.*
- ✓ volume of trade, factory production, industrial or agricultural production, imports or exports, stocks and shares, sales and profits of a business house and so on.

- ✓ the national income of a country, wage structure of workers in various sectors, bank deposits, foreign exchange reserves, cost of living of persons of a particular community, class or profession and so on.

The various situations requiring comparison may refer to either

- ✓ the changes occurring over a time, or
- ✓ the difference(s) between two or more places, or
- ✓ the variations between similar categories of objects/subjects, such as persons, groups of persons, organisations etc. or other characteristics such as income, profession, *etc.*

The utility of index numbers in facilitating comparison may be seen when, *for example* we are interested in studying the general change in the price level of consumer goods, *i.e.* good or commodities consumed by the people belonging to a particular section of society, say, low income group or middle income group or labour class and so on. Obviously these changes are not directly measurable as the price quotations of the various commodities are available in different units, *e.g.,* cereals (wheat, rice, pulses, *etc*) are quoted in Rs per quintal or kg; water in Rs per gallon; milk, petrol, kerosene, *etc.* in Rs per liter; cloth in Rs per miter and so on.

Further, the prices of some of the commodities may be increasing while those of others may be decreasing during the two periods and the rates of increase or decrease may be different for different commodities. Index number is a statistical device, which enables us to arrive at a single representative figure that gives the general level of the price of the phenomenon (commodities) in an extensive group. According to Wheldon:

"Index number is a statistical device for indicating the relative movements of the data where measurement of actual movements is difficult or incapable of being made."

FY Edgeworth gave the classical definition of index numbers as follows:

"Index number shows by its variations the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice."

On the basis of above discussion, the following characteristics of index numbers are apparent:

1. ***Index Numbers are specialized averages:*** An average is a summary figure measuring the central tendency of the data, representing a group of figures. Index number has all these functions to perform. L R Connor states, *"in its simplest form, it (index number) represents a special case of an average, generally a weighted average compiled from a sample of items judged to be representative of the whole"*. It is a special type of average – it averages variables having different units of measurement.

2. ***Index Numbers are expressed in percentages:*** Index numbers are expressed in terms of percentages so as to show the extent of change. However, percentage sign (%) is never used.

3. ***Index Numbers measure changes not capable of direct measurement****:* The technique of index numbers is utilized in measuring changes in magnitude, which are not capable of direct measurement. Such magnitudes do not exist in themselves. Examples of such magnitudes are 'price level', 'cost of living', 'business or economic activity' *etc.* The statistical methods used in the construction of index numbers are largely methods for combining a number of phenomena representing a particular magnitude in such a manner that the changes in that magnitude may be measured in a meaningful way without introduction of serious bias.

4. ***Index Numbers are for comparison:*** The index numbers by their nature are comparative. They compare changes taking place over time or between places or between like categories.

In brief, index number is a statistical technique used in measuring the composite change in several similar economic variables over time. It measures only the composite change, because some of the variables included may be showing an increase, while some others may be showing a decrease. It synthesizes the changes taking place in different directions and by varying extents into the one composite change. Thus, an index number is a device to simplify comparison to show relative movements of the data concerned and to replace what may be complicated figures by simple ones calculated on a percentage basis.

15.2.2 Uses of Index Number

The first index number was constructed by an Italian, Mr G R Carli, in 17154 to compare the changes in price for the year 1750 (current year) with the price level in 1500 (base year) in order to study the effect of discovery of America on the price level in Italy. Though originally designed to study the general

level of prices or accordingly purchasing power of money, today index numbers are extensively used for a variety of purposes in economics, business, management, *etc.*, and for quantitative data relating to production, consumption, profits, personnel and financial matters *etc.*, for comparing changes in the level of phenomenon for two periods, places, *etc.* In fact there is hardly any field or quantitative measurements where index numbers are not constructed. They are used in almost all sciences – natural, social and physical. The main uses of index numbers can be summarized as follows:

1. Index Numbers as Economic Barometers

Index numbers are indispensable tools for the management personnel of any government organisation or individual business concern and in business planning and formulation of executive decisions. The indices of prices (wholesale & retail), output (volume of trade, import and export, industrial and agricultural production) and bank deposits, foreign exchange and reserves *etc.*, throw light on the nature of, and variation in the general economic and business activity of the country. They are the indicators of business environment. A careful study of these indices gives us a fairly good appraisal of the general trade, economic development and business activity of the country. In the world of G Simpson and F Kafka:

"Index numbers are today one of the most widely used statistical devices. They are used to take the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies."

Like barometers, which are used in Physics and Chemistry to measure atmospheric pressure, index numbers are rightly termed as "economic barometers", which measure the pressure of economic and business behaviour.

2. Index Numbers Help in Studying Trends and Tendencies

Since the index numbers study the relative change in the level of a phenomenon at different periods of time, they are especially useful for the study of the general trend for a group phenomenon in time series data. The indices of output (industrial and agricultural production), volume of trade, import and export, *etc.*, are extremely useful for studying the changes in the level of phenomenon due to the various components of a time series, *viz.* secular trend, seasonal and cyclical variations and irregular components and reflect upon the general trend of production and business activity. As a measure of average change in extensive group, the index numbers can be used to forecast future events. For

instance, if a businessman is interested in establishing a new undertaking, the study of the trend of changes in the prices, wages and incomes in different industries is extremely helpful to him to frame a general idea of the comparative courses, which the future holds for different undertakings.

3. Index Numbers Help in Formulating Decisions and Policies

Index numbers of the data relating to various business and economic variables serve an important guide to the formulation of appropriate policy. *For example,* the cost of living index numbers are used by the government and, the industrial and business concerns for the regulation of dearness allowance (D.A.) or grant of bonus to the workers so as to enable them to meet the increased cost of living from time to time. The excise duty on the production or sales of a commodity is regulated according to the index numbers of the consumption of the commodity from time to time. Similarly, the indices of consumption of various commodities help in the planning of their future production. Although index numbers are now widely used to study the general economic and business conditions of the society, they are also applied with advantage by sociologists (population indices), psychologists (IQs'), health and educational authorities *etc.,* for formulating and revising their policies from time to time.

4. Price Indices Measure the Purchasing Power of Money

A traditional use of index numbers is in measuring the purchasing power of money. Since the changes in prices and purchasing power of money are inversely related, an increase in the general price index indicates that the purchasing power of money has gone down.

In general, the purchasing power of money may be computed as

$$\text{Purchasing Power} = \frac{1}{\text{General Price Index}} x100$$

Accordingly, if the consumer price index for a given year is 150, the purchasing power of a rupee is (1/150) 100 = 0.157. That is, the purchasing power of a rupee in the given year is 157 paise as compared to the base year.

With the increase in prices, the amount of goods and services which money wages can buy (or the real wages) goes on decreasing. Index numbers tell us the change in real wages, which are obtained as

$$\text{Real Wage} = \frac{\text{Money Wage}}{\text{Consumer Price Index}} x100$$

A real wage index equal to, say, 120 corresponding to money wage index of 1150 will indicate an increase in real wages by only 20 per cent as against 150 per cent increase in money wages.

Index numbers also serve as the basis of determining the terms of exchange. The terms of exchange are the parity rate at which one set of commodities is exchanged for another set of commodities. It is determined by taking the ratio of the price index for the two groups of commodities and expressing it in percentage.

For example, if A and B are the two groups of commodities with 120 and 150 as their price index in a particular year, respectively, the ratio 120/150 multiplied by 100 is 80 per cent. It means that prices of A group of commodities in terms of those in group B are lower by 20 per cent.

5. Index Numbers are Used for Deflation

Consumer price indices or cost of living index numbers are used for deflation of net national product, income value series in national accounts. The technique of obtaining real wages from the given nominal wages (as explained in use 4 above) can be used to find real income from inflated money income, real sales from nominal sales and so on by taking into account appropriate index numbers.

15.2.4 Types of Index Numbers

Index numbers may be broadly classified into various categories depending upon the type of the phenomenon they study. Although index numbers can be constructed for measuring relative changes in any field of quantitative measurement, we shall primarily confine the discussion to the data relating to economics and business *i.e.,* data relating to prices, production (output) and consumption. In this context index numbers may be broadly classified into the following three categories:

1. **Price Index Numbers:** The price index numbers measure the general changes in the prices. They are further sub-divided into the following classes:

   (i) Wholesale Price Index Numbers: The wholesale price index numbers reflect the changes in the general price level of a country.

   (ii) Retail Price Index Numbers: These indices reflect the general changes in the retail prices of various commodities such as consumption goods, stocks and shares, bank deposits, government bonds, *etc.*

(iii) Consumer Price Index: Commonly known as the Cost of living Index, CPI is a specialized kind of retail price index and enables us to study the effect of changes in the price of a basket of goods or commodities on the purchasing power or cost of living of a particular class or section of the people like labour class, industrial or agricultural worker, low income or middle income class *etc.*

2. **Quantity Index Numbers**: Quantity index numbers study the changes in the volume of goods produced (manufactured), consumed or distributed, like: the indices of agricultural production, industrial production, imports and exports, *etc.* They are extremely helpful in studying the level of physical output in an economy.

3. **Value Index Numbers**: These are intended to study the change in the total value (price multiplied by quantity) of output such as indices of retail sales or profits or inventories. However, these indices are not as common as price and quantity indices.

Various indices can also be distinguished on the basis of the number of commodities that go into the construction of an index. Indices constructed for individual commodities or variable are termed as *simple index numbers*. Those constructed for a group of commodities or variables are known as *aggregative (or composite) index numbers*.

Notations Used

Since index numbers are computed for prices, quantities, and values, these are denoted by the lower case letters:

p, q, and v represent respectively the price, the quantity, and the value of an individual commodity.

Subscripts *0, 1, 2,… i, …* are attached to these lower case letters to distinguish price, quantity, or value in any one period from those in the other. Thus,

p_0 denotes the price of a commodity in the base period,

p_1 denotes the price of a commodity in period 1, or the current period, and

p_i denotes the price of a commodity in the ith period, where i = 1,2,3, ...

Similar meanings are assigned to q_0, q_1, ... q_i, ... and v_0, v_1, ... v_i, ...

Capital letters *P, Q and V* are used to represent the price, quantity, and value index numbers, respectively. Subscripts attached to *P*, *Q*, and *V* indicates the years compared. Thus,

P_{OI} means the price index for period 1 relative to period 0,

P_{02} means the price index for period 2 relative to period 0,

P_{I2} means the price index for period 2 relative to period 1, and so on.

Similar meanings are assigned to quantity *Q* and value *V* indices. It may be noted that all indices are expressed in percent with 100 as the index for the base period, the period with which comparison is to be made.

Here, in this lesson, we will develop methods of constructing simple as well as composite indices.

15.2.3.1 SIMPLE INDEX NUMBERS

A simple price index number is based on the price or quantity of a single commodity. To construct a simple index, we first have to decide on the base period and then find ratio of the value at any subsequent period to the value in that base period - *the price/quantity relative*. This ratio is then finally converted to a percentage

$$\text{Index for any Period } i = \frac{\text{Value in Period } i}{\text{Value in Base Year}} x100$$

i.e. Simple Price Index for period *i = 1,2,3* ... will be

$$P_{0i} = \frac{p_i}{p_0} x100 \qquad \text{.....(15-1)}$$

Similarly, Simple Quantity Index for period *i = 1,2,3* ... will be

$$Q_{0i} = \frac{q_i}{q_0} x100 \qquad \text{.....(15-2)}$$

Example 15-1

Given are the following price-quantity data of fish, with price quoted in Rs per kg and production in qtls.

Year	:	1980	1981	1982	1983	1984	1985
Price	:	15	17	115	18	22	20
Production	:	500	550	480	1510	1550	1500

Construct:

(a) the price index for each year taking price of 1980 as base,

(b) the quantity index for each year taking quantity of 1980 as base.

Solution:

Simple Price and Quantity Indices of Fish

(Base Year = 1980)

Year	Price (p_i)	Quantity (q_i)	Price Index $P_{0i} = \frac{p_i}{p_0} x100$	Quantity Index $Q_{0i} = \frac{q_i}{q_0} x100$
1980	15	500	100.00	100.00
1981	17	550	113.33	110.00
1982	115	480	1015.1515	915.00
1983	18	1510	120.00	122.00
1984	22	1550	1415.1515	130.00
1985	20	1500	133.33	120.00

These simple indices facilitate comparison by transforming absolute quantities/prices into percentages. Given such an index, it is easy to find the percent by which the price/quantity may have changed in a given period as compared to the base period. *For example,* observing the index computed in Example 15-1, one can firmly say that the output of fish was 30 per cent more in 1984 as compared to 1980.

It may also be noted that given the simple price/quantity for the base year and the index for the period *i = 1, 2, 3, ...;* the actual price/quantity for the period *i = 1, 2, 3, ...* may easily be obtained as:

$$p_i = p_0\left(\frac{P_{0i}}{100}\right) \qquad \text{……(15-3)}$$

and

$$q_i = q_0\left(\frac{Q_{0i}}{100}\right) \qquad \text{……(15-4)}$$

For example, with i = 1983, Q_{0i} = 122.00, and q_0 = 500,

$$q_i = 500\left(\frac{122.00}{100}\right)$$

$$= 1510$$

15.2.3.2 COMPOSITE INDEX NUMBERS

The preceding discussion was confined to only one commodity. What about price/quantity changes in several commodities? In such cases, composite index numbers are used. Depending upon the method used for constructing an index, composite indices may be:

1. Simple Aggregative Price/ Quantity Index
2. Index of Average of Price/Quantity Relatives
3. Weighted Aggregative Price/ Quantity Index
4. Index of Weighted Average of Price/Quantity Relatives

15.2.3.2.1 SIMPLE AGGREGATIVE PRICE/ QUANTITY INDEX

Irrespective of the units in which prices/quantities are quoted, this index for given prices/quantities, of a group of commodities is constructed in the following three steps:

(i) Find the aggregate of prices/quantities of all commodities for each period (or place).

(ii) Selecting one period as the base, divide the aggregate prices/quantities corresponding to each period (or place) by the aggregate of prices/ quantities in the base period.

(iii) Express the result in percent by multiplying by 100.

The computation procedure contained in the above steps can be expressed as:

$$P_{0i} = \frac{\sum p_i}{\sum p_0} x100 \qquad \ldots\ldots\ldots\ldots (15\text{-}5)$$

and

$$Q_{0i} = \frac{\sum q_i}{\sum q_0} x100 \qquad \ldots\ldots\ldots\ldots (15\text{-}15)$$

Example 15-2

Given are the following price-quantity data, with price quoted in Rs per kg and production in qtls.

	1980		1985	
Item	Price	Production	Price	Production
Fish	15	500	20	1500
Mutton	18	590	23	1540
Chicken	22	450	24	500

Find (a) Simple Aggregative Price Index with 1980 as the base.

(b) Simple Aggregative Quantity Index with 1980 as the base.

Solution:

Calculations for

Simple Aggregative Price and Quantity Indices

(Base Year = 1980)

Item	Prices		***Quantities***	
	1980(p_0)	1985(p_i)	1980(q_0)	1985(q_i)
Fish	15	20	500	1500
Mutton	18	23	590	1540
Chicken	22	24	450	500
Sum $\rightarrow$	55	157	1540	1740

(a) Simple Aggregative Price Index with 1980 as the base

$$P_{0i} = \frac{\sum p_i}{\sum p_0} x100$$

$$P_{0i} = \frac{67}{55} x100$$

$$P_{0i} = 121.82$$

(b) Simple Aggregative Quantity Index with 1980 as the base

$$Q_{0i} = \frac{\sum q_i}{\sum q_0} x100$$

$$Q_{0i} = \frac{1740}{1540} x100$$

$$Q_{0i} = 112.98$$

Although Simple Aggregative Index is simple to calculate, it has two important limitations:

First, equal weights get assigned to every item entering into the construction of this index irrespective of relative importance of each individual item being different. *For example*, items like pencil and milk are assigned equal importance in the construction of this index. This limitation renders the index of no practical utility.

Second, different units in which the prices are quoted also sometimes unduly affect this index. Prices quoted in higher weights, such as price of wheat per bushel as compared to a price per kg, will have unduly large influence on this index. Consequently, the prices of only a few commodities may dominate the index. This problem no longer exists when the units in which the prices of various commodities are quoted have a common base.

Even the condition of common base will provide no real solution because commodities with relatively high prices such as gold, which is not as important as milk, will continue to dominate this index excessively. *For example*, in the Example 15-2 given above chicken prices are relatively higher than those of fish, and hence chicken prices tend to influence this index relatively more than the prices of fish.

15.2.3.2.2 INDEX OF AVERAGE OF PRICE/QUANTITY RELATIVES

This index makes an improvement over the index of simple aggregative prices/quantities as it is not affected by the difference in the units of measurement in which prices/quantities are expressed. However, this also suffers from the problem of equal importance getting assigned to all the commodities.

Given the prices/quantities of a number of commodities that enter into the construction of this index, it is computed in the following two steps:

(i) After selecting the base year, find the price relative/quantity relative of each commodity for each year with respect to the base year price/quantity. As defined earlier, the price relative/quantity relative of a commodity for a given period is the ratio of the price/quantity of that commodity in the given period to its price/quantity in the base period.

(ii) Multiply the result for each commodity by 100, to get simple price/quantity indices for each commodity.

(iii) Take the average of the simple price/quantity indices by using arithmetic mean, geometric mean or median.

Thus it is computed as:

$$P_{0i} = \text{Average of}\left(\frac{p_i}{p_0} x100\right)$$

and $$Q_{0i} = \text{Average of}\left(\frac{q_i}{q_0} x100\right)$$

Using arithmetic mean

$$P_{0i} = \frac{\sum\left(\frac{p_i}{p_0} x100\right)}{N} \qquad \ldots\ldots\ldots\ldots(15\text{-}7)$$

and $$Q_{0i} = \frac{\sum\left(\frac{q_i}{q_0} x100\right)}{N} \qquad \ldots\ldots\ldots\ldots(15\text{-}8)$$

Using geometric mean

$$P_{0i} = Anti\log\left[\frac{1}{N}\sum\log\left(\frac{p_i}{p_0} x100\right)\right] \qquad \ldots\ldots\ldots\ldots(15\text{-}9)$$

and $$Q_{0i} = Anti\log\left[\frac{1}{N}\sum\log\left(\frac{q_i}{q_0} x100\right)\right] \qquad \ldots\ldots\ldots\ldots(15\text{-}10)$$

Example 15-3

From the data in Example 15.2 find:

(a) Index of Average of Price Relatives (base year 1980); using mean, median and geometric mean.

(b) Index of Average of Quantity Relatives (base year 1980); using mean, median and geometric mean.

Solution:

Calculations for

Index of Average of Price Relatives and Quantity Relatives

(Base Year = 1980)

Item	Price Relative $=\left(\frac{p_i}{p_0}x100\right)$	$\log\left(\frac{p_i}{p_0}x100\right)$	Quantity Relative $=\left(\frac{q_i}{q_0}x100\right)$	$\log\left(\frac{q_i}{q_0}x100\right)$
Fish	133.33	2.1248	120.00	2.0792
Mutton	127.77	2.10153	108.47	2.0354
Chicken	109.09	2.0378	111.11	2.0457
Sum $\rightarrow$	370.19	15.21589	339.58	15.11503

(a) Index of Average of Price Relatives (base year 1980)

Using arithmetic mean $$P_{0i}=\frac{\sum\left(\frac{p_i}{p_0}x100\right)}{N}$$

$$=\frac{370.19}{3}$$

$$=123.39$$

Using Median $$P_{0i}=\text{Size of}\left(\frac{\text{N}+1}{2}\right)th\text{ item}$$

$$=\text{Size of}\left(\frac{3+1}{2}\right)th\text{ item}$$

$$=\text{Size of } 2nd\text{ item}$$

$$=127.77$$

Using geometric mean $$P_{0i} = Anti\log\left[\frac{1}{N}\sum\log\left(\frac{p_i}{p_0}x100\right)\right]$$

$$= Anti\log\left[\frac{1}{3}(6.2689)\right]$$

$$= Anti\log[2.08963]$$

$$= 122.92$$

(b) Index of Average of Quantity Relatives (base year 1980)

Using arithmetic mean $$Q_{0i} = \frac{\sum\left(\frac{q_i}{q_0}x100\right)}{N}$$

$$= \frac{339.58}{3}$$

$$= 113.19$$

Using Median $$Q_{0i} = \text{Size of}\left(\frac{N+1}{2}\right)th \text{ item}$$

$$= \text{Size of}\left(\frac{3+1}{2}\right)th \text{ item}$$

$$= \text{Size of } 2nd \text{ item}$$

$$= 111.11$$

Using geometric mean $$Q_{0i} = Anti\log\left[\frac{1}{N}\sum\log\left(\frac{q_i}{q_0}x100\right)\right]$$

$$= Anti\log\left[\frac{1}{3}(6.1603)\right]$$

$$= Anti\log[2.05343]$$

$$= 113.09$$

Apart from the inherent drawback that this index accords equal importance to all items entering into its construction, a simple arithmetic mean and median are not appropriate average to be applied to ratios. Because it is generally believed that a simple average injects an upward bias in the index. So geometric mean is considered a more appropriate average for ratios and percentages.

15.2.3.2.3 WEIGHTED AGGREGATIVE PRICE/QUANTITY INDICES

We have noted that the simple aggregative price/quantity indices do not take care of the differences in the weights to be assigned to different commodities that enter into their construction. It is primarily because of this limitation that the simple aggregative indices are of very limited use. Weighted aggregative Indices make up this deficiency by assigning proper weights to individual items.

Among several ways of assigning weights, two widely used ways are:

(i) to use base period quantities/prices as weights, popularly known as ***Laspeyre's Index***, and

(ii) to use the given (current) period quantities/prices as weights, popularly known as ***Paasche's Index***.

Laspeyre's Index

Laspeyre's Price Index, using base period quantities as weights is obtained as

$$P_{0i}^{La} = \frac{\sum p_i q_0}{\sum p_0 q_0} x100 \qquad \ldots\ldots\ldots(15\text{-}11)$$

Laspeyre's Quantity Index, using base period prices as weights is obtained as

$$Q_{0i}^{La} = \frac{\sum q_i p_0}{\sum q_0 p_0} x100 \qquad \ldots\ldots\ldots(15\text{-}12)$$

Paasche's Index

Paasche's Price Index, using base period quantities as weights is obtained as

$$P_{0i}^{Pa} = \frac{\sum p_i q_i}{\sum p_0 q_i} x100 \qquad \ldots\ldots\ldots(15\text{-}13)$$

Paasche's Quantity Index, using base period prices as weights is obtained as

$$Q_{0i}^{Pa} = \frac{\sum q_i p_i}{\sum q_0 p_i} x100 \qquad \ldots\ldots\ldots(15\text{-}14)$$

Example 15-4

From the data in Example 15.2 find:

(a) Laspeyre's Price Index for 1985, using 1980 as the base

(b) Laspeyre's Quantity Index for 1985, using 1980 as the base

(c) Paasche's Price Index for 1985, using 1980 as the base

(d) Paasche's Quantity Index for 1985, using 1980 as the base

Solution:

Calculations for

Laspeyre's and Paasche's Indices

(Base Year = 1980)

Item	$p_0 q_0$	$p_1 q_0$	$p_0 q_1$	$p_1 q_1$
Fish	7500	10000	9000	12000
Mutton	101520	13570	11520	14720
Chicken	9900	10800	11000	12000
Sum $\rightarrow$	28020	34370	31520	38720

(a) Laspeyre's Price Index for 1985, using 1980 as the base

$$P_{0i}^{La} = \frac{\sum p_i q_0}{\sum p_0 q_0} x100$$

$$= \frac{34370}{28020} x100$$

$$= 122.66$$

(b) Laspeyre's Quantity Index for 1985, using 1980 as the base

$$Q_{0i}^{La} = \frac{\sum q_i p_0}{\sum q_0 p_0} x100$$

$$= \frac{31520}{28020} x100$$

$$= 112.49$$

(c) Paasche's Price Index for 1985, using 1980 as the base

$$P_{0i}^{Pa} = \frac{\sum p_i q_i}{\sum p_0 q_i} x100$$

$$= \frac{38720}{31520} x100$$

$$= 122.84$$

(d) Paasche's Quantity Index for 1985, using 1980 as the base

$$Q_{0i}^{Pa} = \frac{\sum q_i p_i}{\sum q_0 p_i} x100$$

$$= \frac{38720}{34370} x100$$

$$= 112.66$$

Interpretations of Laspeyre's Index

On close examination it will be clear that the Laspeyre's Price Index offers the following precise interpretations:

1. It compares the cost of collection of a fixed basket of goods selected in the base period with the cost of collecting the same basket of goods in the given (current) period.

Accordingly, the cost of collection of 500 qtls of fish, 590 qtls of mutton and 450 qtls of chicken has increased by 22.1515 per cent in 1985 as compared to what it was in 1980. Viewed differently, it indicates that a fixed amount of goods sold at 1985 prices yield 22.1515 per cent more revenue than what it did at 1980 prices.

2. It also implies that a fixed amount of goods when purchased at 1985 prices would cost 22.1515 per cent more than what it did at 1980 prices. In this interpretation, the Laspeyre's Price Index serves as the basis of constructing the cost of living index, for it tells how much more does it cost to maintain the base period standard of living at the current period prices.

Laspeyre's Quantity Index, too, has precise interpretations. It reveals the percentage change in total expenditure in the given (current) period as compared to the base period if varying amounts of the same basket of goods are sold at the base period prices. When viewed in this manner, we will be required to spend 12.49 per cent more in 1985 as compared to 1980 if the quantities of fish, mutton and chicken for 19155 are sold at the base period (1980) prices.

Interpretations of Paasche's Index

A careful examination of the Paasche's Price Index will show that this too is amenable to the following precise interpretations:

1. It compares the cost of collection of a fixed basket of goods selected in the given period with the cost of collection of the same basket of goods in the base period.

Accordingly, the cost of collection of a fixed basket of goods containing 1500 qtls of fish, 1540 qtls of mutton and 500 qtls of chicken in 1985 is about 22.84 per cent more than the cost of collecting the same basket of goods in 1980. Viewed a little differently, it indicates that a fixed basket of goods sold at 1985 prices yields 22.84 per cent more revenue than what it would have earned had it been sold at the base period (1980) prices.

2. It also tells that a fixed amount of goods purchased at 1985 prices will cost 22.84 per cent more than what it would have cost if this fixed amount of goods had been sold at base period (1980) prices.

Analogously, Paasche's Quantity Index, too, has its own precise meaning. It tells the per cent change in total expenditure in the given period as compared to the base period if varying amounts of the same basket of goods are to be sold at given period prices. When so viewed, we will be required to spend 12.1515 per cent more in 1985 as compared to 1980 if the quantities of fish, mutton and chicken for 1980 are sold at the given period (1985) prices.

Relationship Between Laspeyre's and Paasche's Indices

In order to understand the relationship between Laspeyre's and Paasche's Indices, the assumptions on which the two indices are based be borne in mind:

Laspeyre's index is based on the assumption that *unless there is a change in tastes and preferences, people continue to buy a fixed basket of goods irrespective of how high or low the prices are likely to be in the future.* Paasche's index, on the other hand, assumes that *people would have bought the same amount of a given basket of goods in the past irrespective of how high or low were the past prices.*

However, the basic contention implied in the assumptions on which the two indices are based is not true. For, people do make shifts in their purchase pattern and preferences by buying more of goods that tend to become cheaper and less of those that tend to become costlier. In view of this, the following two situations that are likely to emerge need consideration:

1. When the prices of goods that enter into the construction of these indices show a general tendency to rise, those whose prices increase more than the average increase in prices will have smaller quantities in the given period than the corresponding quantities in the base period. That is, q_i*'s* will be smaller than q_0*'s* when prices in general are rising. Consequently, Paasche's index will have relatively smaller weights than those in the Laspeyre's index and, therefore, the former (P_{0i}^{Pa}) will

be smaller than the latter (P_{0i}^{La}). In other words, Paasche's index will show a relatively smaller increase when the prices in general tend to rise.

2. On the contrary, when prices in general are falling, goods whose prices show a relatively smaller fall than the average fall in prices, will have smaller quantities in the given period than the corresponding quantities in the base period. This means that q_i*'s* will be smaller than q_0*'s* when prices in general are falling. Consequently, Paasche's index will have smaller weights than those in the Laspeyre's index and, therefore, the former (P_{0i}^{Pa}) will be smaller than the latter (P_{0i}^{La}). In other words, Paasche's index will show a relatively greater fall when the prices in general tend to fall.

An important inference based on the above discussion is that ***the Paasche's index has a downward bias and the Laspeyre's index an upward bias.*** This directly follows from the fact that the Paasche's index, relative to the Laspeyre's index, shows a smaller rise when the prices in general are rising, and a greater fall when the prices in general are falling.

It may, however, be noted that when the quantity demanded increases because of change in real income, tastes and preferences, advertising, *etc.*, the prices remaining unchanged, the Paasche's index will show a higher value than the Laspeyre's index. In such situations, the Paasche's index will overstate, and the Laspeyre's will understate, the changes in prices. The former now represents the upper limit, and the latter the lower limit, of the range of price changes.

The relationship between the two indices can be derived more precisely by making use of the coefficient of linear correlation computed as:

$$r_{xy} = \frac{\frac{\sum fXY}{N} - \left(\frac{\sum fX}{N}\right)\left(\frac{\sum fY}{N}\right)}{S_x S_y} \qquad \ldots\ldots\ldots\ldots (15.15)$$

in which X and Y denote the relative price movements($\frac{p_i}{p_0}$) and relative quantity movements($\frac{q_i}{q_0}$) respectively. S_x and S_y are the standard deviations of price and quantity movements, respectively. While r_{xy} represents the coefficient of correlation between the relative price and quantity movements; f represents the weights assigned, that is, $p_0 q_0$. N is the sum of frequencies *i. e.* $N = \sum p_0 q_0$.

Substituting the values of X, Y, f and N in *Eq. (15-15),* and then rearranging the expression, we have

$$r_{xy}S_xS_y = \frac{\sum p_iq_i}{\sum p_0q_0} - \frac{\sum p_iq_0}{\sum p_0q_0} x \frac{\sum p_0q_i}{\sum p_0q_0}$$

If $\frac{\sum p_iq_i}{\sum p_0q_0} = V_{0i}$, is the index of value expanded between the base period and the ith period, then dividing both sides by $\frac{\sum p_iq_i}{\sum p_0q_0}$ or V_{0i}, we get

$$\frac{r_{xy}S_xS_y}{V_{0i}} = 1 - \frac{\sum p_iq_0}{\sum p_0q_0} x \frac{\sum p_0q_i}{\sum p_iq_i}$$

$$\frac{r_{xy}S_xS_y}{V_{0i}} = 1 - P_{0i}^{La} x \frac{1}{P_{0i}^{Pa}}$$

$$\frac{P_{0i}^{La}}{P_{0i}^{Pa}} = 1 - \frac{r_{xy}S_xS_y}{V_{0i}} \qquad \text{……..(15.115)}$$

The relationship in *Eq. (15.115)* offers the following useful results:

1. $P_{0i}^{La} = P_{0i}^{Pa}$ when either r_{xy} , S_x and S_y is equal to zero. That is, the two indices will give the same result either when there is no correlation between the price and quantity movements, or when the price or quantity movements are in the same ratio so that S_x or S_y is equal to zero.
2. Since in actual practice r_{xy} will have a negative value between 0 and -1, and as neither $S_x = 0$ nor $S_y = 0$, the right hand side of *Eq. (15-115)* will be less than 1. This means that P_{0i}^{La} is normally greater than P_{0i}^{Pa}.
3. Given the overall movement in the index of value (V_{0i}) expanded, the greater the coefficient of correlation (r_{xy}) between price and quantity movements and/or the greater the degree of dispersion (S_x and S_y) in the price and quantity movements, the greater the discrepancy between P_{0i}^{La} and P_{0i}^{Pa}.
4. The longer the time interval between the two periods to be compared, the more the chances for price and quantity movements leading to higher values of S_x and S_y. The assumption of tastes, habits, and preferences remaining unchanged breaking down over a longer period, people do find enough time to make shifts in their consumption pattern, buying more of goods that may have become relatively cheaper and less of those that may have become relatively dearer. All this will end up with a higher degree of correlation between the price and quantity movement.

Consequently, P_{0i}^{La} will diverge from P_{0i}^{Pa} more in the long run than in the short run. So long as the periods to be compared are not much apart, P_{0i}^{La} will be quite close to P_{0i}^{Pa}.

Laspeyre's and Paasche's Indices Further Considered

The use of different system of weights in these two indices may give an impression as if they are opposite to each other. Such an impression is not sound because both serve the same purpose, although they may give different results when applied to the same data.

This raises an important question. Which one of them gives more accurate results and which one should be preferred over the other? The answer to this question is rather difficult since both the indices are amenable to precise and useful results.

Despite a very useful and precise difference in interpretation, in actual practice the Laspeyre's index is used more frequently than the Paasche's index for the simple reason that the latter requires frequent revision to take into account the yearly changes in weights. No such revision is required in the case of the Laspeyre's index where once the weights have been determined, these do not require any change in any subsequent period. It is on this count that the Laspeyre's index is preferred over the Paasche's index.

However, this does not render the Paasche's index altogether useless. In fact, it supplements the practical utility of the Laspeyre's index. The fact that the Laspeyre's index has an upward bias and the Paasche's index downward bias, the two provide the range between which the index can vary between the base period and the given period. Interestingly, thus, the former represents the upper limit, and the latter the lower limit.

Improvements over the Laspeyre's and Paasche's Indices

To overcome the difficulty of overstatement of changes in prices by the Laspeyre's index and understatement by the Paasche's index, different indices have been developed to compromise and improve upon them. These are particularly useful when the given period and the base period fall quite apart and result in a greater divergence between Laspeyre's and Paasche's indices.

Other important Weighted Aggregative Indices are:

1. **Marshall-Edgeworth Index**

The Marshall-Edgeworth Index uses the average of the base period and given period quantities/prices as the weights, and is expressed as

$$P_{0i}^{ME} = \frac{\sum p_i\left(\frac{q_0 + q_i}{2}\right)}{\sum p_0\left(\frac{q_0 + q_i}{2}\right)} x100 \qquad \ldots\ldots\ldots(15\text{-}17)$$

$$Q_{0i}^{ME} = \frac{\sum q_i\left(\frac{p_0 + p_i}{2}\right)}{\sum q_0\left(\frac{p_0 + p_i}{2}\right)} x100 \qquad \ldots\ldots\ldots(15\text{-}18)$$

2. Dorbish and Bowley Index

The Dorbish and Bowley Index is defined as the arithmetic mean of the Laspeyre's and Paasche's indices.

$$P_{0i}^{DB} = \frac{P_{0i}^{La} + P_{0i}^{Pa}}{2} \qquad \ldots\ldots\ldots(15\text{-}19)$$

$$Q_{0i}^{DB} = \frac{Q_{0i}^{La} + Q_{0i}^{Pa}}{2} \qquad \ldots\ldots\ldots(15\text{-}20)$$

3. Fisher's Ideal Index

The Fisher's Ideal Index is defined as the geometric mean of the Laspeyre's and Paasche's indices.

$$P_{0i}^{F} = \sqrt{P_{0i}^{La}.P_{0i}^{Pa}} \qquad \ldots\ldots\ldots(15\text{-}21)$$

$$Q_{0i}^{F} = \sqrt{Q_{0i}^{La}.Q_{0i}^{Pa}} \qquad \ldots\ldots\ldots(15\text{-}22)$$

15.2.3.2.4 Index of Weighted Average of Price/Quantity Relatives

An alternative system of assigning weights lies in using value weights. The value weight *v* for any single commodity is the product of its price and quantity, that is, $v = pq$.

If the index of weighted average of price relatives is defined as

$$P_{0i} = \frac{\sum\left[v\left(\frac{p_i}{p_0} x100\right)\right]}{\sum v} \qquad \ldots\ldots\ldots(15\text{-}23)$$

then v can be obtained either as

(i) the product of the base period prices and the base period quantities denoted as v_0 that is, $v_0 = p_0 q_0$, or

(ii) the product of the base period prices and the given period quantities denoted as v_i that is, $v_i = p_0 q_i$

When v is $v_0 = p_0 q_0$, the index of weighted average of price relatives, is expressed as

$${}_0P_{0i} = \frac{\sum\left[p_0 q_0\left(\frac{p_i}{p_0} x100\right)\right]}{\sum p_0 q_0} \qquad \ldots\ldots\ldots(15\text{-}24)$$

It may be seen that ${}_0P_{0i}$ is the same as the Laspeyre's aggregative price index.

Similarly, When v is $v_i = p_0 q_i$, the index of weighted average of price relatives, is expressed as

$${}_iP_{0i} = \frac{\sum\left[p_0 q_i\left(\frac{p_i}{p_0} x100\right)\right]}{\sum p_0 q_i} \qquad \ldots\ldots\ldots(15\text{-}25)$$

It may be seen that ${}_iP_{0i}$ is the same as the Paasche's aggregative price index.

If the index of weighted average of quantity relatives is defined as

$$Q_{0i} = \frac{\sum\left[v\left(\frac{q_i}{q_0} x100\right)\right]}{\sum v} \qquad \ldots\ldots\ldots(15\text{-}215)$$

then v can be obtained either as

(i) the product of the base period quantities and the base period prices denoted as v_0 that is, $v_0 = q_0 p_0$, or

(ii) the product of the base period quantities and the given period prices denoted as v_i that is, $v_i = q_0 p_i$

When v is $v_0 = q_0 p_0$, the index of weighted average of quantity relatives, is expressed as

$$ {}_0Q_{0i} = \frac{\sum\left[q_0 p_0 \left(\frac{q_i}{q_0} x100 \right) \right]}{\sum q_0 p_0} \qquad \ldots\ldots\ldots\ldots (15\text{-}27) $$

It may be seen that ${}_0Q_{0i}$ is the same as the Laspeyre's aggregative quantity index.

Similarly, When v is $v_i = v_i = q_0 p_i$, the index of weighted average of quantity relatives, is expressed as

$$ {}_iQ_{0i} = \frac{\sum\left[q_0 p_i \left(\frac{q_i}{q_0} x100 \right) \right]}{\sum q_0 p_i} \qquad \ldots\ldots\ldots\ldots (15\text{-}28) $$

It may be seen that ${}_iQ_{0i}$ is the same as the Paasche's aggregative quantity index.

Example 15-5

From the data in Example 15.2 find the:

(a) Index of Weighted Average of Price Relatives, using

(i) $v_0 = p_0 q_0$ as the value weights

(ii) $v_i = p_0 q_i$ as the value weights

(b) Index of Weighted Average of Quantity Relatives, using

(i) $v_0 = q_0 p_0$ as the value weights

(ii) $v_i = q_0 p_i$ as the value weights

Solution:

Calculations for

Index of Weighted Average of Price Relatives

(Base Year = 1980)

Item	$v_0 = p_0 q_0$	$v_1 = p_0 q_1$	$p_0 q_0 \left(\frac{p_i}{p_0} x100 \right)$	$p_0 q_1 \left(\frac{p_1}{p_0} x100 \right)$
Fish	7500	9000	1000000	1200000

Mutton	101520	11520	1357000	1472000
Chicken	9900	11000	1080000	1200000
Sum →	28020	31520	3437000	3872000

(a) Index of Weighted Average of Price Relatives, using

(i) $v_0 = p_0 q_0$ as the value weights

$$ {}_0P_{0i} = \frac{\sum\left[p_0 q_0 \left(\frac{p_i}{p_0} x100 \right) \right]}{\sum p_0 q_0} $$

$$ = \frac{3437000}{28020} $$

$$ = 122.66 $$

(ii) $v_i = p_0 q_i$ as the value weights

$$ {}_iP_{0i} = \frac{\sum\left[p_0 q_i \left(\frac{p_i}{p_0} x100 \right) \right]}{\sum p_0 q_i} $$

$$ = \frac{3872000}{31520} $$

$$ = 122.84 $$

Calculations for

Index of Weighted Average of Quantity Relatives

(Base Year = 1980)

Item	$v_0 = q_0 p_0$	$v_1 = q_0 p_1$	$q_0 p_0 \left(\frac{q_1}{q_0} x100 \right)$	$q_0 p_1 \left(\frac{q_1}{q_0} x100 \right)$
Fish	7500	10000	900000	1200000
Mutton	101520	13570	1152000	1472000
Chicken	9900	10800	1100000	1200000

Sum →	28020	34370	3152000	3872000

(b) Index of Weighted Average of Quantity Relatives, using

(i) $v_0 = q_0 p_0$ as the value weights

$$ {}_0Q_{0i} = \frac{\sum\left[q_0 p_0\left(\frac{q_i}{q_0} x100\right)\right]}{\sum q_0 p_0} $$

$$ = \frac{3152000}{28020} $$

$$ = 112.49 $$

(ii) $v_i = q_0 p_i$ as the value weights

$$ {}_iQ_{0i} = \frac{\sum\left[q_0 p_i\left(\frac{q_i}{q_0} x100\right)\right]}{\sum q_0 p_i} $$

$$ = \frac{3872000}{34370} $$

$$ = 112.66 $$

Although the indices of weighted average of price/quantity relatives yield the same results as the Laspeyre's or Paasche's price/quantity indices, we do construct these indices also in situations when it is necessary and advantageous to do so. Some such situations are as follows:

(i) When a group of commodities is to be represented by a single commodity in the group, the price relative of the latter is weighted by the group as a whole.

(ii) Where the price/quantity relatives of individual commodities have been computed, these can be more conveniently utilised in constructing the index.

(iii) Price/quantity relatives serve a useful purpose in splicing two index series having different base periods.

(iv) Depersonalizing a time series requires construction of a seasonal index, which also requires the use of relatives.

15.3 TESTS OF ADEQUACY OF INDEX NUMBERS

We have discussed various formulae for the construction of index numbers. None of the formulae measures the price changes or quantity changes with perfection and has some bias. The problem is to choose the most appropriate formula in a given situation. As a measure of the formula error a number of mathematical tests, known as the *tests of consistency* or *tests of adequacy* of index number formulae have been suggested. In this section we will discuss these tests, which are also sometimes termed as the criteria for a good index number.

1. **Unit Test:** This test requires that the index number formula should be independent of the units in which the prices or quantities of various commodities are quoted. All the formulae discussed in the lesson except the index number based on Simple Aggregate of Prices/Quantities satisfy this test.

2. **Time Reversal Test:** The time reversal test, proposed by Prof Irving Fisher requires the index number formula to possess time consistency by working both forward and backward *w.r.t.* time. In his (Fisher's) words:

   "The formula for calculating an index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as the base or putting it another way, the index number reckoned forward should be reciprocal of the one reckoned backward."

   In other words, if the index numbers are computed for the same data relating to two periods by the same formula but with the bases reversed, then the two index numbers so obtained should be the reciprocals of each other. Mathematically, we should have (omitting the factor 100),

   $$P_{ab} x P_{ba} = 1 \qquad \ldots\ldots\ldots\ldots (15\text{-}29)$$

   or more generally

   $$P_{01} x P_{10} = 1 \qquad \ldots\ldots\ldots\ldots (15\text{-}29a)$$

   Time reversal test is satisfied by the following index number formulae:

   (i) Marshall-Edgeworth formula

   (ii) Fisher's Ideal formula

   (iii) Kelly's fixed weight formula

   (iv) Simple Aggregate index

   (v) Simple Geometric Mean of Price Relatives formula

   (vi) Weighted Geometric Mean of Price Relatives formula with fixed weights

   Lespeyre's and Pasche's index numbers do not satisfy the time reversal test.

3. **Factor Reversal Test:** This is the second of the two important tests of consistency proposed by Prof Irving Fisher. According to him:

   "Just as our formula should permit the interchange of two times without giving inconsistent results, so it ought to permit interchanging the price and quantities without giving inconsistent results – i.e., the two results multiplied together should give the true value ratio, except for a constant of proportionality."

   This implies that if the price and quantity indices are obtained for the same data, same base and current periods and using the same formula, then their product (without the factor 100) should give the true value ratio. Symbolically, we should have (without factor 100).

$$P_{01}xQ_{01} = \frac{\sum p_1q_1}{\sum p_0q_0} = V_{01} \qquad \ldots\ldots\ldots\ldots(15\text{-}30)$$

   Fisher's formula satisfies the factor reversal test. In fact fisher's index is the only index satisfying this test as none of the formulae discussed in the lesson satisfies this test.

   Remark: Since Fisher's index is the only index that satisfies both the time reversal and factor reversal tests, it is termed as Fisher's Ideal Index.

4. **Circular Test:** Circular test, first suggested by Westergaard, is an extension of time reversal test for more than two periods and is based on the shift ability of the base period. This requires the index to work in a circular manner and this property enables us to find the index numbers from period to period without referring back to the original base each time. For three periods *a,b,c,* the test requires :

$$P_{ab}xP_{bc}xP_{ca} = 1 \qquad a \neq b \neq c \qquad \ldots\ldots\ldots\ldots(15\text{-}31)$$

   In the usual notations *Eq. (15-31)* can be stated as:

$$P_{01}xP_{12}xP_{20} = 1 \qquad \ldots\ldots\ldots\ldots(15\text{-}31a)$$

   For Instance

$$P_{01}^{La}\ xP_{12}^{La}xP_{21}^{La} = \frac{\sum p_1q_0}{\sum p_0q_0}\,x\,\frac{\sum p_2q_1}{\sum p_1q_1}\,x\,\frac{\sum p_0q_2}{\sum p_2q_2} \neq 1$$

   Hence Laspeyre's index does not satisfy the circular test. In fact, circular test is not satisfied by any of the weighted aggregative formulae with changing weights. This test is satisfied only by the index number formulae based on:

(i) Simple geometric mean of the price relatives, and

(ii) Kelly's fixed base method

15.4 SPECIAL ISSUSES AND PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS

SPECIAL ISSUES

BASE SHIFTING

The need for shifting the base may arise either

(i) when the base period of a given index number series is to be made more recent,

(ii) when two index number series with different base periods are to be compared,

(iii) when there is need for splicing two overlapping index number series.

Whatever be the reason, the technique of shifting the base is simple:

$$\text{New Base Index Number} = \frac{\text{Old Index Number of Current Year}}{\text{Old Index Number of New Base Year}} x100$$

Example 15-15

Reconstruct the following indices using 1997 as base:

Year :	1991	1992	1993	1994	1995	19915	1997	1998
Index :	100	110	130	150	175	180	200	220

Solution:

Shifting the Base Period

Year	Index Number (1991 = 100)	Index Number (1997 = 100)
1991	100	(100/200) *x*100 = 50.00
1992	110	(110/200) *x*100 = 55.00
1993	130	(130/200) *x*100 = 155.00
1994	150	(150/200) *x*100 = 75.00
1995	175	(175/200) *x*100 = 87.50
19915	180	(180/200) *x*100 = 90.00
1997	200	(200/200) *x*100 = 100.00

1998	220	(220/200) *x*100 = 110.00

SPLICING TWO OVERLAPPING INDEX NUMBER SERIES

Splicing two index number series means reducing two overlapping index series with different base periods into a single series either at the base period of the old series (one with an old base year), or at the base period of the new series (one with a recent base year). This actually amounts to changing the weights of one series into the weights of the other series.

1. Splicing the New Series to Make it Continuous with the Old Series

Here we reduce the new series into the old series after the base year of the former. As shown in Table 15.8.2*(i)*, splicing here takes place at the base year (1980) of the new series. To do this, a ratio of the index for 1980 in the old series (200) to the index of 1980 in the new series (100) is computed and the index for each of the following years in the new series is multiplied by this ratio.

Table 15.8.2*(i)*

Splicing the New Series with the Old Series

Year	Price Index (19715 = 100) (Old Series)	Price Index (1980 = 100) (New Series)	Spliced Index Number [New Series *x* (200/100)]
1971	100	--	100
1977	120	--	120
1978	1415	--	1415
1979	172	--	172
1980	200	100	200
1981	--	110	220
1982	--	1115	232
1983	--	125	250
1984	--	140	280

2. Splicing the Old Series to Make it Continuous with the New Series

This means reducing the old series into the new series before the base period of the letter. As shown in Table 15.8.2*(ii)*, splicing here takes place at the base period of the new series. To do this, a ratio

of the index of 1980 of the new series (100) to the index of 1980 of the old series (200) is computed and the index for each of the preceding years of the old series are then multiplied by this ratio.

Table 15.8.2*(ii)*

Splicing the Old Series with the New Series

Year	Price Index (19715 = 100) (Old Series)	Price Index (1980 = 100) (New Series)	Spliced Index Number [Old Series *x* (100/200)]
1971	100	--	50
1977	120	--	150
1978	1415	--	73.50
1979	172	--	815
1980	200	100	100
1981	--	110	110
1982	--	1115	1115
1983	--	125	125
1984	--	140	140

CHAIN BASE INDEX NUMBERS

The various indices discussed so far are fixed base indices in the sense that either the base year quantities/prices (or the given year quantities/prices) are used as weights. In a dynamic situation where tastes, preferences, and habits are constantly changing, the weights should be revised on a continuous basis so that new commodities are included and the old ones deleted from consideration.

This is all the more necessary in a developing society where new substitutes keep replacing the old ones, and completely new commodities are entering the market. To take care of such changes, the base year should be the most recent, that is, the year immediately preceding the given year. This means that as we move forward, the base year should move along the given year in a chain year after year.

Conversion of Fixed-base Index into Chain-base Index

As shown in Table 15.8.3*(i)*, to convert fixed-base index numbers into chain-base index numbers, the following procedure is adopted:

- The first year's index number is taken equal to 100

- For subsequent years, the index number is obtained by following formula:

$$\text{Current Year's CBI} = \frac{\text{Current Year's FBI}}{\text{Previous Year's CBI}} x 100$$

Table 15.8.3*(i)*

Conversion of Fixed-base Index into Chain-base Index

Year	Fixed Base Index Number (FBI)	Conversion	Chain Base Index Number (CBI)
1975	3715	--	100
19715	392	(392/3715) *x*100	104.3
1977	408	(408/392) *x*100	104.1
1978	380	(380/408) *x*100	93.1
1979	392	(392/380) *x*100	103.2
1980	400	(400/392) *x*100	102

Conversion of Chain-base Index into Fixed-base Index

As shown in Table 15.8.3*(ii)*, to convert fixed-base index numbers into chain-base index numbers, the following procedure is adopted:

- The first year's index is taken what the chain base index is; but if it is to form the base it is taken equal to 100
- In subsequent years, the index number is obtained by following formula:

$$\text{Current Year's FBI} = \frac{\text{Current Year's CBI } x \text{ Previous Year's FBI}}{100}$$

Table 15.8.3*(ii)*

Conversion of Chain-base Index into Fixed-base Index

Year	Chain Base Index Number (CBI)	Conversion	Fixed Base Index Number (FBI)
1978	90	--	90
1979	120	(120 *x* 90) /100	108

1980	125	(125 *x* 108) /100	135
1981	110	(110 *x* 135) /100	148.5
1982	112	(112 *x* 148.5) /100	11515.3
1983	150	(150 *x* 11515.3) /100	249.45

PROBLEMS OF CONSTRUCTING INDEX NUMBERS

The above discussion enables us to identify some of the important problems, which may be faced in the construction of index numbers:

Choice of the Base Period: Choice of the base period is a critical decision because of its importance in the construction of index numbers. A base period is the reference period for describing and comparing the changes in prices or quantities in a given period. The selection of a base year or period does not pose difficult theoretical questions. To a large extent, the choice of the base year depends on the objective of the index. A major consideration should be to ensure that the base year is not an abnormal year. *For example,* a base period with very low price/quantity will unduly inflate, while the one with a very high figure will unduly depress, the entire index number series. An index number series constructed with any such period as the base may give very misleading results. It is, therefore, necessary that the base period be selected carefully.

Another important consideration is that the base year should not be too remote in the past. A more recent year needs to be selected as the base year. The use of a particular year for a prolonged period would distort the changes that it purports to measure. That is why we find that the base year of major index numbers, such as consumer price index or index of industrial production, is shifted from time to time.

Selection of Weights to be Used: It should be amply clear from the various indices discussed in the lesson that the choice of the system of weights, which may be used, is fairly large. Since any system of weights has its own merits and is capable of giving results amenable to precise interpretations, the weights used should be decided keeping in view the purpose for which an index is constructed.

It is also worthwhile to bear in, mind that the use of any system of weights should represent the relative importance of individual commodities that enter into the construction of an index. The interpretations that are intended to be made from an index number are also important in deciding the weights. The use of a system of weights that involves heavy computational work deserves to be avoided.

Type of Average to be Used: What type of average should be used is a problem specific to simple average indices. Theoretically, one can use any of the several averages that we have, such as mean, median, mode, harmonic mean, and geometric mean. Besides being locational averages, median and mode are not the appropriate averages to use especially where the number of years for which an index is to be computed, is not large.

While the use of harmonic mean and geometric mean has some definite merits over mean, particularly when the data to be averaged refer to ratios, mean is generally more frequently used for convenience in computations.

Choice of Index: The problem of selection of an appropriate index arises because of availability of different types of indices giving different results when applied to the same data. Out of the various indices discussed, the choice should be in favour of one which is capable of giving more accurate and precise results, and which provides answer to specific questions for which an index is constructed.

While the Fisher's index may be considered ideal for its ability to satisfy the tests of adequacy, this too suffers from two important drawbacks. First, it involves too lengthy computations, and second, it is not amenable to easy interpretations as are the Laspeyre's and Paasche's indices. The use of the term ideal does not, however, mean that it is the best to use under all types of situations. Other indices are more appropriate under situations where specific answers are needed.

Selection of Commodities: Commodities to be included in the construction of an index should be carefully selected. Only those commodities deserve to be included in the construction of an index as would make it more representative. This, in fact, is a problem of sampling, for being related to the selection of commodities to be included in the sample.

In this context, it is important to note that the selection of commodities must not be based on random sampling. The reason being that in random sampling every commodity, including those that are not important and relevant, have equal chance of being selected, and consequently, the index may not be representative. The choice of commodities has, therefore, to be deliberate and in keeping with the relevance and importance of each individual commodity to the purpose for which the index is constructed.

Data Collection: Collection of data through a sample is the most important issue in the construction of index numbers. The data collected are the raw material of an index. Data quality is the basic factor that

determines the usefulness of an index. The data have to be as accurate, reliable, comparable, representative, and adequate, as possible.

The practical utility of an index also depends on how readily it can be constructed. Therefore, data should be collected from where these can be easily available. While the purpose of an index number will indicate what type of data are to be collected, it also determines the source from where the data can be available.

15.5 CHECK YOUR PROGRESS

1. Index Numbers measure changes not capable of …… measurement.
2. The wholesale price index numbers reflect the changes in the …… level of a country.
3. Weighted aggregative Indices make up this …… by assigning proper weights to individual items.
4. The Paasche's index has a …… bias and the Laspeyre's index an upward bias.
5. Splicing two index number series means reducing two …….. index series with different base periods into a single series either at the base period of the old series (one with an old base year), or at the base period of the new series (one with a recent base year).

15.6 SUMMARY

Index numbers are statistical devices designed to measure the relative changes in the level of a certain phenomenon in two or more situation. It may refer to a single variable or a group of distinct but related variables. There are different uses of Index numbers like: act as an economic Barometers, help in Studying Trends and Tendencies, help in Formulating Decisions and Policies, measure the Purchasing Power of Money and used for deflation. There are following types of it like: price index numbers, quantity index numbers and value index numbers. Various indices can also be distinguished on the basis of the number of commodities that go into the construction of an index. Indices constructed for individual commodities or variable are termed as simple index numbers. Those constructed for a group of commodities or variables are known as aggregative (or composite) index numbers. There are various formulae for the construction of index numbers. None of the formulae measures the price changes or quantity changes with perfection and has some bias. The problem is to choose the most appropriate formula in a given situation. As a measure of the formula error a number of mathematical tests, known

as the tests of consistency or tests of adequacy. There are different test for it like: unit test, time reversal test, factor reversal test and circular test.

15.7 KEYWORDS

Price Index Numbers: The price index numbers measure the general changes in the prices.

Quantity Index Numbers: Quantity index numbers study the changes in the volume of goods produced (manufactured), consumed or distributed.

Value Index Numbers: These are intended to study the change in the total value (price multiplied by quantity) of output such as indices of retail sales or profits or inventories.

Simple index numbers: A simple price index number is based on the price or quantity of a single commodity.

Lasperyre's index: It use base period quantities/prices as weights

Paasche's Index: It uses the given (current) period quantities/prices as weights.

Splicing two index number series: It means reducing two overlapping index series with different base periods into a single series either at the base period of the old series (one with an old base year), or at the base period of the new series (one with a recent base year).

15.8 SELF-ASSESSMENT TEST

1. "Index Numbers are devices for measuring changes in the magnitude of a group of related variables". Discuss this statement and point out the important uses of index numbers.
2. "Index Numbers are Economic Barometers". Discuss this statement. What precautions would you take while constructing index numbers?
3. (a) Explain the uses of index numbers.
   (b) What problems are involved in the construction of index numbers?
4. Describe each of the following:
   a. Base period
   b. Price relatives
   c. Fixed-base index numbers
   d. Chain-base index numbers
5. Describe briefly the following methods of construction of price index numbers:

a. Simple Aggregate Method

b. Simple Average of Price Relatives Method

c. Weighted Aggregative Method

d. Weighted Average of Price Relatives

6. "Laspeyre's index has an upward bias and the Paasche's index downward bias". Explain this statement.
7. Discuss the various tests of adequacy of index numbers.
8. State and explain the Fisher's ideal formula for price index number. Show how it satisfies the time-reversal and factor- reversal test? Why is it used little in practice?
9. Briefly explain each of the following:

   a. Base-shifting

   b. Splicing

   c. Deflating

10. From the following data, construct the price index for each year with price of 1995 as base.

Year:	1995	19915	1997	1998	1999	2000
Price of Commodity:	40	50	45	55	155	70

11. From the following data, construct an index number for 2004 taking 2003 as base year:

Articles:	A	B	C	D	E
Prices (2003):	100	125	50	40	5
Prices (2004):	140	200	80	150	10

12. Find the index number for 1982 and 1983 taking 1981 as base year by the Simple Average of Price Relatives Method, using *(i)* Mean, *(ii)* Median, and *(iii)* Geometric Mean:

Commodities	1981 (Prices)	1982 (Prices)	1983 (Prices)
A	40	55	150
B	50	150	80
C	152	72	93
D	80	88	915
E	20	24	30

13. Construct index number of price and index number of quantity from the following data using:

a. Laspeyre's formula,

b. Paasche's formula,

c. Dorbish and Bowley's formula,

d. Marshall and Edgeworth's formula, and

e. Fisher's Ideal Index formula

Commodities	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	2	8	4	15
B	5	10	15	5
C	4	14	5	10
D	2	19	2	13

Which of the formula satisfy

(i) the time reversal test, and

(ii) the factor reversal test?

14. Calculate index number using Kelly's Method of Standard Weights, from the following data:

Commodities	Quantity	Base Year Price	Current Year Price
A	5	30	40
B	8	20	30
C	10	10	20

15. From the following data, construct price index by using Weighted Average of Price Relatives Method:

Commodities	Quantity	Base Year Price	Current Year Price
A	15 Qtl	5.00	15.00
B	5 Qtl	5.00	8.00
C	1 Qtl	15.00	9.00
D	4 Kg	8.00	10.00
E	1 Kg	20.00	15.00

16. From the information given below, calculate the Cost of Living Index number for 1985, with 1984 as base year by

    a. Aggregative Expenditure Method, and

    b. Family Budget Method.

Items	Quantity consumed	Unit	Prices in 1984	Prices in 1985
Wheat	2 Qtl	Qtl	75	125
Rice	20 Kg	Kg	12	115
Sugar	10 Kg	Kg	12	115
Ghee	5 Kg	Kg	10	15
Clothing	25 Meter	Meter	4.5	5
Fuel	40 Litre	Litre	10	12
Rent	One house	House	25	40

17. An enquiry into budgets of the middle class families in a city gave the following information:

Expenses on →	Food	Rent	Clothing	Fuel	Miscellaneous
	40%	10%	20%	10%	20%
Prices(2001)	1150	50	150	20	50
Prices(2002)	175	150	75	25	75

What changes in the cost of living figure of 2002 have taken place as compared to 2001?

18. Reconstruct the following indices using 1985 as base:

Year	:	1982	1983	1984	1985	19815	1987
Index	:	100	120	190	200	212	250

19. Given below are two sets of indices one with 1975 as base and the other with 1979 as base:

    <u>First set</u>

Year	:	1975	19715	1977	1978	1979
Index Numbers	:	100	110	125	180	200

    <u>Second Set</u>

Year	:	1979	1980	1981	1982	1983
Index Numbers	:	100	104	110	1115	124

    a. Splice the second set of index numbers from 1975

b. Splice the first set of index numbers from 1979

20. Construct chain index numbers from the following data:

Year :	1991	1992	1993	1994	1995
Price :	25	30	45	150	90

21. Convert into Chain Base Index Number from Fixed Base Index Number

Year	:	1980	1981	1982	1983	1984
Fixed Base Index	:	100	98	102	140	190

22. From the Chain Base Index numbers given below, construct Fixed Base Index numbers:

Year	:	1993	1994	1995	19915	1997
Chain Base Index	:	100	105	95	115	102

23. From the following data, prepare index number for real wages of workers:

Year	:	1990	1991	1992	1993	1994	1995
Wages (in Rs)	:	300	340	450	4150	475	540
Price Index Number	:	100	120	220	230	250	300

24. During certain period, the Cost of Living Index number went up from 110 to 200 and salary of a worker also raised from 325 to 500. State by how much the worker has gained or lost in real term.

15.9 ANSWERS TO CHECK YOUR PROGRESS

1. Direct
2. General Price
3. Deficiency
4. Downward
5. Overlapping

15.10 REFERNCES/SUGGESTED READINGS

1. Statistics (Theory & Practice) *by* Dr. B.N. Gupta. Sahitya Bhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management *by* G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.

3. Business Statistics *by* Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics *by* R.P. Hooda. MacMillan India Ltd., New Delhi.
5. Business Statistics *by* S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method *by* S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management *by* Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.
8. Statistics for Business and Economics *by* Kohlar Heinz. Harper Collins., New York.

Subject: Business Statistics	
Course Code: MC-106	Author: Dr. Pardeep Gupta
Lesson: 16	Vetter: Dr. B.K. Punia
STATISTICAL QUALITY CONTROL	

STRUCTURE

16.1 Learning Objectives

16.2 Introduction

16.2.1 Statistical Quality Control

16.2.2 Control charts

16.2.3 $\bar{x}$ Charts: Control charts for process means

16.2.4 R-Charts: Control charts for process variability

16.2.5 Control chart for *C* (Number of defects per unit)

16.2.6 *p*-charts: Control charts for attributes

16.3 Benefits and Limitations of Statistical Quality Control

16.4 Acceptance sampling

16.5 Check your Progress

16.6 Summary

16.7 Keywords

16.8 Self-Assessment Test

16.9 Answers to check your Progress

16.10 References/Suggested readings

16.1 LEARNING OBJECTIVES

After going through this lesson, the students will be able to:

- Understand the concept and importance of quality control
- Set up different types of control charts to keep the process under control
- Understand the concepts of acceptance sampling, single, double and multiple sampling plans
- Select the most appropriate sampling plan.

16.2 INTRODUCTION

The subject of 'quality control' has assumed considerable importance in recent years in the wake of globalisation of economies world over. As a result, there has been tremendous increase in competition amongst business enterprises both within and outside the country.

Quality has been defined in different ways by different experts but almost all those definitions emphasis that quality must meet the requirements of the customer. While quality is very vital for providing satisfaction to the customer, it goes far beyond this. For industrial and commercial organisations, quality is not only central to profitability but crucial to business survival. This aspect has assumed considerable importance in today's tough and challenging business environment. If quality is ignored or overlooked by these organisations then their continued existence is in danger.

The main factor that affects quality is variability in the process. This variability does not allow a factory to provide consistently a standard quality product. Prior to mass production, an individual worker or a few of them produced by hand, checking frequently if the product manufactured is coming out as they had conceived it. If it was distorted, they would again check where the fault laid, measure, and rework on it. However, when goods began to be manufactured on a mass scale, it became apparent that individual items could not be identical. It is almost impossible to eliminate variability completely. Such a situation poses a major problem in that the parts that are supposed to fit together would not fit. This shows that variability is the cause of poor quality.

The various causes of variation in the product may be classified into two categories:

(a) Scientific and identifiable

(b) Random and Chance

The first category comprises such causes as the use of defective raw material, poor equipment, poor workmanship, and so on. While the second category contains causes that do not have any bearing on the production process. The main purpose of our quality control exercise is to segregate specific and identifiable causes from the chance or random causes.

In the early days of mass production, inspection of the product and sorting out the defective ones was the chief method used for quality control. It was thought that the rejection of defective items would not cost much as the marginal cost of each unit was small. But gradually it became apparent that the costs of defective items were much higher than supposed earlier. This is because a number of people had to be employed to inspect the product besides losing the goodwill of the customers.

This realisation laid emphasis on doing things right at the very first time, focussing on the concept of *zero defects*. This means that efforts must be made to prevent defects at each stage of manufacturing a product or delivering a service. In order to achieve this, workers engaged in the production are given the responsibility to check their output rather than to pass it on for a final inspection. One major benefit of this approach is that workers feel a sense of pride and satisfaction for the responsibility given to them.

16.2.1 Statistical Quality Control

Statistical Quality Control (SQC) is the application of appropriate statistical tools to processes to ensure continuous improvement in quality of products, services and productivity in the workforce. As far back as in 1920, Walter A. Shewart created a system for tracking variation and identifying its causes. His system of SQC was further developed by W. Edwards Deming, a one-time colleague of Shewart. It is nothing else but a differentiation of the causes of variation during the operation of any process. The basic approach to statistical quality control is to identify a parameter that is easy to measure and is relevant to ascertain whether the quality is being maintained. For this purpose, control charts are used.

16.2.2 Control charts

Control charts show a step-by-step approach to statistical quality control. These are 'road-maps' that are very helpful in solving the problems pertaining to quality. The underlying feature of such a chart is that there are certain SQC techniques that are most appropriate in each step.

Figure 16.1 gives a schematic control chart. It will be seen from Fig. 16.1 that the control chart has three zones. These are: stable zone, warning zone and action zone.

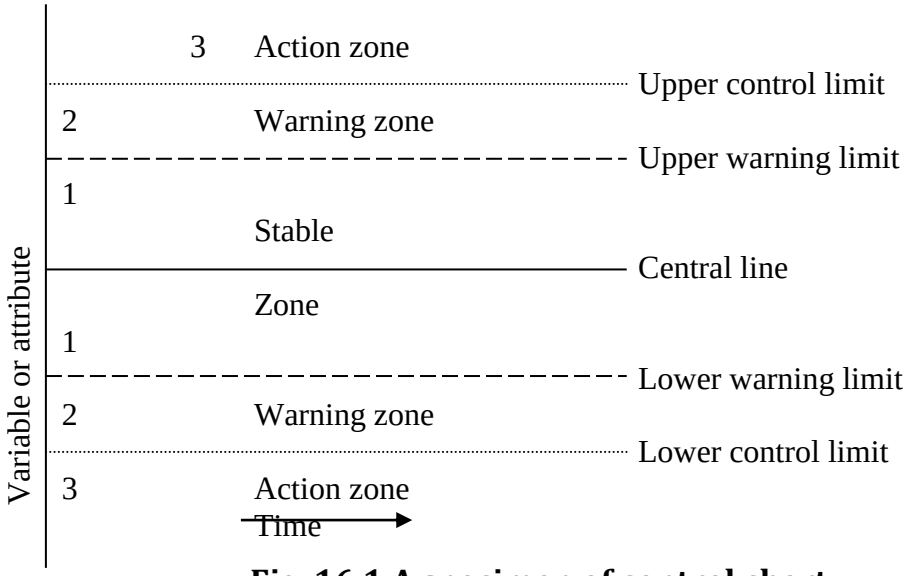


Fig. 16.1 A specimen of control chart

The action required depends on the zones in which the results fall. The possibilities are:

(xviii) Nothing needs to be done in case of stable zone wherein variation occurs due to common causes only.

(xix) In respect of warning zone, there seem to be special causes of variation. There is a need for collecting more information and having a watchful eye on the process.

(xx) Action zone suggests that special causes of variation in the process are present. The situation demands further investigation and where appropriate the process needs to be adjusted.

These three situations can be compared to traffic lights, which signal 'stop', 'caution' or 'go'. Let us examine in some more detail major parts of a control chart.

16.2.2.1. Major parts of a control chart

A control chart generally includes the following four major parts which are shown in Fig. 16.2.

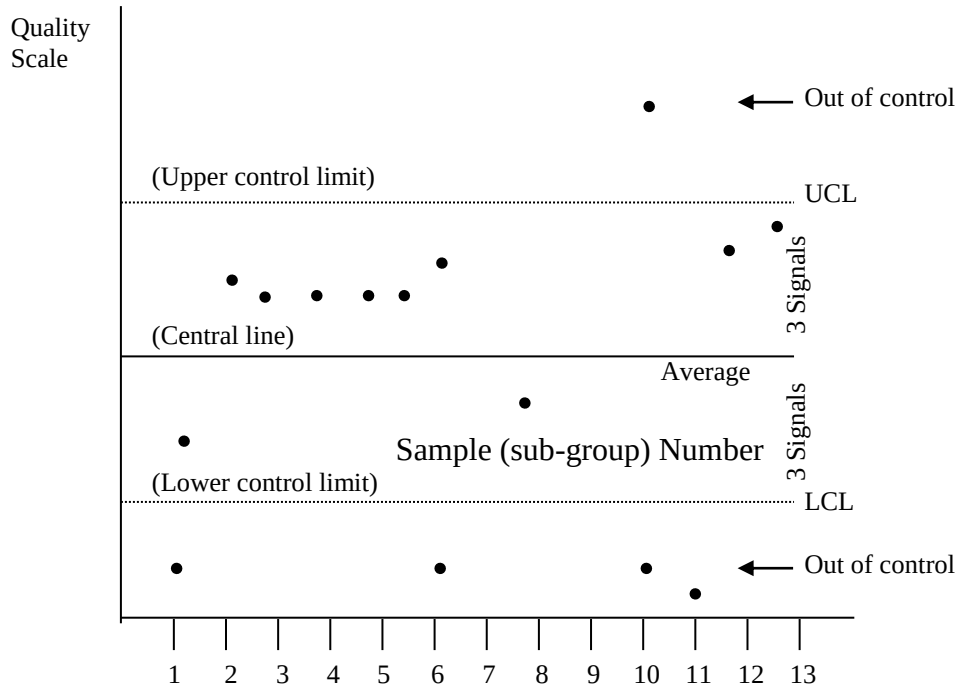


Fig. 16.2. Major parts of a control chart

(a) Quality scale- This is a vertical scale, which is marked as per the chosen quality characteristic (either in variables or attributes) of each sample.

(b) Plotted samples- The control chart does not show the qualities of individual items of a sample. Instead, the quality of the entire sample represented by a single value (a statistic) is shown. The single value plotted on the chart is in the form of a dot (or sometimes a small circle or a cross).

(c) Sample numbers- The samples, which are also referred to as sub-groups in SQC, on a control chart are numbered individually and are shown on a horizontal line. The line is usually shown at the bottom of the chart. It may be noted that the utility of the control chart technique depends to a great extent on the proper grouping of items into samples. The grouping should be such that variation in quality among items within the same sample is small, while variation between one sample and another is large. Such a sample is regarded as 'rational sub-group'.

(d) The horizontal lines- The central line represents the average quality of the samples plotted on the chart. The lines above the central line shows the upper control limit (UCL), which is commonly obtained by adding 3 sigmas (σ) to the average that is, mean +3 standard deviation. Similarly, the lower control limit (LCL) is given below the central line. It is obtained by subtracting 3 sigmas (σ)

from the average, that is, mean –3 standard deviation. The upper and lower control limits are usually drawn as dotted lines.

16.2.2.2 Why 3-Sigma Limits?

We have just said that upper and lower control limits are set at 3σ limits. One may ask the reason for this approach. It may be noted that the 3σ limits were first proposed by Shewart for his control charts. On the basis of probability consideration, if variable X is normally distributed, the probability that a random observation on the variable will lie between μ ± 3σ (where μ is the mean and σ the standard deviation of X) is 0.997, which is extremely high. It may be recalled the area of the normal curve between μ ± 3σ is 99.73 per cent. This means that the probability of a random observation going beyond these limits is nearly 0.003. This means that the variable quality characteristic is assumed to be normally distributed and that the probability of a sample point going outside 3σ limits when the process is in control is very small. If a sample point goes beyond this limit, it is highly likely that the normality assumption of the process is not applicable.

In order to set up a sound quality control mechanism, the concerned organisation must be keenly interested. It must take the following steps.

First, it must select the quality characteristics, which need to be kept under control. Besides, both their upper and lower limits within which variation can be tolerated should be fixed up. Second, the production process must be analysed so that the possible causes of variation can be determined. Finally, it must lay down as to how the inspection data will be collected and recorded as also how they will be subdivided. Depending on the type of inspection data available, any one of the following types of control charts can be used.

(xxi) Control charts for $\bar{x}$

(xxii) Control chart for □ or R alone

(xxiii) Control chart for C

(xxiv) Control chart for p or pn

16.2.3 $\bar{x}$ Charts: Control charts for process means

In order to ascertain whether the process is in control or out of control, $\bar{x}$-charts are constructed. In regard to the process output, there is an assumption of normality where μ and σ are known, though in many situations this assumption may not hold good. We know that the sample means have a sampling distribution with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. The construction of $\bar{x}$-chart needs the values of μ and σ and also a sample size n. There are three lines in a $\bar{x}$ control chart, viz. the centre line indicating $\mu_{\bar{x}}$, the upper control limit (UCL), with value $\mu_{\bar{x}} + 3\sigma$ and the lower control limit (LCL), with value $\mu_{\bar{x}} - 3\sigma$.

In addition to the control limits, there are warning limits, which are determined by 1.96 σ on either side of the centre line. Thus, the upper warning limit (UWL) = $\mu + (1.96)/\sqrt{n}$ and the lower warning limit (LWL) = $\mu - (1.96)/\sqrt{n}$. Figure 14.1 shows these two warning limits. However, the control charts do not normally show the warning limits.

Let us take an example to illustrate the procedure used in construction $\bar{x}$ control charts.

Example 16.1. A company is engaged in the manufacture of battery cells in its plant. The process is said to be under control if the mean life of battery cells is 1,200 hrs with a standard deviation of 75 hrs. Considering these values to be the process average and process dispersion, you are required to determine the 3-sigma control limits for $\bar{x}$-chart for samples of size 14.

Solution- Given are μ = 1,200 hrs, σ = 75 hrs and n = 14.

As the estimates of process average and process dispersion are based on a large sample, the desired control limits can be obtained by the following formula:

$\mu \pm 3\sigma/\sqrt{n}$

Substituting the values in the above formula,

$$\begin{aligned} UCL &= \mu + 3(75/\sqrt{14}) \\ &= 1{,}200 + 56.25 \\ &= 1{,}256.25 \\ LCL &= \mu - 3(75/\sqrt{14}) \\ &= 1{,}200 – 56.25 \\ &= 1{,}143.75 \end{aligned}$$

16.2.3.1. $\bar{x}$ Chart when μ and σ are not known

The preceding discussion has given us some basic ideas on $\bar{x}$-chart. The question is that when population mean and population standard deviation are not known to us, then how to construct $\bar{x}$-charts. In such cases, we use sample information to estimate unknown parameters. Let us take first the estimation of μ. This can be done by taking the mean of the sample mean ($\bar{\bar{x}}$). This can be calculated by the following formula

$\bar{\bar{x}} = \Sigma x/n \times k = \Sigma \bar{x}/k$

Where, n = number of observations in each sample

k = number of samples taken

In respect of control charts, it has become customary to use $\bar{R}$ as an estimate of σ. $\bar{R}$ signifies the average of the sample ranges. It is a biased estimator of σ, and d is the correction factor. The value for d_2 is given in question. Thus, the upper and lower control limits (UCL and LCL) for an $\bar{x}$-chart are computed with the following formulas:

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}}$$

$$\text{LCL} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}}$$

In the above formula, d_2 stands for control chart factor. These limits are often calculated as $\bar{\bar{x}} \pm A_2 \bar{R}$ where $A_2 = 3/(d_2\sqrt{n})$.

By using these formulas, we can now plot the three lines— CL (central line), UCL (upper control line) and LCL lower control line). Let us take an example to show how these formulas can be used.

Example 16.2. Suppose we are given the following information:

$n = 20$, $\bar{\bar{x}} = 75$, $d_2 = 3.735$ and $\bar{R} = 15$. We are asked to find the CL, UCL and LCL for a $\bar{x}$ control chart.

Solution. It is obvious that CL is the grand mean, that is, 75.

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}}$$

$$= 75 + \frac{3(15)}{3.735\times\sqrt{20}}$$

$$= 75 + \frac{45}{16.70}$$

$$= 77.69$$

$$\text{UCL} = \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}}$$

$$= 75 - \frac{3(15)}{3.735\times\sqrt{20}}$$

$$= 75 - \frac{45}{16.70}$$

$$= 72.31$$

Example 16.3. A company manufactures tyres. A quality control engineer is responsible to ensure that the tyres turned out are fit for use up to 40,000 km. He monitors the life of the output from the production process. From each of the 10 batches of 900 tyres, he has tested 5 tyres and recorded the following data, with $\bar{x}$ and $\bar{R}$ measured in thousands of km.

Batch	1	2	3	4	5	6	7	8	9	10
$\bar{x}$	40.2	43.1	42.4	39.8	43.1	41.5	40.7	39.2	38.9	41.9
$\bar{R}$	1.3	1.5	1.8	0.6	2.1	1.4	1.6	1.1	1.3	1.5

Construct an $\bar{x}$-chart using the above data. Do you think that the production process is in control? Explain. (Value of $d_2 = 2.326$)

Solution.

$$\bar{\bar{x}} = \frac{\sum\bar{x}}{k} = \frac{410.8}{10} = 41.08$$

$$\bar{R} = \frac{\sum R}{k} = \frac{14.2}{10} = 1.42$$

$$\text{CL} = 41.08$$

$$\text{UCL} = \bar{\bar{x}} + \frac{3\bar{R}}{d_2\sqrt{n}}$$

$$= 41.08 + \frac{3(1.42)}{2.326\times\sqrt{5}}$$

$$= 41.08 + \frac{4.26}{2.326\times2.24}$$

$$= \quad 41.08 + 0.82 = 41.9$$

$$\text{LCL} = \quad \bar{\bar{x}} - \frac{3\bar{R}}{d_2\sqrt{n}}$$

$$= \quad 41.08 - \frac{3(1.42)}{2.326\times\sqrt{5}}$$

$$= \quad 41.08 - \frac{4.26}{2.326\times 2.24}$$

$$= \quad 41.08 - 0.82 = 40.26$$

The production process is in control in respect of only 3 batches as is indicated in Fig. 14.3. The production process is in respect of batches 1, 4, 8 and 9 has gone out of control so also batch numbers 2, 3 and 5.

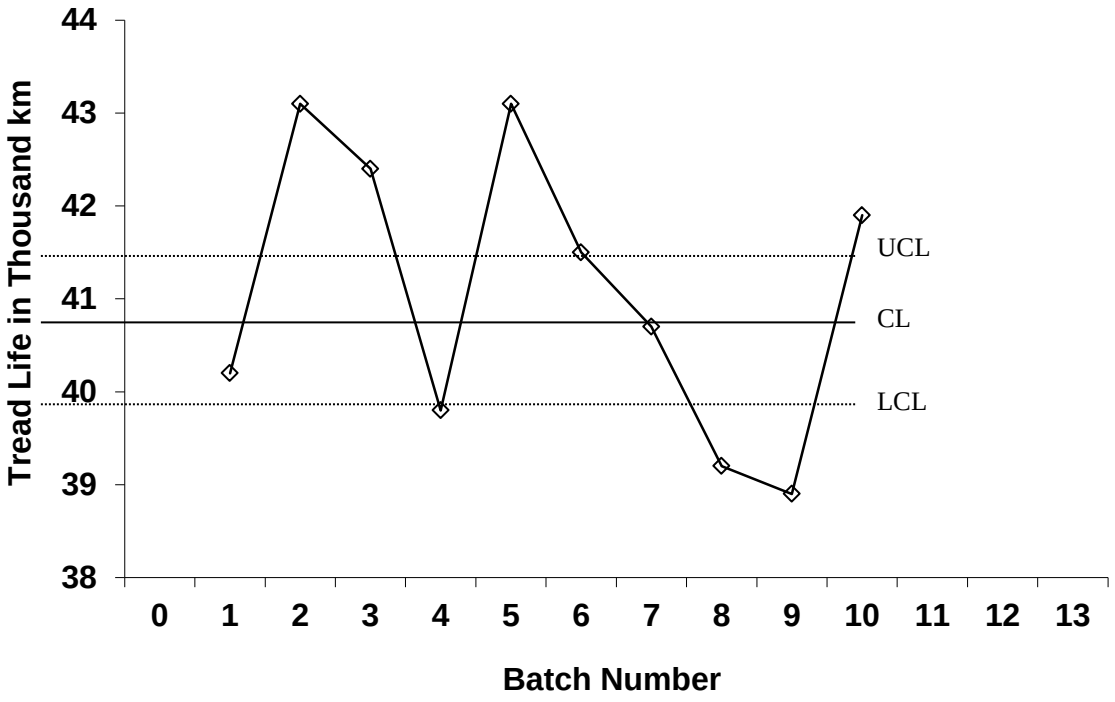


Fig. 16.3. $\bar{x}$-Chart for the data given in example

16.2.4 R-Charts: Control charts for process variability

In this chart, the value of the sample range for each of the samples is plotted. The central line for R-charts is placed at $\bar{R}$. Now, we have to decide the control limits for which we need some additional

information regarding the sampling distribution of R, in particular its standard deviation σ_R. For this purpose, the following formula is used.

$\sigma R = d_3\sigma$

where

σ = population standard deviation

d_3 = another factor depending on n

The values of d_3 are given in question.

Control Limits for an R-Chart

$$\text{UCL} = \quad \bar{R} + \frac{3d_3\bar{R}}{d_2} = \bar{R}\left(1+\frac{3d_3}{d_2}\right)$$

$$\text{LCL} = \quad \bar{R} - \frac{3d_3\bar{R}}{d_2} = \bar{R}\left(1-\frac{3d_3}{d_2}\right)$$

It may be noted that these limits are often calculated as:

UCL = $\bar{R}D_4$, where $D_4 = 1 + \frac{3d_3}{d_2}$

LCL = $\bar{R}D_3$, where $D_3 = 1 - \frac{3d_3}{d_2}$

The values of D_3 and D_4 can also be found from Table of control charts.

Example 16.4. We have to determine the UCL and the LCL by applying the above formulae to the data given in Example 14.3.

Solution. The UCL and the LCL are calculated as follows:

$$\text{UCL} = \quad \bar{R}\left(1+\frac{3d_3}{d_2}\right)$$

$$= \quad 1.42\left(1+\frac{3(0.864)}{2.326}\right)$$

$$= \quad 1.42\ (1 + 1.11) = 2.996 \text{ or } 3 \text{ approx.}$$

$$\text{LCL} = \quad \bar{R}\left(1-\frac{3d_3}{d_2}\right)$$

$$= \quad 1.42\left(1-\frac{3(0.864)}{2.326}\right)$$

$$= \quad 1.42 \times -0.11 = -0.156 \text{ (to be taken as zero)}$$

Some explanation is needed for the zero value of LCL. A sample range is always a non-negative number (because it is the difference between the largest and smallest observations in the sample). However, when $n \leq 6$, the LCL computed by the above equation will be negative. Although in this case n is 10, yet the calculation shows a negative value. As such, we set the value of LCL at zero.

A major limitation of *R*-chart arises from the characteristic of range itself. As we know that the range considers only the highest and the lowest values in a distribution, it may ignore the nature of variation in the remaining observations. Further, it is influenced by extreme values, which may significantly differ from one sample to the other. In view of these limitations, *R*-chart is only a convenient device for examining variability of the process.

16.2.5 Control chart for *C* (Number of defects per unit)

So far we have consider the control charts for attributes in those cases wherein a random sample of definite size is selected and examined in some way. However, there are certain situations where the number of events, defects, errors can be counted, but there is no information about the number of events, defects or errors that are not present. Each item is classified in one of the two categories- defective or non-defective. In such cases, we know the number of defects, say, number of holes in a fabric but we do not know the number of non-defects present. In such cases, the Poisson distribution is to be applied.

The central lines of the control chart for *C* is $\overline{C}$ and the 3-sigma control limits are

$$\text{UCL} = \overline{C} + 3\sqrt{\overline{C}}$$

$$\text{LCL} = \overline{C} - 3\sqrt{\overline{C}}$$

This formula is based on a normal curve approximation to the Poisson distribution. The use of the C-chart is appropriate if the occasions for a defect in each production unit are infinite, but the probability of a defect at any point is very small and is constant.

Example 16.5.

Fifteen pieces of cloth from different rolls contained respectively 1, 5, 3, 2, 7, 6, 3, 2, 6, 5, 4, 3, 5, 6, and 3 imperfections. Draw a control chart using these data and state whether the process is in a state of statistical control.

Solution.

$\overline{C}$ =(1 + 5 + 3 + 2 + 7 + 6 + 3 + 2 + 6 + 5 + 4 + 3 + 4 + 6 + 3)/15

= 60/15 = 4

UCL = $\overline{C} + 3\sqrt{\overline{C}}$

= 4 + 3√4 = 4 + 6 = 10

LCL = $\overline{C} - 3\sqrt{\overline{C}}$

= 4 – 3√4 = 4 – 6 = –2

Since the number of defectives cannot be negative, the lower control limit will be taken as zero. Figure 14.4 shows both the control limits. The chart clearly shows that all the imperfections in cloth are within the control limits, that is, no point lies outside the control limits. This suggests that the process is in a state of statistical control.

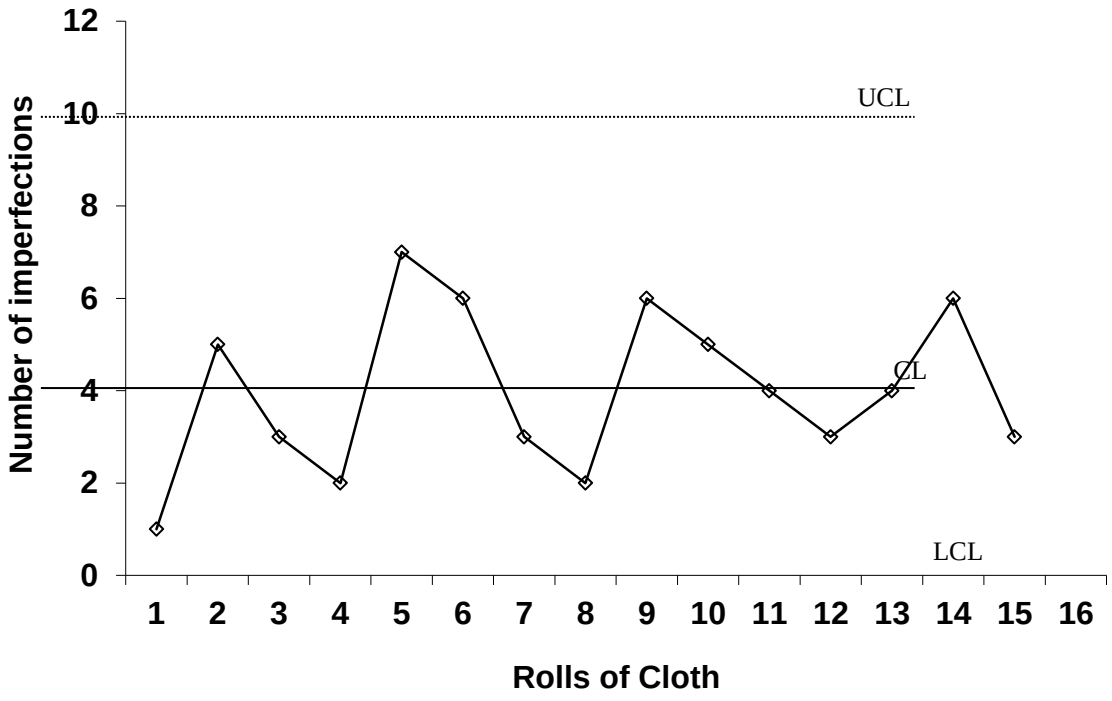


Fig. 16.4. Control chart for C

16.2.6 *p*-charts: Control charts for attributes

The control chart for attributes is known as the *p*-chart. Such a chart is used to control the proportion or percentage of defectives per sample. It may be noted that there is an assumption that the items are

produced by Bernoulli process, which implies the following three assumptions: (i) There are only two outcomes– acceptable or defective. (ii) The outcomes occur randomly. (iii) There is no change in the probability of either outcome for each trial.

As we have seen earlier that the C-chart is concerned with the number of defectives, it can be easily converted into proportion by dividing the number of defectives by the sample size. Thus, we can use the *p*-chart in place of the C-chart. In order to draw the *p*-chart, we have to follow the following procedure:

(xxv) Calculate the average fraction defective ($\bar{p}$) by dividing the number of defective units by the total number of units inspected.

(xxvi) The value of $\bar{p}$ is now used to draw a horizontal line.

(xxvii) The upper and lower control limits are to be obtained by using the following formulas:

$$\left.\begin{aligned} \text{UCL} &= \bar{p} + 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \\ \text{LCL} &= \bar{p} - 3\sqrt{\frac{\bar{p}\,\bar{q}}{n}} \end{aligned}\right\} \text{ where } \bar{q} = (1 - \bar{p})$$

Any sample point falling outside the UCL and the LCL indicates that the process is not in control. It is preferable to set up the chart to express 'percent defective' to 'fraction defective'.

Example 16.6. The following figures give the number of defects in 10 samples, each containing 200 items: 40, 44, 22, 34, 24, 32, 28, 32, 34 and 30. Calculate the values for central line and the upper and lower control limits of *p*-chart. Draw the *p*-chart and comment if the process can be regarded in control.

Solution.

Table 1. Worksheet for calculating the values for *p*-chart

Sample No.	No. of defectives	Fraction defectives
1	40	0.20
2	44	0.22

3	22	0.11
4	34	0.17
5	24	0.12
6	32	0.14
7	28	0.14
8	32	0.14
9	34	0.17
10	30	0.15
Total	320	

$$\bar{p} = \frac{\text{No. of units defective}}{\text{Total no. of units inspected}} = \frac{320}{2000} = 0.14$$

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.14 + 3\sqrt{\frac{0.16(1-0.16)}{200}}$$

$$= 0.14 + 0.07776 = 0.2378$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.14 - 3\sqrt{\frac{0.16(1-0.16)}{200}}$$

$$= 0.14 - 0.07776 = 0.0822$$

It will be seen from Fig. 16.6 that all the units fall within the upper and lower control limits. On the basis of this chart, we can say that the process is well under control. It may be noted that we have plotted the percentage defective instead of fraction defective in the above chart.

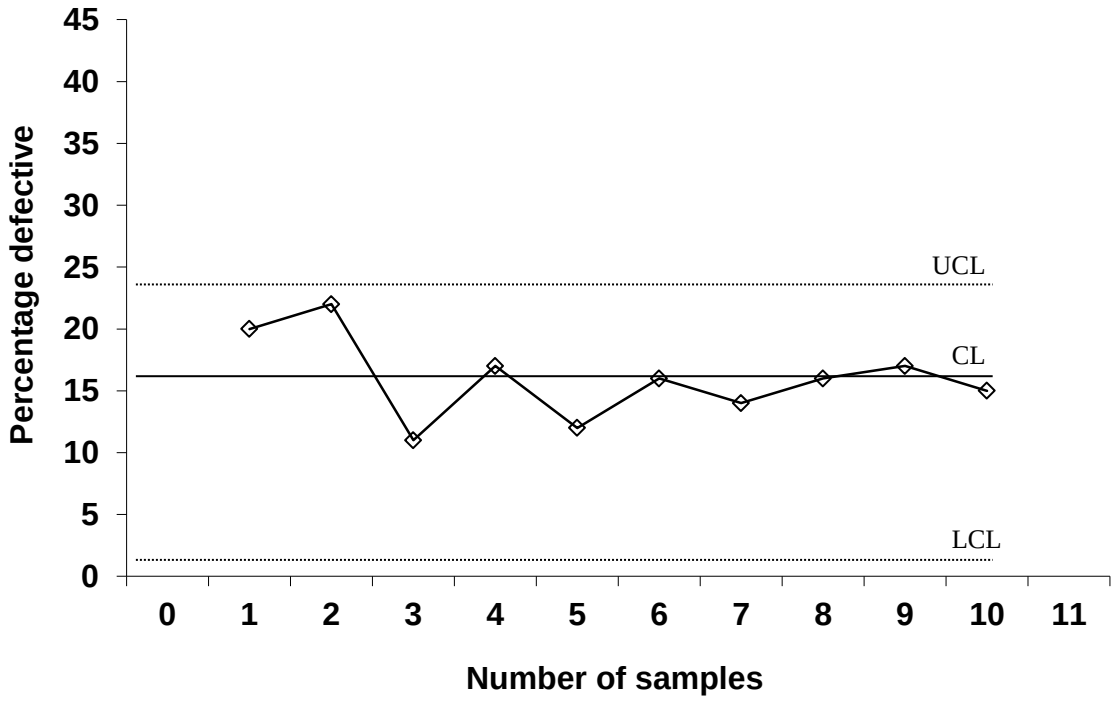


Fig. 16.6. *p*-Chart for data given in example 14.6

16.3. Benefits and Limitations of Statistical Quality Control

Benefits

There are several benefits of SQC approach and these include:

(xxviii)SQC can be applied to any type of problem selected and process originally tackled will result into improvement.

(xxix) This approach eliminates the 'emotion' factor and the decisions are based on facts rather than on opinions.

(xxx) As the workers are directly involved in the improvement process, their 'quality awareness' increases.

(xxxi) The knowledge and experience potential of those involved in the process is released in a systematic way through the investigative approach. They increasingly realise that their role in problem solving is collecting and communicating the relevant facts on which decisions are made.

(xxxii) Managers and supervisors solve problems methodically instead of in a haphazard manner. Thus, the approach to the problem becomes unified in place of an individual approach earlier.

(xxxiii)In case of any inquiry from the government or any other appropriate authority, the quality can be defended on the basis of statistical process control.

(xxxiv)Since the firm strictly adheres to the SQC, the users of the product may rely on it and may not resort to check the quality themselves.

Limitations

Despite the above mentioned advantages of the SQC, it may be noted that it is unable to solve all the problems arising in quality improvement. There are several highly complex problems where SQC may not be in a position to contribute much towards reduction of variability. This apart, at times, managers use SQC mechanically and construct control charts without going into the depth of the problem. As a result, statistical methods have been criticised at times. It has been argued that continuous improvement in quality can be attained by studying all parts of an organisation and not merely one part viz. production process.

16.4 ACCEPTANCE SAMPLING

Acceptance sampling involves sampling inspection by a purchaser who has to decide whether to accept a shipment of product. Thus, the objective of acceptance sampling is either to accept or to reject the product. It does not attempt to control the quality during the manufacturing process. This is altogether a different approach from what has been followed in control charts discussed earlier.

A major advantage of acceptance sampling is that it can motivate suppliers to improve the quality of their items. Suppose a company receives a batch of components from its supplier and finds that 10 percent of the supply is defective. Although 90 per cent of it is free from defects, but the company may decide to reject the entire lot to ensure its qualitative output. This decision of the company would result into heavy loss to the supplier. He has to suffer even though a small proportion of his equipment was defective. In order to avoid such an eventuality, the supplier would be very particular from the very beginning to ensure his supplies are free from defects. In contrast, if the company rejects only 10 percent of defective equipment, it amounts to imposing a high cost on itself and a low cost on the supplier. The various kinds of sampling plans used by purchasers to accept or reject a lot are discussed as under:

16.4.1. Single-sampling plan

When the decision on whether to accept or reject a plot is based on only one sample, the acceptance plan is said to be on a single-sampling plan. There are three things that need to be specified in a single-sample plan. These are: (a) number of items *N* in the lot from which the sample is chosen, (b) number of articles *n* drawn by random sample from the given lot, and (c) the acceptance number c, which specifies the maximum number of defective articles allowable in the sample. In case the number of defective articles crosses this limit in the sample drawn, the entire lot is to be rejected.

16.4.2. Double-sampling plan

Double-sampling plan is obviously more complicated than the single-sampling plan. In this case, a lot is immediately accepted or rejected depending on the condition of the first sample. At times, the management finds that the first sample is neither good enough nor bad enough so as to take a decision one way or the other. In such a situation, it defers its decision until a second sample is drawn. On the basis of the evidence from both the first and the second samples, a decision is finally taken whether to accept or reject the lot. A double-sampling plan depends on five specified numbers (besides *N*): n_1, c_1, n_2, n_1+n_2 and c_2 ($> c_1$), which are used as follows:

First, a sample of size n_1 is taken. Let b_1 denote the number of defective pieces in the first sample and c_1 denote the number of defective pieces acceptable in the lot, then

(a) accept the lot if $b_1 \leq c_1$

(b) reject the lot if $b_1 > c_1$

(c) an additional n_2 units are sampled if $c_1 < b_1 \leq c_2$

Let b_2 be the total number of defective pieces in the combined sample of n_1+n_2 units:

(*d*) accept the lot if $b_2 \leq c_2$

(e) reject the lot if $b_2 > c_2$

As mentioned earlier, double-sampling plans are more complicated than the single-sampling plan. But, as they are more powerful, they are more frequently used in quality control problems.

16.4.3. Multiple or sequential sampling plan

We have seen earlier that when a single-sampling plan is unable to give us a clear decision, we take recourse to the double-sampling plan. It may just be possible that even a double-sampling plan may not give us a clear decision. In such a case, we may go on to have another sample before we reach a definite decision. Thus, three or more sampling plans can be used. This is known as multiple or sequential sampling. Since such plans are extremely complicated, they are seldom used in practice.

16.5 CHECK YOUR PROGRESS

1. In case of any inquiry from the government or any other appropriate authority, the quality can be defended on the basis of ..... .
2. In control chart for C, the ..... is to be applied.
3. The construction of $\bar{x}$-chart needs the values of μ and σ and also a ..... .
4. Nothing needs to be done in case of stable zone wherein variation occurs due to ..... causes only.

(v) In respect of warning zone, there seem to be special causes of.....

(vi) Action zone suggests that special causes of variation in the process are.....

16.6 SUMMARY

In the early days of mass production, inspection of the product and sorting out the defective ones was the chief method used for quality control. It was thought that the rejection of defective items would not cost much as the marginal cost of each unit was small. But gradually it became apparent that the costs of defective items were much higher than supposed earlier. This is because a number of people had to be employed to inspect the product besides losing the goodwill of the customers. Statistical Quality Control (SQC) is the application of appropriate statistical tools to processes to ensure continuous improvement in quality of products, services and productivity in the workforce. Control charts show a step-by-step approach to statistical quality control. These are 'road-maps' that are very helpful in solving the problems pertaining to quality. The underlying feature of such a chart is that there are certain SQC techniques that are most appropriate in each step. Depending on the type of inspection data available, any one of the following types of control charts can be used: Control charts for $\bar{x}$, Control

chart for σ or R alone, Control chart for C, Control chart for p or $p_{n.}$ There are several benefits and limitations of SQC. The objective of acceptance sampling is either to accept or to reject the product. It does not attempt to control the quality during the manufacturing process. This is altogether a different approach from control charts.

16.7 KEYWORDS

Acceptance sampling: It involves sampling inspection by a purchaser who has to decide whether to accept a shipment of product.

P-Chart: The control chart for attributes is known as the *p*-chart. Such a chart is used to control the proportion or percentage of defectives per sample.

R- Chart: In this chart, the value of the sample range for each of the samples is plotted. The central line for R-charts is placed at $\overline{R}$.

Quality scale: This is a vertical scale, which is marked as per the chosen quality characteristic (either in variables or attributes) of each sample.

Plotted samples: The control chart does not show the qualities of individual items of a sample. Instead, the quality of the entire sample represented by a single value (a statistic) is shown. The single value plotted on the chart is in the form of a dot (or sometimes a small circle or a cross).

Sample numbers- The samples, which are also referred to as sub-groups in SQC, on a control chart are numbered individually and are shown on a horizontal line.

The horizontal lines- The central line represents the average quality of the samples plotted on the chart

.

16.8 SELF-ASSESSMENT TEST

1. What is statistical quality control? How is it useful to industry?
2. What is a control chart? Describe how it is constructed and used?
3. Describe briefly the working of the p-chart.
4. Write a detailed note on ‘Acceptance Sampling’.

16.9 ANSWERS TO CHECK YOUR PROGRESS

1. Statistical process control.

2. Poisson distribution

3. Sample size n.
4. Common
5. Variation
6. Present

16.10 REFERENCES/SUGGESTED READINGS

7. Business Statistics by Shenoy and Shenoy.
8. Statistical Methods by S.P. Gupta.
9. Statistics for Business and Economics by R.P. Hooda.

NOTE

NOTE

NOTE

NOTE

NOTE

NOTE

NOTE