Statistical Methods In Economics

DECO504

Edited by:
Dr.Pavitar Parkash Singh

LOVELY PROFESSIONAL UNIVERSITY

STATISTICAL METHODS IN ECONOMICS

Edited By
Dr. Pavitar Parkash Singh

Printed by
USI PUBLICATIONS
2/31, Nehru Enclave, Kalkaji Ext.,
New Delhi-110019
for
Lovely Professional University
Phagwara

SYLLABUS

Statistical Methods in Economics

Objectives:

The course aims to equip the students with statistical tools and concepts that help in decision making. The emphasis is on their application in business.

Sr. No.	Content
1	Definition of Statistics: Importance and scope of statistics and its limitations, Types of data collection: Primary and Secondary: Methods of collecting Primary data, Classification and Tabulation of data: Frequency and cumulative frequency distribution
2	Central Tendency: Mean, Median and Mode and their Properties, Application of Mean, Median and Mode
3	Dispersion: Meaning and characteristics. Absolute and relative measures of dispersion including Range, Quartile deviation, Percentile, Mean deviation, Standard deviation, Skewness and Kurtosis: Karl Pearson, Bowley, Kelly's methods
4	Correlation: Definition, types and its application for Economists, Correlation: Scatter Diagram Method, Karl Pearson's coefficient of correlation, Rank correlation method
5	Linear Regression Analysis: Introduction and lines of Regression, Coefficient of regression method simple, Correlation analysis vs. Regression Analysis
6	Index number: Introduction and Use of index numbers and their types, Methods: Simple (unweighted) Aggregate Method, Weighted aggregate method, Methods: Simple (unweighted) Aggregate Method, Methods: Simple Average of Price Relatives, Methods: Weighted Average of Price Relatives, Test of consistency: Unit test, Time Reversal Test, Factor Reversal Test and Circular, Cost of Living index and its uses. Limitation of Index Numbers
7	Time Series Analysis: Introduction and components of time series, Time Series Methods: Graphic, method of semi-averages, Time Series Methods: Principle of Least Square and its application, Methods of Moving Averages
8	Theory of Probability: Introduction and uses, Additive and Multiplicative law of probability
9	Theory of Estimation: Point estimation, Unbiasedness, Consistency, Efficiency and Sufficiency, Method of point estimation and interval estimation
10	Types of Hypothesis: Null and Alternative, types of errors in testing hypothesis, Level of significance

CONTENTS

Notes

Unit 1: Definition of Statistics, Importance and Scope of Statistics and Its Limitations

CONTENTS

Objectives

After reading this unit students will be able to:

- Know the Definition of Statistics.
- Discuss the Importance and Scope of Statistics.
- Explain the Limitations of Statistics

Introduction

In general sense the word Statistics means facts and figures of a particular phenomenon—Under reference in numerical numbers. In the traditional period the scope of Statistics was very much limited to the collection of facts and figures pertaining to the age-wise and sets-wise distribution of population, wealth etc. But now-a-days we can say that Statistics constitutes an integral part of every scientific and economic inquiry: Social and economic studies without Statistics are useless. Statistics thus play a vital role and as **Tippet** has rightly remarked, "It affects everybody and touches life at many points."

1.1 Definition of Statistics

It has been observed that the word 'Statistics' comes from Latin word 'Status' which means Political State. It has also been believed that the word Statistics comes from Italian word 'Stato'. This word was used in the fifteenth century for the 'State' in actual practice these words were used for Political State or Stateman's art.

Now-a-days Statisticians use statistics both in singular and plural sense. In the singular sense the term Statistics is associated with "A body of methods for making decisions when there is uncertainty arising from incompleteness or the instability of the information available for making such decisions." In its plural sense Statistics refers to numerical Statements of facts such as per capita income, population etc. Thus, some authorities have defined Statistics as Statical data (Plural sense) whereas other as Statistical method (Singular sense). According to **Oxford Concise Dictionary,** "Statistics—(as treated as plural): numerical facts, systematically collected, as Statistics of population, crime etc. (treated as singular): Science of collecting, classifying and using Statistics."

Notes

Definition

The definition of Statistics can be divided into the following two heads:

(A) In Plural Sense,

(B) In Singular Sense.

(A) In Plural Sense: The following are the definitions of Statistics in Plural Sense:

According to **H. Secrist**—"*By Statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.*"

In the words of **L. R. Connor**—"*Statistics are measurements, enumeration or estimater of natural or social phenomena systematically arranged so as to exhibit their interrelations.*"

According to **Yule & Kendall**—"*By Statistics we mean quantitative data affected to a marked extent by multiplicity of causes.*"

In the opinion of **A. L. Bowley**—"*Statistics are numerical statement of facts in any department of enquiry placed in relation to each other.*"

According to **Webster**—"*Classified facts, representing the condition of the people in a State, specially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement.*"

On the basis of the above definitions the following characteristics are there in Statistics:

1. **Statistics are aggregate of facts:** Single and unconnected figures are not Statistics. A single age of 22 years or 37 years is not Statistics but a series relating to the ages of a group of people would be called Statistics. Likewise single figure relating to birth, death, sale, etc. cannot be called Statistics but aggregates of such figures would be Statistics because they can be studied in relation to each other and are capable of comparison.
2. **Statistics are affected to a marked extent by multiplicity of causes:** Usually facts and figures are affected, to a considerable extent, by a number of factors operating together. For example —Statistics of prices are affected by conditions of demand, supply, imports, exports, currency circulation, etc. and various other factors.
3. **Statistics are numerically expressed:** Qualitative expression like good, bad, young, old etc. do not form a part of statisical study unless numerical equivalent is assigned to such expression. If it is said that the production of rice per acre in 1997 was 30 quintals and in the year 2002 it was 50 quintals, we shall be making Statistical statements.
4. **Statistics are enumerated according to reasonable standard of accuracy:** Facts and figures relating to any subject can be derived in two way, example—by actual counting and measurement or by estimates. Estimates cannot be as accurate and precise as actual measurements. For example—If the heights of a group of people are being measured, it is right if the measurements are correct to a centimetre but if are measuring the distance from Agra to Gwalior, a difference of a few kilometres even, can be easily ignored.
5. **Collected in a systematic manner:** If Statistics are collected in a haphazard manner, it might fail to give the accurate result. It is, therefore, essential that statistics must be collected in a systematic manner so that they may *Conform* to reasonable standard of accuracy.
6. **Collected for a pre-determined purpose:** Statistical data are collected and processed for a definite and pre-determined purpose. In general, no data are collected without a pre-determine purpose.
7. **Placed in relation to each other:** The Statistics should be comparable. If they are not comparable, they lose part of their value and thus the efforts in collecting them may not prove to be as useful as the requirements may be. It is necessary that the figures which are collected should be a homogeneous so as to make them comparable and more useful.

   On the basis of the above description it may be said that numerical data cannot be called Statistics hence "All Statistics are numerical statements of facts but all numerical statements of facts may not essentially be Statistics."

(B) In Singular Sense: The following are the definitions of Statistics in Singular Sense. **Lovitt** defines the science as, "*That which deals with the collection, classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomena.*"

Notes

According to **King,** "*The Science of Statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection or estimation.*
According to **Croxton** and **Cowden,** "*Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data.*"
A. L. Bowley tried to define Statistics in this group also. He was of the opinion that, "*Statistics is the science of measurement of social organism, regarded as a whole in all its manifestations.*"
According to ***Seligman,*** "*Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting of numerical data collected to throw some light on any sphere of enquiry.*"
On the basis of the above definitions it may be said that Statistics are numerical statements of facts capable of analysis and interpretation and science of Statistics are a study of principles and the method used in the correction, presentation and analysis of numerical data in any sphere of enquiry.

1.2 Importance and Scope of Statistics

Scope of Statistics

The scope of statistics are concerned with the new dimensions in the definition of statistics. In other words we can say — Are statistics a science or an art or both ? Science is concerned with the systematised body of knowledge. It shows the relationship between cause and effects. So far as art is concerned, it refers to the skill of collecting and handling of data to draw logical inference and arrive at certain results. Statistics may be used as a science and as an art. In this regard the following definitions may be given:
As per **Netter** and **Wanerman —** "*Statistics methods are mathematical techniques used to facilitate the interpretation of numerical data secured from groups of individuals.*"
In the words of **Paden** and **Lindquist —** "*Statistical methods are mathematical techniques used to facilitate the interpretation of numerical data secured from groups of individuals.*"
According to **Kaney** and **Keeping —** "*Statistics has usually meant the science and art concerned with the collection, presentation and analysis of quantitative data so that intelligent judgement may be made upon them.*"
According to **Anderson** and **Bancraft —** "*Statistics is the science and art of the development and application of the most effective method of collecting, tabulating and interpreting quantitative data in such a manner that the fallibility of conclusions and estimates may be assessed by means of inductive reasoning based on mathematics of probability.*"

Did u know? "Statistics are the straw out of which, I like every other economist have to make the bricks."

Importance Or Significance of Statistics

The importance of statistics is now being felt in almost every field of study. In fact, it is difficult to mention a subject which does not have any relation with the science of statistics. As a matter of fact statistical methods are common ways of thinking and hence are used by all types of persons. Suppose a person wants to purchase a car and he goes through the price list of various companies and makers, to arrive at a decision, what he really aims at is to have an idea about the average level and the range within which the prices vary, though he may not know a word about these terms. No doubt to say that statistical methods are so closely connected with the human actions and behaviour that all human activity may be explained by statistical method. The importance of statistics can be shown in the following heads:
Importance in Economics: In the study of economics the use of statistical methods are of great importance. Most of the economic principles and doctrines are based on the study of a large number of units and their analysis. By statistical analysis we can study the ways in which people spend their

Notes

income over food, rent, clothing, entertainment, education etc. For example, if the law of demand is to be analysed then we have to make an idea about the effect of price changes on demand both for an individual and for a market. For this purpose a large number of data and figures would be collected. On the basis of the available information, the demand schedule can be prepared and then the law of demand can be formulated. We thus find that in the field of economics, the use of statistics is indispensable.

Importance in Business Management: Business managers are required to make decision in the face of uncertainty. Modern statistical tools of collection, classification, tabulation, analysis of interpretation of data have been found to aid in making wise decisions at various levels of managerial functions. These tools are relied upon in arriving at correct decision in all these aspects —sales forcasting, price situations, credit position, quality control, inventory control, investment planning, tax planning are some of the areas where statistical techniques help the business management in present and future planning. On the basis of the above discussion it is clear that the use of statistical data and techniques is indispensable in almost all the branches of business management.

Role of Statistics in planning: Today planning cannot be formulated without statistics. The problems like over production, unemployment, low rate of capital formation etc. which are the major characteristics of developing countries can be understood with the help of statistical data. National Sample Survey Scheme was started to collect statistical data for use in planning in India. Economic planning is done to achieve pre-determined objectives and goals. They have to be expressed in quantitative terms. We, thus, find that in the field of economic planning the use of statistics is indispensable.

Statistics in Commerce: Statistics plays a very important role in the development of commerce. The statistical data on some macro variables like income, investment, profits etc. are used for the compilation of national income. Economic barometer are the gifts of statistical methods and businessmen all over the world make extensive use of them. The increasing application of the statistical data and the statistical techniques in accountancy and auditing are supported by the inclusion of a compulsory paper on statistics both in the Charted Accountant's and the Cost Accountant's examinations. Various branches of commerce utilise the services of statistics in different forms. Cost Accounting is entirely statistical in its outlook and it is with the help of this technique that the manufacturers and the producers are in a position to decide about the prices of various commodities. We, thus, find that the science of statistics is of great importance to commerce.

Utility of Bankers, Brokers, Insurance Companies etc: Bankers, Stock Exchange Brokers, insurance companies, investors and public utility concerns all make extensive use of statistical data and technique. A banker has to make a statistical study of business cycles to forecast a probable boom. On the basis of this study a banker decides about the amount of reserves that should be kept.

Statistics are important from the view-point of stock exchange, brokers and investors. They have to be conversant with the prevailing money rate at various centres and have to study their future trends. Likewise insurance companies cannot carry on their business in the absence of statistical data relating to life tables and premium rates. As a matter of fact insurance has been one of the basic branches of commerce and business which has been making use of statistics.

Importance to State: Statistics are very important to a State as statistics help in administration. In all the fields where the State has to keep accurate records and information, statistical systems are adopted. For example, for making the economic plan the State has to collect data or information, it has to estimate the figures of National Income to find out the real position of the country. For this purpose, the state needs statistics for carrying on these works. The state also needs data about the roads, transport and communication, financial affairs, internal and external trade etc.

Importance to Research: Statistical methods and techniques happen to be useful in gathering the public opinion on various problems facing the society. In the field of Industry and Commerce statisticians carry on different types of researches. No researcher, without the use of statistics, can fulfil his targets. Today, the study of statistical method is not only useful but necessary for research. To conclude, statistical methods and techniques have been used in almost all the spheres. Statistical methods are essential to understand the effect to determining the factors of economic development

Notes

in the past, what psychological and sociological factor need to be developed for economic development and for the success of a plan.

1.3 Limitations of Statistics

Though statistics is an important instrument of quantitative method and research in social sciences, physical science and life sciences, it suffers from a number of limitations. The following are the main limitations of statistics:

(1) **Absence of uniformity:** In any statistical inquiry the data obtained are heterogeneous in nature. Statistical methods alone cannot bring in perfect uniformity. Generally results obtained need not be uniform and hence will serve no purpose.

(2) **Statistics does not study individuals:** Statistics deals only with aggregate of facts. Hence, single figures, however important they might be, cannot be taken up within the purview of statistics. For example, the marks obtained by X student of a class are not the subject-matter of statistics but the average marks has statistical relevance.

(3) **Statistical results speak about only average:** Prof. A. L. Bowley has rightly remarked that Statistics is a science of average. It implies that statistical results are true only on average. For example, if we say that per capita income in India is Rs. 12,000 per annum, it does not mean that the per capita income of the members of the Birla's family and the income of the poor fellows who sleep in the slum area are equal. Therefore, averages give only contradicting results.

(4) **Statistics can be misused:** Statistics is misused very often in the sense that a corrupt man can always prove all that he wants to do by using false statistics. In the words of **W. I. King,** "One of the shortcomings of statistics is that they do not, bear on their face the label of their quality."

(5) **Laws are not stable:** The statistical laws are obtained on the basis of information available at one stage need not be true at another stage. The basic data changes and hence the basic laws governing them also change. Moreover, what is applicable to India need not be true in Japan.

(6) **Statistics cannot be applied to qualitative statistics:** The Statistic studies cannot be applied to qualitative attributes like good, bad, beautiful etc. For a whole sum coverage, the statistical tools must be applicable for quantitative and qualitative data.

Self-Assessment

1. Fill in the blanks:

(*i*) The word statistics is used in ..... senses namely ..... and ..... .

(*ii*) The word statistics refers either ..... information or to a method of dealing with ..... information.

(*iii*) Any collection of related observations is called as .....

(*iv*) Applied statistics is divided into two groups, they are ..... and .....

(*v*) All the rules of procedures and general principles which are applicable to all kinds of groups of data are studied under .....

1.4 Summary

- In the traditional period the scope of Statistics was very much limited to the collection of facts and figures pertaining to the age-wise and sets-wise distribution of population, wealth etc. But now-a-days we can say that Statistics constitutes an integral part of every scientific and economic inquiry: Social and economic studies without Statistics are useless. Statistics thus play a vital role and as **Tippet** has rightly remarked, "It affects everybody and touches life at many points."
- Now-a-days Statisticians use statistics both in singular and plural sense. In the singular sense the term Statistics is associated with "A body of methods for making decisions when there is uncertainty arising from incompleteness or the instability of the information available for making such decisions." In its plural sense Statistics refers to numerical Statements of facts such as per capita income, population etc. Thus, some authorities have defined Statistics as Statical data (Plural sense) whereas other as Statistical method (Singular sense). According to **Oxford Concise**

Notes

Dictionary, "Statistics—(as treated as plural): numerical facts, systematically collected, as Statistics of population, crime etc. (treated as singular): Science of collecting, classifying and using Statistics."

- If Statistics are collected in a haphazard manner, it might fail to give the accurate result. It is, therefore, essential that statistics must be collected in a systematic manner so that they may *Conform* to reasonable standard of accuracy.
- The Statistics should be comparable. If they are not comparable, they lose part of their value and thus the efforts in collecting them may not prove to be as useful as the requirements may be. It is necessary that the figures which are collected should be a homogeneous so as to make them comparable and more useful.
- On the basis of the above description it may be said that numerical data cannot be called Statistics hence "All Statistics are numerical statements of facts but all numerical statements of facts may not essentially be Statistics."
- The scope of statistics are concerned with the new dimensions in the definition of statistics. In other words we can say—Are statistics a science or an art or both ? Science is concerned with the systematised body of knowledge. It shows the relationship between cause and effects. So far as art is concerned, it refers to the skill of collecting and handling of data to draw logical inference and arrive at certain results. Statistics may be used as a science and as an art. In this regard the following definitions may be given:
- The importance of statistics is now being felt in almost every field of study. In fact, it is difficult to mention a subject which does not have any relation with the science of statistics. **Alfred Marshall** had mentioned that, "Statistics are the straw out of which, I like every other economist have to make the bricks." As a matter of fact statistical methods are common ways of thinking and hence are used by all types of persons. Suppose a person wants to purchase a car and he goes through the price list of various companies and makers, to arrive at a decision, what he really aims at is to have an idea about the average level and the range within which the prices vary, though he may not know a word about these terms. No doubt to say that statistical methods are so closely connected with the human actions and behaviour that all human activity may be explained by statistical method. The importance of statistics can be shown in the following heads:
- In the study of economics the use of statistical methods are of great importance. Most of the economic principles and doctrines are based on the study of a large number of units and their analysis. By statistical analysis we can study the ways in which people spend their income over food, rent, clothing, entertainment, education etc. For example, if the law of demand is to be analysed then we have to make an idea about the effect of price changes on demand both for an individual and for a market. For this purpose a large number of data and figures would be collected. On the basis of the available information, the demand schedule can be prepared and then the law of demand can be formulated. We thus find that in the field of economics, the use of statistics is indispensable.
- Today planning cannot be formulated without statistics. The problems like over production, unemployment, low rate of capital formation etc. which are the major characteristics of developing countries can be understood with the help of statistical data. National Sample Survey Scheme was started to collect statistical data for use in planning in India. Economic planning is done to achieve pre-determined objectives and goals. They have to be expressed in quantitative terms. We, thus, find that in the field of economic planning the use of statistics is indispensable.
- Bankers, Stock Exchange Brokers, insurance companies, investors and public utility concerns all make extensive use of statistical data and technique. A banker has to make a statistical study of business cycles to forecast a probable boom. On the basis of this study a banker decides about the amount of reserves that should be kept.
- Statistics are very important to a State as statistics help in administration. In all the fields where the State has to keep accurate records and information, statistical systems are adopted. For example, for making the economic plan the State has to collect data or information, it has to estimate the figures of National Income to find out the real position of the country. For this

Notes

purpose, the state needs statistics for carrying on these works. The state also needs data about the roads, transport and communication, financial affairs, internal and external trade etc.

- To conclude, statistical methods and techniques have been used in almost all the spheres. Statistical methods are essential to understand the effect to determining the factors of economic development in the past, what psychological and sociological factor need to be developed for economic development and for the success of a plan.
- In any statistical inquiry the data obtained are heterogeneous in nature. Statistical methods alone cannot bring in perfect uniformity. Generally results obtained need not be uniform and hence will serve no purpose.
- Prof. A. L. Bowley has rightly remarked that Statistics is a science of average. It implies that statistical results are true only on average. For example, if we say that per capita income in India is Rs. 12,000 per annum, it does not mean that the per capita income of the members of the Birla's family and the income of the poor fellows who sleep in the slum area are equal. Therefore, averages give only contradicting results.

1.5 Key-Words

1. Statistics : Statistics is the study of the collection, organization, analysis, interpretation, and presentation of data. It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments.
2. Statistical method : A method of analyzing or representing statistical data; a procedure for calculating a statistic.

1.6 Review Questions

1. Define statistics and explain its importance.
2. Examine the important definitions of statistics. Which in your opinion, is the best ?
3. Discuss the scope of the study of this science.
4. Explain the use of statistics for economic analyses and planning.
5. Discuss the limitations of statistics.

Answers: Self-Assessment

1. (*i*) Two, singular, plural (*ii*) Quantitative, qualitative
   (*iii*) Data (*iv*) Descriptive and inductive statistics
   (*v*) Statistical methods.

1.7 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 2: Types of Data Collection: Primary and Secondary, Methods of Collecting Primary Data

CONTENTS

Objectives

After reading this unit students will be able to:

- Define Primary and Secondary Data.
- Discuss the Methods of Collecting Primary Data.

Introduction

On the basis of the method and source by which the data is collected the data is classified into two types: (1) Primary Data, (2) Secondary Data. Primary data are collected by the investigator first hand, for the first time and are, therefore, original. Secondary data, on the other hand, which have already been collected by some other people or agency may be for some other type of enquiry, in this way it can be seen that secondary data are primary for some. Primary data are in the shape of raw material whereas secondary data are in the shape of finished products.

2.1 Primary Data and Secondary Data

Primary Data

Primary data are those data which are collected directly from the individual respondents for the first time by the investigator for certain purpose of study. They are but the raw materials for an investigation. Primary data are original in character in the sense that they have been recorded as they occurred without having being grounded at all. They simply relate to the collection of original statistical information. They are also current and fresh.

For example, the data collected by National Sample Survey Organization (NSSO) and Central Statistical Organization (CSO) for various surveys are primary in character. If an experiment is conducted to know the effect of a drug on the patients, the observations taken on each patient constitute the primary data. Primary data are collected when fresh data are needed and also when no other statistical information are available.

Notes

Merits

Primary data are of the following merits:

1. They are more accurate.
2. They are reliable.
3. They are suitable for sampling inquiry.
4. They are original in character.
5. They are the latest or current information.

Demerits

Primary data are of the following demerits:

1. They are inconvenient.
2. They are expensive which involves a great amount of planning and supervision.
3. They consume more time and energy.
4. They may prove inaccurate if the enumerators are not trained well.
5. Large number of investigators are needed.
6. There may be personal bias and prejudices.
7. If the respondents are non-responsive, the quality of the collected information suffers a lot.

Secondary Data

Secondary data are those data which have already been collected by somebody for others for some other purposes. They are in the finished form of the investigation.

For example, if the statistical data given in different population census years are again processed to obtain trends of population growth, sex ratio, mortality rate, etc. it is termed secondary data.

Secondary data are collected when adequate and authentic statistical information are already available and when there is waste of time and money to collect fresh statistical information.

Merits

Secondary data are of the following merits:

1. They require only a minimum cost for the collection.
2. They can be used for a quick survey or investigation.
3. They save time, money and energy.
4. The errors occurred can be easily eliminated by the primary investigator.

Demerits

Secondary data are of the following demerits:

1. They may be outdated information.
2. They may have no accuracy of data.
3. They are only secondary in character.
4. In some cases, all particulars may not be available.

Did u know? Secondary data are secondary in character in the sense that those statistical information which have already been processed to a certain extent for a certain purpose. They are expressed in totals, averages or percentages.

Notes

Sources of Secondary Data

The main sources of secondary data are of two categories:

Published sources

Unpublished sources

Published Sources

The various government agencies, international bodies and local agencies generally publish sources of data which are secondary in character. The following are some of the important published sources of secondary data:

1. Official publications brought out by the Central, State and Local Governments such as Pay Commission Reports, Indian Population Census Reports, etc.
2. Official publications brought out by the international bodies like International Monetary Fund (IMF), International Bank for Reconstruction and Development (IBRD), United Nations Organization (UNO), International Labour Organization (ILO), etc.
3. Reserve Bank of India's publications of Bulletin, namely "RBI Bulletin" and the official records of the nationalized commercial banks and the State Bank of India.
4. Publications of Trade Unions, Chambers of Commerce and Industry such as Federation of Indian Chambers of Commerce and Industry, Trade Bulletins issued by Stock Exchanges and large business houses.
5. Publications brought out by individual research scholars, Research centres, reports submitted by economists and statistical organizations.
6. Publications brought out by the National Sample Survey Organization (NSSO) and Central Statistical Organization (CSO).

Unpublished Sources

The following are some of the unpublished sources of data which are secondary in character:

1. Records of government offices and business concerns such as account books.
2. Research undertaken, by research institutions, scholars, site.
3. Unpublished sources of data are also available with Trade Unions, Chambers of Commerce, Labour Bureaus, etc.
4. Central Bureau of Investigation (CBI) Records.
5. District Collectorate Office Records.

Precautions in Using Secondary Data

Secondary data should be used only after careful examination of the data. It is because of the fact that the secondary data may sometimes be unsuitable, inadequate, inaccurate, unreliable and incomparable to serve the purpose of the the present investigation or inquiry.

The following are some of the precautions in using secondary data:

1. The suitability of the data available for the purpose of the present inquiry should be ascertained.
2. The adequacy of the data available for the present analysis should be ascertained.
3. The degree of accuracy desired and actually achieved should also be taken into consideration.
4. The degree of reliability of secondary data is to be assessed from the source, the compiler and his capacity to obtain current statistics for the purpose of interpretation.
5. The question of comparability of the data over a period of time should be assured.
6. The secondary data should be used only when the scope and object of the present inquiry are commensurate with that of the original inquiry.

Notes

7. The definition of units in which data are expressed should be kept stable in the present inquiry also.
8. While using secondary data, general information should also be obtained regarding the type of investigators, editors and tabulators employed in the primary data collection method.
9. The sources and the methods used should be clearly known.

2.2 Methods of Collecting Primary Data

There are various methods of collecting primary data. They are:

(1) Direct Personal Investigation: As the name itself suggests, in this method information is collected personally from the sources concerned. Under this, the investigator personally interviews everyone who is in a position to supply the information he requires. The investigator must possess the following qualities so that genuine data may be collected. *Firstly*, the investigator must be totally unbiased. *Secondly*, the investigator must have the tact of sustaining calm and unflustered enquiring environment so that the truth can be revealed. *Thirdly*, the investigator must have a sociable and pleasing personality so that the interviewers do not run away from him. Fourthly, the interviewer should be neutral in matters of clan, sex, race, religion etc.

Merits of this method: (i) Original data are collected, (ii) Correct and required information is gathered, (iii) The personal presence of the investigator helps in keeping the flexibility in the enquiry depending upon the type of respondent, (iv) Problem of no-answers is solved to a great extent *i.e.*, reluctance of response due to not understanding the question etc. can be avoided. (v) Uniformity in data collection is maintained.

Demerits of this method: (i) Requires a lot of time and personal presence of investigator every time, (ii) The method is very costly due the above reason, (iii) Cannot be used ideally in cases of wide investigating field, (iv) Too much dependence on the skills of investigator, (v) Not suitable when respondents are reluctant to reveal the truth when approached directly.

This method of direct personal investigation is suitable only for intensive investigations.

(2) Indirect Oral Investigation: Sometimes, the requirement of information is such that people are reluctant to answer when approached directly, and sometimes it is not possible for the investigator to be present personally with each respondent, it such is case, method of indirect oral in investigation can be used to collect the primary data. However, in order to ensure genuine data, it is essential that only those persons should be interviewed who: (i) possess full knowledge, (ii) are capable of expressing themselves, (iii) are not prejudiced, (iv) are rational.

In this method, a small list of questions is prepared and is put to different people (known as witnesses) and their answers are recorded. There should be no motivation to give colours to the facts.

Merits of this Method: (i) A wide field may be brought under investigation, (ii) This method saves time and labour and hence is less costly, (iii) There is no need to depend too much on the personal skills of the investigator, (iv) It is easier to deal with all aspects of the problem.

Demerits of this Method: (i) Great care and vigilance is needed in assessing the correct value of information collected, (ii) Due allowance needs to be made for the conscious and unconscious bias of the persons giving information, (iii) There is a possibility that the witnesses colour the information according to their interests, (iv) Dependence of local correspondents increases when the field of enquiry is wide.

(3) Local Reports: This method is generally used by the news-papers, periodicals, news channels etc. The government also collects information about prices, agricultural production etc. by this method.

Merits of this Method: (i) A very wide geographical area can be covered, (ii) Information on specific issues can be obtained, (iii) Regular flow of information over a long period of time can be obtained.

Notes

Demerits of this Method: (i) Reliability of the information is questionable, (ii) Correspondent's personal bias may come in, (iii) Same information with different attitudes may look different, (iv) Chances of errors are many, as the correspondents are not personally interested in the problem.

(4) Schedules and Questionnaires Method: This method is usually used by private individuals, research workers, non-official institutions and even the Government. In this method, a list of questions is prepared and information is collected about the problem. This can be done by —

(a) distributing questionnaires to the persons knowing about the information.

(b) sending the information by post or e-mail to the people from whom information is to be collected.

(c) using enumerators to send the questionnaire to the people, who also help them in filling the questionnaire. This method is adopted when there is a possibility of language problem or the respondent is illiterate or when there is probability of avoidance of answering by the respondents.

Merits of this Method: (i) Wide area of investigation can be covered, (ii) The Method is simple and cheap, (iii) Can be used with least expenses for geographically dispersed respondents, (iv) Original data is collected, (v) Information is given by the respondents themselves, hence the data is free from the bias of the investigation.

Demerits of this Method: (i) Possibility of no-response is quite high, (ii) This method can be used successfully where the respondents are educated, (iii) Information given by the respondents may be false, (iv) Clarification of the questions, supplementary and complimentary questions etc. is not possible, hence the method is inflexible.

From the above, it is quite clear that none of the methods is free from one or the other drawback. In fact, the method to be chosen depends upon the nature of investigation, object and scope of enquiry, budget made for the purpose of data collection, degree of accuracy desired and the time within which the data has to be collected.

Questionnaire Method Or Essentials of a Good Questionnaire

The questionnaire method is the method in which primary data is collected by distributing a list of questions related to the probe to those who are supposed to have knowledge about the problem. In this way, it can be said that, the success of the Statistical enquiry in this case depends, to a large extent on the questionnaire.

Preparation of questionnaire is a highly specialised job. Although, there are no hard and fast set rules to prepare a questionnaire. However, a few broad principles should be followed in order to have a good questionnaire. They are:

1. The questionnaire should be started with a covering letter which should be written in a polite language requesting the respondents to answer to the best of their knowledge. The letter should emphasize the need and usefulness of the information that is being collected. The letter should also ensure that the information obtained from them shall be strictly used only for the said purpose and the information and name of the respondents shall be kept confidential.

   The covering letter may also accompany some small gift etc. to create the acceptance among the respondents so that there is a greater chance of getting a response. Moreover, the letter may also give a promise, that if the respondents so desire, a copy of the results of the survey may be sent to them. This would increase the credibility of the investigator/investigating institution.

2. The questionnaire should not be very long. Unnecessary details in the form of separate questions must be avoided. Although, there is no hard and fast rule about what should be the number of questions in a questionnaire. Much shall depend upon the problem undertaken. However, efforts should be made to frame only relevant questions otherwise, the respondents feel bored or feel answering them to be a waste of time, and correct information will be a casualty.

Notes

3. The questions should be framed in a simple way and in easy language. They should be capable of a straight answer. As far as possible, the questions should be capable of objective answers. A set of possible answers may be accompanied with each question so that the respondents feel easy to give the answers. Questions with 'yes' or 'no' answers are also useful.
4. Asking personal questions should be avoided because it is quite likely that these questions are not answered correctly. For example, perks received, income tax paid etc.
5. Questions which tend to hurt the sentiments of the respondents must be avoided. For example, private-life litigation, indebtness, etc.

   In both these cases it is quite likely that there will be either no-response or the response shall be false.
6. Corroboratory questions should be incorporated in a good questionnaire. These are the questions which are meant for cross-checking the answers given by the respondents for earlier question.
7. Unless and until it is very essential, questions whose answering requires calculations should not be asked. For example, what per cent of your income is spent on your children's schooling will need a series of calculation. And it is quite likely that the respondents does not give an accurate answer.
8. The questionnaire should be attractive and impressive. There should be sufficient space for answering the questions, the quality of paper used and printing on the paper should be good. It always helps if it is so.

Which method is the best in Collecting Primary Data ?

None of the methods can be termed to be best or worst. Following considerations are taken into account while selecting the method which should be used to collect primary data —

(1) **The nature of investigation:** If it is essential to establish personal contacts, 'direct personal investigation' will be appropriate. But if the number of respondents is large and they are educated also, questionnaire method shall be better. But if the area covered is very wide and information is to be gathered on a number of subjects, using enumerators shall be better.

(2) **Object and scope of enquiry:** If the scope of enquiry is limited and is of confidential nature, 'direct personal investigation' should be done. But if the scope extends to a number of subjects, use of questionnaire or enumerators can be made.

(3) **Budget:** If financial resources are strong, personal investigation can be carried out. But if a wide survey is to be done with limited financial resources, questionnaire method should be chosen.

(4) **Degree of accuracy desired:** Highest degree of accuracy is achieved from direct personal investigation and the accuracy is least in case of information collected from correspondents. Now, on the basis of budget and other above requirements, the method can be chosen.

(5) **Time factor:** A large amount of information can be obtained in minimum time by using enumerators and/or correspondents. If there is long-time available, 'direct personal investigation' may be done.

Self-Assessment

1. Fill in the blanks:

(*i*) Secondary data may be ..... or .....

(*ii*) Before finalising a questionnaire ..... is done.

(*iii*) ..... method is the cheapest method of collecting primary data.

(*iv*) In ..... method, much depends upon the skills of the investigator.

(*v*) Least accuracy of information is likely in case of information from the ..... .

Notes

2.3 Summary

- Primary data are those data which are collected directly from the individual respondents for the first time by the investigator for certain purpose of study. They are but the raw materials for an investigation.
- Primary data are original in character in the sense that they have been recorded as they occurred without having being grounded at all. They simply relate to the collection of original statistical information. They are also current and fresh.
- Secondary data are those data which have already been collected by somebody for others for some other purposes. They are in the finished form of the investigation.
- Secondary data are secondary in character in the sense that those statistical information which have already been processed to a certain extent for a certain purpose. They are expressed in totals, averages or percentages.
- Secondary data are collected when adequate and authentic statistical information are already available and when there is waste of time and money to collect fresh statistical information.
- Secondary data should be used only after careful examination of the data. It is because of the fact that the secondary data may sometimes be unsuitable, inadequate, inaccurate, unreliable and incomparable to serve the purpose of the the present investigation or inquiry.
- While using secondary data, general information should also be obtained regarding the type of investigators, editors and tabulators employed in the primary data collection method.
- As the name itself suggests, in this method information is collected personally from the sources concerned. Under this, the investigator personally interviews everyone who is in a position to supply the information he requires. The investigator must possess the following qualities so that genuine data may be collected. *Firstly*, the investigator must be totally unbiased. *Secondly*, the investigator must have the tact of sustaining calm and unflustered enquiring environment so that the truth can be revealed. *Thirdly*, the investigator must have a sociable and pleasing personality so that the interviewers do not run away from him. Fourthly, the interviewer should be neutral in matters of clan, sex, race, religion etc.
- Sometimes, the requirement of information is such that people are reluctant to answer when approached directly, and sometimes it is not possible for the investigator to be present personally with each respondent, it such is case, method of indirect oral in investigation can be used to collect the primary data.
- This method is usually used by private individuals, research workers, non-official institutions and even the Government. In this method, a list of questions is prepared and information is collected about the problem.
- Using enumerators to send the questionnaire to the people, who also help them in filling the questionnaire. This method is adopted when there is a possibility of language problem or the respondent is illiterate or when there is probability of avoidance of answering by the respondents.
- The questionnaire method is the method in which primary data is collected by distributing a list of questions related to the probe to those who are supposed to have knowledge about the problem. In this way, it can be said that, the success of the Statistical enquiry in this case depends, to a large extent on the questionnaire.
- The questionnaire should be started with a covering letter which should be written in a polite language requesting the respondents to answer to the best of their knowledge. The letter should emphasize the need and usefulness of the information that is being collected. The letter should also ensure that the information obtained from them shall be strictly used only for the said purpose and the information and name of the respondents shall be kept confidential.

Notes

- The covering letter may also accompany some small gift etc. to create the acceptance among the respondents so that there is a greater chance of getting a response. Moreover, the letter may also give a promise, that if the respondents so desire, a copy of the results of the survey may be sent to them. This would increase the credibility of the investigator/investigating institution.
- The questions should be framed in a simple way and in easy language. They should be capable of a straight answer. As far as possible, the questions should be capable of objective answers. A set of possible answers may be accompanied with each question so that the respondents feel easy to give the answers. Questions with 'yes' or 'no' answers are also useful.
- The questionnaire should be attractive and impressive. There should be sufficient space for answering the questions, the quality of paper used and printing on the paper should be good. It always helps if it is so.
- If it is essential to establish personal contacts, 'direct personal investigation' will be appropriate. But if the number of respondents is large and they are educated also, questionnaire method shall be better. But if the area covered is very wide and information is to be gathered on a number of subjects, using enumerators shall be better.
- Highest degree of accuracy is achieved from direct personal investigation and the accuracy is least in case of information collected from correspondents. Now, on the basis of budget and other above requirements, the method can be chosen.

2.4 Key-Words

1. Primary Data : Primary research consists of a collection of original primary data. It is often undertaken after the researcher has gained some insight into the issue by reviewing secondary research or by analyzing previously collected primary data. It can be accomplished through various methods, including questionnaires and telephone interviews in market research, or experiments and direct observations in the physical sciences, amongst others.
2. Secondary Data : Secondary data, is data collected by someone other than the user. Common sources of secondary data for social science include censuses, organisational records and data collected through qualitative methodologies or qualitative research. Primary data, by contrast, are collected by the investigator conducting the research.
3. Secondary Data : analysis saves time that would otherwise be spent collecting data and, particularly in the case of quantitative data, provides larger and higher-quality databases that would be unfeasible for any individual researcher to collect on their own. In addition, analysts of social and economic change consider secondary data essential, since it is impossible to conduct a new survey that can adequately capture past change and/or developments.

2.5 Review Questions

1. What preliminary steps ought to be taken by a statistician before starting on with the task of collection of data ?
2. To make collection of data smooth and result-oriented, it is essential for a statistician to carry out some preliminary steps. Justify the statement giving details about these steps.
3. Describe the questionnaire method of collecting Primary data. State the essentials of a good questionnaire.
4. What are the essentials of a good questionnaire ?
5. What are the various methods of collecting statistical data ? Which of these is most suitable and why ?

Notes

Answers: Self-Assessment

1. (*i*) published and unpublished, (*ii*) pre-testing,
   (*iii*) Questionnaire, (*iv*) Direct personal investigation,
   (*v*) correspondents.

2.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 3: Classification and Tabulation of Data: Frequency and Cumulative Frequency Distribution

CONTENTS

Objectives

After reading this unit students will be able to:

- Describe the Classification and Tabulation of Data.
- Understand Frequency Distribution.
- Explain Cumulative Frequency Distribution.

Introduction

The data collection leaves an investigator with a large mass of information. But it is the weakness of human mind that it fails to assimilate a lot of things or information at a time. To remove this difficulty and to make the large mass of data useful to its fullest, classification and tabulation of data is done. By doing so the data are presented in condensed form which helps in making comparisons, analysis and interpretations. Moreover, classification and tabulation segregates the likes from the unlikes. The heterogeneity is removed. The data are classified into classes and sub-classes according to their characteristics. This process is called **classification**. The classified data are presented in prescise and systematic tables. This process is called **tabulation**. By these two processes, the data collected are made simple, easy to understand and carry out analysis and interpretations.

3.1 Classification

Meaning and Definition of Classification

Classification may be defined as the process of arranging the available data in various groups or classes in accordance with their resemblances and similarities and keeping in view some common features and objectives of study. Thus, through classification, an effort is made to achieve homogeneity of the collected information. While classifying, the units with common characteristics are placed together and in this way the whole data is divided into a number of classes and sub-classes. It may be argued that the data collected is as per the requirement, and it is in general homogeneous in nature, then how does classification help ? For example, if a study on 500 students is to be carried out then, the data is homogeneous as it is about students. But this information on 500 students may be classified in terms of different hostels and universities they are coming from, different areas they come from, different subjects they have opted for, and so on. Only by carrying out the one, classification, will the invesitgator be in a position to compare, analyse and interpret the above data.

Notes

Definition

According to **Conner**, "Classification is the process of arranging things (either actually or notionally) in groups or classes according to their resemblances and affinities and gives expression to the unity of attributes that may subsist amongst a diversity of individuals."

From the above definitions it may be said that a group or class has to be determined on the nature of the data and the purpose for which it is going to be used. For example, the data on household may be classified on the basis of age, income, education, occupation, expenditure etc.

Did u know? "Classification is the process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts."

Features of Classification

On the basis of the above dicussion, the chief features of classification can be summarised as under— (i) Classification may be according to attributes, characteristics or measures, (ii) The basis of classification is unity in diversity, (iii) Classification may be actual as notional.

Chief Objects of Classification

Classification helps the investigator and the investigation in a number of ways. The following are the objects of classification. These objects also suggest the *importance of classification*.

(1) **Classification presents the facts in simple form:** The process of classification helps in arranging the data in such a way that the large mass of irrelevant looking data becomes simple and easy to understand, avoiding unnecessary details, making logical sense.

(2) **Classification points out similarities and dissimilarities clearly:** Since classification is done on the basis of characteristics and similarity of data, it helps the investigator in poining out clearly the similarities and dissimilarities so that they can be easily grasped.

(3) **Classification facilitates comparison:** By classification of data, comparison becomes easier, inferences can be drawn logically and confidently and facts can be located with much ease.

(4) **Classification brings out relationship:** The cause—and effect relationship can be located with the help of classification.

(5) **Classification prepares basis for tabulation:** The importance of classification also lies in the fact that it prepares the ground on the basis of which tabulation can be done.

Methods of Classification

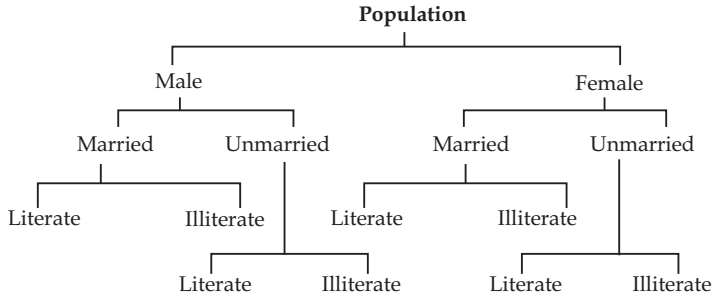
The methods of classification are divided broadly into four types:

(I) **Qualitative Classification:** Here, classification is done in accordance with the attributes or characteristics of the data. Such classification is generally done where data cannot be measured. Under this classification method, the presence or absence of an attribute is the basis of classification. Qualitative classification can be done in two ways:

(1) **Two-fold or Dichomous Classification:** This type of classification is based on the presence or absence of an attribute and the data gets classified in two groups/classes—*one*, possessing that attribute and *two*, not possessing that attribute. For example, on the basis of marital status, the data can be divided into *two* classes, one, married and two, unmarried. On the basis of literacy there can be two classes, one, literate other non-literate.

(2) **Manifold Classification:** Here, the bases of classification are manifold, *i.e.*, more than one attribute. In this method, classes/groups are further classified into sub-classes and sub-groups. For example, population/sample is first classified on the basis of sex, then for each sex (male or female) marital status forms two sub-classes then further these sub-classes are classified as per their literacy state. This can be explained in simple was as below:

Notes



(II) Quantitative Classification: In this method the data are classified on the basis of variables which can be measured for example, age, income, height etc. This kind of classification is done in the form of statistical series. For example, 100 students can be classified in terms of mark obtained by them. This is shown as below:

Marks Obtained	Number of Students
0-20	15
20-40	25
40-60	35
60-80	15
80-100	10

(III) Geographical Classification: In this case, data are classified on the basis of place or location. For example, population is shown on the basis of various states, or students are classified on the basis of the place they belong to etc. Series which are arranged on the basis of place or location are called spatial series.

Name of State	Per capita Income
Punjab	331
Kerala	310
Madhya Pradesh	206
J & K	216
Haryana	320

(IV) Chronological Classification: This is done by arranging the data with respect to time, such series are known as time series. For example,

Year	Sales made (in Rs. crore) by Company 'X'
1995	11,800
1996	12,600
1997	11,200
1998	16,800
1999	17,000
2000	18,200
2001	19,100
2002	20,000
2003	21,000

Notes

2004	21,800
2005	22,200
2006	23,000
2007	25,200
2008	24,600
2009	26,000

On the basis of the above, it may be concluded that depending upon the nature of data, and requirement of the investigation and objectives of study, classification of data facilitates the investigator to compare, analyse and interpret the data, thus helping in using the data scientifically.

3.2 Tabulation of Data

The process of presenting data in the tabular form is termed as tabulation. As per **L. R. Connor**, "Tabulation involves the orderly and systematic presentation of numerical data in a form designed to elucidate the problem under consideration.

Importance of Tabulation

(1) **Simplifies the complex data:** The process of tabulation eliminates unnecessary details and present the complex data concisely in rows and columns. This helps in simplifying the complex data which becomes more meaningful and better understood.

(2) **Presents facts in minimum space:** A large number of facts can be condensed in one table in a much better way than otherwise.

(3) **Facilitates comparison:** Data when depicted in rows and columns, facilitates comparison, and the problem can be better understood.

(4) **Depicts data characteristics:** The important characteristics of data are brought about by the process of tabulation as it is presented concisely but clearly.

(5) **Depicts trends and pattern of data:** Data, in the form of tables, helps in understanding the trends and patterns lying within the figures without much effort. This facilitates better understanding of the problem under study.

(6) **Helps in making references:** Data can be stored perfectly in the form of tables which can be easily identified by its head and footnotes. This can be used for future studies.

(7) **Facilitates statistical analysis:** It is only possible after tabulation, that the data can be subjected to statistical analysis and interpretation. Measures of correlation, regression dispersion etc can be easily calculated when the data is in tabular form.

The above points form the advantages of tabulations to the investigator and investigation as well.

Limitations of Tabulation

Although tabulation is an essential activity in the process of statistical analysis, it is not absolutely free of limitations. The limitations are:

(1) A table does not present any description about the figures expressed. For those who are not familiar, it is not easy to understand facts with the help of tables.

(2) Specialised knowledge is essential to understand a table. It is not a layman's cup of tea.

(3) A table does not lay emphasis on any section of particular importance.

It is because of these limitations that tables are only complementary to textual report. A table only accompanies a text, facilitating better understanding in a concise way.

Essential Parts of a Table

Notes

In order to be complete and most informative, a table should have the following parts:

(1) **Table number:** This is the number by which, the table can be identified and can be used for reference in future. The table number can be given at the top or at the bottom.

(2) **Title of the table:** A title is a brief statement indicating about the nature of the data and the time-span to which the data relate. The geographical distribution of the data, if any, should also be indicated in the title. The title should be in dark letters in comparison to other heads and sub-heads in the table.

(3) **Captions:** A caption is the main heading of the vertical columns. A number of small sub-headings are followed by the caption. Caption should be given in unambiguous language and should be placed at the middle of the column (as shown in the table given below).

(4) **Stubs:** The stubs are the headings of the horizontal rows and are written on the extreme left of the table. The number of stubs depend upon the nature of the data.

(5) **Head note:** The head note refers to the data contained in the major part of the table, and it is placed below the title of the table. It is generally put in brackets. For example, (in percentage) or (in *kg*.) etc.

(6) **Footnote:** They are given at the bottom of the table and are used to clarify about heading, title, stubs and caption etc. It may also be used to give further explanation about the data, or certain terms or figures used in the table. Footnotes are also used to describe the source of the data, if the data is secondary in nature. For example, "the figures in bracket show the per cent rise over previous year" or "the profit above is after tax" etc are some of the information which is given in footnotes.

(7) **Body of the table:** This is the most important and inevitable part of the table which contains the statistical data which have to be presented. The data is arranged in captions and stubs.

Table No.

Title of the Table

Head note

Sub-heading	⟵ Caption ⟶			
	Col. head	Col. ead	Col. head	Col. head
↑ Stub-entires ↓	B	ODY		

Footnotes

Source.

From the above, it becomes very clear that it is only with the help of tabulation, that statistical enquiry can be scientifically carried out. It helps the investigation and the investigator.

A table presents the statistical data in a systematic way in rows and columns which concisely explain the numerical facts. Tabulation is nothing but the process of preparing a table. The preparation of table is a specialised activity and is done through a set of rules.

Rules for Tabulation

Although there are no hard and fast rules regarding preparation of tables as pointed out by Bowley, common sense and experience are a prerequisite while tabulation, however, some general rules may prove to be handy while carrying out tabulation process. According to **Harry Jerome**, "A good statistical table is not mere careless grouping of columns and rows of figures: it is a triumph of

Notes

ingenuity and technique, a masterpiece of economy of space combined with a maximum of clearly presented information. According to **W. M. Harpen**, "The construction of a table is in many ways a work of art." The rules regarding tabulation are not hard and fast, but prove as a guide in tabulation. These rules are divided into two groups:

(A) Rules relating to Table Structure: This includes the rules explaining the preparation of structure of the table. This group includes the following rules:

(1) Table must always have a table number, so that it can be easily identified or can be referred to whenever required.

(2) The table must always have a relevant title, which should be clear, concise and self-explanatory. The title should explain about: (a) the subject area, (b) data or period to which the data belong to, (c) basis and principles used in classification of data, (d) the field to which the data relate to.

(3) The stub and caption should be clear and as brief as possible. Columns should be numbered.

(4) Neat and tidy appearance must be given to the table which can be done by providing proper ruling and spacing as is necessary. If the table continues to the next page, no bottom line must be drawn, as it would indicate the end of the table. Major and minor items must be given space according to their relative importance. Coloured inks, heavy printed titles or sub-titles, thick and thin ruling etc must be used to clarify a complex table.

(5) Use of averages, sub-totals, totals etc must be made if the data so require. In case, it is required, the table should contain sub-totals for each separate classification of data and a general total for all combined classes. For example, data about cost break-up of a particular production process shall require the above. But the data on annual percentage rise in bonus of employees may not require use of these sub-totals and totals.

(6) The body of the table must be as comprehensive as possible, consistent with the purpose. Unnecessary details must be avoided and, items in 'miscellaneous or unclassified columns must be least.

(7) The items in the body of the table must be arranged in some systematic order. Depending upon the type of data and purpose of enquiry, data may be arranged: (a) alphabetically, (b) geographically, (c) progressively, (d) chronologically, (e) ascending or descending order, (f) conventionally or (g) in order of importance etc.

(8) If further clarification about some figures, sub-heads etc is required, it must be given in the footnotes. The important limitation of data can also be specified here if the need is felt by the person who is tabulating.

(9) The source of the data must be specified if the data is secondary.

(10) The units of measurement, if common, must be indicated in the head note, otherwise under each heading and sub-heading.

(B) General Rules: Other than the rules relating to table structure, there are certain rules which should be followed while tabulation, so that, the tabulation can be accomplished successfully. They are:

(1) Table should be precise and easy to understand.

(2) If the data are very large they should not be crowded in a single table. However, it is essential that each table is complete in itself.

(3) The table should suit the size of the paper. The width of the columns should be pre-decided giving due consideration to this.

(4) Those columns whose data are likely to be compared should be preferably kept side-by-side.

(5) Percentages, averages, totals etc must be kept close to the data.

(6) The figures must be approximated to one or two decimals. This must be specified in the footnote.

Notes

(7) 'Zero' quantity must be indicated separately, and in case of unavailability of a particular figure 'NA' (not available) must be indicated clearly. 'Zero' is not equivalent to 'Not available'.

(8) Abbreviations should be avoided. But if the need so arises, it must be clarified in footnotes.

(9) The tabulation should be explicit. Words like 'etc'. must not be used.

(10) The table should be of manageable size.

Notes Tabulation is an art which requires common sense in planning a table and viewing the proposed table from the point of view of the user or the other person. Tabulation is done keeping in mind the purpose of statistical investigation. The rules of tabulation act as guides in preparing a good table.

Seriations of Data

Quantitative classification data is done through seriation of data. If two variable quantities are arranged side by side so that the measurable differences in the one correspond to the measurable differences in the other, the result is formation of a statistical series. For example, marks obtained by a class of students. Here, there are two elements one, the variable (marks) and *two*, the frequency (of students). The number of times a particular variable has repeated is noted down and the total is the frequency of that class. For example, the marks obtained by 25 students out of 10 in a particular subject is as follows:

Marks Obtained

2	4	8	6	8
4	10	6	4	4
8	2	4	6	10
2	6	10	2	6
6	2	6	2	2

On counting how many students obtained 2, 4, 6, 8 and 10 we get the frequencies:

Marks	Tallies	Frequency
2	𝍸 ||	7
4	𝍸	5
6	𝍸 ||	7
8	|||	3
10	|||	3
Total		**25**

Notes

So the discrete data of marks obtained by 25 students is:

Marks Obtained	Number of Students
2	7
4	5
6	7
8	3
10	3

Formation of Continuous Frequency Distribution

In continuous frequency distribution data are divided into class intervals instead of individual values (as is done in case of discrete frequency distribution) class intervals can be formulated like marks from 0 to 10, 10 – 20 and so on. Here the magnitude of class interval is 10 marks. (20 – 10 = 10). But it can be lower, for example, 0 – 5, 5 – 10 and so on or higher, like 0 – 20, 20 – 40 and so on.

Suppose, the marks obtained of 100 students is to be given in continuous series, it can be done as below:

Marks Obtained	Number of Students
0–10	14
10–20	26
20–30	30
30–40	20
40–50	10
	100

This means 14 students obtained marks between '0' (zero) to 10. The marks may lie in any fraction 1/4, $3\frac{3}{4}$, $9\frac{3}{4}$, 9.99 or 10, and likewise. These are called exclusive class intervals.

Sometimes, the data is given as below:

Marks Obtained	Number of Students
10–19	4
20–29	6
30–39	10
40–49	10
	30

This is called inclusive class arrangement. In this case, a question arises as to in which class the student getting 9.5 or 29.5 must be placed ? In such a case to ensure continuity following adjustments must be made:

Marks obtained	Number of Students
9.5–19.5	4
19.5–29.5	6
29.5–39.5	10
39.5–49.5	10
	30

Notes

Difference between Classification and Tabulation

Some important points or differences between classification and tabulation are:

(1) Although both are necessary in statistical investigations, classification is done first and it forms the basis for tabulation.

(2) Tabulation is a mechanical function of classification. But classification is not a mechanical function.

(3) In classification, data are divided/classified in different classes as per similarities and dissimilarities. Under tabulation this classified data is put in rows and columns.

(4) Classification involves analysis of data. Tabulation is the process of presenting the data.

Types of Tables

Following are the important types of tables:

(1) One-way Table: This supplies information about only one characteristic. For example, marks obtained by 100 students can be illustrated in one-way table as below:

Marks Obtained	Number of Students
0–10	14
10–20	26
20–30	30
30–40	20
40–50	10
	100

(2) Two-way Table: If the information of two related characteristics is to be given, it is done by two-way tables. Suppose, in the above example, the students are to be classified in males and females the data can be re-written as below:

Marks Obtained	Number of Students		
	Male	Female	Total
0–10	4	10	14
10–20	16	10	26
20–30	15	15	30
30–40	12	8	20
40–50	6	4	10
Total	**53**	**47**	**100**

(3) Three-way Table: In the above example, the male and female students can be further divided into hostellers and day-scholars, thus providing information about three different characteristics. This is done by a three-way table. The format of the three-way table in this case is given below:

Notes

Number of Students									
Marks		**Males**			**Females**			**Total**	
	Hostellers	Day-scholars	Total	Hostellers	Day-scholars	Total	Hostellers	Day-scholars	Total
0–10	3	1	4	6	4	10	9	5	14
10–20	10	6	16	6	4	10	16	10	26
20–30	9	6	15	10	5	15	19	11	30
30–40	5	7	12	3	5	8	8	12	20
40–50	2	4	6	1	3	4	3	7	10
Total	**29**	**24**	**53**	**26**	**21**	**47**	**55**	**45**	**100**

(4) Higher-order Tables: Tables giving information about more than three characteristics can also be made. They are called higher-order tables. Suppose, in the above example, marital status of the students is also to be informed, the table will have to be made as below:

Marks	**Number of Students**								
	Males			**Females**			**Total**		
	Married	Unmarried	Total	Married	Unmarried	Total	Married	Unmarried	Total
Hostellers									
0–10									
10–20									
20–30									
30–40									
40–50									
Day-Scholars									
0–10									
10–20									
20–30									
30–40									
40–50									
Total									
0–10									
10–20									
20–30									
30–40									
40–50									

(5) General Purpose Tables: They provide information for general use and reference. These are also called repository or reference tables.

Notes

(6) Special Purpose Tables: They are formulated to present some specific information relating to some specific subject under study. Such tables are also called text or summary tables.

Machine Tabulation

Tables are nor prepared with the help of machines which may be either hand-operated or are operated with electricity. Use of 'needle sorting' is one such machine for tabulation. Similarly 'Punch Cards' are also used. The work by these machines is more fast, easy and accurate.

Advantages of machine tabulation are:

(1) Greater accuracy.

(2) Less time required.

(3) Large-scale data can also be handled easily.

(4) Complex procedures like fitting trend lines etc becomes very easy with the help of computer.

(5) No work monotony can be avoided.

(6) Lowers cost by avoiding manual labour.

(7) The results are obtained without much waiting.

3.3 Frequency Distribution

The most important method of organising and summarising statistical data is by constructing a frequency distribution table. In this method, classification is done according to quantitative magnitude. The items are classified into groups or classes according to their increasing order in terms of magnitude and the number of items falling into each group is determined and indicated.

We shall discuss later questions such as how the classes are to be formed and how many classes are to be taken. We consider now how a frequency distribution table is to be constructed in the case of a discrete variable by taking a particular example.

Example 1 (a): Suppose that the marks secured by 60 students of a class are as follows:

46, 67, 23, 5, 12,	53, 38, 58, 26, 43,
36, 63, 26, 48, 76,	45, 66, 74, 16, 86,
56, 31, 58, 90, 32,	43, 36, 66, 46, 58,
36, 59, 54, 48, 21,	36, 64, 58, 45, 76,
58, 84, 68, 65, 59,	74, 48, 64, 58, 50,
46, 53, 64, 57, 65,	58, 95, 56, 66, 44.

Statistical Methods

Construct a frequency distribution table.

Marks obtained are divided into 10 groups or intervals as follows:

Marks below 10, between 11 and 20, between 21 and 30, and so on, between 91 and 100. Represent each mark by a tally (/), for example, corresponding to the mark 46 we put a tally (/) in the group 41 to 50; similarly we continue putting tallies for each mark. We continue upto four tallies and the fifth tally is put crosswise (\) so that it becomes clear at once that the lot contains five tallies, *i.e.* there are five marks. A gap is left after a lot of five tallies, before starting again to mark the tallies after each lot. The number of tallies in a class or group indicates the number of marks falling under that group. This number is known as the frequency of that group or corresponding to that class interval. Proceeding in this way, we get the following frequency table.

Notes

Table 1: Frequency distribution of marks secured by 60 students.

Class interval	Tally	Frequency (No. of students securing marks which fall in the class interval)
0 to 10	|	1
11 to 20	||	2
21 to 30	||||	4
31 to 40	𝍸 ||	7
41 to 50	𝍸 𝍸 ||	12
51 to 60	𝍸 𝍸 𝍸	15
61 to 70	𝍸 𝍸 |	11
71 to 80	||||	4
81 to 90	|||	3
91and above	|	1
Total		**60**

We shall now consider construction of a frequency distribution table of a continuous variable.

Example 1 (a): The heights of 50 students to the nearest centimetre are as given below:

151, 147, 145, 153, 156,	152, 159, 153, 157, 152,
144, 151, 157, 147, 150,	157, 153, 151, 149, 147,
151, 147, 155, 156, 151,	158, 149, 147, 153, 152,
149, 151, 153, 150, 152,	154, 150, 152, 149, 151,
151, 154, 155, 152, 154	152, 156, 155, 154, 150.

Construct a frequency distribution table.

We form the classes as follows: 145-146, 147-148, 149-150, 151-152, 153-154, 155-156, 157-158, 159-160 and construct the following frequency table:

Table 2: Frequency distribution of heights of 50 students

Class interval (Height in *cm*)	Tally	Frequency (Number of students having height)
145-146	||	2
147-148	𝍸	5
149-150	𝍸 |||	8
151-152	𝍸 𝍸 𝍸	15
153-154	𝍸 ||||	9
155-156	𝍸 |	6
157-158	||||	4
159-160	|	1
Total		**50**

Notes

We have given heights in *cms* in whole numbers or heights have been recorded to the nearest centimetre. Thus a height of 144.50 or more but less than 145.5 is recorded as 145; a height of 145.5 or more but less than 146.5 is recorded as 146 and so on. So the class 145-146 could also be indicated by 144.5-146.5 implying the class which includes any height greater than or equal to 144.5 but less than 146.5; the class 147-148 could be indicated by 146.5-148.5, meaning the class which includes any height greater than or equal to 146.5 but less than 148.5. Following this convention, the classes could be represented as: 144.5-146,5, 146.5-148.5, and so on. The above frequency distribution should finally be represented as follows.

Table 3: Frequency distribution of heights of 50 students

Heights (in *cm*)	Frequency (Number of students)
144.5-146.5	2
146.5-148.5	5
148.5-150.5	8
150.5-152.5	15
152.5-154.5	9
154.5-156.5	6
156.5-158.5	4
158.5-160.5	1
Total	**50**

Class intervals, Class limits and Class boundaries

The interval defining a class is known as a *class interval*. For Table: 2 145-146, 147-148, . . . are class intervals. The end numbers 145 and 146 of the class interval 145-146 are known as *class limits*; the smaller number 145 is the *lower class limit* and the larger number 146 is the *upper class limit*.

When we refer to the heights being recorded to the nearest centimetre and consider a height between 144.5.and 146.5 (greater or equal to 144.5 but less than 146.5) as falls in that class and the class is represented as 144.5-146.5, the end numbers are called *class boundaries*, the smaller number 144.5 is known as *lower class boundary* and the larger number 146.5 as *upper class boundary*. The difference between the upper and lower class boundaries is known as the *width* of the class. Here the width is 146.5–144.5 = 2 *cm* and is the same for all the classes. The common width is denoted *by c:* here $c = 2$ *cm*. Note that in certain cases, it may not be possible to have the same width for all the classes (specially the end classes).

Note also that the upper class boundary of a class *coincides* with the lower class boundary of the next class; there is no ambiguity: we have clearly indicated that an observation less than 146.5 will fall in the class 144.5-1465 and an observation equal to 146.5 will fall in the class 146.5-148.5.

3.4 Cumulative Frequency Distribution

Consider the number of all observations which are *less than the upper class boundary* of a given class interval; this number is the sum of the frequencies upto and including that class to which the upper class boundary corresponds. This sum is known as the *cumulative frequency* upto and including that class interval For example, consider Table 2; the cumulative frequency upto and including the class interval 145-146 is 2, that upto and including the next class interval 147-148 is 2 + 5 = 7, that upto and including the next class interval 149-150 is 2 + 5 + 8 = 15 and so on. This implies that two students have heights less than the upper class boundary of the class 145-146, seven students have heights less than the upper class boundary of the class 147-148 and so on. We can thus construct the cumulative frequency table as follows:

Notes

Table 4: Cumulative frequency (less than) table of heights of 50 students

Class (in *cm*) interval	Frequency	Cumulative Frequency (less than)
145-146	2	2
147-148	5	7
149-150	8	15
151-152	15	30
153-154	9	39
155-156	6	45
157-158	4	49
159-160	1	50
Total	**50**	

The cumulative frequency distribution is represented by joining the points obtained by plotting the cumulative frequencies along the vertical axis and the corresponding upper class boundaries along the *x-axis*. The corresponding polygon is known as cumulative frequency polygon (less than) or ogive. By joining the points by a freehand curve we get the cumulative frequency curve ("less than"). Similarly we can construct another cumulative frequency distribution ("more than" type) by considering the sum of frequencies greater than the lower class boundaries of the classes. For example, the total frequency greater than the lower class boundary 158.5 of the class 159-160 is one (1), while the total frequency greater than the lower class boundary 156.5 of the class 157-158 is 1 + 4 = 5, that of the class 155-156 is I + 4 + 6 = 11, and so on. Given below is Table 5 of cumulative frequency distribution ("more than") of the same distribution.

Table 5: Cumulative frequency (more than) table of heights 50 students

Class (in *cm*.) interval	Frequency	Cumulative frequency (more than)
145-146	2	50
147-148	5	48
149–150	8	43
151-152	15	35
153-154	9	20
155-156	6	11
157-158	4	5
159-160	1	1
Total	**50**	

The graph obtained by joining the points obtained by plotting the cumulative frequencies ("more than") along the vertical axis and the corresponding *lower* class boundaries along the *x*-axis is known as *cumulative frequency polygon* (greater than) or *ogive*. By joining the points by a free-hand curve, one gets the *cumulative frequency curve* ("more than" type).

Notes

These two curves are shown in Figure 1.

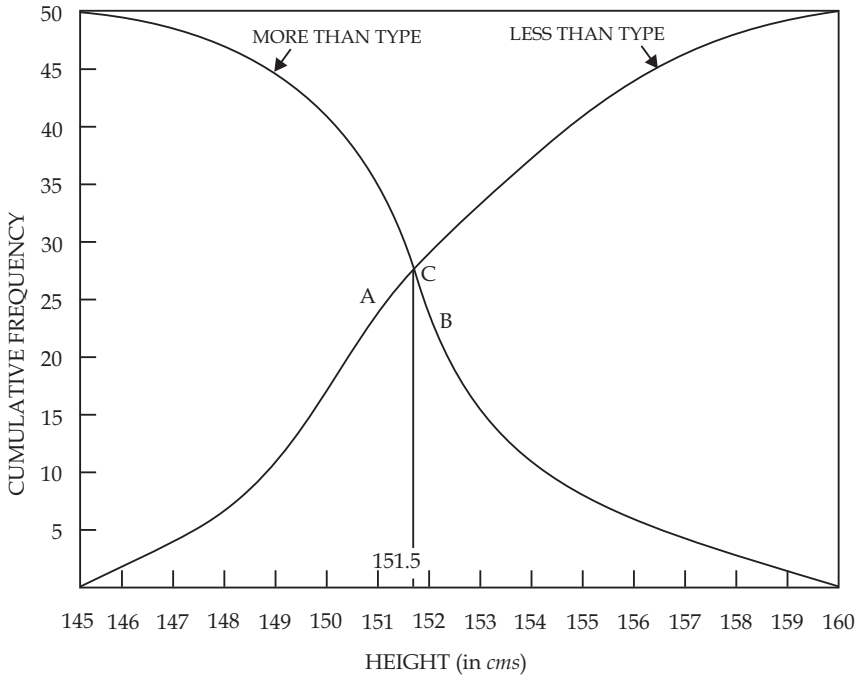


Figure 1: Cumulative frequency curves ('more than' and 'less than' types) of height 1 50 students

PRACTICAL QUESTIONS

1. Transform the following information in continuous series:

Weekly wages of 100 workers (in Rs.) of factory A								
880	230	270	280	960	940	930	860	990
820	240	240	550	990	950	860	820	360
960	390	260	540	1000	560	840	830	860
1020	480	270	260	1000	590	830	840	480
1040	460	300	270	1010	600	890	460	490
1060	330	360	300	1030	700	900	490	500
1040	360	370	400	1060	720	940	500	600
240	390	490	460	1070	760	960	460	670
260	780	500	440	460	790	990	360	680
290	670	560	990	480	800	1020	320	510

Solution: The lowest value is 1230 and highest is 1060. The difference in the highest and lowest value is 830. If we take a class interval of 100, nine classes would be formed. The first class should be 200-300 instead of 230-330.

Notes

Frequency Distribution

Wages	Tally bars	Frequency
200-300	𝍸 𝍸 |	13
300-400	𝍸 𝍸 |	11
400-500	𝍸 𝍸 𝍸 |	18
500-600	𝍸 𝍸	10
600-700	𝍸 |	6
700-800	𝍸	5
800-900	𝍸 𝍸 ||||	14
900-1000	𝍸 𝍸 ||	12
1000-1100	𝍸 𝍸 |	11
Total		**100**

2. If the class mid-points in a frequency distribution of age of a group of persons are 25, 32, 39, 46, 53 and 60, find: (i) the size of the class interval, (ii) the class boundaries and (iii) the class limits, assuming that the age quoted is the age completed on last birthday.

Solution: (i) The size of class interval

= Difference between the mid-values of any two consecutive classes

= 32 – 25 = 39 – 39.....60 – 53 = 7.

(ii) Since the magnitude of the class is 7 and the mid-values of the classes are 25, 32.....60, the corresponding class boundaries for different classes are obtained by adding and subtracting half the magnitude of the class interval *i.e.*, $\frac{7}{2}$ = 3·5 to/from the mid-values to obtain higher and lower class boundaries.

1st Class → 25 – 3·5, 25 + 3·5

2nd Class → 32 – 3·5, 32 + 3·5

3rd Class → 39 – 3·5, 39 + 3·5

4th Class → 46 – 3·5, 46 + 3·5

5th Class → 53 – 3·5, 53 + 3·5

6th Class → 60 – 3·5, 60 + 3·5.

Class Intervals

21·5 – 28·5

28·5 – 35·5

35·5 – 42·5

42·5 – 49·5

49·5 – 56·5

56·5 – 63·5

Notes

(iii) Assuming that the age quoted (X) is the age completed on last birthday then X will be a discreate variable which can take only integral values. Hence the given distribution can be expressed in an inclusive type classes with class interval of magnitude 7, as in the table given below.

Age (on last birthday)	Mid Values
22-28	25
29-35	32
36-42	39
43-49	46
50-56	53
56-63	60

3. Industrial finance in India showed great variations in respect of sources during the first, second and third plans. There were two main sources *viz.*, internal and external. The former had two sources — depreciation and free reserves surplus. The latter had three sources — capital issues, borrowings and 'other sources'. During the first plan, internal and external sources accounted for 62% and 38% of the total and in this depreciation fresh capital and 'other sources formed 29%, 7% and 10·6% respectively.

   During the second plan internal sources decreased by 17·3% compared to the first plan and depreciation was 24·5%. The external finance during the same period consisted of fresh capital 10·9% and borrowings 28·9%. Compared to the second plan, during the third plan, external finance decreased by 4·4% and borrowings and 'other sources' were 29·4% and 14·9% respectively. During the third plan, internal finance increased by 4·4% and free reserves and surplus formed 18·6%.

   Tabulate the above information with the above details as clearly as possible, observing the rules of tabulation.

Solution:

Table Showing Pattern of Industrial Finance (in per cent)

Plan	Sources						
	Internal			External			
	Depreciation	Free reserves and surplus	Total	Capital issues	Borrowings	Other	Total Sources
First	29·0	33·0	62·0	7·0	20·4	10·6	38·0
Second	243·5	20·2	44·7	10·9	28·9	15·5	55·3
Third	30·5	18·6	49·1	6·6	29·4	14·9	50·9

Self-Assessment

1. Fill in the blanks:

(*i*) Classification is the ..... step in tabulation.

(*ii*) When data are observed ..... the type of classification is known as chronological classification.

(*iii*) ..... classification refers to the classification of data according to some characteristics that can be measured.

Notes

(*iv*) A table is a systematic arrangement of statistical data in ..... .

(*v*) In collection and tabulation ..... is the chief requisite and experience the chief ..... .

(*vi*) The number of observations corresponding to a particular class is known as the ..... of that class.

3.5 Summary

- Moreover, classification and tabulation segregates the likes from the unlikes. The heterogeneity is removed. The data are classified into classes and sub-classes according to their characteristics. This process is called **classification**. The classified data are presented in prescise and systematic tables. This process is called **tabulation**. By these two processes, the data collected are made simple, easy to understand and carry out analysis and interpretations.
- Classification may be defined as the process of arranging the available data in various groups or classes in accordance with their resemblances and similarities and keeping in view some common features and objectives of study. Thus, through classification, an effort is made to achieve homogeneity of the collected information. While classifying, the units with common characteristics are placed together and in this way the whole data is divided into a number of classes and sub-classes.
- The process of classification helps in arranging the data in such a way that the large mass of irrelevant looking data becomes simple and easy to understand, avoiding unnecessary details, making logical sense.
- Classification is done in accordance with the attributes or characteristics of the data. Such classification is generally done where data cannot be measured. Under this classification method, the presence or absence of an attribute is the basis of classification.
- The bases of classification are manifold, *i.e.*, more than one attribute. In this method, classes/ groups are further classified into sub-classes and sub-groups. For example, population/sample is first classified on the basis of sex, then for each sex (male or female) marital status forms two sub-classes then further these sub-classes are classified as per their literacy state.
- The process of tabulation eliminates unnecessary details and present the complex data concisely in rows and columns. This helps in simplifying the complex data which becomes more meaningful and better understood.
- It is only possible after tabulation, that the data can be subjected to statistical analysis and interpretation. Measures of correlation, regression dispersion etc can be easily calculated when the data is in tabular form.
- A title is a brief statement indicating about the nature of the data and the time-span to which the data relate. The geographical distribution of the data, if any, should also be indicated in the title. The title should be in dark letters in comparison to other heads and sub-heads in the table.
- Footnotes are also used to describe the source of the data, if the data is secondary in nature. For example, "the figures in bracket show the per cent rise over previous year" or "the profit above is after tax" etc are some of the information which is given in footnotes.
- A table presents the statistical data in a systematic way in rows and columns which concisely explain the numerical facts. Tabulation is nothing but the process of preparing a table. The preparation of table is a specialised activity and is done through a set of rules.
- Neat and tidy appearance must be given to the table which can be done by providing proper ruling and spacing as is necessary. If the table continues to the next page, no bottom line must be drawn, as it would indicate the end of the table. Major and minor items must be given space according to their relative importance. Coloured inks, heavy printed titles or sub-titles, thick and thin ruling etc must be used to clarify a complex table.

Notes

- The body of the table must be as comprehensive as possible, consistent with the purpose. Unnecessary details must be avoided and, items in 'miscellaneous or unclassified columns must be least.
- Tabulation is an art which requires common sense in planning a table and viewing the proposed table from the point of view of the user or the other person. Tabulation is done keeping in mind the purpose of statistical investigation. The rules of tabulation act as guides in preparing a good table.
- Quantitative classification data is done through seriation of data. If two variable quantities are arranged side by side so that the measurable differences in the one correspond to the measurable differences in the other, the result is formation of a statistical series.
- In continuous frequency distribution data are divided into class intervals instead of individual values (as is done in case of discrete frequency distribution) class intervals can be formulated like marks from 0 to 10, 10 – 20 and so on.
- Tables are nor prepared with the help of machines which may be either hand-operated or are operated with electricity. Use of 'needle sorting' is one such machine for tabulation. Similarly 'Punch Cards' are also used. The work by these machines is more fast, easy and accurate.
- The most important method of organising and summarising statistical data is by constructing a frequency distribution table. In this method, classification is done according to quantitative magnitude. The items are classified into groups or classes according to their increasing order in terms of magnitude and the number of items falling into each group is determined and indicated.
- Consider the number of all observations which are *less than the upper class boundary* of a given class interval; this number is the sum of the frequencies upto and including that class to which the upper class boundary corresponds.
- The cumulative frequency distribution is represented by joining the points obtained by plotting the cumulative frequencies along the vertical axis and the corresponding upper class boundaries along the *x-axis*. The corresponding polygon is known as cumulative frequency polygon (less than) or ogive. By joining the points by a freehand curve we get the cumulative frequency curve ("less than"). Similarly we can construct another cumulative frequency distribution ("more than" type) by considering the sum of frequencies greater than the lower class boundaries of the classes.

3.6 Key-Words

1. Tabulation of data : The process of placing classified data into tabular form is known as tabulation. A table is a symmetric arrangement of statistical data in rows and columns. Rows are horizontal arrangements whereas columns are vertical arrangements. It may be simple, double or complex depending upon the type of classification.
2. Frequency distribution : In statistics, a frequency distribution is an arrangement of the values that one or more variables take in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

3.7 Review Questions

1. What do you mean by Tabulation ? What are the objectives and advantages of tabulation?
2. What is the frequency distribution ? Explain how it is formed from raw data.
3. Describe the importance of classification and tabulation in statistical analysis.
4. Describe the various points to be considered in the construction of a frequency table.
5. What are the different parts of a table ? What points should be taken into account while preparing a table ?

Notes

Answers: Self-Assessment

1. (*i*) first (*ii*) over a period of the time
   (*iii*) Quantitative (*iv*) columns and rows
   (*v*) common sense, teacher (*vi*) frequency

3.8 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 4: Central Tendency: Mean, Median and Mode and their Properties

CONTENTS

Objectives

After reading this unit students will be able to:

- Know the Meaning and Definition of Central Tendency.
- Discuss the Mean, Median and Mode and their Properties.

Introduction

For statistical analysis, condensation of data is essential so that the complexity of data is reduced and is made comparable. This can be done by finding the central tendencies of the data or the averages. By this, the large mass of data gets reduced to one figure each and thus comparison becomes much easier. For example, if a comparison of student's results in two different colleges with 200 students each, is to be made, it seems to be impossible to draw any conclusion looking at the results of these 400 students. But if, each of these series is represented by a single figure, comparison becomes very easy. This figure is the one which represents the whole series, and so it neither is the highest nor the lowest value rather, it is the value where most of the items of the series cluster or are nearer. Such figures present the central tendency of the series and are called Measures of central tendency or Averages. Its value lies between maximum and minimum.

4.1 Meaning and Definition of Central Tendency

Measures of central tendency or averages reduce the large number of observations to one figure. Actually the measures of central tendency describe the tendency of items of group around the middle in a frequency distributions of numerical values.

Definitions

According to *L. J. Kaplan* — "One of the most widely used set of summary figures is known as measures of location, which are often referred to as averages, central tendency or central location. The purpose for computing an average value for a set of observations is to obtain a single value which is representative of all the items and which the mind can grasp simply and quickly. The single value is the point of location around which the individual items cluster."

According to *G.P. Watkins,* "average is a representative figure which is gist, if not the substance of statistics."

In the words of *Croxton* and *Cowden,* "An average value is single value within the range of the data that is used to represent all the values in the series."

Notes

Characteristics of a Good Average

The above discussion reveals that an average or the value of central tendency is a representative figure. Therefore, a good average would be the one which has the capability of representing the data most efficiently and effectively. For this, certain are the characteristics of the average so that it can prove to be good. These essential characteristics for an average to prove to be good are:

(1) **It should be rigidly defined:** According to Prof. Yule and Kendall, the average should be defined rigidly so that there is only one possible interpretation and is not subject to observers' own interpretation and bias. For this, the average should be defined in terms of alzebraic formula.

(2) **It should be based on all the observations:** In order to make the data representative it is very essential that it is based on all the observations.

(3) **It should be capable of further alzebraic treatment:** For the average to be good, it is essential that it is capable of further alzebraic treatment, otherwise its use will become very limited. For example, in the absence of this quality, the combined average of two or more series from their individual averages will not be calculated. This would hinder the possibility to study the average relationship of various parts of a variable, if it is expressed as the sum of two or more variables.

(4) **It should not be affected by fluctuations of sampling:** If two independent sample studies are made in any particular field, their averages obtained, should not differ from each other ideally. Practically, it is difficult to obtain no difference, but the average in which this difference, technically called as 'fluctuation of sampling' is least, is considered to be a better average.

(5) **It should be easy to compute:** An average should be capable of being calculated with reasonable ease and within reasonable time. If the time taken is long or the calculations are tedious and complicated, the average shall have only limited use.

(6) **It should be easy to understand:** A good average is the one which is easily understood by the common people. It should neither be abstract nor too mathematical; otherwise its use will again be restricted.

Types of Statistical Averages

The following are the main types of statistical averages:

(1) **Positional Averages:** These include —

(a) Median (represented by M)

(b) Mode (represented by Mo).

(2) **Mathematical Averages:** These include —

(a) Arithmatic Average or Mean (represented by $\bar{X}$)

(b) Geometric Mean (represented by 'G.M.')

(c) Harmonic Mean (represented by 'H.M.')

(d) Quadratic Mean (represented by 'Q.M.')

(3) **Commercial Averages:**

(a) Moving Average.

(b) Progressive Average.

(c) Composite Average.

Measures or types of Central tendency or averages can be shown as in Figure 1.

Notes

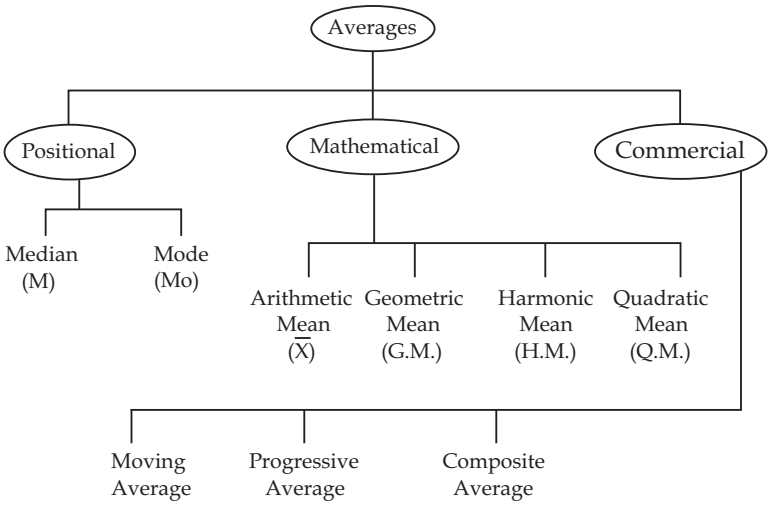


Figure: 1

Symbolically, the above may be shown as:

Mode = Z or M_O

Median = M

A. A. or Mean or $\bar{X}$

Geometric Mean = *g* or G.M.

Harmonic Mean = *h* or H.M.

Quadratic Mean = Q.M.

4.2 Mean, Median and Mode and their Properties

Arithmetic Average or Mean

Arithmetic mean is the most widely used method of calculated averages, so much so that when only 'mean' is indicated it is assumed to be arithmetic mean universally. It is obtained by adding up all the observations and dividing it by number of observations.

Merits of Arithmetic Mean

The merits of Arithmetic Mean are:

(1) Simple to understand,

(2) Easy to compute,

(3) Capable of further mathematical treatment,

(4) Calculated on the basis of all the items of the series,

(5) It gives the value which balances the either side,

(6) Can be calculated even if some values of the series are missing,

(7) It is least affected by fluctuations in sampling.

Demerits of Arithmetic Mean

The demerits of Arithmetic Mean are:

(1) Extreme items have disproportionate effect. For example, average marks obtained of five students are:

Notes

$$\frac{50+10+10+10+10}{5} = \frac{90}{5} = 18.$$

Whereas in reality 4 out of 5 students failed. Therefore, '18' marks cannot be termed as representative.

(2) When data is vast, the calculations become tedious.

(3) In case of open end classes, mean can only be calculated by making some assumptions.

(4) A.M. is not representative if series is asymmetrical.

Purpose or Objectives of Averaging

Central tendency or average is the value by which the data can be represented. The purpose or objectives of calculating this representative figure are —

(1) To present the most important features of a mass of complex data.

(2) To facilitate comparing one set of data with others, so that conclusions can be drawn quickly.

(3) To help in understanding the picture of a complete group by means of sample data.

(4) To trace the mathematical relationship between different groups or classes.

(5) To help in the decision-making.

(6) To facilitate the process of experimentation and research.

Weighted Arithmetic Mean

Weighted arithmetic mean is the method of calculating a more representative central value and takes into consideration the relative importance of the various figures in the series. Whereas in simple arithmetic mean, equal weight or importance is given to each item. If the central value has to more representative and the data is such that few items are more important than other, the method of weighted arithmetic mean is used. This method is generally used in the following situations:

(1) When importance of all the items in the series is not equal.

(2) When the classes of the same group contain widely varying frequencies.

(3) Where there is a change either in the proportion of values or items or in the proportion of frequencies.

(4) When ratios, percentages or rates are being averaged.

(5) When it is desired to calculate the average of series from the average of its component parts.

The formula for the weighted arithmetic average is:

Direct Method : $\overline{X}W = \frac{W_1X_1 + W_2X_2 + ...W_nX_n}{W_1 + X_2 + ...W_n} = \frac{\Sigma WX}{\Sigma W}$

Short-cut Method : $\overline{X}W = A_w + \frac{\Sigma Wd}{\Sigma W}$

where $\overline{X}W$ represents the weighted arithmetic mean.

X represents the variable values *i.e.*, X_1, X_2 ... X_n.

W represents the weights attached to variable values *i.e.*, W_1, W_2, ... W_n, respectively.

ΣWd Sum of the product of the deviations from the assumed mean (AW) multiplied by the respective weights.

AW Assumed (weighted) mean.

Harmonic Mean

An average rate like kilometer per hour, per day items manufactured etc. are required to be found, harmonic mean is calculated. Harmonic mean is a type of average which has limited application only

Notes

that too in a restricted field. The harmonic mean of a series of values is the reciprocal of the arithmetic mean of the reciprocals of the individual values. Reciprocal tables are used with ease for this. The Harmonic Mean is less than the geometric mean of the same observations. The formula to calculate harmonic mean is:

$$\text{H.M.} = \frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}+\ldots\frac{1}{X_n}}$$

or, reciprocal of —

$$\frac{\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}\ldots\frac{1}{X_n}}{N}$$

or, reciprocal of $\frac{(\Sigma \text{Reciprocal of } X)}{N}$

where X, X_1, X_2, ... X_n are the observations or the values of the series.

Merits of H.M.

(1) Harmonic mean is calculated by taking into account all the items of the series.

(2) In series with wide dispersion, this method is useful.

(3) It gives less weight to large items and more weight to small ones (because reciprocals are used).

(4) The method is useful in calculating rate.

(5) While calculating harmonic mean, the values get weight automatically and there is no need to assign weights specifically.

Demerits of H.M.

(1) It is difficult to calculate.

(2) Difficult to be understood by common man.

(3) Harmonic mean cannot be calculated if even one item of the series is missing.

(4) The value of harmonic mean obtained may be a value which is no item in the given series.

Geometric Mean

Geometric mean of '*n*' numbers is defined as the n^{th} root of the product of '*n*' numbers. Symbolically,

$$\text{G.M.} = \sqrt[n]{(X_1)(X_2).(X_3)\ldots(X)_n}$$

where X_1 X_2 ... X_n are the various values of the series.

'*n*' is the number of items. To make the calculations of finding out n^{th} root simpler, logarithms are used. G.M. by using logs is found thus:

$$\text{G.M.} = \text{Antilog of } \frac{\Sigma \log X}{N}.$$

Where, measuring the ratios of change is required, the geometric mean are used. It is also most suitable when large weights have to be given to small items and small weights to large items, which is usually required to study economic and social phenomena.

Merits

(1) Based on all the values of the series.

(2) Capable of further alzebraic treatment.

Notes

(3) It is not much affected by the extreme values in the series.

(4) Capable of being applicable to study social and economic phenomena.

Demerits

(1) If any value is '0' (zero), G.M. cannot be calculated (because (X_1), (X_2) ... (X_n) is to be found. If any of these is zero, the multiplication result will be zero and interpretation would become impossible.

(2) Knowledge of logarithms is essential. Therefore it is found difficult to compute by a non-mathematics background person.

(3) Difficult to locate.

(4) G.M. cannot be calculated if even one value of the series is not available.

(5) The value of G.M. obtained may not be there in the series, therefore, it cannot be termed as the true representative of the data.

Mathematical Properties of the Arithmetic Mean

The following are a few important mathematical properties of the arithmetic mean:

1. The sum of the deviations of the items from the arithmetic mean (taking signs into account) is always zero, *i.e.*, $\Sigma(X - \bar{X}) = 0$*. This would be clear from the following example:

X	$(X - \bar{X})$
10	– 20
20	– 10
30	0
40	+ 10
50	+ 20
$\Sigma X = 150$	$\Sigma(X - \bar{X}) = 0$

Here $\bar{X} = \frac{\Sigma X}{N} = \frac{150}{5} = 30$. When the sum of the deviations from the actual mean, *i.e.*, 30, is taken it comes out to be zero. It is because of this property that the mean is characterised as a *point of balance, i.e.*, the sum of the positive deviations from it is equal to the sum of the negative deviations from it.

2. The sum of the squared deviations of the items from arithmetic mean is minimum, that is less than the sum of the squared deviations of the items from any other value. For example, if the items are 2, 3, 4, 5 and 6 their squared deviation shall be:

X	$(X - \bar{X})$	$(X - \bar{X})^2$
2	– 2	4
3	– 1	1
4	0	0
5	+ 1	1
6	+ 2	4
$\Sigma X = 20$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 10$

Notes

The sum of the squared deviations is equal to 10 in the above case. If the deviations are taken from any other value the sum of the squared deviations would be greater than 10. This is known as least squares property of the arithmetic mean.

3. The standard error of arithmetic mean is less than that of any other measure of central tendency.

4. Since $\bar{X} = \frac{\Sigma X}{N}$

$N\bar{X} = \Sigma X$.

In other words, if we replace each item in the series by the mean, then the sum of these substitutions will be equal to the sum of the individual items. For example, in the discussion of first property ΣX = 150 and the arithmetic mean = 30. If for each item we substitute 30, we get the same total *i.e.*, 150 = [30 + 30 + 30 + 30 + 30].

* Algebraically the property $\Sigma(X - \bar{X}) = 0$ is derived from the fact that $N\bar{X} = \Sigma X$

This property is of great practical value. For example, if we know the average wage in a factory, say, Rs. 200, and the number of workers employed, say, 50, we can compute total wage bill from the relation $N\bar{X} = \Sigma X$. The total wage bill in this case would be 200 × 50 = 10,000 which is equal to ΣX.

5. If we have the arithmetic mean and number of items of two or more than two related groups, we can compute combined average of these groups by applying the following formula:

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$\bar{X}_{12}$ = combined mean of the two groups

$\bar{X}_1$ = arithmetic mean of first group

$\bar{X}_2$ = arithmetic mean of second group

N_1 = number of items in the first group

N_2 = number of items in the second group.

6. If the given observations on X be changed to observation on Y, where Y = a + bX, then $\bar{Y} = a + b\bar{X}$.

The following example shall illustrate the application of the above formula:

Example 1 : A factory employs 100 workers of whom 60 work in the first shift and 40 work in the second shift. The average wage of all the 100 workers is Rs. 38. If the average wage of 60 workers of the first shift is Rs. 40, find the average wage of the remaining 40 workers of the second shift.

Solution : Total no. of employees = 100

No. of employees in the first shift, *i.e.*, N_1 = 60

No. of employees in the second shift, *i.e.*, N_2 = 40

$\bar{X}_{12}$ = 38, $\bar{X}_1$ = 40

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Notes

$$38 = \frac{60(40) + 40\overline{X}_2}{100}$$

$$3800 = 2400 + 40\overline{X}_2$$

$$40\overline{X}_2 = 1400$$

$$\therefore \quad \overline{X}_2 = \frac{1400}{40} = 35$$

Hence the wage of the remaining 40 workers in the second shift is Rs. 35.

If we have to find out the combined mean of the three series, the above formula can be extended as follows:

$$\overline{X}_{123} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2 + N_3\overline{X}_3}{N_1 + N_2 + N_3}$$

Median

Median is one of the measures of central tendency. It is a positional average.

Definition of Median

Median may be defined as 'the middlemost or central value of the series when the values are arranged in ascending or descending order of magnitude'.

If there is an odd number of an item, then the median is found out by taking the middle most items of the series only after arranging the data in order of magnitude.

For example, if daily wages of 7 workers are Rs. 127, Rs. 167, Rs. 154, Rs. 177, Rs. 135, Rs. 160 and Rs. 157 and we wish to know the median wage, the wages must be arranged either in ascending order or descending order of magnitude and the 4^{th} reading will be the median wage.

Arranged in order of magnitude, the wages might be

Rs. 127 135 154 157 160 168 177

and the median wage is Rs. 157, *i.e.* the 4^{th} item.

If there is an even number of items, then the median is half-way between the two middle ones and it is found by taking the average of these two items. For example, if the marks secured by 10 students in an examination are 75, 48, 63, 89, 100, 55, 35, 28, 93 and 79 and we wish to know the median mark, the marks must be arranged either in ascending or descending order of magnitude and then the average of the value of the 5^{th} item and the 6^{th} item will be the median mark.

Arranged in order of magnitude, the marks might be 28, 35, 48, 55, 63, 75, 79, 89, 93, 100 and the median marks is Rs. $\frac{63+75}{2} = \frac{138}{2} = 69$, *i.e.* the average of the 5^{th} item and the 6^{th} item.

Properties of Median

Median is of the following properties:

1. Median may be the middlemost value of the series when the values are arranged in order of magnitude.
2. Median is influenced by the position of items in the array but not by the size of the items.
3. The value of the median of a series may or may not coincide with the value of an existing item.
4. The median cannot readily be located unless the data have been put into an array or into a frequency distribution.

Notes

5. The median of the sum or difference or pairs of corresponding items into two series is not equal to the sum or difference of the medians of the two series.

Mode

The term 'mode' has come from French in which it means 'to be in fashion'. As a statistical language, mode is the value that occurs most frequently in a statistical distribution. Thus 'Mode' is the most representative average and is a position of greatest concentration of values. It has great value conceptually. It is what the doctor means when he describes that a desease of cold and fever usually takes a week to get cured. Similarly, average size of shirt/shoes sold, average family income etc. also cannot be most frequently occurring value.

According to *Tate,* "The mode may be defined as the item which occurs most frequently in a statistical series."

In the words of *Garrett,* "Mode is that single measure or score which occurs most frequently."

Merits

The merits are as follows:

(1) Easy to understand,

(2) Simple to calculate and locate,

(3) Quantitative data in ranking is possible, mode is very useful,

(4) It is the actual value that is in the series,

(5) Mode remains unaffected by dispersion of series,

(6) Not affected by extreme items,

(7) Can be calculated even if extreme values are not known.

Demerits

The demerits are as follows:

(1) Mode cannot be subject to further Mathematical treatment, because it is not obtained from any alzebraic calculations,

(2) It is quite likely that there is no mode for a series,

(3) Cannot be used if relative importance of items have to be considered,

(4) Choice of grouping has a considerable influence on the value of the mode.

Did u know? Harmonic mean is a type of average which has limited application only that too in a restricted field.

Properties of mode

Assuming definedness, and for simplicity uniqueness, the following are some of the most interesting properties.

- All three measures have the following property: If the random variable (or each value from the sample) is subjected to the linear or affine transformation which repalces X by $aX + b$, so are the mean, median and mode.
- However, if there is an arbitrary monotonic transformation, only the median follows; for example, if X is replaced by exp (X), the median changes from m to exp (m) but the mean and mode won't.
- Except for extremely samples, the mode is insentive to "outliers" (such as occasional, rare, false experimental readings). The median is also very robust in the presence of outliers, while the mean is rather sensitive.

Notes

- In continuous unimodal distributions the median lies, as a rule of thumb, between the mean and the mode, about one third of the way going from mean to mode. In a formula, median $\approx$ (2 × mean + mode)/3. This rule, due to Karl Pearson, often applies to slightly non-symmetric distributions that resemble a normal distribution, but it is not always true and in general the three statistics can appear in any order.
- For unimodal distributions, the mode is within $\sqrt{3}$ standard deviations of the mean, and the root mean square deviation about the mode is between the standard deviation and twice the standard deviation.

Self-Assessment

1. Fill in the blanks:

(*i*) The consecutive addition of frequencies is called .....

(*ii*) Below 10, more than 40 are the examples of ..... class-intervals.

(*iii*) Sum of the deviations of the items from the ..... is always zero (taking + *ve* and –*ve* signs).

(*iv*) n^{th} root or '*n*' items of a series is termed as .....

(*v*) ..... of a series is the reciprocal of the arithmetic avrage of the reciprocals of the values of its various items.

4.3 Summary

- Measures of central tendency or averages reduce the large number of observations to one figure. Actually the measures of central tendency describe the tendency of items of group around the middle in a frequency distributions of numerical values.
- For the average to be good, it is essential that it is capable of further alzebraic treatment, otherwise its use will become very limited. For example, in the absence of this quality, the combined average of two or more series from their individual averages will not be calculated. This would hinder the possibility to study the average relationship of various parts of a variable, if it is expressed as the sum of two or more variables.
- An average should be capable of being calculated with reasonable ease and within reasonable time. If the time taken is long or the calculations are tedious and complicated, the average shall have only limited use.
- Arithmetic mean is the most widely used method of calculated averages, so much so that when only 'mean' is indicated it is assumed to be arithmetic mean universally. It is obtained by adding up all the observations and dividing it by number of observations.
- Weighted arithmetic mean is the method of calculating a more representative central value and takes into consideration the relative importance of the various figures in the series. Whereas in simple arithmetic mean, equal weight or importance is given to each item. If the central value has to more representative and the data is such that few items are more important than other, the method of weighted arithmetic mean is used.
- An average rate like kilometer per hour, per day items manufactured etc. are required to be found, harmonic mean is calculated. The harmonic mean of a series of values is the reciprocal of the arithmetic mean of the reciprocals of the individual values. Reciprocal tables are used with ease for this. The Harmonic Mean is less than the geometric mean of the same observations.
- The sum of the squared deviations of the items from arithmetic mean is minimum, that is less than the sum of the squared deviations of the items from any other value.
- Median may be defined as 'the middlemost or central value of the series when the values are arranged in ascending or descending order of magnitude'.
- If there is an odd number of an item, then the median is found out by taking the middle most items of the series only after arranging the data in order of magnitude.

Notes

- If there is an even number of items, then the median is half-way between the two middle ones and it is found by taking the average of these two items. For example, if the marks secured by 10 students in an examination are 75, 48, 63, 89, 100, 55, 35, 28, 93 and 79 and we wish to know the median mark, the marks must be arranged either in ascending or descending order of magnitude and then the average of the value of the 5^{th} item and the 6^{th} item will be the median mark.
- The term 'mode' has come from French in which it means 'to be in fashion'. As a statistical language, mode is the value that occurs most frequently in a statistical distribution. Thus 'Mode' is the most representative average and is a position of greatest concentration of values. It has great value conceptually. It is what the doctor means when he describes that a desease of cold and fever usually takes a week to get cured. Similarly, average size of shirt/shoes sold, average family income etc. also cannot be most frequently occurring value.
- In continuous unimodal distributions the median lies, as a rule of thumb, between the mean and the mode, about one third of the way going from mean to mode. In a formula, median $\approx$ (2 × mean + mode)/3. This rule, due to Karl Pearson, often applies to slightly non-symmetric distributions that resemble a normal distribution, but it is not always true and in general the three statistics can appear in any order.

4.4 Key-Words

1. Central tendency : In statistics, the term central tendency relates to the way in which quantitative data tend to cluster around some value.[1] A measure of central tendency is any of a number of ways of specifying this "central value". In practical statistical analysis, the terms are often used before one has chosen even a preliminary form of analysis: thus an initial objective might be to "choose an appropriate measure of central tendency".
2. Harmonic mean : The reciprocal of the arithmetic mean of the reciprocals of a specified set of numbers

4.5 Review Questions

1. What is meant by measures of central tendency? What are the characteristics of good measure of central tendency?
2. Explain the relative importance of arithmetic mean, median and mode as measures of central tendency in statistical analysis.
3. Define mean, median and mode. Mention its merits and demerits.
4. What are the properties of mean, median and mode?
5. Define Harmonic mean and give a situation in which it is used.

Answers: Self-Assessment

1. (*i*) cumulative frequency (*ii*) open-end
   (*iii*) mean (*iv*) geometric mean
   (*v*) harmonic mean

4.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.

Notes

Unit 5: Application of Mean, Median and Mode

CONTENTS

Objectives

After reading this unit students will be able to:

- Discuss Application of Mean and Median.
- Know the Application of Mode.

Introduction

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

5.1 Application of Mean

The most popular and widely used measure for representing the entire data by one value is what most laymen call an 'average' and what statisticians call the arithmetic mean. Its value is obtained by adding together all the items and by the dividing this total by the number of items. Arithmetic mean may be either (i) simple arithmetic mean, or (ii) weighted arithmetic mean.

Calculation of Arithmetic Mean—Individual Observations

The process of computing mean in case of individual observations (*i.e.* where frequencies are not given) is very simple. Add together the various values of the variable and divide the total by the number of items. Symbolically:

$$\overline{X}^{*} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} = \frac{\Sigma X}{N}$$

$\overline{X}$ = Arithmetic Mean

Notes

$\sum X$ = Sum of all the values of the variable X, *i.e.,* $X_1, X_2, X_3, \ldots\ldots, X_N$.

N = Number of observations.

Steps: The formula suggests two steps in calculating mean:

(i) Add together all the values of the variable X and obtain the total $\sum X$.

(ii) Divide this total by the number of observations.

Example 1: Calculate arithmetic mean from the following data:

Marks obtained by 20 students out of 200			
40	100	144	100
56	106	148	106
68	108	150	108
78	118	156	118
84	128	158	128

Solution:

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

$\sum X$ = Summation of all the items, N = 20.

$\sum X$ = 2202, N = 20

$$\bar{X} = \frac{2202}{20} = 110.1.$$

Answer: Arithmetic mean of the series is = 110.1.

Short-cut Method: The above method is useful only when 'N' is small. Mean of marks cannot be calculated with ease by the above method. Therefore, short-cut method is used. This method is based on the fact that the alzebraic sum of the deviations of a series of individual observations from their mean is always equal to zero.

Arithmetic mean by short-cut is calculated by the following formula:

$$\bar{X} = A + \frac{\sum dx}{N}$$

where, $\bar{X}$ is arithmetic mean, A is assumed mean, N is number of observations, $\sum dx$ is the sum of the deviations from the assumed mean.

Example 2: Using short-cut method, determine the arithmetic mean of the data (given in example 1 taking first 15 students).

Solution:

X	Deviations dx = X – A
40	40 – 100 = – 60
56	56 – 100 = – 44
68	68 – 100 = – 32
78	78 – 100 = – 22
84	84 – 100 = – 16
100	100 – 100 = 0
106	106 – 100 = 6
108	108 – 100 = 8

Notes

118	118 – 100 = 18
128	128 – 100 = 28
144	144 – 100 = 44
148	148 – 100 = 48
150	150 – 100 = 50
156	156 – 100 = 56
158	158 – 100 = 58
N = 15	$\sum dx$ = 142

So let assumed mean A = 100.

A.M. by short-cut method:

$$\bar{X} = A + \frac{\sum dx}{N}$$

$$A = 100 \quad \sum dx = 142, N = 15$$

$$\therefore \quad \bar{X} = 100 + \frac{142}{15}$$

$$= 100 + 9.47$$

$$= 109.47$$

Arithmetic Mean in Discrete Series

A discrete series is obtained from a large number of individual observations. Suppose the marks obtained by 100 students is given. This data can be converted into a discrete series where the marks obtained are accompained by the number of students obtaining it. For example, suppose 10 students obtained 50 marks, 12 students obtained 60, 25 students obtained 78, 3 students obtained 100, 15 students obtained 94, 15 students obtained 82 and 20 students obtained 38. Then instead of writing in form of individual observations, data can be written like this:

Marks Obtained (X)	Number of Students (*f*)
50	10
60	12
78	25
100	3
94	15
82	15
38	20
	Total N = 100

Then this is a discrete series.

Arithmetic Mean by Direct Method

$$\bar{X} = \frac{\sum fx_1 + \sum fx_2 \ldots \sum fx_n}{\sum f} \text{ or } \frac{\sum fx}{\sum f}$$

'*f*' denotes the frequency.

Arithmetic Mean by Short-cut Method

Notes

(i) $$\bar{X} = A + \frac{\Sigma fdx}{\Sigma f}$$

where A is assumed mean.

(ii) $$\bar{X} = A + \frac{\Sigma fdx}{\Sigma f} \times i$$

where i is the common factor of deviations.

Σfdx is the total of the products of the deviations from the assumed average and the respective frequency of the items.

Σf is the summation of all the frequencies.

Example 3: From the following frequency distribution calculate the mean weight of the students.

Weight (in *kgs.*)	64	65	66	67	68	69	70	71	72	73
No. of Students	1	6	10	22	21	17	14	5	3	1

Solution:

Weight (X)	**Number of Students (*f*)**	**Deviation d_2 = (X – A)**	**(*fxdx*) *fdx*.**
64	1	64 – 68 = – 4	– 4 × 1 = – 4
65	6	65 – 68 = – 3	– 3 × 6 = – 18
66	10	66 – 68 = – 2	– 2 × 10 = – 20
67	22	67 – 68 = – 1	– 1 × 22 = – 22
68	21	68 – 68 = 0	0 × 21 = 0
69	17	69 – 68 = 1	1 × 17 = 17
70	14	70 – 68 = 2	2 × 14 = 28
71	5	71 – 68 = 3	3 × 5 = 15
72	3	72 – 68 = 4	4 × 3 = 12
73	1	73 – 68 = 5	5 × 1 = 5
	Σf = 100		Σfdx = 13

Let assumed mean A = 68.

$$\bar{X} = A + \frac{\Sigma fdx}{\Sigma f}$$

$A = 68, \Sigma fdx = 13, \Sigma f = 100.$

$$\therefore \quad \bar{X} = 68 + \frac{13}{100}$$

$$\bar{X} = 68 + 0.13$$

$$\therefore \quad \bar{X} = 68.13$$

Answer: Mean weight of the students = 68.13 *kg*.

Notes

Example 4: Find the arithmetic mean of the following data using step deviation method:

X	1590	1610	1630	1650	1670	1690	1710	1730
f	1	2	9	48	131	102	40	17

Solution:

X	f	dx **= (X – A)**	**Step deviation** dx'	$-fdx'$
1590	1	1590 – 1670 = – 80	– 8	1 × – 8 = – 8
1610	2	1610 – 1670 = – 60	– 6	2 × – 6 = – 12
1630	9	1630 – 1670 = – 40	– 4	9 × – 4 = – 36
1650	48	1650 – 1670 = – 20	– 2	48 × – 2 = – 96
1670	131	1670 – 1670 = 00	0	0 × 131 = 0
1690	102	1690 – 1670 = 20	2	102 × 2 = 204
1710	40	1710 – 1670 = 40	4	40 × 4 = 160
1730	17	1730 – 1670 = 60	6	17 × 6 = 102
	Σf = 350			$\Sigma fdx'$ = 314

Let assumed mean A = 1670.

$$i = 10\ dx' = \frac{dx}{i}$$

$$\bar{x} = A + \frac{\Sigma fdx'}{\Sigma f} \times i$$

$A = 1670, \Sigma fdx' = 314; \Sigma f = 350; i = 10.$

$$\therefore \quad \bar{x} = 1670 + \frac{314}{350} \times 10$$

$$= 1670 + 8.97$$

$$\therefore \quad \bar{x} = 1678.97$$

Answer: The arithmetic mean of the above series is = 1678.97.

Calculation of the Arithmetic Mean in a Continuous Series

The continuous series expresse the data which is very vast. The calculation of arithmetic mean of this series is similar to that of discrete series after calculating the mid point of each segment of the continuous series which is called the class interval. The continuous series may have three types of class intervals: (1) Exclusive class interval for example, 10—20, 20—30, 30—40 .... etc. (2) Inclusive class interval for example, 0—9, 10—19, 20—29, 30—39 ... etc. If the data is given in the form of inclusive class intervals, it is first converted into exclusive class interval, (3) Cumulative class interval for example, more than 10, more than 20 ... etc. or less than 10, less than 20 ... etc.

Example 5: For the following data calculate the mean marks obtained by the students using: (i) Short-cut method, (ii) Step deviation method.

Marks	10—20	20—30	30—40	40—50	50—60
Number of Students	1	2	3	5	7
Marks	60—70	70—80	80—90	90—100	
Number of Students	12	16	10	4	

Solution:

Notes

X	f	Mid value (M)	dx = M – A fdx	Step *devi.* dx'	fdx'
10—20	1	15	15 – 55 = – 40 = – 40	– 4	– 4 × 1 = – 4
20—30	2	25	25 – 55 = – 30 = – 60	– 3	– 3 × 2 = – 6
30—40	3	35	35 – 55 = – 20 = – 60	– 2	– 2 × 3 = – 6
40—50	5	45	45 – 55 = – 10 = – 50	– 1	– 1 × 5 = – 5
50—60	7	55	55 – 55 = 00 = 0	0	0 × 7 = 0
60—70	12	65	65 – 55 = 10 = 120	1	1 × 12 = 12
70—80	16	75	75 – 55 = 20 = 320	2	2 × 16 = 32
80—90	10	85	85 – 55 = 30 = 300	3	3 × 10 = 30
90—100	4	95	95 – 55 = 40 = 160	4	4 × 4 = 16
	$\Sigma f = 60$		$\Sigma fdx = 690$		$\Sigma fdx' = 69$

Let assumed mean A = 55.

(i) $\bar{X}$ (by short-cut method) $= A + \frac{\Sigma fdx}{\Sigma f}$

$A = 55; \Sigma fdx = 690; \Sigma f = 60.$

$$\bar{X} = 55 + \frac{690}{60}$$

$$= 55 + 11.5$$

$$\bar{X} = 66.5$$

(ii) $\bar{X}$ (by step deviation method) $= A + \frac{\Sigma fdx'}{\Sigma f} \times i$

$A = 55; \Sigma fdx' = 69; \Sigma f = 60, i = 10.$

$$\bar{X} = 55 + \frac{69}{60} \times 10$$

$$= 66.5$$

Mean marks obtained by students = 66.5.

$$= 17.5 + \frac{-62}{80}$$

$$= 17.5 - 0.775$$

$$= 16.725 \text{ rupees.}$$

Answer: 16.275 approx.

Example 6: Find out the missing frequency in the following distribution:

Marks	No. of Students
0—10	4
10—20	7
20—30	?

Notes

30—40	17
40—50	6
50—60	4

The mean of the distribution is **30.2** marks.

Solution: Let x be the missing frequency.

Marks	M.V. (m)	frequency (f)	$m.f.$
0—10	5	4	20
10—20	15	7	105
20—30	25	x	$25x$
30—40	35	17	595
40—50	45	6	270
50—60	55	4	220
		$n = 38 + x$	$\Sigma mf = 1210 + 25x$

$$a = \frac{\Sigma mf}{n} \text{ or } 30.2 = \frac{1210 + 25x}{38 + x}$$

or $30.2\ (38 + x) = 1210 + 25\ x$

or $1147.6 + 30.2\ x = 1210 + 25\ x$

or $30.2\ x - 25\ x = 1210 - 1147.6$

or $5.2\ x = 62.4$

$x = 12$

The missing frequency is, therefore, 12.

Answer: 12.

Example 7: Calculate the Geometric Mean of the following two series:

Series A	Series B
173	0.8974
182	0.0570
75	0.0081
5	0.5677
0.8	0.0002
0.08	0.0984
0.8974	0.0854
	0.5672

Solution:

Series A		Series B	
Values	**logs**	**Values**	**logs**
173	2.2380	0.8974	$\underline{1}.9530$
185	2.2601	0.0570	$\underline{2}.7559$
75	1.8751	0.0081	$\underline{3}.9085$
5	$\underline{0}.6990$	0.5677	1.7541

Notes

0.8	$\bar{1}$.9031	0.0002	$\bar{4}$.3010
0.08	$\bar{2}$.9031	0.0984	$\bar{2}$.9930
0.8974	1.9530	0.0854	$\bar{2}$.9315
		0.05672	1.7538
N = 7	Σlogs = 5.8314	N = 8	Σlogs = 10.3508

Series A

$$\text{G.M.} = \text{Antilog}\left(\frac{\Sigma \text{logs}}{N}\right) = \text{A.L.} = \left(\frac{5.8314}{7}\right)$$

G.M. = Antilog 0.8331 = 6.810

Series B

$$\text{G.M.} = \text{Antilog}\left(\frac{\Sigma \text{logs}}{N}\right) = \text{AL}\left(\frac{10.3508}{8}\right)$$

$$\text{G.M.} = \text{A.L.}\left(\frac{16 + 6.3508}{8}\right) = \text{AL}\bar{2}.7938 = 0.0622.$$

5.2 Application of Median

The median by definition is the middle value of the distribution. Whenever the median is given as a measure, one-half of the items in the distribution have a value the size of the median value or smaller and one-half have a value the size of the median value or larger.

As distinct from the arithmetic mean which is calculated from the *value of every* item in the series, the median is what is called a *positional* average. The term 'position' refers to the place of a value in a series. The place of the median in a series is such that an equal number of items lie on either side of it. For example, if the income of five persons is 2,700, 2,720, 2,750, 2,760, 2,780, then the median income would be Rs. 2,750. Changing any one or both of the first two values with any other numbers with value of 2,750 or less, and on changing of the last two values to any other values of 2,760 and more, would not affect the values of the median which would remain 2,750. In contrast, in case of arithmetic mean the change in the value of a single item would cause the value of the mean the changed. Median is thus the central value of the distribution or the value that divides the distribution into two equal parts. If there are even number of items in a series there is no actual value exactly in the middle of the series and as such the median is indeterminate. In such a case the median is arbitrarily taken to be halfway between the two middle items. For example, if there are 10 items in a series, the median position is 5.5, that is, the median value is halfway between the value of the items that are 5th and 6th in order of magnitude. Thus when N is odd the median is an actual value with the remainder of the series in two equal parts on either side of it. If N is even then the median is a derived figure, *i.e.,* half the sum of the two middle values.

Calculation of Median Individual Observation

Steps:

(i) Arrange the data in ascending or descending order of magnitude. (Both arrangements would give the same answer.)

(ii) Apply the formula: Median = Size of $\frac{N+1}{2}$th item.

Notes

Individual series

$$\text{Median} = \text{The size of } \frac{N+1}{2}\text{th item}$$

Example 1: Calculate median from the following data:

80 60 70 55 95 78 43

Solution: Arranging the given data in ascending order, we get

43 55 60 70 78 80 95

$$\text{Median} = \text{The size of } \frac{N+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{7+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{8}{2}\text{th item}$$

Median = The size of 4th item, *i.e.* 70

– Median = 70

Example 2: Compute median from the following data:

74 52 63 45 85 69 55 30

Solution: Arranging the given data in ascending order, we get

30 45 52 55 63 69 74 85

$$\text{Median} = \text{The size of } \frac{N+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{8+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{9}{2}\text{th item}$$

Median = The size of 4.5th item,

$$\text{Median} = \text{The size of } \frac{4\text{th item} + 5\text{th item}}{2}$$

$$\text{Median} = \text{The size of } \frac{55+63}{2}\text{th item}$$

$$\text{Median} = \frac{118}{2}$$

– Median = 59

Discrete series

$$\text{Median} = \text{The size of } \frac{N+1}{2}\text{th item.}$$

Example 3: Find the median wage from the data given below:

Notes

Daily wage (Rs.)	100	150	200	250	300	350	400
No. of workers	10	15	25	30	32	28	10

Solution: Median= The size of $\frac{N+1}{2}$th value

Daily wages (Rs.)	**No. of workers**	
x	*f*	*cf*
100	10	10
150	15	25
200	25	50
250	30	80
300	32	112
350	28	140
400	10	150
	N = 150	

$$\text{Median} = \text{The size of } \frac{N+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{150+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{151}{2}\text{th item}$$

Median = The size of 75.5th item, *i.e.* 250

– Median wage = Rs. 250

Continuous series

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times c$$

where

L = Lower limit of the median class

$\frac{N}{2}$ = Half of the total frequency

cf = Cumulative frequency value lies just above the median class

f = Actual frequency lies on the median class

C = Class interval of the median class.

Notes

Example 4: Calculate median mark from the following data:

Marks	0—20	20—40	40—60	60—80	80—100
No. of students	8	16	24	12	40

Solution: Median= $L + \frac{\frac{N}{2} - cf}{f} \times c$

Marks	*f*	*cf*
0—20	8	8
20—40	16	24
40—60	24	48
60—80	12	60
80—100	40	100
	N = 100	

$$\text{Median class} = \frac{N}{2}\text{th class}$$

$$= \frac{100}{2}\text{th class}$$

= 50th class, *i.e.* 60 – 80

$L = 60$

$\frac{N}{2} = 50$

$cf = 48$

$f = 12$

$c = 20$

Substituting the values in the formula, we get

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times c$$

$$\text{Median} = 60 + \frac{50 - 48}{12} \times 20$$

$$\text{Median} = 60 + \frac{2}{12} \times 20$$

$$\text{Median} = 60 + \frac{40}{12}$$

Median = 60 + 3.33

– Median mark = 63.33.

Notes

Calculation of Median in a Series of Individual Observations

Example 5: Find the median of the following:

Marks obtained by 11 students	40	42	44	48	52	60	68	70	75	80	82

Solution: Median M= Size of $\left(\frac{N+1}{2}\right)^{th}$ item.

$N = 11.$

$\therefore \quad M = \text{Size of}\left(\frac{11+1}{2} = \frac{12}{6} = 6\right)^{th}$ item.

The 6th item is 60.

$\therefore \quad M = 60.$

Answer: The median marks of the above data is 60.

Example 6: Find the value of median from the following:

Marks — 10, 9, 19, 21, 25, 32, 11.

Solution: The above data is first rearranged in the ascending order.

Marks — 9, 10, 11, 19, 21, 25, 32.

$$M = \left(\frac{N+1}{2}\right)^{th} \text{ item.}$$

$N = 7 \quad \therefore M = \left(\frac{7+1}{2} = \frac{8}{2}\right) = 4\text{th item.}$

$\therefore \quad M = 19.$

Answer: The median marks in the above data is 19.

Example 7: Compute median for the following:

X — 9, 19, 21, 6, 12, 18, 17, 20.

Solution: The data is first rearranged in ascending order:

6, 9, 12, 17, 18, 19, 20, 21

$$M = \left(\frac{N+1}{2}\right)^{th} \text{ item. } N = 8.$$

$$\therefore \quad M = \left(\frac{8+1}{2} = \frac{9}{2}\right) = 4.5\text{th item.}$$

$$\therefore \quad M = \frac{\text{Size of 4th item} + \text{Size of 5th item}}{2}$$

4th item = 17, 5th item = 18.

$$M = \frac{17+18}{2} = \frac{35}{2} = 17.5$$

Answer: The median of the above data is 17.5.

Notes

Calculation of Median in Discrete Series

Example 8: Find out the value of median from the following data:

Weekly Wages (Rs.)	100	50	70	110	80
Number of Workers	15	20	15	18	12

Solution: The data is first rearranged in ascending order (with respect to X).

X (ascending order)	*f*	**Cumulative frequency *c.f.***
50	20	20
70	15	20 + 15 = 35
80	12	35 + 12 = 47
100	15	47 + 15 = 62
110	18	62 + 18 = 80
	Σf = 80	

$$M = \left(\frac{N+1}{2}\right)^{th} \text{ item. Here } N = \Sigma f = 80.$$

$$\therefore \quad M = \left(\frac{80+1}{2} = \frac{81}{2}\right) = 40.5^{th} \text{ item.}$$

40.5th item would lie in the cumulative frequency (*c.f.*) 47. Therefore the Median = 80.

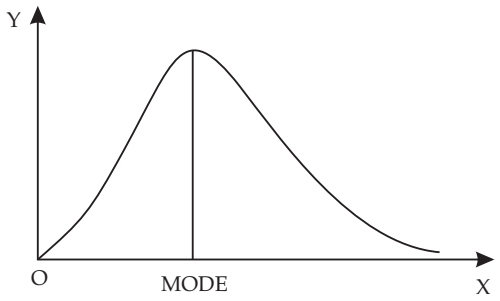
Answer: The median weekly wages = Rs. 80.

5.3 Application of Mode

A third type of "Central value" or "Centre" of the distribution is the value of greatest frequency or, more precisely, of greatest frequency density. Graphically, it is the value on the X-axis below the peak, or highest point of the frequency curve. This average is called the **mode.**

The mode is often said to be the value which occurs most frequently. While this statement is quite helpful in interpreting the mode, it cannot safely be applied to any distribution, because of the vagaries of sampling. Even fairly large samples drawn from a statistical population with a single well-defined mode may exhibit very erratic fluctuations. Hence, mode should be thought as the value which has the greatest density *in its immediate neighbourhood*. For this reason mode is also called the most typical or fashionable value of a distribution.

The following diagram will illustrate the meaning of mode:



The value of the variable at which the curve reaches a maximum is called the mode. It is the value around which the items tend to be most heavily concentrated.

Notes

Although mode is that value which occurs most frequently it does not follow that its frequency represents a majority out of all the total number of frequencies. For example, in the election of college union president the votes obtained by three candidates contesting for presidentship out of a total of 816 votes polled are as follows:

Ramesh	268
Ashok	278
Rakesh	270
Total	816

Mr. Ashok will be elected as president because he has obtained highest votes. But it will be wrong to say that he represents majority because there are more votes against him (268 + 270 = 538) than those for him.

There are many situations in which arithmetic mean and median fail to reveal the true characteristics of data. For example, when we talk of most common wage, most common income, most common height, most common size of shoe or ready-made garments we have in mind mode and the arithmetic mean or median discussed earlier. The mean does not always provide an accurate reflection of the data due to the presence of extreme items. Median may also prove to be quite unrepresentative of the data owing to uneven distribution of the series. For example, the values in the lower half of a distribution range from, say, Rs. 10 to Rs. 100 while the same number of items in the upper half of the series range from Rs. 100 to Rs. 6,000 with most of them near the higher limit. In such a distribution the median value of Rs. 100 will provide little indication of the true nature of the data.

Both these shortcomings may be overcome by the use of mode which refers to the value which occurs most frequently in a distribution. Moreover, mode is simplest to compute since it is the value corresponding to the highest frequency. For example, if the data are:

Size of shoe	5	6	7	8	9	10	11
No. of persons	10	20	25	40	22	15	6

The modal size is '8' since more persons are wearing this size compared to any other size.

Calculation of Mode

Determining the precise value of the mode of a frequency distribution is by no means an elementary calculation. Essentially, it involves fitting mathematically of some appropriate type of frequency curve to the grouped data and the determination of the value on the X-axis below the peak of the curve. However, there are several elementary methods of *estimating* the mode. These methods have been discussed for individual observations, discrete series and continuous series.

Calculation of Mode — Individual Observations

For determining mode count the number of times the various values repeat themselves and the value which occurs the maximum number of times is the modal value. The more often the modal value appears relatively, the more variable the measure is as an average to represent data.

Example 1: Find Mode from the following data:

110, 120, 130, 120, 110, 140, 130, 120, 140, 120

Solution:

Value	Tally Bars	Frequency
110	||	2
120	||||	4
130	||	2
140	||	2
		Total 10

Notes

Since the value 120 occurs the maximum numbers of times, *i.e.*, 4, hence the modal value is 120.

Notes Thus the process of determining mode in case of individual observations essentially involves grouping of data.

When there are two or more values having the same maximum frequency one cannot say which is the modal value and hence mode is said to be ill-defined. Such a series is also known as bimodal or multimodal. For example, observe the following data:

Income (in Rs.) 610, 620, 630, 620, 610, 640, 630, 620, 630, 640.

Size of item	No. of times it occurs
610	2
620	3
630	3
640	2

Mode is ill-defined in this case.

Calculation of Mode — Discrete Series

In discrete series quite often mode can be determined just by inspection, *i.e.,* by looking to that value of the variable around which the items are most heavily concentrated. For example, observe the following data:

Size of garment	No. of persons
28	10
29	20
30	40
31	65
32	50
33	15

From the above data we can clearly say that the modal size is 31 because the value 31 has occurred the maximum number of times, *i.e.,* 65. However, where the mode is determined just by inspection, an error of judgment is possible in those cases where the difference between the maximum frequency and the frequency preceding it or succeeding it is very small and the items are heavily concentrated on either side. In such cases it is desirable to prepare a grouping table and an analysis table. These tables help us ascertaining the modal class.

A grouping table has six columns. In column 1 the maximum frequency is marked or put in a circle; in column 2 frequencies are grouped in two's, in column 3 leave the first frequency and then group remaining in two's; in column 4 group the frequencies in three's; in column 5 leave the first frequency and group the remaining in three's; and in column 6 leave the first two frequencies and then group the remaining in three's. In each of these cases take the maximum total and mark it in a circle or by bold type.

Notes

After preparing the grouping table, prepare an analysis table. While preparing this table put column numbers on the left-hand side and the various probable values of mode on the right-hand side. The values against which frequencies are the highest and marked in the grouping table are then entered by means of a bar in the relevant 'box' corresponding to the values they represent.

The procedure of preparing grouping table and analysis table shall be clear from the following example:

Example 2: From the following data of the weight of 100 persons in a commercial concern determined the modal weight:

Weight (in *kg*)	58	60	61	62	63	64	65	66	68	70
No. of persons	4	6	5	10	20	22	24	6	2	1

Solution:

Grouping Table

Weight in	**Frequency**					
(*kg*)	**Col. 1**	**Col. 2**	**Col. 3**	**Col. 4**	**Col. 5**	**Col. 6**
58	4	10		15		
60	6		11		21	35
61	5	15				
62	10		30	**52**		
63	20	**42**			**66**	
64	22		**46**			**52**
65	**24**	30		32		
66	6		8		9	
68	2	3				
70	1					

Analysis Table

Col. No.	**Weight in *kg*.**										
	58	60	61	62	63	64	65	66	67	68	70
I							1				
II					1	1					
III						1	1				
IV				1	1	1					
V					1	1	1				
VI						1	1	1			
Total				1	3	5	4	1			

Notes

Since the value 64 has been repeated the maximum number of times, *i.e.*, 5, the modal weight is 64 *kg*. It should be noted that by inspection one is likely to say that the modal value is 65 since it has the highest frequency, *i.e.,* 24. But this is incorrect as revealed by the analysis table and grouping table.

Calculation of Mode — Continuous Series

Steps:

(i) By preparing grouping table and analysis table or by inspection ascertain the modal class.

(ii) Determine the value of mode by applying the following formula:

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i \quad \text{... (i)}$$

where L = Lower limit of the modal class; Δ_1 = the difference between the frequency of the modal class the frequency of the pre-modal class, *i.e.,* preceding class (ignoring signs); Δ_2 = the difference between the frequency of the modal class and the frequency of the post-modal class, *i.e.,* succeeding class (ignoring signs); i = the size of the class-interval of the modal class.

Another form of this formula is:

$$M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \quad \text{... (ii)}$$

where L = Lower limit of the modal class; f_1 = frequency of the modal class, f_0 = frequency of the class preceding the modal class; f_2 = frequency of the class succeeding the modal class.

While applying the above formula for calculating mode it is necessary to see that the class intervals are uniform throughout. It they are unequal they should first be made equal on the assumption that the frequencies are equally distributed throughout the class, otherwise we will get misleading results.

Did u know? In the latter case the value of mode cannot be determined by the above formula and hence mode is *ill-defined.*

A distribution having only one mode is called *unimodal*. If it contains more than one mode, it is called *bimodal* or *multimodal*. If collected data produce a bimodal distribution, the data themselves should be questioned. Quite often such a condition is caused when the size of the sample is small; the difficulty can be remedied by increasing the sample size. Another common cause is the use of non-homogeneous data. Instances where a distribution is bimodal and nothing can be done to change it, the mode should not be used as a measure of central tendency.

Where mode is ill-defined, its value may be ascertained by the following formula based upon the relationship between mean, median and mode.

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean} \quad \text{... (iii)}$$

Notes

Example 3: Calculate mode from the following data:

Marks	No. of Students	Marks	No. of Students
Above 0	80	Above 60	28
" 10	77	" 70	16
" 20	72	" 80	10
" 30	65	" 90	8
" 40	55	" 100	0
" 50	43		

Solution: Since this is a cumulative frequency distribution, we first convert it into a simple frequency distribution.

Marks	No. of Students
0 — 10	3
10 — 20	5
20 — 30	7
30 — 40	10
40 — 50	12
50 — 60	15
60 — 70	12
70 — 80	6
80 — 90	2
90 — 100	8

By inspection the modal class is 50 — 60.

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 50, \Delta_1 = (15 - 12) = 3, \Delta_2 = (15 - 12) = 3, i = 10$$

$$M_0 = 50 + \frac{3}{3+3} \times 10 = 50 + 5 = 55.$$

Example 4: From the following data of the weight of 122 persons determine the modal weight by grouping:

Weight (in *lb.*)	No. of persons	Weight (in *lb.*)	No. of persons
100 — 110	4	140 — 150	33
110 — 120	6	150 — 160	17
120 — 130	20	160 — 170	8
130 — 140	32	170 — 180	2

Solution: By inspection it is difficult to say which is the modal class. Hence, we prepare a grouping table and an analysis table.

Notes

Grouping Table

Weight (in *lb.*)	No. of persons					
	Col. 1	**Col. 2**	**Col. 3**	**Col. 4**	**Col. 5**	**Col. 6**
100 – 110	4	10	26	30	58	**85**
110 – 120	6					
120 – 130	20	**52**				
130 – 140	32		**65**			
140 – 150	**33**	50		**82**	**58**	27
150 – 160	17		25			
160 – 170	8	10				
170 – 180	2					

Analysis Table

Col. No.	Class in which mode is expected to lie		
	120 – 130	**130 – 140**	**140 – 150**
I			1
II	1	1	
III		1	1
IV		1	1
V	1	1	1
VI	1	1	1
Total	3	5	5

This is a bimodal series. Hence mode has to be determined indirectly by applying the formula:

$$\text{Mode} = 3\ \text{Median} - 2\ \text{Mean}.$$

$$\text{Median} = \text{Size of } \frac{N}{2}\text{th item} = \frac{122}{2} = 61\text{st item}.$$

Hence median lies in the class 130 – 140.

$$\text{Median} = L + \frac{N/2 - c.f.}{f} \times i$$

$L = 130$, $N/2 = 61$; $c.f. = 30$, $f = 32$, $i = 10$.

$$\text{Median} = 130 + \frac{61 - 30}{32} \times 10 = 130 + \frac{310}{32} = 130 + 9.69 = 139.69\ lb.$$

Notes

Calculation of Mean

Weight in *lb*.	*m*	No. of persons *f*	(*m* – 135)/10 *d*	*fd*
100 — 110	105	4	– 3	– 12
110 — 120	115	6	– 2	– 12
120 — 130	125	20	–1	– 20
130 — 140	135	32	0	0
140 — 150	145	33	+ 1	+ 33
150 — 160	155	17	+ 2	+ 34
160 — 170	165	8	+ 3	+ 24
170 — 180	175	2	+ 4	+ 8
		N = 122		$\sum fd$ = 55

$$\bar{X} = A + \frac{\sum fd}{N} \times i$$

A = 135, $\sum fd$ = 55, N = 122, i = 10

$$\bar{X} = 135 + \frac{55}{122} \times 10 = 135 + 4.51 = 139.51.$$

Mode = 3 Median – 2 Mean.

Mode = (3 × 139.69) – (2 × 139.51) = 419.07 – 279.02 = 140.05

Hence modal weight is 140.05 *lbs*.

Mode when Class Intervals are Unequal

The formula for calculating the value of mode given above is applicable only where there are equal class intervals. If the class intervals are unequal then we must make them equal before we start computing the value of mode. The class interval should be made equal and frequencies adjusted on the assumption that they are equally distributed throughout the class.

Example 5: Calculate the modal income for the following data:

Income (Rs. per month)	No. of Employees
2000 — 2500	8
2500 — 3000	12
3000 — 4000	30
4000 — 4500	3
4500 — 5000	2

Solution: Since class intervals are not equal throughout, we will take 500 as class interval and adjust the frequencies of those classes whose class interval is more than 500. The adjusted frequency distribution is as follows:

Notes

Income (Rs. per month)	No. of Workers
2000 — 2500	8
2500 — 3000	12
3000 — 3500	15
3500 — 4000	15
4000 — 4500	3
4500 — 5000	2

Its is clear from that the mode lies in the class 3000 – 3500.

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 3000, \Delta_1 = (15 - 12) = 3, \Delta_2 = (15 - 15) = 0, i = 20$$

$$M_0 = 3000 + \frac{3}{3+0} \times 500 = 3000 + 500 = 3500$$

Hence modal income = Rs. 3500.

Locating Mode Graphically

In a frequency distribution the value of mode can also be determined graphically. The steps in calculation are:

1. Draw a histogram of the given data.

2. Draw two lines diagonally in the inside of the modal class bar, starting from each upper corner of the bar to the upper corner of the adjacent bar.

3. Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis (horizontal scale) which gives us the modal value.

Example 6: The monthly profits in rupees of 100 shops are distributed as follows:

Profits (Rs.)	No. of shops	Profits (Rs.)	No. of shops
0 — 100	13	300 — 400	20
100 — 200	18	400 — 500	17
200 — 300	27	500 — 600	6

Draw the histogram of the data and hence find the modal value. Check this value by direct calculation.

Solution:

Direct calculation:

Mode lies in the class 200 — 300

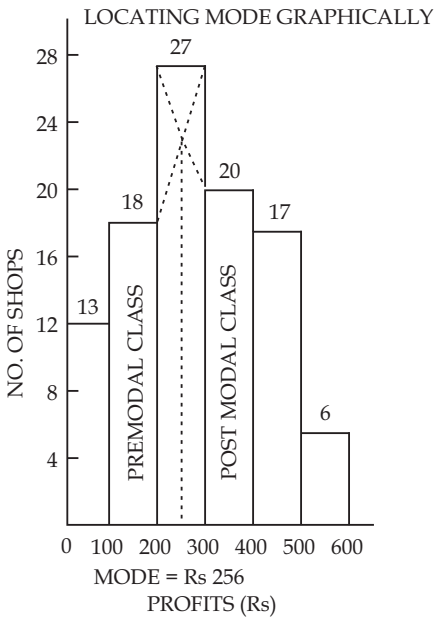
$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 200, \Delta_1 = (27 - 18) = 9, \Delta_2 = (27 - 20) = 7, i = 100.$$

$$M_0 = 200 + \frac{9}{9+7} \times 100 = 200 + 56.25 = 256.25.$$

Notes

Mode can also be determined from a frequency polygon in which case a perpendicular is drawn on the base from the apex of the polygon and the point where it meets the base gives the modal value.



However, graphic method of determining mode can be used only where there is one class containing the highest frequency. If two or more classes have the same highest frequency, mode cannot be determined graphically. For example, for the data given below mode cannot be graphically ascertained.

Size of shoe	No. of persons	Size of shoe	No. of persons
2 — 4	10	8 — 10	8
4 — 6	15	10 — 12	2
6 — 8	15		

Self-Assessment

1. Fill in the Blanks

(*i*) Median is better suited for ..... interval series.

(*ii*) In moderately a symmetrical distributions, the distance between the ..... and the ..... is about ..... the distance between the ..... and the ..... .

(*iii*) Given mean 25, mode 24, the median would be ..... .

(*iv*) The mode of distribution is the value that has the greatest ..... of ..... .

(*v*) In a symmetrical distribution mean ..... median ..... mode.

5.4 Summary

- The most popular and widely used measure for representing the entire data by one value is what most laymen call an 'average' and what statisticians call the arithmetic mean. Its value is obtained by adding together all the items and by the dividing this total by the number of items.

Notes

- A discrete series is obtained from a large number of individual observations. Suppose the marks obtained by 100 students is given. This data can be converted into a discrete series where the marks obtained are accompained by the number of students obtaining it.
- The continuous series expresse the data which is very vast. The calculation of arithmetic mean of this series is similar to that of discrete series after calculating the mid point of each segment of the continuous series which is called the class interval.
- As distinct from the arithmetic mean which is calculated from the *value of every* item in the series, the median is what is called a *positional* average. The term 'position' refers to the place of a value in a series. The place of the median in a series is such that an equal number of items lie on either side of it.
- The mode is often said to be the value which occurs most frequently. While this statement is quite helpful in interpreting the mode, it cannot safely be applied to any distribution, because of the vagaries of sampling. Even fairly large samples drawn from a statistical population with a single well-defined mode may exhibit very erratic fluctuations. Hence, mode should be thought as the value which has the greatest density *in its immediate neighbourhood*. For this reason mode is also called the most typical or fashionable value of a distribution.
- Determining the precise value of the mode of a frequency distribution is by no means an elementary calculation. Essentially, it involves fitting mathematically of some appropriate type of frequency curve to the grouped data and the determination of the value on the X-axis below the peak of the curve. However, there are several elementary methods of *estimating* the mode.
- A distribution having only one mode is called *unimodal*. If it contains more than one mode, it is called *bimodal* or *multimodal*. In the latter case the value of mode cannot be determined by the above formula and hence mode is *ill-defined.* If collected data produce a bimodal distribution, the data themselves should be questioned. Quite often such a condition is caused when the size of the sample is small; the difficulty can be remedied by increasing the sample size. Another common cause is the use of non-homogeneous data. Instances where a distribution is bimodal and nothing can be done to change it, the mode should not be used as a measure of central tendency.
- The formula for calculating the value of mode given above is applicable only where there are equal class intervals. If the class intervals are unequal then we must make them equal before we start computing the value of mode. The class interval should be made equal and frequencies adjusted on the assumption that they are equally distributed throughout the class.

5.5 Key-Words

1. Mean : In statistics, mean has three related meanings:
   (*i*) the arithmetic mean of a sample (distinguished from the geometric mean or harmonic mean).
   the expected value of a random variable.
   the mean of a probability distribution.

   There are other statistical measures of central tendency that should not be confused with means - including the 'median' and 'mode'. Statistical analyses also commonly use measures of dispersion, such as the range, interquartile range, or standard deviation. Note that not every probability distribution has a defined mean; see the Cauchy distribution for an example.
2. Median: In statistics and probability theory, median is described as the numerical value separating the higher half of a sample, a population, or a probability distribution, from the lower

Notes

half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values. A median is only defined on one-dimensional data, and is independent of any distance metric. A geometric median, on the other hand, is defined in any number of dimensions.

In a sample of data, or a finite population, there may be no member of the sample whose value is identical to the median (in the case of an even sample size); if there is such a member, there may be more than one so that the median may not uniquely identify a sample member. Nonetheless, the value of the median is uniquely determined with the usual definition. A related concept, in which the outcome is forced to correspond to a member of the sample, is the medoid…

3. Mode : The mode is the value that appears most often in a set of data.

Like the statistical mean and median, the mode is a way of expressing, in a single number, important information about a random variable or a population. The numerical value of the mode is the same as that of the mean and median in a normal distribution, and it may be very different in highly skewed distributions. The mode is not necessarily unique, since the same maximum frequency may be attained at different values. The most extreme case occurs in uniform distributions, where all values occur equally frequently. The mode of a discrete probability distribution is the value x at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

The mode of a continuous probability distribution is the value x at which its probability density function has its maximum value, so, informally speaking, the mode is at the peak.

5.6 Review Questions

1. Give two examples where arithmetic mean and median would be most appropriate average.
2. Can the value of mean, mode and median be the same in a symmetrical distribution ? If yes, state the situation.
3. Discuss the application mean and median.
4. How do you determine median and mode graphically ?
5. 'The arithmetic mean is the best among all the averages.' Give reasons.

Answers: Self-Assessment

1. (*i*) Positional (*ii*) mean, median, 1/3, mean, mode
   (*iii*) 24.67 (*iv*) concentration, frequencies
   (*v*) is equal to, is equal to

5.7 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 6: Dispersion: Meaning and Characteristics, Absolute and Relative Measures of Dispersion including Range, Quartile Deviation and Percentile

CONTENTS

Objectives

After reading this unit students will be able to:

- Know the Meaning and Characteristics of Dispersion.
- Explain Absolute and Relative Measures of Dispersion.
- Discuss Range, Quartile Deviation and Percentile.

Introduction

Series of date definitely have a great utility but they fail to reveal many facts about the phenomenon. There may be many different series, whose average/mean may come out to be identicle. But when they are studied in depth, they reveal entirely different stories. For example, the income of 5 people is — Rs. 50, 50, 50, 50, 50. Then the average will come out to be Rs. 50. The incomes of another five (5) people are Rs. 20, 80, 25, 25, 100. The average would again come out to be Rs. 50. Now if we consider incomes of another five people, they are Rs. 150, 20, 10, 10, 60. This would again average to Rs. 50 only. But in all the three cases, the average does not seem to represent the data fully. In the first case, the incomes are equal, in the second case, they have less variations in income but in the third case, there is vast variation of income. Therefore concluding about the data only on the basis of averages, considering it to be representative of the series may be misleading. Therefore, there is a need to measure the variations in the data. These variations are also called dispersion. Measures of central tendency are based on items of a series, therefore, they are called 'averages of the first order'. Measures of dispersion, on the other hand, are based upon average of the deviations of the different values from mean.

They are therefore called 'averages of the second order'.

6.1 Meaning and Characteristics of Dispersion

Meaning and Definition of Dispersion

The term dispersion refers to the variability of the size of items. Dispersion explains the size of various items in a series are not uniform rather, they vary. For example, if in a series the lowest and highest values vary only a little, the dispersion is said to be low. But if this variation is very high, dispersion is said to be considerable. In a series of ten students, the marks obtained are 10, 6, 8, 5, 10, 10, 8, 10, 5, 8.

Notes

(the average = 8). In another class, 10 students obtained the following marks. 10, 10, 5, 2, 10, 10, 3, 10, 10 (the average = 8). The dispersion in the second case is more because the size of items in this series vary considerably, inspite of the fact that the averages of the two have come out to be 8. Some of the important definitions of dispersion are — As per Brooks and Dick, "Dispersion or spread is the degree of the scatter or variations of the variable about a central value." A. L. Bowley defines dispersion as — "Dispersion is the measure of variations of the item." In the words of Prof. L. R. Connor, "Dispersion is a measure of the extent to which the individual items vary." According to Spriegel, "The degree to which numerical data tend to spread about an average value is called the variation or dispersion of data."

All the above definitions suggest that the term dispersion refers to the variability in the size of items. This variability is measured with respect to the average of the series. Therefore measures of dispersion are also termed as averages of the second order.

Did u know? "A measure of variation or dispersion describes the degree of scatter shown by observations and is usually measured by comparing the individual values of the variable with the average of all the values and then calculating the average of all the individual differences.

Characteristics of Dispersion

There are four basic characteristics of dispersion:

(1) **To guage the reliability of the average:** Even after making all the efforts to obtain the most representative average, the efforts prove to be successful when the data is homogeneous. In the absence of homogeneity, a measure of dispersion presents a better description about the structure of the distribution and the place of individual items in it. Therefore, in case of heterogeneous data, dispersion is measured to guage the reliability of the average calculated. When the value of dispersion is small, it is concluded that the average closely represents the data but when value of dispersion comes out to be large, it should be concluded that the average obtained is not very reliable.

(2) **To make a comparative study of the variability of series:** The consistency of uniformity of two series can be compared with the help of dispersion. If the value of dispersion measured comes out to be large, it may be concluded that the series lacks uniformity or consistency. Such studies are very useful in many fields like profit of companies, share values, performance individuals, studies related to demand, supply, prices etc.

(3) **To identify the factors causing variability so that it can be controlled:** Another important purpose of calculating dispersion is to identify the nature and causes of variations in a given data so that measures to control these can be suggested. Thus measures of dispersion are not merely supplementary to the averages, describing their reliability rather, they significantly disclose the quality of data in terms of homogeneity and consistency. They help to evaluate the various causes of heterogeneity and inconsistencies and suggest ways to control these. For example, in industrial production, efficient operation requires control of variation, the causes of which are sought through, inspection and quality control programmes." In social sciences, the measurement of inequality in the distribution of income and wealth requires the measures of variation.

(4) **To serve as a basis for further statistical analysis:** Yet another purpose of measures of dispersion is to help the statistician in carrying out further statistical analysis of the data like studying correlation, regression, testing of hypothesis, analysis of time series etc.

On the basis of the above, it can be concluded that due to inconsistencies and lack of uniformity of the data, averages can not prove to be closely representing the data, in most of the cases. In such a situation, dispersion presents a more better picture about the data, and gives logic to

Notes

find out whether the average calculated is reliable or not. It also helps in comparing the two series and also help in finding out ways to control the variations. In this way dispersion is a very strong tool into the hands of statisticians to know about the structure of data more closely and reliably.

Properties of a Good Measure of Dispersion

Just like the properties of a good measure of central tendency, properties of a good measure of dispersion are:

W. A. Sppur and *C. P. Bonim: Statistical Analysis for Business Decision.*

(1) It should be simple to understand.
(2) It must be easy to calculate.
(3) It must be based on all the items of the series.
(4) It should not be unduly affected by the extreme items.
(5) It should be least affected by the fluctuations in sampling.
(6) It should be capable of further statistical treatment.

6.2 Absolute and Relative Measures of Dispersion

Absolute measures of dispersion: When the dispersion of a series is calculated in terms of the absolute or actual figures in the data and the value of dispersion obtained can be expressed in the same units as the items of data are expressed, such measures are called absolute measures of dispersion. For example, if we calculate dispersion of a series indicating the income of group of persons in rupees, and the value of dispersion is obtained in rupees, it is termed as absolute measure of dispersion.

Relative measures of dispersion: When the value of dispersion is calculated as ratio or percentage of the average it is called relative measure of dispersion.

6.3 Range, Quartile Deviation and Percentile

Range

'Range' is the simplest measure of dispersion which is determined by the two extreme values of the observations and it is the difference between the largest and the smallest value in a distribution.

Uses of Range: (1) Range is very useful in quality control measures taken by the production department. It is checked that the quality should not deteriorate beyond the set value of range. Control charts are prepared for the purpose. (2) Another area where 'range' is very useful is the study of fluctuation of data. Variations in the weather forecasts, movement in the prices of securities etc. can be studied effectively and efficiently with the help of range.

Merits/advantages: (1) Simple to understand and easy to calculate, (2) It presents a broad picture of the data.

Demerits/disadvantages: (1) Gets affected by the extreme items, (2) does not take into consideration towards most of the items and their deviations, (3) Does not give reasonable picture of the data, (4) It is influenced by fluctuations of sampling.

Formula

$$\text{Range} = \text{Largest Value} - \text{Smallest Value}$$

or

$$\text{Maximum} - \text{Minimum}$$

or

$$\text{Range} = L - S$$

or

$$M_1 - M_0$$

Notes

$$\text{Coefficient of Range} = \frac{\text{Largest Value} - \text{Smallest Value}}{\text{Largest Value} + \text{Smallest Value}}$$

or

$$\frac{L - S}{L + S}$$

Individual Series

Example 1: Find out absolute and relative dispersion of range from the following observations:

Marks: 63, 68, 71.5, 83, 50, 27, 64, 38, 40

Solution: Absolute measure of Range = Largest – Smallest

= 83 – 27 = 56 marks

$$\text{Relative Measure} = \frac{\text{Largest} - \text{Smallest}}{\text{Largest} + \text{Smallest}}$$

$$= \frac{83 - 27}{83 + 27} = \frac{56}{110} = 0.509$$

Example 2: Calculate Range and its coefficient of the following series:

S. No.	1	2	3	4	5	6	7	8	9	10
Values	391	384	591	407	672	522	777	733	2488	1490

Solution: Largest = 2488, Smallest = 384,

Range = L – S

2488 – 384

= 2104

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

$$= \frac{2488 - 384}{2488 + 384} = \frac{2104}{2872} = 0.7325$$

Example 3: The yearly income of a person for the last ten years in given below. Find the range and its coefficient.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Income ('000 Rs.)	40	30	80	100	80	90	120	110	130	150

Solution: Range = L – S from the data, L = 150, S = 30.

$\therefore$ Range = 150 – 30 = 120

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

or

$$= \frac{150 - 30}{150 + 30} = \frac{120}{180}$$

= 0.66

Notes

Answer: The range for the above data is Rs. 1,20,000 per year and the coefficient of range is 0.66.

Range in Discrete Series

Example 4 : Find out absolute and relative measure of range in the following distribution:

Scores	Frequency
2	6
3	7
4	3
5	11
6	4
7	2
8	5
9	8
10	3

Solution: Absolute Measure of Range = L – S

= 10 – 2 = 8

$$\text{Relative Measure} = \frac{L - S}{L + S}$$

$$= \frac{10 - 2}{10 + 2} = \frac{8}{12} = 0.666$$

Example 5: Calculate Coefficient of Range of the following series:

Size	Frequency
2.5	7
3.5	1
4.5	3
5.5	5
6.5	4
7.5	9
8.5	8
9.5	11
10.5	4

Solution : L = 10.5

S = 2.5

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{10.5 - 2.5}{10.5 + 2.5} = \frac{8}{13} = 0.615$$

Range in Continuous Series

Range can be calculated by the following two methods:

(1) Class mark of the highest class – class mark of the lowest class.

(2) Upper class boundary of the highest class – lower class boundary of the lowest class.

Example 6: Calculate Coefficient of range from the following distribution:

Notes

Marks	No. of Students
25—30	6
30—35	3
35—40	12
40—45	8
45—50	22
50—55	9
55—60	5

Solution: The above question can be calculated by:

(1) Limits Method: L = 60, S = 25

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{60 - 25}{60 + 25} = \frac{35}{85} = 0.411$$

(2) Mid-Value Method: L = 57.5, S = 27.5

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{57.5 - 27.5}{57.5 + 27.5} = \frac{30}{85} = 0.352$$

Example 7: Find out relative dispersion of range from the following frequency table:

Marks	Frequency
5—10	4
10—15	6
15—20	20
20—25	7
25—30	5
30—35	8
35—40	6
40—45	5
45—50	2

Solution: For the calculation of relative dispersion of range, mean, median and mode will be calculated.

Marks	*f*	*c.f.*	*m.v.*	*dx*	*fdx*
5—10	4	4	7.5	– 5	– 20
10—15	6	10	12.5	– 4	– 24
15—20	20	30	17.5	– 3	– 60
20—25	7	37	22.5	– 2	– 14
25—30	5	42	27.5	–1	– 7
30—35	8	50	32.5	0	0
35—40	6	56	37.5	1	6
40—45	5	61	42.5	2	10
45—50	2	63	47.5	3	6
	N = 63				Σfdx = – 103

Notes

$$\text{Mean} = x + \frac{\sum fdx}{n} \times i$$

$$= 32.5 + \frac{-103}{63} \times 5$$

$$= 32.5 + \frac{-515}{63}$$

$$= 32.5 - 8.17 = 24.33 \text{ approx.}$$

$$\text{Median No.} = \frac{N}{2} = \frac{63}{2} = 31.5$$

$$\text{Median} = l + \frac{i}{f}(m - c)$$

$$= 20 + \frac{5}{7}(31.5 - 30)$$

$$= 20 + \frac{7.5}{7}$$

$$= 20 + 1.07 = 21.07$$

$$\text{Mode} = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 15 + \frac{20 - 6}{40 - 6 - 7} \times 5$$

$$= 15 + \frac{14 \times 5}{27}$$

$$= 15 + \frac{70}{27} = 15 + 2.59 = 17.59$$

Coefficient of Range Dispersion

(1) By using Mean $\frac{50 - 5}{24.33} = \frac{45}{24.33} = 1.84$

(2) By using Double the Mean $\frac{50 - 5}{24.33 \times 2} = \frac{45}{48.66} = 0.92$

(3) By using Median $\frac{50 - 5}{21.07} = \frac{45}{21.07} = 2.13$

(4) By using Mode $\frac{50 - 5}{17.59} = \frac{45}{17.59} = 2.55$

(5) By using sum of the extremes $\frac{50 - 5}{50 + 5} = \frac{45}{55} = 0.81$

Quartile Deviation or Semi-Interquartile Range

Notes

Quartile is the location-based measure of dispersion. It measures the average amount by which the first and the third quartiles deviate from the second quartile *i.e.*, median.

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}, \text{ Coefficient of variation} = \frac{\text{Q.D.}}{\text{Median}} \times 100,$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Merits of Q.D.: (1) Easy to compute. (2) It is very useful to know the variability at the centre of the data. (3) It is not much affected by the extreme items. (4) It can be calculated from open end distribution or from a skewed distribution.

Demerits of Q.D.: (1) Based only on the middle part of the data. (2) It is not capable of further mathematical treatment. (3) It is greatly affected by changes in sampling. (4) Gives no indication about variation occurring beyond Q_3 and Q_1.

Third Moment of Dispersion: In this method the deviations of items from mean are cubed, *i.e.,*

$$\text{Third moment of dispersion} = \frac{\sum d^3}{N}$$

$$\text{Coefficient of third moment of dispersion} = \frac{\frac{\sum d^3}{N}}{\sigma}$$

(5) It provides unit of measurement for the normal distribution.

Demerits: (1) If the data is vast, it involves tedious calculations.

Formula

$$\text{Quartile Deviation (Q.D.)} = \frac{Q_3 - Q_1}{2}$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

where Q_1 represents First Quartile

Q_3 represents Third Quartile

Q.D. in Individual Series

Example 8: Calculate Q.D. and its coefficient from the following observations relating to marks of 15 students:

48, 52, 56, 62, 66, 47, 51, 58, 60, 66, 68, 70, 64, 73, 63.

Solution: **Array:** 47, 48, 51, 52, 56, 58, 60, 62, 63, 64, 66, 66, 68, 70, 73

$$Q_1 = \text{Value of the } \left[\frac{N+1}{4}\right]^{th} \text{ item}$$

$$= \left[\frac{15+1}{4}\right]^{th} \text{ item } i.e., \ 4^{th} \text{ item} = 52$$

Notes

$$Q_3 = \text{Value of the } \left[3\left(\frac{N+1}{4}\right)\right]^{th} \text{ item}$$

$$= \left[\frac{3(15+1)}{4}\right]^{th} \text{ item } i.e., 12^{th} \text{ item} = 66$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{66 - 52}{2} = \frac{14}{2} = 7$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{66 - 52}{66 + 52} = \frac{14}{118} = 0.118$$

Q.D. in Discrete Series

Example 9: Calculate quartile deviation and its coefficient from the following data:

Height of students (in *cms*.)	120	122	124	126	130	140	150	160
No. of students	1	3	5	7	10	3	1	1

Solution:

Calculation of Q_3 and Q_1

X	***f***	***c.f.***
120	1	1
122	3	4
124	5	9
126	7	16
130	10	26
140	3	29
150	1	30
160	1	31

$$Q_1 = \text{Size of } \left(\frac{N+1}{4}\right)^{th} \text{ item} = 8^{th} \text{ item}$$

$$\therefore \quad Q_1 = 224.$$

$$Q_3 = \text{Size of } 3\left(\frac{N+1}{4}\right)^{th} \text{ item} = 24^{th} \text{ item}$$

$$\therefore \quad Q_3 = 230$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{230 - 224}{2} = 3 \text{ } cms.$$

Notes

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{230 - 224}{230 + 224} = 0.01321.$$

Answer: Q.D. for the above data is found to be 3 *cms.* and coefficient of Q.D. = 0.01321.

Q.D. in Continuous Series

Example 10: Find quartile deviation and its relative measure:

Variable	Frequency	Variable	Frequency
20—29	306	50—59	96
30—39	182	60—69	42
40—49	144	70—79	32

Solution:

Calculation OF Q.D.

Variable	*f*	*c.f.*
20—29	306	306
30—39	182	488
40—49	144	632
50—59	96	728
60—69	42	770
70—79	34	804

$$Q_1 = \left(\frac{805}{4}\right)^{th} \text{ item} = 201.25^{th} \text{ item}$$

which lies in class 20—29 or class 19.5 – 29.5

$$Q_1 = 19.5 + \left(\frac{10}{306} \times 201\right) = 26.07$$

$$Q_3 = \frac{3(805)^{th}}{4} \text{ item} = 603.75^{th} \text{ item.}$$

which lies in 40—49 class or 39.5 — 49.5.

$$Q_3 = 39.5 + \left(\frac{10}{144} + 155\right) = 47.49$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{47.49 - 26.07}{2} = 10.71$$

Notes

$$\text{Coeff. of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{47.49 - 26.07}{47.49 + 26.07} = 0.2912$$

Answer: For the given data Q.D. = 10.71 and coeff. of Q.D. = 0.2912.

Example 11: Compute Quartile Deviation and its coefficient from the following data:

Mid-Value	3	4	5	6	7	8	9
Frequency	11	14	20	24	20	16	5

Solution:

Class	*f*	*C.f.*
2.5—3.5	11	11
3.5—4.5	14	25
4.5—5.5	20	45
5.5—6.5	24	69
6.5—7.5	20	89
7.5—8.5	16	105
8.5—9.5	5	110
Total	**N = 110**	

$$Q_1 = \text{Size of } \frac{N}{4} = \frac{110}{4} = 27.5^{th} \text{ item}$$

27.5^{th} item which lies in 4.5—5.5 group

$$Q_1 = l_1 + \frac{i}{f}(q_1 - c)$$

$$= 4.5 + \frac{1}{20}(27.5 - 25)$$

$$= 4.5 + \frac{2.5}{20}$$

$$= 4.5 + 0.125 = 4.625$$

$$Q_3 = \text{Size of } \frac{3N}{4} = \frac{3 \times 110}{4} = \frac{330}{4} = 82.5^{th} \text{ item}$$

82.5^{th} item which lies in 6.5 — 7.5 group.

$$Q_3 = l_1 + \frac{i}{f}(q_3 - c)$$

$$= 6.5 + \frac{1}{20}(82.5 - 69)$$

Notes

$$= 6.5 + \frac{13.5}{20}$$

$$= 6.5 + 0.675 = 7.175$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{7.175 - 4.625}{2} = \frac{2.55}{2} = 1.275$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{7.175 - 4.625}{7.175 + 4.625}$$

$$= \frac{2.55}{11.8} = 0.216$$

Example 12: For a distribution, the coefficient of quartile deviation = 0.4, and the difference of two quartiles = 40. Find the values of quartiles.

Solution: Given: C of Q.D. = 0.4, $Q_3 - Q_1 = 40$, $Q_1 = ?$, $Q_3 = ?$

$$\text{C of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

or $$= 0.4 = \frac{40}{Q_3 + Q_1}$$

or $$Q_3 + Q_1 = \frac{40}{0.4} = 100$$

$$Q_3 - Q_1 = 40$$

$$\frac{Q_3 + Q_1 = 100}{2Q_3 = 140}$$

$\therefore$ $$Q_3 = 140/2 = 70, \; Q_1 = 100 - 70 = 30$$

Example 13: From the following data, calculate Quartile Coefficient of Dispersion.

Wages Less than	10	20	30	40	50	60	70
No. of Workers	5	8	15	20	30	33	35

Solution:

Wages in Rs.	**No. of workers *(f)***	***c.f.***
0—10	5	5
10—20	3	8
20—30	7	15
30—40	5	20
40—50	10	30
50—60	3	33
60—70	2	35
Total	**N = 35**	

Notes

$$Q_1 = \text{Size of } \frac{N}{4} = \frac{35}{4} = 8.75^{th} \text{ item}$$

8.75^{th} item which lies in 20—30 group.

$$Q_1 = l_1 + \frac{i}{f}(q_1 - c)$$

$$= 20 + \frac{10}{7}(8.75 - 8)$$

$$= 20 + \frac{10 \times .75}{7}$$

$$= 20 + 1.07 = 21.07$$

$$Q_3 = \text{Size of } \frac{3N}{4} = \frac{3 \times 35}{4} = \frac{105}{4} = 26.25^{th} \text{ item}$$

26.25^{th} item which lies in 40—50 group

$$Q_3 = l_1 + \frac{i}{f}(q_3 - c)$$

$$= 40 + \frac{10}{10}(26.25 - 20)$$

$$= 40 + \frac{10 \times 6.25}{10}$$

$$= 40 + 6.25 = 46.25$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{46.25 - 21.07}{2} = \frac{25.18}{2} = 12.59$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{46.25 - 21.07}{46.25 + 21.07} = \frac{25.18}{67.32}$$

$$= 0.374$$

Percentile

In statistics, **a percentile** (or centile) is the value of a variable below which a certain percent of observations fall. For example, the 20^{th} percentile is the value (or score) below which 20 percent of the observations may be found. The term percentile and the related term percentile rank are often used in the reporting of scores from norm-referenced tests. For example, if a score is in the 86^{th} percentile, it is higher than 85% of the other scores.

The 25^{th} percentile is also known as the first quartile (Q_1), the 50^{th} percentile as the median or second quartile (Q_2), the 75^{th} percentile as the third quartile (Q_3).

Notes

Definition

There is no standard definition of percentile, however all definitions yield similar results when the number of observations is very large.

Nearest rank

One definition of percentile, often given in texts, is that the P-th percentile $(0 \le P < 100)$ of N ordered values (arranged from least to greatest) is obtained by first calculating the (ordinal) rank

$$n = \frac{P}{100} \times N + \frac{1}{2}$$

rounding the result to the nearest integer, and then taking the value that corresponds to that rank.

(Note that the rounded value of n is just the least integer which exceeds $\frac{P}{100} \times N$.)

For example, by this definition, given the numbers

15, 20, 35, 40, 50

the rank of the 30th percentile would be

$$n = \frac{30}{100} \times 5 + \frac{1}{2} = 2.$$

Thus the 30th percentile is the second number in the sorted list, 20.

The 35th percentile would have rank

$$n = \frac{35}{100} \times 5 + \frac{1}{2} = 2.25,$$

so the 35th percentile would be the second number again (since 2.25 rounds down to 2) or 20

The 40th percentile would have rank

$$n = \frac{40}{100} \times 5 + \frac{1}{2} = 2.5,$$

so the 40th percentile would be the third number (since 2.5 rounds up to 3), or 35.

The 100th percentile is defined to be the largest value. (In this case we do not use the above definition with P = 100, because the rank n would be greater than the number N of values in the original list.)

In lists with fewer than 100 values the same number can occupy more than one percentile group.

Linear interpolation between closest ranks

An alternative to rounding used in many applications is to use **linear interpolation** between the two nearest ranks.

In particular, given the N sorted values $v_1 \le v_2 \le v_3 \le \ldots \le v_N$, we define the *percent rank* corresponding to the n^{th} value as:

$$p_n = \frac{100}{N}\left(n - \frac{1}{2}\right).$$

In this way, for example, if N = 5th percent rank corresponding to the third value is

$$p_3 = \frac{100}{5}\left(3 - \frac{1}{2}\right) = 50.$$

Notes

The value v of the P-th percentile may now be calculated as follows:

If $P < p_1$ or $P > p_N$, then we take $v = v_1$ or $v = v_N$, respectively.

If there is some integer k for which $P = p_k$, then we take $v = v_k$.

Otherwise, we find the integer k for which $p_k < P < p_{k+1}$, and take $v = v_k + \frac{P - p_k}{p_{k+1} - p_k}(v_{k+1} + v_k) =$

$v_k + N \times \frac{P - p_k}{100}(v_{k+1} - v_k)$.

Using the list of numbers above, the 40th percentile would be found by linearly interpolating between the 30th percentile, 20, and the 50th, 35. Specifically:

$$v = 20 + 5 \times \frac{40 - 30}{100}(35 - 20) = 27.5$$

This is halfway between 20 and 35, which one would expect since the rank was calculated above as 2.5.

It is readily confirmed that the 50th percentile of any list of values according to this definition of the P-th percentile is just the sample median.

Moreover, when N is even the 25th percentile according to this definition of the P-th percentile is the median of the first $\frac{N}{2}$ values (*i.e.*, the median of the lower half of the data).

Weighted percentile

In addition to the percentile function, there is also a *weighted percentile*, where the percentage in the total weight is counted instead of the total number. There is no standard function for a weighted percentile. One method extends the above approach in a natural way.

Suppose we have positive weights $w_1, w_2, w_3, ..., w_N$ associated, respectively, with our N sorted sample values. Let

$$S_n = \sum_{k=1}^{n} w_k ,$$

the n-th **partial sum** of the weights. Then the formulas above are generalized by taking

$$p_n = \frac{100}{S_N}\left(S_n - \frac{w_n}{2}\right)$$

and

$$v = v_k + \frac{p - p_k}{p_{k+1} - p_k}(v_{k+1} + v_k) .$$

Alternative methods

Some software packages, including **Microsoft Excel** (up to the version 2007) use the following method, noted as an alternative by NIST to estimate the value, v_p, of the P-th percentile of an ascending ordered dataset containing N elements with values $v_1 \le v_2 \le \ldots \le v_N$.

The rank is calculated:

$$n = \frac{P}{100}(N - 1) + 1$$

and then split into its integer component k and decimal component d, such that $n = k + d$.

Then v_p is calculated as:

$$v_p = \begin{cases} v_1, & \text{for } n = 1 \\ v_N, & \text{for } n = N \\ v_k + d(v_{k+1} - v_k), & \text{for } 1 =< n < N \end{cases}$$

The primary method recommended by **NIST** is similar to that given above, but with the rank calculated as

$$n = \frac{P}{100}(N + 1)$$

These two approaches give the rank of the 40th percentile in the above example as, respectively:

$$n = \frac{40}{100}(5 - 1) + 1 = 2.6$$

and

$$n = \frac{40}{100}(5 + 1) = 2.4$$

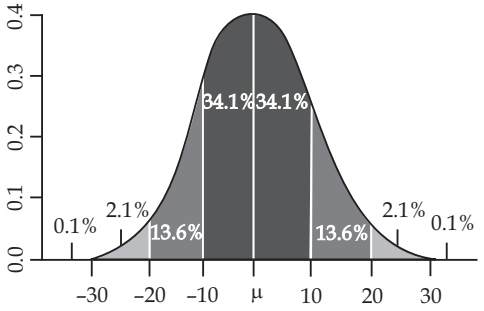
The values are then interpolated as usual based on these ranks, yielding 29 and 26, respectively, for the 40th percentile.

Applications

When **ISPs** bill **"burstable" internet bandwidth,** the 95th or 98th percentile usually cuts off the top 5% or 2% of bandwidth peaks in each month, and then bills at the nearest rate. In this way infrequent peaks are ignored, and the customer is charged in a fairer way. The reason this statistic is so useful in measuring data through put is that it gives a very accurate picture of the cost of the bandwidth. The 95th percentile says that 95% of the time, the usage is below this amount. Just the same, the remaining 5% of the time, the usage is above that amount.

Physicians will often use infant and children's **weight and height percentile** to assess their growth in comparison to national averages.

The normal curve and percentiles



The dark blue zone represents observations within one **standard deviation** (σ) to either side of the **mean** (μ), which accounts for about 68.2% of the population. Two standard deviations from the mean (dark and medium blue) account for about 95.4%, and three standard deviations (dark, medium, and light blue) for about 99.7%.

Notes

Notes

The methods given above are approximations for use in small-sample statistics. In general terms, for very large populations percentiles may often be represented by reference to a **normal curve** plot. The normal curve is plotted along an axis scaled to **standard deviation,** or sigma, units. Mathematically, the normal curve extends to negative **infinity** on the left and positive infinity on the right. Note, however, that a very small portion of individuals in a population will fall outside the – 3 to + 3 range.

In humans, for example, a small portion of all people can be expected to fall above the + 3 sigma height level.

Percentiles represent the area under the normal curve, increasing from left to right. Each standard deviation represents a fixed percentile. Thus, rounding to two decimal places, – 3 is the 0.13^{th} percentile, – 2 the 2.28^{th} percentile, – 1 the 15.87^{th} percentile, 0 the 50^{th} percentile (both the mean and median of the distribution), + 1 the 84.13^{th} percentile, + 2 the 97.72^{nd} percentile, and + 3 the 99.87^{th} percentile. Note that the 0^{th} percentile falls at negative infinity and the 100^{th} percentile at positive infinity.

Self-Assessment

1. Indicate whether the following statements are True or False:

(*i*) A good measure of dispersion is the one which is not defined rigidly.

(*ii*) Range is the best measure of dispersion.

(*iii*) Quartile Deviation is more suitable in case of openend distributions.

(*iv*) Absolute measure of dispersion can be used for purposes of comparison.

6.4 Summary

- The term dispersion refers to the variability of the size of items. Dispersion explains the size of various items in a series are not uniform rather, they vary. For example, if in a series the lowest and highest values vary only a little, the dispersion is said to be low.
- "A measure of variation or dispersion describes the degree of scatter shown by observations and is usually measured by comparing the individual values of the variable with the average of all the values and then calculating the average of all the individual differences.
- Therefore, in case of heterogeneous data, dispersion is measured to guage the reliability of the average calculated. When the value of dispersion is small, it is concluded that the average closely represents the data but when value of dispersion comes out to be large, it should be concluded that the average obtained is not very reliable.
- The consistency of uniformity of two series can be compared with the help of dispersion. If the value of dispersion measured comes out to be large, it may be concluded that the series lacks uniformity or consistency. Such studies are very useful in many fields like profit of companies, share values, performance individuals, studies related to demand, supply, prices etc.
- It can be concluded that due to inconsistencies and lack of uniformity of the data, averages can not prove to be closely representing the data, in most of the cases. In such a situation, dispersion presents a more better picture about the data, and gives logic to find out whether the average calculated is reliable or not. It also helps in comparing the two series and also help in finding out ways to control the variations. In this way dispersion is a very strong tool into the hands of statisticians to know about the structure of data more closely and reliably.
- When the dispersion of a series is calculated in terms of the absolute or actual figures in the data and the value of dispersion obtained can be expressed in the same units as the items of data are expressed, such measures are called absolute measures of dispersion. For example, if we calculate dispersion of a series indicating the income of group of persons in rupees, and the value of dispersion is obtained in rupees, it is termed as absolute measure of dispersion.

Notes

- Quartile is the location-based measure of dispersion. It measures the average amount by which the first and the third quartiles deviate from the second quartile *i.e.*, median.
- In statistics, **a percentile** (or centile) is the value of a variable below which a certain percent of observations fall. For example, the 20th percentile is the value (or score) below which 20 percent of the observations may be found. The term percentile and the related term percentile rank are often used in the reporting of scores from norm-referenced tests. For example, if a score is in the 86th percentile, it is higher than 85% of the other scores.
- In addition to the percentile function, there is also a *weighted percentile*, where the percentage in the total weight is counted instead of the total number. There is no standard function for a weighted percentile. One method extends the above approach in a natural way.
- When **ISPs** bill **"burstable" internet bandwidth,** the 95th or 98th percentile usually cuts off the top 5% or 2% of bandwidth peaks in each month, and then bills at the nearest rate. In this way infrequent peaks are ignored, and the customer is charged in a fairer way. The reason this statistic is so useful in measuring data through put is that it gives a very accurate picture of the cost of the bandwidth. The 95th percentile says that 95% of the time, the usage is below this amount. Just the same, the remaining 5% of the time, the usage is above that amount.
- In general terms, for very large populations percentiles may often be represented by reference to a **normal curve** plot. The normal curve is plotted along an axis scaled to **standard deviation,** or sigma, units. Mathematically, the normal curve extends to negative **infinity** on the left and positive infinity on the right. Note, however, that a very small portion of individuals in a population will fall outside the – 3 to + 3 range.
- Percentiles represent the area under the normal curve, increasing from left to right. Each standard deviation represents a fixed percentile. Thus, rounding to two decimal places, – 3 is the 0.13th percentile, – 2 the 2.28th percentile, – 1 the 15.87th percentile, 0 the 50th percentile (both the mean and median of the distribution), + 1 the 84.13th percentile, + 2 the 97.72nd percentile, and + 3 the 99.87th percentile. Note that the 0th percentile falls at negative infinity and the 100th percentile at positive infinity.

6.5 Key-Words

1. Absolute measures : Absolute measures of Dispersion are expressed in same units in which original data is presented but these measures cannot be used to compare the variations between the two series. Relative measures are not expressed in units but it is a pure number. It is the ratios of absolute dispersion to an appropriate average such as co-efficient of Standard Deviation or Co-efficient of Mean Deviation.

   Relative measures : These measures are calculated for the comparison of dispersion in two or more than two sets of observations. These measures are free of the units in which the original data is measured. If the original data is in dollar or kilometers, we do not use these units with relative measure of dispersion. These measures are a sort of ratio and are called coefficients. Each absolute measure of dispersion can be converted into its relative measure.

2. Quartile deviation : The quartile deviation is a slightly better measure of absolute dispersion than the range. But it ignores the observation on the tails. If we take difference samples from a population and calculate their quartile deviations, their values are quite likely to be sufficiently different. This is called sampling fluctuation. It is not a popular measure of dispersion. The quartile deviation calculated from the sample data does not help us to draw any conclusion (inference) about the quartile deviation in the population.

Notes

6.6 Review Questions

1. Explain with example the term 'Dispersion'. Give its definition and discuss the characteristics of dispersion.
2. What are the uses of 'range' as a method of measuring dispersion ? Give its advantages and disadvantages.
3. Give the merits and demerits of quartile deviation. What is the third movement of dispersion?
4. Define Percentile. Discuss the application of Percentile.

Answers: Self-Assessment

1. (*i*) F (*ii*) F (*iii*) T (*iv*) F

6.7 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Unit 7: Mean Deviation and Standard Deviation

Notes

CONTENTS

Objectives

After reading this unit students will be able to:

- Describe the Mean Deviation.
- Explain the Standard Deviation.

Introduction

The two methods of dispersion discussed earlier in this book, namely, range and quartile deviation, are not measures of dispersion in the strict sense of the term because they do not show the scatterness around an average. However, to study the formation of a distribution we should take the deviations from an average. The two other measures, namely, the average deviation and the standard deviation, help us in achieving this goal.

The average deviation is sometimes called the mean deviation. It is the average difference between the items in a distribution and the median or mean of that series. Theoretically, there is an advantage in taking the deviations from median because *the sum of the deviations of items from median is minimum when signs are ignored.* However, in practice the arithmetic mean is more frequently used in calculating the value of average deviation and this is the reason why it is more commonly called mean deviation. In any case, the average used must be clearly stated in a given problem so that any possible confusion in meaning is avoided.

The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler though practically less robust than the average absolute deviation. A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data.

Did u know? In statistics **standard deviation** (represented by the symbol sigma, σ) shows how much variation or "dispersion" exists from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.

In addition to expressing the variability of a population, standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be

Notes

conducted multiple times. The reported margin of error is typically about twice the standard deviation—the radius of a 95 percent confidence interval. In science, researchers commonly report the standard deviation of experimental data, and only effects that fall far outside the range of standard deviation are considered statistically significant—normal random error or variation in the measurements is in this way distinguished from causal variation. Standard deviation is also important in finance, where the standard deviation on the rate of return on an investment is a measure of the volatility of the investment.

7.1 The Mean Deviation

The average deviation or mean deviation is a measure of dispersion that is based upon all the items in a distribution. It is the arithmetic mean of the deviations of the data from its central value, may it be arithmetic mean, median or mode. While, considering the deviations from its central value, only absolute values are taken into consideration, *(i.e.,* without considering the positive or negative signs). Mean deviation is denoted by δ (delta).

Mean/Average Deviation (denoted by δ)

$$\delta_{\bar{X}} = \frac{\Sigma|d\bar{X}|}{N} \text{ (Deviation taken from Arithmetic Mean)}$$

$$\delta_{M} = \frac{\Sigma|dM|}{N} \text{ (Deviation taken from Median)}$$

$$\delta_{Mo} = \frac{\Sigma|dMo|}{N} \text{ (Deviation taken from Mode)}$$

Coefficient of Dispersion:

$$\text{Coeff. of } \delta_{\bar{X}} = \frac{\delta_{\bar{X}}}{\bar{X}} \ (\bar{X} \text{ is arithmetic mean})$$

$$\text{Coeff. of } \delta_{M} = \frac{\delta_{M}}{M} \text{ (M is Median)}$$

$$\text{Coeff. of } \delta_{Mo} = \frac{\delta_{Mo}}{Mo} \text{ (Mo is Mode)}$$

Example 1: A batch of 10 students obtained the following marks out of 100. Calculate the mean deviation and its coefficient.

Marks: 58, 39, 22, 11, 44, 28, 49, 55, 41 and 42.

Solution: The median value for the series is:

$$\left(\frac{N+1}{2}\right)^{th} \text{ item} = \frac{11}{2} = 5.5^{th} \text{ item.}$$

The series in ascending order:

11, 22, 28, 39, 41, 42, 44, 49, 55, 58

$$\therefore \text{ Median} = \frac{41+42}{2} = \frac{83}{2} = 41.5.$$

Calculation of Deviation from Median

Marks	Deviation (d_M)
11	11 – 41.5 = |30.5|
22	22 – 41.5 = |19.5|
28	28 – 41.5 = |13.5|

Notes

39	39 – 41.5 = |2.5|
41	41 – 41.5 = |.5|
42	42 – 41.5 = |.5|
44	44 – 41.5 = |2.5|
49	49 – 41.5 = |7.5|
55	55 – 41.5 = |13.5|
58	58 – 41.5 = |16.5|
	$\sum d_M = 107$

$$\text{Dispersion} = \frac{\sum d_M}{N},\ N = 10,\ \sum d_M = 107.$$

$$\therefore \quad \text{Dispersion} = \frac{107}{10} = 10.7 \text{ marks.}$$

$$\text{Coefficient} = \frac{\delta M}{M} = \frac{10.7}{41.5} = 0.26 \text{ marks.}$$

Answer: Mean deviation (from median) δ_M for the given data is 10.7 marks and the coefficient is 0.26.

Example 2: Calculate mean deviation from arithmetic mean from the following data:
10.500, 10.250, 10.375, 10.625, 10.750, 10.125, 10.375, 10.625, 10.500, 10.125.

Solution: When the data is in fractions and the mean value comes in fractions the following method may be used to avoid tedious calculations.

$$\delta_{\bar{X}} = \frac{1}{N}\left(\bar{X}_y - \bar{X}_x\right)$$

or

$$\delta_M = \frac{1}{N}\left(M_y - M_x\right)$$

where $\bar{X}_y / M_y$ is sum of items above Mean/Median $\bar{X}_x / M_y$ is sum below Mean/ Median.

Mean for the data given = $\frac{104.25}{10} = 10.425$.

The value of items above mean $\left(\bar{X}_y\right)$

10.500 + 10.500 + 10.625 + 10.625 + 10.750 = 53.

The value of items below mean $\left(\bar{X}_x\right)$

10.125 + 10.125 + 10.250 + 10.375 + 10.375 = 51.25.

$$\text{Mean deviation} = \frac{1}{10}(53 - 51.25)$$

$$= \frac{1.75}{10} = 0.175.$$

(This method is also called short-cut method of calculation Mean deviation).

Notes

Answer: The mean deviation from arithmetic mean of the given data is 0.175.

Example 3: Find the mean deviation for the following data (from median and mean)

Items	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	15	16	21	10	17	8	4	2	1	2	2	0	2

Solution: **To find Median**

Items in ascending order	**Frequency**	**Cumulative Frequency**
0	15	15
1	16	31
2	21	52
3	10	62
4	17	79
5	8	87
6	4	91
7	2	93
8	1	94
9	2	96
10	2	98
11	0	98
12	2	100
	$\Sigma f = 100$	

$$\text{Median} = \frac{(N+1)}{2}\text{th item} = 55.5\text{th item}$$

$\therefore$ Median = 2.

Deviation from Median

Items	f	**Deviation** (d_M)	fd_M
0	15	0 – 2 = |2|	2 × 15 = 30
1	16	1 – 2 = |1|	1 × 16 = 16
2	21	2 – 2 = |0|	00
3	10	3 – 2 = |1|	1 × 10 = 10
4	17	4 – 2 = |2|	2 × 17 = 34
5	8	5 – 2 = |3|	3 × 8 = 24
6	4	6 – 2 = |4|	4 × 4 = 16
7	2	7 – 2 = |5|	5 × 2 = 10
8	1	8 – 2 = |6|	6 × 1 = 6
9	2	9 – 2 = |7|	7 × 2 = 14
10	2	10 – 2 = |8|	8 × 2 = 16
11	0	11 – 2 = |9|	9 × 0 = 00
12	2	12 – 2 = |10|	10 × 2 = 20
			$\Sigma\lvert fd_M\rvert = 196$

Notes

Mean deviation from Median $\delta_M = \frac{\Sigma|fd_M|}{N}$

Here, $\Sigma|fd_M| = 196$, N = 100.

$\therefore \quad \delta_M = \frac{196}{100} = 1.96.$

Mean deviation from mean $\delta_{\bar{X}} = \frac{\Sigma|fd_{\bar{X}}|}{N}$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f}$$

fx = 00, 16, 42, 30, 68, 40, 24, 14, 8, 18, 20, 00, 24.

$\Sigma fx = 304$, $\Sigma f = 100$

$\therefore \quad \bar{X} = \frac{304}{100} = 3.04$

Deviation from Mean

Item	f	deviation $d_{\bar{X}}$	$fd_{\bar{X}}$
0	15	|3.04|	45.6
1	16	|2.04|	32.64
2	21	|1.04|	21.84
3	10	|0.04|	0.4
4	17	|0.96|	16.32
5	8	|1.96|	15.68
6	4	|2.96|	11.84
7	2	|3.96|	7.92
8	1	|4.96|	4.96
9	2	|5.96|	11.96
10	2	|6.96|	13.96
11	0	|7.96|	00
12	2	|8.96|	17.92
			$\Sigma fd_{\bar{X}} = 201.04$

Mean deviation $\delta_{\bar{X}} = \frac{\Sigma|fd_{\bar{X}}|}{N}$

$\Sigma fd_{\bar{X}} = 201.04$, N = 100

$\therefore \quad \delta_{\bar{X}} = \frac{201.04}{100} = 2.01$

Answer: Mean deviation from Median δ_M is equal to 1.96 whereas that from arithmetic mean $\delta_{\bar{X}} = 2.01$.

Notes

Example 4: Calculate mean deviation from the following data:

Class-interval	Frequency	Class-interval	Frequency
10–30	6	90–110	21
30–50	53	110–130	26
50–70	85	130–150	4
70–90	56	150–170	4

Given that the median of the above data is 60.5.

Solution:

Class-interval	f	Mid-points	Deviation d_M	fd_M
10–30	6	20	|40.5|	243.0
30–50	53	40	|20.5|	1086.5
50–70	85	60	|0.5|	42.5
70–90	56	80	|17.5|	1092.0
90–110	21	100	|39.5|	2139.5
110–130	26	120	|59.5|	1547.0
130–150	4	140	|79.5|	318.0
150–170	4	160	|99.5|	398.0
				Σfd_M = 6,866.5

$$\delta_M = \frac{\Sigma|fd_M|}{N}, \quad \Sigma|fd_M| = 6{,}866.5,\ N = 255$$

$$\therefore \quad \delta_M = \frac{6{,}866.5}{255} = 26.93.$$

Example 5: Calculate mean deviation from mean from the following data:

X	0–100	100–200	200–300	300–400	400–500	500–600	600–700
f	6	5	8	15	7	6	3

Solution:

X	f	Mid (M) values	Deviation from $\bar{X}$ $(d_{\bar{X}})$	$fd_{\bar{X}}$
0–100	6	50	|284|	1,704
100–200	5	150	|184|	920
200–300	8	250	|84|	672
300–400	15	350	|16|	240
400–500	7	450	|116|	812
500–600	6	550	|216|	1,296
600–700	3	650	|316|	948
	Σf = 50			$\Sigma fx_{\bar{X}}$ = 6,592

Notes

$$\bar{X} = \frac{\Sigma fM}{\Sigma f} = \frac{16,700}{50} = 334$$

$$\delta_{\bar{X}} = \frac{\Sigma \left| fd_{\bar{X}} \right|}{N}$$

$$\Sigma \left| fd_{\bar{X}} \right| = 6,592, \; N = 50$$

$$\delta_{\bar{X}} = \frac{6,592}{50} = 131.84$$

Merits and Limitations of Mean Deviation

Merits

(i) The outstanding advantage of the average deviation is its relative simplicity. It is simple to understand and easy to compute. Anyone familiar with the concept of the average can readily appreciate the meaning of the average deviation. If a situation requires a measure of dispersion that will be presented to the general public or any group not thoroughly grounded in statistics, the average deviation is very useful.

(ii) It is based on each and every item of the data. Consequently change in the value of any item would change the value of mean deviation.

(iii) Mean deviation is less affected by the values of extreme items than the standard deviation.

(iv) Since deviations are taken from a central value, comparison about, formation of different distributions can easily be made.

Limitations

(i) The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the items. For example if from twenty, fifty is deducted we write 30 and not – 30. This is mathematically wrong and makes the method **non-algebraic.** If the signs of the deviations are not ignored the net sum of the deviations will be zero if the reference point is the mean or approximately zero if the reference point is median.

(ii) This method may not give us very accurate results. The reason is that mean deviation gives us best results when deviations are taken from median. But median is not a satisfactory measure when the degree of variability in a series is very high. And if we compute mean deviation from mean that is also not desirable because the sum of the deviations from mean (ignoring signs) is greater than the sum of the deviations from median (ignoring signs). If mean deviation is computed from mode that is also not scientific because the value of mode cannot always be determined.

(iii) It is not capable of further algebraic treatment.

(iv) It is rarely used in sociological studies.

Because of these limitations its use is limited and it is overshadowed as a measure of variation by the superior standard deviation.

Usefulness of the Mean Deviation: The serious drawbacks of the average deviation should not blind us to its practical utility. Because of its simplicity in meaning and computation, it is especially effective in reports presented to the general public or to groups not familiar with statistical methods. This measure is useful for small samples with no elaborate analysis required. Incidentally it may be mentioned that the National Bureau of Economic Research has found in its work on forecasting business cycles, that the average deviation is the most practical measure of dispersion to use for this purpose.

Notes

7.2 The Standard Deviation

The standard deviation concept was introduced by Karl Pearson in 1893. It is by far the most important and widely used measure of studying dispersion. Its significance lies in the fact that it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion. Standard deviation is also known as *root-mean square deviation* for the reason that it is the square root of the means of the squared deviations from the arithmetic mean. Standard deviation is denoted by the small Greek letter σ (read as sigma).

The standard deviation measures the absolute dispersion or variability of a distribution; the greater the amount of dispersion or variability, the greater the standard deviation, the greater will be the magnitude of the deviations of the values from their mean. A small standard deviation means a high degree of uniformity of the observations as well as homogeneity of a series; a large standard deviation means just the opposite. Thus if we have two or more comparable series with identical or nearly identical means, it is the distribution with the smallest standard deviation that has the most representative mean. Hence standard deviation is extremely useful in judging the representativeness of the mean.

Difference between Average Deviation and Standard Deviation

Both these measures of dispersion are based on each and every item of the distribution. But they differ in the following respects:

(i) Algebraic signs are ignored while calculating mean deviation whereas in the calculation of standard deviations, signs are taken into account.

(ii) Mean deviation can be computed either from median or mean. The standard deviation, on the other hand, is always computed from the arithmetic mean because the sum of the squares of the deviations of items from arithmetic mean is the least.

Calculation of Standard Deviation—Individual Observations

In case of individual observations, standard deviation may be computed by applying any of the following two methods:

1. By taking deviations of the items from the actual mean.
2. By taking deviations of the items from an assumed mean.

1. Deviations taken from Actual Mean: When deviations are taken from actual mean the following formula is applied:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

where $x = \left(X - \bar{X}\right)$ and N = number of observations.

Steps : (i) Calculate the actual mean of the series, *i.e.,* $\bar{X}$.

(ii) Take the deviations of the items from the mean, *i.e.,* find $\left(X - \bar{X}\right)$. Denote these deviation by x.

(iii) Square these deviations and obtain the total Σx^2.

(iv) Divide Σx^2 by the total number of observations, *i.e.,* N, and extract the square-root. This gives us the value of standard deviation.

Example 6: Calculate standard deviation from the following observations of marks of 5 students of a tutorial group:

Marks out of 25 **Notes**

8 12 13 15 22

Solution:

CALCULATION OF STANDARD DEVIATION

X	$(X - \bar{X})$	x^2
8	– 6	36
12	– 2	4
13	– 1	1
15	+ 1	1
22	+ 8	64
$\Sigma X = 70$	$\Sigma x = 0$	$\Sigma x^2 = 106$

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

$$\bar{X} = \frac{\Sigma X}{N} = \frac{70}{5} = 14$$

$$\Sigma x^2 = 106, N = 5$$

$$\sigma = \sqrt{\frac{106}{5}} = \sqrt{21.2} = 4.604.$$

2. Deviations taken from Assumed Mean: When the actual mean is in fractions, say, in the above case 123.674, it would be too cumbersome to take deviations from it and then obtaining squares of these deviations. In such a case, either the mean may be approximated or else the deviations be taken from an assumed mean and the necessary adjustment be made in the value of standard deviation. The former method of approximation is less accurate and, therefore, invariably in such a case deviations are taken from assumed mean.

When deviations are taken from assumed mean the following formula is applied:

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

Steps : (i) Take the deviations of the items from an assumed mean *i.e.*, obtain (X –A). Denote these deviations by d. Take the total of these deviations, *i.e.*, obtain Σd.

(ii) Square these deviations and obtain the total Σd^2.

(iii) Substitute the value of Σd^2, Σd and N in the formula.

Example 7: Following figures give the income of 10 persons in rupees. Find the standard deviation.

227, 235, 255, 269, 292, 299, 312, 321, 333, 348

Notes

Solution:

CALCULATION OF STANDARD DEVIATION

X	(X–280)d	d^2
227	– 53	2809
235	– 45	2025
255	– 25	625
269	– 11	121
292	+ 12	144
299	+ 19	361
312	+ 32	1024
321	+ 41	1681
333	+ 53	2809
348	+ 68	4624
N = 10	$\sum d$ = 91	$\sum d^2$ = 16223

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

$\sum d^2$ = 16223, N = 10, $\sum d$ = 91

$$\sigma = \sqrt{\frac{16223}{10} - \left(\frac{91}{10}\right)^2} = \sqrt{1622.3 - 82.81} = 39.24.$$

Calculation of Standard Deviation—Discrete Series

For calculating standard deviation in discrete series any of the following methods may be applied:

1. Actual mean method.
2. Assumed mean method.
3. Step deviation method.

1. Actual Mean Method: When this method is applied deviations are taken from the actual mean, *i.e.*, we find $\left(X - \bar{X}\right)$ and denote these deviations by x. These deviations are then squared and multiplied by the respective frequencies. The following formula is applied:

$$\sigma = \sqrt{\frac{\sum fx^2}{N}}$$

where $x = \left(X - \bar{X}\right)$.

However, in practice this method is rarely used because if the actual mean is in fractions the calculations take a lot of time.

2. Assumed Mean Method: When this method is used, the following formula is applied.

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

Notes

where $d = (X - A)$.

Steps : (i) Take the deviations of the items from an assumed mean and denote these deviations by d.

(ii) Multiply the deviations by the respective frequencies and obtain the total, Σfd.

(iii) Obtain the squares of the deviations, *i.e.*, calculate d^2.

(iv) Multiply the squared deviations by respective frequencies and obtain the total, Σfd^2.

Substitute the values in the above formula.

Example 8: Calculate the standard deviation from the data given below:

Size of item	Frequency	Size of item	Frequency
3.5	3	7.5	85
4.5	7	8.5	32
5.5	22	9.5	8
6.5	60		

Solution:

CALCULATION OF STANDARD DEVIATION

X Size of item	f	$(X-6.5)d$	fd	fd^2
3.5	3	– 3	– 9	27
4.5	7	– 2	– 14	28
5.5	22	– 1	– 22	22
6.5	60	0	0	0
7.5	85	+ 1	+ 85	85
8.5	32	+ 2	+ 64	128
9.5	8	+ 3	+ 24	72
	N = 217		Σfd = 128	Σfd^2 = 362

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

$\Sigma fd^2 = 362$, $\Sigma fd = 128$, N = 217

$$\sigma = \sqrt{\frac{362}{217} - \left(\frac{128}{217}\right)^2} = \sqrt{1.67 - (.59)^2}$$

$$= \sqrt{1.67 - 0.35} = 1.149.$$

3. Step Deviation Method: When this method is used we take a common factor from the given data. The formula for computing standard deviation is:

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

Notes

where $d = \frac{(X - A)}{i}$ and i = class interval

The use of the above formula simplifies calculations.

Example 9: Find the standard deviation for the following distribution:

X	4.5	14.5	24.5	34.5	44.5	54.5	64.5
f	1	5	12	22	17	9	4

Solution:

Calculation of Standard Deviation

X	f	(X–34.5)/10 d	fd	fd^2
4.5	1	– 3	– 3	9
14.5	5	– 2	– 10	20
24.5	12	– 1	– 12	12
34.5	22	0	0	0
44.5	17	+ 1	+ 17	17
54.5	9	+ 2	+ 18	36
64.5	4	+ 3	+ 12	36
	N = 70		Σfd = 22	Σfd^2 = 130

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

Here Σfd^2 = 130, Σfd = 22, N = 70, i = 10.

$$\sigma = \sqrt{\frac{130}{70} - \left(\frac{22}{70}\right)^2} \times 10 = \sqrt{1.857 - .1} \times 10 = 1.326 \times 10$$

= 13.26.

Calculation of Standard Deviation—Continuous Series

In continuous series any of the methods discussed above for discrete frequency distribution can be used. However, in practice it is the step deviation method that is mostly used. The formula is:

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$$d = \frac{(m - A)}{i}; \; i = \text{Class interval.}$$

Steps : (i) Find the mid-points of various classes.

(ii) Take the deviations of these mid-points from an assumed mean and divide by the class interval. Denote these deviations by d.

Notes

(iii) Multiply the frequency of each class with these deviations and obtain Σfd.

(iv) Square the deviations and multiply them with the respective frequencies of each class and obtain Σfd^2.

Thus, the only difference in procedure in case of continuous series is to find mid-points of the various classes.

Example 10: The following table gives the distribution of income of 100 families in a village. Calculate standard deviation:

Income (Rs.)	No. of families
0–1000	18
1000–2000	26
2000–3000	30
3000–4000	12
4000–5000	10
5000–6000	4

(B.Com., Kerala Univ., 1996)

Solution:

Calcultion of Standard of Diviation

Income (Rs.)	*m.p.* *m*	*f*	(*m*–2500)/1000 *d*	*fd*	fd^2
0–1000	500	18	– 2	– 36	72
1000–2000	1500	26	– 1	– 26	26
2000–3000	2500	30	0	0	0
3000–4000	3500	12	+ 1	+ 12	12
4000–5000	4500	10	+ 2	+ 20	40
5000–6000	5500	4	+ 3	+ 12	36
		N = 100		Σfd = – 18	Σfd^2 = 186

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$$= \sqrt{\frac{186}{100} - \left(\frac{-18}{100}\right)^2} \times 1000 = \sqrt{1.86 - .0324} \times 1000$$

$$= 1.3519 \times 1000 = 1351.9.$$

Mathematical Properties of Standard Deviation

Standard deviation has some very important mathematical properties which considerably enhance its utility in statistical work.

1. **Combined standard deviation:** Just as it is possible to compute combined mean of two or more than two groups, similarly we can also compute combined standard deviation of two or more groups. Combined standard deviation is denoted by σ_{12} and is computed as follows:

Notes

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2\sigma_2^2}{N_1 + N_2}}$$

where σ_{12} = combined standard deviation; σ_1 = standard deviation of first group; σ_2 = standard deviation of second group; $d_1 = \left(\overline{X}_1 - \overline{X}_{12}\right)$; $d_2 = \left(\overline{X}_2 - \overline{X}_{12}\right)$.

The above formula can be extended to find out the standard deviation of three or more groups. For example, combined standard deviation of three groups would be:

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1d_1^2 + N_2d_2^2 + N_3d_3^2}{N_1 + N_2 + N_3}}$$

$$d_1 = \left|\overline{X}_1 - \overline{X}_{123}\right|,\ d_2 = \left|\overline{X}_2 - \overline{X}_{123}\right|,\ d_3 = \left|\overline{X}_3 - \overline{X}_{123}\right|.$$

Example 11: The number examined, the mean weight and the standard deviation in each group of examination by two medical examiners is given below. Find the mean weight and standard deviation of both the groups taken together.

A	50	113	6.5
B	60	120	8.2

Solution:

$$\overline{X}_{12} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

$$N_1 = 50,\ N_2 = 60,\ \overline{X}_1 = 113,\ \overline{X}_2 = 120$$

$$\overline{X}_{12} = \frac{(50 \times 113) + (60 \times 120)}{50 + 60} = \frac{5650 + 7200}{110} = \frac{12850}{110} = 116.82$$

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$$N_1 = 50,\ \sigma_1 = 6.5,\ N_3 = 60,\ \sigma_2 = 8.2$$

$$d_1 = \left|\overline{X}_1 - \overline{X}_{12}\right| = (113 - 116.82) = -3.82$$

$$d_2 = \left|\overline{X}_2 - \overline{X}_{12}\right| = (120 - 116.82) = 3.18.$$

Substituting the values

$$\sigma_{12} = \sqrt{\frac{50(6.5)^2 + 60(8.2)^2 + 50(-3.82)^2 + 60(3.18)^2}{50 + 60}}$$

$$= \sqrt{\frac{2112.5 + 4034.4 + 729.62 + 606.744}{110}} = \sqrt{\frac{7483.264}{100}}$$

$$= \sqrt{68.03} = 8.25.$$

Example 12: The number of workers employed, the mean wage (in Rs.) per month and the standard deviation (in Rs.) in each section of a factory are given below. Calculate the mean

wage and standard deviation of all the workers taken together.

Notes

Section	No. of workers employed	Mean wage in Rs.	Standard deviation in Rs.
A	50	113	6
B	60	120	7
C	90	115	8

Solution:

$$\bar{X}_{123} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3}{N_1 + N_2 + N_3}$$

$$= \frac{(50\times 113) + (60\times 120) + (90\times 115)}{60 + 60 + 90}$$

$$= \frac{5650 + 7200 + 10350}{200} = \frac{23200}{200} \text{ Rs. } 116.$$

Combined standard deviation of three series.

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1d_1^2 + N_2d_2^2 + N_3d_3^2}{N_1 + N_2 + N_3}}$$

$$d_1 = \left|\bar{X}_1 - \bar{X}_{123}\right| = \left|113 - 116\right| = 3$$

$$d_2 = \left|\bar{X}_2 - \bar{X}_{123}\right| = \left|120 - 116\right| = 4$$

$$d_3 = \left|\bar{X}_3 - \bar{X}_{123}\right| = \left|115 - 116\right| = 1$$

$$\sigma_{123} = \sqrt{\frac{50(6)^2 + 60(7)^2 + 90(8)^2 + 50(3)^2 + 60(4)^2 + 90(-1)^2}{50 + 60 + 90}}$$

$$= \sqrt{\frac{1800 + 2940 + 5760 + 450 + 960 + 90}{200}} = \sqrt{\frac{12000}{200}} = \sqrt{60} = 7.75.$$

2. **Standard deviation of *n* natural numbers:** The standard deviation of the first *n* natural numbers can be obtained by the following formula:

$$\sigma = \sqrt{\frac{1}{12}\left(N^2 - 1\right)}$$

Thus, the standard deviation of natural numbers 1 to 10 will be

$$\sigma = \sqrt{\frac{1}{12}\left(10^2 - 1\right)} = \sqrt{\frac{1}{12}\times 99} = \sqrt{8.25} = 2.872.$$

Note: The answer would be the same when direct method of calculating standard deviation is used. But this holds good only for natural numbers from 1 to *n* in continuation without gaps.

3. The sum of the squares of the deviations of items in the series from their arithmetic mean is minimum. In other words, the sum of the squares of the deviations of items of any series from a value other than the arithmetic mean would always be greater. This is the reason why standard deviation is always computed from the arithmetic mean.

Notes

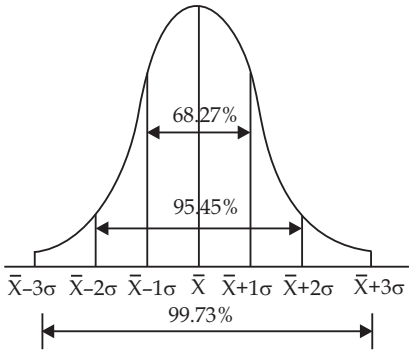
4. For a symmetrical distribution, the following area relationship holds good:

Mean $\pm 1\sigma$ covers 68.27% items.

Mean $\pm 2\sigma$ covers 95.45% items.

Mean $\pm 3\sigma$ covers 99.73% items.

This can be illustrated by the following diagram:



Relation between Measures of Dispersion

In a normal distribution there is a fixed relationship between the three most commonly used measures of dispersion. The quartile deviation is smallest, the mean deviation next and the standard deviation is largest, in the following proportion:

$$\text{Q.D.} = \frac{2}{3}\sigma \text{ ; M.D.} = \frac{4}{5}\sigma$$

These relationships can be easily memorised because of the sequence 2, 3, 4, 5. The same proportions tend to hold true for many distributions that are quite normal. They are useful in estimating one measure of dispersion when another is known, or in checking roughly the accuracy of a calculated value. If the computed σ differs very widely from its value estimated from Q.D. or M.D. either an error has been made or the distribution differs considerably from normal.

Another comparison may be made of the proportion of items that are typically included within the range of one Q.D., M.D. or S.D. measured both above and below the mean. In a normal distribution:

$\bar{X} \pm \text{Q.D.}$ includes 50 per cent of the items.

$\bar{X} \pm \text{M.D.}$ includes 57.51 per cent of the items.

$\bar{X} \pm \sigma$ includes 68.27 per cent of the items.

Coefficient of Variation

The Standard deviation discussed above is an absolute measure of dispersion. The corresponding relative measure is known as the *coefficient of variation.* This measure developed by Karl Pearson is the most commonly used measure of relative variation. It is used in such problems where we want to compare the variability of two or more than two series. That series (or group) for which the coefficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable or less homogeneous. On the other hand, the series for which coefficient of variation is less is

Notes

said to be less variable or more consistent, more uniform more stable or more homogeneous. Coefficient of variation is denoted by the symbol C.V. and is obtained as follows:

Coefficient of variation or C.V. = $\frac{\sigma}{\bar{X}} \times 100$.

It may be pointed out that although any measure of dispersion can be used in conjunction with any average in computing relative dispersion, statisticians, in fact, almost always use the standard deviation as the measure of dispersion and the arithmetic mean as the average. When the relative dispersion is stated in terms of the arithmetic mean and the standard deviation, the resulting percentage is known as the coefficient of variation of coefficient of variability.

A distinction is sometimes made between coefficient of variation and coefficient of standard deviation.

The former is always a percentage, the latter is just the ratio of standard deviation to mean, *i.e.*, $\left(\frac{\sigma}{\bar{X}}\right)$.

Example 13: The scores of two batsmen A and B in ten innings during a certain season are:

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Find (using coefficient of variation) which of the batsmen A, B is more consistent in scoring. **(B.Com., Calcutta Univ., 1996)**

Solution:

Calculation of Coefficient of Variation

X	**(X – 46)** x	x^2	**Y**	$(Y - \bar{Y})$ y	y^2
32	– 14	196	19	– 31	961
28	– 18	324	31	– 19	361
47	+ 1	1	48	– 2	4
63	+ 17	289	53	+ 3	9
71	+ 25	625	67	+ 17	289
39	– 7	49	90	+ 40	1600
10	– 36	1296	10	– 40	1600
60	+ 14	196	62	+ 12	144
96	+ 50	2500	40	– 10	100
14	– 32	1024	80	+ 30	900
$\Sigma X = 460$	$\Sigma x = 0$	$\Sigma x^2 = 6500$	$\Sigma Y = 500$	$\Sigma y = 0$	$\Sigma y^2 = 5968$

Batsman A

$$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

$$\bar{X} = \frac{460}{10} = 46$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{6500}{10}} = 25.5$$

Batsman B

$$C.V. = \frac{\sigma}{\bar{Y}} \times 100$$

$$\bar{Y} = \frac{500}{10} = 50$$

$$\sigma = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{5968}{10}} = 24.43$$

Notes

$$\text{C.V.} = \frac{25.5}{46} \times 100 = 55.43 \qquad \text{C.V.} = \frac{24.43}{50} \times 100 = 48.86$$

Since coefficient of variation is less for batsman B hencebatsman B is more consistent.

Example 14: A panel of two judges P and O graded seven dramatic performances by independently awarding marks as follows:

Performance:	1	2	3	4	5	6	7
Marks by P:	46	42	44	40	43	41	45
Marks by O:	40	38	36	35	39	37	41

Find out coefficient of variation in the marks awarded by two judges and interpret the result.

Solution:

Coefficient of Variation of the Marks Obtained by P and O

Marks by P **X**	$(\mathbf{X} - \bar{\mathbf{X}})$ x	x^2	**Marks by O** **Y**	$(\mathbf{Y} - \bar{Y})$ y	y^2
46	+ 3	9	40	+ 2	4
42	–1	1	38	0	0
44	+ 1	1	36	– 2	4
40	– 3	9	35	– 3	9
43	0	0	39	+ 1	1
41	– 2	4	37	– 1	1
45	+ 2	4	41	+ 3	9
$\Sigma X = 301$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma Y = 266$	$\Sigma y = 0$	$\Sigma y^2 = 28$

Marks by P

$$\bar{X} = \frac{\Sigma X}{N} = \frac{301}{7} = 43$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{28}{7}} = 2$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{2}{43} \times 100 = 4.65$$

Marks by O

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{266}{7} = 38$$

$$\sigma = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{28}{7}} = 2$$

$$\text{C.V.} = \frac{\sigma}{\bar{Y}} \times 100 = \frac{2}{38} \times 100 = 5.26.$$

Notes

The average marks obtained by P are higher. Hence his performance is better. The coefficient of variation is lower in case of P hence he is a more consistent student.

Example 15: Suppose that samples of polythene bags two manufactures, A and B are tested by prospective buyer for bursting pressure, with the following results:

Bursting Pressure (*lb.*)	Number of bags	
	A	B
5.0–9.9	2	9
10.0–14.9	9	11
15.0–19.9	29	18
20.9–24.9	54	32
25.0–29.9	11	27
30.0–34.9	5	13
	110	110

Which set of the bags has the highest average bursting pressure ? Which has more uniform pressure ? If prices are the same, which manufacture's bags would be preferred by the buyer ? Why ?

Solution: For determining the set of bags having average bursting pressure, calculate arithmetic mean and for finding out set of bags having more uniform pressure compute coefficient of variation.

Manufacturer A

Calculation of Mean and Standard Deviation

Bursting pressure (*lb.*)	m	f	$\left(\frac{m-17.45}{5}\right)$ d	fd	fd^2
4.95–9.95	7.45	2	– 2	– 4	8
9.95–14.95	12.45	9	– 1	– 9	9
14.95–19.95	**17.45**	29	0	0	0
19.95–24.95	22.45	54	+ 1	+ 54	54
24.95–29.95	27.45	11	+ 2	+ 22	44
29.95–34.95	32.45	5	+ 3	+ 15	45
	N = 110			Σfd = 78	Σfd^2 = 160

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i$$

Here $A = 17.45$, $\Sigma fd = 78$, $N = 110$, $i = 5$

$$\bar{X} = 17.45 + \frac{78}{110} \times 5 = 17.45 + 3.55 = 21.$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{160}{110} - \left(\frac{78}{110}\right)^2} \times 5$$

Notes

$$= \sqrt{1.455 - 0.503} \times 5 = \sqrt{0.952} \times 5 = 0.976 \times 5 = 4.88$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{4.88}{21} \times 100 = 23.24\%.$$

Manufacturer B

Calculation of Mean and Standard Deviation

Bursting pressure (*lb.*)	m	f	$\left(\frac{m - 17.45}{5}\right)$ d	fd	fd^2
4.95–9.95	7.45	9	– 2	–18	36
9.95–14.95	12.45	11	– 1	– 11	11
14.95–19.95	**17.45**	18	0	0	0
19.95–24.95	22.45	54	+ 1	+ 54	54
24.95-29.95	27.45	27	+ 2	+ 54	108
29.95–34.95	32.45	13	+ 3	+ 39	117
		N = 110		Σfd = 96	Σfd^2 = 304

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 17.45 + \frac{96}{110} \times 5 = 17.45 + 4.36 = 21.18$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{304}{110} - \left(\frac{96}{110}\right)^2} \times 5$$

$$= \sqrt{2.764 - 0.762} \times 5 = 1.4149 \times 5 = 7.075$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{7.075}{21.81} \times 100 = 32.44\%$$

Since the average bursting pressure is higher for manufacturer B, the bags of manufacturer B have a higher bursting pressure. The bags of manufacturer A have more uniform pressure since the coefficient of variation is less for manufacturer A. If princes are the same, the bags of manufacturer A should be preferred by the buyer because they have more uniform pressure.

Variance

The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1918. The concept of variance is highly important in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variation in the original series. Variance is defined as follows:

$$\text{Variance} = \frac{\Sigma\left(X - \bar{X}\right)^2}{N}.$$

Thus, variance is nothing but the square of the standard deviation, *i.e.,*

$$\text{Variance} = \sigma^2$$

or $$\sigma = \sqrt{\text{Variance.}}$$

Notes

In a frequency distribution where deviations are taken from assumed mean, variance may directly be computed as follows:

$$\text{Variance} = \left\{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2\right\} \times i$$

where $= \frac{(X - A)}{i}$ and i = class interval.

Example 16: The weights of a number of packages are given as follows:

16.1, 15.9, 15.8, 16.3, 16.2, 16.0, 15.9, 16.0, 16.1, 16.0, 15.9, 16.1, 16.0, 16.0.

From a frequency table. Find the standard deviation and the variance.

Solution:

Weight X	Tally Bar	Frequency f	$(X - A)$ d	fd	fd^2
15.8	|	1	– .3	– .3	.09
15.9	|||	3	– .2	– .6	.12
16.0	||𝍸	5	– .1	– .5	.05
16.1	|||	3	0	0	0
16.2	|	1	+ .1	+ .1	01
16.3	|	1	+ .2	+ .2	04
		N = 14		Σfd = – 1.1	Σfd^2 = 0.31

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

$\Sigma fd^2 = 0.31$, $\Sigma fd = -1.1$, $N = 14$

$$= \sqrt{\frac{.31}{14} - \frac{(-1.1)^2}{14}} = \sqrt{0.022 - .0062} = 0.126$$

Variance $= \sigma^2 = (.126)^2 = 0.0159$.

Merits and Limitations of Standard Deviation

Merits

(i) The standard deviation is the best measure of variation because of its mathematical characteristics. It is based on every item of the distribution. Also it is amenable to algebraic treatment and is less affected by fluctuations of samlping than most other measures of dispersion.

(ii) It is possible to calculate the combined standard deviation of two or more groups. This is not possible with any other measure.

(iii) For comparing the variability of two or more distributions coefficient of variation is considered to be most appropriate and this is based on mean and standard deviation.

(iv) Standard deviation is most prominently used in further statistical work. For example, in computing skewness, correlation, etc., use is made of standard deviation. It is a key note in sampling and provides a unit of measurement for the normal distribution.

Notes

Limitations

(i) As compared to other measures it is difficult to compute. However, it does not reduce the importance of this measure because of high degree of accuracy of results it gives.

(ii) It gives more weight to extreme items and less to those which are near the mean. It is because of the fact that the squares of the deviations which are big in size would be proportionately greater than the squares of those deviations which are comparatively small. The deviations 2 and 8 are in the ratio of 1: 4 but their sqaures, *i.e.*, 4 and 64, would be in the ratio of 1: 16.

Correcting Incorrect Values of Standard Deviation

Mistakes in calculations are always possible. Sometimes it so happens that while calculating mean and standard deviation we unconsciously copy out wrong items. For example, an item 21 may be copied as 12. Similarly one item 127 may be taken as only 27. In such cases if the entire calculations are done again, it would become too tedious a task. By adopting a very simple procedure we can correct the incorrect values of mean and standard deviation. For obtaining correct mean we find out correct ΣX by deducting from the original ΣX the wrong items and adding to it the correct items.

Similarly for calculating correct standard deviation we obtain the value of correct ΣX^2. The following illustration shall clarify the calculations.

Example 17: A student obtained the mean and standard deviation of 100 observations as 40 and 5.1 respectively. It was later found that one observation was wrongly copied as 50, the correct figure being 40. Find the correct mean and standard deviation.

Solution:

Correct Mean:

We are given $\overline{X} = 40$, $\sigma = 5.1$, N = 100

$$\overline{X} = \frac{\Sigma X}{N}$$

$$40 = \frac{\Sigma X}{100} \text{ or } \Sigma X = 4000$$

But correct ΣX = ΣX – Wrong items + Correct items = 4000 – 50 + 40 = 3990.

$$\text{Correct } \overline{X} = \frac{\text{Correct } \Sigma X}{N} = \frac{3990}{100} = 39.9.$$

Correct Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\overline{X}\right)^2}$$

$$5.1 = \sqrt{\frac{\Sigma X^2}{100} - (40)^2}$$

Squaring, we get

$$26.01 = \frac{\Sigma X^2}{100} - 1600$$

$$2601 = \Sigma X^2 - 1{,}60{,}000 \text{ or } \Sigma X^2 = 2601 + 1, 60, 000 = 1, 62, 601.$$

Correct ΣX^2 = Incorrect ΣX^2 – wrong item square + Correct item square

Notes

$$\text{Correct } \Sigma X^2 = 162601 - (50)^2 + (40)^2 = 162601 - 2500 + 1600 = 161701$$

$$\text{Correct } \sigma = \sqrt{\frac{\text{Correct } \Sigma X^2}{N} - \left(\text{Correct } \bar{X}\right)^2}$$

$$= \sqrt{\frac{161701}{100} - (39.9)^2} = \sqrt{1617.01 - 1592.01} = \sqrt{25} = 5.$$

Self-Assessment

1. Indicate whether the following statements are True or False:

(*i*) Mean deviation can be calculated from arithmetic mean, median or mode.

(*ii*) Mean deviation ignores the sings of deviations.

(*iii*) Standard deviation is an absolute measure of dispersion.

(*iv*) Standard deviations of more than two component parts cannot be combined in one.

(*v*) Mean deviation is least when deviations are taken from median.

7.3 Summary

- The average deviation is sometimes called the mean deviation. It is the average difference between the items in a distribution and the median or mean of that series. Theoretically, there is an advantage in taking the deviations from median because *the sum of the deviations of items from median is minimum when signs are ignored*. However, in practice the arithmetic mean is more frequently used in calculating the value of average deviation and this is the reason why it is more commonly called mean deviation. In any case, the average used must be clearly stated in a given problem so that any possible confusion in meaning is avoided.
- In statistics **standard deviation** (represented by the symbol sigma, σ) shows how much variation or "dispersion" exists from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.
- The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler though practically less robust than the average absolute deviation. A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data.
- The average deviation or mean deviation is a measure of dispersion that is based upon all the items in a distribution. It is the arithmetic mean of the deviations of the data from its central value, may it be arithmetic mean, median or mode. While, considering the deviations from its central value, only absolute values are taken into consideration, *(i.e.,* without considering the positive or negative signs). Mean deviation is denoted by δ (delta).
- The outstanding advantage of the average deviation is its relative simplicity. It is simple to understand and easy to compute. Anyone familiar with the concept of the average can readily appreciate the meaning of the average deviation. If a situation requires a measure of dispersion that will be presented to the general public or any group not thoroughly grounded in statistics, the average deviation is very useful.
- The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the items. For example if from twenty, fifty is deducted we write 30 and not – 30. This is mathematically wrong and makes the method **non-algebraic.** If the signs of the deviations are not ignored the net sum of the deviations will be zero if the reference point is the mean or approximately zero if the reference point is median.
- The serious drawbacks of the average deviation should not blind us to its practical utility. Because of its simplicity in meaning and computation, it is especially effective in reports

Notes

presented to the general public or to groups not familiar with statistical methods. This measure is useful for small samples with no elaborate analysis required. Incidentally it may be mentioned that the National Bureau of Economic Research has found in its work on forecasting business cycles, that the average deviation is the most practical measure of dispersion to use for this purpose.

- The standard deviation concept was introduced by Karl Pearson in 1893. It is by far the most important and widely used measure of studying dispersion. Its significance lies in the fact that it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion. Standard deviation is also known as *root-mean square deviation* for the reason that it is the square root of the means of the squared deviations from the arithmetic mean. Standard deviation is denoted by the small Greek letter σ (read as sigma).
- The standard deviation measures the absolute dispersion or variability of a distribution; the greater the amount of dispersion or variability, the greater the standard deviation, the greater will be the magnitude of the deviations of the values from their mean. A small standard deviation means a high degree of uniformity of the observations as well as homogeneity of a series; a large standard deviation means just the opposite. Thus if we have two or more comparable series with identical or nearly identical means, it is the distribution with the smallest standard deviation that has the most representative mean. Hence standard deviation is extremely useful in judging the representativeness of the mean.
- Mean deviation can be computed either from median or mean. The standard deviation, on the other hand, is always computed from the arithmetic mean because the sum of the squares of the deviations of items from arithmetic mean is the least.
- When the actual mean is in fractions, say, in the above case 123.674, it would be too cumbersome to take deviations from it and then obtaining squares of these deviations. In such a case, either the mean may be approximated or else the deviations be taken from an assumed mean and the necessary adjustment be made in the value of standard deviation.
- The sum of the squares of the deviations of items in the series from their arithmetic mean is minimum. In other words, the sum of the squares of the deviations of items of any series from a value other than the arithmetic mean would always be greater. This is the reason why standard deviation is always computed from the arithmetic mean.
- In a normal distribution there is a fixed relationship between the three most commonly used measures of dispersion. The quartile deviation is smallest, the mean deviation next and the standard deviation is largest, in the following proportion:

$$\text{Q.D.} = \frac{2}{3}\sigma \text{ ; M.D.} = \frac{4}{5}\sigma$$

- These relationships can be easily memorised because of the sequence 2, 3, 4, 5. The same proportions tend to hold true for many distributions that are quite normal. They are useful in estimating one measure of dispersion when another is known, or in checking roughly the accuracy of a calculated value. If the computed σ differs very widely from its value estimated from Q.D. or M.D. either an error has been made or the distribution differs considerably from normal.
- The Standard deviation discussed above is an absolute measure of dispersion. The corresponding relative measure is known as the *coefficient of variation.* This measure developed by Karl Pearson is the most commonly used measure of relative variation. It is used in such problems where we want to compare the variability of two or more than two series. That series (or group) for which the coefficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable or less homogeneous.
- The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1918. The concept of variance is highly important in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variation in the original series.

Notes

- The standard deviation is the best measure of variation because of its mathematical characteristics. It is based on every item of the distribution. Also it is amenable to algebraic treatment and is less affected by fluctuations of samlping than most other measures of dispersion.
- Mistakes in calculations are always possible. Sometimes it so happens that while calculating mean and standard deviation we unconsciously copy out wrong items. For example, an item 21 may be copied as 12. Similarly one item 127 may be taken as only 27. In such cases if the entire calculations are done again, it would become too tedious a task. By adopting a very simple procedure we can correct the incorrect values of mean and standard deviation. For obtaining correct mean we find out correct ΣX by deducting from the original ΣX the wrong items and adding to it the correct items.

7.4 Key-Words

1. Mean Deviation : The mean deviation or the average deviation is defined as the mean of the absolute deviations of observations from some suitable average which may be the arithmetic mean, the median or the mode. The difference ( ) is called deviation and when we ignore the negative sign, this deviation is written as and is read as mod deviations.
2. Standard Deviation : In statistics and probability theory, standard deviation (represented by the symbol sigma, ?) shows how much variation or "dispersion" exists from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values

7.5 Review Questions

1. Discuss 'mean deviation' method of measuring dispersion giving its merits and demerits.
2. Explain the concept of Standard deviation. What are its merits and demerits ?
3. Discuss the mathematical Properties of Standard deviation.
4. Distinguish between mean deviation and Standard deviation.

Answers: Self-Assessment

1. (*i*) T (*ii*) T (*iii*) T (*iv*) F (*v*) T

7.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 8: Skewness and Kurtosis: Karl Pearson, Bowley, Kelly's Methods

CONTENTS

Objectives

After reading this unit students will be able to:

- Describe the Meaning, Definition and Types of Skewness.
- Know the Measures of Skewness.
- Explain Karl Pearson, Bowley and Kelly's Methods.
- Understand Kurtosis.

Introduction

Measuring of central tendencies reveal the concentration of frequencies towards the central value of the series and methods of dispersion reveal the dispersal of values in relation to the central value. But the nature of dispersal of values on either sides of an average is not known by measuring dispersion. Similarly, Kurtosis is yet another measure which tells us about the form of a distribution. Thus, it can be said that the central tendencies and dispersion measures should be supplemented by measures of skewness and kurtosis so that a more elaborate picture about the distribution given can be obtained. The study becomes more important in subjects of economics, sociology and other social sciences where normal distribution in a series usually does not occur. However, studies hold importance in biological sciences and other physical sciences as well.

8.1 Meaning, Definition and Types of Skewness

Skewness—Meaning and Definition

The word 'skewness' is the opposite of symmetry and its presence tells us that a particular distribution is not symmetrical or in other words it is skewed. The word 'skewness' can be understood by the following definitions given by eminent statisticians, economists and mathematicians.

(1) As per *Croxten* and *Cowden*, "When a series is not symmetrical it is skewed."

(2) In the words of *Simpson* and *Kafka*, "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the arithmetic mean, median and mode are identicle. The more the mean moves away from mode, the larger the asymmetry or skewness."

Notes

(3) According to *Garrett*, "A distribution is said to be skewed when the mean and median fall at different points in the distribution and the balance or centre of gravity is shifted to one side or the other.

(4) *Riggleman* and *Frisbee* have defined skewness as, "Skewness is the lack of symmetry. When a frequency distribution is plotted on a chart, skewness present in items tends to the disperse chart more on one side of the mean than on other."

Did u know? The measures of 'Skewness' tell about the pattern of dispersal of items from an average, whether it is symmetrical or not. The nature of distribution is further studied deeply by calculating 'Moments' which reveals whether the symmetrical curve is normal, more flat than normal or more peaked than normal.

From the above discussion it is clear that skewness is the lack of symmetry. Measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with symmetrical or normal distribution. In a symmetrical distribution, frequencies go on increasing upto a point and then begin to decrease in the same fashion. There are various possible patterns of symmetrical distribution and normal distribution which is bell-shaped is one of these. Some of the possible patterns of the symmetrical distribution are:

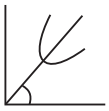


Figure 1

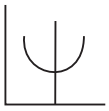


Figure 2

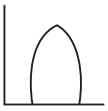


Figure 3

Figure 4

Symmetrical but not bell shaped

Figure 5

Normal distribution

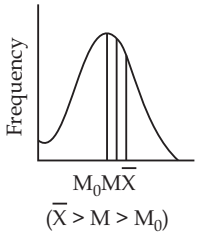
Figure 5: (Symmetrical bell-shaped distribution)

In a symmetrical distribution, mean = median = mode and they lie at the centre of the distribution. When symmetry is disturbed, these values are pulled apart.

Types of Skewness

The skewness may be broadly of two types:

(a) Positive skewness: A distribution in which more than half of the area under the curve is to the right side of the mode, it is said to be a positively skewed distribution. In this type of skewness the right tail is longer than the left tail. In this case, mean is greater than median and the median is greater than the mode and $Q_3 - M > M - Q_1$. Diagrammatically,



$(\bar{X} > M > M_0)$

Figure: 6

Notes

(b) Negative skewness: A distribution in which more than half of the area under the distribution curve is to the left side of the mode, it is said to be a negatively skewed distribution. In this case, the elongated tail is to the left and mean is less than the median which is less than mode and $Q_3 - M < M - Q_1$.

Diagrammatically, the negative skewness can be explained as below:

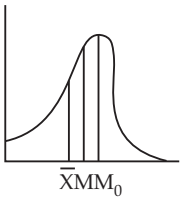


Figure: 7

8.2 Karl Pearson's, Bowley and Kelly's Methods

The following are the main methods of measuring *Skewness* of data:

(1) Karl Pearson's Method

(2) Bowley's Method

(3) Kelly's Method

(1) Karl Pearson's Method

The method of skewness given by Karl Pearson is also called as First Measure of Skewness. This method is based on the difference between the 'mean' and 'mode'. Thus,

$S_k = \bar{X} - Z$, [where S_k = skewness; $\bar{X}$ = arithmetic mean; Z = mode]

In a symmetrical distribution, mean and mode coincide, so skewness will be zero. If $\bar{X} > Z$, the skewness will be positive and will have positive sign. If $\bar{X} < Z$, the skewness will be negative and will have negative sign.

- **Karl Pearson's Co-efficient of Skewness**

  Karl Pearson has given a formula for relative measure of skewness. It is also known as Karl Pearson's of coefficient Skewness of Pearsonian Coefficient of Skewness. The formula is that the difference between the mean and mode is divided by the standard deviation.

  $$\text{Coefficient of } S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{\bar{X} - Z}{\sigma} \quad \text{... (1)}$$

  Mode = 3 Median – 2 Mean

  Or $Z = 3M - 2\bar{X}$

  Substituting the value of modes in equation (1)

  $$\text{Coefficient of } S_k = \frac{\bar{X} - (3M - 2\bar{X})}{\sigma} = \frac{3(\bar{X} - M)}{\sigma}$$

  In a distribution, we have more than one mode, *i.e.*, mode is ill-defined, we cannot apply the above state formula. Then we have the following alternative formula:

  $$\text{Coefficient of } S_k = \frac{3(\bar{X} - M)}{\sigma}$$

Notes

- **Coefficient of Skewness in Individual Observations**

Example 1: Calculate the co-efficient of skewness of the following data by using Karl Pearson's method.

Marks:	2	4	4	6	7

Solution :

Marks (X)	$x = (X - A)$ **A = 4**	x^2
2	– 2	4
4	0	0
4	0	0
6	2	4
7	3	9
$\Sigma X = 23$		$\Sigma x^2 = 17$

$$\text{Mean } (\bar{X}) = \frac{\Sigma X}{N} = \frac{23}{5} = 4.6$$

As 4 is repeated twice therefore mode = 4.

Now $$\text{S.D. } (\sigma) = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{17}{5}} = \sqrt{3.40} = 1.844$$

Thus, co-efficient of skewness $$S_k = \frac{\bar{X} - \text{Mode}}{\sigma} = \frac{4.6 - 4}{1.844} = \frac{0.6}{1.844} = 0.325$$

- **Co-efficient of Skewness in Continuous Series**

Example 2: Find out the coefficient of skewness for the following distribution:

Class:	0–10	10–20	20–30	30–40	40–50
Frequency:	14	23	27	21	15

Solution:

Class (x)	**Frequency** (f)	**Mid-value** (m)	**Deviation** $d = (m - A)$ **A = 25**	fd	fd^2
0–10	14	5	– 20	– 280	5,600
10–20	23	15	– 10	– 230	2,300
20–30	27	25	0	0	0
30–40	21	35	+ 10	+ 210	2,100
40–50	15	45	+ 20	+ 300	6,000
	$\Sigma f = 100$			$\Sigma fd = 0$	$\Sigma fd^2 = 16000$

Notes

$$\bar{X} = A + \frac{\sum fd}{\sum f} = 25 + \frac{0}{100} = 25$$

Mode (Z) is located in the class interval 20–30.

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 20 + \frac{27 - 23}{2 \times 27 - 23 - 21} \times 10$$

$$= 20 + \frac{4}{10} \times 10 = 20 + 4 = 24$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{16000}{100} - \left(\frac{0}{100}\right)^2}$$

$$= \sqrt{160} = 12.65 \text{ Approx.}$$

$$\text{Coefficient of Skewness} = \frac{\bar{X} - Z}{\sigma} = \frac{25 - 24}{12.65} = \frac{1}{12.65} = 0.08$$

Example 3: The Karl Pearsons coefficient of skewness of a distribution is 0.32. The Standard Deviation is 6.5 and Mean is 29.6. Find mode.

Solution: Given $S_k = 0.32$, $\bar{X} = 29.6$, $\sigma = 6.5$, $Z = ?$

As we know $S_k = \frac{\bar{X} - Z}{\sigma}$

Substituting the values, we get

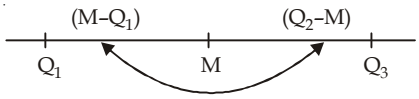
$$0.32 = \frac{29.6 - Z}{6.5}$$

$$\Rightarrow \quad 29.6 - Z = 6.5 \times 0.32$$

$$\Rightarrow \quad Z = 29.6 - 2.08 = 27.52$$

(2) Bowley's Method

This method is called Second Measure of Skewness, which is propounded by Dr. A.L. Bowley. This method is based on relative positions of the median and the two quartiles. In a symmetrical distribution, the upper and lower quartiles are equidistant from median *i.e.*,



Mathematically,

$$Q_3 - M = M - Q_1 \text{ or } Q_3 + Q_1 = M + M \text{ or } Q_3 + Q_1 - 2M = 0$$

Thus, here skewness is absent.

But in an asymmetrical distribution first and third quartiles are not equidistant from median, *i.e.*,

$$Q_3 + Q_1 - 2M \neq 0$$

Notes

In this case skewness is present. It is important to note that if Q_1 is farther away from median than the Q_3, then skewness will be negative and if the case is opposite the skewness will be positive. Dr. Bowley has given the following method of skewness.

$$S_k = (Q_3 - M) - (M - Q_1) = Q_3 + Q_1 - 2M$$

Coefficient of Skewness is

$$\text{Coefficient of } S_k = \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)}$$

$$\text{Coefficient of } S_k = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

The calculation of median and quartiles in the case of individual, discrete and continuous series is already explained in Unit - 4.

Example 4: If sum and difference of two quartiles are 22 and 8 respectively. Find the co-efficient of skewness when the median is 10.5.

Solution: Given $Q_3 - Q_1 = 8$; $Q_3 + Q_1 = 22$ and M = 10.5

$$\text{Now, Coefficient of } S_k = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{22 - 2(10.5)}{8} = \frac{22 - 21}{8} = \frac{1}{8} = 0.125$$

Example 5: If Bowley's co-efficient of skewness is – 0.36, Q_1 = 8.6 and median = 12.3. What is the quartile co-efficient of dispersion ?

Solution: Given, Bowley's Coefficient of S_k = – 0.36, Q_1 = 8.6, M = 12.3 Coefficient of Q.D = ?

$$\text{Coefficient of } S_k = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} \text{ or } -0.36 = \frac{Q_3 + 8.6 - 2 \times 12.3}{Q_3 - 8.6}$$

or $-0.36(Q_3 - 8.6) = Q_3 + 8.6 - 24.6$

or $-0.36Q_3 + 3.096 = Q_3 - 16$

or $-0.36Q_3 - Q_3 = -16 - 3.096$

or $-1.36Q_3 = -19.096$

or $Q_3 = \frac{19.095}{1.36} = 14.04$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{14.04 - 8.6}{14.04 + 8.6} = \frac{5.44}{22.64} = 0.24$$

(3) Kelly's Method

Prof. Kelly has given a formula which is based on deciles or percentiles. It is defined as

or $S_k = P_{90} + P_{10} - 2P_{50}$

or $S_k = D_9 + D_1 - 2D_5$

Coefficient of skewness is defined as

Notes

$$\text{Coefficient of } S_k = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})} = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$$

$$\text{Coefficient of } S_k = \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

Example 6: Find out the Kelly's co-efficient of skewness of the data given below:

Class:	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency:	3	10	17	7	6	4	2	1

Solution:

Class	**Frequency (*f*)**	**Cumulative Frequency (*c.f.*)**
0–10	3	3
10–20	10	13
20–30	17	30
30–40	7	37
40–50	6	43
50–60	4	47
60–70	2	49
70–80	1	50
	$\Sigma f = 50$	

P_{10} = Size of $\frac{10N}{100}$ th item = $\frac{10(50)}{100}$ th item = 5th item

P_{10} = Size of 5th item which lies in 10–20 group

$$P_{10} = l_1 + \frac{\left(\frac{10N}{100} - c.f.\right)}{f} \times i$$

$$P_{10} = 10 + \frac{(5-3)}{10} \times 10 = 10 + 2 = 12$$

P_{50} = Size of $\frac{50 \times N}{100}$ th item = $\frac{50 \times 50}{100}$ th item = 25th item.

P_{50} = Size of 25th item which lies in 20-30 group.

$$P_{50} = l_1 + \frac{\left(\frac{50N}{100} - c.f.\right)}{f} \times i$$

$$P_{50} = 20 + \frac{(25-13)}{17} \times 10 = 20 + \frac{120}{17} = 27.06$$

Notes

$$P_{90} = \text{Size of } \frac{90 \times N}{100} \text{ th item} = \frac{90 \times 50}{100} \text{ th item} = 45\text{th item.}$$

Now P_{90} = Size of 45th item which lies in 50–60 group.

$$P_{90} = l_1 + \frac{\frac{90N}{100} - c.f.}{f} \times i$$

$$P_{90} = 50 + \frac{(45-43)}{4} \times 10 = 50 + \frac{20}{4} = 55$$

Measures of Skewness at a Glance

Methods	Formula
1. Karl Pearson's Method	
(a) Absolute Skewness	$S_k = \bar{X} - Z$
When mode is ill-defined	$S_k = 3(\bar{X} - M)$
(b) Coefficient of Skewness	Co-efficient of $S_k = \frac{\bar{X} - Z}{\sigma} = \frac{3(\bar{X} - M)}{\sigma}$
When mode is ill-defined	
2. Bowley's Method	
(a) Absolute Skewness	$S_k = (Q_3 - M) - (M - Q_1) = Q_3 + Q_1 - 2M$
(b) Coefficient of Skewness	$= \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$
3. Kelly's Method	
(a) Absolute Skewness	$= P_{90} + P_{10} - 2P_{50}$ or $D_9 + D_1 - 2D_5$
(b) Coefficient of Skewness	$= \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$ or $= \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$

8.3 Kurtosis

Besides averages, variation and skewness, a fourth characteristic used for description and comparison of frequency distributions is the peakedness of the distribution. Measures of peakedness are known as measures of **kurtosis.**

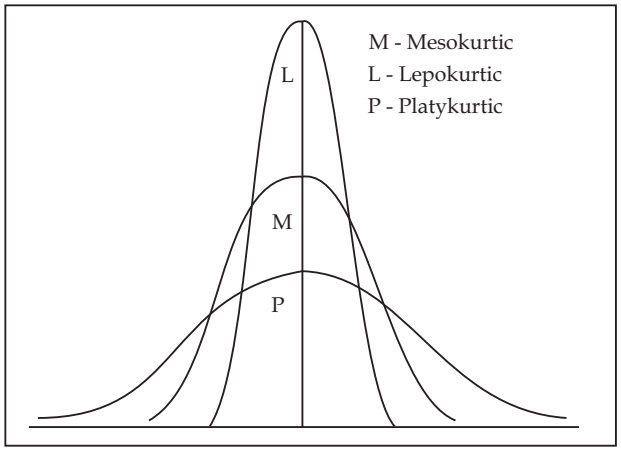
Notes Kurtosis in Greek means "*bulginess*". In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve.

In other words, measures of kurtosis tell us the extent to which a distribution is more peaked or flat-topped than the normal curve. If a curve is more peaked than the normal curve, it is called '*leptokurtic*'.

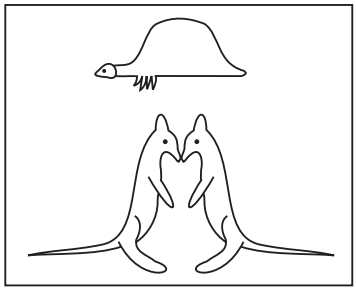
Notes

In such a case the items are more closely bunched around the mode. On the other hand, if a curve is more flat-topped than the normal curve, it is called '*platykurtic*'. The normal curve itself is known as '*mesokurtic*'. The condition of peakedness or flat-toppedness itself is known as kurtosis or excess. The concept of kurtosis is rarely used in elementary statistics.

The following diagram illustrates the shapes of three different curves mentioned above:



The above diagram clearly shows that these curves differ widely with regard to convexity, an attribute which Karl Pearson referred to as 'kurtosis'. Curve M is a normal one and is called 'mesokurtic'. Curve L is more peaked than M and is called 'leptokurtic'. Curve P is less peaked (or more flat-topped) than curve M and is called 'platykurtic'.



A famous British statistician Willian S. Gosset ("Student") has very humorously pointed out the nature of these curves in the sentence, "Platykurtic curves, like the platypus, are squat with short tails; lepto-kurtic curves are high with long tails like the kangaroos noted for lapping." Gosset's little sketch is reproduced above.

Measures of Kurtosis

The most important measure of kurtosis is the value of the coefficient β_2. It is defined as:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \text{ where } \mu_4 = \text{4th moment and } \mu_2 = \text{2nd moment.}$$

For a normal curve the value of $\beta_2 = 3$. When the value of β_2 is greater than 3 the curve is more peaked than the normal curve, *i.e.*, leptokurtic. When the value of β_2 is less than 3 the curve is less peaked than the normal curve, *i.e.*, platykurtic. The normal curve and other curves with $\beta_2 = 3$ are called mesokurtic.

Sometimes γ_2, the derivative of β_2, is used as a measure of kurtosis, γ_2 is defined as

$$\gamma_2 = \beta_2 - 3.$$

Notes

For a normal distribution $\gamma_2 = 0$. If γ_2 is positive, the curve is leptokurtic and if γ_2 is negative, the curve is platykurtic.

Example 7: The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the skewness and kurtosis of the distribution.

Solution:

Testing Skewness

We are given $\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$ and $\mu_4 = 18.75$

Skewness is measured by the coefficient β_1

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

Here $\mu_2 = 2.5$, $\mu_3 = 0.7$

Substituting the values, $\beta_1 = \frac{(0.7)^2}{(2.5)^3} = +0.031$

Since $\beta_1 = +0.031$, the distribution is slightly skewed.

Testing Kurtosis:

For testing kurtosis we compute the value of β_2. When a distribution is normal or symmetrical, $\beta_2 = 3$. When a distribution is more peaked than the normal, β_2 is more than 3 and when it is less peaked than the normal, β_2 is less than 3.

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$\mu_4 = 18.75$, $\mu_2 = 2.5$

$$\therefore \quad \beta_2 = \frac{18.75}{(2.5)^2} = \frac{18.75}{6.25} = 3$$

Since β_2 is exactly three, the distribution is mesokurtic.

Self-Assessment

1. Fill in the blanks:

(*i*) If $Q_3 = 30$, $Q_1 = 20$, Med = 25, Coeff. of Sk. shall be .....

(*ii*) If $\overline{X} = 50$, Mode = 48, $\sigma = 20$, the Coefficient of Skewness shall be .....

(*iii*) If Coeff. of Sk. = 0.8, Median = 35, $\sigma = 12$, the mean shall be .....

(*iv*) In a symmetrical distribution the coefficient of skewness is .....

(*v*) The limits for Bowley's coefficient of skewness are .....

8.4 Summary

- The nature of distribution is further studied deeply by calculating 'Moments' which reveals whether the symmetrical curve is normal, more flat than normal or more peaked than normal.

Notes

Similarly, Kurtosis is yet another measure which tells us about the form of a distribution. Thus, it can be said that the central tendencies and dispersion measures should be supplemented by measures of skewness and kurtosis so that a more elaborate picture about the distribution given can be obtained. The study becomes more important in subjects of economics, sociology and other social sciences where normal distribution in a series usually does not occur. However, studies hold importance in biological sciences and other physical sciences as well.

- In the words of *Simpson* and *Kafka*, "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the arithmetic mean, median and mode are identicle. The more the mean moves away from mode, the larger the asymmetry or skewness."
- Measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with symmetrical or normal distribution. In a symmetrical distribution, frequencies go on increasing upto a point and then begin to decrease in the same fashion. There are various possible patterns of symmetrical distribution and normal distribution which is bell-shaped is one of these.
- A distribution in which more than half of the area under the curve is to the right side of the mode, it is said to be a positively skewed distribution. In this type of skewness the right tail is longer than the left tail. In this case, mean is greater than median and the median is greater than the mode and $Q_3 - M > M - Q_1$.
- A distribution in which more than half of the area under the distribution curve is to the left side of the mode, it is said to be a negatively skewed distribution. In this case, the elongated tail is to the left and mean is less than the median which is less than mode and $Q_3 - M < M - Q_1$.
- Kurtosis in Greek means "*bulginess*". In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve. In other words, measures of kurtosis tell us the extent to which a distribution is more peaked or flat-topped than the normal curve. If a curve is more peaked than the normal curve, it is called '*leptokurtic*'. In such a case the items are more closely bunched around the mode. On the other hand, if a curve is more flat-topped than the normal curve, it is called '*platykurtic*'. The normal curve itself is known as '*mesokurtic*'. The condition of peakedness or flat-toppedness itself is known as kurtosis or excess. The concept of kurtosis is rarely used in elementary statistics.
- A famous British statistician Willian S. Gosset ("Student") has very humorously pointed out the nature of these curves in the sentence, "Platykurtic curves, like the platypus, are squat with short tails; lepto-kurtic curves are high with long tails like the kangaroos noted for lapping."

8.5 Key-Words

1. Skewness and Kurtosis : Skewness and kurtosis are terms that describe the shape and symmetry of a distribution of scores. Unless you plan to do inferential statistics on your data set skewness and kurtosis only serve as descriptions of the distribution of your data. Be aware that neither of these measures should be trusted unless you have a large sample size.

   Skewness refers to whether the distribution is symmetrical with respect to its dispersion from the mean. If on one side of the mean has extreme scores but the other does not, the distribution is said to be skewed. If the dispersion of scores on either side of the mean are roughly symmetrical (i.e. one is a mirror reflection of the other, the distribution is said to be not skewed.

2. Kelly's Methods : In probability theory, the Kelly criterion, or Kelly strategy or Kelly formula, or Kelly bet, is a formula used to determine the optimal

Notes

size of a series of bets. In most gambling scenarios, and some investing scenarios under some simplifying assumptions, the Kelly strategy will do better than any essentially different strategy in the long run. It was described by J. L. Kelly, Jr in 1956.[1] The practical use of the formula has been demonstrated.

8.6 Review Questions

1. What do you understand by skewness ? Give various definitions. What are the various methods of measuring it.
2. Give the concept of kurtosis.
3. Distinguish between Pearson's and Bowley's measure of skewness.
4. State the formula for calculating Karl Pearson's coefficient of skewness.

Answers: Self-Assessment

1. (*i*) 0 (*ii*) 0.1 (*iii*) 108.2 (*iv*) zero (*v*) ± 1

8.7 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 9: Correlation: Definition, Types and its Application for Economists

CONTENTS

Objectives

After reading this unit students will be able to:

- Know Correlation and Types of Correlation.
- Discuss the Application of Correlation for Economists.

Introduction

Correlation means a relation between two groups. In statistics, it is the measure to indicate the relationship between two variables in which, with changes in the values of one variable, the values of other variable also change. These variables may be related to one item or may not be related to one item but have dependence on the other due to some reason. For example, the data on height and weights of a group of people would relate to each member of the group but prices of sugar and sugarcane are two different series altogether but there would be some relation between the values of the two, prices of sugar depending upon the prices of sugarcane. This technique provides a tool into the hands of decision-makers because it provides better understanding of the trends and their dependance on other factors so that the range of uncertainities associated with decision-making is reduced.

9.1 Definition and Types of Correlation

Definition of Correlation

The term correlation indicates the relationship between two variables in which with changes in the value of one variable, the values of the other variable also change. Correlation has been defined by various eminent statisticians, mathematicians and economists. Some of the important definitions of correlation are given below:

(1) According to *La Yun Chow*, "Correlation analysis attempts to determine the degree of relationship between variables."

(2) As per *W. I. King*, "Correlation means that between two series or groups of data there exists some casual connections. ..... If it is proved true that in a large number of instances two variables tend always to fluctuate in the same or in opposite directions, we consider that the fact is established and that a relationship exists. This relationship is called correlation."

(3) In the words of *L. R. Conner*, "If two more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other/others then they are said to be correlated.

Notes

(4) *Croxton* and *Lowden* define correlation as, "When the relationship of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

(5) As per *Prof. Boddington,* "Whenever some definite connection exists between two or more groups, classes or series of data, there is said to be correlation."

From the above definitions it can be said that correlation is a statistical tool which requires about the relationship between two or more variables.

Utility: Correlation has immense utility in various fields of knowledge. Some of the important areas where correlation has been used successfully are:

(1) In the field of genetics: Galton and Pearson developed a method of assessing correlation which was used in studying many problems of biology and genetics.

(2) In the field of management: Basically, management is all about making decisions. Correlation technique presents a strong tool into the hands of the manager which reduces the range of uncertainity associated with decision-making. Moreover, it also helps in identifying the stabilishing factors for a disturbed economic situation.

(3) Other field of social sciences: Correlation helps in determining the interrelationships between different variables and in this way it is very helpful in promoting research and opening new frontiers of knowledge.

In this way it can be said that correlation has immense utility in various fields in promoting research and opening new frontiers of knowledge.

Did u know? Correlation is very useful in understanding the economic behaviour. It helps in locating those variables on which other variables depend. In this way various economic events can be analysed.

"Correlation" and "cause and effect relationship"

Correlation measures a degree of the relationship between two or more variables but it does not indicate any kind of cause and effect relationship between the variables. If, high degree of correlation is found exist between two variables, it implies that there must be a reason for such close relationship, but the cause and effect relation can be revealed specifically when other knowledge of the factor involved being brought to bear on the situation. This means, to establish a 'functional relationship' between two or more variables, one has to go beyond the confines of statistical analysis to other factors. (Functional relationship means that two or more factors are interdependent. In fact, although, high degree of correlation may mean that two or more variables are mutually dependent, but at the same time, this high degree of correlation may be due to many other reasons like:

(1) The two variables are being affected by a third variable or by more than one variable.

(2) The two variables might be mutually affecting each other and neither of them is the cause or the effect.

(3) The high degree of correlation between two variables comes out just by chance or by sheer coincidence.

Therefore, although high degree of correlation does not necessarily indicate the cause and effect relationship. The quantitative tool requires the support of proper knowledge and logic about the variables on the basis of which the results should be interpreted. In this way, although 'correlation' in a strong tool it needs to be used carefully by those who have knowledge otherwise its misuse is quite likely.

Types of Correlation

Correlation can be classified as given ahead:

(1) Positive and negative correlation: When the values of the two variables move in the same direction, *i.e.,* an increase in one is associated with an increase in other, or *vice versa,* the correlation

Notes

is said to be positive. If the values of two variables move in the opposite directions *i.e.,* an increase in the value of one variable is associated with fall in other, or *vice versa,* the correlation is said to be negative. For example, the price and supply are positively correlated but price and demand are negatively correlated.

(2) **Linear and non-linear correlation:** If, in response to a unit change in the value of one variable, there is a constant change in the value of the other variable, the correlation between them is said to be linear. This means, the relation between variables fits in $Y = a + bX$. But when no constant change in variable is registered for a given unit change in other variable, non-linear or curvilinear correlation is said to exist.

(3) **Simple, multiple and partial correlation:** When relation between two variables is studied, it is simple correlation. When three or more factors are studied together to find relationships, it is called multiple correlation. In partial correlation, two or more factors are agreed to be involved but correlation is studied between only two factors, considering other factors to be constant.

9.2 Application of Correlation for Economists

The cause and effect relation existing between economic events is especially difficult to ascertain because of the presence of innumerable variable elements. In solving his problems the economist can not, like the physicist or chemist, eliminate all causes except one and then by experiment determine the effect of that one. Causes must be dealt with *en masse.* Since any effect is the result of many combined causes the economist is never sure that a given effect will follow a given cause. In stating an economic law he always has to postulate "other things remaining the same," with, perhaps, little appreciation of what the other things may be. It is rarely, if ever, possible for the economist to state more than "such and such a cause *tends* to produce such and such an effect." Events can only be stated to be more or less probable. He is dealing mainly, therefore, with correlation and not with simple causation.

The problems of economics are similar to certain problems of biology, such as the effect of environment and heredity upon the individual. In dealing with the question of heredity Karl Pearson says: "Taking our stand then on the observed fact that a knowledge neither of parents nor of the whole ancestry will enable us to predict with certainty in a variety of important cases the character of the individual offspring we ask: What is the correct method of dealing with the problem of heredity in such cases ? The causes A, B, C, D, E , . . . which we have as yet succeeded in isolating and defining are not always followed by the effect X, but by any one of the effects U, V, W, X, Y, Z. We are therefore not dealing with causation but correlation, and there is therefore only one method of procedure possible; we must collect statistics of the frequency with which U, V, W, X, Y, Z, respectively, follow on A, B, C, D, E . . . From these statistics we know the most *probable* result of the causes A, B, C, D, E and the frequency of each deviation from this most probable result. The recognition that in the existing state of our knowledge the true method of approaching the problem of heredity is from the statistical side, and that the most that we can hope at present to do is to give the *probable* character of the offspring of a given ancestry, is one of the great services of Francis Galton to biometry."

Just as the biologists cannot predict a man's height or color of eyes or temper or combativeness by knowing those qualities in his ancestors, so economists cannot predict that a definite call rate in Wall Street will go with a given percentage of reserves to deposits in New York banks or that a given supply of wheat will result in a definite price per bushel. But, on the other hand, just as it has been observed that there *is* a relation existing between a man's stature and the stature of his ancestors, so it has been observed that a relation *does* exist between bank reserves and call rates and between supply of wheat and its price per bushel.

In order to deal in a satisfactory way with such questions as those given above it is necessary to accumulate statistics of the supposedly related phenomena. In order to have those statistics indicate anything it is necessary to obtain a method of measuring the extent of correlation between the phenomena.

The commonly used method of measuring the amount of correlation between any two series of economic statistics is to represent the two series graphically upon the same sheet of cross-section paper and then compare the fluctuations of one series with those of the other. The quantity theory of

Notes

prices has been tested in this way by Dr. E. W. Kemmerer. Dr. Kemmerer builds up the following price equation:

$$P_s = \frac{MR+CR_c}{NE+N_cE_c}$$

in which:

P_s = the average price (weighted by the total flows) of all commodities sold for money and deposit currency during a unit of time.

M = the total currency in circulation during the unit of time
R = the average number of times each unit of currency changes hands during the unit of time.
} = the flow of currency

NE = the flow of goods exchanged for currency.

C = the volume of deposit currency exchanged for goods.
R_c = the average rate of turnover of such deposit currency.
} = flow of deposit currency.

N_cE_c = the flow of goods exchanged for deposit currency.

Dr. Kemmerer then attempts to find the answer that facts give to the following questions:

1. Do the bank reserves vary directly with the money supply ?
2. Does the proportion of bank reserves to check circulation vary directly with the degree of business distrust existing in the country ?
3. Is "a relative increase in the circulating media accompanied by a corresponding and proportionate increase in general prices and a relative decrease in the circulating media, by a corresponding and proportionate decrease in general prices," or, in the language of the formula, is

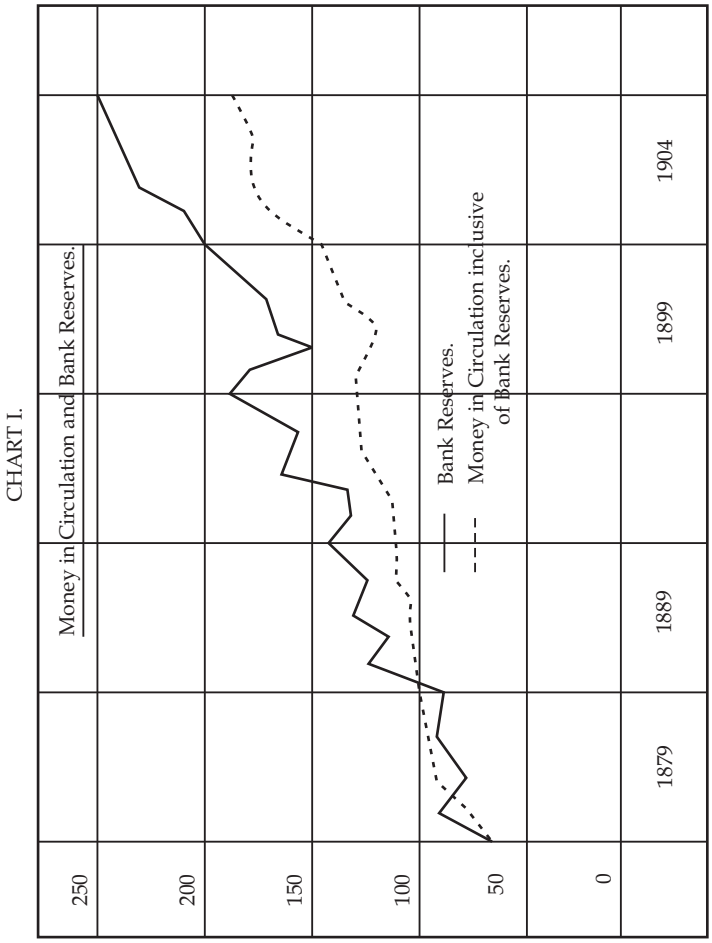
$$P_s = \frac{MR+CR_c}{NE+N_cE_c}$$

borne out by the facts ?

All of the questions to be tested by the statistics collected are questions of correlation. Dr. Kemmerer makes the tests graphically, as has been stated, by comparing the fluctuations of the two curves based upon the pair of series of statistics being considered. The charts presented by Dr. Kemmerer from which his conclusions are drawn are given below.

In the case of the correlation of bank reserves and money in circulation, inclusive of bank reserves, Dr. Kemmerer concludes, "There can be no question but that when due allowance is made for fluctuations in business confidence, the evidence of Chart I strongly supports the contention that there exists a close relationship between the amount of money in circulation and the amount of the country's bank reserves." In the case of the correlation of business distrust and the ratio of bank reserves to check circulation the conclusion is, "the chart substantiates the contention . . . that the ratio of check circulation to bank reserves is a function of business confidence . . . "The final test of the quantity theory is the amount of correlation between the figures for the right and left-hand sides of the equation $P_s = \frac{MR+CR_c}{NE+N_cE_c}$.

Notes



Upon examination of the curves plotted from the two series of statistics representing general prices and relative circulation (the left and right-hand sides, respectively, of the price equation) Dr. Kemmerer concludes, "The general movement of the two curves taken as a whole is the same, while the individual variations from year to year exhibit a striking similarity."

The graphic method of comparing fluctuations is well enough as a preliminary, *but does it enable anyone to tell anything of the extent of the correlation between the series of figures being considered* ? Is Dr. Kemmerer warranted in deducing his conclusions from observation of the charts ? It seems to the writer that one opposing the quantity theory might draw opposite conclusions with as much (or as little) reason. *The charts do not answer the questions proposed.* The painstaking collection of statistics to test correlation is useless if there be no more reliable method to measure correlation. A numerical measure of the correlation must be found if we wish to determine the *extent* to which the fluctuations of one series synchronize with the fluctuations of another series.

A second illustration of a conclusion based upon graphic representation is that of Ira Cross in his study of strike statistics. He says, upon consideration of data taken from the Twenty-first Annual Report of the United States Bureau of Labor, "the percentage of successful strikes decreases during periods of business prosperity and increases during 'hard times.' " In the accompanying charts the per cent. of establishments in which strikes were successful is plotted, first, with the per capita exports and imports and second, with index numbers of wholesale prices. The foreign trade and the price statistics are taken as indicative of the activity of business, as indices of prosperity.

A third illustration of a conclusion relating to correlation is taken from the *London Statist* of April 4, 1908, where the proposition is made that, "When commodities advance prices of Stock Exchange securities recede; when commodities recede Stock Exchange securities advance." The proposition is

Notes

supported by reference to the following chart showing the yearly average price of consols and Sauerbeck's index numbers of prices.

The foregoing illustrations show the need by economists of a quantitative measure of correlation. Such a measure has been widely used in biological statistics and used to a limited extent in economic statistics. G.U. Yule has used the measure in his study of "Pauperism ;" R.H. Hooker has used it in his "Correlation of the Weather and Crops;" J. P. Norton applied it in his study of the "New York Money Market." This measure, the coefficient of correlation, will be computed for the data upon which the conclusions quoted above are based. The formula for the coefficient of correlation is

$$r = \frac{\sum xy}{n\sigma_1\sigma_2};$$

where:

x = deviation from arithmetic mean = $X - M_1$

y = deviation from arithmetic mean= $Y - M_2$

σ_1 = standard deviation of X series

σ_2 = standard deviation of Y series

n = number of items.

The coefficient of correlation "serves as a measure of any statement involving two qualifying adjectives, which can be measured numerically, such as tall men have tall sons,' 'wet springs bring dry summers,' 'short hours go with high wages.' " It is not the purpose in what follows to go through the mathematical derivation of the coefficient of correlation, but to test the formula empirically in order to ascertain how it actually varies for given series of statistics and to point out some of its features.

However, it should be noted at this point that the coefficient of correlation is not empirical but was derived by *a priori* reasoning. It was found by assuming that a large number of independent causes operate upon each of the two series X and Y, producing normal distributions in both cases. Upon the assumption that the set of causes operating upon the series X is *not independent* of the set of causes operating upon the series Y the value $r = \frac{\sum xy}{n\sigma_1\sigma_2}$ is obtained. This value becomes zero when the operating causes are absolutely independent. Hence the value of r was taken as a measure of correlation. In what follows *no assumption concerning the type of distribution of the X and Y series will be made.*

Some appreciation of the meaning of the coefficient of correlation can be obtained by the consideration of a few simple applications. Suppose that we consider the two series of measurements:

X = l, 2, 3, 4, 5 $\quad M_1 = 3$

Y = 6, 8, 10, 12, 14 $\quad M_2 = 10$

Deviations.		Square of Deviation.		Product of Deviations.	
x	y	x^2	y^2	xy	
– 2	– 4	4	16	8	$\sigma_1 = \sqrt{2}$
– 1	– 2	1	4	2	
0	0	0	0	0	$\sigma_2 = 2\sqrt{2}$
+ 1	+ 2	1	4	2	$r = \frac{20}{5\sqrt{2}.2\sqrt{2}} = 1$
+ 2	+ 4	4	16	8	

Notes

In the above illustration the numbers were chosen so that for an increase of 1 unit in the X series there is an increase of 2 units in the Y series. Thus the correlation is perfect and r equals +1. If the Y series had been 14, 12, 10, 8, 6 (the X series remaining the same) the value of r would have been −1. Thus −1 stands for perfect *negative* correlation, an increase in one series corresponding to a decrease in the other. It should also be noted in this connection that the coefficient of correlation (r) cannot be less than −1 nor more than + l.

The above illustration suggests the question, "Will a linear relationship between X and Y *always* give perfect correlation ?"

Assume the linear relationship

$$Y = aX + b$$

Since $y = Y - M_2$ and $x = X - M_2$

$$M_2 + y = a(x + M_1) + b \text{ or } y = ax$$

(since $b - aM_1 - M_2 = 0$)

and $$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}} = \frac{\sum ax^2}{\sqrt{\sum x^2 . \sum a^2 x^2}} = \frac{a\sum x^2}{\sqrt{a^2\left(\sum x^2\right)^2}} = \pm 1$$

(The sign of r depends upon the sign of a.)

Therefore a linear relationship between two variables will give a correlation coefficient of +1 or −1 depending upon whether large values of one occur with large values of the other or large values of one occur with small values of the other.

The converse of the above proposition is likewise true, *i.e.,* if the coefficient of correlation (r) equals 1 then the relationship between the X and Y series is linear.

Assume $r = 1$

then $\left(\sum xy\right)^2 - \sum x^2 . \sum y^2 = 0$

Letting $x_1 = \lambda_1 y_1, x_2 = \lambda_2 y_2 \ldots x_n = \lambda_n y_n$ the above expression becomes

$$y_1^2 y_2^2 (\lambda_1 - \lambda_2)^2 + y_1^2 y_3^2 (\lambda_1 - \lambda_3)^2 + \ldots + y_r^2 y_s^2 (\lambda_r - \lambda_s)^2 + \ldots = 0$$

The only way in which this expression can equal zero is by having

$$\lambda_1 = \lambda_2 = \lambda_3 = \ldots = \lambda_n$$

and it follows that

$$x_1 = \lambda_1 y_1, x_2 = \lambda_1 y_2 \ldots x_n = \lambda_1 y_n$$

or

$$x = \lambda_1 y$$

which denotes a linear relationship between X and Y.

That any relation other than a linear one will not lead to r = l is illustrated by the following:

Let $Y = X^2$

$X = 1, 2, 3, 4, 5,$ $\quad M_1 = 3$

$Y = 1, 4, 9, 16, 25,$ $\quad M_2 = 11$

Notes

x	y	x^2	y^2	xy	
– 2	– 10	4	100	20	
– 1	– 7	1	49	7	σ_1 = 1.41
0	– 2	0	4	0	σ_2 = 8.65
+ 1	+ 5	1	25	5	r = 0.981
+ 2	+ 14	4	196	28	
Total		10	374	60	

Although the two series increase regularly, so that deviations of like signs always correspond, yet the correlation is not perfect *because a linear relation does not exist between X and Y.*

If the number of items in each series be increased to 11 and the Y items remain squares of the X's the value of *r* will be 0.974.

If there be no law connecting the X and Y series the products of the deviations (*xy*) are as apt to be negative as positive. The expression Σxy will therefore tend to approach zero. With a very large number of measurements the correlation coefficient will approximate zero.

From the condition of no relationship to the condition of a linear relationship existing between the pair of series of measurements the correlation coefficient varies from 0 to ± 1.

Suppose that we are investigating the relation existing between two series of measurements X and Y. Let points be plotted on cross-section paper whose coordinates are corresponding measurements X_1 and Y_1. If there be a relationship existing between the two series, the points thus located will not lie chaotically all over the plane, but they will range themselves about some curve or locus. This curve, which has been called the *curve of regression,* is illustrated in the accompanying diagram. The straight line best fitting the points is called the line of regression.

For example suppose we consider the two series of index numbers for the period 1879-1904 inclusive, representing (1) money in circulation in the United States inclusive of bank reserves, and (2) bank reserves. Let points be located with abscissas proportionate to the money in circulation and with ordinates proportionate to the bank reserves of the same year. The chart on the next page shows that these points lie near a straight line, the line of regression.

The coefficient of correlation (*r*) *is a measure of the closeness of the grouping of the points about this line of regression.* If the points should all range themselves on a line then *r* would equal + 1 or — 1 depending upon whether, looking left to right, the line sloped upward or downward.

We will now derive the equation of the line of regression. Let X and Y be associated measurements and *x* and *y* be associated deviations from the respective arithmetic means. A *linear* relation between the measurements is of the form

$$Y = a_1X + b_1$$

The relation between the deviations will be of form

$$y = a_1x \text{ or } y - a_1x = 0$$

Since all of the points are not located exactly upon a straight line the substitution of the values x_1, y_1, x_2, y_2, etc. in the equations will give residues v_1, v_2, etc. as follows:

$$y_1 - a_1x_1 = v_1$$

$$y_2 - a_1x_2 = v_2$$

$$y_n - a_1x_n = v_n$$

Notes

The values $\frac{v_1}{\sqrt{1+a_1^2}}, \frac{v_2}{\sqrt{1+a_1^2}}, \ldots . \frac{v_n}{\sqrt{1+a_1^2}}$ equal the distances of the various points to the straight line $y = a_1x$.

The equation of a line such that the sum of the squares of the distances from the given points is a minimum will now be found. In other words that value of a_1 will be taken which makes $v_1^2 + v_2^2 + \ldots . v_n^2$ = a minimum. To find the value of a_1, for which $(y_1 - a_1x_1)^2 + (y_2 - a_1x_2)^2 + \ldots . + (y_n - a_1x_n)^2$ will be a minimum, differentiate with respect to a_1 and obtain $-2x_1(y_1 - a_1x_1) - 2x_2(y_2 - a_1x_2) \ldots . -2x_n(y_n - a_1x_n)$. In order that the original function be a minimum, this derivative must equal zero. We will then have

$$(x_1y_1 - a_1x_1^2) + (x_2y_2 - a_1x_2^2) \ldots . + (x_ny_n - a_1x_n^2) = 0, \text{ or}$$

$$\Sigma xy - a_1 \Sigma x^2 = 0$$

$$a_1 = \frac{\Sigma xy}{\Sigma x^2}$$

Similarly if $x = a_2y$, then $\Sigma xy - a_2 \Sigma y^2 = 0$ will give the value of a_2 for which the sum of the squares of the distances of the given points to the straight line $X = a_2Y + b_2$ is a minimum, or

$$a_2 = \frac{\Sigma xy}{\Sigma y^2}$$

Let $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = \frac{\Sigma xy}{n\sigma_1\sigma_2} \text{ and } \Sigma x^2 = n\sigma_1^2 \Sigma y^2 = n\sigma_2^2.$$

The equations between the deviations are:

$$y = r \frac{\sigma_2}{\sigma_1} x$$

$$x = r \frac{\sigma_1}{\sigma_2} y$$

It may seem that the two equations just given are inconsistent. But it must be remembered that these equations do not give the relationship existing between *any* corresponding pair of deviations unless all of the points lie exactly on a straight line and there be perfect correlation. For all cases of imperfect correlation a *given* deviation x will occur with several different deviations y (if we have a large number of measurements). If these deviations y are distributed according to the normal law of distribution then the given value x substituted in the first equation will give the mean of the deviations occuring with the deviation x and if a given value y be substituted in the second equation the value of x resulting will be the mean of the deviations of the associated characteristics.

Since $$y = Y - M_2 \text{ and } x = X - M_1$$

$$Y = M_2 + r\frac{\sigma_2}{\sigma_1}(X - M_1)$$

Notes

and $$X = M_1 + r\frac{\sigma_1}{\sigma_2}(Y - M_2)$$

The coefficients $r\frac{\sigma_2}{\sigma_1}$ and $r\frac{\sigma_1}{\sigma_2}$ are called the coefficients of regression of Y upon X and of X upon Y respectively. The first coefficient $\left(r\frac{\sigma_2}{\sigma_1}\right)$ and the reciprocal of the second $\left(\frac{\sigma_2}{r\sigma_1}\right)$ are the *slopes* of the lines of regression. If X and Y be measured in terms of their respective standard deviations as units the slopes of the lines of regression will be *r* and $\frac{1}{r}$. *In other words, the slope of the line of regression of Y upon X, each series being measured in terms of its standard deviation, is equal to the coefficient of correlation for the two series. For perfect positive correlation the line would make an angle of 45° with the X axis for perfect negative correlation the line would make an angle of 135° with the x axis, and for no correlation the line would be parallel to the x axis.*

The correlation coefficients show that there is a very great difference in the degree of correlation of different pairs of series of statistics. The full significance of the "probable error," which is used as a measure of unreliability of any determination, cannot be developed at this point. It is sufficient to note that, "When *r* is not greater than its probable error we have no evidence that there is any correlation, for the observed phenomena might easily arise from totally unconnected causes; but, when *r* is greater than, say, six times its probable error, we may be practically certain that the phenomena are not independent of each other, for the chance that the observed results would be obtained from unconnected causes is practically zero."

The high degree of correlation (+0.98) between money in circulation inclusive of bank reserves and bank reserves is due to the tendency of the two items to vary together during the long time period and not due to correspondence of minor fluctuations. The reasons for the great increase of money in circulation in the United States during the period 1879-1904 are the great increase of population and the industrial expansion. Likewise the number of banks increased in order to serve the increased population and this meant an increase of total reserves. It is self-evident that the long time tendency of the two series of statistics must be upward in a growing country. It seemed to me that the bank reserves during the 26 years, 1879-1904, would be as closely correlated with the *population* as with total circulation. The computation of the correlation coefficient between bank reserves and population gave +0.98. It is the variation upwards of both series during the entire period that causes the high coefficient.

The correlation coefficient between the index numbers of business distrust and the rates of bank reserves to check circulation for the same years is 0.53. When the index numbers of business distrust for one year are correlated with the ratio of bank reserves to check circulation the following year the coefficient is 0.72. As Dr. Kemmerer has suggested (but not verified), there is a closer correlation "when proper allowance is made for the time required for alterations in business confidence to exert their influence on bank reserves."* The lowest correlation (+ 0.23), that between relative circulation and general prices, is not high enough to warrant a conclusion that the items vary together. The smallness of the correlation indicated may have resulted either because the quantity theory is in error or because the statistics are not adequate to test the theory. Whatever may be the fact, the statistics and the method of measuring correlation presented by Dr. Kemmerer do not demonstrate that general prices move in sympathy with relative circulation.

The amount of correlation indicated in each case is small—considering the number of years taken, so small that no conclusion as to the connection between the two series can be drawn. The correlation coefficient in the last instance, *i. e.,* between per cent. of successful strikes and business distrust, suggests an opposite conclusion to that indicated by the other coefficients and that of Mr. Cross. The

Notes

analysis shows that the conclusion that there is negative correlation between *general* prosperity and per cent. of successful strikes is not warranted.

Finally, what is the degree of correlation between the prices of British Consols and Sauerbeck's index numbers of the prices of commodities ? The chart on indicates a greater degree of correlation (negative) between the *minor* fluctuations of the two series than shown by any of the pairs of series that we have considered. The coefficient of correlation based upon statistics for the 57 years from 1851 to 1907, inclusive, is -0.58 ± 0.06. A correlation coefficient of -0.58 based upon 57 pairs of items warrants the conclusion that the two series have inverse movements.

The relations between the *average* deviations, x and y, of the two series of statistics being considered are:

$$y = -1.465x \text{ and } x = -0.2295x$$

The equations of regression are:

$$Y = 225.6 - 1.465\ X \text{ and } X = 19.439 - 0.2295\ Y$$

For certain pairs of time-series (corresponding items occur at same time) of measurements a correlation coefficient approximating zero may be obtained even though graphs of the statistics show that the up-and-down fluctuations occur together. This result will come about if the *long-time* variations show opposite tendencies, as, for instance, in the statistics of marriages and bank clearings in the United Kingdom. On the other hand, a *high* correlation coefficient may be obtained for two series having the same long-time tendency regardless of the non-correspondence of the short-time fluctuations. For example, the coefficient for the two series, population and bank reserves, came out to be 0.98. This high coefficient comes from the fact that the long-time variation of both series is the same. Consequently, before it is legitimate to draw any conclusions as to the meaning of a lack of correlation, or amount of correlation between two series of measurements it is necessary to ascertain the periodic and the secular variations in the two series. This correlation coefficient may be large through the correspondence of either secular or periodic variation, or both. It may be null because one variation covers up the other.

Three methods have been used for isolating the short-time variations of time-series of measurements. They will now be considered.

1. If upon plotting the two series being compared with time as abscissa and the measurements as ordinates, *periodic* variations appear at approximately equal intervals of time the curve may be "smoothed" and the secular variations may be eliminated as follows:
   (a) Ascertain the length of the wave by finding the number of time units between corresponding parts of the waves, *i. e.,* crest to crest, or hollow to hollow. Let 1 represent the number of time units found.
   (b) Average groups of 1 consecutive measurements, placing the points, determined by these averages at the middle of each group of measurements. Take enough groups so that the points obtained will indicate the general tendency of the series.
   (c) Draw a smooth curve through the points located by the process described in (b). This curve shows the secular tendency.
   (d) Subtract (this can be done graphically on cross-section paper) the ordinates of the "smoothed" curve from those of the original curve in order to obtain the series of measurements of the periodic fluctuation. Let d stand for any one of these differences.
   (e) The coefficients computed for corresponding ordinates of two smoothed curves, and for corresponding differences, d and d', give measures of the secular and periodic correlation, respectively.

   The method described above has been applied by Mr. R. H. Hooker in his paper "On the Correlation of the Marriage-Rate with Trade," and by Mr. G. U. Yule in his study of " Changes in Marriage and Birth-Rates in England and Wales during the Past Half Century. " The following table gives the correlation coefficients computed in the articles named for the *periodic* variation:

Notes

Series	Period	Deviations from	Coefficient of Correlation
{Marriage rate ..... / Imports plus exports per capita .....}	1861–1895	9 yr. means	+ 0.86
{Marriage rate ..... / Amount of bank clearings per capita}	1876–1895	9 yr. means	+ 0.47
{Marriage rate ..... / Sauerbeck's index numbers of prices.}	1865–1896	11yr. means	+ 0.795
{Marriage rate ..... / Hartley's index numbers of unemployment}	1870-1895	11 yr. means	– 0.873

The effect of using the deviations rather than the original series in computing the coefficient is shown by the comparison of the first correlation coefficient of + 0.86, given above, with the correlation coefficient of + 0.18, obtained for the same two series of *original* measurements for the same period, 1861–1895.

Using the deviation-method, Mr. Yule computed the correlation coefficients between *first,* the marriage rate of one year (m), and *second,* exports (e), imports (i), total trade (t), the price of wheat (w), and bank clearings (c) for the same year, and for each of several preceding years in order to answer the question, "does the maximum amount of correlation occur when corresponding items are of same year or when the marriage rate of one year is paired with the business item for a preceding year ?"

Mr. Yule says, "Fitting a parabola to the three values thus determined, a maximum correlation of about 0.482 must subsist between the birth-rate and the marriage-rate of 2.17 (two years and two months) previously."

Further analysis leads Mr. Yule to the conclusion that birth-rate is independently (not through marriage-rate only) sensitive to short-time economic changes and that the birth-rate is lowered after a depression, not only because of a decrease in the number of marriages during such depression, but also to a decrease in fertility.

2. In case the statistics show a long-time tendency with no *regular periodic* fluctuation Mr. R. H. Hooker has suggested that the "differences between successive values of the two variables, instead of the differences from the arithmetic means"* be correlated. Put into mathematical symbols we have:

Letting $\begin{Bmatrix} X_0, X_1 .... X_n \\ X'_0, X'_1 .... X'_n \end{Bmatrix}$ represent two series of measurements, and $\begin{Bmatrix} d_1, d_2 ... d_n \\ d'_1, d'_2 ... d'_n \end{Bmatrix}$ represent differences between any two consecutive measurements, and $\begin{Bmatrix} d_m \\ d'_m \end{Bmatrix}$ represent the respective means of these differences,

then $$d_m = \frac{X_n - X_o}{n} = \frac{\Sigma d}{n}, \text{ and}$$

$$d'_m = \frac{X'_n - X'_o}{n} = \frac{\Sigma d'}{n};$$

and the standard deviations of the differences are

$$\delta = \sqrt{\frac{\Sigma (d - d_m)^2}{n}}$$

Notes

$$\delta' = \sqrt{\frac{\Sigma(d' - d'_m)^2}{n}};$$

and the coefficient of correlation is

$$\rho = \frac{\Sigma(d - d_m)(d' - d'_m)}{n\delta'\delta} = \frac{\Sigma dd' - nd_m d'_m}{\sqrt{\left(\Sigma d^2 - nd_m^2\right)\left(\Sigma d'^2 - nd'^2_m\right)}}$$

Comparing this method of differences with the method described in (1) Mr. Hooker says, "Correlation of the deviations from an instantaneous average (or trend) may be adopted to test the similarity of more or less marked periodic influences, correlation of the difference between successive values will probably prove most useful where the similarity of the shorter rapid changes (with no apparent periodicity) are the subject of investigation, or where the normal level of one or both series of observations does not remain constant." He finds that the ordinary correlation coefficient (r) for the price of corn in Iowa and total production in the United States for the period 1870 – 1899 is – 0.28, while ρ = – 0.84.

Notes *The coefficient of correlation* (r) *is a measure of the closeness of the grouping of the points about this line of regression.* If the points should all range themselves on a line then r would equal + 1 or — 1 depending upon whether, looking left to right, the line sloped upward or downward.

I have computed ρ for the statistics of corn production in the United States and the average farm price on December 1* for the period 1866 – 1906 and finds ρ = – 0.833 ± 0.034. Letting x represent the production *difference* in millions of bushels, and y represent the price *difference* in cents per bushel, the equations of regression are

$$y = -0.0256\,x + 1.132$$

$$x = -27.05\,y + 46.42$$

A graphic representation of the points whose abscissas and ordinates are the corresponding production and price differences, respectively, and the line of regression is given. The lack of correlation between the original pair of series is shown by the chart.

From the equations of regression such statements as the following can be made:

(i) For no change in corn production there is an increase in price of 1.132 cents per bushel.

(ii) For an increase in production of 100 million bushels the price decreases 1.43 cents per bushel.

(iii) For a decrease in production of 100 million bushels the price increases 3.69 cents per bushel.

(iv) For a stationary price the production must increase 46 million bushels per year.

It seemed to me that if *percentage* changes in price and production were used instead of absolute changes a still closer correlation might result. The computation of ρ from such percentages, however, gave – 0.794.

In the preceding illustrations the amount of correlation between the differences was greater than that between the original series. The method of differences has also been used by the writer for Kemmerer's statistics (considered on page 15 of this article) of (1) money in circulation, and (2) bank reserves for the period 1879 – 1904 with the result ρ = + 0.392, whereas the value of r is 0.98. This shows that there is a lack of correspondence of the short-time variations in these two series.

Notes

3. A third method of eliminating the long-time tendency and thus isolating the short-time fluctuations is to assume some curve, represented by an algebraic equation, which "fits" the statistics in question. The first step in the process is to select some curve, which, for *a priori* or other reasons is considered the best representation of the "growth element."* The second step is to fit the curve to the statistics; stated algebraically, to determine the constants in the equation of the curve by use of the actual data. Finally the deviations of the original measurements from the smooth curve (called by Norton "the growth axis") are computed. The accuracy with which one law, the geometric, $y = bc^x$, describes the population of the United States, and consequently many things that depend upon population is shown by the following diagram. The full points are fixed by the actual population according to each of the censuses from 1850. The smooth line is the graph of the equation

$$y = 24{,}086{,}000\ (1.0238)^x,$$

which equation was determined from the actual population.

Prof. J. P. Norton has applied the method here described to determine the correlation existing between percentage of reserves to deposits of New York Associated Banks and call rates.* Weekly statistics were taken for the period 1885 – 1900. The growth axes assumed were the geometric curve, $y = bc^x$, and the straight line $y = a$, respectively. (y = the measurement, x = time measured in weeks, while a, b and c are constants to be determined from the data.) The typical periodic fluctuations of *percentage* deviations of reserves and loans were also correlated by this method, using $y = bc^x$ as the growth function in both cases. The following table gives the correlation coefficients, ρ.

Series	ρ
Reserve deviations and discount rate	-0.37 ± 0.02
(a) Reserve and (b) Loan Periods Immediate.....	$+0.49 \pm 0.07$
(a) precedes (b) by one week	$+0.62 \pm 0.06$
(a) precedes (b) by two weeks	$+0.87 \pm 0.02$
(a) precedes (b) by three weeks	$+0.96 \pm 0.01$
(a) precedes (b) by four weeks	$+0.91 \pm 0.05$

The conclusion from this study is that "the loan period is really the shadow of the reserve period" ... and apparently follows the latter by "an interval of approximately three weeks."

Up to this point the problem before us has been the measurement of the amount of correlation between two variables. This is the simplest case of the general problem of the measurement of the amount of correlation between one series of measurements, and a group of any number of series of measurements. The solution of the general problem leads to very complex relations, § and it will not be taken up here. The case of three variables will be considered briefly.

Messrs. R. H. Hooker and G. U. Yule have considered the problem, To find the relation between the production of wheat in India during the period 1890 – 1904 (years ending March 31), the price of wheat (calendar years), and the exports of the subsequent twelve months, 1891 – 1905 (years ending March 31). The correlation of the annual differences according to the method described in (2) of gives the following results:

Series Correlated	Coefficient Correlation
1. Exports and Production	+ 0.77
2. Exports and Price	+ 0.86
3. Production and Price combined in the ratio 1: 1, and Exports ...	+ 0.90
4. Production and Price combined in the ratio 3: 1, and Exports ...	+ 0.81
5. Production and Price in the ratio 1: 3, and Exports ...	+ 0.58

Notes

The table indicates that exports depend upon production and price, and depend equally upon them.

Messrs. Hooker and Yule give the following general solution of the special problem just considered:

To find the maximum correlation coefficient between x_1 and $x_2 + bx_3$ that results from considering b a variable, where x_1, x_2, and x_3 are the deviations of the series X_1, X_2, and X_3 from their respective arithmetic averages.

Let $x_2 + bx_3 = z$

then $\Sigma(x_1 z) = \Sigma x_1 x_2 + \Sigma b x_1 x_3 = n\left(r_{12}\sigma_1\sigma_2 + br_{13}\sigma_1\sigma_3\right)$

and $\Sigma z^2 = n\left(\sigma_2{}^2 + b^2\sigma_3{}^2 + 2br_{23}\sigma_2\sigma_3\right)$

Hence $\sqrt{x_1 z} = \dfrac{r_{12}\sigma_2 + br_{13}\sigma_2}{\sqrt{\sigma_2{}^2 + b^2\sigma_3{}^2 + 2br_{23}\sigma_2\sigma_3}}$

To find the value of b for which this is a maximum, differentiate with respect to b and equate to zero; then

$$b = \frac{\left(r_{13} - r_{12}r_{23}\right)\sigma_2}{\left(r_{12} - r_{13}r_{23}\right)\sigma_3}$$

which gives the maximum value

$$\sqrt{x_1 z} = \sqrt{\frac{r_{12}{}^2 + r_{13}{}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}{}^2}}$$

Computing $\sqrt{x_1 z}$ from the data of Indian production, price, and exports of wheat the value 0.905 is obtained.

Mr. G. U. Yule, in the paper already referred to,* has worked out the general solution of the problem of the correlation between three variables. In the course of the solution the problem just considered is solved incidentally. The argument is similar to that used in the case of two variables and so it will not be repeated here. A concrete notion of the results secured by Mr. Yule can be obtained from the following explanation taken from Mr. Hooker's article on the "Correlation of the Weather and the Crops."

"I have in the first place formed the ordinary coefficient $r = \dfrac{\Sigma(xy)}{\sqrt{n\sigma_1\sigma_2}}$ between the crop and (a) rainfall, (b) accumulated temperature above 42°. But rainfall and temperature are themselves correlated; hence an apparent influence of, say, rainfall upon a crop may really be due to rainfall conditions being dependent upon temperature, or *vice versa.* Hence it seemed desirable to calculate the *partial* or *net* correlation coefficients, *i.e.* (following the notation given in Mr. Yule's paper of 1897).

$$\rho_{12} = \frac{r_{12} - r_{12}r_{23}}{\sqrt{\left(1 - r_{13}{}^2\right)\left(1 - r_{23}{}^2\right)}}, \quad \rho_{13} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{\left(1 - r_{23}{}^2\right)\left(1 - r_{12}{}^2\right)}}$$

"This partial coefficient (ρ) may be regarded as a truer indication of the connection between the crop and each factor alone, inasmuch as, speaking approximately, we may say that the effect of the other factor is eliminated. It may be observed, moreover, that the relative influence of rainfall and temperature upon the crop is given by $\dfrac{\rho_{12}}{\rho_{13}}$; or, more accurately, this fraction measures the relative

Notes

effect of changes equal in amount to their respective standard deviations in the rainfall and temperature. In discussing the figures in the tables I shall accordingly utilize the partial correlation coefficients rather than the others. Finally, I have worked out what Mr. Yule calls the coefficient of double correlation between the crop and rainfall and accumulated temperature above 42°,

$$R = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\cdot r_{23}\cdot r_{13}}{\left(1 - r_{23}^2\right)}},$$

or as it may also be written,

$$R = \sqrt{1 - \left(1 - r_{12}^2\right)\left(1 - \rho_{13}^2\right)},$$

a form which is quicker to calculate. This may be regarded as a measure of the joint influence of the rainfall and the temperature upon the crop. For the sake of brevity, I shall speak of R as measuring the effect of the 'weather,' using this term in the strictly limited sense of consisting only of these two factors. . . .

"I propose to regard a coefficient between 0.3 and 0.5 as *suggestive* of dependence. Values below 0.3 I shall, as a rule, ignore, in the absence of any corroborative evidence. Perhaps I may remark that I believe that some statisticians would consider themselves justified in drawing deductions from lower coefficients than those I have adopted as my limits."*

Mr. Yule notes that the partial or net correlation coefficient retains three of the chief properties of the ordinary coefficients: " (1) it can only be zero if both net regressions are zero; (2) it is a symmetrical function of the variables (*i.e.*, $\rho_{12} = \rho_{21}$); (3) it cannot be greater than unity."

The various illustrations which have been cited show the importance of questions of correlation in economics. The ordinary graphic method of measuring correlation is inadequate. The coefficient of correlation is simple and yet is sensitive to small changes. It has been used in many fields of statistics by Galton, Pearson, Yule, Hooker, Elderton and others. The experience of these writers warrants the adoption of the coefficient of correlation by economists as one of their standard averages.

Self-Assessment

1. Fill in the blanks:

(*i*) If more than one items is assigned the same rank ..... adjustment is made.

(*ii*) Probable error for coefficient of correlation can be found by the formula ..... .

(*iii*) Coefficient of determination is ..... .

(*iv*) Limits of correlation is ..... to ..... .

(*v*) The relationship between three or more variables is studied with the help of ..... correlation.

9.3 Summary

- Correlation means a relation between two groups. In statistics, it is the measure to indicate the relationship between two variables in which, with changes in the values of one variable, the values of other variable also change. These variables may be related to one item or may not be related to one item but have dependence on the other due to some reason.
- The term correlation indicates the relationship between two variables in which with changes in the value of one variable, the values of the other variable also change. Correlation has been defined by various eminent statisticians, mathematicians and economists.
- Correlation is very useful in understanding the economic behaviour. It helps in locating those variables on which other variables depend. In this way various economic events can be analysed. Moreover, it also helps in identifying the stabilishing factors for a disturbed economic situation.
- Correlation measures a degree of the relationship between two or more variables but it does not indicate any kind of cause and effect relationship between the variables. If, high degree of

Notes

correlation is found exist between two variables, it implies that there must be a reason for such close relationship, but the cause and effect relation can be revealed specifically when other knowledge of the factor involved being brought to bear on the situation. This means, to establish a 'functional relationship' between two or more variables, one has to go beyond the confines of statistical analysis to other factors. (Functional relationship means that two or more factors are interdependent.

- When the values of the two variables move in the same direction, *i.e.,* an increase in one is associated with an increase in other, or *vice versa,* the correlation is said to be positive. If the values of two variables move in the opposite directions *i.e.,* an increase in the value of one variable is associated with fall in other, or *vice versa,* the correlation is said to be negative. For example, the price and supply are positively correlated but price and demand are negatively correlated.
- When relation between two variables is studied, it is simple correlation. When three or more factors are studied together to find relationships, it is called multiple correlation. In partial correlation, two or more factors are agreed to be involved but correlation is studied between only two factors, considering other factors to be constant.
- The cause and effect relation existing between economic events is especially difficult to ascertain because of the presence of innumerable variable elements. In solving his problems the economist can not, like the physicist or chemist, eliminate all causes except one and then by experiment determine the effect of that one. Causes must be dealt with *en masse.* Since any effect is the result of many combined causes the economist is never sure that a given effect will follow a given cause. In stating an economic law he always has to postulate "other things remaining the same," with, perhaps, little appreciation of what the other things may be. It is rarely, if ever, possible for the economist to state more than "such and such a cause *tends* to produce such and such an effect." Events can only be stated to be more or less probable. He is dealing mainly, therefore, with correlation and not with simple causation.
- Just as the biologists cannot predict a man's height or color of eyes or temper or combativeness by knowing those qualities in his ancestors, so economists cannot predict that a definite call rate in Wall Street will go with a given percentage of reserves to deposits in New York banks or that a given supply of wheat will result in a definite price per bushel. But, on the other hand, just as it has been observed that there *is* a relation existing between a man's stature and the stature of his ancestors, so it has been observed that a relation *does* exist between bank reserves and call rates and between supply of wheat and its price per bushel.
- The commonly used method of measuring the amount of correlation between any two series of economic statistics is to represent the two series graphically upon the same sheet of cross-section paper and then compare the fluctuations of one series with those of the other. The quantity theory of prices has been tested in this way by Dr. E. W. Kemmerer. Dr. Kemmerer builds up the following price equation:
- In the case of the correlation of bank reserves and money in circulation, inclusive of bank reserves, Dr. Kemmerer concludes, "There can be no question but that when due allowance is made for fluctuations in business confidence, the evidence of Chart I strongly supports the contention that there exists a close relationship between the amount of money in circulation and the amount of the country's bank reserves."
- The graphic method of comparing fluctuations is well enough as a preliminary, *but does it enable anyone to tell anything of the extent of the correlation between the series of figures being considered*? Is Dr. Kemmerer warranted in deducing his conclusions from observation of the charts ? It seems to the writer that one opposing the quantity theory might draw opposite conclusions with as much (or as little) reason. *The charts do not answer the questions proposed.* The painstaking collection of statistics to test correlation is useless if there be no more reliable method to measure correlation. A numerical measure of the correlation must be found if we wish to determine the *extent* to which the fluctuations of one series synchronize with the fluctuations of another series.

Notes

- Report of the United States Bureau of Labor, "the percentage of successful strikes decreases during periods of business prosperity and increases during 'hard times.' " In the accompanying charts the per cent. of establishments in which strikes were successful is plotted, first, with the per capita exports and imports and second, with index numbers of wholesale prices. The foreign trade and the price statistics are taken as indicative of the activity of business, as indices of prosperity.
- The coefficient of correlation "serves as a measure of any statement involving two qualifying adjectives, which can be measured numerically, such as tall men have tall sons,' 'wet springs bring dry summers,' 'short hours go with high wages.' " It is not the purpose in what follows to go through the mathematical derivation of the coefficient of correlation, but to test the formula empirically in order to ascertain how it actually varies for given series of statistics and to point out some of its features.
- The correlation coefficients show that there is a very great difference in the degree of correlation of different pairs of series of statistics. The full significance of the "probable error," which is used as a measure of unreliability of any determination, cannot be developed at this point. It is sufficient to note that, "When *r* is not greater than its probable error we have no evidence that there is any correlation, for the observed phenomena might easily arise from totally unconnected causes; but, when *r* is greater than, say, six times its probable error, we may be practically certain that the phenomena are not independent of each other, for the chance that the observed results would be obtained from unconnected causes is practically zero."
- The amount of correlation indicated in each case is small—considering the number of years taken, so small that no conclusion as to the connection between the two series can be drawn. The correlation coefficient in the last instance, *i. e.,* between per cent. of successful strikes and business distrust, suggests an opposite conclusion to that indicated by the other coefficients and that of Mr. Cross. The analysis shows that the conclusion that there is negative correlation between *general* prosperity and per cent. of successful strikes is not warranted.
- The coefficient for the two series, population and bank reserves, came out to be 0.98. This high coefficient comes from the fact that the long-time variation of both series is the same. Consequently, before it is legitimate to draw any conclusions as to the meaning of a lack of correlation, or amount of correlation between two series of measurements it is necessary to ascertain the periodic and the secular variations in the two series. This correlation coefficient may be large through the correspondence of either secular or periodic variation, or both. It may be null because one variation covers up the other.
- For a stationary price the production must increase 46 million bushels per year.

  It seemed to me that if *percentage* changes in price and production were used instead of absolute changes a still closer correlation might result. The computation of ρ from such percentages, however, gave – 0.794.
- In the preceding illustrations the amount of correlation between the differences was greater than that between the original series. The method of differences has also been used by the writer for Kemmerer's statistics (considered on page 15 of this article) of (1) money in circulation, and (2) bank reserves for the period 1879 – 1904 with the result ρ = + 0.392, whereas the value of *r* is 0.98. This shows that there is a lack of correspondence of the short-time variations in these two series.
- Mr. G. U. Yule, in the paper already referred to,* has worked out the general solution of the problem of the correlation between three variables. In the course of the solution the problem just considered is solved incidentally. The argument is similar to that used in the case of two variables and so it will not be repeated here. A concrete notion of the results secured by Mr. Yule can be obtained from the following explanation taken from Mr. Hooker's article on the "Correlation of the Weather and the Crops."
- The ordinary graphic method of measuring correlation is inadequate. The coefficient of correlation is simple and yet is sensitive to small changes. It has been used in many fields of statistics by Galton, Pearson, Yule, Hooker, Elderton and others. The experience of these writers

Notes

warrants the adoption of the coefficient of correlation by economists as one of their standard averages.

9.4 Key-Words

1. Correlation : The correlation coefficient a concept from statistics is a measure of how well trends in the predicted values follow trends in past actual values. It is a measure of how well the predicted values from a forecast model "fit" with the real-life data.
2. Galton : An explorer and anthropologist, Francis Galton is known for his pioneering studies of human intelligence. He devoted the latter part of his life to eugenics, i.e. improving the physical and mental makeup of the human species by selected parenthood.

9.5 Review Questions

1. Define correlation. What is its utility ?
2. Explain the meaning of the term 'correlation'. Does it always signify cause and effect relationship?
3. Discuss the various types of correlation.
4. Describe the application of correlation for economists.
5. Explain, how correlation is a powerful statistical tool. Can it be used to establish cause and effect relationship ?

Answers: Self-Assessment

1. (*i*) $\frac{1}{12}\left(m^3 - m\right)$ (*ii*) $\frac{1-r^2}{\sqrt{N}}$ (*iii*) r^2

   (*iv*) r + P.E, to r – P.E (*v*) multiple

9.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 10: Correlation: Scatter Diagram Method, Karl Pearson's Coefficient of Correlation

CONTENTS

Objectives

After reading this unit students will be able to:

- Discuss Scatter Diagram Method.
- Explain Karl Pearson's Coefficient of Correlation.

Introduction

A scatter diagram is used to show the relationship between two kinds of data. It could be the relationship between a cause and an effect, between one cause and another, or even between one cause and two others. To understand how scatter diagrams work, consider the following example.

Suppose you have been working on the process of getting to work within a certain time period. The control chart you constructed on the process shows that, on average, it takes you 25 minutes to get to work. The process is in control. You would like to decrease this average to 20 minutes. What causes in the process affect the time it takes you to get to work ? There are many possible causes, including traffic, the speed you drive, the time you leave for work, weather conditions, etc. Suppose you have decided that the speed you drive is the most important cause. A scatter diagram can help you determine if this is true.

In this case, the scatter diagram would be showing the relationship between a "cause" and an "effect." The cause is the speed you drive and the effect is the time it takes to get to work. You can examine this cause and effect relationship by varying the speed you drive to work and measuring the time it takes to get to work. For example, on one day you might drive 40 *mph* and measure the time it takes to get to work. The next day, you might drive 50 *mph*. After collecting enough data, you can then plot the speed you drive versus the time it takes to get to work. Figure 1 is an example of a scatter diagram for this case. The cause (speed) is on the *x*-axis. The effect (time it takes to get to work) is on the *y*-axis. Each set of points is plotted on the scatter diagram.

In statistics, the Pearson product-moment correlation coefficient (*r*) is a common measure of the correlation between two variables X and Y. When measured in a population the Pearson Product Moment correlation is designated by the Greek letter rho (?). When computed in a sample, it is designated by the letter "*r*" and is sometimes called "Pearson's *r*." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from + 1 to – 1. A correlation of + 1 means that there is a perfect positive linear relationship between variables. A correlation of – 1 means that there is a perfect negative linear relationship between variables. A correlation of 0 means there is

Notes

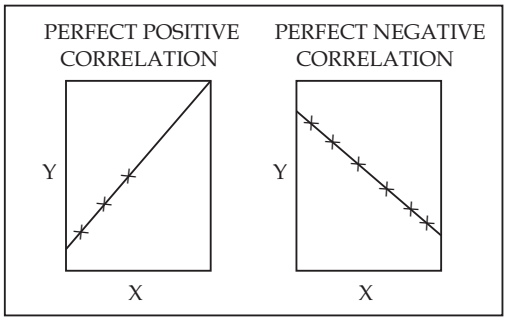
no linear relationship between the two variables. Correlations are rarely if ever 0, 1, or – 1. If you get a certain outcome it could indicate whether correlations were negative or positive.

Mathematical Formula

The quantity *r*, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson.

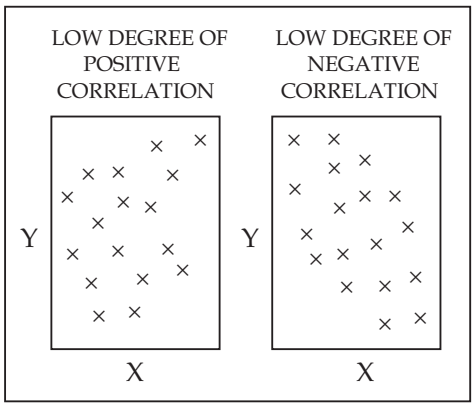
10.1 Scatter Diagram Method

The simplest device for determining relationship between two variables is a special type of dot chart called scatter diagram. When this method is used the given data are plotted on a graph paper in the form of dots, *i.e.,* for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not. The more the plotted points "scatter" over a chart, the less relationship there is between the two variables. The more nearly the points come to falling on a line, the higher the degree of relationship. If all the points lie on a straight line falling from the lower left-hand corner to the upper right corner, correlation is said to be perfectly positive *(i.e., r* = + l) (diagram I).



I **II**

On the other hand, if all the points are lying on a straight line rising from the upper left hand corner to the lower right-hand corner of the diagram correlation is said to be perfectly negative, (*i.e.*, *r* = –1) (diagram II). If the plotted points fall in a narrow band there would be a high degree of correlation between the variables—correlation shall be positive if the points show a rising tendency from the lower left-hand corner to the upper right-hand corner (diagram III) and negative if the points show

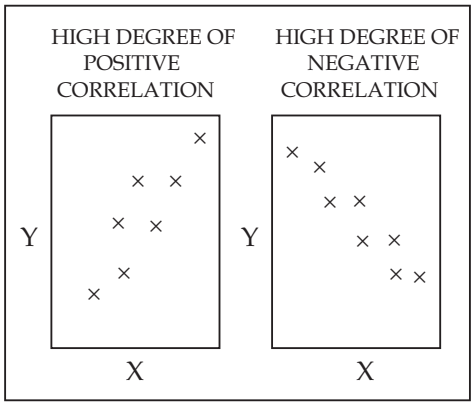


III **IV**

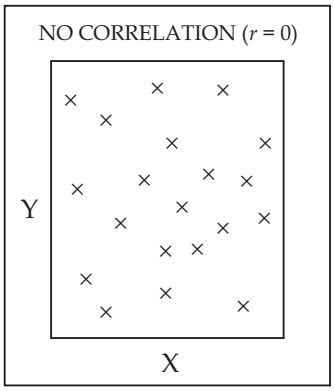
a declining tendency from the upper left-hand corner to the lower right-hand corner of the diagram (diagram IV). On the other hand, if the points are widely scattered over the diagram it is the indication

Notes

of very little relationship between the variables—correlation shall be positive if the points are rising from the lower left-hand corner to the upper right hand corner (diagram V) and negative if the points are running from the upper left-hand side to the lower right-hand side of the diagram (diagram VI). If the plotted points lie on a straight line parallel to the *x*-axis or in a haphazard manner, it shows the absence of any relationship between the variables, *(i.e., r* = 0) as shown by diagram VII.



V **VI**



VII

Example 1: Given the following pairs of value of the variables X and Y*:*

X:	2	3	7	6	8	9
Y*:*	6	5	5	8	12	11

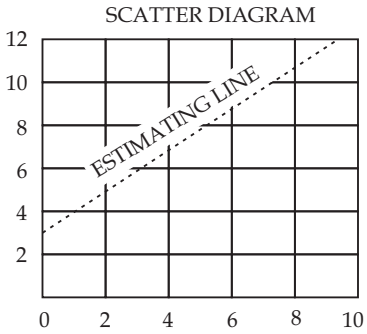
(a) Make a scatter diagram.

(b) Do you think that there is any correlation between the variables X and Y ? Is it positive or negative ? Is it high or low ?

(c) By graphic inspection, draw an estimating line.

Solution: By looking at the following scatter diagram we can say that the variables X and Y are correlated. Further, correlation is positive because the trend of the points is upward rising from the lower left-hand corner to the upper right-hand corner of the diagram. The diagram also indicates that the degree of relationship is high because the plotted points are near to the line which shows perfect relationship between the variables.

Notes



Merits and Limitations of the Method

Merits

1. It is a simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.
2. It is not influenced by the size of extreme items whereas most of the mathematical methods of finding correlation are influenced by extreme items.
3. Making a scatter diagram usually is the first step in investigating the relationship between two variables.

Limitations

By applying this method we can get an idea about the direction of correlation and also whether it is high or low. But we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical methods.

10.2 Karl Pearson's Coefficient of Correlation

Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely used in practice. The Pearsonian coefficient of correlation is denoted by the symbol *r*. It is one of the very few symbols that is used universally for describing the degree of correlation between two series. The formula for computing Pearsonian *r* is:

$$r = \frac{\sum xy}{N\sigma_x\sigma_y} \quad \text{... (i)}$$

Hence $x = \left(X - \overline{X}\right)$, $y = \left(Y - \overline{Y}\right)$

σ_x = Standard deviation of series X

σ_y = Standard deviation of series Y

N = Number of paired observations.

This method is to be applied only when the deviations of items are taken from *actual* means and *not* from assumed means.

The value of the coefficient of correlation as obtained by the above formula shall always lie between ± 1. When *r* = + 1, it means there is perfect positive correlation between the variables. When *r* = – 1, it means there is perfect negative correlation between the variables. When *r* = 0, it means there is no relationship between the two variables. However, in practice, such values of *r* as + 1, – 1 and 0 are rare. We normally get values which lie between + 1 and – 1 such as + .1, – .4, etc. The coefficient of

Notes

correlation describes not only the magnitude of correlation but also its direction. Thus, + .8 would mean that correlation is positive because the signs of r is + and the magnitude of correlation is .8.

The above formula for computing Pearsonian coefficient of correlation can be transformed in the following form which is easier to apply:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} \quad \text{... (ii)}$$

where $x = \left(X - \bar{X}\right)$ and $y = \left(Y - \bar{Y}\right)$.

It is obvious that while applying this formula we have not to calculate separately the standard deviation of X and Y series as is necessary while applying formula (i). This simplifies greatly the task of calculating correlation coefficient.

Steps

(i) Take the deviation of X series from the mean of X and denote the deviations by x.

(ii) Square these deviations and obtain the total, *i.e.*, $\sum x^2$.

(iii) Take the deviations of Y series from the mean of Y and denote these deviations by y.

(iv) Square these deviations and obtain the total, *i.e.*, $\sum y^2$.

(v) Multiply the deviation of X and Y series and obtain the total, *i.e.*, $\sum xy$.

(vi) Substitute the values of $\sum xy$, $\sum x^2$ and $\sum y^2$ in the above formula.

The following examples will illustrate the procedure:

Example 2: Calculate Karl Pearson's coefficient of correlation from the following data:

X:	6	8	12	15	18	20	24	28	31
Y:	10	12	15	15	18	25	22	26	28

Solution:

Calculation of Karl Pearson's Correlation Coefficient

X	(X – 18) x	x^2	Y	(Y – 19) y	y^2	xy
6	– 12	144	10	– 9	81	+ 108
8	– 10	100	12	– 7	49	+ 70
12	– 6	36	15	– 4	16	+ 24
15	– 3	9	15	– 4	16	+ 12
18	0	0	18	– 1	1	0
20	+ 2	4	25	+ 6	36	+ 12
24	+ 6	36	22	+ 3	9	+ 18
28	+ 10	100	26	+ 7	49	+ 70
31	+ 13	169	28	+ 9	81	+ 117
$\sum X = 162$	$\sum x = 0$	$\sum x^2 = 598$	$\sum Y = 171$	$\sum y = 0$	$\sum y^2 = 338$	$\sum xy = 431$

Notes

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$\Sigma xy = 431, \Sigma x^2 = 598, \Sigma y^2 = 338$$

$$r = \frac{431}{\sqrt{598 \times 338}} = \frac{431}{449.582} = +0.959.$$

Example 3: Find coefficient of correlation for the following:

Cost (Rs.)	39	65	62	90	82	75	25	98	36	78
Sales (Rs.)	47	53	58	86	62	68	60	91	51	84

Solution: **Calculation of Karl Peason's Correlation Coefficient**

X	**(X – 65)** x	x^2	**Y**	**(Y – 66)** y	y^2	xy
39	– 26	676	47	– 19	361	+ 494
65	0	0	53	– 13	169	0
62	– 3	9	58	– 8	64	+ 24
90	+ 25	625	86	+ 20	400	+ 500
82	+ 17	289	62	– 4	16	– 68
75	+ 10	100	68	+ 2	4	+ 20
25	– 40	1600	60	– 6	36	+ 240
98	+ 33	1089	91	+ 25	625	+ 825
36	– 29	841	51	– 15	225	+ 435
78	+ 13	169	84	+ 18	324	+ 234
$\Sigma X = 650$	$\Sigma x = 0$	$\Sigma x^2 = 5398$	$\Sigma Y = 660$	$\Sigma y = 0$	$\Sigma y^2 = 2224$	$\Sigma xy = 2704$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{650}{10} = 65, \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{660}{10} = 66$$

Since actual means of x and y are whole numbers, apply the actual mean method of finding correlation

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$r = \frac{2704}{\sqrt{5398 \times 2224}} = \frac{2704}{3464.85} = +0.78$$

We can also solve the question with the help of Logarithms. This method is easy where calculators are not allowed.

$$r = \frac{2704}{\sqrt{5398 \times 2224}}$$

Notes

$$\log r = \log 2704 - \frac{1}{2}\ [\log 5398 + \log 2224]$$

$$= 3.4320 - \frac{1}{2}\ [3.7322 + 3.3472]$$

$$= 3.4320 - 3.5397 = \bar{1}.8923 = 0.78$$

Thus the answer is the same.

Example 4: Making use of the data summarised below, calculate the coefficient of correlation, r_{12}:

Case	X_1	X_2	Case	X_1	X_2
A	10	9	E	12	11
B	6	4	F	13	13
C	9	6	G	11	8
D	10	9	H	9	4

Solution: **Calculation of Coefficient of Correlation**

Case	X_1	$(X_1 - \bar{X}_1) x_1$	x_1^2	X_2	$(X_2 - \bar{X}_2) x_2$	x_2^2	$x_1 x_2$
A	10	0	0	9	+ 1	1	0
B	6	– 4	16	4	– 4	16	16
C	9	– 1	1	6	+ 2	1	2
D	10	0	0	9	+ 1	1	0
E	12	+ 2	4	11	+ 3	+ 3	6
F	13	+ 3	9	13	+ 5	25	15
G	11	+ 1	1	8	0	0	0
H	9	– 1	1	4	– 4	16	4
N = 8	$\Sigma X_1 = 80$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 32$	$\Sigma X_2 = 64$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 72$	$\Sigma x_1 x_2 = 43$

$$\bar{X}_1 = \frac{\Sigma X_1}{N} = \frac{80}{8} = 10; \qquad \bar{X}_2 = \frac{\Sigma X_2}{N} = \frac{64}{8} = 8.$$

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \times \Sigma x_2^2}}$$

$$\Sigma x_1 x_2 = 43, \Sigma x_1^2 = 32, \Sigma x_2^2 = 72.$$

Substituting the values

$$r_{12} = \frac{43}{\sqrt{32 \times 72}} = \frac{43}{\sqrt{2304}} = \frac{43}{48} = 0.896.$$

Note: It should be noted that the above formula is the same as given earlier, *i.e.*,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

Notes

The only difference is that of the symbols. Since in this question we were given series $\bar{X}_1$ and $\bar{X}_2$ we changed the symbols in the formula accordingly.

Calculation of Correlation Coefficient when Change of Scale and Origin is made

Since *r* is a pure number, shifting the origin and changing the scale of series do not affect its value.

Example 5: Find the coefficient of correlation from the following data:

X :	300	350	400	450	500	550	600	650	700
Y :	800	900	1000	1200	1300	1400	1500	1600	

Solution: In order to simplify calculations, let us divide each value of the variable X by 50 and each value of variable Y by 100.

CALCULATION OF CORRELATION COEFFICIENT

X	$\frac{X}{50}$ X_1	$(X_1 - \bar{X}_1)$ $\bar{X}_1 = 10$ x	x^2	Y	$\frac{Y}{100}$ Y_1	$(Y_1 - \bar{Y}_1)$ $\bar{Y}_1 = 12$ y	y^2	xy
300	6	– 4	16	800	8	– 4	16	16
350	7	– 3	9	900	9	– 3	9	9
400	8	– 2	4	1,000	10	– 2	4	4
450	9	– 1	1	1,100	11	– 1	1	1
500	10	0	0	1,200	12	0	0	0
550	11	+ 1	1	1,300	13	+ 1	1	1
600	12	+ 2	4	1,400	14	+ 2	4	4
650	13	+ 3	9	1,500	15	+ 3	9	9
700	14	+ 4	16	1,600	16	+ 4	16	16
	$\Sigma X_1 = 90$	$\Sigma x = 0$	$\Sigma x^2 = 60$		$\Sigma Y_1 = 108$	$\Sigma y = 0$	$\Sigma y^2 = 60$	$\Sigma xy = 60$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$\Sigma xy = 60, \Sigma x^2 = 60, \Sigma y^2 = 60$$

$$r = \frac{60}{\sqrt{60 \times 60}} = \frac{60}{60} = 1.$$

When Deviations are taken from an Assumed Mean

When actual means are in fractions, say the actual means of X and Y series are 20.167 and 29.23, the calculation of correlation by the method discussed above would involve too many calculations and would take a lot of time. In such cases we make use of the assumed mean method for finding out correlation. When deviations are taken from an assumed mean the following formula is applicable:

Notes

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_X)^2}{N}}\sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

where d_x refers to deviations of X series from an assumed mean, *i.e.*, (X – A); d_y refers to deviations of Y series from an assumed mean, *i.e.*, (Y – A); $\Sigma d_x d_y$ = sum of the product of the deviations of X and Y series from their assumed means; Σd_x^2 = sum of the squares of the deviations of X series from an assumed means; Σd_y^2 = sum of the squares of the deviations of Y series from an assumed mean; Σd_x = sum of the deviations of X series from an assumed mean; Σd_y = sum of the deviations of Y series from an assumed mean.

It may be pointed out that there are many variations of the above formula. For example, the above formula may be written as:

$$r = \frac{N\Sigma d_x d_y - \{(\Sigma d_x)(\Sigma d_y)\}}{\sqrt{N\ \Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\ \Sigma d_y^2 - (\Sigma d_y)^2}}$$

But the form given above is the easiest to apply.

Note: While applying assumed mean method, any value can be taken as the assumed mean and the answer will be the same. However, the nearer the assumed mean to the actual mean, the lesser will be the calculations.

Steps

(i) Take the deviations of X series from an assumed mean, denote these deviations by d_x and obtain the total, *i.e.*, Σd_x.

(ii) Take the deviations of Y series from an assumed mean, denote these deviations by d_y and obtain the total, *i.e.*, Σd_y.

(iii) Square d_x and obtain the total Σd_x^2.

(iv) Square d_y and obtain the total Σd_y^2.

(v) Multiply d_x and d_y and obtain the total $\Sigma d_x d_y$.

(vi) Substitute the values of $\Sigma d_x d_y$, Σd_x, Σd_y, Σd_x^2 and Σd_y^2 in the formula given above.

The following examples shall illustrate the procedure:

Notes

Example 6: Compute Karl Pearson's correlation coefficient for the data given below:

X :	45	55	56	58	60	65	68	70	75	80	85
Y :	56	50	48	60	62	64	65	70	74	82	90

Solution: Since means of X and Y are in fractions we will apply the assumed mean method of calculating correlation. Taking 65 as the assumed mean in case of X and 66 in case of Y:

Calculation of Coefficient of Correlation

X	(X – 65) d_x	d_x^2	Y	(Y – 66) d_y	d_y^2	$d_x d_y$
45	– 20	400	56	– 10	100	+ 200
55	– 10	100	50	– 16	256	+ 160
56	– 9	81	48	– 18	324	+ 162
58	– 7	49	60	– 6	36	+ 42
60	– 5	25	62	– 4	16	+ 20
65	0	0	64	– 2	4	0
68	+ 3	9	65	– 1	1	– 3
70	+ 5	25	70	+ 4	16	+ 20
75	+ 10	100	74	+ 8	64	+ 80
80	+ 15	225	82	+ 16	256	+ 240
85	+ 20	400	90	+ 24	576	+ 480
$\Sigma X = 717$	$\Sigma d_x = +2$	$\Sigma d_x^2 = 1414$	$\Sigma Y = 721$	$\Sigma d_y = -5$	$\Sigma d_y^2 = 1649$	$\Sigma d_x d_y = 1401$

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}\sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$\Sigma d_x d_y = 1401, \Sigma d_x = +2, \Sigma d_y = -5, \Sigma d_x^2 = 1414, \Sigma d_y^2 = 1649$, N = 11

$$r = \frac{1401 - \frac{(2)(-5)}{11}}{\sqrt{1414 - \frac{(2)^2}{11}}\sqrt{1649 - \frac{(-5)^2}{11}}}$$

$$= \frac{1401.91}{\sqrt{1414 - .364}\sqrt{1649 - 2.273}}$$

$$= \frac{1401.91}{\sqrt{1413.636}\sqrt{1646.727}} = \frac{1401.91}{37.598 \times 40.5799} = \frac{1401.91}{1525.723} = +0.919$$

Note: We can simplify considerably the calculation by using logarithms.

Notes

Let $$r = \frac{1401.91}{\sqrt{1413.636}\sqrt{1646.727}}$$

$$\log r = \log 1401.91 - \frac{1}{2}[\log 1413.636 + \log 1646.727]$$

$$= 3.1467 - \frac{1}{2}[3.1504 + 3.2167]$$

$$= 3.1467 + \frac{1}{2}[6.3671] = 3.1467 - 3.1836 = 1.9631$$

r = AL 1.9631 = + 0.919.

Example 7: The following table gives the distribution of the total population and those who are wholly or partially blind among them. Find out if there is any relation between age and blindness.

Age	No. of persons (in thousands)	Blind
0—10	100	55
10—20	60	40
20—30	40	40
30—40	36	40
40—50	24	36
50—60	11	22
60—70	6	18
70—80	3	15

Solution: For facilitating comparison we must determine the number of blinds in terms of a common denominator, say 1 lakh. The first figure would remain as it is because 55 persons are blind out of 100 thousand, *i.e.*, 1 lakh. The second value would be obtained like this.

Out of 60,000 persons number of blinds = 40

Out of 1,00,000 persons number of blinds = $\frac{40}{60,000} \times 1,00,000 = 67$

and so on.

Age	Mid-points X	(X – 35)/10 d_x	d_x^2	Blind persons per lakh Y	(Y – 185) d_y	d_y^2	$d_x d_y$
0—10	5	– 3	9	55	– 130	16,900	+ 390
10—20	15	– 2	4	67	– 118	13,924	+ 236
20—30	25	– 1	1	100	– 85	7,225	+ 85
30—40	35	0	0	111	– 74	5,476	0
40—50	45	+ 1	1	150	– 35	1,225	– 35
50—60	55	+ 2	4	200	+ 15	225	+ 30

Notes

60—70	65	+ 3	9	300	+ 115	13,225	+ 345
70—80	75	+ 4	16	500	+ 315	99,225	+ 1,260
N = 8		Σd_x	Σd_x^2		Σd_y	Σd_y^2	$\Sigma d_x\, d_y$
		= 4	= 44		= +3	= 1,57,425	= 2,311

$$r = \frac{\Sigma d_x\, d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}\sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$N = 8,\ \Sigma d_x\, d_y = 2,311,\ \Sigma d_x = 4,\ \Sigma d_y = 3,\ \Sigma d_x^2 = 44,$

$\Sigma d_y^2 = 1,57,425$

Substituting these values

$$r = \frac{2311 - (4)(3)}{\sqrt{44 - \frac{(4)^2}{8}}\sqrt{157425 - \frac{(3)^2}{8}}} = \frac{2309.5}{\sqrt{42}\sqrt{157423.88}}$$

$$= \frac{2309.5}{6.4807 \times 396.77} = \frac{2309.5}{2571.34} = 0.898.$$

Correlation of Grouped Data

When the number of observations of X and Y variables is large, the data are often classified into two-way frequency distribution called a correlation table. The class intervals for Y are listed in the captions or column headings, and those for X are listed in the stubs at the left of the table (the order can also be reversed). The frequencies for each cell of the table are determined by either tallying or sorting just as in the case of a frequency distribution of a single variable.

The formula for calculating the coefficient of correlation is:

$$r = \frac{\Sigma fd_x\, d_y - \frac{(\Sigma fd_x)(\Sigma fd_y)}{N}}{\sqrt{\Sigma fd_x^2 - \frac{(\Sigma fd_x)^2}{N}}\sqrt{\Sigma fd_y^2 - \frac{(\Sigma fd_y)^2}{N}}}$$

Note: The formula is the same as the one discussed above for assumed mean. The only difference is that here the deviations are also multiplied by the frequencies.

Steps

(i) Take the step deviations of variable X and denote these deviations by d_x.

(ii) Take the step deviations of the variable Y and denote these deviations by d_y.

(iii) Multiply $d_x\, d_y$ and the respective frequency of each cell and write the figure obtained in the right hand upper corner of each cell.

(iv) Add together all the cornered values as calculated in step (iii) and obtain the total $\Sigma f_x d_y$.

(v) Multiply the frequencies of the variable X by the deviations of X and obtain the total Σfd_x.

Notes

(vi) Take the squares of the deviations of the variable X and multiply them by the respective frequencies and obtain Σfd_x^2.

(vii) Multiply the frequencies of the variable Y by the deviations of Y and obtain the total Σfd_y.

(viii) Take the squares of the deviations of the variable Y and multiply them by the respective frequencies and obtain Σfd_y^2.

(ix) Substitute the values of $\Sigma fd_x d_y$, Σfd_x, Σfd_x^2, Σfd_y and Σfd_y^2 in the above formula and obtain the value of *r*.

Example 8: Calculate Karl Pearson's coefficient of correlation and its probable error between the ages of 100 mothers and daughters from the following data:

Age of mothers in years	Age of daughters in years					
	5–10	10–15	15–20	20–25	25–30	Total
15–25	6	3	–	–	–	9
25–35	3	16	10	–	–	29
35–45	–	10	15	7	–	32
45–55	–	–	7	10	4	21
55–65	–	–	–	4	5	9
Total	**9**	**29**	**32**	**21**	**9**	**100**

Solution: Let age of daughters be denoted by X and that of mothers by Y.

Calculation of Coefficient of Correlation

Y \ X		*m*	5–10 7.5	10–15 12.5	15–20 17.5	20–25 22.5	25–30 27.5				
		d_y \ d_x	– 2	– 1	0	1	2	*f*	fd_y	fd_y^2	fd_xd_y
15–25	*m* 20	– 2	6 |24	3 | 6	—	—	—	9	– 18	36	20
25–35	30	– 1	3 | 6	16 |16	10 | 0	—	—	29	– 29	29	22
35–45	40	0	—	10 | 0	15 | 0	7 | 0	—	32	0	0	0
45–55	50	1	—		7 | 0	10 |10	4 | 8	21	21	21	18
55–65	60	2	—			4 | 8	5 |20	9	18	36	28
Total		*f*	9	29	32	21	9	N = 100	Σfd_y = – 8	Σfd_y^2 = 122	Σfd_xd_y = 98
		fd_x	– 18	– 29	0	21	18	Σfd_x = – 8			
		fd_x^2	36	29	0	21	36	Σfd_x^2 = 122			
		fd_xd_y	30	22	0	18	28	Σfd_xd_y = 98			

Notes

$$r = \frac{\Sigma fd_x\, d_y - \frac{(\Sigma fd_x)(\Sigma fd_y)}{N}}{\sqrt{\Sigma fdx^2 - \frac{(\Sigma fdx)^2}{N}}\sqrt{\Sigma fdy^2 - \frac{(\Sigma fd_y)^2}{N}}}$$

$\Sigma fd_x\, d_y = 98, \Sigma fd_x = -8, \Sigma fd_y = -8, \Sigma fd_x^2 = 122, \Sigma fd_y^2 = 122, N = 100$

Substituting the values in the above formula

$$r = \frac{98 - \frac{(-8)(-8)}{100}}{\sqrt{122 - \frac{(-8)^2}{100}}\sqrt{122 - \frac{(-8)^2}{100}}}$$

$$= \frac{98 - .64}{\sqrt{121.36}\,\sqrt{121.36}} = \frac{97.36}{121.36} = +0.802$$

P.E. $$r = 0.6745\frac{1 - r^2}{\sqrt{N}}$$

$r = +0.802, N = 100$

$\therefore$ $$P.E._r = 0.6745\frac{1 - (.802)^2}{\sqrt{100}} = 0.6745\frac{1 - 0.6432}{10}$$

$$= 0.6745 \times 0.03568 = 0.024.$$

Assumptions of the Pearsonian Coefficient

Karl Pearson's coefficient of correlation is based on the following assumptions:

1. There is linear relationship between the variables, *i.e.,* when the two variables are plotted on a scatter diagram straight line will be formed by the points so plotted.
2. The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc., are affected by such forces that a normal distribution is formed.
3. There is a cause-and-effect relationship between the forces affecting the distribution of the items in the two series. If such a relationship is not formed between the variables, *i.e.,* if the variables are independent, there cannot be any correlation. For example, there is no relationship between income and height because the forces that affect these variables are not common.

Merits and Limitations of the Pearsonian Coefficient

Amongst the mathematical methods used for measuring the degree of relationship, Karl Pearson's method is most popular. The correlation coefficient summarises in one figure not only the degree of correlation but also the direction, *i.e.,* whether correlation is positive or negative.

However, the utility of this coefficient depends in part on a wide knowledge of the meaning of this 'yardstick', together with its limitations. The chief *limitations* of the method are:

1. The correlation coefficient always assumes linear relationship regardless of the fact whether that assumption is correct or not.
2. Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted.

Notes

3. The value of the coefficient is unduly affected by the extreme items.
4. As compared with some other methods this method is more time consuming.

Interpreting the Coefficient of Correlation

The coefficient of correlation measures the degree of relationship between two sets of figures. As the reliability estimate depends upon the closeness of the relationship, it is imperative that utmost care is taken while interpreting the value of coefficient of correlation, otherwise fallacious conclusion may be drawn.

Unfortunately, the interpretation of the coefficient of correlation depends very much on experience. The full significance of *r* will only be grasped after working out a number of correlation problems and seeing the kinds of data that give rise to various values of *r*. The investigator must know his data thoroughly in order to avoid errors of interpretation and emphasis. He must be familiar, or become familiar, with all the relationships and theory which bear upon the data and should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. However, the following general rules are given which would help in interpreting the value of *r*.

1. When $r = +1$ it means there is perfect positive relationship between the variables.
2. When $r = -1$ it means there is perfect negative relationship between the variables.
3. When $r = 0$ it means that there is no relationship between the variables, *i.e.,* the variables are uncorrelated.
4. The closer *r* is to + 1 or – 1, the closer the relationship between the variables and the closer *r* is to 0, the less close the relationship. Beyond this it is not safe to go. The full interpretation of *r* depends upon circumstances one of which is the size of the sample. All that can really be said is that when estimating the value of one variable from the value of another, the higher the value of *r* the better the estimate.
5. The closeness of the relationship is not proportional to *r*. If the value of *r* is 0.8 it does not indicate a relationship twice as close as one of 0.4. It is in fact very much closer.

Coefficient of Correlation and Probable Error

The probable error of the coefficient of correlation helps in interpreting its value. With the help of probable error it is possible to determine the reliability of the value of the coefficient in so far as it depends on the conditions of random sampling. The probable error of the coefficient of correlation is obtained as follows:

$$\text{P.E.} = 0.6745\,\frac{1-r^2}{\sqrt{N}}$$

where *r* is the coefficient of correlation and N the number of pairs of items.

1. If the value of *r* is less than the probable error there is no evidence of correlation, *i.e.,* the value of *r* is not at all significant.
2. If the value of *r* is more than six times the probable error, the existence of correlation is practically certain, *i.e.,* the value of *r* is significant.
3. By adding and subtracting the value of probable error from the coefficient of correlation we get respectively the upper and lower limits within which coefficient of correlation in the population can be expected to lie. Symbolically,

$$\rho = r \pm \text{P.E.}$$

ρ (rho) denotes correlation in the population.

Carrying out the computation of the probable error, assuming a coefficient of correlation of 0.80 computed from a sample of 16 pairs of items, we have

Notes

$$\text{P.E.}_r = 0.6745\frac{1-.8^2}{\sqrt{16}} = .06$$

The limits of the correlation in the population would be $r \pm \text{P.E.}$, *i.e.*, .8 ± .06 or .74—.86.

Instances are quite common wherein a correlation coefficient of 0.5 or even 0.4 has been considered to be a fairly high degree of correlation by a writer or research worker. Yet a correlation coefficient of 0.5 means that only 25 per cent of the variation is explained. A correlation coefficient of 0.4 means that only 16 per cent of the variations is explained.

Conditions for the Use of Probable Error

The measure of probable error can be properly used only when the following three conditions exist:

1. The data must approximate a normal frequency curve (bell-shaped curve).
2. The statistical measure for which the P.E. is computed must have been calculated from a sample.
3. The sample must have been selected in an unbiased manner and the individual items must be independent.

   However, these conditions are generally not satisfied and as such the reliability of the correlation coefficient is determined largely on the basis of exterior tests of reasonableness which are often of a statistical character.

Example 9: If r = 0.6 and N = 64, find out the probable error of the coefficient of correlation and determine the limits for population r.

Solution:

$$\text{P.E.}_r = 0.6745\frac{1-r^2}{\sqrt{N}}$$

r = 0.6 and N = 64

$$\text{P.E.}_r = 0.6745\frac{1-(.6)^2}{\sqrt{64}} = \frac{0.6745\times 0.64}{8} = 0.054$$

Limits of population correlation

$$= 0.6 \pm 0.054 = 0.546—0.654.$$

Coefficient of Determination

One very convenient and useful way of interpreting the value of coefficient of correlation between two variables is to use the square of coefficient of correlation, which is called coefficient of determination. The coefficient of determination thus equals r^2. The coefficient r^2 expresses the proportion of the variance in y determined by x; that is, the ratio of the explained variance to total variance. Therefore, the coefficient of determination expresses the proportion of the total variation that has been 'explained', or the relative reduction in variance when measured about the regression equation rather than about the mean of the dependent variable. If the value of r = 0.9, r^2 will be 0.81 and this would mean that 81 per cent of the variation in the dependent variable has been explained by the independent variable. The maximum value of r^2 is unity because it is possible to explain all of the variation in Y, but it is not possible to explain more than all of it.

It is much easier to understand the meaning of r^2 than r and, therefore, the coefficient of determination is to be preferred in presenting the results of correlation analysis. Tuttle has beautifully pointed out that "the coefficient of correlation has been grossly overrated and is used entirely too much. Its square, the coefficient of determination, is a much more useful measure of the linear covariation of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before coming to any conclusion about the extent of the linear relationship between the two correlated variables."

Notes

The relationship between r and r^2 may be noted—as the value of r decreases from its maximum value of 1, the value of r^2 decreases much more rapidly. r will, of course, always be larger than r^2, unless $r^2 = 0$ or 1.

r	r^2
0.90	0.81
0.80	0.64
0.70	0.49
0.60	0.36
0.50	0.25

Thus the coefficient of correlation is 0.707 when just half the variance in Y is due to X.

It should be clearly noted that the fact that a correlation between two variables has a value of $r = 0.60$ and the correlation between two other variables has a value of $r = 0.30$ does not demonstrate that the first correlation is twice as strong as the second. The relationship between the two given values of r can better be understood by computing the value of r^2. When $r = 0.6$, $r^2 = 0.36$ and when $r = 0.30$, $r^2 = 0.09$.

The coefficient of determination is a highly useful measure. However, it is often misinterpreted. The term itself may be misleading in that it implies that the variable X stands in a determining or causal relationship to the variable Y. The statistical evidence itself never establishes the existence of such causality. All that statistical evidence can do is to define covariation, that term being used in a perfectly neutral sense. Whether causality is present or not, and which way it runs if it is present, must be determined on the basis of evidence other than the quantitative observations.

Properties of the Coefficient of Correlation

The following are the important properties of the correlation coefficient r:

1. The coefficient of correlation lies between – 1 and + 1. Symbolically, $-1 \le r \le +1$ or $|r| \le 1$.
2. The coefficient of correlation is independent of change of scale and origin of the variables X and Y.
3. The coefficient of correlation is the geometric mean of two regression coefficients.

   Symbolically,

$$r = \sqrt{b_{xy} \times b_{yx}}$$

Self-Assessment

1. Indicate whether the following statements are True or False:

(*i*) There are no limits to the value of r.

(*ii*) If r is negative both the variable are decreasing.

(*iii*) If the values of X variable are 1, 2, 3, 4, 5 and those of Y 4, 6, 8, 10, 12 the Karl Pearson and the Rank method would give the same answer.

(*iv*) Pearsonian coefficient is the best under all situations.

(*v*) Karl Pearson's coefficient of correlation always lies between 0 and + 1.

10.3 Summary

- A scatter diagram is used to show the relationship between two kinds of data. It could be the relationship between a cause and an effect, between one cause and another, or even between one cause and two others.
- In statistics, the Pearson product-moment correlation coefficient (r) is a common measure of the correlation between two variables X and Y. When measured in a population the Pearson Product

Notes

Moment correlation is designated by the Greek letter rho (?). When computed in a sample, it is designated by the letter "*r*" and is sometimes called "Pearson's *r*." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from + 1 to – 1. A correlation of + 1 means that there is a perfect positive linear relationship between variables. A correlation of – 1 means that there is a perfect negative linear relationship between variables. A correlation of 0 means there is no linear relationship between the two variables. Correlations are rarely if ever 0, 1, or – 1. If you get a certain outcome it could indicate whether correlations were negative or positive.

- The simplest device for determining relationship between two variables is a special type of dot chart called scatter diagram. When this method is used the given data are plotted on a graph paper in the form of dots, *i.e.,* for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not. The more the plotted points "scatter" over a chart, the less relationship there is between the two variables. The more nearly the points come to falling on a line, the higher the degree of relationship. If all the points lie on a straight line falling from the lower left-hand corner to the upper right corner, correlation is said to be perfectly positive *(i.e., r* = + l) (diagram I).
- It is a simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.
- Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely used in practice. The Pearsonian coefficient of correlation is denoted by the symbol *r*. It is one of the very few symbols that is used universally for describing the degree of correlation between two series.
- When the number of observations of X and Y variables is large, the data are often classified into two-way frequency distribution called a correlation table. The class intervals for Y are listed in the captions or column headings, and those for X are listed in the stubs at the left of the table (the order can also be reversed). The frequencies for each cell of the table are determined by either tallying or sorting just as in the case of a frequency distribution of a single variable.
- The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc., are affected by such forces that a normal distribution is formed.
- There is a cause-and-effect relationship between the forces affecting the distribution of the items in the two series. If such a relationship is not formed between the variables, *i.e.,* if the variables are independent, there cannot be any correlation. For example, there is no relationship between income and height because the forces that affect these variables are not common.
- Amongst the mathematical methods used for measuring the degree of relationship, Karl Pearson's method is most popular. The correlation coefficient summarises in one figure not only the degree of correlation but also the direction, *i.e.,* whether correlation is positive or negative.
- The coefficient of correlation measures the degree of relationship between two sets of figures. As the reliability estimate depends upon the closeness of the relationship, it is imperative that utmost care is taken while interpreting the value of coefficient of correlation, otherwise fallacious conclusion may be drawn.
- The probable error of the coefficient of correlation helps in interpreting its value. With the help of probable error it is possible to determine the reliability of the value of the coefficient in so far as it depends on the conditions of random sampling.
- One very convenient and useful way of interpreting the value of coefficient of correlation between

Notes

two variables is to use the square of coefficient of correlation, which is called coefficient of determination. The coefficient of determination thus equals r^2. The coefficient r^2 expresses the proportion of the variance in y determined by x; that is, the ratio of the explained variance to total variance. Therefore, the coefficient of determination expresses the proportion of the total variation that has been 'explained', or the relative reduction in variance when measured about the regression equation rather than about the mean of the dependent variable. If the value of r = 0.9, r^2 will be 0.81 and this would mean that 81 per cent of the variation in the dependent variable has been explained by the independent variable. The maximum value of r^2 is unity because it is possible to explain all of the variation in Y, but it is not possible to explain more than all of it.

- The coefficient of determination is a highly useful measure. However, it is often misinterpreted. The term itself may be misleading in that it implies that the variable X stands in a determining or causal relationship to the variable Y. The statistical evidence itself never establishes the existence of such causality. All that statistical evidence can do is to define covariation, that term being used in a perfectly neutral sense. Whether causality is present or not, and which way it runs if it is present, must be determined on the basis of evidence other than the quantitative observations.

10.4 Key-Words

1. Scatter Diagram : A scatter diagram is a tool for analyzing relationships between two variables. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis. The pattern of their intersecting points can graphically show relationship patterns. Most often a scatter diagram is used to prove or disprove cause-and-effect relationships. While the diagram shows relationships, it does not by itself prove that one variable causes the other. In addition to showing possible cause-and-effect relationships, a scatter diagram can show that two variables are from a common cause that is unknown or that one variable can be used as a surrogate for the other.
2. Coefficient determination : In statistics, the coefficient of determination, denoted R2, is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. R2 is most often seen as a number between 0 and 1.0, used to describe how well a regression line fits a set of data. An R2 near 1.0 indicates that a regression line fits the data well, while an R2 closer to 0 indicates a regression line does not fit the data very well. It is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model.

10.5 Review Questions

1. What is Scatter diagram? How do you interprat a Scatter diagram?
2. What is a 'Scatter diagram'? How does it help us in studying the correlation between two variables in respect of both its nature and extent ?
3. How does a scatter diagram help in ascertaining the degree of correlation between two variables?
4. State the properties of Pearson's coefficient of correlation. How do you interpret a calculated value of r ? Explain the term 'Probable error of r'.
5. State the assumptions of Karl Pearson's Correlation.

Notes

Answers: Self-Assessment

1. (*i*) F (*ii*) F (*iii*) T (*iv*) F (*v*) F

10.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 11: Rank Correlation Method

CONTENTS

Objectives

After reading this unit students will be able to:

- Explain Rank Correlation Method.
- Know the Merits and Limitations of Rank Method.

Introduction

When a group of individuals are arranged according to their degree of possession of a character, they are said to be ranked. The ordinal number of an individual in the arrangement is called its *rank* the arrangement as a whole is called a *ranking*.

When there are two series of ranks for the same set of individuals, corresponding to two different characters or two judges assigning ranks for the same character, one may be interested to know if the two series are associated. The association between two series of ranks for the same set of individuals is called *rank correlation.*

11.1 Rank Correlation Method

This method of finding out covariability or the lack of it between two variables was developed by the British psychologist Charles Edward Spearman in 1904. This measure is especially useful when quantitative measures for certain factors (such as in the evaluation of leadership ability or the judgment of female beauty) cannot be fixed, but the individuals in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank in the group. In any event, the rank correlation coefficient is applied to a set of ordinal rank numbers, with 1 for the individual ranked first in quantity, or quality, and so on, to *n* for the individual ranked last in a group of *n* individuals (or *n* pairs of individuals). Spearman's rank correlation coefficient is defined as:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

where R denotes rank coefficient of correlation and D refers to the difference of ranks between paired items in two series.

Did u know? The association between two series of ranks for the same set of individuals is called *rank correlation.*

Notes

The value of this coefficient also ranges between + 1 and – 1. When R is + 1 there is complete agreement in the order of the rank and the ranks are in the same direction. When R is – 1 there is complete agreement in the order of the ranks and they are in opposite directions.

In rank correlation we may have two types of problems:

A. Where actual ranks are given.

B. Where ranks are not given.

A. Where Actual Ranks are Given

Where actual ranks are given to us the steps required for computing rank correlation are:

(i) Take the differences of the two ranks, *i.e.*, $(R_1 - R_2)$ and denote these differences by D.

(ii) Square these differences and obtain the total $\sum D^2$.

(iii) Apply the formula:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

Example 1 : Two judges in a beauty competition rank the 12 entries as follows:

X :	1	2	3	4	5	6	7	8	9	10	11	12
Y:	12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between the judgment of the two judges ?

Solution:

CALCULATION OF RANK CORRELATION COEFFICIENT

X R_1	**Y** R_2	$(R_1 - R_2)$ **D**	D^2
1	12	– 11	121
2	9	– 7	49
3	6	– 3	9
4	10	– 6	36
5	3	+ 2	4
6	5	+ 1	1
7	4	+ 3	9
8	7	+ 1	1
9	8	+ 1	1
10	2	+ 8	64
11	11	0	0
12	1	+ 11	121
			$\sum D^2 = 416$

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

$\sum D^2 = 416$, N = 12

$$R = 1 - \frac{6 \times 416}{12^3 - 12} = 1 - \frac{2496}{1716} = 1 - 1.454 = - 0.454.$$

Notes

Example 2: Ten competitors in a beauty contest are ranked by three judges in the following order:

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	9	1	6	9
3rd Judge	6	4	9	8	1	6	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty.

Solution: In order to find out which pair of judges has the nearest approach to common tastes in beauty we compare Rank Correlation between the judgments of:

(i) 1st Judge and 2nd Judge; (ii) 2nd Judge and 3rd Judge; (iii) 1st Judge and 3rd Judge.

COMPUTATION OF RANK CORRELATION

Rank by 1st Judge R_1	**Rank by 2nd Judge R_2**	**Rank by 3rd Judge R_3**	**$(R_1 - R_2)^2$ D^2**	**$(R_2 - R_3)^2$ D^2**	**$(R_1 - R_3)^2$ D^2**
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
N = 10	N = 10	N = 10	$\sum D^2 = 200$	$\sum D^2 = 214$	$\sum D^2 = 60$

Rank correlation between the judgments of 1st and 2nd Judge:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

$\sum D^2 = 200$, N = 10

Here we have directly calculated D^2 because D's are not required in applying formula.

$$\therefore \quad R = 1 - \frac{6 \times 200}{10^3 - 10}$$

(I and II)

$$= 1 - \frac{1200}{990} = 1 - 1.212 = -0.212$$

Rank correlation between the judgments of 2nd and 3rd Judge:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

Notes

(II and III)

$$\sum D^2 = 214, N = 10$$

$$= 1 - \frac{6 \times 214}{10^3 - 10} = 1 - \frac{1284}{990} = 1 - 1.297 = -0.297.$$

Rank correlation between the judgments of the 1st and 3rd Judge:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

(I and III)

$$\sum D^2 = 60, N = 10$$

$$= 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{360}{990} = 1 - .364 = +0.636$$

Thus we find the first and third judges have the nearest approach to common tastes in beauty.

B. Where Ranks are not Given ?

When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value we must follow the same method in case of both the variables.

Example 3: (a) Calculate Spearman's coefficient of rank correlation for the following data:

X :	53	98	95	81	75	61	59	55
Y :	47	25	32	37	30	40	39	45

Solution:

Calculation of Rank Correlation Coefficient

X	**R_1**	**Y**	**R_2**	**$(R_1 - R_2)^2$ D^2**
53	1	47	8	49
98	8	25	1	49
95	7	32	3	16
81	6	37	4	4
75	5	30	2	9
61	4	40	6	4
59	3	39	5	4
55	2	45	7	25
				$\sum D^2 = 160$

$$R = 1 - \frac{6\sum D^2}{N^3 - N}; \sum D^2 = 160, N = 8$$

$$R = 1 - \frac{6 \times 160}{8^3 - 8} = 1 - \frac{960}{504} = 1 - 1.905 = -0.905.$$

Notes

Example 4: (b) Find the rank correlation coefficient for the following distribution:

Marks in Statistics	48	60	72	62	56	40	39	52	30
Marks in Accountancy	62	78	65	70	38	54	60	32	31

Solution: We first rank the given data:

Calculation of Rank Correlation Coefficient

Marks in Statistics	R_1	Marks is Accountancy	R_2	$(R_1 - R_2)^2$ D^2
48	4	62	6	4
60	7	78	9	4
72	9	65	7	4
62	8	70	8	0
56	6	38	3	9
40	3	54	4	1
39	2	60	5	9
52	5	32	2	9
30	1	31	1	0
				$\sum D^2 = 40$

$$R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6 \times 40}{9^3 - 9} = 1 - \frac{240}{720} = +0.667.$$

Equal Ranks

In some cases it may be found necessary to rank two or more individuals or entries as equal. In such a case it is customary to give each individual an average rank. Thus if two individuals are ranked equal at fifth place, they are each given the rank $\frac{5+6}{2}$ that is 5.5 while if three are ranked equal at fifth place they are given the rank $\frac{5+6+7}{3}$ = 6. In other words, where two or more individuals are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of the ranks which these individuals would have not got had they differed even slightly from each other.

Where equal ranks are assigned to some entries an adjustment in the above formula for calculating the rank coefficient of correlation is made.

The adjustment consists of adding $\frac{1}{12}\left(m^3 - m\right)$ to the value of $\sum D^2$, where *m* stands for the number of items whose ranks are common. If there are more than one such group of items with common rank, this value is added as many times as the number of such groups. The formula can thus be written:

$$R = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + ...\right\}}{N^3 - N}$$

Notes

Example 5: (a) Calculate the coefficient of rank correlation from the following data:

X :	48	33	40	9	16	16	65	24	16	57
Y :	13	13	24	6	15	4	20	9	6	19

Solution: **Calculation of Rank Correlation Coefficient**

X	R_1	Y	R_2	$(R_1 - R_2)^2$ D^2
48	8	13	5.5	6.25
33	6	13	5.5	0.25
40	7	24	10	9.00
9	1	6	2.5	2.25
16	3	15	7	16.00
16	3	4	1	4.00
65	10	20	9	1.00
24	5	9	4	1.00
16	3	6	2.5	0.25
57	9	19	8	1.00
				$\sum D^2 = 41$

$$R = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right)\right\}}{N^3 - N}$$

Since item 6 is repeated 3 times in series X, $m = 3$. Since items 13 is repeated twice and 6 is repeated twice in case of series Y, m shall be 2 in each case. Hence:

$$R = 1 - \frac{6\left\{41 + \frac{1}{12}\left(3^3 - 3\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right)\right\}}{10^3 - 10}$$

$$= 1 - \frac{6(41 + 2 + 0.5 + 0.5)}{990} = 1 - \frac{264}{990} = 1 - 0.267 = 0.733.$$

Example 6: (b) Calculate the rank coefficient of correlation of the following data:

X :	80	78	75	75	68	67	60	59
Y :	12	13	14	14	14	16	15	17

Solution:

Calculation of Rank Correlation

X	R_x	Y	R_y	$(R_x - R_y)^2$ D^2
80	8	12	1	49.00
78	7	13	2	25.00
75	5.5	14	4	2.25

Notes

75	5.5	14	4	2.25
68	4	14	4	0.00
67	3	16	7	16.00
60	2	15	6	16.00
59	1	17	8	49.00
				$\sum D^2 = 159.5$

$$R = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right)\right\}}{N^3 - N}$$

$$= 1 - \frac{6\left\{159.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)\right\}}{8^3 - 8}$$

$$= 1 - \frac{6\{159.5 + 5 + 2\}}{504}$$

$= 1 - 1.929 = - 0.929.$

11.2 Merits and Limitations of the Rank Method

Merits:

1. This method is simpler to understand and easier to apply compared to the Karl Pearson's method. The answer obtained by this method and the Karl Pearson's method will be the same provided no value is repeated, *i.e.*, all the items are different.
2. Where the data is of a qualitative nature like honesty, efficiency, intelligence, etc., this method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and degree of correlation established by applying this method.
3. This is the only method that can be used where we are given the ranks and not the actual data.
4. Even where actual data are given, rank method can be applied for ascertaining degree of correlation.

Limitations:

1. This method cannot be used for finding out correlation in a grouped frequency distribution.
2. Where the number of items exceeds 30 the calculations become quite tedious and require a lot of time. Therefore, this method should not be applied where N is exceeding 30 unless we are given the ranks and not actual values of the variable.

When to use Rank Correlation Coefficient

The rank method has two principal uses:

(1) The initial data are in the form of ranks.

(2) If N is fairly small (say, not large than 25 or 30), rank method is sometimes applied to interval data as an approximation to the more time-consuming *r*. This requires that the interval data be transferred to rank orders for both variables. If N is much in excess of 30, the labour required in ranking the scores becomes greater than is justified by the anticipated saving of time through the rank formula.

Notes

Self-Assessment

1. Indicate whether the following statements are True or False:

(*i*) Rank method can be used for finding correlation coefficient even when actual data is given.

(*ii*) If two items are to be assigned equal ranks, rank method of correlation coefficient cannot be used.

(*iii*) The rank correlation coefficient was developed by spearman.

(*iv*) Rank correlation coefficient is obtained by the formula:

$R = 1 - 6\sum D^2 / N^3 - N$.

(*v*) If difference of ranks in each pair is zero, show that the rank correlation coefifcient is+ 1.

11.3 Summary

- When there are two series of ranks for the same set of individuals, corresponding to two different characters or two judges assigning ranks for the same character, one may be interested to know if the two series are associated. The association between two series of ranks for the same set of individuals is called *rank correlation.*
- This method of finding out covariability or the lack of it between two variables was developed by the British psychologist Charles Edward Spearman in 1904. This measure is especially useful when quantitative measures for certain factors (such as in the evaluation of leadership ability or the judgment of female beauty) cannot be fixed, but the individuals in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank in the group.
- The value of this coefficient also ranges between + 1 and – 1. When R is + 1 there is complete agreement in the order of the rank and the ranks are in the same direction. When R is – 1 there is complete agreement in the order of the ranks and they are in opposite directions.
- When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value we must follow the same method in case of both the variables.
- If two individuals are ranked equal at fifth place, they are each given the rank $\frac{5+6}{2}$ that is 5.5 while if three are ranked equal at fifth place they are given the rank $\frac{5+6+7}{3}$ = 6. In other words, where two or more individuals are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of the ranks which these individuals would have not got had they differed even slightly from each other.
- Where the data is of a qualitative nature like honesty, efficiency, intelligence, etc., this method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and degree of correlation established by applying this method.
- If N is fairly small (say, not large than 25 or 30), rank method is sometimes applied to interval data as an approximation to the more time-consuming *r*. This requires that the interval data be transferred to rank orders for both variables. If N is much in excess of 30, the labour required in ranking the scores becomes greater than is justified by the anticipated saving of time through the rank formula.

Notes

11.4 Key-Words

1. Rank correlation : In statistics, a rank correlation is the relationship between different rankings of the same set of items. A rank correlation coefficient measures the degree of similarity between two rankings, and can be used to assess its significance.

11.5 Review Questions

1. Define rank correlation. Derive formula for correlation coefficient.
2. If the difference of rank in each pair is zero, show that the rank correlation coefficient is + 1.
3. State the merits and demerits of rank correlation method.
4. Give the features of spearman's coefficient of correlation.

Answers: Self-Assessment

1. (*i*) T (*ii*) F (*iii*) T (*iv*) F (*v*) T

11.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 12 : Linear Regression Analysis : Introduction and Lines of Regression

CONTENTS

Objectives

After reading this unit students will be able to :

- Introduce Linear Regression Analysis
- Discuss Line of Regression.

Introduction

In the previous unit we discussed correlation analysis, which seeks to determine the degree of linear relationship or correlation between two variables in a bivariate distribution. The coefficient of correlation indicates whether the variables are linearly related and if so, how strong the relationship is. In scatter diagram method of determining correlation, we observed that if $r = \pm 1$, then all the points lie exactly on a straight line showing a linear relationship between the two variables. Also for high positive or negative value of correlation coefficient, we observed that the point in a scatter diagram lie near about a straight line. In case $r = 0$, the scatter of points is considerable and the linear trend disappears. Such observations give rise to questions : what is the straight line in the scatter diagram, how can this line be obtained and finally what is the usefulness of this line. *Statistical methods used to answer such questions are the subject matter of* regression analysis. *The* regression analysis *is concerned with the formulation and determination of algebraic expressions for the relationship between the two variables. We use the general form 'regression lines' for these algebraic expressions. These regression lines or the exact algebraic forms of the relationship are then used for predicting the value of one variable from that of the other. Here, the variable whose value is to be predicted is called* **dependent** or **explained variable** *and the variable used for prediction is called* **independent** or **explanatory variable.**

Did u know? **Galton** *termed the line describing the average relationship between the two variables as the line of regression.*

The word regression, which means reversion, was first introduced by **Sir Francis Galton** in the study of heredity. *His study on the heights of fathers and sons revealed an interested relationship. He showed that the heights of sons tended or reverted towards the average rather than two extreme values, i.e., tall fathers tend*

Notes

to have tall sons and short fathers short sons, but the average height of the sons of a group of short fathers is greater than that of fathers. Thus, by regression we mean the average relationship between two or more variables which can be used for estimating the value of one variable from the given values of one or more variables. However, in a bivariate distribution, the analysis is restricted to only two variables only.

12.1 Introduction to Linear Regression Analysis

The study of regression has special importance in statistical analysis. We know that the mutual relationship between two series is measured with the help of correlation. Under correlation, the direction and magnitude of the relationship between two variables is measured. But it is not possible to make the best estimate of the value of a dependent variable on the basis of the given value of the independent variable by correlation analysis. Therefore, to make the best estimates and future estimation, the study of regression analysis is very important and useful.

Meaning and Definition

According to **Oxford English Dictionary,** the word 'regression' means "Stepping back" or "Returning to average value". The term was first of all used by a famous Biological Scientist in 19th century, **Sir Francis Galton** relating to a study of hereditary characteristics. He found out an interesting result by making a study of the height of about one thousand fathers and sons. His conclusion was that (i) Sons of tall fathers tend to be tall and sons of short fathers tend to be short in height (ii) But mean height of the tall fathers is greater than the mean height of the sons, whereas mean height of the short sons is greater than the mean height of the short fathers. The tendency of the entire mankind to **twin** back to average height, was termed by **Galton** 'Regression towards Mediocricity' and the line that shows such type of trend was named as 'Regression Line'.

In statistical analysis, the term 'Regression' is taken in wider sense. **Regression is the study of the nature of relationship between the variables so that one may be able to predict the unknown value of one variable for a known value of another variable.** In regression, one variable is considered as an independent variable and another variable is taken as dependent variable. With the help of regression, possible values of the dependent variable are estimated on the basis of the values of the independent variable. For example, there exists a functional relationship between demand and price, *i.e.*, D = *f* (P). Here, demand (D) is a dependent variable, and price (P) is an independent variable. On the basis of this relationship between demand and price, probable values of demand can be estimated corresponding to the different values of price.

Definition of Regression

Some important definitions of regression are as follows :

1. Regression is the measure of the average relationship between two or more variables.

   — M.M. Blair

2. Regression analysis measures the nature and extent of the relation between two or more variables, thus enables us to make predictions. **— Hirsch**

In brief, regression is a statistical method of studying the nature of relationship between two variables and to make prediction.

Utility of Regression

The study of regression is very useful and important in statistical analysis, which is clear by the following points :

(1) Nature of Relationship : Regression analysis explains the nature of relationship between two variables.

(2) Estimation of Relationship : The mutual relationship between two or more variables can be measured easily by regression analysis.

(3) Prediction : By regression analysis, the value of a dependent variable can be predicated on the basis of the value of an independent variable. For example, if price of a commodity rises, what will be the probable fall in demand, this can be predicted by regression.

Notes

(4) Useful in Economic and Business Research : Regression analysis is very useful in business and economic research. With the help of regression, business and economic policies can be formulated.

12.2 Line of Regression

If the variables in a bivariate frequency distribution are correlated, we observe that the points in a scatter diagram cluster around a straight line called the **line of regression.** In a bivariate study, we have **two lines of regression,** namely :

1. Regression of Y on X.
2. Regression of X on Y.

Regression of Y on X

The line of regression of Y on X *is used to predict or estimate the value of* Y *for the given value of the variable* X. Thus, Y is the dependent variable and X is an independent variable in this case. The *algebraic form of the line* **line of regression of Y on X** is of the form :

$$Y = a + bX \quad \ldots (1)$$

where, *a* and *b* are unknown constants to be determined by observed data on the two variables X and Y. Let (X_1, Y_1), (X_2, Y_2)..., (X_N, Y_N) be N pairs of observations on the variable X and Y. *Then, for determining a and b in equation* (1) *we make use of the following* **normal equations :**

$$\Sigma Y = Na + b\Sigma X \quad \ldots (2)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \ldots (3)$$

The values ΣY, ΣX, ΣX^2 and ΣXY can be obtained from the given data.

These normal equations are obtained by minimising the error sum of squares according to the principle of least squares. Solving equations (2) and (3) for *a* and *b,* the line of regression of Y on X is completely determined.

Alternatively

There is another way of finding the algebraic form of line of the regression of Y on X. Line of regression of Y on X *can also be written in the following form* :

$$\left(Y - \overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right) \quad \ldots (4)$$

or

$$\left(Y - \overline{Y}\right) = b_{YX}\left(X - \overline{X}\right) \quad \ldots (5)$$

Here,

$\overline{Y}$ = the mean of Y

$\overline{X}$ = the mean of X

σ_Y = the S.D. of Y

σ_X = the S.D. of X

r = the correlation coefficient between X and Y

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$ = the regression coefficient of Y on X

From observed bivariate data $[(X_i, Y_i); i = 1, 2, \ldots N]$ the regression coefficient of Y on X, b_{YX}, can be computed from any of the following formula :

Notes

$$b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \quad \ldots (6)$$

or

$$b_{YX} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2} \quad \ldots (7)$$

Here, $d_x = (X - A)$ and $d_y = (Y - B)$ are deviations from assumed means A and B respectively.

Regression of X and Y

The line of regression of X and Y *is used to estimate or predict the value of X for a given value of the variable Y*. In this case X is the dependent variable and Y is the independent variable. The standard algebraic form of the line of regression of X on Y is :

$$X = c + dY \quad \ldots (8)$$

where c and d are unknown constant which are determined from the following two normal equations:

$$\Sigma X = Nc + d\Sigma Y \quad \ldots (9)$$

$$\Sigma XY = c\Sigma Y + d\Sigma Y^2 \quad \ldots (10)$$

The values of ΣX, ΣY, ΣXY and ΣY^2 can be obtained from observed data. The normal equations (9) and (10) are also obtained by minimising error sum of squares according to the method of least squares. Solving (9) and (10) for c and d and putting these values in (8), the form of regression of X on Y is completely determined.

Alternatively

Like regression of Y on X, the line of regression of X on Y also has an alternative form as

$$(X - \bar{X}) = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}) \quad \ldots (11)$$

or

$$(X - \bar{X}) = b_{XY}(Y - \bar{Y}) \quad \ldots (12)$$

Here, $b_{XY} = r\frac{\sigma_X}{\sigma_Y}$ is called the regression coefficient of X on Y. For observed data, the value of b_{XY} can be computed from any of the two formulae :

$$b_{XY} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2} \quad \ldots (13)$$

or

$$b_{XY} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_y^2 - (\Sigma d_y)^2} \quad \ldots (14)$$

The determination of both lines of regression will be clear from the following examples :

Example 1: From the following data, obtain the two regression equations.

X	6	2	10	4	8
Y	9	11	5	8	7

Solution: First we obtain the regression lines by using the method of least squares.

Notes

Table showing calculations

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$	$\Sigma XY = 214$

Regression of X and Y :

Let line of regression of Y on X be

$$Y = a + bX \quad \text{... (i)}$$

The normal equations giving the values of a and b are

$$\Sigma Y = Na + b\Sigma X \quad \text{... (ii)}$$

$$\Sigma XY = a\Sigma X + \Sigma bX^2 \quad \text{... (iii)}$$

Putting the values from the table, one gets

$$40 = 5a + 30b \quad \text{... (iv)}$$

$$214 = 30a + 220b \quad \text{... (v)}$$

Multiplying equation (iv) by 6, we get

$$240 = 360a + 180b \quad \text{... (vi)}$$

Subtracting (v) from (vi), we have

$$26 = -40b \therefore b = -0.65$$

Thus, from (iv), one gets

$$40 = 5a - 30 \times 0.65 \text{ or } 5a = 40 + 19.5 \therefore a = 11.9$$

Putting the values of a and b in (i), the regression of Y on X becomes

$$Y = 11.9 - 0.65\,X \text{ or } Y + 0.65\,X = 11.9$$

Regression of X on Y :

Let the line of regression of X on Y be

$$X = c + dY \quad \text{... (vii)}$$

The normal equations giving the value of c and d are

$$\Sigma X = Nc + d\Sigma Y \quad \text{... (viii)}$$

$$\Sigma XY = c\Sigma Y + d\Sigma Y^2 \quad \text{... (ix)}$$

Putting the values in (viii) and (ix), one gets

$$30 = 5c + 40d \quad \text{... (x)}$$

$$214 = 40c + 340d \quad \text{... (xi)}$$

Multiplying equation (x) by 8, we have

$$240 = 40c + 320d \quad \text{... (xii)}$$

Subtracting equation (xii) from equation (xi), one gets

Notes

$-26 = 20d \quad \therefore \quad d = -1.3$

Putting $d = -1.3$ in equation (x),

$30 = 5c + 40 \times (-1.3)$ or $5c = 30 + 52 = 82$

$\therefore \quad c = 16.4$

Putting the values of c and d in (vii), the line of regression of X on Y is :

$X = 16.4 - 1.3\,Y$

$\therefore \quad X + 1.3\,Y = 16.4$

Example 2: Obtain regression lines for the data in example 1 by computing regression coefficients.

Solution:

Regression of Y on X :

Computation of regression coefficients

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$	$\Sigma XY = 214$

The line of regression of Y on X using its regression coefficient can be written as :

$$\left(Y - \overline{Y}\right) = b_{YX}\left(X - \overline{X}\right) \quad \text{... (i)}$$

Here,
$$\overline{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

and
$$b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)}$$

$$= \frac{5 \times 214 - 30 \times 40}{5 \times 220 - (30)^2}$$

$$= \frac{1070 - 1200}{1100 - 900} = \frac{-130}{200} = -0.65$$

Putting the values of $\overline{X}$, $\overline{Y}$ and b_{YX} in equation (i), one gets the line of regression of Y on X as :

$(Y - 8) = -0.65\,(X - 6)$ or $Y = 8 - 0.65\,X + 3.9$

or $Y + 0.65\,X = 11.9$

which is the same as obtained in example 1.

Notes

Regression of X on Y :

The line of regression of X on Y using its regression coefficient is :

$$\left(X - \overline{X}\right) = b_{YX}\left(Y - \overline{Y}\right) \quad \text{... (ii)}$$

Here,
$$b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{5 \times 214 - 30 \times 40}{5 \times 340 - (40)^2} = \frac{1070 - 1200}{1700 - 1600} = -\frac{130}{100} = -1.30$$

Putting the value of $\overline{X}$, $\overline{Y}$ and b_{YX} in equation (ii), the line of regression of X and Y becomes

$$(X - 6) = -1.30\ (Y - 8) \text{ or } X = 6 - 1.30\ Y + 10.4$$

or
$$X + 1.30\ Y = 16.4$$

which is also the same as obtained in example 1.

Remark

The two expression lines can be obtained by any of the above two methods if the values of the pairs of observations are not very large. *However, when the data are large, short cut method for computing regression coefficient should be applied to avoid huge calculations.* Calculations are further reduced if the deviations of the two variables are taken from their respective means, *i.e.,* when Σd_x and Σd_y are zero. *In this case, the regression coefficients* b_{YX} *and* b_{XY} *are obtained by using formula* (7) *and* (14). The example will on the next page clarify the point.

Example 3: On the basis of following data, obtain regressions of (i) Y on X and (ii) X on Y.

X	15	27	27	30	38	46
Y	12	15	15	18	22	26

Solution:

Calculation for regression equations

X	d_x **= (X – 30)**	d_x^2	**Y**	d_y **= (Y – 18)**	d_y^2	$d_x d_y$
15	– 15	225	12	– 6	36	90
27	– 3	9	15	– 3	9	9
27	– 3	9	15	– 3	9	9
30	0	0	18	0	0	0
38	8	64	22	4	16	32
46	16	256	26	8	64	128
N = 6	$\Sigma d_x = 3$	$\Sigma d_x^2 = 563$		$\Sigma d_y = 0$	$\Sigma d_y^2 = 134$	$\Sigma d_x d_y = 228$

Thus,
$$\overline{X} = A + \frac{\Sigma dX}{N}\ ; \qquad \overline{Y} = B + \frac{\Sigma dY}{N}$$

Notes

$$= 30 + \frac{3}{6} = 30.5 \quad = 18 + \frac{0}{6} = 18.0$$

$$b_{YX} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{6 \times 268 - 3 \times 0}{6 \times 563 - (3)^2} = \frac{1608}{3378 - 9} = \frac{1608}{3369} = 0.4773$$

$$b_{YX} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$$= \frac{6 \times 268 - 3 \times 0}{6 \times 134 - (0)^2} = \frac{1608}{804} = 2.00$$

Therefore, the regression of Y on X is :

$$(Y - \overline{Y}) = b_{YX}(X - \overline{X}) \text{ or } (Y - 18) = 0.4773\ (X - 30.5)$$

or $Y = 18 + 0.4773X - 14.5577$

or $Y = 0.4773X + 3.4424$

and regression of X on Y is :

$$(X - \overline{X}) = b_{YX}(Y - \overline{Y}) \text{ or } (X - 30.5) = 2.0\ (Y - 18)$$

or $X = 30.5 + 2.0Y - 36.0$

or $X = 2.0Y - 5.50$

Example 4: In the estimation of regression equations of two variables X and Y, the following results were obtained :

$$\overline{X} = 20,\ \overline{Y} = 30,\ N = 10,\ \Sigma X^2 = 6360,\ \Sigma Y^2 = 9860,\ \Sigma XY = 5900$$

obtain the two equations.

Solution: Given that : $\overline{X} = 20,\ \overline{Y} = 30,\ N = 10,\ \Sigma X^2 = 6360,\ \Sigma Y^2 = 9860,\ \Sigma XY = 5900$

$\therefore \quad \Sigma X = N\overline{X} = 10 \times 20 = 200;\ \Sigma Y = 10 \times 30 = 300$

$\therefore \quad b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$

$$= \frac{10 \times 5900 - 200 \times 300}{10 \times 6360 - (200)^2} = \frac{59000 - 60000}{63600 - 40000}$$

$$= \frac{-1000}{23600} = -0.042$$

and $b_{XY} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$

Notes

$$= \frac{10 \times 5900 - 60000}{10 \times 9860 - (300)^2} = \frac{-1000}{8600} = -0.116$$

Regression of Y on X

$$(Y - \bar{Y}) = b_{YX}(X - \bar{X})$$

or $(Y - 30) = -0.042\,(X - 20)$ or $Y = 30 - 0.042X + 0.84$

or $Y + 0.042X = 30.84$

Regression of X on Y

$$(X - \bar{X}) = b_{XY}(Y - \bar{Y})$$

or $(X - 20) = -0.116\,(Y - 30)$ or $X = 20 - 0.116Y + 3.48$

or $X + 0.116Y = 23.48$

Did u know? Statistical methods used to answer such questions are the subject matter of regression analysis. *The* regression analysis *is concerned with the formulation and determination of algebraic expressions for the relationship between the two variables.*

Properties of the Regression Lines

1. The regression lines of Y on X is used to estimate or predict the best value (in least squares sense) of Y for a given value of the variable X. Here Y is dependent and X is an independent variable.
2. The regression line of X on Y is used to estimate to best value of X for a given value of the variable Y. Here X is dependent and Y is an independent variable.
3. The two lines of regression cut each other at the points $(\bar{X}, \bar{Y})$. *Thus, on solving the two lines of regression, we get the values of means of the variables in the bivariate distribution.*
4. In a bivariate study, there are two lines of regression. *However, in case of perfect correlation that is when r* = + 1 on – 1 *we have only one regression line as both the regression lines coincide in this case.*
5. When *r* = 0, *i.e.*, if correlation exists between X and Y, the two lines of regression become perpendicular to each other.

Self-Assessment

1. Which of the following statements is true or false :

(*i*) The term 'regression' was first used by Karl Pearson in the year 1900.

(*ii*) Regression analysis reveals average relationship between two variables.

(*iii*) The regression line cut each other at the point of average of X and Y.

(*iv*) In regression analysis b_{xy} stand for regression coefficient of X on Y.

(*v*) The regression line of Y on X minimises total of the squares of the vertical deviations.

12.3 Summary

- *We use the general form 'regression lines' for these algebraic expressions. These regression lines or the exact algebraic forms of the relationship are then used for predicting the value of one variable from that of the other. Here, the variable whose value is to be predicted is called* **dependent** or **explained variable** *and the variable used for prediction is called* **independent** or **explanatory variable.**

Notes

- The word regression, which means reversion, was first introduced by **Sir Francis Galton** in the study of heredity. *His study on the heights of fathers and sons revealed an interested relationship. He showed that the heights of sons tended or reverted towards the average rather than two extreme values, i.e., tall fathers tend to have tall sons and short fathers short sons, but the average height of the sons of a group of short fathers is greater than that of fathers.* **Galton** *termed the line describing the average relationship between the two variables as the line of regression. Thus, by regression we mean the average relationship between two or more variables which can be used for estimating the value of one variable from the given values of one or more variables.*
- Under correlation, the direction and magnitude of the relationship between two variables is measured. But it is not possible to make the best estimate of the value of a dependent variable on the basis of the given value of the independent variable by correlation analysis. Therefore, to make the best estimates and future estimation, the study of regression analysis is very important and useful.
- In statistical analysis, the term 'Regression' is taken in wider sense. **Regression is the study of the nature of relationship between the variables so that one may be able to predict the unknown value of one variable for a known value of another variable.** In regression, one variable is considered as an independent variable and another variable is taken as dependent variable. With the help of regression, possible values of the dependent variable are estimated on the basis of the values of the independent variable. For example, there exists a functional relationship between demand and price, *i.e.*, D = *f* (P). Here, demand (D) is a dependent variable, and price (P) is an independent variable. On the basis of this relationship between demand and price, probable values of demand can be estimated corresponding to the different values of price.
- By regression analysis, the value of a dependent variable can be predicated on the basis of the value of an independent variable. For example, if price of a commodity rises, what will be the probable fall in demand, this can be predicted by regression.
- If the variables in a bivariate frequency distribution are correlated, we observe that the points in a scatter diagram cluster around a straight line called the **line of regression.**
- The line of regression of X and Y *is used to estimate or predict the value of X for a given value of the variable Y*. In this case X is the dependent variable and Y is the independent variable.

12.4 Key-Words

1. Linear Regression Analysis : In statistics, regression analysis is a statistical technique for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables - that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function, which can be described by a probability distribution.

Notes

12.5 Review Questions

1. What is meant by linear regression analysis ? Explain clearly the significance of linear regression analysis.
2. What are regression line ? Explain their uses.
3. Discuss the introduction of linear regression analysis.
4. What are the properties of regression line.
5. Explain the concept of lines of regression. State why there are two lines of regression.

Answers: Self-Assessment

1. (*i*) F (*ii*) T (*iii*) T (*iv*) F (*v*) T

12.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Unit 13: Coefficient of Simple Regression Method

Notes

CONTENTS

Objectives

After reading this unit students will be able to:

- Explain Regression Equations
- Discuss the coefficients of Simple Regression Method.

Introduction

After having established the fact that two variables are closely related, we may be interested in estimating (predicting) the value of one variable given the value of another. For example, if we know that advertising and sales are correlated, we may find out the expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales. Similarly, if we know that the yield of rice and rainfall are closely related, we may find out the amount of rain required to achieve a certain production figure. The *statistical tool with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable is called* **regression.** With the help of regression analysis,* we are in a position to find out the *average probable* change in one variable given a certain amount of change in another.

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. For example, if we know that two variables, price (X) and demand (Y), are closely related, we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of commodity are closely related, we can find out the expected price for a certain amount of tax levy. Thus we find that the study of regression is of considerable help to the economists and businessmen.

13.1 Regression Equations

Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations–the regression of X on Y is used to describe the variation in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

Regression Equation of Y on X

The regression equation of Y on X is expressed as follows:

$$Y_c = a + bX$$

Notes

In this equation *a* and *b* are two unknown constants (fixed numerical values) which determine the position of the line completely. These constants are called the parameters of the line. If the value of either or both of them is changed, another line is determined. The parameter '*a*' determines the *level* of the fitted line (*i.e.,* the distance of the line directly above or below the origin). The parameter '*b*' determines the *slope* of the line, *i.e.*, the change in Y per unit change in X. The symbol Y_c stands for the value of X computed from the relation for a given X.

If the values of the constants '*a*' and '*b*' are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of Least Squares which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the deviation of the actual Y values from the computed Y values is the least or, in other words, in order to obtain a line which fits the points best $\sum (Y - Y_c)^2$ should be minimum. Such a line is known as the line of 'best fit'.

With a little algebra and differential calculus it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters *a* and *b* such that the least squares requirement is fulfilled:

$$\sum Y = Na + b \sum X$$

$$\sum XY = a \sum X + b \sum X^2$$

These equations are usually called the *normal equations*. In these equations $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$ indicate totals which are computed from the observed pairs of values of two variables X and Y to which the least squares estimating line is to be fitted and N is the number of observed pairs of values.

Notes The dictionary meaning of the term 'regression' is that act of returning or going back. The term 'regression' was first used by Francis Galton towards the end of nineteenth century while studying the relationship between the height of fathers and sons. This term was introduced by him in the paper 'Regression towards Mediocrity in Hereditary Stature'. His study of height of about one thousand fathers and sons revealed a very interesting relationship, *i.e.,* tall fathers tend to have tall sons and short fathers short sons, but the average height of the sons of a group of tall fathers is less than that of the fathers and the average height of the sons of a group of short fathers is greater than that of the fathers. The line describing the tendency to regress or going back was called by Galton a 'Regression Line'. The term is still used to describe that line drawn for a group of points to represents the trend present, but it no longer necessarily carries the original implication that Galton intended. These days is growing tendency of the modern writers to use the term *estimating line* instead of *regression line* because the expression estimating line is more clarificatory in character.

Regression Equation of X on Y

The regression equation of X on Y is expressed as follows:

$$X_c = a + bY$$

To determine the value of *a* and *b* the following two normal equations are to be solved simultaneously:

$$\sum X = Na + b \sum Y$$

$$\sum XY = a \sum Y + b \sum Y^2$$

Example 1: From the following data obtain the two regression equations:

Notes

X	6	2	10	4	8
Y	9	11	5	8	7

Solution: **Obtaining Regression equations**

X	**Y**	**XY**	**X^2**	**Y^2**
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\sum X = 30$	$\sum Y = 40$	$\sum XY = 214$	$\sum X^2 = 220$	$\sum Y^2 = 340$

Regression Equation of Y on X

$$Y_c = a + bX$$

To determine the value of a and b the following two normal equations are to be solved:

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values,

$$40 = 5a + 30b \quad \text{... (i)}$$

$$214 = 30a + 220b \quad \text{... (ii)}$$

Multiplying Eqn. (i) by 6

$$240 = 30a + 180b \quad \text{... (iii)}$$

$$214 = 30a + 220b \quad \text{... (iv)}$$

Subtracting Eqn. (iv) from (iii)

$$-40b = +26$$

$$b = -0.65$$

Substituting the value of b in Eqn. (i)

$$40 = 5a + 30(-0.65)$$

$$5a = 40 + 19.5 = 59.5$$

$$a = 11.9.$$

Putting the values of a and b in the equation, the regression line of Y on X is

$$Y = 11.9 - 0.65X$$

Regression Line of X on Y

$$X_c = a + bY$$

and the two normal equations are:

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2 \quad \text{... (i)}$$

$$30 = 5a + 40b \quad \text{... (ii)}$$

$$214 = 40a + 34b$$

Multiplying Eqn. (i) by 8

$$240 = 40a + 320b \quad \text{... (iii)}$$

Notes

$$214 = 40a + 340b \quad \text{... (iv)}$$

Deducting Eqn. (iv) from (iii)

$$-20b = 26$$

$$\therefore \quad b = -1.3$$

Substituting the value of b in Eqn. (i)

$$30 = 5a + 40(-1.3)$$

$$5a = 30 + 52 = 82$$

$$a = 16.4$$

Putting the values of a and b in the equation, the regression line of X on Y is

$$X = 16.4 - 1.3\,Y.$$

Deviations taken from Arithmetic Means of X and Y

The calculations can be very much simplified if instead of dealing with the actual values of X and Y we take the deviations of X and Y series from their respective means. In such a case the equation $Y_c = a + bX$ is changed to

$$Y - \bar{Y} = b\left(X - \bar{X}\right)$$

or simply $y = bx$

where $y = \left(Y - \bar{Y}\right)$ and $x = \left(X - \bar{X}\right)$

The value of b can be easily obtained as follows:

$$b = \frac{\sum xy}{\sum x^2}$$

The two normal equations which we had written earlier when changed in terms of x and y become

$$\sum y = Na + b\sum x \quad \text{... (i)}$$

$$\sum xy = a\sum x + b\sum x^2 \quad \text{... (ii)}$$

Since $\sum x = \sum y = 0$ (deviations being taken from means)

Equation (i) reduces to $Na = 0 \quad \therefore \quad a = 0$

Equation (ii) reduces to $\sum xy = b\sum x^2 \quad \therefore \quad b = \frac{\sum xy}{\sum x^2}$

After obtaining the value of b the regression equation can easily be written in terms of X and Y by substituting for y, $\left(Y - \bar{Y}\right)$ and for x, $\left(X - \bar{X}\right)$.

Similarly the regression equation $X_c = a + b\,Y$ is reduced to $x = 0$ and the value of b is obtained as follows:

$$b = \frac{\sum xy}{\sum y^2}$$

Example 2: From the following data obtain the regression equation of X on Y, and also that of Y on X:

X	6	2	10	4	8
Y	9	11	5	8	7

Solution: **Notes**

Calculation of Regression Equations

X	(X – 6) x	x^2	Y	(Y – 8) y	y^2	xy
6	0	0	9	+ 1	1	0
2	– 4	16	11	+ 3	9	– 12
10	+ 4	16	5	– 3	9	– 12
4	– 2	4	8	0	0	0
8	+ 2	4	7	– 1	1	– 2
$\Sigma X = 30$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma Y = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

Regression Equation of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}\left(Y - \bar{Y}\right)$$

$$\bar{X} = \frac{30}{5} = 6,\ \bar{Y} = \frac{40}{5} = 8,\ r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{20} = -1.3$$

$$X - 6 = -1.3\,(Y - 8)$$

$$X - 6 = -1.3\,Y + 10.4 \text{ or } X = 16.4 - 1.3\,Y$$

Regression Equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma x}\left(X - \bar{X}\right)$$

$$r\frac{\sigma_y}{\sigma x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{-26}{40} = -0.65$$

$$Y - 8 = -0.65\,(X - 6)$$

$$Y - 8 = -0.65\,X + 3.9$$

$$Y = 11.9 - 0.65\,X.$$

Deviations taken from Assumed Means

When actual means of X and Y variables are in fractions, the calculations can be simplified by taking the deviations from the assumed mean. The value of *b*, *i.e.*, the regression coefficient, will be calculated as follows:

Regression Equation of X and Y

$$X - \bar{X} = b_{xy}\left(Y - \bar{Y}\right)$$

$$\sigma_{xy} = \frac{N\,\Sigma\, d_x d_y - \left(\Sigma\, d_x\right)\left(\Sigma\, d_y\right)}{N\Sigma\, d_y^2 - \left(\Sigma\, d_y\right)^2}$$

Notes

Regression Equation of Y on X

$$Y - \bar{Y} = b_{yx}\left(X - \bar{X}\right)$$

$$b_{yx} = \frac{N \Sigma d_x d_y - \left(\Sigma d_x\right)\left(\Sigma d_y\right)}{N \Sigma d_x^2 - \left(\Sigma d_x\right)^2}$$

Once the value of b_{xy} and b_{yx} are determined in the above manner the regression equations can be obtained very easily.

Example 3: From the data of Example 1, obtain regression equations taking deviations from 5 in case of X and 7 in case of Y.

Solution:

X d_x	(X – 5) d_x^2		Y	(Y – 7) d_y	d_y^2	$d_x d_y$
6	+ 1	1	9	+ 2	4	+ 2
2	– 3	9	11	+ 4	16	– 12
10	+ 5	25	5	– 2	4	– 10
4	– 1	1	8	+ 1	1	– 1
8	+ 3	9	7	0	0	0
$\Sigma X = 30$	$\Sigma d_x = +5$	$\Sigma d_x^2 = 45$	$\Sigma Y = 40$	$\Sigma d_y = 5$	$\Sigma d_y^2 = 25$	$\Sigma d_x d_y = -21$

Regression Equation of X on Y

$$X - \bar{X} = b_{xy}\left(Y - \bar{Y}\right)$$

$$b_{xy} = \frac{N \Sigma d_x d_y - \left(\Sigma d_x\right)\left(\Sigma d_y\right)}{N \Sigma d_y^2 - \left(\Sigma d_y\right)^2}$$

$$= \frac{(5)(-21) - (5)(5)}{(5)(25) - (5)^2} = \frac{-105 - 25}{125 - 25} = -1.3$$

$$\bar{X} = \frac{30}{5} = 6, \ \bar{Y} = \frac{40}{5} = 8$$

Hence the regression equation is

$$X - 6 = -1.3\,Y\,(-8)$$

$$X - 6 = -1.3\,Y + 10.4$$

or

$$X = 16.4 - 1.3\,Y$$

Regression Equation of Y on X

$$Y - \bar{Y} = b_{yx}\left(X - \bar{X}\right)$$

Notes

$$b_{yx} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{(5)(-21) - (5)(5)}{(5)(45) - (5)^2} = \frac{-105 - 25}{225 - 25} = \frac{-130}{200} = -0$$

So the regression equation is

$$Y - 8 = -0.65 (X - 6)$$
$$Y - 8 = -0.65 X + 3.9$$
$$Y = 11.9 - 0.65 X.$$

13.2 Coefficients of Simple Regression Method

The quantity b in the regression equation is called the regression coefficient. Since there are two regression equations there are also two regression coefficients–regression coefficient of X on Y and regression coefficient of Y on X.

Regression Coefficient of X on Y

The regression coefficient of X and Y is represented by the symbol b_{xy} or b_1. It measure the change in X corresponding to a unit change in Y. The regression coefficient of X on Y is given by

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

where deviations are taken from means of X and Y, the regression coefficient is obtained by:

$$b_{xy} * = \frac{\Sigma xy}{\Sigma y^2}$$

$$* r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}},\ \sigma_x = \sqrt{\frac{\Sigma x^2}{N}} \text{ and } \sigma_y = \sqrt{\frac{\Sigma y^2}{N}}$$

Substituting these values

$$b_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \times \frac{\sqrt{\frac{\Sigma y^2}{N}}}{\sqrt{\frac{\Sigma x^2}{N}}} = \frac{\Sigma xy}{\Sigma y^2}$$

where deviations are taken from assumed means, the value of b_{xy} is obtained as follows:

$$b_{xy} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

Regression Coefficient of Y on X

The regression coefficient of Y on X is represented by b_{yx} or b_2. It measures the change in Y corresponding to unit change in X. The value of b_{yx} is given by

Notes

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

When deviations are taken from actual means of X and Y

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

When deviations are taken from assumed means of X and Y

$$b_{yx} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

Calculating Correlation from Regression Coefficients

It should be interesting to note that the underroot of the product of the two regression coefficients gives us the value of the coefficient of correlation. Symbolically:

$$r = \sqrt{b_1 \times b_2} \text{ or } \sqrt{b_{xy} \times b_{yx}}$$

Proof:

$$b_1 \text{ or } b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$b_2 \text{ or } b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_1 \times b_2 = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2$$

$$\therefore \quad r = \sqrt{b_1 \times b_2}$$

Since the value of the coefficients of correlation (r) cannot exceed one, one of the regression coefficients must be less than one, or in other words, both the regression coefficients cannot be greater than one. For example, if $b_{yx} = 1.2$ and $b_{xy} = 1.4$, r would be $\sqrt{1.2 \times 1.4} = 1.29$ which is not possible. Further, the regression coefficient which may exceed one should also be such in value that when multiplied by the other coefficient the underroot of the product of the two coefficients does not exceed one. Also both the regression coefficients will have the same sign, *i.e.,* they will be either positive or negative. The coefficient of correlation (r) will have the same sign as that of regression coefficients, *i.e.,* if regression coefficients have a negative sign, r will also have negative sign and if regression coefficients have a positive sign, r will also have positive sign.

For example, if $b_{xy} = -0.8$ and $b_{yx} = -1.2$, r would be

$$\sqrt{-0.8 \times -1.2} = -0.98$$

Since

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

we can find out any of the four values, given the other three. For example, if we know that $r = 0.6$, $\sigma_x = 4$ and $\sigma_{xy} = 0.8$, we can find σ_y

Notes

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$0.8 = \frac{0.6 \times 4}{\sigma_y}$$

$$\sigma_y = \frac{2.4}{0.8} = 3.$$

Example 4: Obtain the value of the correlation coefficient through the method of regression analysis from the data given below first by taking deviation from the actual means of X and Y and secondly from assumed means 2 and 18 for series X and Y respectively.

X	1	2	3	4	5
Y	10	20	15	25	30

Solution:

(a) Calculation of Regression Coefficient from the actual means

X	(X – 3) x	x^2	Y	(Y – 20) y	y^2	xy
1	– 2	4	10	– 10	100	20
2	– 1	1	20	0	0	0
3	0	0	15	– 5	25	0
4	+ 1	1	25	+ 5	25	5
5	+ 2	4	30	+ 10	100	20
$\Sigma X = 15$	$\Sigma x = 0$	$\Sigma x^2 = 10$	$\Sigma Y = 100$	$\Sigma y = 0$	$\Sigma y^2 = 250$	$\Sigma xy = 45$

$$\bar{X} = \frac{15}{5} = 3, \bar{Y} = \frac{100}{5} = 20$$

Regression Coefficient of X on Y

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{45}{250} = 0.18$$

Regression Coefficient of Y on X

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{45}{10} = 4.5$$

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.18 \times 4.5} = \sqrt{0.81} = 0.9.$$

Notes

(b) Calculation of Regression coefficients from the assumed means 2 and 18

X	$(X-2)$		Y	$(Y-18)$			
d_x	d_x^2			d_y	d_y^2	$d_x d_y$	
1	– 1	1		10	– 8	64	+ 8
2	0	0		20	+ 2	4	0
3	+ 1	1		15	– 3	9	– 3
4	+ 2	4		25	+ 7	49	+ 14
5	+ 3	9		30	+ 12	144	+ 36
$\Sigma X = 15$	$\Sigma d_x = 5$	$\Sigma d_x^2 = 15$		$\Sigma Y = 100$	$\Sigma d_y = 10$	$\Sigma d_y^2 = 270$	$\Sigma d_x d_y = 55$

$$b_{xy} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$$\Sigma d_x d_y = 55,\ \Sigma d_x = 5,\ \Sigma d_y = 10,\ \Sigma d_y^2 = 270,\ N = 5$$

$$= \frac{(5)(55) - (5)(10)}{(5)(270) - (10)^2} = \frac{275 - 50}{1,350 - 100} = \frac{225}{1,250} = 0.18$$

$$b_{yx} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{5(55) - (5)(10)}{(5)(15) - (5)^2} = \frac{275 - 50}{75 - 25} = \frac{225}{50} = 4.5$$

Thus the value of regression coefficients is the same by both the methods.

Example 5: Find the regression coefficient of X on Y and Y on X for the following data:

X	3	2	– 1	6	4	– 2	5
Y	5	13	12	– 1	2	20	0

Solution: **Calculation of Regression Coefficients**

X	$(X-2)$		Y	$(Y-7)$		
	d_x	d_x^2		d_y	d_y^2	$d_x d_y$
3	+ 1	1	5	– 2	4	– 2
2	0	0	13	+ 6	36	0
– 1	– 3	9	12	+ 5	25	– 15
6	+ 4	16	– 1	– 8	64	– 32
4	+ 2	4	2	– 5	25	– 10
– 2	– 4	16	20	+13	169	– 52
5	+ 3	9	0	– 7	49	– 21
$\Sigma X = 17$	$\Sigma d_x = +3$	$\Sigma d_x^2 = 55$	$\Sigma Y = 51$	$\Sigma d_y = +2$	$\Sigma d_y^2 = 372$	$\Sigma d_x d_y = -132$

Regression Coefficient of X on Y

Notes

$$b_{xy} = \frac{N \Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N \Sigma d_y^2 - (\Sigma d_y)^2}$$

$$\Sigma d_x d_y = -132,\ \Sigma d_x = +3,\ \Sigma d_y = 2,\ \Sigma d_y^2 = 372,\ N = 7$$

$$b_{xy} = \frac{(7)(-132) - (3)(2)}{(7)(372) - (2)^2} = \frac{-924 - 6}{2,604 - 4} = -0.353$$

Regression Coefficient of Y on X

$$b_{yx} = \frac{N \Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N \Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{(7)(-132) - (3)(2)}{(7)(55) - (3)^2} = \frac{-924 - 6}{385 - 9} = -2.473.$$

Example 6: Given the bivariate data:

X	1	5	3	2	1	1	7	3
Y	6	1	0	0	1	2	1	3

(a) Fit a regression line of Y on X and thence predict Y if X = 5.

(b) Fit the regression line of X on Y and thence predict X if Y = 2.5.

Solution:

X	(X – 3) d_x	d_x^2	Y	(Y – 2) d_y	d_y^2
1	– 2	4	6	+ 4	16
5	+ 2	4	1	– 1	1
3	0	0	0	– 2	4
2	– 1	1	0	– 2	4
1	– 2	4	1	– 1	1
1	– 2	4	2	0	0
7	+ 4	16	1	– 1	1
3	0	0	5	+ 3	9
$\Sigma X = 23$	$\Sigma d_x = -1$	$\Sigma d_x^2 = 33$	$\Sigma Y = 16$	$\Sigma d_y = 0$	$\Sigma d_y^2 = 36$

Regression Equation of X on Y

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

Notes

$$b_{xy} = \frac{N \Sigma d_x d_y - \Sigma d_x d_y}{N \Sigma d_y^2 - \left(\Sigma d_y\right)^2}$$

$$b_{xy} = \frac{(8)(-10) - (-2)(0)}{(8)(36) - (0)^2} = \frac{-80}{288} = -0.278$$

$$\overline{X} = \frac{23}{8} = 2.875, \ \overline{Y} = \frac{16}{8} = 2, \ b_{xy} = -0.278.$$

Substituting the values in the equation

$$X - 2.875 = -0.278 (Y - 2)$$
$$X - 2.875 = -0.278 Y + 0.556$$
$$X = 3.431 - 0.278 Y$$

If $Y = 2.5$, X is equal to $3.431 + (0.278 \times 2.5)$

$$= 3.431 - 0.695 = 2.736.$$

Regression Equation of Y on X

$$Y - \overline{Y} = b_{yx}\left(X - \overline{X}\right)$$

$$b_{xy} = \frac{N \Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N \Sigma d_y^2 - \left(\Sigma d_y\right)^2}$$

$$= \frac{(8)(-10) - (-1)(0)}{8(33) - (-1)^2} = \frac{-80}{264 - 1} = -0.304$$

$$Y - 2 = -0.304 (X - 2.875)$$
$$Y - 2 = -0.304 X + 0.874$$
$$Y = 2.874 - 0.304 X$$

$X = 5$, Y is equal to $2.874 - (0.304 \times 5) = 2.874 - 1.52$

$$= 1.354$$

$$r = \sqrt{b_{xy} \times b_{yx}} = -\sqrt{(-0.278)(-0.304)} = -0.291.$$

Example 7: Given that the means of X and Y are 65 and 67, their standard deviations are 2.5 and 3.5 respectively, and the coefficient of correlation between them is 0.8.

(i) Write down the two regression lines.

(ii) Obtain the best estimate of X when Y = 70.

(iii) Using the estimated value of X as the given value of X, estimate the corresponding value of Y.

Solution:

(i) ***Regression Line of Y on X***

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

$\overline{Y} = 67, \ \overline{X} = 65, \ \sigma_y = 3.5, \ \sigma_x = 2.5, \ r = 0.8$

Notes

$$Y - 67 = 0.8\frac{3.5}{2.5}(X - 65)$$

$$Y - 67 = 1.12 (X - 65)$$

$$Y - 67 = 1.12 X - 72.8$$

$$Y = 1.12 X - 5.8 \text{ or } Y = -5.8 + 1.12 X$$

Regression Line of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$X = 65 = 0.571 (Y - 67)$$

$$X - 65 = 0.571 Y - 38.257$$

$$X = 0.571 Y + 26.743 \text{ or } X = 26.743 + 0.571 Y$$

(ii) Best estimate of X when Y = 70 can be obtained from the regression equation of X on Y.

$$X = 26.743 + 0.571 (70) = 26.743 + 39.97 = 66.713$$

(iii) When X = 66.713, Y will be

$$X = 1.12 (66.713) - 5.8 = 74.72 - 5.8 = 68.92.$$

Self-Assessment

1. Which of the following statements are True or False (T/F):

(*i*) If both the regression coefficients are negative, the correlation coefficient would be negative.

(*ii*) The under root of two regression coefficients gives us the value of correlation coefficient.

(*iii*) Regression coefficients are independent of change of scale and origin.

(*iv*) Regression coefficient of Y on X measures the change in X corresponding to a unit change in Y.

(*v*) The regression coefficient of Y on X is denoted by the symbol *bxy*.

13.3 Summary

- The *statistical tool with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable is called* **regression.** With the help of regression analysis,* we are in a position to find out the *average probable* change in one variable given a certain amount of change in another.
- Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. For example, if we know that two variables, price (X) and demand (Y), are closely related, we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of commodity are closely related, we can find out the expected price for a certain amount of tax levy. Thus we find that the study of regression is of considerable help to the economists and businessmen.
- Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations–the regression of X on Y is used to describe the variation in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.
- If the values of the constants '*a*' and '*b*' are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of Least Squares

Notes

which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the deviation of the actual Y values from the computed Y values is the least or, in other words, in order to obtain a line which fits the points best $\sum (Y - Y_c)^2$ should be minimum. Such a line is known as the line of 'best fit'.

- The quantity *b* in the regression equation is called the regression coefficient. Since there are two regression equations there are also two regression coefficients–regression coefficient of X on Y and regression coefficient of Y on X.
- the regression coefficient which may exceed one should also be such in value that when multiplied by the other coefficient the underroot of the product of the two coefficients does not exceed one. Also both the regression coefficients will have the same sign, *i.e.,* they will be either positive or negative. The coefficient of correlation (*r*) will have the same sign as that of regression coefficients, *i.e.,* if regression coefficients have a negative sign, *r* will also have negative sign and if regression coefficients have a positive sign, *r* will also have positive sign.

13.4 Key-Words

1. Regression : A statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

13.5 Review Questions

1. What are regression coefficients ? State the properties of regression coefficients
2. What are regression equations ? Explain regression equation of Y on X and X on Y.
3. If one of the regression coefficients is negative what type of variation would you expect in the original series of pairs of observations ?
4. Give the various properties and characteristics of regression coefficients.
5. If two regression coefficients are 0.8 and 0.6, what would be the coefficient of correlation? [r = 0.693]

Answers: Self-Assessment

1. (*i*) T (*ii*) T (*iii*) F (*iv*) F (*v*) F

13.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Unit 14: Correlation Analysis Vs. Regression Analysis

Notes

CONTENTS

Objectives

After reading this unit students will be able to:

- Describe Correlation and Regression Analysis.
- Explain Correlation Analysis Vs. Regression Analysis.

Introduction

Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Values of the correlation coefficient are always between – 1 and + 1. A correlation coefficient of + 1 indicates that two variables are perfectly related in a positive linear sense, a correlation coefficient of – 1 indicates that two variables are perfectly related in a negative linear sense, and a correlation coefficient of 0 indicates that there is no linear relationship between the two variables. For simple linear regression, the sample correlation coefficient is the square root of the coefficient of determination, with the sign of the correlation coefficient being the same as the sign of b 1, the coefficient of x l in the estimated regression equation.

Neither regression nor correlation analyses can be interpreted as establishing cause-and-effect relationships. They can indicate only how or to what extent variables are associated with each other. The correlation coefficient measures only the degree of linear association between two variables. Any conclusions about a cause-and-effect relationship must be based on the judgment of the analyst.

When once a relationship between two variables is ascertained, it is quite likely that estimating the value of one for some given value of other is expected. This can be done with the help of regression. It measures the average relationship between two or more variables in terms of original units of the data. The dictionary meaning of the term 'Regression' is to revert or return back. The term was used for the first time by Sir Francis Galton in 1877. In statistics, the technique of Regression is used in all those fields where two or more variables have the tendency to go back to the mean. While correlation measures the direction and strength of the relationship between two or more variables, regression involves methods by which estimates are made of the values of a variable from the knowledge of the values of one or more other variables. Along with this, measurement of the error involved in the estimation process are also included. This means, the regression technique can be used for the prediction on the basis of the average relationship.

Notes

14.1 Correlation Analysis

So far we have studied problems relating to one variable only. In practice, we come across a large number of problems involving the use of two or more than two variables. If two quantities vary in such a way that movements in one are accompanied by movements in the other, these quantities are correlated. For example, there exists some relationship between age of husband and age of wife, price of a commodity and amount demanded, increase in rainfall up to a point and production of rice, an increase in the number of television licences and number of cinema admissions, etc. The statistical tool with the help of which these relationships between two or more than two variables are studied is called **correlation.** The measure of correlation, called the correlation coefficient, summarizes in one figure the direction and degree of correlation. Thus correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables. A very simple definition of correlation is that given by A.M. Tuttle. He defines correlation as: "An analysis of the covariation of two or more variables is usually called *correlation*".

The problem of analysing the relation between different series can be broken down into three steps:

(1) Determining whether a relation exists and, if does, measuring it.

(2) Testing whether it is significant.

(3) Establishing the cause and effect relation, if any.

In this unit, only the first aspect will be discussed. For second aspect a reference may be made in the unit on Tests of Significance. The third aspect in the analysis, that of establishing the cause-effect relation, is beyond the scope of statistical analysis. An extremely high and significant correlation between the increase in smoking and increase in lung cancer would not prove that smoking causes lung cancer. The proof of a cause and effect relation can be developed only by means of an exhaustive study of the operative elements themselves.

It should be noted that the detection and analysis of correlation *(i.e.*, covariation) between two statistical variables requires relationships of some sort which associate the observations in pairs, one of which pair being a value of each of the two variables. In general, the pairing relationship may be of almost any nature, such as observations at the same time or place over a period of time or different places.

Notes The computation concerning the degree of closeness is based on the regression equation. However, it is possible to perform correlation analysis without actually having a regression equation.

Utility of the Studt of Correlation

The study of correlation is of immense use in practical life because of the following reasons:

1. Most of the variables show some kind of relationship. For example, there is relationship between price and supply, income and expenditure, etc. With the help of correlation analysis, we can measure in one figure the degree of relationship existing between the variables.
2. Once we know that two variables are closely related, we can estimate the value of one variable given the value of another. This is done with the help of regression equations discussed in Unit 8.
3. Correlation analysis contributes to the economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connection by which disturbances spread and suggest to him the paths through which stabilizing forces become effective.

   In business, correlation analysis enables the executive to estimate costs, sales, prices and other variables on the basis of some other series with which these costs, sales, or prices may be functionally related. Some guesswork can be removed from decisions when the relationship between a variable to be estimated and the one or more other variables on which it depends are close and reasonably invariant.

Notes

However, it should be noted that coefficient of correlation is one of the most widely used and also one of the most widely *abused* of statistical measures. It is abused in the sense that one sometimes overlooks the fact that *r* measures nothing but the strength of *linear* relationships and that it does not necessarily imply a cause-effect relationship.

4. Progressive development in the methods of science and philosophy has been characterised by increase in the knowledge of relationships or correlations. Nature has been found to be a multiplicity of interrelated forces.

Correlation and Causation

Correlation analysis helps in determining the degree of relationship between two or more variables—it does not tell us anything about cause and effect relationship. Even a high degree of correlation does not necessarily mean that a relationship of cause and effect exists between the variables or, simply stated, correlation does not necessarily imply causation of functional relationship though the existence of causation always implies correlation. By itself it establishes only *covariation*. The explanation of a significant degree of correlation may be due to any one or a combination of the following reasons:

1. **The correction may be due to pure chance, especially in a small sample:** We may get a high degree of correlation between two variables in a sample but in the universe there may not be any relationship between the variables at all. This is especially so in case of small samples. Such a correlation may arise either because of pure random sampling variation or because of the bias of the investigator in selecting the sample. The following example shall illustrate the point:

Income (Rs.)	Weight (*lb.*)
10,000	120
20,000	140
30,000	160
40,000	180
50,000	200

The above data show a perfect positive relationship between income and weight, *i.e.*, as the income is increasing the weight is increasing and the ratio of change between two variables is same.

2. **Both the correlated variables may be influenced by one or more other variables:** It is just possible that a high degree of correlation between variables may be due to the same causes affecting each variable or different causes affecting each with the same effect. For example, a high degree of correlation between the yield per acre of rice and tea may be due to the fact that both are related to the amount of rainfall. But none of the two variables is the cause of the other.
3. **Both the variables may be mutually influencing each other so that neither can be designated as cause and the other the effect:** There may be a high degree of correlation between the variables but it may be difficult to pin point as to which is the cause and which is the effect. This is especially likely to be so in case of economic variables. For example, such variables as demand and supply, price and production, etc., mutually interact. To take a specific case, it is a well known principle of economics that as the price of a commodity increases its demand goes down and so price is the cause, and demand the effect. But it is also possible that increased demand of a commodity due to growth of population or other reasons may force its price up. Now the cause is the increased demand, the effect the price. Thus at times it may become difficult to explain from the two correlated variables which is the cause and which is the effect because both may be reacting on each other.

The above points clearly bring out the fact that correlation does not manifest causation of functional relationship. By itself, it establishes only covariation. Correlation observed between variables that could not conceivably be causally related are called *spurious* or *non-sense correlation*. More appropriately we should remember that it is the *interpretation* of the degree of correlation that is spurious, not the

Notes

degree of correlation itself. The high degree of correlation indicates only the mathematical result. We should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. It should also be noted that errors in correlation analysis include not only reading causation into spurious correlation but also interpreting spuriously a perfectly valid association.

14.2 Regression Analysis

Meaning and Definition

The term regression was for the first time used by Sir Francis Galton in 1877 while studying the relationship between the height of fathers and sons. He carried out a study on height of one thousand fathers and sons and revealed that tall fathers tend to have tall sons and short fathers short sons, but the average height of the sons of a group of tall fathers is less than that of the fathers and the average height of the sons of a group of short fathers is greater than that of the fathers. This line describing the tendency to regress or going back was called as a regression line by Galton. The term is still used to describe that line drawn for a group of points to represent the trend present, however, today it does not necessarily have the original implication of stepping back (for which Galton had used this term). In modern times term 'estimating line' is coming to be used instead of 'regression line'. On examining a few definitions, the term regression as used in statistics can be clearly described.

(1) As described by *Morris Hamburg,* "The term regression analysis refers to the methods by which estimates are made of the values of a variable from a knowledge of values of one or more other variables and to the measurement of the errors involved in this estimation process."

(2) According to *Taro Yamane,* "One of the most frequently used techniques in economics and business research, to find, relation between two or more variables that are related casually, is regression analysis."

On the basis of the above definitions, it has become very clear that regression analysis is done for estimating or predicting the unknown value of one variable from the known value of the other variable. The variable which is used to predict the variable of interest is called the independent variable or explanatory variable and the variable we are trying to predict is called the dependent or explained variable.

Did u know? In the words of *Ya Lum Chou,* "Regression analysis attempts to establish the nature of the relationship between variables, that is, to study the functional relationship between the variable and thereby provide a mechanism for prediction or forecasting."

14.3 Correlation Analysis Vs. Regression Analysis

Most of the times, the correlation and regression analysis are confused with one another, probably because of the fact that both of them study about the relationship between two variables. By studying the points of differentiation, this would become clear:

(1) Correlation coefficient measures the degree of covariability between X and Y. The regression analysis, on the other hand, studies the nature of relationship between X and Y so that one may be predicted on the basis of the other.

(2) Correlation only ascertains the degree of relationship between two variables and it not be made clear that one variable is the cause and the other is the effect. But in regression analysis, one variable is taken as dependent while other as independent so that the cause and effect relationship can be studied.

(3) In correlation $r_{xy} = r_{yx}$ but regression coefficients b_{xy} is never equal to b_{yx}.

$$b_{xy} \neq b_{yx}.$$

(4) Correlation may be found to exist between two variables by chance with no practical relevance. But in regression the results are never by chance.

Notes

(5) Correlation coefficient is independent of origin and change of scale. Regression coefficient is independent of change of scale but not of origin.

Some Similarities: (1) Coefficient of correlation for two variables shall take the same sign as regression coefficients. (2) If, at a given level of significance, the value of regression coefficients is significant, the value of correlation coefficient shall also be significant at that level.

On examining the various definitions, it reveals that regression is a tool which helps in estimating or predicting the unknown value of one variable from the known value of the other variable. It differs from correlation as the later only tell the direction and extent of relationship between two variables whereas regression is a step further.

Self-Assessment

1. Indicate whether the following statements are True or False [T/F]:

(*i*) Correlation always signifies a cause and effect relationship between the variables.

(*ii*) If *r* is negative both the variables are decreasing.

(*iii*) Regression analysis reveals average relationship between two variables.

(*iv*) The terms 'dependent' and 'independents' do not imply that there is necessarily any cause and effect relationship between the variables.

(*v*) In regression analysis b_{xy} stands for regression coefficient of X on Y.

14.4 Summary

- Correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables. A very simple definition of correlation is that given by A.M. Tuttle. He defines correlation as: "An analysis of the covariation of two or more variables is usually called *correlation"*.
- The computation concerning the degree of closeness is based on the regression equation. However, it is possible to perform correlation analysis without actually having a regression equation.
- Correlation analysis contributes to the economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connection by which disturbances spread and suggest to him the paths through which stabilizing forces become effective.
- Progressive development in the methods of science and philosophy has been characterised by increase in the knowledge of relationships or correlations. Nature has been found to be a multiplicity of interrelated forces.
- Correlation observed between variables that could not conceivably be causally related are called *spurious* or *non-sense correlation.* More appropriately we should remember that it is the *interpretation* of the degree of correlation that is spurious, not the degree of correlation itself. The high degree of correlation indicates only the mathematical result. We should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. It should also be noted that errors in correlation analysis include not only reading causation into spurious correlation but also interpreting spuriously a perfectly valid association.
- In modern times term 'estimating line' is coming to be used instead of 'regression line'. On examining a few definitions, the term regression as used in statistics can be clearly described.
- The variable which is used to predict the variable of interest is called the independent variable or explanatory variable and the variable we are trying to predict is called the dependent or explained variable.

Notes

14.5 Key-Words

1. Correlation : In the world of finance, a statistical measure of how two securities move in relation to each other. Correlations are used in advanced portfolio management.

14.6 Review Questions

1. Define Correlation. Discuss its uses.
2. What is regression ? Explain clearly the significance of this concept with the help of an example.
3. Distinguish clearly between regression and correlation analysis giving suitable examples.
4. What are the similarities between correlation and regression analysis ?
5. Describe Correlation and Causation.

Answers: Self-Assessment

1. (*i*) F (*ii*) F (*iii*) T (*iv*) T (*v*) T

14.7 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 15 : Index Number–Introduction and Use of Index Numbers and their Types

CONTENTS

Objectives

After reading this unit students will be able to :

- Discuss the Introduction of Index Number.
- Know the Use of Index Number.
- Explain the Types of Index Number.

Introduction

Any change in the level of a phenomenon with respect to time, geographical location etc. is measured with the help of a statistical device called Index numbers. It was for the first time used to compare the changes in prices for the year 1750 with the price level of year 1500 in Italy. It was constructed by Carli. However, today it is used to measure the change in level of any phenomena may it be changes national income, expenditure, cost of living, incidences of crimes, number of accidents and so on. Index numbers are said to be barometers which measure the charge in the level of a phenomenon.

15.1 Introduction to Index Numbers

Index numbers have become today one of the most widely used statistical devices. Though originally developed for measuring the effect of change in prices, there is hardly any field today where index numbers are not used. Newspapers headline the fact that prices are going up or down, that industrial production is rising or falling, that imports are increasing or decreasing, that crimes are rising in a particular period compared to the previous period as disclosed by index numbers. They are used to feel the pulse of the economy and they have to be used as indicator of inflationary or deflationary tendencies. In fact, they are described as *barometers of economic activity, i.e.,* if one wants to get an idea as to what is happening to an economy he should look to important indices like the index number of industrial production, agricultural production, business activity, etc.

An index number may be described as a specialized average designed to measure the change in the level of a phenomenon with respect to time, geographic location or other characteristics such as income, etc. Thus, when we say that the index number of wholesale prices is 125 for the period Dec. 2005 compared to Dec. 2004, it means there is a net increase in the prices of wholesale commodities to the extent of 25 per cent.

Notes

For a proper understanding of the term index number, the following points are worth considering:

(1) Index numbers are specialized averages : As explained in unit 4 central value, an average is a single figure representing a group of figures. However, to obtain an average, items must be comparable; for example, the average weight of men, women and children of a certain locality has no meaning at all. Furthermore, the unit of measurement must be the same for all the items. Thus an average of the weight expressed in *kg*., *lb*., etc., has no meaning. However, this is not so with index numbers. Index numbers are used for purposes of comparison in situations where two or more series are expressed in different units or the series are composed of different types of items. For example, while constructing a consumer price index the various items are divided into broad heads, namely (i) Food, (ii) Clothing, (iii) Fuel and Lighting, (iv) House Rent, and (v) Miscellaneous. These items are expressed in different units : thus under the head 'food' wheat and rice may be quoted per quintal, ghee per *kg*., etc.

Similarly, cloth may be measured in terms of metres or yards. An average of all these items expressed in different units is obtained by using the technique of index numbers.

(2) Index numbers measure the change in the level of a phenomenon : Since index numbers are essentially averages they describe in one simple figure the increase or decrease in the level of a phenomenon under study. Thus if the consumer price index of working class for Delhi has gone up to 125 in 2005 compared to 100 in 2004 it means that there is a net increase of 25% in the prices of commodities included in the index. Similarly, if the index of industrial production is 108 in 2005 compared to 100 in 2004 it means that there is a net increase in industrial production to the extent of 8%. It should be carefully noted that even where an index is showing a net increase, it may include some items which have actually decreased in value and others which have remained constant.

(3) Index numbers measure the effect of change over a period of time : Index numbers are most widely used for measuring changes over a period of time. Thus we can find out the net change in agricultural prices from the beginning of first plan period to the Ninth plan period, *i.e.,* 1997-2002. Similarly, we can compare the agricultural production, industrial production, imports, exports, wages, etc., at two different times. However, it should be noted the index numbers not only measure changes over a period of time but also compare economic conditions of different locations, different industries, different cities or different countries. But since the basic problems are essentially the same and since most of the important index numbers published by the Government and private research organisations refer to data collected at different times, we shall consider in this chapter index numbers measuring changes relative to time only. However, methods described can be applied to other cases also.

Did u know? "Index numbers are devices for measuring differences in the magnitude of a group of related variables."

Meaning and Definition

Index numbers are the specialised averages designed to measure the changes in a group of related variables over a period of time. Some important definitions of index number are given below. They would help in understanding about what index numbers are :

(1) According to *Spiegel,* "An index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographic location or other characteristics such as income, profession etc."

(2) As per *Morris Homburg,* "In its simplest form, an index number is nothing more than a relative number or a 'relative' which expresses the relationship between two figures, where one of the figures is used as base."

(3) *A. M. Tuttle* suggests, "Index number is a single ratio (usually in percentages) which measures the combined (*i.e.,* averaged) change of several variables between two different times, places or situations."

Notes

(4) As per *Patterson,* "In its simplest form, an index number is the ratio of two index numbers expressed as a per cent. An index number is a statistical measure — a measure designed to show changes in one variable or in a group of related variables over time or with respect to geographic location or other characteristic."

From the above definitions it is very clear that the index numbers are specialised averages designed to measure change in a group of related variables over a period of time.

Features of Index Number

To understand what an index number is the following features are worth considering :

(1) Index numbers are specialised averages,

(2) Index numbers measure the effect of changes over a period of time,

(3) Index numbers measure the net change in a group of related variables.

15.2 Purpose or Use of Index Number

The definitions and features of index number stated above makes it very clear that index numbers measure changes. In this way they are indispensible fools in the hands of economists and business analysts who constantly work for change, of course, towards betterment. The various uses of index numbers are highlighted below :

(1) Index numbers reveal trends and tendencies : The trend of the phenomenon under study can be obtained by measuring the changes over a period of time. On the basis of this analysis can be done. For example, by examining the index numbers of industrial production, business activity etc. their trend can be understood and analysed.

(2) Index numbers help in measuring suitable policies : The above use of index number which reveals the trends and tendencies help in framing suitable policies so as to achieve the said goal. For example, by knowing about the rising trend of imports, suitable policy can be formulated to prevent it.

(3) Index numbers are used in deflating : Index numbers are very useful in deflating *i.e.,* they are used to adjust the original data for price changes, or to adjust wages for cost of living changes and thus transform nominal wages into real wages. Moreover nominal income can be transformed into real income and nominal sales into real sales through appropriate index numbers.

(4) Index number are used in forecasting future economic activity : Along with studying the past and the present variables of the economy, index numbers are also useful in estimating the future economic activity. The long-term variations, trends etc. help in estimating the coming problems.

On the basis of the above, it may be concluded that index numbers are strong tools into the hands of the economists and business analysts with the help of which they can measure changes in certain phenomenon over a period of time or through a geographical location. This enables them to form suitable policies to obtain the desired results. Past, present and future trends and tendencies are revealed with the help of index numbers and they are also useful in deflating. In the words of *Kafka* and *Simpson,* "Index numbers are today one of the most widely used statistical devices. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies."

Problems in the Construction of Index Numbers

Index numbers are constructed in the form of specialised averages with definite purpose and with some base period. Before constructing index numbers, it is necessary to know about the various problems which arise in its construction so that they can be minimised. Some of the important problems are :

(1) Selection of base period : Base period is the one against which comparisons are made. It may be year a month or a day. The index for base period is always taken to be 100. It is essential to choose an appropriate base before constructing index numbers. The base is selected as per the object of the index, but following considerations should also be made — (a) The base period

Notes

should not be a very old period. It must be a fairly recent period, so that comparisons are between similar set of circumstances. (b) The base period selected should be a normal period *i.e.,* that period should be free from abnormalities like war, earthquake, boom or depression etc. (c) It has to be predecided whether fixed base or chain base index has to be prepared. In fixed base, the base year remains fixed, while in chain-base, the base year for each time period is the index of the preceeding time period *i.e.,* the base is not fixed, it changes with each time period (a year, a month a week or a day).

(2) **Purpose of Index numbers :** While constructing index number the purpose for which it is constructed must be clearly defined. This is because there are no all-purpose index and every index is of limited and particular use. Lack of clearity of purpose would lead to confusion and wastage of time with no fruitful results.

(3) **Selection of number of items :** To study the change in a certain phenomenon, it is not possible to include each and every item leading to change. For example, while constructing price index, change in price of each and every item cannot be included. The selection of commodities should be such that they are representative of the tastes, preferences and habits etc. of the group of people regarding whom the index is constructed. In this way, it becomes a big problem as to which items to be included and which to be excluded ? Again, by having purpose of index properly defined, the selection of items can be eased. Index numbers would give false results if at one time one set of quantities are used and at other time other set of quantities are used.

(4) **Choice of average :** Another problem faced while constructing an index is the decision as to which average, mean, median, mode, geometric mean or harmonic mean, should be used for constructing the index. Each one has its own advantages and drawbacks. Theoretically, geometric mean is considered to be most suitable because of the following reasons — (a) While constructing index numbers ratios of relative changes are taken into consideration and geometric mean gives equal weights to equal ratio of change, (b) index numbers that use geometric mean can be reversed which makes base shifting easily possible, (c) geometric mean is not influenced much by violent fluctuations in the values of individual items. However, use of arithmetic mean is also popular because it is simpler to compute than geometric mean.

(5) **Selection of appropriate weights :** The term 'weight' refers to the relative importance of the different items in the construction of the index. All items are not of equal importance and hence it is necessary to device some suitable method which is done by allocating weights. In case of weighted indices, specific weights are assigned. Implicit or explicit methods of assigning weights can be used. Quantity or value weights can be assigned. If aggregative method is used, quantities are used as weights and in averaging of price relatives method of constructing index, value weights are used. Moreover, it has to be decided whether the weights would be fixed or fluctuating. In this way selection of appropriate weights forms a very crucial problem in constructing index.

(6) **Selecting appropriate formula :** There is a large range of formulae available to construct index numbers. Choosing among these to prepare index is another problem while constructing index numbers. The choice of formula should depend not only on purpose but also on the availability of the data. Theoretically, Fisher's method is the 'ideal' method for constructing index numbers. However, depending upon the purpose and availability of data, other methods may also be used.

(7) **Obtaining quotations :** The change in the level of certain phenomenon can be measured only when proper data regarding quotations is available. For example, while preparing price index, it is essential to obtain proper price quotations of the selected commodities. Or while preparing expenditure index, information about expenditure should be available. In the absence of real information, the results may be misleading. This poses to be another problem while constructing index numbers.

On the basis of the above it can be said that most of the problems arise as a result of availability of alternatives. That is, in the presence of alternatives it becomes difficult to choose the best possible alternative, for example, which average to use, which method to weight must be used,

Notes

which base year to be selected, what items should be selected, which formula should be chosen and so on. If clearity of purpose is assured most of these problems can be solved. Regarding obtaining proper quotations, efforts should be made to choose reliable sources, and items chosen should be standardised and graded, so that quotations are easily and accurately obtained.

15.3 Types of Index Numbers

Price Relatives

One of the simplest types of index numbers is a *price relative.* It is the ratio of the price of a *single* commodity in a given period or point of time to its price in another period or point of time, called the *reference period* or *base period.* If prices for a period, instead of a point of time, are considered, then suitable price average for the period is taken and these prices are expressed in the same units.

If p_0 and p_n denote the price of a commodity during the base period or reference period (0) and the given period (n) then the *price* relative of the period n with respect to (w.r.t) the base period 0 is

defined by price relative in percentage (of period n w.r.t. period 0) = $\frac{p_n}{p_0} \times 100$ ... (1)

and is denoted by $p_{0/n}$.

For example, if the retail price of fine quality of rice in the year 1980 was Rs. 3.75 and that for the year 1983 was Rs. 4.50 then

$$p_{1980\,|\,1983} = \frac{\text{Rs. } 4.50}{\text{Rs. } 3.75} \times 100 = 120\%$$

For example, the exchange rate of a U.S. dollar was Rs. 10.00 in July 1984 (period J) and was Rs. 12.50 in December 1984 (period D) then the price relative of a dollar in December w.r.t. that in July is given by

$$p_{J|D} = \frac{\text{Rs.}12.50}{\text{Rs.}10.00} \times 100 = 125\%$$

Quantity Relatives

Another simple type of index numbers is a quantity relative, when we are interested in changes in quantum or volumes of a commodity such as quantities of production or sale or consumption. Here the commodity is used in a more general sense. It may mean the volume of goods (in tonnes) carried by roadways, the number of passenger miles travelled, or the volume of export to or import from a country. In such cases we consider of quantity or volume relatives. If quantities or volumes are for a period instead of a point of time, a suitable average is to be taken and the quantities or volumes are to be expressed in the same units. If q_0 and q_n denote the quantity or volume produced, consumed or transacted during the base period (0) and the given period (n) then *quantity relative* of the period n w.r.t. the base period 0 is defined by quantity relative in percentage (of period n w.r.t. period 0) =

$\frac{q_n}{q_0} \times 100$ ... (2)

and is denoted by $q_{0|n}$.

Value Relatives

A value relative is another type of simple index number, usable when we wish to compare changes in the money value of the transaction, consumption or sale in two different periods or points of time. Multiplication of the quantity q by the price p of the commodity produced, transacted or sold gives the total money value pq of the production, transaction or sale. If instead of point of time, period of time is considered, a suitable average is to be taken and is to be expressed in the same units.

Notes

If p_0 and q_0 denote the price and the quantity of the commodity during the base period (0) and if p_n and q_n denote the corresponding price and quantity during a given period (n), then the total value $v_0 = p_0 q_0$ and $v_n = p_n q_n$, and the value relative of the period n w.r.t. the base period 0 is defined by

$$\text{value relative in percentage (of period } n \text{ w.r.t. period 0)} = \frac{v_n}{v_0} \times 100$$

$$= \frac{p_n q_n}{p_0 q_0} \times 100 \quad \ldots (3)$$

and is denoted by $v_{0|n}$.

Properties of Relatives

Let $p_a, p_b, p_c, \ldots$ denote the prices, in the periods $a, b, c, \ldots$ respectively, $q_a, q_b, q_c, \ldots$ denote the quantities and $v_a, v_b, v_c, \ldots$ the volumes for the corresponding periods.

The relatives satisfy some properties which are directly obtained from the definitions. We shall state the results for price relatives and write similar results for the quantity and value relatives :

1. Identity property :

   The price relative of a given period w.r.t. the same period is 1, that is

   $$p_{a/a} = 1.$$

2. Time reversal property :

   If the base period and the reference period are interchanged, then the product of the corresponding relatives is unity (one is the reciprocal of the other). That is :

   $$p_{a|b} \times p_{b|a} = 1$$

   or
   $$p_{b|a} = \frac{1}{p_{a|b}}$$

   Here

   $$p_{a|b} = \frac{p_b}{p_a} \text{ and } p_{b|a} = \frac{p_a}{p_b} \text{ and so the result follows.}$$

3. Circular or Cyclic property :

   We have

   $$p_{a|b} = \frac{p_b}{p_a}, \; p_{b|c} = \frac{p_c}{p_b}, \; p_{c|a} = \frac{p_a}{p_c}$$

   and so
   $$p_{a|b} \times p_{b|c} \times p_{c|a} = 1$$

   That is, if the periods *a, b, c,* are in cyclic order then the product of the three relatives w.r.t. the preceding period as base period is unity.

   This holds for *any* number of periods in cyclic order.

4. Modified Circular or Cyclic property :

   We have

   $$p_{a|b} \times p_{b|c} = \frac{p_b}{p_a} \times \frac{p_c}{p_b} = \frac{p_c}{p_a} = p_{a|c}$$

   More generally

   $$p_{a|b} \times p_{b|c} \times p_{c|d} = p_{a|d}.$$

Self-Assessment

Notes

1. Fill in the blanks :

(*i*) Index numbers are ..... averages.

(*ii*) Historically the first index was constructed in ..... .

(*iii*) Theoretically the best average in the construction of index number is ..... .

(*iv*) The index numbers are descriptive measures or ..... .

(*v*) Index numbers are ..... of economic activity.

15.4 Summary

- Index numbers have become today one of the most widely used statistical devices. Though originally developed for measuring the effect of change in prices, there is hardly any field today where index numbers are not used. Newspapers headline the fact that prices are going up or down, that industrial production is rising or falling, that imports are increasing or decreasing, that crimes are rising in a particular period compared to the previous period as disclosed by index numbers. They are used to feel the pulse of the economy and they have to be used as indicator of inflationary or deflationary tendencies. In fact, they are described as *barometers of economic activity, i.e.,* if one wants to get an idea as to what is happening to an economy he should look to important indices like the index number of industrial production, agricultural production, business activity, etc.
- to obtain an average, items must be comparable; for example, the average weight of men, women and children of a certain locality has no meaning at all. Furthermore, the unit of measurement must be the same for all the items. Thus an average of the weight expressed in *kg*., *lb*., etc., has no meaning. However, this is not so with index numbers. Index numbers are used for purposes of comparison in situations where two or more series are expressed in different units or the series are composed of different types of items.
- Index numbers are most widely used for measuring changes over a period of time. Thus we can find out the net change in agricultural prices from the beginning of first plan period to the Ninth plan period, *i.e.,* 1997-2002. Similarly, we can compare the agricultural production, industrial production, imports, exports, wages, etc., at two different times. However, it should be noted the index numbers not only measure changes over a period of time but also compare economic conditions of different locations, different industries, different cities or different countries. But since the basic problems are essentially the same and since most of the important index numbers published by the Government and private research organisations refer to data collected at different times, we shall consider in this chapter index numbers measuring changes relative to time only. However, methods described can be applied to other cases also.
- The trend of the phenomenon under study can be obtained by measuring the changes over a period of time. On the basis of this analysis can be done. For example, by examining the index numbers of industrial production, business activity etc. their trend can be understood and analysed.
- Index numbers are very useful in deflating *i.e.,* they are used to adjust the original data for price changes, or to adjust wages for cost of living changes and thus transform nominal wages into real wages. Moreover nominal income can be transformed into real income and nominal sales into real sales through appropriate index numbers.
- It may be concluded that index numbers are strong tools into the hands of the economists and business analysts with the help of which they can measure changes in certain phenomenon over a period of time or through a geographical location. This enables them to form suitable policies to obtain the desired results. Past, present and future trends and tendencies are revealed

Notes

with the help of index numbers and they are also useful in deflating. In the words of *Kafka* and *Simpson,* "Index numbers are today one of the most widely used statistical devices. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary or deflationary tendencies."

- Index numbers are constructed in the form of specialised averages with definite purpose and with some base period. Before constructing index numbers, it is necessary to know about the various problems which arise in its construction so that they can be minimised.
- Base period is the one against which comparisons are made. It may be year a month or a day. The index for base period is always taken to be 100. It is essential to choose an appropriate base before constructing index numbers.
- To study the change in a certain phenomenon, it is not possible to include each and every item leading to change. For example, while constructing price index, change in price of each and every item cannot be included. The selection of commodities should be such that they are representative of the tastes, preferences and habits etc. of the group of people regarding whom the index is constructed. In this way, it becomes a big problem as to which items to be included and which to be excluded ? Again, by having purpose of index properly defined, the selection of items can be eased. Index numbers would give false results if at one time one set of quantities are used and at other time other set of quantities are used.
- The term 'weight' refers to the relative importance of the different items in the construction of the index. All items are not of equal importance and hence it is necessary to device some suitable method which is done by allocating weights. In case of weighted indices, specific weights are assigned. Implicit or explicit methods of assigning weights can be used. Quantity or value weights can be assigned. If aggregative method is used, quantities are used as weights and in averaging of price relatives method of constructing index, value weights are used. Moreover, it has to be decided whether the weights would be fixed or fluctuating. In this way selection of appropriate weights forms a very crucial problem in constructing index.
- The change in the level of certain phenomenon can be measured only when proper data regarding quotations is available. For example, while preparing price index, it is essential to obtain proper price quotations of the selected commodities. Or while preparing expenditure index, information about expenditure should be available.
- One of the simplest types of index numbers is a *price relative.* It is the ratio of the price of a *single* commodity in a given period or point of time to its price in another period or point of time, called the *reference period* or *base period.* If prices for a period, instead of a point of time, are considered, then suitable price average for the period is taken and these prices are expressed in the same units.
- Simple type of index numbers is a quantity relative, when we are interested in changes in quantum or volumes of a commodity such as quantities of production or sale or consumption. Here the commodity is used in a more general sense. It may mean the volume of goods (in tonnes) carried by roadways, the number of passenger miles travelled, or the volume of export to or import from a country. In such cases we consider of quantity or volume relatives. If quantities or volumes are for a period instead of a point of time, a suitable average is to be taken and the quantities or volumes are to be expressed in the same units.
- A value relative is another type of simple index number, usable when we wish to compare changes in the money value of the transaction, consumption or sale in two different periods or points of time. Multiplication of the quantity q by the price p of the commodity produced, transacted or sold gives the total money value pq of the production, transaction or sale. If instead of point of time, period of time is considered, a suitable average is to be taken and is to be expressed in the same units.

Notes

15.5 Key-Words

1. Index Number : In economics and finance, an index is a statistical measure of changes in a representative group of individual data points. These data may be derived from any number of sources, including company performance, prices, productivity, and employment. Economic indices (index, plural) track economic health from different perspectives. Influential global financial indices such as the Global Dow, and the NASDAQ Composite track the performance of selected large and powerful companies in order to evaluate and predict economic trends.

15.6 Review Questions

1. What are index numbers ? How are they constructed ?
2. Define index numbers. Analyse the use of index numbers.
3. What are the various problems faced in construction of index numbers ?
4. "Index numbers are devices for measuring differences in the magnitude of a group of related variables". Discuss this statement and point out the use of index numbers.
5. What are the various types of index numbers.

Answers: Self-Assessment

1. (*i*) Specialised (*ii*) 1764
   (*iii*) Geometric mean (*iv*) Changes (*v*) Barometers

15.7 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 16: Methods—Simple (Unweighted) Aggregate Method and Weighted Aggregate Method

CONTENTS

Objectives

After reading this unit students will be able to:

- Discuss Simple (Unweighted) Aggregate Method.
- Explain Weighted Aggregate Method.

Introduction

A large number of formulae have been devised for constructing index numbers. Broadly speaking, they can be grouped under two heads:

(a) Unweighted indices, and (b) Weighted indices.

In the unweighted indices weights are not expressly assigned where again the weighted indices weights are assigned to the various items. Each of these types may further be divided under two heads:

(1) Simple Index Numbers

Simple Index Number is that Index number in which all the items are assigned equal importance. In other words, weights are not assigned to the different commodities and as such it is also called unweighted Index Number.

There are two methods of calculating Simple Index Number.

(i) Simple aggregate method

(ii) Simple average of price relative method.

(2) Weighted Index Numbers

In constructing simple index numbers, all commodities are given equal importance but in practice, all commodities don't have equal importance. For example, for a consumer, wheat is more important than vegetable or pulse. Similarly, clothes are more important than a video. To express the relative importance of different commodities, weights on some definite basis are used. When index numbers are constructed taking into consideration the importance of different commodities, then they are called weighted index numbers. There are two methods of constructing weighted index numbers:

(i) Weighted Aggregative Method

(ii) Weighted Average of Price Relative Method.

16.1 Simple (Unweighted) Aggregate Method

Notes

This is the simplest method of constructing Index Number. In this method *the total of current year prices for the various commodities is divided by the total of base year prices,* the resultant so obtained is multiplied by 100 to get the Index Numbers for the current year in terms of percentage.

Simbolically,

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

Where, P_{01} = Current year price Index Number based upon base year;

ΣP_1 = Sum total of current year prices; and ΣP_0 = Sum total of base year prices.

In Index Number 0 is used for base year and 1 is used for current year.

Example 1: Given the following data, and assuming 1991 as the base year, find out index value of the prices of different commodities for the year 1995.

Commodity	A	B	C	D	E
Prices in 1991 (Rs.)	50	40	10	5	2
Prices in 1995 (Rs.)	80	60	20	10	6

Solution: Construction of a Simple Index Number-Simple Aggregate Method

Commodities	**1991 (or Base Year) P_0 (Rs.)**	**1995 (or Current Year) P_1 (Rs.)**
A	50	80
B	40	60
C	10	20
D	5	10
E	2	6
Total	**ΣP_0 = 107**	**ΣP_1 = 176**

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100 = \frac{176}{107} \times 100 = 164.48$$

Thus, Price Index No. = 164.48

It means that prices, in general has increased by 64.48%.

Example 2: From the following data construct an index for 2005 taking 2004 as base.

Commodities	A	B	C	D	E
Prices in 2004 (Rs.)	50	40	80	110	20
Prices in 2005 (Rs.)	70	60	90	120	20

Solution: Construction of Price index

Commodities	**Prices in 2004 P_0 (Rs.)**	**Prices in 2005 P_1 (Rs.)**
A	50	70

Notes

B	40	60
C	80	90
D	110	120
E	20	20
Total	**$\Sigma P_0 = 300$**	**$\Sigma P_1 = 360$**

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

$$\Sigma P_1 = 360, \ \Sigma P_0 = 300$$

$$P_{01} = \frac{360}{300} \times 100 = 120$$

This means that as compared to 2004, in 2005 there is a net increase in the prices of commodities included in the index to the extent of 20%.

Merits and Demerits of Simple Aggregate method: Simple aggregative method of index number construction is very easy but it can be applied only when the prices of all commodities have been expressed in the same unit. If units are different, the results will be misleading:

Limitations of Simple Aggregate method: There are two main limitations of the simple aggregative index.

(i) In this type of index, the items with the large unit. Prices exert the greatest influence.

(ii) No consideration is given to the relative importance of the commodities.

16.2 Weighted Aggregative Index Numbers

These indices are of the simple aggregative type with the fundamental difference that weights are assigned to the various items included in the index. There are various methods of assigning weights and consequently a large number of formulae for constructing index numbers have been devised of which some of the more important ones are:

1. Laspayres method.
2. Paasche method.
3. Dorbish and Bowley method.
4. Fisher's ideal method, and
5. Marshall-Edgeworth method.

All these methods carry the name of persons who have suggested them.

1. Laspeyres Method: It is the most important of all types of index numbers. In this method the base year quantities are taken as weights. The formula for constructing the index is:

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

Steps:

(i) Multiply the current year prices of various commodities with base year weights and obtain $\Sigma p_1 q_0$.

(ii) Multiply the base year prices of various commodities with base year weights and obtain $\Sigma p_0 q_0$.

(iii) Divide $\Sigma p_1 q_0$ by $\Sigma p_0 q_0$ and multiply the quotient by 100. This gives us the price index.

Notes

Laspeyres Index attempts to answer the question: What is the change in aggregate value of the base period list of goods when valued at given period prices ? This index is very widely used in practical work.

2. **Paasche Method:** In this method the *current year* quantities are taken as weights. The formula for constructing the index is:

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

Steps:

(i) Multiply the current year prices of various commodities with base year weights and obtain $\Sigma p_1 q_1$.

(ii) Multiply the base year prices of various commodities with the base year weights and obtain $\Sigma p_0 q_1$.

(iii) Divide $\Sigma p_1 q_1$ by $\Sigma p_0 q_1$ and multiply the quotient by 100.

In general this formula answers the question: What would be the value of the given period list of goods when valued at base-period prices ?

Comparison of Laspeyres and Paasche Methods. From a practical point of view, Laspeyres index is often preferred to Paasche's for the simple reason that in Laspeyres index weights (q_0) are the base-year quantities and do not change from one year to the next. On the other hand, the use of Paasche index requires the continuous use of new quantity weights for each period considered and in most cases these weights are difficult and expensive to obtain.

An interesting property of Laspeyres and Paasche indices is that the former is generally expected to *overestimate* or to leave an upward bias, whereas the latter tends to *underestimate, i.e.*, shows a downward bias. When the prices increase there is usually a reduction in the consumption of those items for which the increase has been the most pronounced, and hence, by using base year quantities we will be giving too much weight to the prices that have increased the most and the numerator of the Laspeyres index will be too large. When the prices go down, consumers often shift their preference to those items which have declined the most and, hence, by using base-period weights in the numerator of the Laspeyres index we shall not be giving sufficient weights to the prices that have gone down the most and the numerator will again be too large. Similarly because people tend to spend less on goods when their prices are rising the use of the Paasche or current weighting produces an index which tends to underestimate the rise in prices, *i.e.*, it has a downward bias. But the above arguments do not imply that Laspeyres index must necessarily be larger than the Paasche's.

Unless drastic changes have taken place between the base year and the given year, the difference between the Laspeyre's and Paasche's will generally be small and either could serve as a satisfactory measure. In practice, however, the base year weighted Laspeyre's type index remains the most popular for reasons of its practicability. The Paasche type index can only be constructed when up-to-date data for the weights are available. Furthermore, the price index of a given year can be compared only with the base year. For example, let P_{82} =102, P_{83} = 130, and P_{84} = 145. Then P_{83} and P_{84} are using different weights and cannot be compared with each other.

If these indices had been obtained by the Laspeyre's formula they could be compared because in that case the weights are the same base-year weights (q_0). For these reasons, in practice the Paasche formula is usually not used and the Laspeyre type index remains most popular for reasons of its practicability.

3. **Dorbish and Bowley's Method:** Dorbish and Bowley have suggested simple arithmetic mean of the two indices (Laspeyres and Paasche) mentioned above so as to take into account the influence of both the periods, *i.e.*, current as well as base periods. The formula for constructing the index is:

Notes

$$P_{01} = \frac{L + P}{2}$$

where L = Laspeyre's Index

P = Paasche's Index

or

$$P_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100$$

4. **Fisher's Ideal Index:** Prof. Irving Fisher has given a number of formulae for constructing index numbers and of these he calls one as the 'ideal' index. The Fisher's Ideal Index is given by the formula:

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

or

$$P_{01} = \sqrt{L \times P}$$

It shall be clear from the above formula that Fisher's Ideal Index is the geometric mean of the Laspeyre and Paasche indices. Thus in the Fisher method we average geometrically formulae that are in opposite directions.

Did u know? Dorbish and Bowley have suggested simple arithmetic mean of the two indices (Laspeyres and Paasche) mentioned above so as to take into account the influence of both the periods, *i.e.*, current as well as base periods.

The above formula is known as 'Ideal' because of the following reasons:

(i) It is based on the geometric mean which is theoretically considered to be the best average for constructing index numbers.

(ii) It takes into account both current year as well as base year prices and quantities.

(iii) It satisfies both the time reversal test as well as the factor reversal test as suggested by Fisher.

(iv) It is free from bias. The two formulae (Laspeyre's and Paasche's) that embody the opposing type and weight biases are, in the ideal formula, crossed geometrically, *i.e.*, by the averaging process that of itself has no bias. The result is the complete cancellation of biases of the kinds revealed by time reversal and factor reversal tests.

It is not, however, a practical index to compute because it is excessively laborious. The data, particularly for the Paasche segment of index, are not really available. In practice, statisticians will continue to rely upon simple, although perhaps less exact, index number formulae.

5. **Marshall-Edgeworth Method:** In this method also both the current year as well as base year prices and quantities are considered. The formula for constructing the Index is:

$$P_{01} = \frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} \times 100$$

Notes

or, opening the brackets

$$P_{01} = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

It is a simple, readily constructed measure, giving a very close approximation to the results obtained by the ideal formula.

Example 3: Construct index numbers of price from the following data by applying (1) Laspeyre's method, (2) Passche method, (3) Bowley method, (4) Fisher's Ideal method, and (5) Marshall-Edgeworth method.

	2004		2005	
Commodities	**Price**	**Quantity**	**Price**	**Quantity**
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution:

Calculation of Various Indices

Commodities	**2004**		**2005**					
	Price p_0	**Quantity** q_0	**Price** p_1	**Quantity** q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	2	8	4	6	32	16	24	12
B	5	10	6	5	60	50	30	25
C	4	14	5	10	70	56	50	40
D	2	19	2	13	38	38	26	26
					Σp_1q_0 = 200	Σp_0q_0 = 160	Σp_1q_1 = 130	Σp_0q_1 = 103

1. Laspeyre's Method: $P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

$\Sigma p_1q_0 = 200,\ \Sigma p_0q_0 = 160$

$P_{01} = \frac{200}{160} \times 100 = 125$

2. Paasche Method: $P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$

$\Sigma p_1q_1 = 130,\ \Sigma p_0q_1 = 103$

$P_{01} = \frac{130}{103} \times 100 = 126.21$

3. Bowley's Method: $P_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100$

$= \frac{\frac{200}{160} + \frac{130}{103}}{2} \times 100$

Notes

$$= \frac{1.25 + 1.262}{2} \times 100 = \frac{2.512}{2} \times 100 = 125.6$$

or $$P_{01} = \frac{L + P}{2} = \frac{125 + 126.21}{2} = 125.61$$

4. Fisher's Ideal Method:

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{200}{160} \times \frac{130}{103}} \times 100$$

$$= \sqrt{1.578} \times 100 = 1.256 \times 100 = 125.6$$

5. Marshall-Edgeworth Method:

$$P_{01} = \frac{\Sigma (q_0 + q_1) p_1}{\Sigma (q_0 + q_1) p_0} \times 100$$

$$= \frac{200 + 130}{160 + 103} \times 100 = \frac{330}{263} \times 100 = 125.48.$$

Example 4: Using appropriate formula, construct Index Numbers for the year 1994 on the basis of year 1992 of the following data:

Year	Article 1		Article II		Article III	
	Price	Quantity	Price	Quantity	Price	Quantity
1992	5	10	8	6	6	3
1994	4	12	7	7	5	4

Solution: Since we are given price and quantity data for base as well as current year, the suitable index will be the fisher's Ideal Index.

	1992 Base Year		1994 Current Year					
Article	Price p_0	Quantity q_0	Price p_1	Quantity q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
I	5	10	4	12	50	60	40	48
II	8	6	7	7	48	56	42	49
III	6	3	5	4	18	24	15	20
Total					Σp_0q_0 = 116	Σp_0q_1 = 140	Σp_1q_0 = 97	Σp_1q_1 = 117

According to Fisher's Ideal Formula,

Index Number for 1994

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\frac{97}{116} \times \frac{117}{140}} \times 100$$

$$= \sqrt{0.6969} \times 100 = 83.6$$

Example 5: From the following data, calculate price index numbers for the current year by using.

(a) Paasche's Method

(b) Laspeyre's Method

Notes

Commodity	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	8	50	20	60
B	2	15	6	10
C	1	20	2	8
D	2	10	5	8
E	1	40	3	30

Solution:

Commodities	Base year		Current year					
	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	8	50	20	60	400	480	1000	1200
B	2	15	6	10	30	20	90	60
C	1	20	2	8	20	8	40	16
D	2	10	5	8	20	16	50	40
E	1	40	3	30	40	30	120	90
					510	554	1300	1406

(a) Paasche's Method

$$P_{01} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 = \frac{1406}{554} \times 100 = 253.79$$

(b) Laspeyre's Method

$$P_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = \frac{1300}{510} \times 100 = 254.90$$

Example 6: Calculate the weighted price Index from the following data:

Materials required	Unit	Quantity Required	1990 (Rs.)	1995 (Rs.)
Cement	100 Qtl.	500 *lb.*	5.0	8.0
Timber	*c.ft.*	2,000 *c.ft.*	9.5	14.2
Steel sheets	*c.wt.*	50 *c.wt.*	34.0	42.0
Bricks	per' 000	20,000	12.0	24.0

Solution: Since the weight are fixed, we apply Kelly's method for computing Index.

Notes

Calculation of Weighted Price Index

Material	Unit	Quantity	Price During 1990	Price During 1995		
		q	P_0	P_1	p_0q	p_1q
Cement	100 Qtl.	5	5.0	8.0	25	40
Timber	*c.ft.*	2,000	9.5	14.2	19,000	28,400
Steel Sheet	*c.wt.*	50	34.0	42.0	1,700	2,100
Bricks	per' 000	20	12.0	24.0	240	480
					$\sum p_0q$ = 20,965	$\sum p_1q$ = 31,020

$$P_{01} = \frac{\sum p_1q}{\sum p_0q} \times 100 ; \sum p_1q = 31,020; \quad \sum p_0q = 20,965$$

$$\therefore \quad P_{01} = \frac{31,020}{20,965} \times 100 = 147.96$$

Example 7: Construct index numbers of price from the following data by applying:

1. Laspeyres method,
2. Paasche method,
3. Bowley's method,
4. Fisher's Ideal method, and
5. Marshall-Edgeworth method.

Commodity	2006		2007	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution:

Commodity	2006		2007					
	Price p_0	Qty. q_0	Price p_1	Qty. q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	2	8	4	6	32	16	24	12
B	5	10	6	5	60	50	30	25
C	4	14	5	10	70	56	50	40
D	2	19	2	13	38	38	26	26
					$\sum p_1q_0$ = 200	$\sum p_0q_0$ = 160	$\sum p_1q_1$ = 130	$\sum p_0q_1$ = 103

Notes

1. Laspeyres Method: $P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$; where $\sum p_1 q_0 = 200$, $\sum p_0 q_0 = 160$

$$P_{01} = \frac{200}{160} \times 100 = 125$$

2. Paasche's Method: $P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$; where $\sum p_1 q_1 = 130$, $\sum p_0 q_1 = 103$

$$P_{01} = \frac{130}{103} \times 100 = 126.21$$

3. Bowley's Method: $P_{01} = \frac{\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}}{2} \times 100$

$$= \frac{\frac{200}{160} + \frac{130}{103}}{2} \times 100$$

$$= \frac{1.25 + 1.262}{2} \times 100 = \frac{2.512}{2} \times 100 = 125.6$$

$$P_{01} = \frac{L + P}{2} = \frac{125 + 126.2}{2} = 125.6$$

4. Fisher's Ideal Method:

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100 = \sqrt{\frac{200}{160} \times \frac{130}{103}} \times 100$$

$$= \sqrt{1.578} \times 100 = 1.2561 \times 100 = 125.61$$

5. Marshall-Edgeworth Method:

$$P_{01} = \frac{\sum (q_0 + q_1) p_1}{\sum (q_0 + q_1) p_0} \times 100 = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1}$$

$$= \frac{200 + 130}{160 + 103} \times 100 = \frac{330}{263} \times 100 = 125.48$$

Example 8: Compute Laspeyres, Paasche's, Fisher's and Marshall-Edgeworth's Index Numbers from the following data:

Item	Base Year		Current Year	
	Price (Rs.)	Quantity	Price (Rs.)	Quantity
A	5	25	6	30
B	3	8	4	10
C	2	10	3	8
D	10	4	3	5

Notes

Solution: Computation of Laspeyres, Paasche's, Fisher's Index

Item	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	5	25	6	30	150	125	180	150
B	3	8	4	10	32	24	40	30
C	2	10	3	8	30	20	24	16
D	10	4	3	5	12	40	15	50
					$\sum p_1q_0$ = 224	$\sum p_0q_0$ = 209	$\sum p_1q_1$ = 259	$\sum p_0q_1$ = 246

Laspeyres Index: $P_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = \frac{224}{209} \times 100 = 107.17$

Paasche's Index: $P_{01} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 = \frac{259}{246} \times 100 = 105.28$

Fisher's Index: $P_{01} = \sqrt{L \times P} = \sqrt{107.17 \times 105.28} = 106.22$

Marshall-Edgeworth's Index: $P_{01} = \frac{\sum p_1(q_0 + q_1)}{\sum p_0(q_0 + q_1)} \times 100 = \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \times 100$

$= \frac{224 + 259}{209 + 246} \times 100 = 106.15$

Example 9 : Prepare Index Number for 2010 on the basis of 2005, where the following information is given:

Year	Article I		Article II		Article III	
	Price	Quantity	Price	Quantity	Price	Quantity
2005	5	10	8	6	6	4
2010	4	12	7	7	5	3

Solution : Fisher's Ideal Index Number for 2010 — (Base 2005)

Article	2005 (Base Year)		2010		p_0q_0	p_1q_0	p_0q_1	p_1q_1
	Price p_0	Qnty. q_0	Price p_1	Qnty. q_1				
I	5	10	4	12	50	40	60	48
II	8	6	7	7	48	42	56	49
III	6	4	5	3	24	20	18	15
					122 ($\sum p_0q_0$)	102 ($\sum p_1q_0$)	134 ($\sum p_0q_1$)	112 ($\sum p_1q_1$)

Fisher's Ideal Index Number for 2010 **Notes**

or $$P_{01} = \sqrt{\frac{\sum p_1q_0 \times \sum p_1q_1}{\sum p_0q_0 \times \sum p_0q_1}} \times 100$$

$$= \sqrt{\frac{102}{122} \times \frac{112}{134}} \times 100$$

$$= \sqrt{.84 \times .84} \times 100 = .84 \times 100 = 84$$

Self-Assessment

1. Which of the following statements is True or False:
   (*i*) The base period should be a normal period.
   (*ii*) Unweighted indices are actually implicitly weighted.
   (*iii*) Bowley's index is the geometric mean of Laspeyre and Paasche Index.
   (*iv*) Marshall-Edgeworth's method satisfies the time reversal test.
   (*v*) In the Paasche's Price index, the weights are determined by quantities in the base period.

16.3 Summary

- *Simple Index Number is that Index number in which all the items are assigned equal importance.* In other words, weights are not assigned to the different commodities and as such it is also called unweighted Index Number.
- In constructing simple index numbers, all commodities are given equal importance but in practice, all commodities don't have equal importance. For example, for a consumer, wheat is more important than vegetable or pulse. Similarly, clothes are more important than a video. To express the relative importance of different commodities, weights on some definite basis are used. When index numbers are constructed taking into consideration the importance of different commodities, then they are called weighted index numbers.
- This is the simplest method of constructing Index Number. In this method *the total of current year prices for the various commodities is divided by the total of base year prices,* the resultant so obtained is multiplied by 100 to get the Index Numbers for the current year in terms of percentage.
- Simple aggregative method of index number construction is very easy but it can be applied only when the prices of all commodities have been expressed in the same unit.
- These indices are of the simple aggregative type with the fundamental difference that weights are assigned to the various items included in the index. There are various methods of assigning weights and consequently a large number of formulae for constructing index numbers have been devised of
- From a practical point of view, Laspeyres index is often preferred to Paasche's for the simple reason that in Laspeyres index weights (q_0) are the base-year quantities and do not change from one year to the next. On the other hand, the use of Paasche index requires the continuous use of new quantity weights for each period considered and in most cases these weights are difficult and expensive to obtain.
- When the prices increase there is usually a reduction in the consumption of those items for which the increase has been the most pronounced, and hence, by using base year quantities we will be giving too much weight to the prices that have increased the most and the numerator of the Laspeyres index will be too large.
- Unless drastic changes have taken place between the base year and the given year, the difference between the Laspeyre's and Paasche's will generally be small and either could serve as a satisfactory measure. In practice, however, the base year weighted Laspeyre's type index remains

Notes

the most popular for reasons of its practicability. The Paasche type index can only be constructed when up-to-date data for the weights are available. Furthermore, the price index of a given year can be compared only with the base year. For example, let P_{82} =102, P_{83} = 130, and P_{84} = 145. Then P_{83} and P_{84} are using different weights and cannot be compared with each other.

- Prof. Irving Fisher has given a number of formulae for constructing index numbers and of these he calls one as the 'ideal' index.
- It shall be clear from the above formula that Fisher's Ideal Index is the geometric mean of the Laspeyre and Paasche indices. Thus in the Fisher method we average geometrically formulae that are in opposite directions.
- A practical index to compute because it is excessively laborious. The data, particularly for the Paasche segment of index, are not really available. In practice, statisticians will continue to rely upon simple, although perhaps less exact, index number formulae.

16.4 Key-Words

1. Aggregate method : The term Aggregate Method refers the way price and volume data are handled when daily prices are gathered into weekly, monthly or even longer-term aggregate files. These settings are included in UA Preferences. Click "Aggregate Method" to see your current setting and adjust as desired.
2. Weighted index number : When all commodities are not of equal importance. We assign weight to each commodity relative to its importance and index number computed from these weights is called weighted index numbers.

16.5 Review Questions

1. Discuss the Simple aggregative method of constructing Price index number.
2. What are weighted index numbers ? Describe various types of weighted aggregative index numbers.
3. Why Fisher's formula for computing index numbers is said to be ideal ?
4. Distinguish between Laspeyre's and Paasche's index.
5. Explain the situations in which weighted and unweighted index numbers are useful.

Answers: Self-Assessment

1. (*i*) T (*ii*) T (*iii*) F (*iv*) T (*v*) F

16.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 17: Methods: Simple (Unweighted) Aggregate Method

CONTENTS

Objectives

After reading this unit students will be able to:

- Explain Simple Aggregate Method.

Introduction

Simple Index Number is that Index number is which all the items are assigned equal importance. In other words, weights are not assigned to the different commodities and as such it is also called unweighted Index Number.

There are two methods of calculating Simple Index Number.

- Simple aggregate method.
- Simple average of price relative method.

17.1 Simple Aggregate Method

This is the simplest method of constructing Index Number. In this method the total of current year prices for the various commodities is divided by the total of base year prices, the resultant so obtained is multiplied by 100 to get the Index Numbers for the current year in terms of percentage.

Simbolically,

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

Where, P_{01} = Current year pric Index Number based upon base year,

ΣP_1 = Sum total of current year prices; ΣP_0 = Sum total of base year prices.

In Index Number 0 is used for base year and 1 is used for current year.

Example 1: Given the following data, and assuming 1991 as the base year, find out index value of the prices of different commodities for the year 1995.

Commodity	A	B	C	D	E
Prices in 1991 (Rs.)	50	40	10	5	2
Prices in 1995 (Rs.)	80	60	20	10	6

Notes

Solution: Construction of a Simple Index Number-Simple Aggregate Method:

Commodities	1991 (or Base Year) P_0 (Rs.)	1995 (or Current Year) P_1 (Rs.)
A	50	80
B	40	60
C	10	20
D	5	10
E	2	6
Total	$\Sigma P_0 = 107$	$\Sigma P_1 = 176$

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100 = \frac{176}{107} \times 100 = 164.48$$

Thus, Price Index No. = 164.48

It means that prices, in general has inceased by 64.48%.

Example 2: Construct price index number for 1990 based on 1981 using Simple Agregative Method:

Commodities	Price in 1981 (in Rs.)	Price in 1990 (in Rs.)
A	50	80
B	40	60
C	10	20
D	5	10
E	2	8

Solution:

Construction of Price Index Number

Commodities	Price in 1981 (P_0)	Price in 1990 (P_1)
A	50	80
B	40	60
C	10	20
D	5	10
E	2	8
Total	$\Sigma P_0 = 107$	$\Sigma P_1 = 178$

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100 = \frac{176}{107} \times 100 = 166.48$$

Notes

Merits and Demerits of Simple Aggregate Method: Simple aggregative method of index number construction is very easy but it can be applied only when the prices of all commodities have been expressed in the same unit. If units are different, the retults will be misleading?

Example 3: Given the following data, and assuming 1991 as the base year, find out index value of the prices of different commodities for the year 1995.

Commodity	A	B	C	D	E
Prices in 1991 (Rs.)	50	40	10	5	2
Prices in 1995 (Rs.)	80	60	20	10	6

Solution: Construction of a Simple Index Number-Simple Aggregate Method

Commodities	**1991 (or Base Year) P_0 (Rs.)**	**1995 (or Current Year) P_1 (Rs.)**
A	50	80
B	40	60
C	10	20
D	5	10
E	2	6
Total	**ΣP_0 = 107**	**ΣP_1 = 176**

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100 = \frac{176}{107} \times 100 = 164.48$$

Thus, Price Index No. = 164.48

It means that prices, in general has increased by 64.48%.

Example 4: From the following data construct an index for 2005 taking 2004 as base.

Commodities	A	B	C	D	E
Prices in 2004 (Rs.)	50	40	80	110	20
Prices in 2005 (Rs.)	70	60	90	120	20

Solution: Construction of Price index

Commodities	**Prices in 2004 P_0 (Rs.)**	**Prices in 2005 P_1 (Rs.)**
A	50	70
B	40	60
C	80	90
D	110	120
E	20	20
Total	**ΣP_0 = 300**	**ΣP_1 = 360**

Notes

$$P_{01} = \frac{\Sigma P_1}{\Sigma P_0} \times 100$$

$$\Sigma P_1 = 360, \ \Sigma P_0 = 300$$

$$P_{01} = \frac{360}{300} \times 100 = 120$$

This means that as compared to 2004, in 2005 there is a net increase in the prices of commodities included in the index to the extent of 20%.

Merits and Demerits of Simple Aggregate method: Simple aggregative method of index number construction is very easy but it can be applied only when the prices of all commodities have been expressed in the same unit. If units are different, the results will be misleading:

Limitations of Simple Aggregate method: There are two main limitations of the simple aggregative index.

(i) In this type of index, the items with the large unit. Prices exert the greatest influence.

(ii) No consideration is given to the relative importance of the commodities.

Self-Assessment

1. Tick (√) the correct statements

(*i*) Simple index number is that number in which all the items are assigned equal.

(*ii*) In Lasreyers method, the base year quantities are taken as weights.

(*iii*) In pass the method the current year quantities are taken as weights.

(*iv*) In Marshall-Edgeworth Method both the current year as well as base year prices ae considered.

17.2 Summary

- Simple aggregative method of index number construction is very easy but it can be applied only when the prices of all commodities have been expressed in the same unit.
- Simple Index Number is that Index number is which all the items are assigned equal importance. In other words, weights are not assigned to the different commodities and as such it is also called unweighted Index Number.

  There are two methods of calculating Simple Index Number.

  (*i*) Simple aggregate method.

  (*ii*) Simple average of price relative method.
- This is the simplest method of constructing Index Number. In this method the total of current year prices for the various commodities is divided by the total of base year prices, the resultant so obtained is multiplied by 100 to get the Index Numbers for the current year in terms of percentage.

17.3 Key-Words

1. Index Number	:	Index Number is that Index number is which all the items are assigned equal importance. In other words, weights are not assigned to the different commodities and as such it is also called unweighted Index Number.
2. Price index	:	Index that tracks inflation by measuring price changes. Examples include the Consumer Price Index and the Producer Price Index.
3. Consumer price index (CPI)	:	A measure of changes in the purchasing-power of a currency and the rate of inflation. The consumer price index expresses the current prices of a basket of goods and services in terms of the

Notes

prices during the same period in a previous year, to show effect of inflation on purchasing power. It is one of the best known lagging indicators. See also producer price index.

4. Producer price index (PPI) : Relative measure of average change in price of a basket of representative goods and services sold by manufacturers and producers in the wholesale market. A family of three indices (finished goods, intermediate goods, and raw materials or crude commodities), it is used as an indicator of rate of inflation or deflation. In contrast to the consumer price index (CPI) which measures price changes from the consumer's perspective, PPI measures them from the seller's perspective. Older name wholesale price index.

17.4 Review Questions

1. What do you mean by Simple Index Number? Discuss its methods.
2. What is Simple aggregate method? Explain with examples.

Answers: Self-Assessment

1. (*i*) √ (*ii*) √ (*iii*) √ (*iv*) √

17.5 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 18: Methods—Simple Average of Price Relatives

CONTENTS

Objectives

After reading this unit students will be able to:

- Explain Simple Average of Price Relatives.
- Know the Merits and Limitations of Simple Average of Price Relatives Method.

Introduction

When this method is used to construct a price index first of all price relatives are obtained for the various items included in the index and then an average of these relatives is obtained using anyone of the measures of central value.

18.1 Simple Average of Price Relatives Method

In this method, we can use either Arithmetic Mean or Geometric Mean as the average of relatives.

(a) Using Arithmetic Mean: The arithmetic average has the advantage of simplicity but it is too much affected by the extreme values. It gives too much weight to increasing prices and little to decreasing ones. According to this method, we first find out price relative for each commodity and then take simple average of all price relatives. A price relative is the percentage ratio of the price of a variable in the current year to the price in the base year. Thus,

$$P_{01} = \frac{\Sigma\left(\frac{P_1}{P_0} \times 100\right)}{N}$$

$$\Rightarrow P_{01} = \frac{\Sigma P}{N}$$

[where, $\frac{P_1}{P_0} \times 100$ = Price Relative = P; N = Number of Commodities; P_1 = Current Year's Price; P_0 = Base Year's Price.]

Example 1: Given the following data and using the Price Relative method, construct an Index for the year 1993 in relation to 1983 price.

Notes

Commodities	Wheat (Per Qt.)	Ghee (Per *kg.*)	Milk (Per *kg.*)	Rice (Per Qt.)	Sugar (Per *kg.*)
1983 Prices (Rs.)	100	8	0.50	200	1
1993 Prices (Rs.)	200	40	4	800	6

Solution:

Construction of a Simple Index Number Average of Price Relative Method

Commodities	Base year 1983 Price P_0	1993 Price P_1	Price Relative of 1993 in relation to 1983 $P = \frac{P_1}{P_0} \times 100$
Wheat	100 (Per Qt.)	200 (Per Qt.)	$\frac{200}{100} \times 100 = 200$
Ghee	8 (Per *kg.*)	40 (Per *kg.*)	$\frac{40}{8} \times 100 = 500$
Milk	0.50 (Per *kg.*)	4 (Per *kg.*)	$\frac{4}{0.5} \times 100 = 800$
Rice	200 (Per Qt.)	800 (Per Qt.)	$\frac{800}{200} \times 100 = 400$
Sugar	1 (Per *kg.*)	6 (Per *kg.*)	$\frac{6}{1} \times 100 = 600$
N = 5			$\Sigma\left(\frac{P_1}{P_0} \times 100\right) = \Sigma P = 2500$

$$P_{01} = \frac{\Sigma\left(\frac{P_1}{P_0} \times 100\right)}{N} = \frac{\Sigma P}{N} = \frac{2500}{5} = 500$$

(b) Using Geometric Mean: *The Geometric Mean is used when items in a group are considered from the view point of their relative difference rather than that of their absolute difference*. For example, if the price of a commodity increases by 50% and that of another falls by 50%, the arithmetic average of relatives will neither rise nor fall implying that there has been no change in the price level. But in fact both the prices have changed. The Geometric Mean of relatives would in this case show that there has been a change in the price.

When Geometric mean is used, then the following formula is used:

$$\log P_{01} = \frac{\Sigma \log\left[\frac{P_1}{P_0} \times 100\right]}{N}, \text{ then}$$

Notes

$$P_{01} = \frac{\text{Antilog}\,\Sigma\log\left[\frac{P_1}{P_0}\times 100\right]}{N}$$

If $\left(\frac{P_1}{P_0}\times 100\right)$ is represented by P, then

$$P_{01} = \text{Antilog}\,\frac{\Sigma\log P}{N}$$

The following example will illustrate the application of above rules.

Example 2: In the above example 2, Calculate Index Number using Geometric Mean as Average of Relatives.

Solution:

Commodities	1983 Base Year Price P_0	1993 Current Year Price P_1	$P = \frac{P_1}{P_0}\times 100$	log P
Wheat	100	200	$\frac{200}{100}\times 100 = 200$	2.3010
Ghee	8	40	$\frac{40}{8}\times 100 = 500$	2.6990
Milk	0.50	4	$\frac{4}{0.5}\times 100 = 800$	2.9031
Rice	200	800	$\frac{800}{200}\times 100 = 400$	2.6021
Sugar	1	6	$\frac{6}{1}\times 100 = 600$	2.7782
				$\Sigma\log P = 13.2834$

$$\log P_{01} = \frac{\Sigma\log P}{N}$$

$$\log P_{01} = \frac{13.2834}{5}$$

$$\log P_{01} = \text{Antilog}\,[2.6567]$$

$$P_{01} = 453.63$$

Example 3: From the following data construct an Index for 2005 taking 2004 as base by the average of relatives methods using (a) arithmetic mean, and (b) geometric mean for averaging relatives.

Notes

Commodities	Price in 2004 (Rs.)	Price in 2005 (Rs.)
A	50	70
B	40	60
C	80	90
D	100	120
E	20	20

Solution:

(a) INDEX NUMBER USING ARITHMETIC MEAN OF PRICE RELATIVES

Commodities	Price in 2004 (Rs.) P_0	Price in 2005 (Rs.) P_1	Price $\frac{P_1}{P_0}\times 100$
A	50	70	140.0
B	40	60	150.0
C	80	90	112.5
D	110	120	109.1
E	20	20	100.0
			$\Sigma\frac{P_1}{P_0}\times 100 = 611.6$

$$P_{01} = \frac{\Sigma\frac{P_1}{P_0}\times 100}{N} = \frac{611.6}{5} = 122.32$$

(b) INDEX NUMBR USING GEOMETRIC MEAN OF PRICE RELATIVES

Commodities	Price in 2004 P_0	Price in 2005 P_1	Price Relatives P	Log P
A	50	70	140.0	2.1461
B	40	60	150.0	2.1761
C	80	90	112.5	2.0512
D	110	120	109.1	2.0378
E	20	20	100.0	2.0000
				log P = 10.4112

$$P_{01} = \text{Antilog}\left[\frac{\Sigma\log P}{N}\right]$$

$$= \text{Antilog}\left[\frac{10.4112}{5}\right] = \text{Antilog } 2.0822 = 120.9$$

Notes

Although arithmetic mean and geometric mean have both been used, the arithmetic mean is often preferred because it is easier to compute and much better known. Some economists, notably F.Y. Edgeworth, have preferred to use the median which is not affected by a single extreme value. Since the argument is important only when an index is based on a very small number of commodities, it generally does not carry much weight and the median is seldom used in actual practice.

18.2 Merits and Limitations of Simple Average of Price Relative Method

Merits

This method has the following two advantages over the previous method:

1. Extreme items do not influence the index. Equal importance is given to all the items.
2. The index is not influenced by the units in which prices are quoted or by the absolute level of individual prices. Relatives are pure numbers and are, therefore, divorced from the original units. Consequently, index numbers computed by the relatives method would be the same regardless of the way in which prices are quoted. This simple average of price relatives is said to meet what is called the *units test*.

Limitations

Despite these merits this method is not very satisfactory because of two reasons:

1. Difficulty is faced with regard to the selection of an appropriate average. The use of the arithmetic mean is considered as questionable sometimes because it has an upward bias. The use of geometric mean involves difficulties of computation. Other averages are almost never used while constructing index numbers.
2. The relatives are assumed to have equal importance. This is again a kind of concealed weighting system that is highly objectionable since economically same relatives are more important than others.

Self-Assessment

1. Fill in the Blanks:

(*i*) Theoretically the best average in the construction of index number is.

(*ii*) A price relative is the percentage ratio of the price of a variable in the ..... year to the price in the ..... year.

(*iii*) Weighted average of relatives can be combined to form a new ..... .

(*iv*) The index is not influenced by the unit in which ..... are quoted.

(*v*) The average may be arithmetic mean, median, mode or ..... .

18.3 Summary

- The arithmetic average has the advantage of simplicity but it is too much affected by the extreme values. It gives too much weight to increasing prices and little to decreasing ones. According to this method, we first find out price relative for each commodity and then take simple average of all price relatives. A price relative is the percentage ratio of the price of a variable in the current year to the price in the base year.
- *The Geometric Mean is used when items in a group are considered from the view point of their relative difference rather than that of their absolute difference*. For example, if the price of a commodity

Notes

increases by 50% and that of another falls by 50%, the arithmetic average of relatives will neither rise nor fall implying that there has been no change in the price level. But in fact both the prices have changed. The Geometric Mean of relatives would in this case show that there has been a change in the price.

- Although arithmetic mean and geometric mean have both been used, the arithmetic mean is often preferred because it is easier to compute and much better known. Some economists, notably F.Y. Edgeworth, have preferred to use the median which is not affected by a single extreme value. Since the argument is important only when an index is based on a very small number of commodities, it generally does not carry much weight and the median is seldom used in actual practice.
- The index is not influenced by the units in which prices are quoted or by the absolute level of individual prices. Relatives are pure numbers and are, therefore, divorced from the original units. Consequently, index numbers computed by the relatives method would be the same regardless of the way in which prices are quoted. This simple average of price relatives is said to meet what is called the *units test*.
- Difficulty is faced with regard to the selection of an appropriate average. The use of the arithmetic mean is considered as questionable sometimes because it has an upward bias. The use of geometric mean involves difficulties of computation. Other averages are almost never used while constructing index numbers.
- The relatives are assumed to have equal importance. This is again a kind of concealed weighting system that is highly objectionable since economically same relatives are more important than others.

18.4 Key-Words

1. Arithmetic mean : In mathematics and statistics, the arithmetic mean, or simply the mean or average when the context is clear, is the central tendency of a collection of numbers taken as the sum of the numbers divided by the size of the collection. The collection is often the sample space of an experiment. The term "arithmetic mean" is preferred in mathematics and statistics because it helps distinguish it from other means such as the geometric and harmonic mean.
2. Geometric mean : In mathematics, the geometric mean is a type of mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the nth root (where n is the count of numbers) of the product of the numbers.

18.5 Review Questions

1. Discuss steps of simple average of price relative method of constructing index numbers.
2. What are the merits and limitations of simple average of price relative method.
3. Explain the role of weights in the construction of general price index numbers.
4. What is simple average of price relative method of constructing index numbers ? Explain by using arithmetic mean.
5. What is simple average of price relative method of constructing index numbers? Explain by using geometric mean.

Notes

Answers: Self-Assessment

1. (*i*) Geometric mean (*ii*) Current, base (*iii*) Index
   (*iv*) Prices (*v*) Geometric mean

18.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Unit 19: Methods—Weighted Average of Price Relatives

Notes

CONTENTS

Objectives

After reading this unit students will be able to:

- Describe Weighted Average of Price Relatives.
- Explain Quantity Index Number.

Introduction

In the weighted aggregative methods discussed earlier price relatives were not computed. However, like unweighted relative method it is also possible to compute weighted average of relatives. For the purpose of averaging we may use either the arithmetic mean or the geometric mean.

19.1 Weighted Average of Price Relative Method

In order to compute index number by Weighted Average of Relatives Method, following steps are necessarily be taken: (1) Express each item of the period for which the index number is being calculated as a percentage of the same item in the base period. (2) Multiply the percentage as obtained in step (1) for each item by the weight that has been assigned to that item. (3) Add the results obtained in step (2), (4) Divide the sum obtained in step (3) by the sum of weights used to obtain the index number.

When arithmetic mean is used,

$$P_{01} = \frac{\sum PV}{\sum V}$$

where P is price relative $\frac{p_1}{p_0} \times 100$ and V is value weights $p_0 q_0$.

When geometric mean is used,

$$P_{01} = \text{Antilog}\left[\frac{\sum V \log P}{\sum V}\right]$$

where $P = \frac{p_1}{p_0} \times 100$, $V = p_0 q_0$

Notes

Example 1: From the following data compute price index by applying weighted average of Price relative method using:

(a) arithmetic mean, and

(b) geometric mean.

Commodities	p_0 Rs.	q_0	p_1 Rs.
Sugar	6.0	10 *kg.*	8.0
Rice	3.0	20 *kg.*	3.2
Milk	2.0	5 *lt.*	3.0

Solution:

(a) Index number using weighted arithmetic mean of Price Relatives

Commodities	p_0	q_0	p_1	$p_0 q_0$ V	$\frac{p_1}{p_0} \times 100$ P	PV
Sugar	Rs. 6.0	10 *kg.*	Rs. 8.0	60	$\frac{8}{6} \times 100$	8,000
Rice	Rs. 3.0	20 *kg.*	Rs. 3.2	60	$\frac{3.2}{3} \times 100$	6,400
Milk	Rs. 2.0	5 *lt.*	Rs. 3.0	10	$\frac{3}{2} \times 100$	1,500
				$\Sigma V = 130$		$\Sigma PV = 15,900$

$$P_{01} = \frac{\Sigma PV}{\Sigma V} = \frac{15,900}{1300}$$

$$= 122.31.$$

This means that there has been a 22.3 percent increase in prices over the base level.

(b) Index Number using Geometric mean of Price Relatives

Commodities	p_0	q_0	p_1	V	P	Log P	V. Log P
Sugar	Rs. 6.0	10 *kg.*	Rs. 8.0	60	133.3	2.1249	127.494
Rice	Rs. 3.0	20 *kg.*	Rs. 3.2	60	106.7	2.0282	121.692
Milk	Rs. 2.0	5 *lt.*	Rs. 3.0	10	150.0	2.1761	21.761
				$\Sigma V = 130$			$\Sigma V. \log P = 270.947$

$$P_{01} = \text{Antilog}\left[\frac{\Sigma V. \log P}{\Sigma V}\right]$$

$$= \text{Antilog}\left[\frac{270.947}{130}\right] = \text{Antilog } 2.084 = 121.3.$$

Notes

The result obtained by applying the Laspeyre's method would come out to be the same as obtained by weighted arithmetic mean of price relative method (as shown below):

PRICE INDEX BY LASPEYRE'S METHOD

Commodities	p_0	q_0	p_1	p_1q_0	p_0q_0
Sugar	Rs. 6.0	10 *kg.*	Rs. 8.0	80	60
Rice	Rs. 3.0	20 *kg.*	Rs. 3.2	64	60
Milk	Rs. 2.0	5 *lt.*	Rs. 3.0	15	10
				$\sum p_1q_0 = 159$	$\sum p_0q_0 = 130$

$$P_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100 = \frac{159}{130} \times 100 = 122.31$$

The answer is the same as that obtained by weighted arithmetic mean of price relatives method.

Merits of Weighted Average of Relative Indices

The following are the special advantages of weighted average of relative indices over weighted aggregative indices:

(1) When different index numbers are constructed by the average of relatives method, all of which have the same base, they can be combined to form a new index.

(2) When an index is computed by selecting one item from each of the many sub-groups of items, the values of each sub-group may be used as weights. Then only the method of weighted average of relatives is appropriate.

(3) When a new commodity is introduced to replace the one formerly used, the relative for the new item may be spliced to the relative for the old one, using the former value weights.

(4) The price or quantity relatives for each single item in the aggregate are, in effect, themselves a simple index that often yields valuable information for analysis.

Did u know? Price index numbers measure and permit comparison of the price of certain goods; quantity index numbers, on the other hand, measure and permit comparison of the physical volume of goods produced or distributed or consumed.

19.2 Quantity Index Numbers

Price index numbers measure and permit comparison of the price of certain goods; quantity index numbers, on the other hand, measure and permit comparison of the physical volume of goods produced or distributed or consumed. Though price indices are more widely used, production indices are highly significant as indicators of the level of output in the economy or in parts of it.

In constructing quantity index numbers, the problems confronting the statistician are analogous to those involved in price indices. We measure changes in quantities, and when we weigh we use prices or values as weights. Quantity indices can be obtained easily by changing p to q and q to p in the various formulae discussed above.

Thus when Laspeyre's method is used

$$Q_{01} = \frac{\sum q_1p_0}{\sum q_0p_0} \times 100$$

Notes

When Paasche's formula is used

$$Q_{01} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

When Fisher's formula is used

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$$

These formulae represent the *quantity index* in which the quantities of the different commodities are weighted by their prices. However, any other suitable weights can be used instead.

Example 2: Compute by suitable method the index number of quantity from the data given below:

Commodities	2004		2005	
	Price	Total Value	Price	Total Value
A	8	80	10	110
B	10	90	12	108
C	16	256	20	340

Solution : Since we are given the value and the price we can obtain quantity figure by dividing value by price for each commodity. We can then apply Fisher's method for finding out quantity index.

COMPUTATION OF QUANTITY INDEX BY FISHER'S METHOD

Commodities	2004		2005					
	p_0	q_0	p_1	q_1	q_1p_0	q_0p_0	q_1p_1	q_0p_1
A	8	10	10	11	88	80	110	100
B	10	9	12	9	90	90	108	108
C	16	16	20	17	272	256	340	320
					$\sum q_1p_0$ = 450	$\sum q_0p_0$ = 426	$\sum q_1p_1$ = 558	$\sum q_0p_1$ = 528

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$$

$$Q_{01} = \sqrt{\frac{450}{426} \times \frac{558}{528}} \times 100$$

$$= \sqrt{1.116} \times 100 = 1.0564 \times 100$$

$$= 105.64.$$

Example 3: Compute Price index by applying weighted average of price relatives:

Commodities	p_0	q_0	p_1
Sugar	10	6 *kg.*	15
Rice	20	10 *kg.*	25
Milk	10	8 *kg.*	14

Solution: Computing price index by applying weighted average of Price Relatives.

Notes

Commodities	p_0	q_0	p_1	p_1q_0	p_0q_0
Sugar	10	6	15	90	60
Rice	20	10	25	250	200
Milk	10	8	14	112	80
				$\sum p_1q_0 = 452$	$\sum p_0q_0 = 340$

$$p_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100$$

$$= \frac{452}{340} \times 100$$

$$= 132.94$$

Example 4: Using the following data construct in index for 2010 taking 2009 as base by the average of relatives method using arithmetic and geometric mean:

Commodity	Price in 2009	Price in 2010
A	500	700
B	400	600
C	800	900
D	110	120
E	20	20

Solution:

Commodity	p_0	p_1	$p_1/p_0 \times 100$	$\log p_1/p_0 \times 100$
A	500	700	140	2.1461
B	400	600	150	2.1761
C	800	900	112.5	2.0512
D	110	120	109.1	2.0378
E	20	20	100	2.0000
				$\frac{\sum p_1}{p_0} \times 100 = 611.6$

$$\sum \log \frac{p_1}{p_0} \times 100 = 10.4112$$

Using arithmetic mean, $P_{01} = \frac{\sum \left(\frac{p_1}{p_0} \times 100 \right)}{N} = \frac{611.6}{5} = 122.32.$

Using geometric mean,

Notes

$$P_{01} = \text{antilog}\left[\frac{\Sigma \log \frac{p_1}{p_0} \times 100}{N}\right] = \text{antilog}\left[\frac{10.4112}{5}\right].$$

$$= \text{antilog } 2.0822$$

$$P_{01} = 120.9.$$

Example 5: From the following data construct a price index number of the group of four commodities using the appropriate formula:

Commodity	Base Year		Current Year	
	Price per unit	Expenditure (Rs.)	Price per unit	Expenditure (Rs.)
A	2	40	5	75
B	4	16	8	40
C	1	10	2	24
D	5	25	10	60

Solution: Since we are given base year and current year price and expenditure fishers ideal formula is appropriate for index.

Commodity	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	2	20	5	15	100	40	75	30
B	4	4	8	5	32	16	40	20
C	1	10	2	12	20	10	24	12
D	5	5	10	6	50	25	60	30
					Σp_1q_0 = 202	Σp_0q_0 = 91	Σp_1q_1 = 199	Σp_0q_1 = 92

Quantity q is calculated by the following method:

$$q = \frac{\text{Expenditure}}{\text{Price per unit}}$$

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \cdot \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

$$= \sqrt{\frac{202}{91} \times \frac{199}{92}} \times 100$$

$$= 2.1912 \times 100 = 219.12.$$

Example 6: From the following data, compute price index by supplying weighted average of price relatives method using: (a) arithmetic mean, (b) geometric mean.

Commodity	p_0 (Rs.)	q_0	p_1 (Rs.)
A	3.0	20 *kg.*	4.0
B	1.5	40 *kg.*	1.6
C	1.0	10 *lt.*	1.5

Solution: **(a) Index number using weighted arithmetic mean of price relatives** **Notes**

Commodity	p_0	q_0	p_1	$V = p_0q_0$	$P = p_1/p_0 \times 100$	PV
A	3	20 *kg.*	4.0	60	4/3 × 100	8,000
B	1.5	40 *kg.*	1.6	60	1.6/1.5 × 100	6,400
C	1.0	10 *lt.*	1.5	10	1.5/1 × 100	1,500
				$\Sigma V = 130$		$\Sigma PV = 15,900$

$$P_{01} = \frac{\Sigma PV}{\Sigma V} = \frac{15,900}{130} = 122.31.$$

This means that there is a 122.31% increase in price over base year.

(b) Index number using geometric mean of price relatives.

Commodity	p_0	q_0	p_1	$V = p_0q_0$	$P = \frac{p_1}{p_0} \times 100$	log P	V log P
A	3	20 *kg.*	4.0	60	133.33	2.1249	127.404
B	1.5	40 *kg.*	1.6	60	106.7	2.0282	121.602
C	1.0	10 *lt.*	1.5	10	150.0	2.1761	21.761
				$\Sigma V = 130$			$\Sigma V \log P = 270.947$

$$P_{01} = \text{Antilog}\left[\frac{\Sigma V \cdot \log P}{\Sigma V}\right].$$

$$= \text{Antilog}\left[\frac{270.947}{130}\right].$$

= Antilog 2.084 = 120.9.

Self-Assessment

1. Fill in the blanks:

(*i*) Laepeyre's index is based on ..... .

(*ii*) Fisher's ideal index is ..... .

(*iii*) If with a rise of 10% in prices the wages are increased by 20%, the real wage increase is by ..... .

(*iv*) ..... index is known as the 'Ideal' formula for constructing index numbers.

(*v*) The reference period is the period against which ..... are made.

19.3 Summary

- In the weighted aggregative methods discussed earlier price relatives were not computed. However, like unweighted relative method it is also possible to compute weighted average of relatives. For the purpose of averaging we may use either the arithmetic mean or the geometric mean.

Notes

- When an index is computed by selecting one item from each of the many sub-groups of items, the values of each sub-group may be used as weights. Then only the method of weighted average of relatives is appropriate.
- The price or quantity relatives for each single item in the aggregate are, in effect, themselves a simple index that often yields valuable information for analysis.
- Though price indices are more widely used, production indices are highly significant as indicators of the level of output in the economy or in parts of it.
- In constructing quantity index numbers, the problems confronting the statistician are analogous to those involved in price indices. We measure changes in quantities, and when we weigh we use prices or values as weights. Quantity indices can be obtained easily by changing p to q and q to p in the various formulae discussed above.

19.4 Key-Words

1. Relative method : Index numbers measure changes or differences and are used in a variety of contexts. The Office for National Statistics (ONS) produces index numbers principally in the field of economics. Economists are interested in how changes in the monetary value of economic transactions can be attributed to changes in price (to measure inflation) and changes in quantity (to measure sales volume or economic output). Index numbers typically measure these changes over time. However, index numbers can also be used to make other comparisons, such as between regions of the UK.

19.5 Review Questions

1. Describe weighted Average of Price Relatives method to compute index numbers.
2. What are quantity or volume index numbers ?
3. What are the merits of weighted average of price relative method.
4. What is weighted average of price relative. Method of compute index number? Explain by using weighted arithmetic mean of Price Relatives.
5. Explain weighted average of Price Relative Method by using Geometric mean of Price Relatives.

Answers: Self-Assessment

1. (*i*) base year quantities (*ii*) geometric mean (*iii*) less than 10%
   (*iv*) Fisher's ideal index (*v*) comparisons

19.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Unit 20: Test of Consistency: Unit Test, Time Reversal Test, Factor Reversal Test and Circular Test

Notes

CONTENTS

Objectives

After reading this unit students will be able to:

- Define Unit Test and Time Reversal Test.
- Know Factor Reversal Test and Circular Test.

Introduction

We have read in previous unit 19 that the relatives have certain important properties. What is true for an individual commodity should also be true for a group of commodities. The index number as an aggregative relative should also satisfy the same set of properties.

A number of mathematical criteria for judging the adequacy of an Index Number formula have been developed by statisticians. In fact, the problem is that of selecting the most appropriate one in a given situation. The following test are suggested for selecting an appropriate index.

(i) Unit Test

(ii) Time Reversal Test

(iii) Factor Reversal Test

(iv) Circular Test.

20.1 Unit Test

This test requires that the formula for constructing an index should be independent of the units in which the prices are quoted. All formulae of weighted aggregate method except simple aggregative method satisfy this test.

20.2 Time Reversal Test

Prof. Fisher has stated Time Reversal Test. 'The test is that the formula for calculating an Index Number should be such that will give the same ratio between one point of comparison and the other, no matter which of the two is taken as base. Time Reversal means that if we change the base year to the current year and *vice versa* then the product of the indices should be equal to unity. In other

Notes

words, Simple Aggregative Method does not satisfy this test. The index number reckoned forward should be the reciprocal of that reckoned backward'. Thus, an ideal Index Number formula should work both ways, *i.e.*, forward as well as backward.

Mathematically, the following relation should be satisfied:

$$\frac{P_{01}}{100} \times \frac{P_{10}}{100} = 1 \text{ or } P_{01} \times P_{10} = 1$$

Laspeyre's Method

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \text{ and } P_{10} = \frac{\sum p_0 q_1}{\sum p_1 q_1}; \; P_{01} \times P_{10} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \neq 1$$

$\because$ $P_{01} \times P_{10} \neq 1$ hence, test is not satisfied.

Paasche's Method

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \text{ and } P_{10} = \frac{\sum p_0 q_0}{\sum p_1 q_0}$$

Here, also $P_{01} \times P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0} \neq 1$, hence, test is not satisfied.

The test is not satisfied by Laspeyre's and the Paasche's method. However, Fisher's Method satisfies the test.

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \text{ and } P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}}$$

i.e., $P_{01} \times P_{10} = \sqrt{1}$

or $P_{01} \times P_{10} = 1$

Since, $P_{01} \times P_{10} = 1$, the Fisher's Ideal Index satisfies the test.

20.3 Factor Reversal Test

This is another test suggested by Fisher. According to Fisher Just as our formula should permit the interchange of two factors without giving inconsistent results, so without to permit interchange of prices and quantities without giving inconsistent results, *i.e.*, the two results multiplied together should give the true value ratio.

Mathematically,

$$\frac{P_{01}}{100} \times \frac{Q_{01}}{100} = \frac{\sum p_1 q_1}{\sum p_0 q_0} \text{ or } P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Laspeyre's & Paasche's method does not satisfy this test also like the above test.

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \text{ and } Q_{01} = \frac{\sum q_1 p_0}{\sum q_0 p_0}$$

Notes

Here $P_{01} \times Q_{01} \neq \frac{\sum p_1 q_1}{\sum p_0 q_0}$

Paasche's Method

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \text{ and } Q_{01} = \frac{\sum q_1 p_1}{\sum q_0 p_1}$$

Here also $P_{01} \times Q_{01} \neq \frac{\sum p_1 q_1}{\sum p_0 q_0}$

This test is also satisfied by Fisher's Method only.

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}}$$

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum q_1 p_0}{\sum p_0 q_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

$$\sqrt{\left(\frac{\sum p_1 q_1}{\sum p_0 q_0}\right)^2} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \text{Value Index}$$

Hence, Fisher's Formula satisfies this test.

20.4 Circular Test

Another test of adequacy is circular test. This test is an extension of the time reversal test. It requires that if an index is constructed for the year '*b*' on year '*a*' and for the year '*c*' on base year '*b*' we should get the same result as if we calculate directly an Index for '*c*' on base year '*a*' without going through *b* as an intermediary.

Mathematically,

Let year *a*, *b* and *c* are denoted by 0, 1 and 2 respectively.

Condition of the test:

$$P_{01} \times P_{12} \times P_{20} = 1$$

For example, suppose prices have doubled in year 1 as compared to year 0, and again prices have doubled in year 2 as compared to year 1, in such a case if we correlated the prices of year 2 and 0 then we will find that prices of year 0 were $\frac{1}{4}$ of the prices of year 2.

Mathematically,

$$P_{01} = 2;\ P_{12} = 2;\ P_{20} = \frac{1}{4}$$

Here, $P_{01} \times P_{12} \times P_{20} = 2 \times 2 \times \frac{1}{4} = 1$

Notes

The test is not satisfied by any of the Laspeyre's, Paasche's or Fisher's Method. However, the simple aggregative method and the fixed weight aggregative method (Kelly's method) satisfy this test.

Example 1: Compute Fisher's ideal number and prove that it satisfies factor reversal and time reversal test.

Commodity	Price		Quantity	
	2002	2003	2002	2003
A	10	12	12	15
B	7	5	15	20
C	5	9	24	20
D	16	14	5	5

Solution:

Commodity Item	Price		Quantity		$p_n q_0$	$p_0 q_0$	$p_n q_n$	$p_0 q_n$
	2002	2003	2002	2003				
A	10	12	12	15	144	120	180	150
B	7	5	15	20	75	105	100	140
C	5	9	24	20	216	120	180	100
D	16	14	5	5	70	80	70	80
				Total	$\sum p_n q_0$	$\sum p_0 q_0$	$\sum p_n q_n$	$\sum p_0 q_n$
					505	425	530	470

Fisher's price index

$$F_p = P_{on}^F = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum p_n q_n}{\sum p_0 q_n}} \times 100$$

$$= \sqrt{\frac{505}{425} \times \frac{530}{470}} \times 100$$

$$= \sqrt{1.188 \times 1.128} \times 100$$

$$= 115.8$$

(i) Time reversal test

$$P_{no} = \sqrt{\frac{\sum p_0 q_n}{\sum p_n q_n} \cdot \frac{\sum p_0 q_0}{\sum p_n q_0}}$$

$$= \sqrt{\frac{470}{530} \times \frac{525}{505}}$$

$$\therefore \quad P_{no} \times P_{on} = \sqrt{\frac{505}{425} \cdot \frac{530}{470} \cdot \frac{470}{530} \cdot \frac{429}{505}}$$

$$= \sqrt{1} = 1$$

(ii) Factor reversal test

Notes

$$P_{on} = \sqrt{\frac{505}{425} \times \frac{530}{470}}$$

$$Q_{on} = \sqrt{\frac{470}{425} \times \frac{530}{505}}$$

$$P_{no} \times P_{on} = \frac{\sum p_n q_n}{\sum p_0 q_0}$$

Hence, the data satisfies both time reversal and factor reversal tests.

Example 2: The following figures relate to the prices and quantities of certain commodities. Construct an appropriate index number and find out whether it satisfies the time reversal test.

Commodities	2004		2005	
	Price	Quantity	Price	Quantity
A	30	50	32	50
B	25	40	30	35
C	18	50	16	55

Solution:

INDEX NUMBER BY FISHER'S IDEAL METHOD

Commodities	2004		2005					
	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	30	50	32	50	1,600	1,500	1,600	1,500
B	25	40	30	35	1,200	1,000	1,050	875
C	18	50	16	55	800	900	880	990
					$\sum p_1q_0$ = 3,600	$\sum p_0q_0$ = 3,400	$\sum p_1q_1$ = 3,530	$\sum p_0q_1$ = 3,365

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{3,600}{3,400} \times \frac{3,530}{3,365}} \times 100 = \sqrt{1.111} \times 100 = 1.054 \times 100 = 105.4$$

Time reversal test is satisfied when

$$P_{01} \times P_{10} = 1$$

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} = \sqrt{\frac{3,600}{3,400} \times \frac{3,530}{3,365}}$$

$$P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} = \sqrt{\frac{3,365}{3,530} \times \frac{3,400}{3,600}}$$

Notes

$$P_{01} \times P_{10} = \sqrt{\frac{3,600}{3,400} \times \frac{3,530}{3,365} \times \frac{3,365}{3,530} \times \frac{3,400}{3,600}} = \sqrt{1} = 1.$$

Hence time reversal test is satisfied by the above formula.

Example 3: From the following data calculate Fisher's ideal index and prove that it satisfies both the time reversal and factor reversal tests.

Commodity	2004		2005	
	Price	Qty.	Price	Qty.
A	4	8	5	8
B	5	10	6	12
C	3	6	4	7
D	8	5	10	4

Solution:

Calculation of Fishers's Ideal Index

Commodity	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	4	8	5	8	40	32	40	32
B	5	10	6	12	60	50	72	60
C	3	6	4	7	24	18	28	21
D	8	5	10	4	50	40	40	32
					Σp_1q_0 = 174	Σp_0q_0 = 140	Σp_1q_1 = 180	Σp_0q_1 = 145

Fisher's Ideal Index, *i.e.*, $P_{01} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$

$\Sigma p_1q_0 = 174$, $\Sigma p_0q_0 = 140$, $\Sigma p_1q_1 = 180$, $\Sigma p_0q_1 = 145$

Substituting the values:

$$P_{01} = \sqrt{\frac{174}{140} \times \frac{180}{145}} \times 100 = 1.2429 \times 100 = 124.29$$

Time Reversal Test

Time Reversal Test is satisfied if $P_{01} \times P_{10} = 1$

$$P_{10} = \sqrt{\frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}} = \sqrt{\frac{145}{180} \times \frac{140}{174}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{174}{140} \times \frac{180}{145} \times \frac{145}{180} \times \frac{140}{174}} = \sqrt{1} = 1.$$

Hence Time Reversal Test is satisfied.

Factor Reversal Test

Notes

Factor Reversal Test is satisfied when:

$$P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \sqrt{\frac{145}{140} \times \frac{180}{174}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{174}{140} \times \frac{180}{145} \times \frac{145}{140} \times \frac{180}{174}} = \frac{180}{140}$$

$$\frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{180}{140}.$$

Hence Factor Reversal Test is satisfied.

Example 4: From the following data construct Fisher's Ideal Index Number and show how it satisfies Time Reversal Test and Factor Reversal Test.

Items	Base Year		Current Year	
	Price Per unit Rs.	Total Expenditure Rs.	Price per unit Rs.	Total Expenditure Rs.
1	2	40	5	75
2	4	16	8	40
3	1	10	2	24
4	5	25	10	60

Solution: Divide expenditure by price to get quantity figures and then calculate Fisher's Ideal Index.

Item	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
1	2	20	5	15	100	40	75	30
2	4	4	8	5	32	16	40	20
3	1	10	2	12	20	10	24	12
4	5	5	10	6	50	25	60	30
					$\sum p_1q_0$ = 202	$\sum p_0q_0$ = 91	$\sum p_1q_1$ = 199	$\sum p_0q_1$ = 92

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{202}{91} \times \frac{199}{92}} \times 100 = 2.1912 \times 100 = 219.12.$$

Notes

Time Reversal Test

Time reversal test is satisfied if $P_{01} \times P_{10} = 1$

$$P_{10} = \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{\sum p_1 q_0}} = \sqrt{\frac{92}{199} \times \frac{91}{202}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{202}{91} \times \frac{199}{92} \times \frac{92}{199} \times \frac{91}{202}} = \sqrt{1} = 1.$$

Hence time reversal test is satisfied.

Factor Reversal Test

Factor reversal test is satisfied if:

$$P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \sqrt{\frac{92}{91} \times \frac{199}{202}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{202}{91} \times \frac{199}{92} \times \frac{92}{91} \times \frac{199}{202}} = \frac{199}{91}$$

Hence factor reversal test is satisfied.

Self-Assessment

1. Tick the correct Answer

(*i*) A good index number is one that satisfies

(*a*) Unit test (*b*) time reversal test

(*c*) factor reversal test (*d*) circular test

(*e*) all the test form.

(*ii*) Time reversal test is satisfied when

(*a*) $P_{01} \times P_{10} = 0$ (*b*) $P_{01} \times P_{10} = 1$

(*c*) $P_{01} \times P_{10} > 1$ (*d*) $P_{01} \times P_{10} < 1$

(*iii*) The circular test is satisfied when

(*a*) $P_{12} \times P_{23} \times P_{31} = 0$ (*b*) $P_{12} \times P_{23} \times P_{13} = 1$

(*c*) $P_{12} \times P_{23} \times P_{31} = 1$ (*d*) $P_{12} \times P_{32} \times P_{31} = 0$

(*iv*) The circular test is an extension of the

(*a*) unit test (*b*) time reversal test

(*c*) Factor reversal test (*d*) None of these

(*v*) Factor reversal test is satisfied when

(*a*) $P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$ (*b*) $\sum p_1 q_0$

(*c*) $P_{01} \times Q_{10} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$ (*d*) None of these

Notes

20.5 Summary

- A number of mathematical criteria for judging the adequacy of an Index Number formula have been developed by statisticians. In fact, the problem is that of selecting the most appropriate one in a given situation.
- This test requires that the formula for constructing an index should be independent of the units in which the prices are quoted. All formulae of weighted aggregate method except simple aggregative method satisfy this test.
- Time Reversal means that if we change the base year to the current year and *vice versa* then the product of the indices should be equal to unity. In other words, Simple Aggregative Method does not satisfy this test. The index number reckoned forward should be the reciprocal of that reckoned backward'. Thus, an ideal Index Number formula should work both ways, *i.e.*, forward as well as backward.
- According to Fisher Just as our formula should permit the interchange of two factors without giving inconsistent results, so without to permit interchange of prices and quantities without giving inconsistent results, *i.e.*, the two results multiplied together should give the true value ratio.
- Another test of adequacy is circular test. This test is an extension of the time reversal test. It requires that if an index is constructed for the year '*b*' on year '*a*' and for the year '*c*' on base year '*b*' we should get the same result as if we calculate directly an Index for '*c*' on base year '*a*' without going through *b* as an intermediary.

20.6 Key-Words

1. Coefficient of variation (CV) : The standard deviation divided by the mean.
2. Collinearity : The condition in which the independent variables are (usually highly) correlated with each other.
3. Column totals : The total number of observations occurring in a column of a contingency table.

20.7 Review Questions

1. Explain time reversal test and factor reversal tests. Show that Fisher's Ideal index number satisfies both.
2. What do you mean by time reversal test for index numbers ? Show that laspeyre and paasche index numbers do not satisfy it and that Fisher's Ideal index does.
3. What are the various tests of adequacy of index number formulae ? Describe each briefly.
4. Distinguish between Laspeyre's and Paasche's index.
5. What are the differences between time reversal test and factor reversal test?

Answers: Self-Assessment

1. (*i*) (*e*) (*ii*) (*b*) (*iii*) (*c*) (*iv*) (*b*) (*v*) (*a*)

20.8 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.

Notes

Unit 21: Cost of Living Index and Its Uses and Limitation of Index Numbers

CONTENTS

Objectives

After reading this unit students will be able to:

- Explain cost of Living Index and its Uses.
- Describe the Limitation of Index Numbers.

Introduction

One of the main types of index numbers in use is the cost of living index number (CLI). This is also known as consumer price index number (CPI). Gradually the expression CLI is being replaced by CPI; it is a special index number of retail prices in which only prices of selected commodities are considered which enter into the consumption pattern of a particular group of people. Thus different items enter into the "market basket of goods'', of different groups. Different groups of people have different CLI numbers. The market basket of goods includes goods and services needed for maintaining a certain standard of living for that group over a period of time. The CLI measures changes in the cost of maintaining the standard of living for that group.

In India CLI numbers are being constructed for three groups of people. These index numbers are

(1) The working class cost of living index numbers

(2) The middle-class cost of living index numbers

(3) The cost of living index numbers of the Central Government employees.

Did u know? The commodities of selected items for the group constitute what is known as "market basket of goods" for that group.

We shall describe them later. The basket of goods is divided into five major groups-food, housing, fuel and light, clothing and other goods and services.

In the U.S.A. a "Consumer Price Index for Urban Wage Earners and Clerical Workers" is constructed regularly. The commodities are divided into 8 major groups — food, housing, dress, transportation, medical care, personal care, reading and recreation and other goods and services, The special problems that arise in the construction of cost of living index numbers for a group lie in determining the market basket of goods and services needed for a person of the group for maintaining a certain standard of living. While transport may be an item for city dwellers, this may not be so in the case of villagers in a developing country.

Notes

The following points need be considered in the selection of items for a group of people: (1) items taken should be such as to represent the habits, tastes and traditions of the average person in the group; (2) the economic and social importance of the goods and services are also to be examined; (3) items should be such that they are not likely to vary in quality in appreciable degree over two different places or different periods of time; and (4) items should be fairly large in number so as to represent adequately the standard of living for the groups). A fairly reasonable number should be selected. Again after determining the items to be included in the basket, the question that arises is the determination of suitable weights for different items in the basket.

To determine the weights, a proper study of consumption habits of persons of that group is to be made. The usual procedure is to conduct "family budget surveys". Such surveys help in determining the items to be entered into the consumption pattern of the group and also help in determining the weights to be assigned to different items. One problem may occur regarding different *qualities* of the same type of commodity. Another problem may concern items of common use which do not occur in both the base period and the given period.

For a detailed account of this topic refer to Banerjee (1975).

21.1 Cost of Living Index and Its Uses

Meaning and Need

The consumer price index numbers, also known as cost of living index number, are generally intended to represent the average change over time in the prices paid by the ultimate consumer of a specified basket of goods and services. The need for constructing consumer price indices arises because the general index numbers fail to give an exact idea of the effect of the change in the general price level on the cost of living of different classes of people, since a given change in the level of prices affects different classes of people in different manners. Different classes of people consume different types of commodities and even the same type of commodities are not consumed in the same proportion by different classes of people. For example, the consumption pattern of rich, poor and middle class people varies widely. Not only this, the consumption habits of the people of the same class differ from place to place. For example, the mode of expenditure of a lower division clerk living in Delhi may differ widely from that of another clerk of the same category living in, say, Chennai. The consumer price index helps us in determining the effect of rise and fall in prices on different classes of consumers living in different areas. The construction of such an index is of great significance because very often the demand for a higher wage is based on the cost of living index and the wages and salaries in most countries are adjusted in accordance with the consumer price index.

The consumer price index numbers were earlier known as cost of living index numbers. But this name was not a happy one since the cost of living index does not measure the actual cost of living nor the fluctuations in the cost of living due to causes other than the change in the price level ; its object is to find out how much the consumers of a particular class have to pay more for a certain basketful of goods and services in a given period compared to the base period. At present, the three terms, namely, cost of living index, consumer price index and retail price index, are in use in different countries with practically no difference in their connotation. However, the term 'consumer price index' is the most popular of the three.

Notes To bring out clearly this fact, the Sixth International Conference of Labour Statisticians recommended that the term 'cost of living index' should be replaced in appropriate circumstances by the terms *'Price of living index'*, *'cost of living price index'*, or *'consumer price index'*.

It should be clearly understood at the very outset that two different indices representing two different geographical areas cannot be used to compare actual living costs of the two areas. A higher index for one area than for another with the same period is no indication that living costs are higher in the one

Notes

than in the other. All it means is that as compared with the base periods, prices have risen in one area than in another. But actual costs depend not only on the rise in prices as compared with the base period, but also on the actual cost of living for the base period which will vary for different regions and for different classes of population.

Utility of the Cost of Living Index

The Consumer Price Indices are of great significance as can be seen from the following:

(1) The most common use of these indices is in wage negotiations and wage contracts. Automatic adjustments of wage or dearness allowance component of wages are governed in many countries by such indices.

(2) At Governmental level, the index numbers are used for wage policy, price policy, rent control, taxation and general economic policies.

(3) The index numbers are also used to measure changing purchasing power of the currency, real income, etc.

(4) Index numbers are also used for analysing markets for particular kinds of goods and services.

Construction of a Consumer Price Index or Cost of Living Index:

The following are the steps in constructing a consumer price index:

(1) Decision about the class of people for whom the index is meant: It is absolutely essential to decide clearly the class of people for whom the index is meant, *i.e.*, whether it relates to industrial workers, teachers, officers, etc. The scope of the index must be clearly defined. For example, when we talk of teachers, we are referring to primary teachers, middle class teachers, etc., or to all the teachers taken together. Along with the class of people it is also necessary to decide the geographical area covered by the index. Thus in the example taken above it is to be decided whether all the teachers living in Delhi are to be included or those living in a particular locality of Delhi, say, Chandni Chowk or Karol Bagh, etc.

Task What do you mean by cost of living index?

(2) Conducting family budget enquiry: Once the scope of the index is clearly defined the next step is to conduct a family budget enquiry covering the population group for whom the index is to be designed. The object of conducting a family budget enquiry is to determine the amount that an average family of the group included in the index spends on different items of consumption. While conducting such an enquiry, therefore, the quantities of commodities consumed and their prices are taken into account. The consumption pattern can thus be easily ascertained. It is necessary that the family budget enquiry amongst the class of people to whom the index series is applicable should be conducted during the base period. The Sixth International Conference of Labour Statisticians held in Geneva in 1946 suggested that the period of enquiry of the family budgets and the base periods should be identical as far as possible.

The enquiry is conducted on a random basis. By applying lottery method some families are selected from the total number and their family budgets are scrutinized in detail. The items on which the money is spent are classified into certain well accepted groups, namely,

(i) Food (ii) Clothing (iii) Fuel and Lighting

(iv) House Rent (v) Miscellaneous.

Each of these groups is further divided into sub-groups. For example, the broad group 'food' may be divided into wheat, rice, pulses, sugar, etc. The commodities included are those which are generally consumed by people for whom the index is meant. Through family budget enquiry an average budget is prepared which is the standard budget for that class of people. While constructing the index only such commodities should be included as are not subject to wide variations in quality or to wide seasonal alterations in supply and for which regular and comparable quotations of prices can be obtained.

Notes

(3) Obtaining price quotations: The collection of retail prices is a very important and, at the same time, very tedious and difficult task because such prices may vary from place to place, shop to shop and person to person. Price quotations should be obtained from the localities in which the class of people concerned reside or from where they usually make their purchases. Some of the principles recommended to be observed in the collection of retail price data required for purposes of construction of cost of living indices are described below

(a) The retail prices should relate to a fixed list of items and for each item, the quality should be fixed by means of suitable specification.

(b) Retail prices should be those actually charged to consumers for cash sales.

(c) Discount should be taken into account if it is automatically given to all customers.

(d) In a period of price control or rationing, where illegal prices are charged openly, such prices should be taken into account along with the controlled prices.

The most difficult problem in practice is to follow principle (a), *i.e.*, the problem of keeping the weights assigned and qualities of the basket of goods and services constant with a view to ensuring that only the effect of price change is measured. To conform to uniform qualities, the accepted method is to draw up detailed descriptions for specifications of the items priced for the use of persons furnishing or collecting the price quotations.

Since prices form the most important component of cost of living indices, considerable attention has to be paid to the methods of price collection and to the price collection personnel. Prices are collected usually by special agents or through mailed questionnaire or in some cases through published price lists. The greatest reliance can be placed on the price collection through special agents as they visit the retail outlets and collect the prices from them. However, these agents should be properly selected and trained and should be given a manual of instructions as well as manual of specifications of items to be priced.

Notes Appropriate methods of price verification should be followed such as *'check pricing'* in which price quotations are verified by means of duplicate prices obtained by different agents or *'purchase checking'* in which actual purchases of goods are made.

After quotations have been collected from all retail outlets an average price for each of the items included in the index has to be worked out. Such averages are first calculated for the base period of the index and later for every month if the index is maintained on a monthly basis. The method of averaging the quotations should be such as to yield unbiased estimates of average prices as being paid by the group as a whole. This, of course, will depend upon the method of selection of retail outlets and also the scope of the index.

In order to convert the prices into index numbers the prices or their relatives must be weighted. The need for weighting arises because relative importance of various items for different classes of people is not the same. For this reason, the cost of living index is always a weighted index. While conducting the family budget enquiry the amount spent on each commodity by an average family is ascertained and these constitute the weights. Percentage expenditures on the different items constitute the individual weights' allocated to the corresponding price relative and the percentage expenditure on the five groups constitute the 'group weight'.

Methods of Constructing the Index

After the above mentioned problems are carefully decided the index may be constructed by applying any of the following methods:

(1) Aggregate Expenditure Method or Aggregative Method; and

(2) Family Budget method or The Method of Weighted Relatives.

1. Aggregate Expenditure Method: When this method is applied the quantities of commodities consumed by the particular group in the base year are estimated which constitute the weights. The prices of commodities for various groups for the current year are multiplied by the quantities

Notes

consumed in the base year and the aggregate expenditure incurred in buying those commodities is obtained. In a similar manner the prices of the base year are multiplied by the quantities of the base year and aggregate expenditure for the base period is obtained. The aggregate expenditure of the current year is divided by the aggregate expenditure of the base year and the quotient is multiplied by 100. Symbolically,

$$\text{Consumer Price Index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

There is in fact the Laspeyre's method discussed earlier. This method is the most popular method for consumer price index.

2. **Family Budget Method:** When this method is applied the family budgets of a large number of people for whom the index is meant are carefully studied and the aggregate expenditure of an average family on various items is estimated. These constitute the weights. The weights are thus the value weights obtained by multiplying the prices by quantities consumed (*i.e.*, p_0q_0). The price relatives for each commodity are obtained and these price relatives are multiplied by the value weight for each item and the product is divided by the sum of the weight. Symbolically,

$$\text{Consumer Price Index} = \frac{\sum PV}{\sum V}$$

where, $P = \frac{p_1}{p_0} \times 100$ for each item

V = Value weights, *i.e.*, p_0q_0.

This method is the same as the weighted average of price relative method discussed earlier.

It should be noted that the answer obtained by applying the aggregate expenditure method and the family budget method shall be the same.

Example 1: Construct the consumer price index number of 2005 on the basis from the following data using (i) the average expenditure method, and (ii) the family budget method:

Commodity	Quantity consumed in 2004	Unit	Price in 2004		Price in 2005	
			Rs.	Paise	Rs.	Paise
A	6 Quintal	Quintal	5	75	6	0
B	6 "	"	5	0	8	0
C	1 "	"	6	0	9	0
D	6 "	"	8	0	10	0
E	4 Kg.	Kg.	2	0	1	50
F	1 Quintal	Quintal	20	0	15	0

Solution:

Computation of Consumer Price Index Number for 2005
(Base 2004 = 100) By the Aggregate Expenditure Method

Commodities	Quantities consumed q_0	Unit	Price in 2004 p_0	Price in 2005 p_1	p_1q_0	p_0q_0
A	6 Qtl.	Qtl.	5.75	6.00	36.00	34.50
B	6 "	"	5.00	8.00	48.00	30.00

Notes

C	1 "	"	6.00	9.00	9.00	6.00
D	6 "	"	8.00	10.00	60.00	48.00
E	4 Kg.	Kg.	2.00	1.50	6.00	8.00
F	1 Qtl.	Qtl.	20.00	15.00	15.00	20.00
					$\Sigma p_1 q_0$ = 174	$\Sigma p_0 q_0$ = 146.50

$$\text{Consumer Price Index} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \frac{174}{146.5} \times 100 = 118.77$$

Construction of consumer Price Index Number for 2005 (Base 2004 = 100)

By the Family Budget Method

Articles	**Quantities consumed** q_0	**Unit**	**Price in 2004** p_0	**Price in 2005** p_1	$\frac{p_1}{p_0} \times 100$ p	$p_0 q_0$ V	**PV**
A	6 Qtl.	Qtl.	5.75	6.0	104.35	34.5	3,600
B	6 "	"	5.00	8.0	160.00	30.0	4,800
C	1"	"	6.00	9.0	150.00	6.0	900
D	6 "	"	8.00	10.0	125.00	48.0	6,000
E	4 Kg.	Kg.	2.00	1.5	75.00	8.0	600
F	1 Qtl.	Qtl.	20.00	15.0	75.00	20.0	1,500
						ΣV = 146.5	ΣPV = 17,400

$$\text{Consumer Price Index} = \frac{\Sigma PV}{\Sigma V} = \frac{17,400}{146.5} = 118.77$$

Thus, the answer is the same by both methods. However, the reader should prefer the aggregate expenditure method because it is far more easier to apply compared to the family budget method.

Example 2: An enquiry into the budgets of the middle class families in a city in India gave the following information:

Expenses on:	**Food** *35%*	**Rent** 15%	**Clothing** 20%	**Fuel** 10%	**Misc.** 20%
Price in 2004:	450	90	225	75	120
Price in 2005:	435	90	195	69	135

What change in the cost of living figures of 2005 has taken place as compared to 2004?

Notes

Solution:

Construction of Cost of Living Index for 2005 with 2004 the Base

Expenses on	Price in Rs.		Price Relative	W	PW
	2004	2005	$\frac{p_1}{p_0}\times 100$ P		
Food	450	435	96.67	35	3383.45
Rent	90	90	100.00	15	1500.00
Clothing	225	195	86.67	20	1733.40
Fuel	75	69	92.00	10	920.00
Misc.	120	135	112.50	20	2250.00
				$\Sigma W = 100$	$\Sigma PW = 9786.85$

$$\text{Cost of Living Index} = \frac{\Sigma PW}{\Sigma W}$$

$$\Sigma PW = 9786.85, \Sigma W = 100$$

$$\text{Index} = \frac{9786.85}{100} = 97.87$$

Thus a fall of (100 – 97.87), *i.e.*, 2.13% has taken place in 2005 as compared to 2004.

Example 3 : Construct the cost of living index number from the data given below:

Group	Index	Expenditure
1. Food	550	46%
2. Clothing	215	10%
3. Fuel and Lighting	220	7%
4. House Rent	150	12%
5. Miscellaneous	275	25%

Solution:

Construction of Cost of Living Index Number

Group	Index Number I	Expenditure W	IW
Food	550	46	25,300
Clothing	215	10	2,150
Fuel and Lighting	220	7	1,540
House Rent	150	12	1,800
Miscellaneous	275	25	6,875
		$\Sigma W = 100$	$\Sigma IW = 37,665$

$$\text{Cost of Living Index} = \frac{\Sigma IW}{\Sigma W} = \frac{37,665}{100} = 376.65.$$

Precautions while using Consumer Price Index of Cost of Living Index

Notes

Quite often the consumer price indices are misinterpreted. Hence while using these indices the following points should be kept in mind

1. As pointed out earlier the consumer price index measures changes in the retail prices only in the given period compared to base period —it does not tell us anything about variations in living standard at two different places. Thus, if the cost of living index for working class for Mumbai is 175 and Delhi 150 for the same period and for the same class of people, it does not necessarily mean that living costs are higher in Mumbai as compared to Delhi.
2. While constructing the index it is assumed that the quantities of the base year are constant and hold good for current year also. But this assumption does not appear to be very logical because the pattern of consumption, goes on changing with the change in fashion, introduction of new commodities in the market, etc. It is desirable, therefore, that while constructing the index the current year quantities are taken into account. But this is a difficult task. The Sixth International Conference of Labour Statisticians recommended that the pattern of consumption should be examined and the weights adjusted, if necessary, at intervals of not more than ten years to correspond changes in the consumption pattern. The index also does not take into account changes in qualities. Unlike changes in consumption pattern changes in qualities of goods and services are more frequent and when a marked change in the quality of items occurs appropriate adjustment should be made to ensure that the index takes into account changes is qualities also. But in practice it is a difficult proposition to follow and, therefore, constant qualities are assumed at two different dates which again is a shaky assumption.
3. Like any other index the consumer price index is based on a sample. While constructing the index, sampling is used at every stage in the selection of commodities, in obtaining price quotations, selecting families for family budget enquiry, etc. The accuracy of the index thus hinges upon the use of sampling methods. The consumption pattern derived from the expenditure data of a sample of households covered in the course of family budget enquiry has to be representative of all the items in the average budget, the localities from which price data are collected have to be representative of all the localities from which the population group make purchases, the retail outlets from which prices are collected have to be representative of all the retail outlets patronised by the population group, etc. However, it is often difficult to ensure perfect representativeness and in the absence of this the index may fail to provide the real picture.

Example 4: The following are the group index numbers and the group weights of the budget of an average working class family. Construct the cost of living index number.

Group	Index	Weight
Food	352	48
Fuel and Lighting	220	10
Clothing	230	8
Rent	160	12
Miscellaneous	190	15

Solution:

Construction of Cost of Living Index

Group	Index I	Weight W	IW
Food	352	48	16,896
Fuel and Lighting	220	10	2,200

Notes

Clothing	230	8	1,840
Rent	160	12	1,920
Miscellaneous	190	15	2,850
		$\Sigma W = 93$	$\Sigma IW = 25,706$

$$\text{Cost of Living Index} = \frac{\Sigma IW}{\Sigma W} = \frac{25,706}{93} = 276.41.$$

Example 5: The percentage expenses on different commodities consumed by the middle class families of a certain city and the group index numbers in 2005 as compared with the base year 2004 are as follows:

Commodities	% Expenses	Index
Food	45	410
Rent	15	150
Clothing	12	343
Fuel and Lighting	8	248
Miscellaneous	20	285

Calculate the consumer index for the year 2005.

Solution:

Calculation of Consumer Price Index

Commodities	% Expenses W	Index I	IW
Food	45	410	18,450
Rent	15	150	2,250
Clothing	12	343	4,116
Fuel and Lighting	8	248	1,984
Miscellaneous	20	285	5,700
	$\Sigma W = 100$	$\Sigma IW = 32,500$	

$$\text{Consumer Price Index} = \frac{\Sigma IW}{\Sigma W} = \frac{32,500}{100} = 325.$$

Example 6: Construct the cost of living index number from the following group data:

Group		Weights	Group Index No.
(1)	Food	47	247
(2)	Fuel and Lighting	7	293
(3)	Clothing	8	289
(4)	House Rent	13	100
(5)	Miscellaneous	14	236

Solution: Construction of Cost of Living Index

Notes

Group	Weights w	Group Index I	IW
Food	47	247	11,609
Fuel and Lighting	7	293	2,051
Clothing	8	289	2,312
House Rent	13	100	1,300
Miscellaneous	14	236	3,304
	$\sum W = 89$		$\sum IW = 20,576$

$$\text{Cost of Living Index} = \frac{\sum IW}{\sum W} = \frac{20,576}{89} = 231.19.$$

Example 7: The data below show the percentage increases in price of a few selected food items and the weights attached to each of them. Calculate the index number for the food group.

Food Items	Rice	Wheat	Dal	Ghee	Oil	Spices	Milk	Fish	Vegetables	Refreshments
Weight	33	11	8	5	5	3	7	9	9	10
Percentage increase in price	180	202	115	212	175	517	260	426	332	279

Using the above food index and the information given below, calculate the cost of living index number.

Group	Food	Clothing	Fuel & Light	Rent & Rates	Miscellaneous
Index	—	310	220	150	300
Weight	60	5	8	9	18

Solution:

Calculations for Food Index

Food Items	Weight W	Percentage Increase	Current Index* I	IW
Rice	33	180	280	9240
Wheat	11	202	302	3322
Dal	8	115	215	1720
Ghee	5	212	312	1560
Oil	5	175	275	1375
Spaces	3	517	617	1851
Milk	7	260	360	2520
Fish	9	426	526	4734
Vegetables	9	332	432	3888
Refreshments	10	279	379	3790
	$\sum W = 100$			$\sum IW = 34000$

Notes

$$\text{Food Index} = \frac{\sum IW}{\sum W} = \frac{34000}{100} = 340$$

Construction of Cost of Living Index

Group	Index I	Weights W	IW
Food	340	60	20,400
Clothing	310	5	1,550
Fuel and Light	220	8	1,760
Rent and Rates	150	9	1,350
Miscellaneous	300	18	5,400
		$\sum IW = 100$	$\sum IW = 30,460$

$$\text{Cost of Living Index} = \frac{\sum IW}{\sum W} = \frac{30460}{100} = 304.6$$

* Current index has been obtained by adding 100 to the percentage increase in the various food items.

Example 8: Calculate the Cost of Living Index Number from the following data:

Items	Price		Weights
	Base Year	Current Year	
Food	30	47	4
Fuel	8	12	1
Clothing	14	18	3
House Rent	22	15	2
Miscellaneous	25	30	1

Solution:

Construction of Cost of Living Index

Items	p_0	p_1	$\frac{p_1}{p_0} \times 100$ p	W	PW
Food	30	47	156.67	4	626.68
Fuel	8	12	150.00	1	150.00
Clothing	14	18	128.57	3	385.71
House Rent	22	15	68.18	2	136.36
Miscellaneous	25	30	120.00	1	120.00
				$\sum W = 11$	$\sum PW = 1418.75$

$$\text{Cost of Living Index} = \frac{\sum PW}{\sum W} = \frac{1418.75}{11} = 128.98.$$

Example 9: (a) From the chain base index numbers given below, prepare fixed base index numbers:

Notes

Year:	2001	2002	2003	2004	2005
Index:	110	150	140	200	150

(b) From the chain base index number given below, construct fixed base index numbers:

Year:	2001	2002	2003	2004	2005
Chain Base Indices:	80	110	120	90	140

Solution:

CONSTRUCTION OF COST OF LIVING INDEX

Expenses	**2004** p_0	**2005** p_1	$\frac{p_1}{p_0} \times 100$	**W**	**PW**
Food	150	174	116	35	4060
Rent	50	60	120	15	1800
Clothing	100	125	125	20	2500
Fuel	20	25	125	10	1250
Misc	60	90	150	20	3000
				$\Sigma W = 100$	$\Sigma PW = 12610$

Cost of Living Index $\frac{\Sigma PW}{\Sigma W} = \frac{12610}{100} = 126.1$. Thus as compared to 2004 the cost of living index has risen by 26.1 per cent in 2005.

21.2 Limitations of Index Numbers

Though the index numbers are of great significance, the reader must also be aware of their limitations so that he avoids errors of interpretation. The chief limitations of index numbers are:

1. Since index numbers are generally based on a sample, it is not possible to take into account each and every item in the construction of the index.
2. While taking the sample random sampling is seldom used. This is so because to sample from a population of literally millions of commodities and services, the random procedure could neither be practical nor representative. Typically, indices are constructed from samples deliberately selected. This is likely to introduce errors and every effort must be made to minimise these errors.
3. It is often difficult to take into account changes in the quality of products. With the passage of time tastes and habits of people also change with the result that very often old commodities go out of use and new commodities are introduced. In a really typical index, qualities of commodities should remain the same over a period of time because differences in quality would mean differences in prices also. But very often it is not practicable and it makes comparisons over long periods less reliable.
4. A large number of methods have been designed for constructing index numbers and different methods of computation give different results. Very often the selection of an appropriate formula creates problems and in the interest of comparability, it is necessary to ensure that the same

Notes

formula is adopted over a period of time for constructing a particular index. There is no index number method which is most satisfactory from all the various points of view which may logically or practically be taken. Index numbers are averages, and all averages are basically compromises between opposing extremes or forces.

5. Just like other statistical tools, index numbers can also be manipulated in such a manner as to draw the desired conclusions. Choosing a freak year is a favourite trick of those who use statistics to mislead. A dishonest capitalist could choose a record year of profits as base and so "prove" subsequent profits to be pitifully low. Similarly, in order to prove that the current prices are intolerably high a dishonest trade unionist may choose a year of exceptionally low prices as base.
6. Since in the construction of index numbers a large number of factual questions are involved, lack of adequate and accurate data in most cases becomes a serious limitation of the index itself. In most of the cases one cannot collect the data himself and, therefore, one has to rely on a published source. Ordinarily, we draw upon many sources of data which are geographically dispersed. Problems of comparability and reliability thus multiply and the chances of spurious results are increased. One mistake may "bias" the index such as including the price of one commodity for one time period, or the price of a slightly different commodity for another period, or taking the manufacturer's price at one time and wholesaler's or retailer's price another time.
7. Comparisons over long periods are not reliable.

Self-Assessment

1. Fill in the blanks:

(*i*) Theoretically the best average in the cost of living index numbers is .....

(*ii*) Cost of living index = $\frac{\text{.....}}{\sum p_0 q_0} \times 100$.

(*iii*) Kelly's method of constructing index involves the formula P_{01} = ... where q = ...

(*iv*) Cost of living index help in determaining ..... wages.

(*v*) Cost of living index help in determing the purchasing power of ..... .

21.3 Summary

- One of the main types of index numbers in use is the cost of living index number (CLI). This is also known as consumer price index number (CPI). Gradually the expression CLI is being replaced by CPI; it is a special index number of retail prices in which only prices of selected commodities are considered which enter into the consumption pattern of a particular group of people. The commodities of selected items for the group constitute what is known as "market basket of goods" for that group. Thus different items enter into the "market basket of goods'', of different groups. Different groups of people have different CLI numbers. The market basket of goods includes goods and services needed for maintaining a certain standard of living for that group over a period of time. The CLI measures changes in the cost of maintaining the standard of living for that group.
- In the U.S.A. a "Consumer Price Index for Urban Wage Earners and Clerical Workers" is constructed regularly. The commodities are divided into 8 major groups — food, housing, dress, transportation, medical care, personal care, reading and recreation and other goods and services, The special problems that arise in the construction of cost of living index numbers for a group lie in determining the market basket of goods and services needed for a person of the group for maintaining a certain standard of living. While transport may be an item for city dwellers, this may not be so in the case of villagers in a developing country.
- To determine the weights, a proper study of consumption habits of persons of that group is to be made. The usual procedure is to conduct "family budget surveys". Such surveys help in

Notes

determining the items to be entered into the consumption pattern of the group and also help in determining the weights to be assigned to different items. One problem may occur regarding different *qualities* of the same type of commodity. Another problem may concern items of common use which do not occur in both the base period and the given period.

- The consumer price index numbers, also known as cost of living index number, are generally intended to represent the average change over time in the prices paid by the ultimate consumer of a specified basket of goods and services. The need for constructing consumer price indices arises because the general index numbers fail to give an exact idea of the effect of the change in the general price level on the cost of living of different classes of people, since a given change in the level of prices affects different classes of people in different manners. Different classes of people consume different types of commodities and even the same type of commodities are not consumed in the same proportion by different classes of people.
- The consumer price index numbers were earlier known as cost of living index numbers. But this name was not a happy one since the cost of living index does not measure the actual cost of living nor the fluctuations in the cost of living due to causes other than the change in the price level ; its object is to find out how much the consumers of a particular class have to pay more for a certain basketful of goods and services in a given period compared to the base period. To bring out clearly this fact, the Sixth International Conference of Labour Statisticians recommended that the term 'cost of living index' should be replaced in appropriate circumstances by the terms *'Price of living index'*, *'cost of living price index'*, or *'consumer price index'*. At present, the three terms, namely, cost of living index, consumer price index and retail price index, are in use in different countries with practically no difference in their connotation. However, the term 'consumer price index' is the most popular of the three.
- The consumer price index numbers were earlier known as cost of living index numbers. But this name was not a happy one since the cost of living index does not measure the actual cost of living nor the fluctuations in the cost of living due to causes other than the change in the price level ; its object is to find out how much the consumers of a particular class have to pay more for a certain basketful of goods and services in a given period compared to the base period. To bring out clearly this fact, the Sixth International Conference of Labour Statisticians recommended that the term 'cost of living index' should be replaced in appropriate circumstances by the terms *'Price of living index'*, *'cost of living price index'*, or *'consumer price index'*. At present, the three terms, namely, cost of living index, consumer price index and retail price index, are in use in different countries with practically no difference in their connotation. However, the term 'consumer price index' is the most popular of the three.
- The scope of the index must be clearly defined. For example, when we talk of teachers, we are referring to primary teachers, middle class teachers, etc., or to all the teachers taken together. Along with the class of people it is also necessary to decide the geographical area covered by the index. Thus in the example taken above it is to be decided whether all the teachers living in Delhi are to be included or those living in a particular locality of Delhi, say, Chandni Chowk or Karol Bagh, etc.
- The consumption pattern can thus be easily ascertained. It is necessary that the family budget enquiry amongst the class of people to whom the index series is applicable should be conducted during the base period. The Sixth International Conference of Labour Statisticians held in Geneva in 1946 suggested that the period of enquiry of the family budgets and the base periods should be identical as far as possible.
- The commodities included are those which are generally consumed by people for whom the index is meant. Through family budget enquiry an average budget is prepared which is the standard budget for that class of people. While constructing the index only such commodities should be included as are not subject to wide variations in quality or to wide seasonal alterations in supply and for which regular and comparable quotations of prices can be obtained.

Notes

- Price quotations should be obtained from the localities in which the class of people concerned reside or from where they usually make their purchases. Some of the principles recommended to be observed in the collection of retail price data required.
- Since prices form the most important component of cost of living indices, considerable attention has to be paid to the methods of price collection and to the price collection personnel. Prices are collected usually by special agents or through mailed questionnaire or in some cases through published price lists. The greatest reliance can be placed on the price collection through special agents as they visit the retail outlets and collect the prices from them. However, these agents should be properly selected and trained and should be given a manual of instructions as well as manual of specifications of items to be priced. Appropriate methods of price verification should be followed such as *'check pricing'* in which price quotations are verified by means of duplicate prices obtained by different agents or *'purchase checking'* in which actual purchases of goods are made.
- In order to convert the prices into index numbers the prices or their relatives must be weighted. The need for weighting arises because relative importance of various items for different classes of people is not the same. For this reason, the cost of living index is always a weighted index. While conducting the family budget enquiry the amount spent on each commodity by an average family is ascertained and these constitute the weights. Percentage expenditures on the different items constitute the individual weights' allocated to the corresponding price relative and the percentage expenditure on the five groups constitute the 'group weight'.
- The Sixth International Conference of Labour Statisticians recommended that the pattern of consumption should be examined and the weights adjusted, if necessary, at intervals of not more than ten years to correspond changes in the consumption pattern. The index also does not take into account changes in qualities. Unlike changes in consumption pattern changes in qualities of goods and services are more frequent and when a marked change in the quality of items occurs appropriate adjustment should be made to ensure that the index takes into account changes is qualities also. But in practice it is a difficult proposition to follow and, therefore, constant qualities are assumed at two different dates which again is a shaky assumption.
- The consumption pattern derived from the expenditure data of a sample of households covered in the course of family budget enquiry has to be representative of all the items in the average budget, the localities from which price data are collected have to be representative of all the localities from which the population group make purchases, the retail outlets from which prices are collected have to be representative of all the retail outlets patronised by the population group, etc. However, it is often difficult to ensure perfect representativeness and in the absence of this the index may fail to provide the real picture.
- While taking the sample random sampling is seldom used. This is so because to sample from a population of literally millions of commodities and services, the random procedure could neither be practical nor representative. Typically, indices are constructed from samples deliberately selected. This is likely to introduce errors and every effort must be made to minimise these errors.
- A large number of methods have been designed for constructing index numbers and different methods of computation give different results. Very often the selection of an appropriate formula creates problems and in the interest of comparability, it is necessary to ensure that the same formula is adopted over a period of time for constructing a particular index. There is no index number method which is most satisfactory from all the various points of view which may logically or practically be taken. Index numbers are averages, and all averages are basically compromises between opposing extremes or forces.

21.4 Key-Words

1. Combinations : The number of ways objects can be selected without regard to order.
2. Combinatorics : The branch of mathematics dealing with the number of different ways objects can be selected or arranged.

Notes

21.5 Review Questions

1. What is meant by cost of living Index number ? What are its uses ?
2. How are cost of living Index number constructed ?
3. Explain briefly the various methods of construction of cost of living index number.
4. What are the limitations of index numbers.
5. What do you understand by cost of living index numbers ? Describe briefly the various steps involved in their construction.

Answers: Self-Assessment

1. (*i*) Median (*ii*) $\sum p_1 q_0$

   (*iii*) $\frac{\sum p_1 q}{\sum p_0 q} \times 100$, where $q = \frac{q_1 + q_2 \ldots . q_n}{n}$

   (*iv*) read (*v*) money

21.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 22: Time Series Analysis—Introduction and Components of Time Series

CONTENTS

Objectives

After reading this unit students will be able to:

- Know the Introduction to Time Series Analysis.
- Discuss Components of Time Series.
- Explain an Illustration Involving All Components.

Introduction

One of the major managerial responsibilities is the design and implementation of policies for the achievement of the short-term and long-term goals of the business firm. Previous performances must be studied so as to generate or forecast future business activity. Given a projection of the pattern and the level of future business activity, the desirability of alternative actions can then be investigated. For example, we may be interested to project sales activity levels with maintenance of adequate but not excessive inventory levels. Labour and material requirements must be projected. Need of working capital must be anticipated, and appropriate arrangements for financing investigated. The suitability and timing of capital intensive projects must be carefully evaluated. And lastly, once a strategy has been selected, control procedures must be incorporated to enable the firm to reassess the validity of the original projected values and the extent to which the actual results vary on a continuous basis. The quality of the forecasts or projections the management can make is strongly related to the information that can be extracted and used from past data. Time series analysis is one of the quantitative methods used to determine the patterns in data collected over a period of time. Thus, a time series consists of a set of chronological observations of a statistical series recorded either at successive points in time or over successive periods of time.

22.1 Introduction to Time Series Analysis

A series of observations recorded over time is known as a *time series*. The data on the population of a country over equidistant time points constitute a time series, e.g. the population of India recorded at the ten-yearly censuses. Some other examples of time series are: annual production of a crop, say, rice over a number of years, the wholesale price index over a number of months, the turn-over of a firm over a number of months, the sales of a business establishment over a number of weeks, the daily maximum temperature of a place over a number of days, and so on. In fact, economic data are, in general, recorded over time and are released at regular intervals. These constitute economic time series.

Notes

The objectives of analysis of a time series are (l) to give a general description of the past behaviour of the series, (2) to analyse the past behaviour, and (3) to attempt to forecast the future behaviour on the basis of the past behaviour; however, great caution is needed to do this.

A *forecast* does not tell what will happen but indicates what *would* happen if the past behaviour (as reflected in trends etc.) continues.

The techniques of time series analysis have largely been developed by economists. Empirical investigations dealing with economic theory are largely dependent on time series analysists. Social scientists, in general, do not have the privilege of conducting studies through laboratory experimentation. Studies are to be based on time series data collected over time in such cases. For example, trade cycles are important to economists and others in business and commerce. The exact behaviour of the cycles and their causes are of interest to them. Various theories explaining the phenomena are put forward. Analysis of time series provides an important tool for testing the theories and the explanations. Consumer behaviour is studied mainly with the help of time series data.

The analysis of time series plays an important role in empirical investigations, leading to quantitative revolution, in economics and in several other areas of social sciences and even in biological sciences. Thus political economy (usually known as economics) has been described as 'the oldest of the arts, the newest of the sciences-indeed the queen of the social sciences.'

We shall now discuss the techniques used in the analysis of time series. We begin with the main components or characteristic movements in a time series.

Did u know? The analysis of time series is of interest in several areas, such as economics, commerce, business, sociology, geography, meteorology, demography, public health, biology, and so on.

Objectives or Importance of Time Series Analysis

In the words of **Prof. Hirsch,** *"The main objective in analysing of time series is to understand, interpret and evaluate changes in economic phenomena in the hope of most correctly anticipating the course of future events."*

Following are the main objectives of time series analysis.

(1) **Study of the Past Behaviour of the Data:** The purpose of time series analysis is to study the past behaviour of the data to easily understand what changes have taken place in the past.

(2) **To Forecast Future Behaviour:** The second objective of time series analysis is to predict the future behaviour of a particular variable. Time series can play an important role not in making short range estimates for a year or two a head but also estimating the probable seasonal variations within a year.

(3) **Comparison with other Series:** Time series analysis is helpful in making a comparison between the behaviour of different time series. We can make this comparison by knowing the causes of variations in two time series.

(4) **Study of Present Fluctuations:** Time series analysis is helpful in studying the present fluctuations in the economic variables like, national income, cost, prices, production, etc. It enables us to know achievements and failures regarding a particular variable.

(5) **Estimation of Trade Cycles:** The basic objective of time series analysis is to estimate the trade cycles. The businessmen can avoid their losses and get profits with the help of the estimation of trade cycles.

22.2 Components of Time Series

Variations in Time Series

The term time senes are used to refer to any group of statistical information collected at regular intervals of time.

Notes

There are four kinds of changes, or variations, involved in time series analysis. They are:

(i) Secular trend

(ii) Cyclical fluctuation (variation)

(iii) Seasonal variation

(iv) Irregular variation

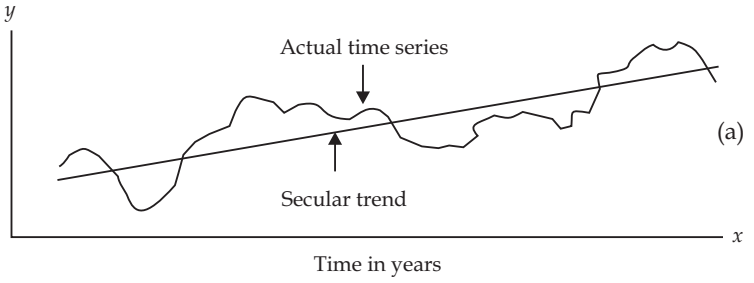
With the secular trend, the value of the variable tends to increase or decrease over a long period of time. The steady increase in the cost of living recorded by the consumer price index is an example of secular trend. From year to year, the cost of living varies a great deal; but, if we consider a long-term period, we see that the trend is towards steady increase. Other examples of secular trend are steady increase of population over a period of time, steady growth of agricultural food production in India over the last ten to fifteen years of time. Figure 1 *(a)* shows a secular trend in an increasing but fluctuating time series.

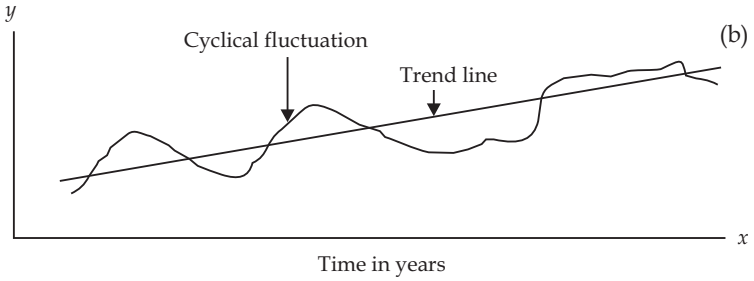
The second type of variation that can be observed in a time series is cyclical fluctuation. The most common example of cyclical fluctuation is the business cycle. Over a period of time, there are years when the business cycle has a peak above the trend line, and at other times, the business activity is likely to slump, touching a low point below the trend line. The time between touching peaks or failing to low points is generally 3 to 5 years, but it can be as many as 15 to 20 years. Figure 1 *(b)* illustrates a typical pattern of cyclical fluctuation. It should be noted that the cyclical movements do not follow any definite trend but move in a somewhat unpredictable manner.

The third kind of fluctuation that can occur in a time series data is the seasonal variation. Seasonal variation involves patterns of change within a year, that tend to be repeated from year to year. For example, sale of umbrellas is on the increase during the months of June and July every year because of the seasonal requirement. Since these are regular patterns, they are useful in forecasting the future production runs. Figure 1 *(c)* gives the seasonal variation in time series.

Irregular variation is the fourth type of change that can be observed in a time series data. These variations may be due to (i) random fluctuations: irregular random fluctuations refer to a large number of minute environmental influences (some uplifting, some depressing) operating on a series at any one time–no one of which is significantly important in and of itself to warrant singling out for individual treatment, and (ii) non-recurring irregular influences that exert a significant one time impact on the behaviour of a time series and as such must be explicitly recognized. The events included in this category are floods, strikes, wars, and so on, which influence the time series data.

The above four variations are generally considered as interacting in a multiplicative manner to produce observed values of the overall time series:





Notes

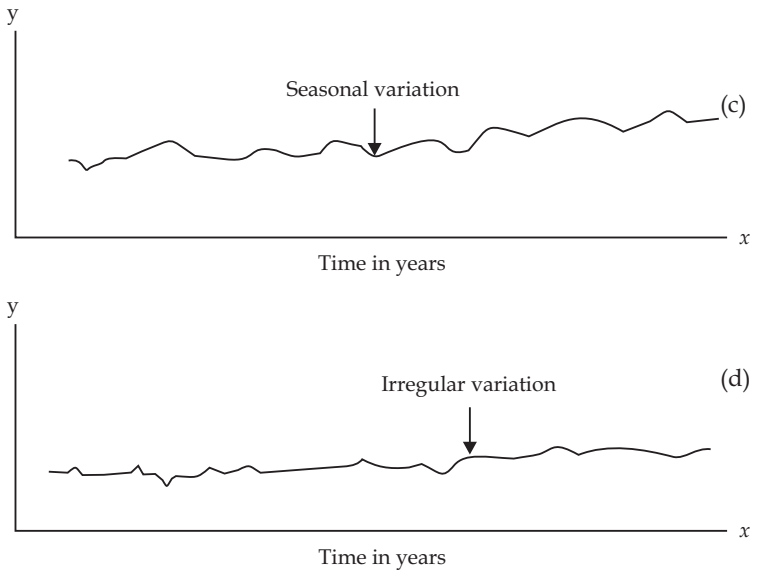


Figure 1: Time Series Variations

Multiplicative model: $O = T \times C \times S \times I$.

where

O = observed value of time series

T = trend component

C = cyclical component

S = seasonal component

I = irregular component

Other types of models that are possible are:

Additive model : $O = T + C + S + I$

Combination model : $O = T \times C \times S + I$

$O = (T + C) \times S \times I$

However, we shall restrict our discussion to the multiplicative model.

Trend Analysis (Secular trend)

Secular trend represents the long-term variation of the time series. One way to describe the trend component in a time series data is to fit a line to a set of points on a graph. An approach to fit the trend line is by the method of least squares.

Following are the three major reasons for studying secular trend in a time series data:

1. Study of secular trend allows us to describe a historical pattern in the data. There are many situations when we can use past trend to evaluate the success of a previous management policy. For example, a multinational organisation may evaluate the effectiveness of the recruitment policy by examining its past enrollment trends.
2. Studying secular trend permits us to project past patterns, or trends into the future. Information of the past can tell us a great deal about the future. Examination of the growth of industrial production in the country, for example, help us to estimate the production for some future years.
3. In many situations, studying the secular trend of time series allows us to eliminate the trend component from the series. This makes it easier for us to study the other components of the time series. If we want to determine the seasonal variation in the sale of shoes, the elimination of the trend component gives us more accurate idea of the seasonal component.

Notes

Trends can be linear, or they can be curvilinear, *i.e.*, parabola. A straight line has a constant rate of change, while a parabola represents a changing rate of change. In fact, parabola shows a trend that is increasing at an increasing rate. The decision as to whether the trend should be a straight line or a parabola is an important one. It certainly is a subjective matter, and one has to be careful to choose the right trend to represent the data. Once the decision is made to the implied nature of the trend (linear or non-linear), one computes the rate of change and any errors in the choice of the trend curve will affect the results. Computation of the rate of change and measuring the trend is a process known as "fitting a curve to the data".

Before we discuss curve fitting, however, let us first consider the general types of trends that are available to represent the data. We shall identify three types of trends which are very popular in the analysis:

The Linear Trend (Straight Line, Constant Rate of Change)

This type of trend is represented in Figure 2, which shows two linear trend curves, (1) sloping upwards and (2) sloping downwards. The mathe-matical model for these linear trend is

$$Y = a + bX \quad \text{... (1)}$$

where X and Y are variables, a is the Y-intercept (*i.e.*, the value of Y when X is equal to zero), and b is the rate of change of Y for unit change in X, or the constant rate of change.

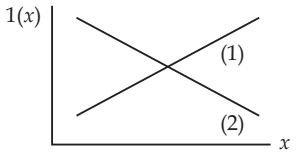


Figure 2: Pattern of Linear Trend.

The Parabolic Trend (Changing Rate)

Figure 3 represents the pattern of this trend. In this pattern (1) the trend is slopping upwards, indicating that the trend is increasing at an increasing rate of change, while in (2) the trend is slopping downwards, indicating that the trend is increasing at a decreasing rate. The mathematical representation of this curve is given by

$$Y = a + bX + cX^2 \quad \text{... (2)}$$

where X and Y are variables, a is the Y-*intercept* (*or the value of* Y when X is zero), and b and c are the rates of change of Y at given values of X which vary with different values of X.

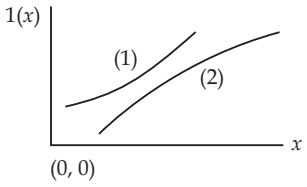


Figure 3:

The Exponential or Logarithmic Trend (Constant Percentage Rate of Change)

The mathematical model for this curve is*:*

$$Y = ab^X \quad \text{... (3)}$$

It is known as an exponential curve because the variable X appears as an exponent in the expression — unlike the other two cases stated earlier, where the variable X appeared as a factor. Plotting this curve on a semi-log graph sheet would produce a straight line (Figure 4), while plotting it on an ordinary graph sheet produces Figure 5. Taking logarithms of both sides of the expression (3), we get

Notes

$$\log Y = \log a + X \log b \quad \ldots (4)$$

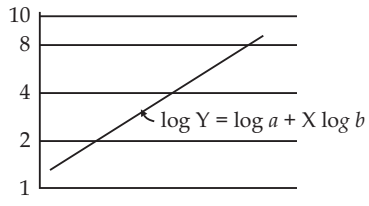


Figure 4: Exponential Trend on a Semi-log Graph.

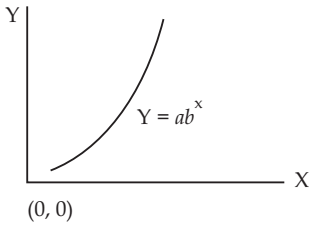


Figure 5: Exponential Trend on an Ordinary Graph.

If (4) is plotted on an ordinary graph sheet, it will produce a straight line (Figure 6), where log a is again the Y intercept and log b is the rate of change. In this case, since this expression is a straight line in a semi-logarithmic chart which shows percentage change, b (the rate of change) is a constant percentage rate of change.

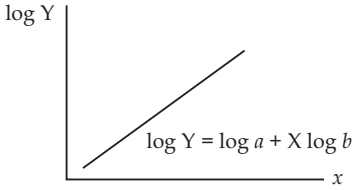


Figure 6: Logarithmic of Exponential Trend on an Ordinary Graph.

We shall study each of the curves individually and develop methods for computing their rates of change. Basically, there are four methods for fitting the treand in time series. These are:

(i) Free hand method

(ii) Method of semi-averages

(iii) Method of moving averages

(iv) Method of least squares.

These four types of secular trend are discussed in units 23, 24 and 25.

Uses of Secular Trend

The following are the main uses of secular trend:

(a) Basis of Fluctuations: The trend values are regarded as normal values. These normal values provide the basis for determining the nature of fluctuations. In other words, it can be found whether the fluctuations are regular or irregular. So general tendency of the data can be analysed with the help of secular trend.

(b) Help in Forecasting: The trend values help in business forecasting and planning of the future operations. This becomes possible since trend values describe the underlined behaviour pattern of the series in past.

(c) Effect of Short-term Variations: By eliminating the trend component in a time series, we can study the effect of short-term variations.

(d) It Facilitates Comparison: The trend analysis facilitates the comparison of two or more time series over different period of time. It helps us in drawing important conclusions about them.

Cyclical Variation

Cyclical variation is that component of a time series that tends to oscillate above and below the secular trend line for periods longer than one year and that they do not ordinarily exhibit regular periodicity. The periods and amplitudes may be quite irregular.

Notes

In a time series of annual data, only the secular trend, cyclical and irregular components are considered since the seasonal variation makes a complete, regular cycle within each year and thus do not affect the annual data. As secular trend can be described by a trend line, it is possible to isolate the remaining, cyclical and irregular components from the trend. For simplicity, we shall assume that the cyclical component explains most of the variations left unexplained by the trend component. However, in situations where this assumption does not hold good, methods such as Fourier analysis and spectral analysis can be used to analyse cyclical component in a time series. (These advanced techniques are beyond the scope of this text.)

If we use a time series composed of annual data, we can find the fraction of the trend by dividing the original value (Y) by the corresponding trend value (Y_{cal}) for each observation in the time series. We then take the percentage of this value by multiplying by 100. This gives us the measure of cyclical variation as a per cent of trend. Mathematically, this can be expressed as:

$$\text{Per cent of trend} = \frac{Y}{Y_{cal}} \times 100 \quad \ldots (5)$$

where Y is the actual observation of the time series data, and Y_{cal} is the calculated trend value in the time series.

The above procedure used to identify cyclical variation is called the residual method.

Example 1: Let us consider the data given in Table 1 which refers to the yield per hectare of a certain foodgrain in an Indian state during 1970 to 1978. The third column in this table refers to the values of linear trend for each time period. The trend line has been developed using the method discussed in the previous section. It can be noted from the graph drawn (see Figure 7) with the actual (Y) and the trend (Y_{cal}) values for the nine years, the actual values move above and below the trend line.

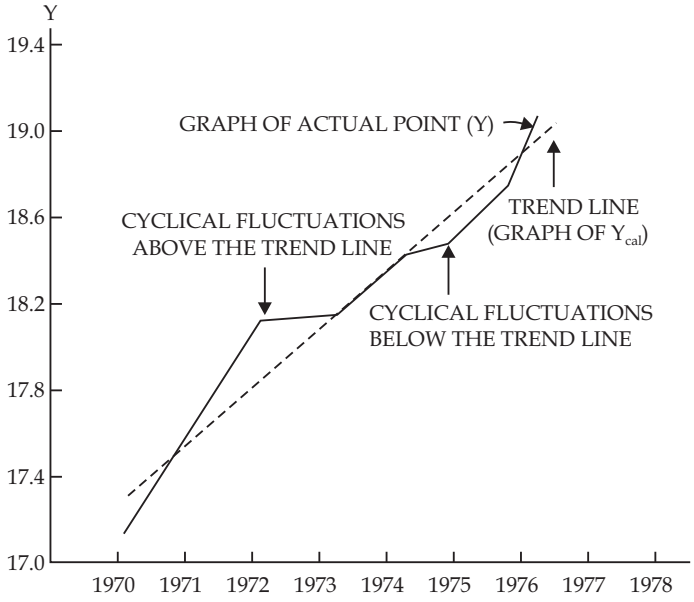


Figure 7: Cyclical Fluctuations Around the Trend Line.

From column 4 of Table 22.1 we can see the variation in actual yield around the estimated trend (98.8 to 101.7). We can attribute these cyclical variations to such factors as rainfall, humidity, etc. However, since these factors are relatively unpredictable, we cannot forecast any specific patten of variation using the method of residuals.

Notes

The above two methods of cyclical variation, per cent of trend and relative cyclical residual, are percentages of the trend. For example, in 1976 the per cent of trend indicated that the actual yield was 99.5 per cent of the expected yield for that year, while for the same year, the relative cyclical residual indicated that the actual yield was 0.5 per cent short of the expected yield during the year. It must be noted here that the methods described above are only used for describing the past cyclical variations and not for predicting future cyclical variations.

Use of Cyclical Variations

The following are the main uses of cyclical variations:

(a) **Aid to Policy Formation:** The study of cyclical variations is extremely useful in framing suitable policies for stabilizing the level of business activity. One can avoid the periods of booms and depressions since both are bad for an economy.

(b) **Helps in Studying Fluctuations of Business:** The cyclical variations are very helpful in studying the characteristics of fluctuations of a business. One can come to know how sensitive is the business to general cyclical influences ? The general pattern of a particular firm's production, profits, sales, raw material prices, etc. can also be known.

(c) **Helps in Forecasting:** The cyclical variations are helpful in forecasting and estimating about the future behaviour. Accurate forecasting is a prerequisite for successful business.

(d) **Knowledge of Irregular Fluctuations:** The study of cyclical variations is helpful in analysing and isolating the effects of irregular fluctuations. One can come to know either the variations are unpredictable or are caused by other isolated special occurrences like floods, earthquakes, strikes, wars, etc.

Seasonal Variation

Seasonal variations are those forces affecting time series that are the result of man made or physical phenomena. The major characteristic of seasonal variations is that they are repetitive and periodic, the period is less than one year, say a week, a month or a quarter. Seasonal variations can affect a time during November is normal, it can examine the seasonal pattern in the previous years and get the information it needs.

1. It is possible to establish the pattern of past changes. This helps us to compare two time intervals that would otherwise be too dissimilar. For example, if a business house wants to know whether the slump in sales during November is normal, it can examine the seasonal pattern in previous years and get the information it needs.
2. Seasonal variations help us to project past patterns into the future. In the case of long range decisions, secular trend analysis may be adequate. However, for short-run decisions, the ability to predict seasonal fluctuations is essential. For example, consider the case of a wholesale food dealer who wants to maintain a minimum adequate stock of all food items. The ability to predict short-run patterns, such as the demand of food items during Diwali, or at Christmas, or during the summer, is very useful to the management of the store.
3. Once the existence of the seasonal pattern has been established, it is possible to eliminate its effects from the time series. This elimination helps us to calculate the cyclical variation that takes place each year. When the effect of the seasonal variation has been eliminated from the time series, we have deseasonalized time series.

In order to measure the seasonal variation, we use the ratio-to-moving average method. This method provides an index that describes the degree of seasonal variation. The index is based on a mean of 100, with the degree of seasonality measured by variations away from the base.

The method of the ratio-to-moving average for computing the indices of seasonal variation is a procedure whereby the different components in the series are measured and are isolated or eliminated. Subsequently, the seasonal effect is identified and expressed in percentage form. We first take a series in which seasonal pattern is suspected and plot this series on a graph to identify the recurrence

Notes

of the pattern. To identify the seasonal component, the data could be in quarters or months or any other time period less than a year.

Example 2 : Consider the data given in Table 3. Compute the seasonal index of quarterly sales of the departmental store by the method of ratio-to-moving average.

The first step in computing the seasonal index is to calculate the four-quarter moving totals for the series. This total is written in between quarters II and III in column 4 of Table 3. However, it could be "dropped down" one line to avoid the problem of having data between the lines.

Secondly, we compute the four-quarter moving average by dividing each of the four-quarter totals by four. We then find the centred four-quarter moving average so as to centre the moving averages against the periods. The seasonal and irregular components have thus been smoothened out. Figure 8 demonstrates how the moving average has smoothened the peaks and troughs of the original time series. The dotted line represents the cyclical and trend components.

Table 3: Computation of Ratios to Moving Averages to Quarterly Sales of a Departmental Store

(Rs. in lakhs)

Year	Quarter	Sales	4-quarter moving total	4-quarter moving average	Centred 4-quarter moving average	Ratio-to-moving average in percentage
1974	I	6.83				
	II	6.26	25.53	6.38	6.35	96.2
	III	6.11	25.24	6.31	6.25	101.3
	IV	6.33	24.77	6.19	6.13	106.7
1975	I	6.54	24.30	6.08	6.05	95.7
	II	5.79	24.12	6.08	6.06	93.1
	III	5.64	24.39	6.10	6.16	99.8
	IV	6.15	24.87	6.22	6.30	108.1
1976	I	6.81	25.51	6.38	6.45	97.2
	II	6.27	26.06	6.52	6.54	96.0
	III	6.28	26.26	6.57	6.52	102.8
	IV	6.70	25.90	6.48	6.39	109.7
1977	I	7.01	25.25	6.31	6.25	94.6
	II	5.91	24.71	6.18	6.04	93.2
	III	5.63	23.65	5.91	5.86	105.1
	IV	6.16	23.22	5.81	5.74	103.7
1978	I	5.95	23.70	5.74	6.00	91.3
	II	5.48	24.34	6.08		
	III	6.11				
	IV	6.80				

The next step is to calculate the percentage of the actual value to the moving average value for each quarter in the time series having a four-quarter moving average entry.

Notes

One should note at this stage that the first two and the last two quarters do not have the corresponding moving averages. This step allows us to recover the seasonal component for the quarters.

The fourth step is to collect all the percentages of actual to moving average values and to arrange them by quarter. We then calculate the mean for each quarter. These computations have been shown in Table 4.

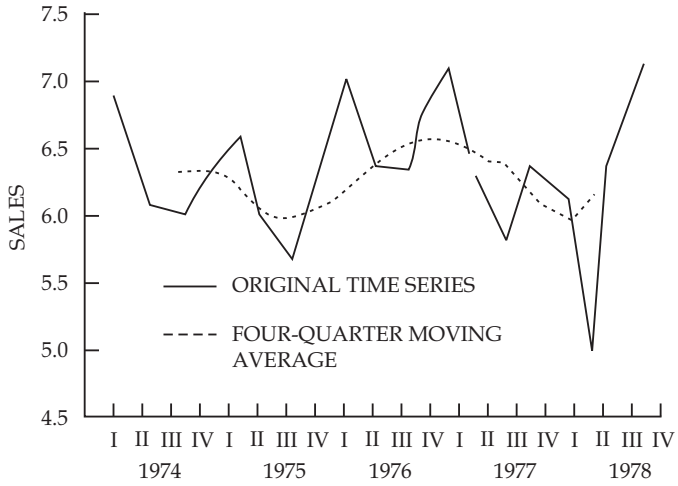


Figure 8: Moving Average: Smoothening the Original Data in Table 3

Table 4: Computation of Seasonal Index for Data Given in Table 3

Year	Quarter I	Quarter II	Quarter III	Quarter lV
1974	—	—	96.2	101.3
1975	106.7	95.7	93.1	99.8
1976	108.1	97.2	96.0	102.8
1977	109.7	94.6	93.2	105.1
1978	103.7	91.3	—	—
Mean	**107.1**	**94.7**	**94.6**	**102.2**
Seasonal Index	**107.5**	**95.1**	**95.0**	**102.5**

The final step is to adjust these quarterly means slightly. One can find out that the sum of the four quarterly means is 398.5 (from Table 4). However, the base for an index is 100 and therefore the four quarterly means should total 400 so that their mean is 100. To correct this error, we multiply each of the quarterly indices in Table 4 by an adjusting constant, the constant being computed by dividing the desired sum of indices (400) by the actual sum (398.5). In this case, this constant is 1.0038. Thus, we get the seasonal indices for each of the quarters. This has been given in the last row of Table 4.

There are a few other methods for computing the seasonal index. One of them is the link-relative method. However, the computation of seasonal indices by this method is slightly complicated. The steps involved in this method are as follows:

1. Calculate the seasonal link relatives for each seasonal value by the following formula:

$$\text{Link relative} = \frac{\text{Current seasons' value}}{\text{Previous season's value}} \times 100$$

2. Calculate the average of the link relatives for each season.

Notes

3. Convert these averages into chain relatives on the basis of the first season.
4. Calculate the chain relative of the first season on the basis of the last season.
5. A correction is applied to each of the relatives that have been computed in the earlier step. For this correction, the chain relative of the first season calculated by the first method is deducted from the chain relative of the first season calculated by the second method. The difference is divided by the number of seasons in a year. The resulting figure multiplied by 1, 2, 3 etc. is deducted respectively from the chain relatives of the 2nd, 3rd, 4th, etc.
6. The seasonal indices are obtained when the corrected chain relatives are expressed as percentage of their relative averages.

Example 3 : Calculate the seasonal index for the data given in Table 3 by the link-relative method.

Solution : The computation of seasonal indices has been explained below:

Table 5: Computation of Seasonal Indices by Link Relative Method

	Quarters			
Year	**I**	**II**	**III**	**IV**
1974	—	91.7	97.6	103.6
1975	103.3	88.5	97.4	109.0
1976	110.7	92.1	100.2	106.7
1977	104.6	84.3	95.3	109.3
1978	96.6	92.1	111.5	111.3
Mean chain	103.8	89.7	100.4	108.0
Relatives	100	$\frac{100 \times 89.7}{100}$ = 89.7	$\frac{89.7 \times 104.4}{100}$ = 90.1	$\frac{90.1 \times 108.0}{100}$ = 97.3
Corrected chain relatives	100.0	89.7 – 0.25 = 87.45	90.1 – 0.5 = 89.6	97.3 – 0.75 = 96.55
Corrected Seasonal index	100.0	$\frac{87.45}{93.4} \times 100$ = 93.6	$\frac{89.6}{93.4} \times 100$ = 95.9	$\frac{96.55}{93.4} \times 100$ = 103.4

The correction factor has been calculated as follows:

Chain relative of the first quarter on the basis of first quarter = 100.0

Chain-relative on the basis of the last quarter $= \frac{103.8 \times 97.3}{100}$

= 101.0

The difference between these chain relatives = 101.0 –100.0

= 1.0

Difference per quarter $= \frac{1.0}{4}$

= 0.25

Notes

Seasonal indices have been corrected as follows:

$$\text{Average of chain relatives} = \frac{100.0 + 87.45 + 89.6 + 96.55}{4}$$

$$= 93.4$$

$$\text{Corrected seasonal index} = \frac{\text{Corrected chain relative}}{93.4} \times 100$$

The other alternative method for determining the seasonal indices is the ratios-to-trend method. This method assumes that seasonal variation for a given season is a constant fraction of the trend. With the basic multiplicative model, O = TSCI, it is argued that the trend can be eliminated by dividing each observation by its corresponding trend value. The ratios resulting from this computation compose SCI. Each of these ratios to trend is a one-based relative, that is pure number with a unity base. Next, an average is computed for each season. This averaging process eliminates cyclical and random (irregular) fluctuations from the ratios to trend. Thus, these averages of ratios to trend contain only the seasonal component. These averages, therefore, constitute the seasonal indices. However, slight corrections can be incorporated in order to adjust these ratios to average to unity.

A simplest but a crude method of computing a seasonal index is to calculate the average value for each season, and express the averages as percentages so that all the seasonal percentages can add up to 100 multiplied by the number of seasons.

The seasonal indices are used to remove the seasonal effects from a time series. Before identifying either the trend or cyclical components of a time series, one must eliminate the seasonal variation. To do this, we divide each of the actual values in the series by the appropriate seasonal index. Once the seasonal effect has been eliminated, the deseasonalized values that remain in the series reflect only the trend, cyclical, and irregular components of the time series. With the help of the deseasonalized values we can project the future.

Uses of Seasonal Variations

The following are the main uses of seasonal variations:

(a) **Knowledge of the Pattern of Change:** A study of the seasonal variations helps in determining the pattern of the change. It can be known whether the change is stable or gradual or abrupt.

(b) **Helps in the Study of Cyclical Fluctuations:** The cyclical and irregular variations can be accurately studied only after eliminating seasonal components from a time series.

(c) **Aid of Policy Decisions:** The seasonal variations aid in formulating policy decisions. These are also useful in planning future variations. For example, the manufacturer may decide to cut the prices during slack season and providing incentives in the off-season. They may also incur huge expenditure in advertising off-seasonal use of the product.

(d) **Knowledge of the Nature of Change:** The study of seasonal variations provides a better understanding of the nature of variations. For example, in the absence of the knowledge of seasonal variations a seasonal upswing may be mistaken as an indication of better business conditions. Similarly, a seasonal slump may be mistaken as an indication of deterioration in business conditions. While in fact both these changes are seasonal and not of permanent nature.

Notes

Distinction between Cyclical and Seasonal Variations

The following are the main points of distinction between seasonal and cyclical variations:

(i) **Duration of Variations:** Cyclical variations have a duration of two to fifteen years, whereas seasonal variations have a duration of one year only.

(ii) **Degree of Accuracy:** Cyclical variations cannot be accurately estimated because of lack of their regularity whereas seasonal variations can be estimated with a high degree of accuracy.

(iii) **Regularity:** There is no regularity in the periodicity of cyclical variations whereas there is regular periodicity in seasonal variations.

(iv) **Causes of Variations:** The main causes of cyclical variations are economic whereas seasonal variations take place because of weather conditions and customs and traditions.

(v) **Activities of Preceding Variations:** Cyclical variations depend upon the activities of the preceding period whereas seasonal variations do not depend on the activities of preceding period.

Irregular Variation

The last component of a time series is the irregular variation. After eliminating the trend, cyclical, and seasonal variations from a time series, we have an unpredictable element left in the series. Irregular variation, generally, occurs over a short interval of time period and follows a random pattern. For example, a strike in an industrial unit may push down its production and consequently, the sales. Some other causes for these variations are flood draught, fire, war or other unforeseeable events.

Because of the unpredictability of irregular variation, attempt has not been made to study it mathematically. However, we can often isolate its causes, although in some situations it is difficult to identify such causes. But it should be noted that over a period of time, these random fluctuations tend to counteract each other and thus we may have a time series free of irregular variation.

22.3 An Illustration Involving all Components

As an illustration for studying all the components of a time series, we shall work out a problem involving all the components. An engineering firm producing farm equipments wants to predict future sales based on the analysis of its past sales pattern. The sales effected by the firm during the past five years is given in Table 6.

Table 6: Quarterly Sales of an Engineering Firm during 1975 to 1979

(Rs. in lakhs)

	Quarters			
Year	**I**	**II**	**III**	**IV**
1975	5.5	5.4	7.2	6.0
1976	4.8	5.6	6.3	5.6
1977	4.0	6.3	7.0	6.5
1978	5.2	6.5	7.5	7.2
1979	6.0	7.0	8.4	7.7

The procedure involved in this study consists of:

1. deseasonalizing the time series,
2. fitting the trend line, and
3. identifying the cyclical variation around the trend line.

The steps involved in deseasonalizing the time series are given in Table 7 and Table 8. These steps have been already discussed in Seasonal variation.

Notes

Table 7: Computation of Ratio-to-Moving Averages to Quarterly Sales Data of Table 6

Year	Quarter	Actual sales	Centred 4-quarter moving total	Centred 4-quarter moving average	Ratio-to moving average in percentage
1975	I	5.5			
	II	5.4			
	III	7.2	23.8	6.0	120.0
	IV	6.0	23.5	5.9	101.7
1976	I	4.8	23.2	5.8	82.8
	II	5.6	22.5	5.6	100.0
	III	6.3	21.9	5.5	114.5
	IV	5.6	21.9	5.5	101.8
1977	I	4.0	22.6	5.7	70.2
	II	6.3	23.4	5.9	06.8
	III	7.0	24.4	6.1	114.8
	IV	6.5	25.1	6.3	103.2
1978	I	5.2	25.5	6.4	81.3
	II	6.5	26.1	6.5	100.0
	III	7.5	26.8	6.7	111.9
	IV	7.2	27.5	6.9	104.3
1979	I	6.0	28.2	7.1	84.5
	II	7.0	28.9	7.2	97.2
	III	8.4			
	IV	7.7			

Table 8: Computation of Seasonal Indices for Quarterly Sales Data of Table 6.

	Quarters			
Year	**I**	**II**	**III**	**IV**
1975	–	–	120.0	101.7
1976	82.8	100.0	114.5	101.8
1977	70.2	106.8	114.8	103.2
1978	81.3	100.0	111.9	104.3
1979	84.5	97.2	–	–
Mean	**79.7**	**101.0**	**115.3**	**102.8**
Seasonal index	**79.9**	**101.3**	**115.6**	**103.2**

Notes

Table 9: Deseasonalized Sales of the Engineering Firm

Year	Quarter	Actual sales	Seasonal index 100	Deseasonalized sales
1975	I	5.5	0.799	6.9
	II	5.4	1.013	5.3
	III	7.2	1.156	6.2
	IV	6.0	1.032	5.8
1976	I	4.8	0.799	6.0
	II	5.6	1.013	5.5
	III	6.3	1.156	5.4
	IV	5.6	1.032	5.4
1977	I	4.0	0.799	5.0
	II	6.3	1.013	6.2
	III	7.0	1.156	6.0
	IV	6.5	1.032	6.3
1978	I	5.2	0.799	6.5
	II	6.5	1.013	6.4
	III	7.5	1.156	6.5
	IV	7.2	1.032	7.0
1979	I	6.0	0.799	7.5
	II	7.0	1.013	6.9
	III	8.4	1.156	7.3
	IV	7.7	1.032	7.5

The second step in identifying the components of the time series is to develop the trend line. For this purpose, we use the least squares technique to the deseasonalized time series. Table 9 gives the deseasonalized time series.

With the values from Table 9. we now find the equation for the linear trend. These computations have been shown in Table 10.

Table 10: Computations of the Trend from the Data of Table 9.

Year	Quarter	Deseasonalized sales (Y)	X*	X^2	XY
1975	I	6.9	– 19	361	– 131.1
	II	5.3	– 17	289	– 90.1
	III	6.2	– 15	225	– 93.0
	IV	5.8	– 13	169	– 75.4
1976	I	6.0	– 11	121	– 66.0
	II	5.5	– 9	81	– 49.5
	III	5.4	– 7	49	– 37.8
	IV	5.4	– 5	25	– 27.0
1977	I	5.0	– 3	9	– 15.0
	II	6.2	– 1	1	– 6.2
	III	6.0	1	1	6.0
	IV	6.3	3	9	18.9

Notes

1978	I	6.5	5	25	32.5
	II	6.4	7	49	44.8
	III	6.5	9	81	58.5
	IV	7.0	11	121	77.0
1979	I	7.5	13	169	97.5
	II	6.9	15	225	103.5
	III	7.3	17	289	124.1
	IV	7.5	19	361	142.5
Total		**125.6**		**2660**	**114.2**

* Since there are even number of periods in the series, we have assigned modified values to X.

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{114.2}{2660.0} = 0.04$$

$$a = \frac{\Sigma Y}{n} = \frac{125.6}{20} = 6.3$$

Thus, the straight line trend equation is: $Y = a + bX$

that is, $Y_{cal} = 6.3 + 0.04\ X$

We have now been able to identify the seasonal and trend components in the time series. It is now required to find the cyclical variation around the trend line. This component is identified by measuring deseasonalized variation around the trend line. The cyclical variation is computed with the help of the residual method and the same is given in Table 11.

Table 11: Computations of Cyclical Variation

Year	Quarter	Deseasonalized sales (Y)	$Y_{cal} = a + bX$*	Per cent of trend $\frac{Y}{Y_{cal}} \times 100$
1975	I	6.9	5.54	124.5
	II	5.3	5.62	94.3
	III	6.2	5.70	108.8
	IV	5.8	5.78	100.3
1976	I	6.0	5.86	102.4
	II	5.5	5.94	92.6
	III	5.4	6.02	89.7
	IV	5.4	6.10	88.5
1977	I	5.0	6.18	80.9
	II	6.2	6.26	99.0
	III	6.0	6.34	94.6
	IV	6.3	6.42	98.1
1978	I	6.5	6.50	100.0
	II	6.4	6.58	97.3
	III	6.5	6.66	97.6
	IV	7.0	6.74	103.9

Notes

1979	I	7.5	6.82	111.0
	II	6.9	6.90	100.0
	III	7.3	6.98	104.6
	IV	7.5	7.06	106.2

The irregular variation is assumed to be short-term and relatively insignificant. We have, thus, described the time series in this problem using the trend, cyclical, and seasonal components. Figure 22.9 represents the original time series, its four-quarter moving average (containing the trend and cyclical components), and the trend line.

* Appropriate value of X is taken as given in Table 22.10.

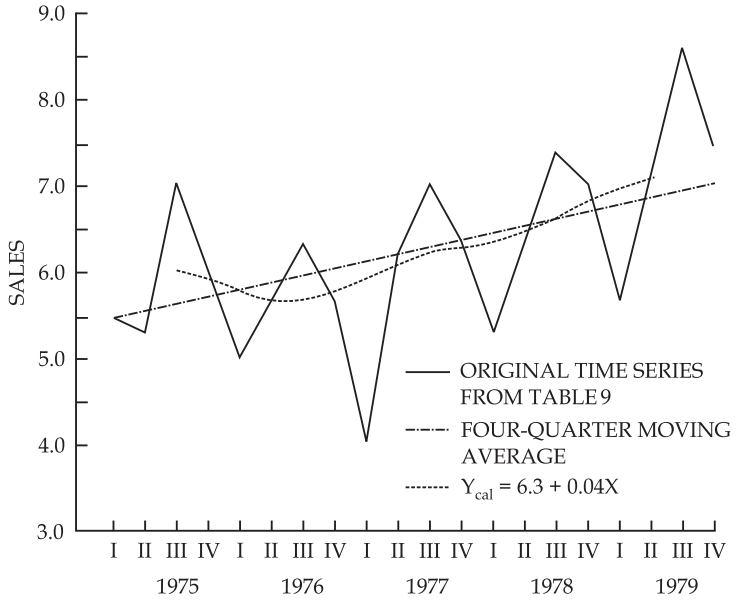


Figure 9: Original Time Series. Trend Line, and Four Quarter Moving Average for Sales Data of Table 6.

Suppose, now, the management of the engineering firm is interested in estimating the sales for the second and third quarters of 1980. Following procedure, then, will have to be adopted to estimate these figures:

$$Y_{cal} = a + bX$$

i.e. $Y = 6.3 + 0.04\ (23)$

(2^{nd} quarter, 1980)

$= 7.22$

and $Y = 6.3 + 0.04\ (25)$

(3^{rd} quarter, 1980)

$= 7.30$

Thus, the deseasonalized sales estimates for second and third quarters of 1980 are Rs. 7.22 lakhs and Rs. 7.30 lakhs, respectively. These estimates will now have to be seasonalized for the second and third quarters respectively. This is done in the following way:

1980-second quarter:

Seasonalized sales estimate $= 7.22 \times 1.013$

$= 7.31$

Notes

1980-third quarter:

Seasonalized sales estimate = 7.30 × 1.156

= 8.44

On the basis of the above analysis, the sales estimates of the engineering firm for the second and third quarters of 1980 are Rs. 7.31 lakhs and Rs. 8.44 lakhs, respectively. It should be noted here that these estimates have been obtained by taking the trend and seasonal variations into account. Further the cyclical and irregular components have not been taken into account in these estimates.

The procedure described earlier for the cyclical variation will only help us to study the past behaviour and does not help us in predicting the future behaviour. As stated in the earlier section, the irregular variations cannot be studied mathematically.

Notes Time series analysis is helpful in studying the present fluctuations in the economic variables like, national income, cost, prices, production, etc. It enables us to know achievements and failures regarding a particular variable.

Self-Assessment

1. Fill in the blanks:

(*i*) A time series consists of data arranged ..... .

(*ii*) The four components of time series are ..... , ... , ... , and ..... .

(*iii*) The additive model of components is ..... .

(*iv*) Secular trend is referred for ..... trend.

(*v*) Forces of rythmic nature cause ..... .

22.4 Summary

- A series of observations recorded over time is known as a *time series*. The data on the population of a country over equidistant time points constitute a time series, e.g. the population of India recorded at the ten-yearly censuses. Some other examples of time series are: annual production of a crop, say, rice over a number of years, the wholesale price index over a number of months, the turn-over of a firm over a number of months, the sales of a business establishment over a number of weeks, the daily maximum temperature of a place over a number of days, and so on.
- The analysis of time series is of interest in several areas, such as economics, commerce, business, sociology, geography, meteorology, demography, public health, biology, and so on. The techniques of time series analysis have largely been developed by economists. Empirical investigations dealing with economic theory are largely dependent on time series analysists. Social scientists, in general, do not have the privilege of conducting studies through laboratory experimentation. Studies are to be based on time series data collected over time in such cases. For example, trade cycles are important to economists and others in business and commerce. The exact behaviour of the cycles and their causes are of interest to them. Various theories explaining the phenomena are put forward. Analysis of time series provides an important tool for testing the theories and the explanations. Consumer behaviour is studied mainly with the help of time series data.
- The second objective of time series analysis is to predict the future behaviour of a particular variable. Time series can play an important role not in making short range estimates for a year or two a head but also estimating the probable seasonal variations within a year.
- With the secular trend, the value of the variable tends to increase or decrease over a long period of time. The steady increase in the cost of living recorded by the consumer price index is an

Notes

example of secular trend. From year to year, the cost of living varies a great deal; but, if we consider a long-term period, we see that the trend is towards steady increase. Other examples of secular trend are steady increase of population over a period of time, steady growth of agricultural food production in India over the last ten to fifteen years of time.

- Seasonal variation involves patterns of change within a year, that tend to be repeated from year to year. For example, sale of umbrellas is on the increase during the months of June and July every year because of the seasonal requirement. Since these are regular patterns, they are useful in forecasting the future production runs.
- Secular trend represents the long-term variation of the time series. One way to describe the trend component in a time series data is to fit a line to a set of points on a graph. An approach to fit the trend line is by the method of least squares.
- In many situations, studying the secular trend of time series allows us to eliminate the trend component from the series. This makes it easier for us to study the other components of the time series. If we want to determine the seasonal variation in the sale of shoes, the elimination of the trend component gives us more accurate idea of the seasonal component.
- The trend values are regarded as normal values. These normal values provide the basis for determining the nature of fluctuations. In other words, it can be found whether the fluctuations are regular or irregular. So general tendency of the data can be analysed with the help of secular trend.
- The trend analysis facilitates the comparison of two or more time
- Cyclical variation is that component of a time series that tends to oscillate above and below the secular trend line for periods longer than one year and that they do not ordinarily exhibit regular periodicity. The periods and amplitudes may be quite irregular.
- Another method used to measure the cyclical variation is the relative cyclical residual method. In this method, the percentage deviation from the trend is found for each value.
- The cyclical variations are very helpful in studying the characteristics of fluctuations of a business. One can come to know how sensitive is the business to general cyclical influences ? The general pattern of a particular firm's production, profits, sales, raw material prices, etc. can also be known.
- The study of cyclical variations is helpful in analysing and isolating the effects of irregular fluctuations. One can come to know either the variations are unpredictable or are caused by other isolated special occurrences like floods, earthquakes, strikes, wars, etc.
- Seasonal variations are those forces affecting time series that are the result of man made or physical phenomena. The major characteristic of seasonal variations is that they are repetitive and periodic, the period is less than one year, say a week, a month or a quarter. Seasonal variations can affect a time during November is normal, it can examine the seasonal pattern in the previous years and get the information it needs.
- Seasonal variations help us to project past patterns into the future. In the case of long range decisions, secular trend analysis may be adequate. However, for short-run decisions, the ability to predict seasonal fluctuations is essential. For example, consider the case of a wholesale food dealer who wants to maintain a minimum adequate stock of all food items. The ability to predict short-run patterns, such as the demand of food items during Diwali, or at Christmas, or during the summer, is very useful to the management of the store.
- The method of the ratio-to-moving average for computing the indices of seasonal variation is a procedure whereby the different components in the series are measured and are isolated or eliminated. Subsequently, the seasonal effect is identified and expressed in percentage form. We first take a series in which seasonal pattern is suspected and plot this series on a graph to identify the recurrence of the pattern. To identify the seasonal component, the data could be in quarters or months or any other time period less than a year.
- The seasonal indices are used to remove the seasonal effects from a time series. Before identifying either the trend or cyclical components of a time series, one must eliminate the seasonal variation.

Notes

To do this, we divide each of the actual values in the series by the appropriate seasonal index. Once the seasonal effect has been eliminated, the deseasonalized values that remain in the series reflect only the trend, cyclical, and irregular components of the time series. With the help of the deseasonalized values we can project the future.

- A study of the seasonal variations helps in determining the pattern of the change. It can be known whether the change is stable or gradual or abrupt.
- The seasonal variations aid in formulating policy decisions. These are also useful in planning future variations. For example, the manufacturer may decide to cut the prices during slack season and providing incentives in the off-season. They may also incur huge expenditure in advertising off-seasonal use of the product.
- The last component of a time series is the irregular variation. After eliminating the trend, cyclical, and seasonal variations from a time series, we have an unpredictable element left in the series. Irregular variation, generally, occurs over a short interval of time period and follows a random pattern. For example, a strike in an industrial unit may push down its production and consequently, the sales. Some other causes for these variations are flood draught, fire, war or other unforeseeable events.

22.5 Key-Words

1. Conditional odds : The odds of success given some level of another variable.
2. Conditional probability : The probability of one event given the occurrence of some other event.
3. Confidence interval : An interval, with limits at either end, with a specified probability of including the parameter being estimated.
4. Confidence limits : An interval, with limits at either end, with a specified probability of including the parameter being estimated.

22.6 Review Questions

1. What is time series ? What is the need to analyse the time series ?
2. Define time series. What are the preliminary adjustments that should be made before analysing time series ?
3. What are the various components of time series ? Explain.
4. What is secular trends ? Point out the significance of its study.
5. What do you mean by cyclical variation ? What are the methods of measuring such variations.

Answers: Self-Assessment

1. (*i*) Chronologically
   (*ii*) Secular trend (T), Seasonal variations (S), Cyclical variations (C) and Irregular variations (I)
   (*iii*) T + S + C + I
   (*iv*) Cyclical variations (*v*)Seasonal variation

22.7 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 23 : Time Series Methods—Graphic, Method of Semi-averages

CONTENTS

Objectives

After reading this unit students will be able to:

- Describe Graphic Method.
- Explain Semi-average Method.

Introduction

Fitting a trend curve involves assuming that a given time series exhibits a certain trend movement which, were it not for cyclical, irregular and seasonal fluctuations would have been a linear or non-linear form. Therefore, we first assume that the data to be plotted on a graph exhibit a certain trend form (linear, parabolic or exponential) and then an attempt is made to measure this trend. Measuring a trend actually means computing the constants of the equation that we have chosen to be representative of the trend in the data. However, it should be remembered that if we choose a wrong curve for the data, then the constants of the equation computed would be wrong, and any forecasting made on the basis of this equation would be wrong.

23.1 Graphic

Freehand method is also called the graphic method in the sense that the trend line is determined by inspecting the graph of the series. According to this method, the trend values are determined by drawing freehand straight line through the time series data that is judged by the analyst to represent adequately the long-term movement in the series. Once the freehand trend line is drawn, a trend equation for the line can be approximated. This is done by first reading off the trend values of the first and the last period from the chart with reference to the freehand line. For this purpose, the first period is usually considered the origin. Thus, the trend value for the first period is the value of a for the equation. Then the difference between the trend values of the first and the last period is obtained. This difference represents the total change in variable Y throughout the whole duration of the series. Therefore, when this difference is divided by the number of periods in the series, the result represents the average change in Y per unit time period. This is the value of b in the equation. The trend line for many series may be satisfactorily drawn provided the fluctuations around the general drift are so small that the path of the trend is clearly defined. The main trouble with this method is that it is too subjective. Even an expert in this subject may draw different lines at different times for the same series. There is no formal statistical criterion whereby the adequacy of such a line can be judged.

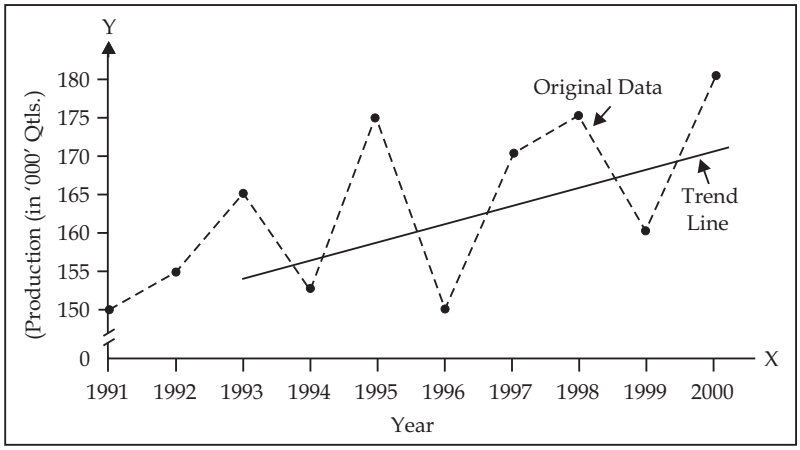
Notes

Furthermore, although this method appears simple and direct, in actuality. it is very time consuming to construct a freehand trend curve if a careful and conscientious job is to be done. For these reasons, the freehand method is not recommended for fitting a trend line.

Example 1: Draw a time series graph relating to the following data and fit the trend by freehand method.

Year :	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Production ('000 Qtls) :	150	155	165	152	174	150	170	175	160	178

Solution:



Merits of Graphic Method

(1) Simplest : This method of estimating trend is the simplest of all the methods of measuring trend. It involves no calculation at all since it is purely non-mathematical.

(2) Easy to Fit Trend : This method makes possible rapid approximations of trend that are relatively reliable. It gives a better expression of the secular movements.

(3) Flexible : This method is more flexible than rigid mathematical function, hence fits the curve more closely to the data *i.e* this method can be used even in cases where the size of the series is lengthy.

Demerits of Graphic Method

(1) Trend Values are not Definite : This method involves no calculation or mathematical formulae, hence the trend value cannot be definite. Different persons can draw different trend lines from the same original data.

(2) Subjective Method : This method is a subjective method. Being subjective, it has little value as a basis of future analysis of time series.

(3) Lack of Accuracy : This method lacks accuracy. Therefore, it is not suitable where a high degree of accuracy is desired. This method gives us an approximate picture of the tendency in the long run. This method should therefore be used only by experienced persons.

Limitations of Freehand Method

1. This method is highly subjective because the trend line depends on the personal judgment of the investigator and, therefore, different persons may draw different trend lines from the same set of data. Moreover, the work cannot be left to clerks and it must be handled by skilled and experienced people who are well conversant with the history of the particular concern.
2. Since freehand curve fitting is subjective it cannot have much value if it is used as basis for predictions.

Notes

3. Although this method seems to be quite simple, in actual practice it is very time-consuming to construct a freehand trend if a careful and conscientious job is to be done. It is only after long experience in trend fitting that a person should attempt to fit a trend line by inspection.

Notes To determine the trend values by the semi-average method, the series in question is first divided into two equal segments; then the arithmetic mean for each part is computed.

23.2 Semi-averages Method

To determine the trend values by the semi-average method, the series in question is first divided into two equal segments; then the arithmetic mean for each part is computed. Lastly, a straight line passing through these two averages is drawn to provide the trend for the series. Each average provides the trend value for the middle time period of the corresponding segment. When the time series includes an odd number of periods, there are three methods for separating the series :

(a) Add half of the value of the middle period to the total value of each part.

(b) Add the total value of the middle period to the total value of each part.

(c) Drop the value of the middle period from the computations of the averages.

With the semi-average method, the middle time unit is considered as the origin, and the values of the Y-intercept and the slope of the straight line are derived by applying the following equations :

$$a = \frac{S_1 + S_2}{t_1 + t_2} \quad \text{... (1)}$$

$$b = \frac{S_2 - S_1}{t_1(n - t_2)} \quad \text{... (2)}$$

where t_1 and t_2 refer to the number of time units for the first and second segments in the series; S_1 and S_2 refer to the corresponding partial sums respectively; and n is the total number of periods in the series.

Example 2: For the purpose of example the above method, we shall use the data given in Table 1. This series contains 15 years and is divided into two parts with 7 years in each, the middle year being dropped.

Table 1 : Computation of Trend by Semi-average Method for the Data Relating to Number of Persons Registered in an Employment Exchange in an Indian State During 1965-1979 (Figures Given in Thousands)

Year	X	No. of persons	Semi-average	Trend value
1965	– 7	10.5		11.6
1966	– 6	15.3		12.3
1967	– 5	13.5		12.9
1968	– 4	12.9	13.6	13.6
1969	– 3	11.1		14.3
1970	– 2	15.9		14.9
1971	– 1	16.0		15.6
1972	0	16.5		16.3
1973	1	16.0		17.0
1974	2	16.4		17.6
1975	3	19.9		18.3
1976	4	21.7	19.0	19.0

Notes

1977	5	18.7		19.6
1978	6	18.6		20.3
1979	7	21.5		21.0

The arithmetic mean for the first half is 13.6 and that for the second is 19.0. The straight line trend drawn through these two points is given in Figure.

Furthermore, with these two average values, we get :

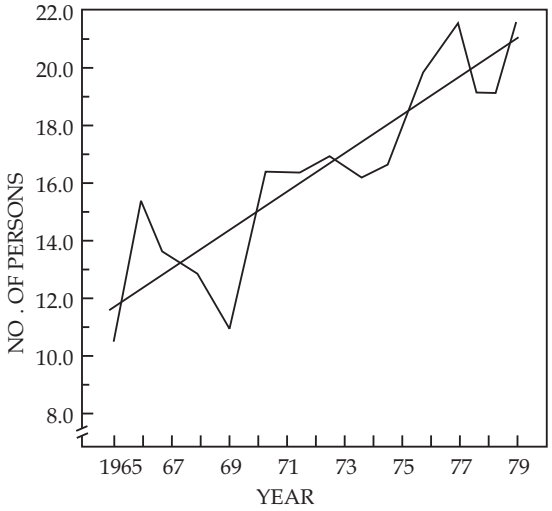
$$a = \frac{95.2 + 132.8}{7 + 7} = 16.3$$

$$b = \frac{132.8 - 95.2}{7(15 - 7)} = 0.67$$

Thus, the trend becomes

$$Y_{cal} = 16.3 + 0.67\ X$$

Trend by the Method of Semi-Averages



Trend value for each year can now be determined by substituting the X value for that period in the above trend equation. The trend values have been calculated for all the years and the same has been given in the last column of the Table 23.1.

This method of determining the trend is not a subjective one. The slope of the trend line now depends upon the difference between the averages that are computed from the original values, with each average as typical of the level of that segment of the data. However, this method is not entirely free from drawbacks. The major drawback here is due to the arithmetic mean which can be unduly affected by the extreme values in the series. If one part of the series contains more depressions or fewer prosperities than the other, then the trend line is not a true representation of the secular movements of the series. Therefore, the trend values obtained by this method are not accurate enough for the purpose either of forecasting the future trend or of eliminating the trend from the original data.

Example 3: Fit a trend line to the following data by the method of semi-averages :

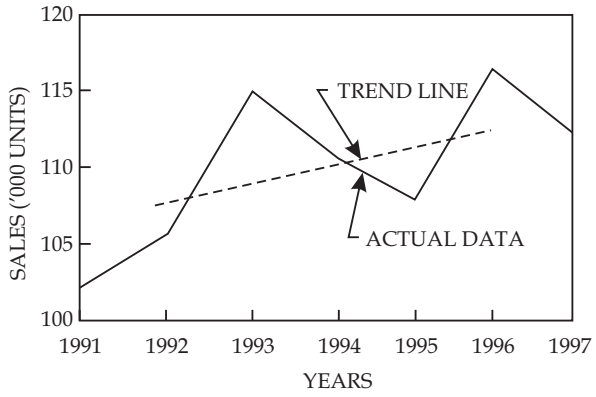
Notes

Year	Sales of Firm A (Thousand Units)
1991	102
1992	105
1993	114
1994	110
1995	108
1996	116
1997	112

Solution: Since seven years are given the middle year shall be left out and an average of first three years (1991-93) and the last three years (1995-97) shall be obtained. The average of the first three years is $\frac{102+105+114}{3} = \frac{321}{3} = 107$ and the average of the last three years is $\frac{108+116+112}{3} = \frac{336}{3} = 112$. Thus we get two points 107 and 112 which shall be plotted corresponding to their respective middle years, *i.e.*, 1992 and 1996. By joining these two points we shall obtain the required trend line. The line can be extended and can be used either for prediction or for determining intermediate values.

The actual data used and the trend line are also shown on the following graph :

TREND BY THE METHOD OF SEMI-AVERAGES



When there are even number of years like 6, 8, 10, etc., two equal parts can easily be formed and an average of each part obtained. However, when the average is to be centred there would be some problem in case the number of years is 8, 12, etc. For example, if the data relates to 1994, 1995, 1996 and 1997 which would be the middle year ? In such a case the average will be centered corresponding to 1st July 1995, *i.e.*, middle of 1995, and 1996. The following example shall illustrate the point.

Example 4: Fit a trend line by method of semi-averages for the data given below. Estimate the population for 1998. If the actual figure for that year is 520 million, account for the difference between the two figures.

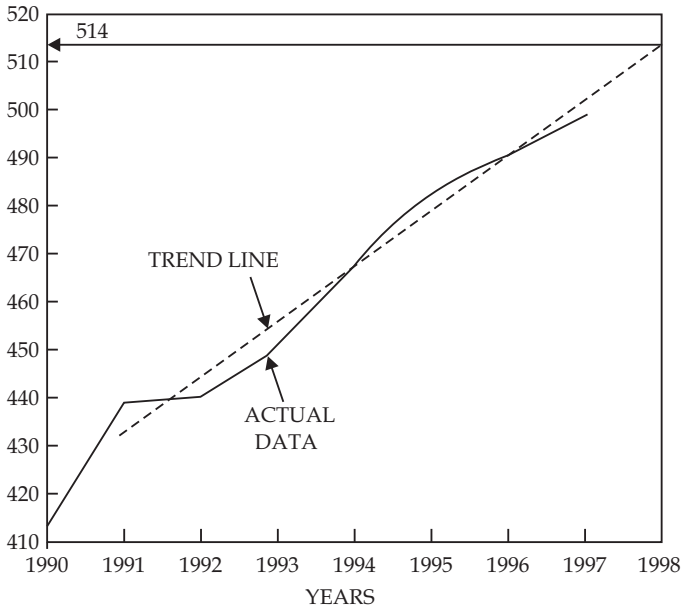
Notes

Year	Population (in million)	
1990	412	
1991	438	
1992	444	$\frac{1748}{4} = 437$
1993	454	
1994	470	
1995	482	
1996	490	$\frac{1942}{4} = 485.5$
1997	500	

Solution: The average of the first four years is 437 and that of the last four years 485.5. These two points shall be taken corresponding to the middle periods, *i.e.*, 1st July, 1991 and 1st July, 1995.

The estimate of population for 1998 by projecting the semi-average trend line is 514 million. The actual figure given to us is 520 million. The difference is due to the fact that time series analysis helps us to get the best possible estimates on certain assumptions which may come out to be true or not depending upon how far those assumptions have been realised in practice.

Trend by the Method of Semi-Averages



Example 5: The sale of a commodity in tonnes varied from January 1997 to December 1997 in the following manner :

280	300	280	280	270	240
230	230	220	200	210	200

Fit a trend line by the method of semi-averages.

Solution: **Notes**

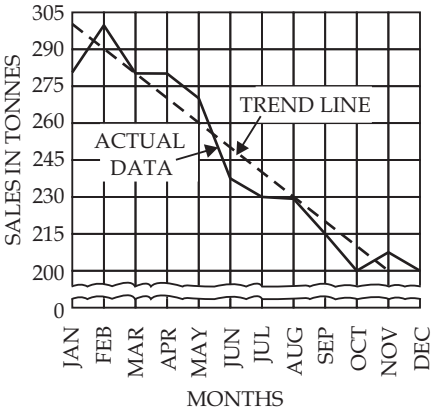
Calculation of Trend Values by the Method of Semi-Averages

Months		Sales in tonnes
January	280	
February	300	1,650, (Total of
March	280	first six months)
April	280	
May	270	
June	240	
July	230	
August	230	1,290, (Total of
September	220	last six months)
October	200	
November	210	
December	200	

Average of the first half = $\frac{1650}{6}$ = 275 tonnes.

Average of the second half = $\frac{1290}{6}$ = 215 tonnes.

These two figures, namely, 275 and 215, shall be plotted at the middle of their respective periods, *i.e.*, at the middle of March-April and that of September-October, 1997. By joining these two points we get a trend line which describes the given data.



Merits of Semi-average Method

Merits

1. This method is simple to understand compared to the moving average method and the method of least squares.
2. This is an objective method of measuring trend as everyone who applies the method is bound to get the same result (of course, leaving aside the arithmetic mistakes).

Notes

Demerits of Semi-average Method

1. **Extreme Values :** This method is based on arithmetic mean. Therefore, it is greatly affected by extreme values of the items.
2. **Straight-line Relationship :** This method assumes a straight-line relationship between the two points plotted on the graph, regardless of the fact whether such relationship exists or not.
3. **Influence of Cycle :** In this method, there is no assurance that the influence of cycle is eliminated. The danger is greater when the time period represented by average is small.

   In short, despite of above demerits, this method is definitely better than freehand curve method.

Limitations of Semi-average Method

1. This method assumes straight line relationship between the plotted points regardless of the fact whether that relationship exists or not.
2. The limitations of arithmetic average shall automatically apply. If there are extremes in either half or both halves of the series, then the trend line is not a true picture of the growth factor. This danger is greatest when the time period represented by the average is small. Consequently, trend values obtained are not precise enough for the purpose either of forecasting the future trend or of eliminating trend from original data.

   For the above reasons if the arithmetic averages of the data are to be used in estimating the secular movement, it is sometimes better to use moving averages than semi-averages.

Self-Assessment

1. Fill in the blanks:

(*i*) Graphic method is flexible and can be used for linear as well as ..... trends.

(*ii*) The values of the time series are plotted on a graph paper in the form of a ..... .

(*iii*) In graphic method, the seasonal and cyclical and irregular variations are ..... .

(*iv*) Semi-average means average of the two ..... of the series.

(*v*) Semi-average method is based on arithmetic ..... .

23.3 Summary

- Measuring a trend actually means computing the constants of the equation that we have chosen to be representative of the trend in the data. However, it should be remembered that if we choose a wrong curve for the data, then the constants of the equation computed would be wrong, and any forecasting made on the basis of this equation would be wrong.
- Freehand method is also called the graphic method in the sense that the trend line is determined by inspecting the graph of the series. According to this method, the trend values are determined by drawing freehand straight line through the time series data that is judged by the analyst to represent adequately the long-term movement in the series. Once the freehand trend line is drawn, a trend equation for the line can be approximated. This is done by first reading off the trend values of the first and the last period from the chart with reference to the freehand line. For this purpose, the first period is usually considered the origin. Thus, the trend value for the first period is the value of *a* for the equation. Then the difference between the trend values of the first and the last period is obtained.
- The trend line for many series may be satisfactorily drawn provided the fluctuations around the general drift are so small that the path of the trend is clearly defined. The main trouble with this method is that it is too subjective. Even an expert in this subject may draw different lines at different times for the same series. There is no formal statistical criterion whereby the adequacy of such a line can be judged. Furthermore, although this method appears simple and direct, in

Notes

actuality. it is very time consuming to construct a freehand trend curve if a careful and conscientious job is to be done. For these reasons, the freehand method is not recommended for fitting a trend line.

- This method of estimating trend is the simplest of all the methods of measuring trend. It involves no calculation at all since it is purely non-mathematical.
- This method is more flexible than rigid mathematical function, hence fits the curve more closely to the data *i.e* this method can be used even in cases where the size of the series is lengthy.
- This method lacks accuracy. Therefore, it is not suitable where a high degree of accuracy is desired. This method gives us an approximate picture of the tendency in the long run. This method should therefore be used only by experienced persons.
- This method is highly subjective because the trend line depends on the personal judgment of the investigator and, therefore, different persons may draw different trend lines from the same set of data. Moreover, the work cannot be left to clerks and it must be handled by skilled and experienced people who are well conversant with the history of the particular concern.
- Although this method seems to be quite simple, in actual practice it is very time-consuming to construct a freehand trend if a careful and conscientious job is to be done. It is only after long experience in trend fitting that a person should attempt to fit a trend line by inspection.
- Lastly, a straight line passing through these two averages is drawn to provide the trend for the series. Each average provides the trend value for the middle time period of the corresponding segment. When the time series includes an odd number of periods.
- This method of determining the trend is not a subjective one. The slope of the trend line now depends upon the difference between the averages that are computed from the original values, with each average as typical of the level of that segment of the data. However, this method is not entirely free from drawbacks. The major drawback here is due to the arithmetic mean which can be unduly affected by the extreme values in the series. If one part of the series contains more depressions or fewer prosperities than the other, then the trend line is not a true representation of the secular movements of the series. Therefore, the trend values obtained by this method are not accurate enough for the purpose either of forecasting the future trend or of eliminating the trend from the original data.
- This method of determining the trend is not a subjective one. The slope of the trend line now depends upon the difference between the averages that are computed from the original values, with each average as typical of the level of that segment of the data. However, this method is not entirely free from drawbacks. The major drawback here is due to the arithmetic mean which can be unduly affected by the extreme values in the series. If one part of the series contains more depressions or fewer prosperities than the other, then the trend line is not a true representation of the secular movements of the series. Therefore, the trend values obtained by this method are not accurate enough for the purpose either of forecasting the future trend or of eliminating the trend from the original data.
- This method is simple to understand compared to the moving average method and the method of least squares.
- This is an objective method of measuring trend as everyone who applies the method is bound to get the same result (of course, leaving aside the arithmetic mistakes).
- This method assumes a straight-line relationship between the two points plotted on the graph, regardless of the fact whether such relationship exists or not.
- In this method, there is no assurance that the influence of cycle is eliminated. The danger is greater when the time period represented by average is small.

Notes

23.4 Key-Words

1. Measures of location : Another term for measures of central tendency.
2. Median (Med) : The score corresponding to the point having 50% of the observations below it when observations are arranged in numerical order.
3. Median location : The location of the median in an ordered series.

23.5 Review Questions

1. Explain briefly the graphic method for determining the trend.
2. What are the merits and demerits of graphic method ?
3. Describe the Semi-average method. How the method of Semi-averages help analysing a Time Series.
4. What are the limitations of graphic method ?
5. Explain the Semi-average method of determing trend.

Answers: Self-Assessment

1. (*i*) non-linear (*ii*) histogram (*iii*) eliminated
   (*iv*) halves (*v*) mean

23.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 24: Time Series Methods—Principle of Least Square and Its Application

CONTENTS

Objectives

After reading this unit students will be able to:

- Explain the Principle of Least Square and its Applications.
- Know Merits, Demerits and Limitations of the Method of Least Square.

Introduction

The earlier discussed methods for trend analysis have certain defects, particularly in providing a satisfactory projection for the future. To overcome this defect, a convenient method is to follow a mathematical approach. The device for getting an objective fit of a straight line to a series of data is the least squares method. It is perhaps the most commonly employed and a very satisfactory method to describe the trend. The Mark off theorem states that for a given condition, the line fitted by the method of least squares is the line of "best" fit in a well-defined sense. The term "best" is used to mean that the estimates of the constants *a* and *b* are the best linear unbiased estimates of those constants.

24.1 Principle of Least Square and its Application

In the least squares method, the sum of the vertical deviations of the observed values from the fitted straight line is zero. Secondly, the sum of the squares of all these deviations is less than the sum of the squared vertical deviations from any other straight line. The method of least squares can be used for fitting linear and non-linear trends as well.

To determine the values of *a* and *b* in a linear equation by the least squares method, we are required to solve the following two normal equations simultaneously:

$$\sum Y = an + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

In the case of time series analysis, the solution of *a* and *b* from these two equations is simplified by using the middle of the series as the origin. Since the time units in a series are usually of uniform duration and are consecutive numbers, when the middle point is taken as the origin, the sum of time units, *i.e.*, $\sum X$, will be zero. As a result, the above two normal equations reduce to:

$$\sum Y = an$$

$$\sum XY = b\sum X^2$$

Notes

Therefore, we can get

$$a = \frac{\sum Y}{n} \quad \ldots (1)$$

$$b = \frac{\sum XY}{\sum X^2} \quad \ldots (2)$$

From the above expressions, we can state that a, the value of Y at the origin, is the arithmetic mean of the Y variable. The value of b, of course, is the average amount of change in the trend values per unit of time.

It should be noted that in computing the trend it is convenient to use the middle of the series as the origin. If the series contains an odd number of periods, the origin is the middle of the given period. If an even number of periods is involved, the origin is set between the two middle periods.

It is important to recognize that the least squares technique requires that the type of line desired be specified. Once this has been done, the technique generates the line of best fit of that type, that is, the least squares line. Thus, for any set of data, one could generate: (a) a least squares straight line, (b) a geometric least squares straight line, and (c) a least squares second-degree parabola. If one were to compute each of the above and determine which of the three has the least sum of squared deviations, the least squares criterion could also be used to indicate which type of line provides the best fit.

In case of a straight line trend, for any set of data, it is sufficient to compute $\sum Y$, $\sum X$, $\sum XY$, and $\sum X^2$. We shall generate these values for the data in Example 1.

Example 1: Fit a linear trend to the data given in Table 24.1. In this Example, the middle point, *i.e.*, the year 1973, is taken as the origin, thus simplifying our computations. Therefore, the coefficients for the least squares straight line are:

$$a = \frac{\sum Y}{n} = \frac{206.53}{11} = 18.78$$

$$b = \frac{\sum XY}{\sum X^2} = \frac{-23.21}{110} = -0.21$$

Table 24.1: Annual Sales of Electronic Calculators by an Indian Manufacturer

(Rs. in lakhs)

Year	Y	X	X^2	XY
1968	20.15	– 5	25	– 100.75
1969	19.49	– 4	16	– 77.96
1970	19.41	– 3	9	– 58.23
1971	19.54	– 2	4	– 39.08
1972	18.74	– 1	1	– 18.74
1973	18.00	0	0	0
1974	18.44	1	1	18.44
1975	18.81	2	4	37.62
1976	18.29	3	9	54.87
1977	17.68	4	16	70.72
1978	17.98	5	25	89.90
Total	**206.53**	**0**	**110**	**– 23.21**

Notes

To obtain the value of the trend line for any given period, we simply substitute the value of X for that year in the equation $Y_{cal} = 18.78 - 0.21\ X$, where Y_{cal} is the calculated value of Y. As an example, for the year 1973, X = 0.

Thus

$$Y_{cal}\,(1973) = 18.78 - 0.21 \times 0$$
$$= 18.78$$

As has been stated earlier, our interest in trend analysis is not only confined to determining the growth pattern of a series in the past but also concerned with forecasting future trend values. To forecast the trend values, we use the same technique that was used to compute the trend values for the periods covered by the series. For example, if we are interested in forecasting the trend value of annual sale of electronic calculators for 1980, we observe that 1980 is 7 years ahead of the origin of the trend equation we have established above, and thus

$$Y_{est}\,(1980) = 18.78 - 0.21 \times 7$$
$$= 17.31$$

We may, thus, state that, under the assumption that the same trend factors that produced the trend equation for 1968 and 1978 will remain operative, the average annual sales of electronic calculators by an Indian manufacturer will have a trend value of Rs. 17.31 lakhs during 1980.

Notes The straight line trend studied earlier is appropriate when there is reason to believe that the time series is changing, on the average, by equal absolute amounts in each time period.

However, in many situations the growth is such that the absolute amounts of the Y variable increases more rapidly in later time periods than in earlier ones. When this is the case, if a trend is to be fitted to the original data on the natural graph, a curvilinear instead of the simple linear description would be necessary. However, the same data when plotted on a semilogarithmic scale, may very often reveal linear average relationship. In other words, a straight line would seem to describe the trend of the series plotted on a semilogarithmic chart, such a growth pattern in the series can be expressed by a geometric straight line.

To fit a geometric trend line, trend values are computed from the logarithms of the data (Y) instead of the original data. Thus, the trend values obtained by this method will be logarithms of trend values and could be related to logarithms of the data. However, in practice, it is more useful to take the antilogs which will give the trend values in natural numbers and that can be compared well with the original values of the series.

In fitting a straight line to the logarithms of the data by the least squares method, the procedure used is identical with that for an arithmetic straight line studied earlier. The only difference is that logarithms of the original data instead of the natural numbers are used throughout. The normal equations now are:

$$\Sigma(\log Y) = n \log a + \log b\,(\Sigma X)$$

$$\Sigma(X \log Y) = \log a(\Sigma X) + \log b\left(\Sigma X^2\right)$$

By setting the origin at the middle of the series, the formulae for the Y intercept and the slope become

$$\log a = \frac{\Sigma \log Y}{n} \qquad \ldots (3)$$

Notes

$$\log b = \frac{\sum X \log Y}{\sum X^2} \quad \text{... (4)}$$

With the values of *a* and *b*, the logarithmic straight line equation can now be written as

$$\log Y = \log a + X \log b \quad \text{... (5)}$$

The property of *b* for a geometric straight line is of considerable significance. When it has been converted into a pure percentage it defines the annual rate of growth or decline of the series. Being an abstract measure, it allows comparison of trends all described by straight lines on ratio paper, of series with different original units. Series, such as production of all kinds, population or national income, become immediately comparable, and conclusions about the direction and magnitude of economic activities can be easily drawn. This measure provides an effective device for the study of the socio-economic changes in a country.

Did u know? Mathematical curves are useful to describe the general movement of a time series, but it is doubtful whether any analytical significance should be attached to them, except in special cases.

So far we have studied the method of fitting an arithmetic or geometric straight line to a time series. But many time series are best described by curves, and not by straight lines. In such situations, the linear model does not adequately describe the change in the variable as time changes. To study such cases, we often use a parabolic curve, which is described mathematically by a second degree equation. Such a curve is given in Figure 1. The general mathematical form of the second-degree parabolic trend is

$$Y = a + bX + cX^2 \quad \text{... (6)}$$

where *a* is still the Y intercept, *b* is the slope of the curve at the origin, and *c* is the rate of change in the slope. It should be noted that, just as *b* is a constant in the first-degree curve *c* is a constant in the second-degree curve.

The *c* value determines whether the curve is concave or convex and the extent to which the curve departs from linearity.

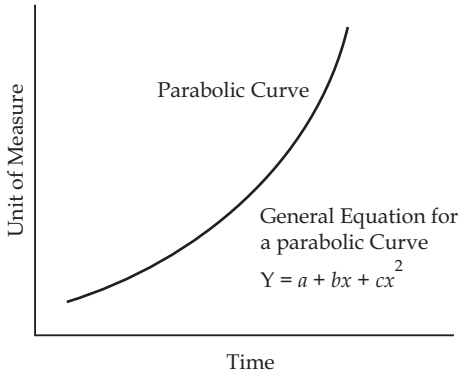


Figure 1: Form and Equation for a Parabolic Curve.

We use the least squares method to determine the second-degree equation to describe the best fit. The three normal equations are

$$\sum Y = na + b\sum X + c\sum X^2$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^2$$

$$\sum X^2 Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

Notes

If the middle of the time series is taken as origin, as done earlier, the above normal equations reduce to

$$\sum Y = na + c\sum X^2$$

$$\sum XY = b\sum X^2$$

$$\sum X^2Y = a\sum X^2 + c\sum X^4$$

Simplifying, we get:

$$c = \frac{n\sum X^2Y - \sum X^2 \sum Y}{n\sum X^4 - \left(\sum X^2\right)^2} \quad \ldots (7)$$

$$a = \frac{\sum Y - c\sum X^2}{n} \quad \ldots (8)$$

$$b = \frac{\sum XY}{\sum X^2} \quad \ldots (9)$$

Example 2: Consider India's exports of engineering goods during the years 1980 to 1986 given in Table 24.2. We shall fit a parabolic trend to describe the exports of engineering goods.

Table 24.2: India's Exports of Engineering Goods

(in crores of rupees)

Year	Exports Y	X	X^2	X^4	XY	X^2Y	Y_{cal}
1980	116.6	– 3	9	81	– 349.8	1049.4	120.28
1981	126.0	– 2	4	16	– 252.0	504.0	112.71
1982	130.0	– 1	1	1	– 130.0	130.0	137.10
1983	176.0	0	0	0	0	0.0	193.45
1984	299.0	1	1	1	299.0	299.0	281.76
1985	404.0	2	4	16	808.0	1616.0	402.03
1986	550.0	3	9	81	1650.0	4950.0	554.26
Total	**1801.6**	**0**	**28**	**196**	**2025.2**	**8548.4**	

Substituting the values from Table 24.2 into expressions for *a*, *b* and *c*, we get

$$c = \frac{7 \times 8548.4 - 28 \times 1801.6}{7 \times 196 - 28 \times 28} = 15.98$$

$$a = \frac{1801.6 - 15.98 \times 28}{7} = 193.45$$

$$b = \frac{2025.2}{28} = 72.33$$

With these values, the trend equation becomes

$$Y_{cal} = 193.45 + 72.33X + 15.98X^2$$

Notes

The trend values are obtained by substituting the X's and X²'s into the trend equation. These values have been given in the last column of Table 24.2.

Suppose we want to forecast India's exports of engineering goods for the year 1989, we observe that 1989 is 6 years ahead of the origin for the equation established above. Thus, when this value of X (= 6), is substituted into the second degree equation, we get

$$Y_{cal}\ (1989) = 193.45 + 72.33\ X + 15.98\ X^2$$
$$= 193.45 + 72.33\ (6) + 15.98\ (6)^2$$
$$= 1202.70$$

Based upon the past trend, we can conclude that India's exports of engineering goods during 1989 would be Rs. 1202.70 crores. This extra-ordinarily large forecast suggests, however, that we must be more careful in forecasting with a parabolic curve than when using a linear trend. The slope of the second degree equation is continually increasing. Therefore, the parabolic curve may become a poor estimator as we attempt to predict further into the future.

Example 3: Below are given the figures of production in (thousand quintals) of a sugar factory:

Year:	1992	1993	1994	1995	1996	1997	1998
Production (in '000 quintals):	80	90	92	83	94	99	92

(i) Fit a straight line trend to these figures.

(ii) Plot these figures on a graph and show the trend line.

(iii) Estimate the production in 2001.

Solution:

(i) FITTING THE STRAIGHT LINE TREND

Year X	**Production ('000 *qtls.*) Y**	**X**	**XY**	**X^2**	**Trend values Y_c**
1992	80	– 3	– 240	9	84
1993	90	– 2	– 180	4	86
1994	92	– 1	– 92	1	88
1995	83	0	0	0	90
1996	94	+ 1	+ 94	1	92
1997	99	+ 2	+ 198	4	94
1998	92	+ 3	+ 276	9	96
N = 7	$\Sigma Y = 630$	$\Sigma X = 0$	$\Sigma XY = 56$	$\Sigma X^2 = 28$	$\Sigma Y_c = 630$

The equation of the straight line trend is

$$Y_c = a + bX$$

Since $\Sigma X = 0$

$$a = \frac{\Sigma Y}{N},\ b = \frac{\Sigma XY}{\Sigma X^2}$$

Here $\Sigma Y = 630$, N = 7, $\Sigma XY = 56$, $\Sigma X_2 = 28$,

$$\therefore \qquad a = \frac{630}{7} = 90$$

and $$b = \frac{56}{28} = 2$$

Hence the equation of the straight line trend is

$$Y_c = 90 + 2X$$

For $X = -3, Y_c = 90 + 2(-3) = 84$

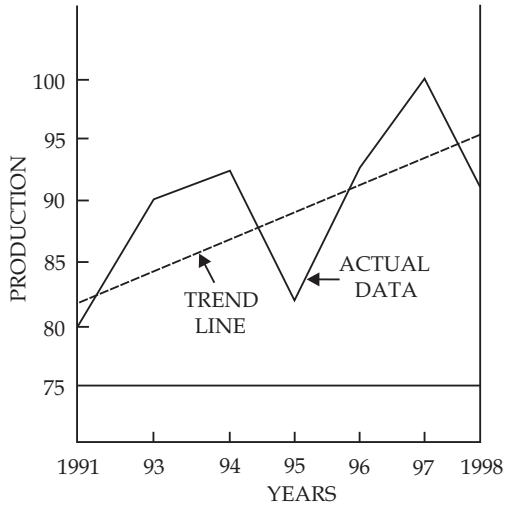
For $X = -2, Y_c = 90 + 2(-2) = 86$

For $X = -1, Y_c = 90 + 2(-1) = 88$

Similarly, by putting X = 0, 1, 2, 3, we can obtain other trend values. However, since the value of *b* is constant, only first trend value need be obtained and then if the value of *b* is positive we may continue adding the value of *b* to every preceding value. For example, in the above case for 1992 the calculated value of Y is 84. For 1993 it will be 84 + 2 = 86; for 1994 it will be 86 + 2 = 88, and so on. If *b* is negative then instead of adding we will deduct.

(ii) The graph of the above data is given below.

Linear Trend by the Method of Least Squares



For 2001 X would be + 6

$Y_{2001} = 90 + 2(6) = 102$ thousand quintals.

Example 4: Apply the method of least squares to obtain the trend value from the following data and show that $\Sigma(Y - Y_c) = 0$:

Year	Sales (in lakh tonnes)
1993	100
1994	120
1995	110
1996	140
1997	80

Also predict the sales for the year 1999.

Notes

Solution:

Calculation of Trend Values by the Method of Least Squares

Year	Sales	Deviations from middle year X	XY	X^2	Y_c	$(Y - Y_c)$
1993	100	– 2	– 200	4	114	– 14
1994	120	– 1	– 120	1	112	+ 8
1995	110	0	0	0	110	0
1996	140	+ 1	+ 140	1	108	+ 32
1997	80	+ 2	+ 160	4	106	– 26
N = 5	$\Sigma Y = 550$	$\Sigma X = 0$	$\Sigma XY = -20$	$\Sigma X^2 = 10$		$\Sigma(Y - Y_c) = 0$

1992 93 94 96 97 1998

The equation of the straight line trend is

$$Y_c = a + bX$$

Since $\Sigma X = 0$,

$$a = \frac{\Sigma Y}{N}, b = \frac{\Sigma XY}{\Sigma X^2}$$

$\Sigma Y = 550$, N = 5, $\Sigma XY = -20$, $\Sigma X^2 = 10$.

Substituting the values

$$a = \frac{550}{5} = 110$$

$$b = -\frac{20}{10} = -2$$

The required equation is

For 1993, *i.e.*, X = – 2, Y = 110 – 2 (– 2) = 114.

Since *b* is negative the other trend values will be obtained by *deducting* the value of *b* from the preceding value. Thus for 1994 the trend value will be 114 – 2 = 112 (since the value of *b* is negative). For 1999 likely sales = 100 lakh tonnes (since X would be + 5 for 1999).

Example 5: Fit a straight line trend by the method of least squares to the following data. Assuming that the same rate of change continues. What would be the predicted earnings for the year 1998 ?

Year	1987	1988	1989	1990	1991	1992	1993	1994
Earnings (Rs. lakhs.)	38	40	65	72	69	60	87	95

Do not plot the trend values on the graph.

Solution:

Notes

Fitting of Straight Line Trend by the Method of Least Squares

Year	Earnings (Rs. Lakhs) Y	Deviations from 1990-5	Deviations multiplied by 2 X	XY	X^2
1987	38	– 3.5	– 7	– 266	49
1988	40	– 2.5	– 5	– 200	25
1989	65	– 1.5	– 3	– 195	9
1990	72	– 0.5	1	– 72	1
1991	69	+ 0.5	+ 1	+ 69	1
1992	60	+ 1.5	+ 3	+ 180	9
1993	87	+ 2.5	+ 5	+ 435	25
1994	95	+ 3.5	+ 7	+ 665	49
N = 8	$\sum Y = 526$		$\sum X = 0$	$\sum XY = 616$	$\sum X^2 = 168$

$$Y_c = a + bX$$

$$a = \frac{\sum Y}{N} = \frac{526}{8} = 65.75$$

$$b = \frac{\sum XY}{\sum X^2} = \frac{616}{168} = 3.67$$

$$Y = 65.75 + 3.67\ X.$$

For 1998 X will be + 15

When X is + 15, Y will be

$$Y = 65.75 + 3.67\ (15)$$
$$= 65.75 + 55.05 = 120.8.$$

Thus the estimated earnings for the year 1998 are Rs. 120.8 lakhs.

The same result will be obtained if we do not multiply the deviations by 2. But in that case our computations would be more difficult as would be seen below

Year	Sales in thousands of rupees Y	Deviations from 1991-5 X	XY	X^2
1987	38	– 3.5	– 133.00	12.25
1988	40	– 2.5	– 100.00	6.25
1989	65	– 1.5	– 97.50	2.25
1990	72	– 0.5	– 36.00	0.25
1991	69	+ 0.5	+ 34.50	0.25
1992	60	+ 1.5	+ 90.00	2.25
1993	87	+ .5	+ 217.50	6.25
1994	95	+ 3.5	+ 332.50	12.25
N = 8	$\sum Y = 526$	$\sum X = 0$	$\sum XY = 308$	$\sum X^2 = 42.00$

Notes

$$a = \frac{\Sigma Y}{N} = \frac{526}{8} = 65.75$$

$$b = \frac{\Sigma XY}{\Sigma X^2} = \frac{308}{42} = 7.33$$

The advantage of this method is that the value of *b* gives annual increment of charge rather than 6 monthly increment, as in the first method discussed above. Hence, we will not have to double the value of *b* to obtain yearly increment. It is clear from the above illustration that in the first case the value of *b* is half of what we obtain from the second method. (*b*) was 3.67 in the first case and 7.33 in the second case.

24.2 Merits, Demerits and Limitations of the Method of Least Squares

Merits

1. This is a mathematical method of measuring trend and hence there is no possibility of subjectiveness.
2. The line obtained by this method is called *the line of best fit* because it is this line from where the sum of the positive and negative deviations is zero and sum of the squares of the deviations least, *i.e.*, $(Y - Y_c) = 0$ and $(Y - Y_c)^2$ least.

Demerits

Mathematical curves are useful to describe the general movement of a time series, but it is doubtful whether any analytical significance should be attached to them, except in special cases. It is seldom possible to justify on theoretical grounds any real dependence of a variable on the passage of time. Variables do change in a more or less systematic manner over time, but this can usually be attributed to the operation of other explanatory variables. Thus many economic time series show persistent upward trends over time due to a growth of population or to a general rise in prices, *i.e.*, national income and the trend element can to a considerable extent be eliminated by expressing these series per capita or in terms of constant purchasing power. For these reasons mathematical trends are generally best regarded as tools for describing movements in time series rather than as theories of the causes of such movements that follow, that it is extremely dangerous to use trends forecast future movements of a time series. Such forecasting, involving as it does extrapolation, can be valid only if there is theoretical justification for the particular trend as an expression of a functional relationship between the variable under consideration and the time. But if the trend is purely descriptive of past behaviour, it can give few clues about future behaviour. Sometimes the projection of a trend leads to absurd results which is *prima facie* evidence that the trend could not be maintained.

Hence, mathematical methods of fitting trend are not foolproof. In fact, they can be the source of some of the most serious errors that are made in statistical work. They should never be used unless rigidly controlled by a separate logical analysis. Trend fitting depends upon the judgement of the statistician, and a skilfully made free-hand sketch is often more practical than a refined mathematical formula.

Self-Assessment

1. Gompertz curve is a ..... curve which denoted as ..... .
2. Equation for non-linear curve is ..... .
3. The two normal equations to calculate the values of '*a*' and '*b*' in $Y = a + bX$ are ..... and ..... .
4. A polynomial of the form $Y = a + bY + CX^2$ is called a ..... .
5. The line obtained by method of least squares is known as the line of ..... .

Notes

24.3 Summary

- The device for getting an objective fit of a straight line to a series of data is the least squares method. It is perhaps the most commonly employed and a very satisfactory method to describe the trend. The Mark off theorem states that for a given condition, the line fitted by the method of least squares is the line of "best" fit in a well-defined sense. The term "best" is used to mean that the estimates of the constants *a* and *b* are the best linear unbiased estimates of those constants.
- In the least squares method, the sum of the vertical deviations of the observed values from the fitted straight line is zero. Secondly, the sum of the squares of all these deviations is less than the sum of the squared vertical deviations from any other straight line. The method of least squares can be used for fitting linear and non-linear trends as well.
- It should be noted that in computing the trend it is convenient to use the middle of the series as the origin. If the series contains an odd number of periods, the origin is the middle of the given period. If an even number of periods is involved, the origin is set between the two middle periods.
- It is important to recognize that the least squares technique requires that the type of line desired be specified. Once this has been done, the technique generates the line of best fit of that type, that is, the least squares line.
- The line obtained by this method is called *the line of best fit* because it is this line from where the sum of the positive and negative deviations is zero and sum of the squares of the deviations least, *i.e.*, $(Y - Y_c) = 0$ and $(Y - Y_c)^2$ least.
- It is seldom possible to justify on theoretical grounds any real dependence of a variable on the passage of time. Variables do change in a more or less systematic manner over time, but this can usually be attributed to the operation of other explanatory variables. Thus many economic time series show persistent upward trends over time due to a growth of population or to a general rise in prices, *i.e.*, national income and the trend element can to a considerable extent be eliminated by expressing these series per capita or in terms of constant purchasing power.
- Hence, mathematical methods of fitting trend are not foolproof. In fact, they can be the source of some of the most serious errors that are made in statistical work. They should never be used unless rigidly controlled by a separate logical analysis. Trend fitting depends upon the judgement of the statistician, and a skilfully made free-hand sketch is often more practical than a refined mathematical formula.

24.4 Key-Words

1. Homogeneity of regression : The assumption that the regression line expressing the dependent variable as a function of a covariate is constant across several groups or conditions.
2. Homogeneity of variance : The situation in which two or more populations have equal variances.

24.5 Review Questions

1. Discuss the method of least squares for the measurement of trend.
2. Write the normal equations to determine the values of *a* and *b* in the trend equation $y = a + bx$, given the *n* observations.
3. Explain the principle of least square method and its application.
4. What are the limitations of least square method ?
5. Discuss the merits and demerits of least square method.

Notes

Answers: Self-Assessment

1. (*i*) $Y_c = a + bx + cx^2$ (*ii*) $Y = a + bx$

   (*iii*) $\sum Y = Na + b\sum X$ and $\sum XY = a\sum X + b\sum X^2$

   (*iv*) halves (*v*) Best fit

24.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods – An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods – Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 25 : Methods of Moving Averages

CONTENTS

Objectives

After reading this unit students will be able to :

- Discuss the Methods of Moving Averages.
- Know Merits, Demerits and Limitations of Moving Average Method.

Introduction

When a trend is to be determined by the method of moving averages, the average value for a number of years (or month or weeks) is secured, and this average is taken as the normal or trend value for the unit of time folling at the middle of the period covered in the calculation of the average.

This method may be considered as an artificially constructed time series in which each periodic figure is replaced by the mean of the value of that period and those of a number of preceding and succeeding periods. The computation of moving averages is simple and straight-forward. The properties and the utility of discussed below.

25.1 Method of Moving Averages

Moving average method is quite simple and is used for smoothing the fluctuations in curves. The trend values obtained by this method are very much accurate. Like semi-average method, this method also employs arithmetic means of items. But here we find out the moving averages from the time series. A moving average of a time series is a new series obtained by finding out successively the average of a number of the original successive items choosen on the basis of periodicity of fluctuations, dropping off one item and adding on the next at each stage.

The moving average may be for three, four, five, six, seven, ... years and so on according to the size and the periodicity of fluctuations of the data. Suppose moving average is to be calculated for three years. We will take the average of first three years and will place it against the middle year of the three. Now leave the first item and add the next item of the series and take the average of these items and place it against the middle year of the three. We will go on in this way, taking the average after leaving one preceeding year, till the end. The formula for calculating moving averages is 3-yearly moving average.

$$\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}, \frac{d+e+f}{3}, \ldots$$

Notes

Example 1 : The data given in Table 25.1 refers to a hypothetical data assumed to have a uniform cyclical duration of 5 years and equal amplitude of 2 units. Three-year and five-year moving averages are fitted to the data. The procedure for calculating three-year moving averages is explained below :

(a) Compute the three-year moving totals. This is done by adding up the values of the first three years and centering it at the second year. This is the first three-year moving total. Then the first year value is deleted and fourth year value is included to form the second three-year moving total, which is centred at the third year. In a similar way, the computation moves through the end of the series. The three-year moving totals are entered in column 3 of Table 1.

Table 1 : Computation of Three-year and Five-year Moving Averages for the Hypothetical Data

Year (1)	Original value (2)	Three-year moving total (3)	Three-year moving average (4)	Five-year moving total (5)	Five-year moving average (6)
1	3				
2	4	12	4.0		
3	5	13	4.3	19	3.8
4	4	12	4.0	20	4.0
5	3	11	3.7	21	4.2
6	4	12	4.0	22	4.4
7	5	15	5.0	23	4.6
8	6	16	5.3	24	4.8
9	5	15	5.0	25	5.0
10	4	14	4.7	26	5.2
11	5	15	5.0	27	5.4
12	6	18	6.0	28	5.6
13	7	19	6.3	29	5.8
14	6	18	6.0	30	6.0
15	5	17	5.7	31	6.2
16	6	18	6.0	32	6.4
17	7	21	7.0	33	6.6
18	8	22	7.3	34	6.8
19	7	21	7.0		
20	6				

(b) The three-year moving averages are obtained by dividing each of the three-year moving totals by 3. These values for our hypothetical example is given in column 4 of Table 1.

A similar procedure is used to compute the five-year moving averages. In a five-year moving average, the value of each year is replaced by the mean of the value of the five successive years of which two precede and two succeed the given year. Both five-year moving totals and moving averages are centred in the middle of the respective five-year periods, with the first five-year moving total and the moving average entered in the third year. It is to be noted that in computing the moving averages for an even number of periods, the procedure is slightly more complicated. For example, the calculation of a 12-period (year or month) moving, average starts with adding up the first 12-period values in the series to form a 12-period moving total. The second moving

Notes

total is obtained by dropping the value of the first period from, and adding the value of the thirteenth period to the first moving total, and so on until all the moving total have been obtained from the series. Then each moving total is divided by 12 to get the 12-period moving averages. The moving totals and moving averages so obtained fall in between two periods. However, data that are typical of a period should be centred at the middle of the period. Thus, to compute moving averages when an even number of time units is used, an additional step is required to centre the averages at the middle of each time unit. For a 12-period moving average, centred averages are obtained by adding the two averages at a time and dividing each sum by 2. Thus, the first centred 12-period moving average would fall on the seventh period of the series, the second on the eighth period, and so on.

The results of the computations made earlier are plotted with the original data in Figure 1. Both the sets of moving averages may be considered as the statistical expression of secular trend of our hypothetical data.

With the help of Table 25.1 and Figure 1, we can make the following conclusions about the characteristics of moving average :

1. A moving average of equal length period will completely eliminate the periodic fluctuations.
2. A moving average of equal length will be linear if the series changes on the average by a constant per time unit and its fluctuations are periodic.
3. Even when the data show periodic fluctuations, a moving average of unequal length period, no matter how small the difference is between the duration of periodicity of original series, and the length of the moving average, the moving average cannot completely remove the periodic variations in the original series. The averaging process then only tends to smooth out somewhat the short-run highs and lows.

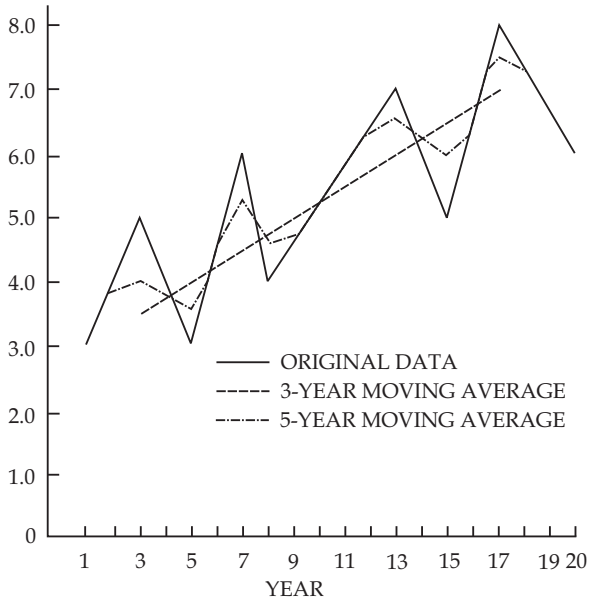


Figure 1 : Three-year and Five-year Moving Average Trends Fitted to Data in Table 1

Thus, we can see that the moving average may constitute a satisfactory trend for a series that is basically linear and that is regular in duration and amplitude. Further, there are two other disadvantages in the use of the moving average as a measure of trend. First, in computing moving averages, we loose some years at the beginning and end of the series. The second drawback is that the moving average is not represented by some mathematical formulae and, therefore, is not capable of objective future projections. Since one of the major objectives of trend analysis is that of forecasting, the moving average is no longer in wide use as a trend measure. However, the method of moving averages is a very useful technique in analyzing a time series data. First of all, in problems in which

Notes

the trend of the time series is clearly not linear and in which we are concerned only with the general movements of the time series, whether it is a trend or a cycle or both, it is customary to study the smoothing behaviour of the series by the use of moving average. Secondly, the characteristic of a moving average is the basis of seasonal analysis, which shall be discussed in a subsequent section.

Selections of Period in Moving Average Method

The most important point in the average method is the selection of period. The selection of period depends upon the periodicity of data. The period of moving average method can be divided into two parts :

(A) Odd Period of Moving Average : Odd period is the period of three, five, seven, nine years and so on. In case of odd period moving average, no difficulty is faced while placing the computed average. The determination of trend in odd period moving average can be made clear with the help of following example :

Example 2 : Calculate trend values by 3-yearly moving average from the following data :

Year :	1980	1981	1982	1983	1984	1985	1986
Sales ('000 Units) :	5	7	9	12	11	10	8

Solution :

Years	Sales (000′ Units)	3-yearly totals	3-yearly moving average
1980	5	–	–
1981	7	21	21/3 = 7.00
1982	9	28	28/3 = 9.33
1983	12	32	32/3 = 10.67
1984	11	33	33/3 = 11.00
1985	10	29	29/3 = 9.67
1986	8	–	–

The trend value is shown in Figure 2.

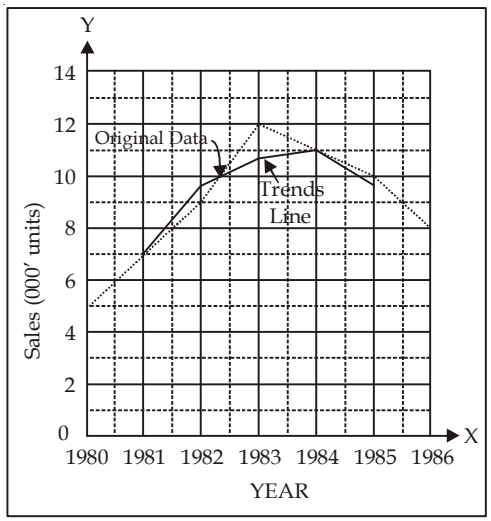


Figure 2 :

(B) Even Period of Moving Average : The procedure of calculation of moving average of even number of years, say, four, six, eight, and so on, is different from the procedure of odd number of years. Suppose moving average is to be calculated for four years. We will take the total of

Notes

first four years and will be placed in between second and third years, *i.e.*, in the middle of four years. Leaving the first year, calculate the total of next four years and so on. It is important to note that these calculated totals are placed between two years. Then we adjust these moving totals. For this we compute two yearly moving totals of four yearly moving totals. We take the total of first and second four years moving totals and write against the third year and then the second and third four yearly moving totals are totalled and written against the fourth year and so on. The two yearly moving totals of four yearly moving totals are then divided by eight and this gives us the centered four yearly moving average. The series so obtained is known as an estimate of the trend.

Did u know? Moving Averages method is that it is difficult to determine the proper period of moving average. If a wrong period is selected, there is every likelihood that conclusions may be misleading.

Characteristics of Moving Averages

The main characteristics of moving averages are as under :

(i) If the original data, when plotted on a graph, give a straight line, the moving averages will simply reproduce the original line.

(ii) If the original series gives a curve which is concave, the moving average curve will be below it.

(iii) If the original series gives a convex curve, the moving average curve will be above it.

(iv) In a series having regular fluctuations, the moving average completely eliminates them, if the period selected for it coincides with the period when the fluctuations repeat themselves.

(v) No particular period of a moving average will eliminate the fluctuations completely. But greater the period, the greater will be the reduction in the irregular fluctuations. Because the duration of business cycles always remain changing.

25.2 Merits, Demerits and Limitations of Moving Average Method

Merits of Moving Average Method

(i) Simple and Easy : This method is quite simple and easy to calculate as compared with all the mathematical methods of fitting a trend.

(ii) Flexible : This method is highly flexible in the sense that if a few more figures are added to the series, entire calculations are not changed. Only thing, we have to do is to extend the process so as to calculate further trend values.

(iii) High Degree of Accuracy : This method is associated with a high degree of accuracy and it can be made the basis for further analysis of time series.

(iv) Automatic Elimination of Fluctuations : If and when the period of moving average is equivalent to the period of the fluctuations, the cyclical fluctuations are completely eliminated.

(v) Irregular Trend : This method is considered to be the most effective method if the trend of a series is very irregular.

(vi) Practical Method : This method is most commonly used and in a very long series, it is the only practical method.

Demerits of Moving Average Method

(i) Period of Moving Average : The main limitation or demerit of this method is that it is difficult to determine the proper period of moving average. If a wrong period is selected, there is every likelihood that conclusions may be misleading.

Notes

(ii) **Ignores the Extreme Figures :** This method ignores the extreme figures *i.e.* figures for first and last few years. Then the trend value for all the years cannot be computed. Hence, perfectly accurate trend line cannot be drawn.

(iii) **Extreme Values :** In this method, moving averages are calculated by using the arithmetic average. Thus, like arithmetic average it is also affected by extreme values of the series.

(iv) **Non-linear Trend :** If the basic trend in the data is not linear this method will produce a bias in the trend. According to **Wough**, *"If the trend line is concave downward (like the side of a bowl), the value of the moving average will always be too high; if the trend line is concave downward (like the side of a derby pot), the value of the moving average will always be too low."*

(v) **Forecasting :** Since moving average is not represented by a mathematical function, this method is of little use in forecasting. Thus, it does not fulfil the basic objective of trend analysis.

(vi) **Conditions :** There are certain conditions for the use of this method. But these conditions are seldom met.

In short, despite of above cited demerits, it is becoming popular in the analysis of seasonal variations.

Limitations of Moving Average

1. Trend values cannot be computed for all the years. The longer the period of moving average, the greater the number of years for which trend values cannot be obtained. For example, in a three-yearly moving average, trend value cannot be obtained for the first year and last year, in a five-yearly moving average for the first two years and the last two years, and so on. It is often these extreme years in which we are most interested.
2. Great care has to be exercised in selecting the period of moving average. No hard and fast rules are available for the choice of the period and one has to use his own judgment.
3. Since the moving average is not represented by a mathematical function this method cannot be used in forecasting which is one of the main objectives of trend analysis.
4. Although theoretically we say that if the period of moving average happens to coincide with the period of cycle, the cyclical fluctuations are completely eliminated, but in practice. Since the cycles are by no means perfectly periodic, the lengths of the various cycles in any given series will usually vary considerably and, therefore, no moving average can completely remove the cycle. The best results would be obtained by a moving average whose period is equal to the average length of all the cycles in the given series. However, it is difficult to determine the average length of the cycle until the cycles are isolated from the series.
5. Finally, when the trend situation is not linear (a straight line) the moving average lies either above or below the true sweep of the data. Consequently, the moving average is appropriate for trend computations only when :
   (a) the purpose of investigation does not call for current analysis or forecasting.
   (b) the trend is linear, and
   (c) the cyclical variations are regular both in period and amplitudes.

Self-Assessment

Fill in the blanks–

1. The period of moving average is to be decided in the light of the ..... of the cycle.
2. Moving average method is used for smoothing the ..... in the curve.
3. The formula for calculating moving averages is ..... moving average.
4. Moving average method is ..... as compared to the method of least squares.
5. Moving averages are calculated by using the ..... average.

25.3 Summary

Notes

This method may be considered as an artificially constructed time series in which each periodic figure is replaced by the mean of the value of that period and those of a number of preceding and succeeding periods. The computation of moving averages is simple and straight-forward.

- Moving average method is quite simple and is used for smoothing the fluctuations in curves. The trend values obtained by this method are very much accurate. Like semi-average method, this method also employs arithmetic means of items. But here we find out the moving averages from the time series. A moving average of a time series is a new series obtained by finding out successively the average of a number of the original successive items choosen on the basis of periodicity of fluctuations, dropping off one item and adding on the next at each stage.
- The most important point in the average method is the selection of period. The selection of period depends upon the periodicity of data.
- Odd period is the period of three, five, seven, nine years and so on. In case of odd period moving average, no difficulty is faced while placing the computed average.
- The procedure of calculation of moving average of even number of years, say, four, six, eight, and so on, is different from the procedure of odd number of years. Suppose moving average is to be calculated for four years. We will take the total of first four years and will be placed in between second and third years, *i.e.*, in the middle of four years. Leaving the first year, calculate the total of next four years and so on. It is important to note that these calculated totals are placed between two years. Then we adjust these moving totals. For this we compute two yearly moving totals of four yearly moving totals. We take the total of first and second four years moving totals and write against the third year and then the second and third four yearly moving totals are totalled and written against the fourth year and so on. The two yearly moving totals of four yearly moving totals are then divided by eight and this gives us the centered four yearly moving average. The series so obtained is known as an estimate of the trend.
- No particular period of a moving average will eliminate the fluctuations completely. But greater the period, the greater will be the reduction in the irregular fluctuations. Because the duration of business cycles always remain changing.
- This method is highly flexible in the sense that if a few more figures are added to the series, entire calculations are not changed. Only thing, we have to do is to extend the process so as to calculate further trend values.
- The main limitation or demerit of this method is that it is difficult to determine the proper period of moving average. If a wrong period is selected, there is every likelihood that conclusions may be misleading.
- The main limitation or demerit of this method is that it is difficult to determine the proper period of moving average. If a wrong period is selected, there is every likelihood that conclusions may be misleading.
- In this method, moving averages are calculated by using the arithmetic average. Thus, like arithmetic average it is also affected by extreme values of the series.
- Trend values cannot be computed for all the years. The longer the period of moving average, the greater the number of years for which trend values cannot be obtained. For example, in a three-yearly moving average, trend value cannot be obtained for the first year and last year, in a five-yearly moving average for the first two years and the last two years, and so on. It is often these extreme years in which we are most interested.
- Although theoretically we say that if the period of moving average happens to coincide with the period of cycle, the cyclical fluctuations are completely eliminated, but in practice. Since the cycles are by no means perfectly periodic, the lengths of the various cycles in any given series will usually vary considerably and, therefore, no moving average can completely remove the cycle. The best results would be obtained by a moving average whose period is equal to the

Notes

average length of all the cycles in the given series. However, it is difficult to determine the average length of the cycle until the cycles are isolated from the series.

25.4 Key-Words

1. Independent variables : Those variables controlled by the experimenter.
2. Independent events : Events are independent when the occurrence of one has no effect on the probability of the occurrence of the other.
3. Inferential statistics : That branch of statistics that involves drawing inferences about parameters of the population(s) from which you have sampled.

25.5 Review Questions

1. Describe the moving average method for smoothing time series data. Explain how this method is applied in the isolation of trend, if an appropriate period is choosen.
2. Explain the concept and advantages of moving average method of obtaining secular trend.
3. What do you understand by time series analysis ? How the method of moving averages help analysing a time series ?
4. What are the merits and Demerits of moving average method ?
5. What are the limitations of moving average method.

Answers: Self-Assessment

1. (*i*) length (*ii*) fluctuations (*iii*) 3-yearly
   (*iv*) simple (*v*) arithmetic

25.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002

Unit 26: Theory of Probability: Introduction and Uses

Notes

CONTENTS

Objectives

After reading this unit students will be able to:

- Introduce Theory of Probability.
- Discuss the Uses of Theory of Probability.

Introduction

In day-to-day life, we all make use of the word 'probability'. But generally people have no definite idea about the meaning of probability. For example, we often hear or talk phrases like, "Probability it may rain today"; "It is likely that the particular teacher may not come for taking his class today"; "there is a chance that the particular student may stand first in the university examination"; "it is possible that the particular company may get the contract which it bid last week"; "most probably I shall be returning within a week"; "it is possible that he may not be able to join his duty". In all the above statements, the terms—possible, probably, likely, chance, etc., convey the same meaning, *i.e.*, the events are not certain to take place. In other words, there is involved an element of uncertainty or chance in all these cases. A numerical measure of uncertainty is provided by the theory of probability. The aim of the probability theory is to provide a measure of uncertainty. The theory of probability owes its origin to the study of games of chance like games of cards, tossing coins, dice, etc. But in modern times, it has great importance is decision making problems.

26.1 Introduction to Theory of Probability

We have understood the difference between descriptive and inferential statistics. The study of probability provides a basis for inferential statistics. Inferential statistics involves sample selection, computing sample statistic on the basis of the concerned sample, and then inferring population parameter on the basis of the sample statistic. We do this exercise because population parameter is unknown. We try to estimate the unknown population parameter on the basis of the known sample statistic. This procedure works on uncertainty. By applying some defined statistical rules and procedures, an analyst can assign the probability of obtaining a result. To make rational decisions, a decision maker must have a deep understanding of probability theory. This understanding enhances his capacity to make optimum decisions in an uncertain environment. This unit focuses on the basic concept of probability which will serve as the foundation of probability distributions. A sound knowledge of probability and probabilistic distributions also helps in developing probabilistic decision models.

Notes

Concept of Probability

We live in a world dominated by uncertainty. Change is the only permanent phenomenon. We can never predict the nature and direction of change in our lives. Sometimes change is planned, but more often, change is unplanned. Even in cases of planned change, it is not possible to avoid uncertainty. There is a perceived need to be accurate (up to an extent) and prepared in this uncertain environment. Our need to cope with this unavoidable uncertainty of life has led to the study of **probability theory**. There might have been many occasions when we have said that the chances are 50-50 or there is a 70% chance of India winning the match, and so on. By making these statements, we try to attach some probability of the event happening or not happening. If we look at the wider picture, all these statements are related to the concept of probability. Therefore, there is a general understanding about the concept of probability, but there is a problem in terms of its proper application-oriented understanding.

In simple words, **probability** is the likelihood or chance that a particular event will or will not occur. The theory of probability provides a quantitative measure of uncertainty or likelihood of occurrence of different events resulting from a random experiment, in terms of quantitative measures ranging from 0 to 1. This means that the probability of a certain event is 1 and the probability of an impossible event is 0. In other words, a probability near 0 indicates that an event is unlikely to occur whereas a probability near 1 indicates that an event is almost certain to occur. For example, suppose an event is the success of a new product launched. A probability 0.90 indicates that the new product is likely to be successful whereas a probability of 0.15 indicates that the product is unlikely to be successful in the market. A probability of 0.50 indicates that the product is just as likely to be successful as not.

Notes Probability is a concept that we all understand. In our daily life, we use words like chance, possibility, likelihood, and of course, probability.

Some Basic Concepts

Before we give definition of the word probability, it is necessary to define the following basic concepts and terms widely used in its study:

(1) An Experiment

When we conduct a trial to obtain some statistical information, it is called an experiment.

Examples: (i) Tossing of a fair coin is an experiment and it has two possible outcomes: Head (H) or Tail (T).

(ii) Rolling a fair die is an experiment and it has six possible outcomes: appearance of 1 or 2 or 3 or 4 or 5 or 6 on the upper most face of a die.

(iii) Drawing a card from a well shuffled pack of playing cards is an experiment and it has 52 possible outcomes.

(2) Events

The possible outcomes of a trial/experiment are called events. Events are generally denoted by capital letters A, B, C, etc.

Examples: (i) If a fair coin is tossed, the outcomes - head or tail are called events.

(ii) If a fair die is rolled, the outcomes 1 or 2 or 3 or 4 or 6 appearing up are called events.

(3) Exhaustive Events

The total number of possible outcomes of a trial/experiment are called exhaustive events. In other words, if all the possible outcomes of an experiment are taken into consideration, then such events are called exhaustive events.

Notes

Examples: (i) In case of tossing a die, the set of six possible outcomes, *i.e.*, 1, 2, 3, 4, 5 and 6 are exhaustive events.

(ii) In case of tossing a coin, the set of two outcomes, *i.e.*, H and T are exhaustive events.

(iii) In case of tossing of two dice, the set of possible outcomes are 6 × 6 = 36 which are given below:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

(4) Equally-Likely Events

The events are said to be equally-likely if the chance of happening of each event is equal or same. In other words, events are said to be equally likely when one does not occur more often than the others.

Examples: (i) If a fair coin is tossed, the events H and T are equally-likely events.

(ii) If a die is rolled, any face is as likely to come up as any other face. Hence, the six outcomes -1 or 2 or 3 or 4 or 5 or 6 appearing up are equally likely events.

(5) Mutually Exclusive Events

Two events are said to be mutually exclusive when they cannot happen simultaneously in a single trial. In other words, two events are said to be mutually exclusive when the happening of one excludes the happening of the other in a single trial.

Examples: (i) In tossing a coin, the events Head and Tail are mutually exclusive because both cannot happen simultaneously in a single trial. Either head occurs or tail occurs. Both cannot occur simultaneously. The happening of head excludes the possibility of happening of tail.

(ii) In tossing a die, the events 1, 2, 3, 4, 5 and 6 are mutually exclusive because all the six events cannot happen simultaneously in a single trial. If number 1 turns up, all the other five *(i.e.,* 2, 3, 4, 5, or 6) cannot turn up.

(6) Complementary Events

Let there be two events A and B. A is called the complementary event of B and B is called the complementary event of A if A and B are mutually exclusive and exhaustive.

Examples: (i) In tossing a coin, occurrence of head (H) and tail (T) are complementary events.

(ii) In tossing a die, occurrence of an even number (2, 4, 6) and odd number (1, 3, 5) are complementary events.

(7) Simple and Compound Events

In case of simple events, we consider the probability of happening or not happening of single events.

Example: If a die is rolled once and A be the event that face number 5 is turned up, then A is called a simple event

In case of compound events, we consider the joint occurrences of two or more events.

Notes

Example: If two coins are tossed simultaneously and we shall be finding the probability of getting two heads, then we are dealing with compound events.

(8) Independent Events

Two events are said to be independent if the occurence of one does not affect and is not affected by the occurence of the other.

Examples: (i) In tossing a die twice, the event of getting 4 in the 2nd throw is independent of getting 5 in the first throw.

(ii) In tossing a coin twice, the event of getting a head in the 2nd throw is independent of getting head in the 1st throw.

(9) Dependent Events

Two events are said to be dependent when the occurence of one does affect the probability of the occurence of the other events.

Examples: (i) If a card is drawn from a pack of 52 playing cards and is not replaced, this will affect the probability of the second card being drawn.

(ii) The probability of drawing a king from a pack of 52 cards is $\frac{4}{52}$ or $\frac{1}{13}$. But if the card drawn (king) is not replaced in the pack, the probability of drawing again a king is $\frac{3}{51}$.

Definition of Probability

The probability is defined in the following three different ways:

(1) Classical or Mathematical Definition

(2) Empirical or Relative Frequency Definition

(3) Subjective Approach.

(1) Classical or Mathematical Definition

This is the oldest and simplest definition of probability. This definition is based on the assumption that the outcomes or results of an experiment are equally likely and mutually exclusive.

According to Laplace, "Probability is the ratio of the favourable cases to the total number of equally likely cases". From this definition, it is clear that in order to calculate the probability of an event, we have to find the number of favourable cases and it is to be divided by the total number of cases. For example, if a bag contains 6 green and 4 red balls, then the probability of getting a green ball will be 6/4 + 6 = 6/10 because the total number of balls are 10 and the number of green balls is 6.

Symbolically,

$$P(A) = p = \frac{\text{Number of Favourable Cases}}{\text{Total Number of Equally Likely Cases}} = \frac{m}{n}$$

Where, P(A) = Probability of occurrence of an event A

m = Number of favourable cases

n = Total number of equally likely cases

Similarly,

$$P(\bar{A}) = q = 1 - P(A) = 1 - \frac{m}{n}$$

Where, $P(\bar{A}) = q$ = Probability of non-occurrence of an event A.

Notes

From the above definition, it is clear that the sum of the probability of happening of an event called success *(p)* and the probability of non-happening of an event called failure *(q)* is always one (1), *i. e.,* $p + q = 1$. If p is known, we can find q and if q is known, then we can find p. In practice, the value of p lies between 0 and 1, *i.e.,* $0 \le p \le 1$. To quote **Prof. Morrison, "If an event can happen in *m* ways and fail to happen in *n* ways, then probability of happening is** $\frac{m}{m+n}$ **and that of its failure to happen is** $\frac{n}{m+n}$**".**

Limitations of Classical Definition

Following are the main limitations of classical definition of probability:

(1) If the various outcomes of the random experiment are not equally-likely, then we cannot find the probability of the event using classical definition.

(2) The classical definition also fails when the total number of cases are infinite.

(3) If the actual value of N is not known, then the classical definition fails.

(2) Empirical or Relative Frequency Definition

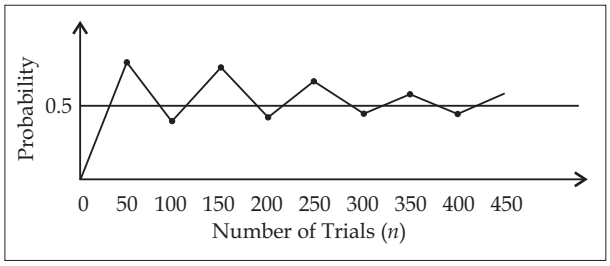
This definition of probability is not based on logic but past experience and experiments and present conditions. If vital statistics gives the data that out of 100 newly born babies, 55 of them are girls, then the probability of the girl birth will be 55/100 or 55%. **To quote Kenny and Keeping, "If event has occured *r* times in a series of *n* independent trials, all are made under the same identical conditions, the ratio *r/n* is the relative frequency of the event. The limit *of r/n* as *n* tends to infinity is the probability of the occurence of the event".**

Did u know? **Probability is the limit of the relative frquency of success in infinite sequences of trials.**

Symbolically,

$$P(A) = \underset{n \to \infty}{\text{Limit}} \frac{r}{n}$$

For example, if a coin is tossed 100 times and the heads turn up 55 times, then the relative frequency of head will be $\frac{55}{100}$ = 0.55. Similarly, if a coin is tossed 1000 times and if the head turns up 495 times, then the relative frequency will be $\frac{495}{1000}$ = 0.495. In 10,000 tosses, the head turns up 5085 then the relative frequency will be 0.5085. Thus as we go on increasing the number of trials, there is a tendency mat the relative frequency of head would approach to 0.50. The following figure illustrate the idea:



Notes

From the above figure, it is clear that as the number of trials increases, the probability of head tends to approach 0.5 and when the number of trials is infinite, *i.e.*, $n \to \infty$, the probability of getting head is equal to 0.5.

(3) Subjective Approach

According to this approach, probability to an event is assigned by an individual on the basis of evidence available to him. Hence probability is interpreted as a measure of degree of belief or confidence that a particular individual reposes in the occurrence of an event. But the main problem here is that different persons may differ in their degree of confidence even when same evidence is offered.

26.2 Uses of Theory of Probability

The theory of probability has its origin in the games of chance related to gambling such as tossing a die, tossing a coin, drawing a card from a deck of 52 cards and drawing a ball of a particular colour from a bag. But in modern times, it is widely used in the field of statistics, economics, commerce and social sciences that involve making predictions in the face of uncertainty. The importance of probability is clear from the following points:

(1) Probability is used in making economic decision in situations of risk and uncertainty by sales managers, production managers, etc.

(2) Probability is used in theory of games which is further used in managerial decisions.

(3) Various sampling tests like Z-test, t-test and F-test are based on the theory of probability.

(4) Probability is the backbone of insurance companies because life tables are based on the theory of probability.

Thus, probability is of immense utility in various fields.

Probability Scale

The probability of an event always lies between 0 and 1, *i.e.*, $0 \le p \le 1$. If the event cannot take place, *i.e.*, impossible event, then its probability will be zero, *i.e.*, P(E) = 0 and if the event is sure to occur, then its probability will be one, *i.e.*, P(E) = 1.

Calculation of Probability of an Event

The following steps are to be followed while calculating the probability of an event:

(1) Find the total number of equally likely cases, *i.e.*, n

(2) Obtain the number of favourable cases to the event,. *i.e.*, m

(3) Divide the number of favourable cases to the event (m) by the total number of equally likely cases (n). This will give the probability of an event.

Symbolically,

Probability of occurrence of an event E is:

$$P(E) = \frac{\text{Number of favourable cases to E}}{\text{Total number of equally likely cases}} = \frac{m}{n}$$

Similarly, Probability of non-occurrence of event E is:

$$P(\bar{E}) = 1 - P(E)$$

The following examples will illustrate the procedure:

Example 1: Find the probability of getting a head in a tossing of a coin.

Solution: When a coin is tossed, there are two possible outcomes - Head or Tail.

Total number of equally likely cases = n = 2

Number of cases favourable to H = m = 1

Notes

$\therefore$ $P(H) = \frac{m}{n} = \frac{1}{2}$

Example 2: What is the probability of getting an even number in a throw of an unbiased die ?

Solution: When a die is tossed, there are 6 equally likely cases, *i.e.*, 1, 2, 3, 4, 5, 6.

Total number of equally likely cases = $n = 6$

Number of cases favourable to even points (2, 4, 6) = $m = 3$

$\therefore$ Probability of getting an even number = $\frac{3}{6} = \frac{1}{2}$

Example 3: What is the probability of getting a king in a draw from a pack of cards ?

Solution: Number of exhaustive cases = $n = 52$

There are 4 king cards in an ordinary pack.

$\therefore$ Number of favourable cases = $m = 4$

$\therefore$ Probability of getting a king = $\frac{4}{52} = \frac{1}{13}$

Example 4: From a bag containing 5 red and 4 black balls. A ball is drawn at random. What is the probability that it is a red ball ?

Solution: Total No. of balls in the bag = 5 + 4 = 9

No. of red balls in the bag = 5

$\therefore$ Probability of getting a red ball = $\frac{5}{9}$

Example 5: A bag contains 5 black and 10 white balls. What is the probability of drawing (i) a black ball, (ii) a white ball ?

Solution: Total number of balls = 5 + 10 = 15

(i) P (black ball) = $\frac{\text{No. of black balls}}{\text{Total No. of balls}} = \frac{5}{15} = \frac{1}{3}$

(ii) P (white ball) = $\frac{\text{No. of white balls}}{\text{Total No. of balls}} = \frac{10}{15} = \frac{2}{3}$

Example 6: In a lottery, there are 10 prizes and 90 blanks. If a person holds one ticket, what are the chances of

(i) getting a prize

(ii) not getting a prize

Solution: Total No. of tickets = 10 + 90 = 100

(i) Probability of getting a prize:

No. of prizes = 10

$\therefore$ No. of favourable cases = 10

Total No. of cases = 100

Required Probability = $\frac{10}{100} = \frac{1}{10} = 0.1$

(ii) The probability of not getting a prize:

No. of Blanks = 90

$\therefore$ Number of favourable cases = 90

Notes

Total Number of cases = 100

Required Probability = $\frac{90}{100}$ = 0.9

Example 7: What is the probability of getting a number greater than 4 with an ordinary die ?

Solution: Number greater than 4 in a die are 5 and 6.

$\therefore$ Number of favourable cases = 2

Total number of cases = 6

Required Probability = $\frac{2}{6} = \frac{1}{3}$

Example 8: Find the probability of drawing a face card in a single random draw from a well shuffled pack of 52 cards.

Solution: There are 52 cards in a pack of cards.

Total number of cases = 52

Number of favourable cases (face cards include the Jack, Queen and King in each) = 12

Required Probability = $\frac{12}{52} = \frac{3}{13}$

Example 9: A card is drawn from an ordinary pack of playing cards and a person bets that it is a spade or an ace. What are odds against his winning this bet ?

Solution: Total number of cases = 52

Since there are 13 spades and 3 aces (one ace is also present in spades), Therefore the favourable cases = 13 + 3 = 16

The probability of winning the bet = $\frac{16}{52} = \frac{4}{13}$

The probability of losing the bet = $1 - \frac{4}{13} = \frac{9}{13}$

Hence, odds against winning the bet = $\frac{9}{13} : \frac{4}{13}$ = 9: 4

Example 10: A single letter is selected at random from the word 'PROBABILITY'. What is the probability that it is a vowel ?

Solution: There are 11 letters in the word 'PROBABILITY' out of which 1 is be selected.

$\therefore$ Total No. of words = 11

There are four vowels *viz.* O, A, I, I. Therefore favourable number of cases = 4

Hence, the required probability = $\frac{4}{11}$

Example 11: Find the probability of drawing an ace from a set of 52 cards.

Solution: Number of exhaustive cases (n) = 52

There are 4 ace cards in an ordinary pack.

$\therefore$ Favourable cases (n) = 4

$\therefore$ Probability of getting an ace = $\frac{4}{52} = \frac{1}{13}$

Example 12: What is the probability that a leap year selected at random will contain 53 Sundays ?

Notes

Solution: Total number of days in a leap year = 366

Number of weeks in a year = $\frac{366}{7} = 52\ \frac{2}{7}$

= 52 weeks and 2 days

Following may be the 7 possible combinations of these two extra days:

(i) Monday and Tuesday (ii) Tuesday and Wednesday

(iii) Wednesday and Thursday (iv) Thursday and Friday

(v) Friday and Saturday (vi) Saturday and Sunday

(vii) Sunday and Monday

A selected leap year can have 53 Sundays if these two extra days happen to be a Sunday

Total possible outcomes of 2 days = $n = 7$

Number of cases having Sundays = $m = 2$

$\therefore$ The required probability = $\frac{2}{7}$

Use of Bernoulli's Theorem in Theory of Probability

Bernoulli's theorem is very useful in working out various probability problems. This theorem states that if the probability of happening of an event in one trial or experiment is known, then the probability of its happening exactly, 1, 2, 3, ... r times in n trials can be determined by using the formula:

$$P(r) = {}^{n}C_{r}\, p^{r} \cdot q^{n-r} \qquad r = 1, 2, 3, \ldots n$$

where,

$P(r)$ = Probability of r successes in n trials.

p = Probability of success or happening of an event in one trial.

q = Probability of failure or not happening of the event in one trial.

n = Total number of trials.

The following examples illustrate the applications of this theorem:

Example 13: The chance that a ship safely reaches a port in 1/5. Find the probability that out of 5 ships expected at least one would arrive safely.

Solution: Given, $n = 5$, $p = \frac{1}{5}$, $q = 1 - \frac{1}{5} = \frac{4}{5}$

P (at least one ship arriving safely) = 1 – 1 (none arriving safely)

$$= 1 - \left[{}^{5}C_{0} (p)^{0} \cdot (q)^{5} \right]$$

$$= 1 - \left[{}^{5}C_{0} \left(\frac{1}{5}\right)^{0} \left(\frac{4}{5}\right)^{5} \right] = 1 - \left(\frac{4}{5}\right)^{5}$$

$$= 1 - \left(\frac{4}{5} \times \frac{4}{5} \times \frac{4}{5} \times \frac{4}{5} \times \frac{4}{5}\right) = 1 - \frac{1024}{3125} = \frac{2101}{3125}$$

Notes

Example 14: Find the probability of throwing 6 at least once in six throws with a single die.

Solution: p = probability of throwing 6 with a single die = $\frac{1}{6}$

$$q = 1-\frac{1}{6} = \frac{5}{6}$$

$$n = 6, p = \frac{1}{6}, q = \frac{5}{6}$$

p (at least one six)= 1 – P [none six in 6 throws]

$$= 1-\left[{}^{6}C_{0}\left(\frac{1}{6}\right)^{0}\cdot\left(\frac{5}{6}\right)^{6}\right] = 1-\left(\frac{5}{6}\right)^{6}$$

Example 15: Three dice are thrown. What is the probability that at least one of the numbers turning up being greater than 4 ?

Solution: p = probability of a number greater than 4 (*i.e.,* 5 and 6) in a throw of one die

$$= \frac{1}{6}+\frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

$$q = 1-\frac{1}{3} = \frac{2}{3}$$

$$\therefore \quad n = 3, p = \frac{1}{3}, q = \frac{2}{3}$$

P (at least one number greater than 4) = 1 – P (none of the number greater than 4)

$$= 1-\left[{}^{3}C_{0}\left(\frac{1}{3}\right)^{0}\cdot\left(\frac{2}{3}\right)^{3}\right]$$

$$= 1-\left(\frac{2}{3}\right)^{3} = 1-\frac{2}{3}\times\frac{2}{3}\times\frac{2}{3} = \frac{19}{27}$$

Example 16: A and B play for a prize of Rs. 1000. A is to throw a die first and is to win if he throws 6. If he fails B is to throw and is to win if throws 6 or 5. If he fails A is to throw again and to win if he throws 6, 5 or 4 and so on. Find their respective expectations.

Solution: Probability of A's winning in the 1st throw (*i.e.,* he throws 6) = $\frac{1}{6}$

Probability of B's winning in the 2nd throw (*i.e.,* he throws 6 or 5) = $\frac{5}{6}\times\frac{2}{6} = \frac{5}{18}$

Probability of A's winning in the 3rd throw (6 or 5 or 4) = $\frac{5}{6}\times\frac{4}{6}\times\frac{3}{6} = \frac{5}{18}$

Probability of B's winning in the 4th throw (6 or 5 or 4 or 3) = $\frac{5}{6}\times\frac{4}{6}\times\frac{3}{6}\times\frac{4}{6} = \frac{5}{27}$

Probability of A's winning in the 5th throw (6 or 5 or 4 or 3 or 2)

Notes

$$= \frac{5}{6}\times\frac{4}{6}\times\frac{3}{6}\times\frac{2}{6}\times\frac{5}{6} = \frac{25}{324}$$

Probability of B's winning in the 6th throw (6 or 5 or 4 or 3 or 2 or 1)

$$= \frac{5}{6}\times\frac{4}{6}\times\frac{3}{6}\times\frac{2}{6}\times\frac{1}{6}\times\frac{6}{6} = \frac{5}{324}$$

A's total chances of success = $\frac{1}{6}+\frac{5}{18}+\frac{25}{324} = \frac{169}{324}$

B's total chances of success = $\frac{5}{18}+\frac{5}{27}+\frac{5}{324} = \frac{155}{324}$

For a prize of Rs. 1,000

A's expectation = $p \times m = \frac{169}{324}\times 1{,}000$ = Rs. 521.6

B's expectation = $p \times m = \frac{155}{324}\times 1{,}000$ = Rs. 478.4

Example 17: A and B play for a prize of Rs. 99. The prize is to be won by a player who first throws 6 with one die. A first throws and if he fails B throws and if he fails A again throws and so on. Find their respective expectations.

Solution: The probability of throwing 6 with a single die = $\frac{1}{6}$

The probability of not throwing 6 with single die = $1-\frac{1}{6} = \frac{5}{6}$

If A is to win, he should throw 6 in the 1st, 3rd, or 5th...throws

If B is to win, he should throw 6 in the 2nd, 4th, 6th...throws

A's chance of success in given by

$$= \frac{1}{6}+\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right)+...\infty$$

$$= \frac{1}{6}\cdot\left[1+\left(\frac{5}{6}\right)^2+\left(\frac{5}{6}\right)^4+...\infty\right] \qquad [\text{Infinite GP series: S} = 1+a+a^2+...\infty\ ]$$

$$= \frac{1}{6}\cdot\left[\frac{1}{1-\left(\frac{5}{6}\right)^2}\right] = \frac{1}{6}\times\frac{36}{11} = \frac{6}{11} \qquad \left(\therefore S_\infty = \frac{1}{1-a} = \frac{\text{First Term}}{1-\text{Common Ratio}}\right)$$

A's expectation = $p \times m = \frac{6}{11}\times 99$ = Rs. 54

B's chance of success is given by

$$= \left(\frac{5}{6}\right)\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)^3\left(\frac{1}{6}\right)+\left(\frac{5}{6}\right)^5\cdot\left(\frac{1}{6}\right)+...\infty$$

Notes

$$= \frac{5}{6} \times \frac{1}{6} \cdot \left[1 + \left(\frac{5}{6}\right)^2 + \left(\frac{5}{6}\right)^4 + \ldots \infty \right]$$

$$= \frac{5}{6} \times \frac{1}{6} \cdot \left[\frac{1}{1 - \left(\frac{5}{6}\right)^2} \right] = \frac{5}{6} \times \frac{1}{6} \times \frac{36}{11} = \frac{5}{11}$$

B's expectation = Rs. $99 \times \frac{5}{11}$ = Rs. 45.

Example 18: A bag contains 6 black and 9 white balls. A person draws out 2 balls. If on every black ball he gets Rs. 20 and on every white ball Rs. 10, find out his expectation.

Solution: There may be the following three options for drawing 2 balls:

(i) Both are white, (ii) Both are black, (iii) One is white and other is black.

(i) Both balls are white

$$P(2W) = p = \frac{{}^{9}C_2}{{}^{12}C_2} = \frac{12}{35}$$

$$\text{Expectation} = p \times m = \frac{12}{35} \times 10 \times 2 = \text{Rs. } 6.86$$

(ii) Both balls are black

$$P(2B) = p = \frac{{}^{6}C_2}{{}^{15}C_2} = \frac{1}{7}$$

$$\text{Expectation} = p \times m = \frac{1}{7} \times 20 \times 2 = \text{Rs. } 5.71$$

(iii) One ball is white and the other is black

$$P(1W\ 1B) = p = \frac{{}^{6}C_1 \times {}^{9}C_1}{{}^{15}C_2} = \frac{18}{35}$$

$$\text{Expectation} = p \times m = \frac{18}{35} \times (20 + 10) = \text{Rs. } 15.43$$

Total Expectation = 6.86 + 5.71 + 15.43 = Rs. 28

Example 19: If it rains, a taxi driver can earn Rs. 1000 per day. If it is fair, he can lose Rs. 100 per day. If the probability of rain in 0.4, what is his expectation ?

Solution: The distribution of earnings (X) is given as:

X	$X_1 = 1000$	$X_2 = -100$
P	$P_1 = 0.4$	$P_2 = 1 - 0.4 = 0.6$

$\therefore$ $E(X) = P_1 X_1 + P_2 X_2$

$= 0.4 \times 1000 + 0.6 \times (-100) = \text{Rs. } 340$

Example 20: A petrol pump dealer sells an average petrol of Rs. 80,000 on a rainy day and an average of Rs. 95,000 at a clear day. The probability of clear weather is 76% on Tuesday. What will be the expected sale ?

Solution: The distribution of earnings (X) is given as:

Notes

X	X_1 = 80,000	X_2 = 95,000
P	1 – 0.76 = 0.24	0.76

E (X) = 80,000 × 0.24 + 95,000 × 0.76

= Rs. 91,400

Example 21: A player tosses 3 fair coins. He wins Rs. 12 if 3 heads appear, Rs. 8 if 2 heads appear and Rs. 3 if 1 head appears. On the otherhand, he loses Rs. 25 if 3 tails appear. Find the expected gain of the player.

Solution: If *p* denotes the probability of getting a head and X denotes the corresponding amount of winning, then the distribution of X is given by:

Heads:	0H	1H	2H	3H
Favourable Events	TTT	HTT, THT, TTH	HHT, HTH, THH	HHH
P	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}$	$\frac{1}{8}+\frac{1}{8}+\frac{1}{8}=\frac{3}{8}$	$\frac{1}{8}+\frac{1}{8}+\frac{1}{8}=\frac{3}{8}$	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}$
X Winning amount	– 25	3	8	12

The expected gain of the player is given by:

$$E(X) = \frac{1}{8}(-25)+\frac{3}{8}(3)+\frac{3}{8}(8)+\frac{1}{8}(12)$$

$$= \frac{-25+9+24+12}{8} = \frac{20}{8} = \frac{5}{2} = \text{Rs. } 2.50.$$

Example 22: A player tosses two fair coins. He wins Rs. 5 if 2 heads appear, Rs. 2 if one head appear and Rs. 1 if no head appear. Find his expected gain of the player.

Solution: If *p* denotes the probability of getting a head and X denotes the corresponding amount of winning, then the probability distribution of X is given by:

Heads:	0H	1H	2H
Favourable Events	TT	HT, TH	HH
P	$\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$	$\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$	$\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$
X	1	2	5

The expected gain of the player is given by:

$$E(X) = P_1X_1 + P_2X_2 + P_3X_3$$

$$= \frac{1}{4}\times 1+\frac{1}{2}\times 2+\frac{1}{4}\times 5 = \text{Rs. } 2.50$$

Notes

Example 23: A survey conducted over the last 25 years indicated that in 10 years, the winter was mild, in 8 years it was cold and in the remaining 7 it was very cold. A company sells 1000 woolen coats in a mild year, 1300 in a cold year and 2000 in a very cold year. If a woolen coat costs Rs. 173 and is sold for Rs. 248, find the yearly expected profit of the company.

Solution:

State of Nature	Prob. P (X)	Sale of woollen coat	Profit (X)
Mild winter	$\frac{10}{25} = 0.4$	1000	1000 × (248 – 173)
Cold winter	$\frac{8}{25} = 0.32$	1300	1300 × (248 – 173)
Very cold winter	$\frac{7}{25} = 0.28$	2000	2000 × (248 – 173)

$\therefore$ Expected profit is given by

$$E(X) = 1000 \times 75 \times 0.4 + 1300 \times 75 \times 0.32 + 2000 \times 75 \times 0.28$$

$$= 30,000 + 31,200 + 42,000 = \text{Rs. } 1,03,200$$

Self-Assessment

1. Which of the following statements are true or false:

(*i*) The classical approach to probability is the oldest and simplest.

(*ii*) The probability of throwing eight with a single dice is 1/6.

(*iii*) The modern probability theory has been developed automatically in which probability is an undefined concept.

(*iv*) In most field of research, a priori probability is employed.

(*v*) Dependent events are those in which the outcome of one does not affect and is not affected by the other.

26.3 Summary

- "It is likely that the particular teacher may not come for taking his class today"; "there is a chance that the particular student may stand first in the university examination"; "it is possible that the particular company may get the contract which it bid last week"; "most probably I shall be returning within a week"; "it is possible that he may not be able to join his duty". In all the above statements, the terms—possible, probably, likely, chance, etc., convey the same meaning, *i.e.,* the events are not certain to take place. In other words, there is involved an element of uncertainty or chance in all these cases. A numerical measure of uncertainty is provided by the theory of probability. The aim of the probability theory is to provide a measure of uncertainty. The theory of probability owes its origin to the study of games of chance like games of cards, tossing coins, dice, etc. But in modern times, it has great importance is decision making problems.
- In simple words, **probability** is the likelihood or chance that a particular event will or will not occur. The theory of probability provides a quantitative measure of uncertainty or likelihood of occurrence of different events resulting from a random experiment, in terms of quantitative measures ranging from 0 to 1. This means that the probability of a certain event is 1 and the probability of an impossible event is 0. In other words, a probability near 0 indicates that an event is unlikely to occur whereas a probability near 1 indicates that an event is almost certain

Notes

to occur. For example, suppose an event is the success of a new product launched. A probability 0.90 indicates that the new product is likely to be successful whereas a probability of 0.15 indicates that the product is unlikely to be successful in the market. A probability of 0.50 indicates that the product is just as likely to be successful as not.

- **According to Laplace, "Probability is the ratio of the favourable cases to the total number of equally likely cases".** From this definition, it is clear that in order to calculate the probability of an event, we have to find the number of favourable cases and it is to be divided by the total number of cases.
- The theory of probability has its origin in the games of chance related to gambling such as tossing a die, tossing a coin, drawing a card from a deck of 52 cards and drawing a ball of a particular colour from a bag.

26.4 Keywords

1. Dependent variables : The variable being measured. The data or score.
2. Depth : Cumulative frequency counting in from the nearer end.
3. Design matrix : A matrix of coded or dummy variables representing group membership.

26.5 Review Questions

1. Define Probability. Discuss the importance of probability in decision-making.
2. Give the classical definition of probability and state its limitations.
3. State and prove the theorem of total probability for mutually exclusive events.
4. Describe the uses of theory of probability.
5. What are the use of Bernoulli's theorem in theory of probability ?

Answers: Self-Assessment

1. (*i*) T (*ii*) F (*iii*) T
   (*iv*) T (*v*) F

26.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 27: Additive and Multiplicative Law of Probability

CONTENTS

Objectives

After reading this unit students will be able to:

- Discuss Additive Rule of Probability.
- Explain Multiplicative Rule of Probability: Conditional Probability.

Introduction

Often it is easier to compute the probability of an event from known probabilities of other events. This can be well observed if the given event can be represented as the **union of two other events** or as the complement of an event. Two such rules used for simplifying the computation of probabilities of events are:

1. Addition Rule of Probability
2. Multiplication Rule of Probability

27.1 Addition Rule of Probability

For any two events A and B

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{... (1)}$$

or

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \quad \text{... (2)}$$

In case A and B are mutually exclusive events, then $P(A \cap B) = 0$ and the addition rule of probability in (1) becomes

$$P(A \cup B) = P(A) + P(B) \quad \text{... (3)}$$

Proof: For any two events A and B, we can write

$$A = (A \cap B)\left(A \cap \overline{B}\right)$$

$$\therefore \quad P(A) = P(A \cap B) + P\left(A \cap \overline{B}\right) \quad \text{... (a)}$$

Using axiom (iii) of probability as the events $(A \cap B)$ and $\left(A \cap \overline{B}\right)$ are mutually exclusive.

Similarly

Notes

$$B = (A \cap B) \cup (\overline{A} \cap B)$$

$$\therefore \quad P(B) = P(A \cap B) + P(\overline{A} \cap B) \quad \ldots (b)$$

The events $(A \cap B)$ and $(\overline{A} \cap B)$ being mutually exclusive. Thus, from (a) and (b), one gets

$$P(A) + P(B) = P(A \cap B) + P(A \cap \overline{B}) + P(A \cap B) + P(\overline{A} \cap B) \quad \ldots (c)$$

Now the last three terms on R.H.S. of (c), *i.e.*, $P(A \cap \overline{B}) + P(A \cap B) + P(\overline{A} \cap B)$ represent the probability of occurrence of the events A or B or both A and B, *i.e.*, $P(A \cup B)$. Thus, replacing these three terms by $P(A \cup B)$, equation (c) can be written as

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

or

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

The rule in (d) is called the **addition of rule probability.**

If A and B are mutually exclusive, $P(A \cap B) = 0$ and the addition rule of probability becomes

$$P(A \cup B) = P(A) + P(B) \quad \ldots (e)$$

Example 1: What is the probability of getting an odd number in tossing a die ?

Solution: There are three odd numbers on a die, *i.e.*, 1, 3 and 5. Let A, B and C be the respective events of getting 1, 3 and 5. Thus, P (A) = 1/6, P (B) = 1/6 and P (C) = 1/6. Since A, B and C are mutually exclusive therefore

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

Example 2: An urn contains 4 white and 2 red balls. Two balls are drawn randomly with replacement. Find the probability that

(i) both balls are white

(ii) both balls will be of the same colour.

Solution: Here total number of balls = 6

Two balls can be drawn out of 6 in ${}^6C_2$ ways *i.e.*, 15 ways. Let A be the event that both balls are white. The number (i) of ways of selecting 2 balls out of 4 is ${}^4C_2$ *i.e.*, 6 ways.

$$\therefore \quad P(A = \text{both balls white}) = \frac{6}{15}$$

(ii) The number of ways of selecting 2 balls out of 4 white balls is ${}^4C_2 = 6$ ways

The number of ways of selecting 2 both out of 2 red balls is ${}^2C_2 = 1$ way

$$\therefore \quad P(\text{both balls will be of the same colour}) = \frac{{}^4C_2}{{}^6C_2} + \frac{{}^2C_2}{{}^6C_2}$$

Notes

$$= \frac{6}{15} + \frac{1}{15}$$

$$= \frac{7}{15}$$

Example 3: Find the probability of getting more than 4 in tossing a die.

Solution: The numbers more than 4 on a die are 5 and 6.

Let A and B be the respective events of getting 5 and 6.

Thus, $P(A) = \frac{1}{6}$, $P(B) = \frac{1}{6}$

Also A and B are mutually exclusive.

Thus, $P(A \text{ or } B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$

Example 4: A card is drawn at random from a well-shuffled pack of 52 cards. Find the probability of getting an ace or a spade.

Solution: Let A be the event of getting an ace and B of getting a spade. Then A = set of all aces, B = set of all spades and $A \cap B$ = set of an ace of spade.

Clearly, $n(A) = 4$, $n(B) = 13$ and $n(A \cap B) = 1$.

Also, $n(S) = 52$. Therefore,

$$P(A) = \frac{n(A)}{n(S)} = \frac{4}{52}, \quad P(B) = \frac{n(B)}{n(S)} = \frac{13}{52} \text{ and } P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{52}$$

Thus, the required probability

$$P(\text{an ace or a spade}) = P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

Example 5: A construction company is bidding for two contracts, A and B. The probability that the company will get contract A is 3/5, will get contract B is 1/4 and the probability that the company gets both the contracts is 1/8. What is the probability that the company will get contract A or B ?

Solution: Let A and B be the respective events of getting the contracts A and B. Then, we are given that

$$P(A) = 3/5, (B) = 1/4 \text{ and } P(A \cap B) = 1/8$$

Thus, the required probability that the company will get a contract A or B is

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

$$= \frac{3}{5} + \frac{1}{4} - \frac{1}{8} = \frac{29}{40}$$

Example 6: A bag contains 30 balls numbered from 1 to 30. One ball is drawn at random. Find the probability that number of the drawn ball is a multiple of (i) 4 or 9 (ii) 5 or 6.

Solution: (i) Let A be the event that the drawn number is a multiple of 4, then A = {4, 8, 12, 16, 20, 24, 28}. Further let B be the event that the drawn number is a multiple of 9, *i.e.*, B = {9, 18, 27}.

Notes

Also $A \cap B = \phi$, null set and $n(S) = 30$.

$\therefore$ $P(A) = \frac{n(A)}{n(S)} = \frac{7}{30}$, $P(B) = \frac{n(B)}{n(S)} = \frac{3}{30}$, $P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{0}{30} = 0$

Thus, the probability of the desired event

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

$$= \frac{7}{30} + \frac{3}{30} - \frac{0}{30} = \frac{10}{30} = \frac{1}{3}$$

(ii) Let A be the event that the drawn number is a multiple of 5 and B that the number is a multiple of 6.

$\therefore$ $A = \{5, 10, 15, 20, 25, 30\}$, $B = \{6, 12, 18, 24, 30\}$.

and $A \cap B = \{30\}$. Thus, $n(A) = 6$, $n(B) = 5$, $n(A \cap B) = 1$, $n(S) = 30$

$\therefore$ $P(A) = \frac{n(A)}{n(S)} = \frac{6}{30}$, $P(B) = \frac{n(B)}{n(S)} = \frac{5}{30}$,

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{1}{30}$$

$\therefore$ P (multiple of 5 or 6) = P (A or B) = $P(A \cup B)$

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

$$= \frac{6}{30} + \frac{5}{30} - \frac{1}{3} = \frac{10}{30} = \frac{1}{3}.$$

Example 7: Two boxes contain respectively 6 brown, 8 blue, 1 black balls and 3 brown, 7 blue and 5 black balls. One ball is drawn from each box. What is the probability that both the balls drawn are of the same colour.

Solution: In all there are 15 balls in one box and 15 balls in another box.

One brown ball from each box may be drawn in $\frac{{}^6C_1}{{}^{15}C_1} \times \frac{{}^3C_1}{{}^{15}C_1}$ ways

One blue ball from each box may be drawn in $\frac{{}^8C_1}{{}^{15}C_1} \times \frac{{}^7C_1}{{}^{15}C_1}$ ways

One black ball from each box may be drawn in $\frac{{}^1C_1}{{}^{15}C_1} \times \frac{{}^5C_1}{{}^{15}C_1}$ ways

All the three cases are mutually exclusive and thus the required probability

$$= \frac{{}^6C_1}{{}^{15}C_1} \times \frac{{}^3C_1}{{}^{15}C_1} + \frac{{}^8C_1}{{}^{15}C_1} \times \frac{{}^7C_1}{{}^{15}C_1} + \frac{{}^1C_1}{{}^{15}C_1} \times \frac{{}^5C_1}{{}^{15}C_1}$$

$$= \frac{6 \times 3}{225} + \frac{8 \times 7}{225} + \frac{1 \times 5}{225}$$

$$= \frac{18 + 56 + 5}{225} = \frac{79}{225}$$

Notes

27.2 Multiplicative Rule of Probability: Conditional Probability

Suppose that two dice are thrown. Then there are 36 sample points and the event A that the first die shows a five consists of 6 sample points (5, 1), (5, 2), (5, 3), (5, 4), (5, 5) and (5, 6). Thus P (A) = $\frac{6}{36}$ = $\frac{1}{6}$. The event B that the sum (total) of numbers in the two dice is 9 consists of the sample points (3, 6), (4, 5), (5, 4), (6, 3) and P (B) = $\frac{1}{9}$. Now suppose that we are given the information that the first die shows a five, then the event that the sum of the numbers shown on the faces of the two dice is nine is a *conditional event,* and this conditional event is denoted by (B | A). The probability of the conditional event is called *conditional probability.* For a conditional event, instead of the whole sample space we have only the sample points comprising of the enent A, *i.e.* the 6 sample points (5, 1), (5, 2), (5, 3), (5, 4), (5, 5) and (5, 6), and the conditional probability of each of these is $\frac{1}{6}$. Conditioned by the event A, *i.e.* that the first die shows a five, the (conditional) event that the sum is nine comprises of only one sample point (5, 4) *i.e.* the sample point common to both A and B. Here the conditional probability of getting a sum of nine given that the first die shows a five is $\frac{1}{6}$, or in symbols P (B | A) = $\frac{1}{6}$.

General formula: Consider a conditional event (B | A) *i.e.* the event B given that A has actually happend. Then for the happening of the event (B | A), the sample space is restricted to the sample points comprising the event A. The conditional probability P (B | A) is given by

$$P(B \mid A) = \frac{i}{j}, \text{ where}$$

i = number of sample points common to both A and B,

j = number of sample points comprising A,

n = total number of points in the whole (unrestricted) sample space S.

Thus, $P(AB) = \frac{i}{n}$, $P(A) = \frac{j}{n}$ and $P(B \mid A) = \frac{i}{j}$

Dividing both the numerator and the denominator of $P(B \mid A) = \frac{i}{j}$ by n, the total number of sample points in the sample space S, we get

$$P(B \mid A) = \frac{i}{j} = \frac{i}{n} \Big/ \frac{j}{n}$$

$$= \frac{P(AB)}{P(A)} \quad \ldots (4)$$

It follows that

$$P(AB) = P(B \mid A)\, P(A). \quad \ldots (5)$$

Notes

Note: (1) These two results hold only if P (A) > 0.

(2) The results are proved under the tacit assumption that n is finite and that each sample point has equal probability $\frac{1}{n}$. It can be shown that the results hold in the general case (without these restrictions).

Example 8: Two dice are thrown. Find the probability that the sum of the numbers in the two dice is 10, given that the first die shows *six*.

Let A be the event that the sum of numbers in two dice is 10,

B be the event that the first die shows 6.

Then AB is the event that the sum is 10 and the first die shows 6 or which is equivalent to the event that first die shows 6 and the second 4.

We have $P(B) = \frac{1}{6}$, $P(AB) = \frac{1}{36}$

Thus the required probability = $P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{1}{6}$. Thus the conditional probability of getting a sum of 10, given that the first shows 6 is $\frac{1}{6}$. The unconditional probability of getting a sum of 10 is $\frac{1}{12}$.

Example 9: Two coins are tossed. What is the conditional probability of getting two heads (event B) given that at least one coin shows a head (event A) ?

Event A comprises of 3 sample points (HH), (HT), (TH) so that $P(A) = \frac{3}{4}$; the event AB comprises of only one point (HH), so that $P(AB) = \frac{1}{4}$. Thus the conditional probability is

$$P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

Example 10: A box contains 5 black and 4 white balls. Two balls are drawn one by one *without replacement*, *i.e.* the first ball drawn is not returned to the box. Given that the first ball drawn is black, what is the probability that both the balls drawn will be black ?

Before the first draw the sample space consists of 9 points each with probability $\frac{1}{9}$. After the first draw the number of sample points reduces to 8 (as one ball is already out of the box) and the probability of each sample point is $\frac{1}{8}$.

Let A be the event that the first ball drawn is black then $P(A) = \frac{5}{9}$, since there are 5 black balls.

Notes

Let B be the event that the second ball drawn is black. Then the conditional event (B | A) implies drawing a black ball from the box which contains 4 black and 4 white balls.

Thus $P(B \mid A) = \frac{4}{8} = \frac{1}{2}$.

The event implies that both the balls drawn are black.

We are required to find P (AB). We have

$$P(AB) = P(B \mid A) \cdot P(A)$$

$$= \frac{5}{9} \cdot \frac{1}{2} = \frac{5}{18}.$$

In the above example, consider that the ball drawn in the first draw is returned to the box, so that the composition of the box remains the same before each drawing. Here it is drawing *with replacement*. Now the conditional event (B | A), that the second ball is black is not affected by the event that the first draw resulted in a black ball since the ball drawn was returned. Thus the event B and the event A are *independent* and the knowledge of one does not affect the other. Then P (B | A) = P (B), independent of P (A). We have then

$$P(B) = \frac{P(AB)}{P(A)}$$

Or, $P(AB) = P(A)\,P(B)$ ... (6)

which gives the multiplication rule for independent events, A and B.

Example 11: What is the probability that in 2 throws of a die, six appears in both the dice ?

Let A and B be the event that 6 appears in the first and the second throws respectively; these events are independent. Then the event AB implies that six appears in both the dice.

$$P(AB) = P(A)\,P(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

Example 12: The probability of getting HHT in tossing 3 coins is thus $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$, since the three events Head in first throw, Head in second and Tail in third are independent.

Example 13: Find the probability that in a family of 2 children (i) both are boys, (ii) both are of the same sex, assuming that the probability of a child being a boy or a girl is equal $\left(\text{equal to } \frac{1}{2}\right)$

(i) Let A_1 , A_2 be the events respectively that the first and the second child is a boy. Then since A_1, A_2 are independent,

$$P(A_1A_2) = P(A_1)\,P(A_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Similarly if B_1 , B_2 are the events respectively that the first and the second child is a girl, then B_1, B_2 are independent and

$$P(B_1B_2) = P(B_1)\,P(B_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Notes

(ii) Now the event that both children are of the same sex is equivalent to the event that both are either boys or girls and these are mutually exclusive events. Thus

the required probability = $\frac{1}{4}+\frac{1}{4} = \frac{1}{2}$.

Mutually exclusive events and independent events

These ideas are not equivalent ideas. We discuss them to bring out the difference between the two. When the happening of one event precludes the happening of the other event, the two events are mutually exclusive (or (disjoint). For two mutually exclusive events A and B

$$P(AB) = 0.$$

When the happening of one event has no effect on the probability of occurrence of happening of the other event, the two events are independent. For two independent events A and B,

$$P(AB) = P(A)\,P(B).$$

Two events can be mutually exclusive and not independent. Again two events can be independent and not mutually exclusive.

Suppose two coins are tossed. The events {H, H} $\equiv$ A (head on both coins) and the event {T, T} $\equiv$ B are mutually exclusive (because if A happens B cannot happen) and

$$P(AB) = 0$$

But $P(A) = \frac{1}{4}$, $P(B) = \frac{1}{4}$ and so $P(AB) = 0 \neq P(A)\,P(B)$ and hence A and B are not independent.

Consider the events A, B that 6 appears in the first and the second die respectively in throwing 2 dice together (example 11). The events A and B are independent, as

$$P(AB) = \frac{1}{36} = P(A)\,P(B) = \frac{1}{6}\times\frac{1}{6}.$$

Here $P(AB) \neq 0$ and hence the events are not mutually exclusive.

The only way that two mutually exclusive events A, B can be independent is when both

$$P(AB) = 0 \text{ and } P(AB) = P(A)\,P(B)$$

hold simultaneously. Both of the above can hold simultaneously if at least one of P (A) or P (B) is zero. In case at least one of P (A) or P (B) is zero then the two events A and B mutually exclusive events as well as independent events.

Number of sample points in a combination of events or sets

Let N (A) denote the number of points in the set A.

Then, using Venn diagrams, it can be easily seen that

$$N(A \cup B) = N(A) + N(B) - N(AB). \quad \ldots (7)$$

Example 14: Suppose that students in an Institution can enrol for one, two or none of the language courses, French (A), German (B). If 30% are enrolled for French, 20% for German and 10% for both French and German, then the number (in percentage) enrolled for at least one of the courses is given by

$$N(A \cup B) = N(A) + N(B) - N(AB)$$
$$= 30 + 20 - 10$$
$$= 40$$

and the percentage not enrolled for any of the courses is

$$100 - 40 = 60$$

Thus the probability that a student selected at random is (i) enrolled for at least one of the courses is 0.4 and (ii) not enrolled for any of the courses is 1 – 0.4 = 0.6. Given that

Notes

a student is enrolled for at least one of the courses, the (conditional) probability that he is enrolled for French is

$$\frac{30}{40} = 0.75$$

Discrete Sample Space

So far we considered cases where the sample space contains a finite number of points. We consider the following example where this is not the case.

Example 15: A coin is tossed until a head appears. Describe the sample space. Find the probability that the coin will be tossed (a) exactly 4 times (b) at the most, 4 times. (c) What is the probability that head will appear if the coin is tossed an infinite number of times ?

The head may appear at the

(i) very first throw (H)

(ii) second throw, the first toss resulting in a tail (TH)

(iii) third throw, the first two tosses resulting in tails (TTH)

(iv) fourth throw, the first three tosses resulting in tails (TTTH)

and so on: an infinite number of throws may be needed to get a head. The sample space consists of an *infinite* number of the *sample* points

H, TH, TTH, TTTH, TTTTH, ...

The trials are independent. Assume that the coin is fair (unbiased).

The probability of the event H = $\frac{1}{2}$

The probability of the event TH = $\frac{1}{2}\cdot\frac{1}{2} = \left(\frac{1}{2}\right)^2$

The probability of the event TTH = $\left(\frac{1}{2}\right)^2\cdot\frac{1}{2} = \left(\frac{1}{2}\right)^3$

The probability of the event TTTH = $\left(\frac{1}{2}\right)^3\cdot\frac{1}{2} = \left(\frac{1}{2}\right)^4$

and so on.

(a) The probability that to get a head, the coin will be tossed exactly 4 times is $\left(\frac{1}{2}\right)^4$.

(b) The event that the coin will be tossed at most 4 times is a compositive event comprising of the 4 sample events H, TH, TTH and TTTH. Thus the required probability

$$= \frac{1}{2}+\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^3+\left(\frac{1}{2}\right)^4 = \frac{15}{16}.$$

(c) Suppose that the coin is tossed as many times as is necessary to get a head. The required probability is

$$\frac{1}{2}+\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^3+ \ ...$$

Notes

$$= \frac{\frac{1}{2}}{1-\frac{1}{2}} = 1$$

Thus it is certain that a head will ultimately appear if the coin is tossed indefinitely. The result holds even if the coin is biased.

Incidentally, it is verified that P (S) = 1.

Note: In this example we find that though the number of points are infinite, these can be arranged according to the sequence of natural numbers (such that there is one-one correspondence between the natural numbers and the sample points); such an infinity of numbers is called *denumerable infinity* (or *countable infinity*) of numbers.

Discrete Sample space

A sample space that consists of a finite number of sample points or a denumerably infinite number of sample points is called a *discrete sample space.*

Self-Assessment

1. Tick the Correct Answer:

(*i*) Addition theorem states that if two events A and B are mutually exclusive, the probability of occurrence of either A or B is given by:

(*a*) P (A) + P (B) (*b*) P (A) × P (B)

(*c*) P (A) – P (B) (*d*) P (A) × (B) – P (AB)

(*ii*) If two events A and B are independent, the probability that they will both occur is given by:

(*a*) P (A) + P (B) (*b*) P (A) × P (B)

(*c*) P (A) – P (B) (*d*) P (A) × (B) + P (AB)

(*iii*) If two events A and B are dependent, the conditional probability of B given A, *i.e.*, P (B | A) is calculated as:

(*a*) P (AB) | P (B) (*b*) P (A) | P (B)

(*c*) P (AB) | P (A) (*d*) P (A) | P (AB)

(*iv*) If two events A and B are dependent, the conditional probability of A given B, *i.e.*, P (A | B) is calculated as:

(*a*) P (B | A) (AB) (*b*) P (B) | P (A)

(*c*) P (AB) | P (A) (*d*) P (AB) | P (B)

(*v*) ${}^5C_2$ is equal to

(*a*) 20 (*b*) 10

(*c*) 30 (*d*) 100

27.3 Summary

- It is easier to compute the probability of an event from known probabilities of other events. This can be well observed if the given event can be represented as the **union of two other events** or as the complement of an event.
- The probability of the conditional event is called *conditional probability.* For a conditional event, instead of the whole sample space we have only the sample points comprising of the enent A, *i.e.* the 6 sample points (5, 1), (5, 2), (5, 3), (5, 4), (5, 5) and (5, 6), and the conditional probability

Notes

of each of these is $\frac{1}{6}$. Conditioned by the event A, *i.e.* that the first die shows a five, the (conditional) event that the sum is nine comprises of only one sample point (5, 4) *i.e*. the sample point common to both A and B. Here the conditional probability of getting a sum of nine given that the first die shows a five is $\frac{1}{6}$, or in symbols P (B | A) = $\frac{1}{6}$.

- Consider a conditional event (B | A) *i.e*. the event B given that A has actually happend. Then for the happening of the event (B | A), the sample space is restricted to the sample points comprising the event A. The conditional probability P (B | A) is given by

$$P(B \mid A) = \frac{i}{j}.$$

- These ideas are not equivalent ideas. We discuss them to bring out the difference between the two. When the happening of one event precludes the happening of the other event, the two events are mutually exclusive (or (disjoint). For two mutually exclusive events A and B P (AB) = 0.
- When the happening of one event has no effect on the probability of occurrence of happening of the other event, the two events are independent. For two independent events A and B,

  P (AB) = P (A) P (B).
- Two events can be mutually exclusive and not independent. Again two events can be independent and not mutually exclusive.

  Suppose two coins are tossed. The events {H, H} $\equiv$ A (head on both coins) and the event {T, T} $\equiv$ B are mutually exclusive (because if A happens B cannot happen) and P (AB) = 0. But P (A) = $\frac{1}{4}$, P (B) = $\frac{1}{4}$ and so P (AB) = 0 $\neq$ P (A) P (B) and hence A and B are not independent.
-
- A sample space that consists of a finite number of sample points or a denumerably infinite number of sample points is called a *discrete sample space*.

27.4 Key-Words

1. Effective sample size : The sample size needed in equal-sized groups to achieve the power when we have groups of unequal sizes. It will generally be less than the total number of subjects in the unequal groups.
2. Efficiency : The degree to which repeated values for a statistic cluster around the parameter.

27.5 Review Questions

1. State and prove the multiplicative theorem of probability. How is the result modified when the events are independents ?
2. State and prove the addition rule of probability.
3. Differentiate between the circumstances when the probabilities of two events (i) added, and (ii) multiplied.
4. Discuss the general rule for probability. What is its form if the concerned events are mutually exclusive ?
5. Distinguish between addition and multiplicative rule of probability.

Answers: Self-Assessment **Notes**

1. (*i*) (*a*) (*ii*) (*b*) (*iii*) (*c*) (*iv*) (*d*) (*v*) (*b*)

27.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 28: Theory of Estimation: Point Estimation, Unbiasedness, Consistency, Efficiency and Sufficiency

CONTENTS

Objectives

After reading this unit students will be able to:

- Explain Point Estimation and Unbiasedness.
- Discuss Consistency, Efficiency and Sufficiency.
- Know the Application of Point Estimation.

Introduction

The topic of estimation in Statistics deals with estimation of population parameters like mean of a statistical distribution. It is assumed, that the concerned variable of the population follows a certain distribution with some parameter(s). For instance, it may be assumed that the life of the electric bulbs follows a normal distribution which has two parameters *viz.* mean (m) and standard deviation (σ).

While one of the parameters, say, standard deviation is known to be equal to 200 hours from past experience, the other parameter, *viz.* the mean life of the bulbs, is not known, and which we wish to estimate.

Given a sample of observations $x_1, x_2, x_3, ..., x_n$, one is required to determine with the aid of these observations, an estimate in the form of a specific number like 2500 *hrs*., in the above case. This number can be taken to be the best value of the unknown mean. Such single value estimate is called **'Point'** estimate. The estimation could also be in the form of an interval, say 2,300 to 2,700 *hrs*. This can be taken to include the value of the unknown mean. This called **'Interval Estimation'**. An example of point and interval estimation could be provided from our day-to-day conversation when we talk about commuting time to office. We do make statements like "It takes about 45 minutes ranging from 40 to 50 minutes depending on the traffic conditions." The statistical details of these two types of estimation are described below.

28.1 Point Estimation

A point estimate is a single value, like 10, analogous to a point in a geometrical sense. It is used to estimate a population parameter, like mean, with the help of a sample of observations.

It may be noted that the observations $x_1, x_2, x_3, ... x_n$ are random variables, and therefore, any function of these observations will also be a random variable. Any function of the sample observations is

Notes

called a **Statistic.** For example, the arithmetic mean $\bar{x}$ of the sample $x_1, x_2, x_3, ..., x_n$, is also a random variable, as also a Statistic. This is illustrated by the numerical example given below:

Let the population comprise only 3 values, say 1, 2 and 3. If a sample of size 2 is taken, then there are 3 possible samples *viz.* 1 & 2, 1 & 3, 2 & 3.

It may be noted from the following Table 1 that the sample means are much closer to each other (in the range from 1.5 to 2.5) than the population values (in the range from 1 to 3). This is quantified by the variance calculated in both the cases. While the variance of the population values is 2/3, the variance of sample means is only 1/6.

Table 1: Variance of Sample Means

Population Values	Arithmetic Mean of Population	Variance	Samples of Two values	Arithmetic Mean of the Three Samples	Variance of the Three Sample Means
1	2	2/3	1, 2	1.5	1/6
2			1, 3	2.0	
3			2, 3	2.5	

In general, if the variance of the population with finite units is σ^2, the variance of the sample means from the population is {(N – *n*)/(N – 1)} $\left(\sigma^2/n\right)$, where *n* is the size of each sample and N is the population size. In the above case, N = 3, and *n* = 2. Therefore, variance of sample mean = {(3 – 2)/ (3 – 1)} {(2/3)/2} = 1/3 × 1/2 = 1/6.

However, if the population size is large as compared to the sample, then the variance of the sample mean is simply σ^2/n.

Incidentally, the standard deviation of the sample mean is known as the **standard error** of the mean. It is a measure of the extent to which sample means could be expected to vary from sample to sample. No statistic can be guaranteed to provide a close value of the parameter on each and every occasion, and for every sample. Therefore, one has to be content with formulating a rule/method which provides good results in the long run or which has a high probability of success.

Did u know? Incidentally, while the method or rule of estimation is called an estimator like sample mean, the value which the method or rule gives in a particular case is called an estimate.

Between two estimators, the estimator with lesser variance is preferred as a value obtained through any sample is more likely to be near the actual value of the parameter. For example, in Figure 1, the estimator 'A' is preferred as its variation is lesser than 'B'.

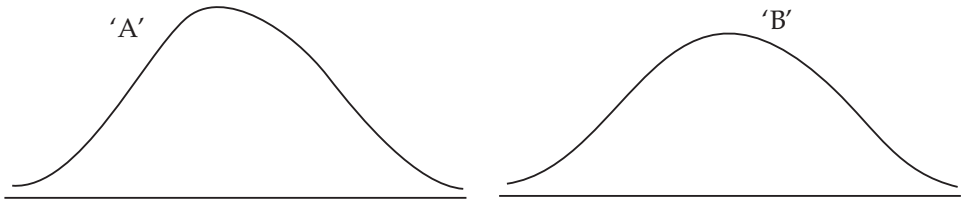


Figure 1: Distributions of Estimators 'A' and 'B'

The real exercise in estimation is to find an estimator. The merit of an estimator is judged by the distribution of estimates to which it gives rise *i.e.* by the properties of its sampling distribution as pointed above.

Notes

28.2 Unbiasedness, Consistency, Efficiency and Sufficiency

There can be more than one estimators of a population parameter. For example, the population mean (μ) may be estimated either by sample mean $(\overline{X})$ or by sample median (M) or by sample mode (Z), etc. Similarly, the population variance (σ^2) may be estimated either by the sample variance (s^2), sample S.D. (s), sample mean deviation, etc. Therefore, it becomes necessary to determine a good estimator out of a number of available estimators. A good estimator is one which is as close to the true value of the parameter as possible. A good estimator possess the following characteristics or properties:

(1) Unbiasedness
(2) Consistency
(3) Efficiency
(4) Sufficiency

Let us consider them in detail:

(1) Unbiased Estimator: An estimator $\hat{\theta}$ is said be unbiased estimator of the population parameter θ if the mean of the sampling distribution of the estimator $\hat{\theta}$ is equal to the corresponding population parameter θ. Symbolically,

$$\mu_{\hat{\theta}} = \theta$$

In terms of mathematical expectation, $\hat{\theta}$ is an unbiased estimator of θ if the expected value of the estimator is equal to the parameter being estimated. Symbolically,

$$E(\hat{\theta}) = \theta$$

Example 1: Sample mean $\overline{X}$ is an unbiased estimate of the population mean μ because the mean of the sampling distribution of the means $\mu_{\overline{X}}$ or $E(\overline{X})$ is equal to the population mean μ. Symbolically,

$$\mu_{\overline{X}} = \mu \text{ or } E(\overline{X}) = \mu$$

Example 2: Sample variance s^2 is a biased estimate of the population variance σ^2 because the mean of the sampling distribution of variance is not equal to the population variance. Symbolically,

$$\mu_s^2 \neq \sigma^2 \text{ or } E(s^2) \neq \sigma^2$$

However, the modified sample variance $(\hat{s}^2)$ is unbiased estimate of the population variance σ^2 because

$$E(\hat{s}^2) = \sigma^2 \text{ where, } \hat{s}^2 = \frac{n}{n-1} \times s^2$$

Example 3: Sample proportion p is an unbiased estimate of the population proportion P because the mean of the sampling distribution of proportion is equal to the population proportion. Symbolically,

Notes

μ_P = P or E (P) = P

(2) Consistent Estimator: An estimator is said to be consistent if the estimator approaches the population parameter as the sample size increases. In other words, an estimator $\hat{\theta}$ is said to be consistent estimator of the population parameter θ, if the probability that $\hat{\theta}$ approaches θ is 1 an n becomes large and larger. Symbolically,

$$P\left(\hat{\theta} \to \theta\right) \to 1 \text{ as } n \to \infty$$

Note: A consistent estimator need not to be unbiased

A sufficient condition for the consistency of an estimator is that

(i) $E\left(\hat{\theta}\right) \to \theta$

(ii) $\text{Var}\left(\hat{\theta}\right) \to 0$ as $n \to \infty$

Example 4: Sample mean $\overline{X}$ is a consistent estimator of the population mean μ because the expected value of the sample mean approaches the population mean and the variance of the sample mean approaches zero as the size of the sample is sufficiently increased. Symbolically,

(i) $E\left(\overline{X}\right) \to \mu$

(ii) $\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$

Example 5: Sample median is also consistent estimator of the population mean because:

(i) $E(M) \to \mu$

(ii) Var (M) $\to$ 0 as $n \to \infty$

(3) Efficient Estimator: Efficiency is a relative term. Efficiency of an estimator is generally defined by comparing it with another estimator. Let us to take two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$. The estimator $\hat{\theta}_1$ is called an efficient estimator of θ if the variance of $\hat{\theta}_1$ is less than the variance of $\hat{\theta}_2$. Symbolically,

$$\text{Var}\left(\hat{\theta}_1\right) < \text{Var}\left(\hat{\theta}_2\right)$$

Then, $\hat{\theta}_1$ is called an efficient estimator.

Example 6: Sample mean $\overline{X}$ is an unbiased and efficient estimator of the population mean (or true mean) than the sample median M because the variance of the sampling distribution of the means is less than the variance of the sampling distribution of the medians.

The relative efficiency of the two unbiased estimators is given below:

We know that, $\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$, Var (M) = $\frac{\pi}{2} \cdot \frac{\sigma^2}{n}$

Notes

$$\text{Efficiency} = \frac{\text{Var}(\bar{X})}{\text{Var}(M)} = \frac{\frac{\sigma^2}{n}}{\frac{\pi\sigma^2}{2n}} = \frac{2}{\pi} = \frac{14}{22} = \frac{7}{11} = 0.64 \left[\because \pi = \frac{22}{7}\right]$$

$$\therefore \quad \text{Var}(\bar{X}) = 0.64 \text{ Var (M)}$$

Therefore, sample mean $\bar{X}$ is 64% more efficiency than the sample median.

Hence, the sample mean is more efficient estimator of the population mean as compared to sample median.

(4) Sufficient Estimator: The last property that a good estimator should possess is sufficiency. An estimator $\hat{\theta}$ is said to be a 'sufficient estimator' of a parameter θ if it contains all the informations in the sample regarding the parameter. In other words, a sufficient estimator utilises all informations that the given sample can furnish about the population. Sample means $\bar{X}$ is said to be a sufficient estimator of the population mean.

28.3 Application of Point Estimation

The applications relating to point estimation are studied under two headings:

(1) Point Estimation in case of Single Sampling

(2) Point Estimation in case of Repeated Sampling.

(1) Point Estimation in case of Single Sampling: When a single independent random sample is drawn from the unknown population, the point estimate of the population parameter can be illustrated by the following examples:

Example 7: A sample of 10 measurements of the diameter of a sphere gave a mean $\bar{X}$ = 4.38 inches and a standard deviation = .06 inches. Determine the unbiased and efficient estimates of (a) the true mean (*i.e.*, population mean) and (b) the true variance (*i.e.*, population variance).

Solution: We are given: n = 10, $\bar{X}$ = 4.38, s = .06

(a) The unbiased and efficient estimate of the true mean μ is given by:

$$\bar{X} = 4.38$$

(b) The unbiased and efficient estimate of the true variance σ^2 is:

$$\hat{s}^2 = \frac{n}{n-1} \cdot s^2$$

Putting the values, we get

$$\hat{s}^2 = \frac{10}{10-1} \times .06 = 1.11 \times 0.06 = .066$$

Thus, $\mu = 4.38$, $\sigma^2 = 0.666$

Example 8: The following five observations constitute a random sample from an unknown population:

Notes

6.33, 6.37, 6.36, 6.32 and 6.37 centimeters.

Find out unbiased and efficient estimates of (a) true mean, and (b) true variance.

Solution: (a) The unbiased and efficient estimate of the true mean (*i.e.* population mean) is given by the value of

$$\bar{X} = \frac{\Sigma X}{n} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = \frac{31.75}{5} = 6.35$$

(b) The unbiased and efficient estimate of the true variance (*i.e.,* population variance) is:

$$\hat{s}^2 = \frac{\Sigma\left(X - \bar{X}\right)^2}{n-1}$$

where, $\hat{s}^2$ = modified sample variance.

$$= \frac{(6.33 - 6.35)^2 + (6.37 - 6.35)^2 + (6.36 - 6.35)^2}{5-1}$$

$$\underline{+(6.32 - 6.35)^2 + (6.37 - 6.35)^2}$$

$$= \frac{.0022}{4} = .00055\ cm^2$$

Example 9: The following data relate to a random sample of 100 students in Kurukshetra University classified by their weights (*kg*):

Weight (*kg*):	60—62	63—65	66—68	69—71	72—74
No. of Students:	5	18	42	27	8

Determine unbiased and efficient estimates of (a) population mean and (b) population variance.

Solution:

Calculation of Mean and Variance

Weight	**No. of Students (*f*)**	**M.V. (*m*)**	**A = 67 *d* = *m* – A**	***d′* = *d*/3**	***fd′***	***fd′*²**
60—62	5	61	– 6	– 2	– 10	20
63—65	18	64	– 3	– 1	– 18	18
66—68	42	67	0	0	0	0
69—71	27	70	+ 3	+ 1	+ 27	27
72—74	8	73	+ 6	+ 2	+ 16	32
	n = 100				$\Sigma fd'$ = 15	$\Sigma fd'^2$ = 97

(a) The unbiased and efficient estimate of the population mean is given by the value:

$$\bar{X} = A + \frac{\Sigma fd'}{n} \times i$$

Notes

$$= 67 + \frac{15}{100} \times 3 = 64 + (0.45) = 67.45$$

(b) The unbiased and efficient estimate of the population variance is:

$$\hat{s}^2 = \frac{n}{n-1} \cdot s^2$$

where, $$s^2 = \frac{\Sigma fd'^2}{n} - \left(\frac{\Sigma fd'}{n}\right)^2 \times i^2$$

$$= \left[\frac{97}{100} - \left(\frac{15}{100}\right)^2\right] \times 3^2$$

$$= [0.97 - .0225] \times 9 = 8.5275$$

Now, $$\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{100}{99} \times 8.5275 = 8.6136$$

Thus, $\mu = 67.45$, $\sigma^2 = 8.6136$

(2) Point Estimation in Case of Repeated Sampling: When large number of random samples of same size are drawn from the population with or without replacement, then the point estimates of the population parameter can be illustrated by the following examples:

Example 10: A population consists of five values: 3, 4, 5, 6 and 7. List all possible samples of size 3 without replacement from this population and calculate the mean $\bar{X}$ of each sample. Verify that sample mean $\bar{X}$ is an unbiased estimate of the population mean.

Solution: The population consists of the five values: 3, 4, 5, 6, 7. The total number of possible samples of size 3 without replacement are ${}^5C_3 = 10$ which are shown in the following table:

Sample No. (1)	Sample Values (2)	Sample Mean $(\bar{X})$ (3)
1	(3, 4, 5)	$\frac{1}{3}(3+4+5) = \frac{12}{3} = 4$
2	(3, 4, 6)	$\frac{1}{3}(3+4+6) = \frac{13}{3} = 4.33$
3	(3, 4, 7)	$\frac{1}{3}(3+4+7) = \frac{14}{3} = 4.67$
4	(3, 5, 6)	$\frac{1}{3}(3+5+6) = \frac{14}{3} = 4.67$
5	(3, 5, 7)	$\frac{1}{3}(3+5+7) = \frac{15}{3} = 5.0$
6	(3, 6, 7)	$\frac{1}{3}(3+6+7) = \frac{16}{3} = 5.33$

Notes

7	(4, 5, 6)	$\frac{1}{3}(4+5+6) = \frac{15}{3} = 5.00$
8	(4, 5, 7)	$\frac{1}{3}(4+5+7) = \frac{16}{3} = 5.33$
9	(4, 6, 7)	$\frac{1}{3}(4+6+7) = \frac{17}{3} = 5.67$
10	(5, 6, 7)	$\frac{1}{3}(5+6+7) = \frac{18}{3} = 6.00$
Total	k **= 10**	$\Sigma\overline{X}$ **= 50**

Mean of Sampling Distribution of Means = $\mu_{\overline{X}} = \frac{\Sigma\overline{X}}{k} = \frac{50}{10} = 5.$

Population Mean = $\mu = \frac{3+4+5+6+7}{5} = 5$

Since, $\mu_{\overline{X}} = \mu$, sample mean $\overline{X}$ is an unbiased estimate of the population mean μ.

Example 11: Consider a hypothetical population comprising three values: 1, 2, 3. Draw all possible samples of size 2 with replacement. Calculate the mean $\overline{X}$ and variance s^2 for each sample. Examine whether the two statistics ($\overline{X}$ and s^2) are unbiased and efficient for the corresponding parameters.

Solution: The population consists of three values: 1, 2 and 3. The total number of possible samples of size 2 with replacement are $N^n = 3^2 = 9$ which are given by

Sample No.	**Sample Values**	**Sample Mean** $(\overline{X})$	**Sample Variance** $s^2 = \frac{1}{2}\left[(x_1-\bar{x})^2+(x_2-\bar{x})^2\right]$	**Modified Sample Variance** $\left(\hat{s}^2 = \frac{n}{n-1}s^2\right)$
1.	(1, 1)	$\frac{1}{2}(1+1) = 1.0$	$\frac{1}{2}\cdot\left[(1-1)^2+(1-1)^2\right] = 0.00$	0.00
2.	(1, 2)	$\frac{1}{2}(1+2) = 1.5$	$\frac{1}{2}\cdot\left[(1-1.5)^2+(2-1.5)^2\right] = 0.25$	0.50
3.	(1, 3)	$\frac{1}{2}(1+3) = 2.0$	$\frac{1}{2}\cdot\left[(1-2)^2+(3-2)^2\right] = 1.00$	2.00
4.	(2, 1)	$\frac{1}{2}(2+1) = 1.5$	$\frac{1}{2}\cdot\left[(2-1.5)^2+(1-1.5)^2\right] = 0.25$	0.5
5.	(2, 2)	$\frac{1}{2}(2+2) = 2.0$	$\frac{1}{2}\cdot\left[(2-2)^2+(2-2)^2\right] = 0.00$	0.00
6.	(2, 3)	$\frac{1}{2}(2+2) = 2.5$	$\frac{1}{2}\cdot\left[(2-2.5)^2+(3-2.5)^2\right] = 0.25$	0.50

Notes

7.	(3, 1)	$\frac{1}{2}(3+1) = 2.0$	$\frac{1}{2}\cdot\left[(3-2)^2+(1-2)^2\right] = 1.00$	2.00
8.	(3, 2)	$\frac{1}{2}(3+2) = 2.5$	$\frac{1}{2}\left[(3-2.5)^2+(2-2.5)^2\right] = 0.25$	0.50
9.	(3, 3)	$\frac{1}{2}(3+3) = 3.0$	$\frac{1}{2}\left[(3-3)^2+(3-3)^2\right] = 0.00$	0.00
Total	k **= 9**	$\Sigma\overline{X}$ **= 18**		$\Sigma\hat{s}^2$ **= 6**

(a) Mean of Sampling Distribution of Means = $\mu_{\overline{x}} = \frac{\Sigma\overline{X}}{k} = \frac{18}{9} = 2$. Here, K = No. of samples.

Population Mean $\mu = \frac{1+2+3}{3} = 2.$

Since, $\mu_{\overline{x}} = \mu$, sample mean $\overline{X}$ is an unbiased estimate of the population mean μ.

(b) Mean of the Sampling Distribution of Variance = $\mu_{s^2} = \frac{\Sigma s^2}{k} = \frac{3}{9} = \frac{1}{3}$

Population Variance $\sigma^2 = \frac{(1-2)^2+(2-2)^2(3-2)^2}{3} = \frac{2}{3}$

Since, $\mu_{s^2} \neq \sigma^2$, sample variance s^2 is not an unbiased estimate of the population variance $\left(\sigma^2\right)$.

But the modified sample variance defined as $\hat{s}^2 = \frac{n}{n-1}s^2$ will be unbiased estimate of the population variance σ^2 because:

$$\mu_{\hat{s}^2} = \frac{\Sigma\hat{s}^2}{k} = \frac{6}{9} = \frac{2}{3}$$

$$\sigma^2 = \frac{2}{3}$$

$$\therefore \quad \mu_{\hat{s}^2} = \sigma^2$$

Since $\mu_{\hat{s}^2} = \sigma^2$, the modified sample variation is an unbiased estimate of the population variance.

Example 12: Show that the sample mean $\left(\overline{X}\right)$ is an unbiased estimate of the population mean.

or

An independent random sample $x_1, x_2, x_3, ..., x_n$ is drawn from a population with mean μ. Prove that the expected value of the sample mean $\overline{X}$ equals the population mean μ.

Solution: A random sampling is one where each sample has an equal chance of being selected. We draw a random sample of size 'n'.

Notes

Then,

$$E(\bar{x}) = E\left[\frac{x_1 + x_2 + ... + x_n}{n}\right]$$ Where x_1 is the sample observation.

$$= \frac{1}{n}\cdot\left[E(x_1) + E(x_2) + ...E(x_n)\right]$$

Now the expected values of x_i (a member of the population) is population mean μ.

$$\therefore\ E(\bar{x}) = \frac{1}{n}[\mu + \mu + ... + \mu] \quad \left[\because E(x_1)=E(x_2)=...E(x_n)=\mu\right]$$

$$= \frac{1}{n}\cdot[n\mu] = \mu \qquad \left[\because \Sigma C = C_1 + C_2 + ...C_n = nC\right]$$

Thus, sample mean $\bar{X}$ is an unbiased estimate of population mean.

Self-Assessment

1. Fill in the Blanks:

(*i*) The two types of estimates are ..... and ..... .

(*ii*) The numerical value of a sample mean is said to be an estimate of the population ..... figure.

(*iii*) A point estimate is a single number which is used as an estimate of the unknown ..... parameter.

(*iv*) Point estimate provides one single value of the ..... .

(*v*) Parameter of a sample denoted by ..... .

28.4 Summary

- The topic of estimation in Statistics deals with estimation of population parameters like mean of a statistical distribution. It is assumed, that the concerned variable of the population follows a certain distribution with some parameter(s). For instance, it may be assumed that the life of the electric bulbs follows a normal distribution which has two parameters *viz.* mean (*m*) and standard deviation (σ). While one of the parameters, say, standard deviation is known to be equal to 200 hours from past experience, the other parameter, *viz.* the mean life of the bulbs, is not known, and which we wish to estimate.
- An example of point and interval estimation could be provided from our day-to-day conversation when we talk about commuting time to office. We do make statements like "It takes about 45 minutes ranging from 40 to 50 minutes depending on the traffic conditions." The statistical details of these two types of estimation are described below.
- A point estimate is a single value, like 10, analogous to a point in a geometrical sense. It is used to estimate a population parameter, like mean, with the help of a sample of observations.
- It may be noted that the observations x_1, x_2, x_3, ... x_n are random variables, and therefore, any function of these observations will also be a random variable. Any function of the sample observations is called a **Statistic.**
- the standard deviation of the sample mean is known as the **standard error** of the mean. It is a measure of the extent to which sample means could be expected to vary from sample to sample.

Notes

- No statistic can be guaranteed to provide a close value of the parameter on each and every occasion, and for every sample. Therefore, one has to be content with formulating a rule/method which provides good results in the long run or which has a high probability of success.
- **Incidentally, while the method or rule of estimation is called an estimator like sample mean, the value which the method or rule gives in a particular case is called an estimate.**
- An estimator $\hat{\theta}$ is said be unbiased estimator of the population parameter θ if the mean of the sampling distribution of the estimator $\hat{\theta}$ is equal to the corresponding population parameter θ.
- An estimator is said to be consistent if the estimator approaches the population parameter as the sample size increases. In other words, an estimator $\hat{\theta}$ is said to be consistent estimator of the population parameter θ, if the probability that $\hat{\theta}$ approaches θ is 1 an n becomes large and larger.
- Efficiency is a relative term. Efficiency of an estimator is generally defined by comparing it with another estimator. Let us to take two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$. The estimator $\hat{\theta}_1$ is called an efficient estimator of θ if the variance of $\hat{\theta}_1$ is less than the variance of $\hat{\theta}_2$.
- The last property that a good estimator should possess is sufficiency. An estimator $\hat{\theta}$ is said to be a 'sufficient estimator' of a parameter θ if it contains all the informations in the sample regarding the parameter. In other words, a sufficient estimator utilises all informations that the given sample can furnish about the population. Sample means $\overline{X}$ is said to be a sufficient estimator of the population mean.

28.5 Key-Words

1. Deviation scores : Data in which the mean has been subtracted from each observation.
2. Descriptive statistics : Statistics which describe the sample data without drawing inferences about the larger population.

28.6 Review Questions

1. What is Estimation ? How many types of estimates are possible ?
2. Explain the properties of a good estimator ?
3. What do you understand by point estimator ?
4. Discuss the application of point estimation.
5. Distinguish between consistency and efficiency.

Answers: Self-Assessment

1. (*i*) Point estimate, interval estimate (*ii*) Mean (*iii*) Population
   (*iv*) Parameter (*v*) θ.

28.7 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 29: Methods of Point Estimation and Interval Estimation

CONTENTS

Objectives

After reading this unit students will be able to:

- Discuss the Methods of Point Estimation.
- Explain the Interval Estimation.

Introduction

The object of sampling is to study the features of the population on the basis of sample observations. A carefully selection sample is expected to reveal these features, and hence we shall infer about the population from a statistical analysis of the sample. This process is known as *Statistical Inference.*

There are two types of problems. Firstly, we may have no information at all about some characteristics of the population, especially the values of the parameters involved in the distribution, and it is required to obtain estimates of these parameters. This is the problem of *Estimation.* Secondly, some information or hypothetical values of the parameters may be available, and it is required to test how far the hypothesis is tenable in the light of the information provided by the sample. This is the problem of *Test of Hypothesis* or *Test of Significance.*

Suppose we have a random sample $x_1, x_2, \ldots x_n$ on a variable x, whose distribution in the population involves an unknown parameter θ. It is required to find an estimate of θ on the basis of sample values. The estimation is done in two different ways: (i) *Point Estimation,* and (ii) *Interval Estimation.* In point estimation, the estimated value is given by a single quantity, which is a function of sample observations (*i.e.* statistic). This function is called the '*estimator*' of the parameter, and the value of the estimator in a particular sample is called an '*estimate*'. In interval estimation, an interval within which the parameter is expected to lie is given by using two quantities based on sample values. This is known as *Confidence Interval,* and the two quantities which are used to specify the interval, are known as *Confidence Limits.*

29.1 Methods of Point Estimation

(1) Method of Maximum Likelihood

This is a convenient method for finding an estimator which satisfies most of the criteria discussed earlier. Let $x_1, x_2, \ldots x_n$ be a random sample from a population with p.m.f. (for discrete case) or p.d.f. (for continuous case) $f(x,\theta)$, where θ is the parameter. Then the joint distribution of the sample observations viz.

Notes

$$L = f(x_1, \theta), f(x_2, \theta), ... f(x_n, \theta) \quad ... (1)$$

is called the *Likelihood Function* of the sample.

The *Method of Maximum Likelihood* consists in choosing as an estimator of θ that statistic, which when substituted for θ, maximises the likelihood function L. Such a statistic is called a *maximum likelihood estimator* (m.l.e.). We shall denote the m.l.e. of θ by the symbol θ_0.

Since log L is maximum when L is maximum, in practice the m.l.e. of θ is obtained by maximising log L. This is achieved by differentiating log L partially with respect to θ, and using the two relations

$$\left[\frac{\partial}{\partial \theta} \log L\right]_{\theta=\theta_0} = 0, \left[\frac{\partial^2}{\partial \theta^2} \log L\right]_{\theta=\theta_0} < 0 \quad ... (2)$$

Properties of maximum likelihood estimator (m.l.e.)

(1) The m.l.e. is consistent, most efficient, and also sufficient, provided a sufficient estimator exists.

(2) The m.l.e. is not necessarily unbiased. But when the m.l.e. is biased, by a slight modification, it can be converted into an unbiased estimator.

(3) The m.l.e. tends to be distributed normally for large samples.

(4) The m.l.e. is invariant under functional transformations. This means that if T is an m.l.e. of θ, and $g(\theta)$ is a function of θ, then g (T) is the m.l.e. of $g(\theta)$.

Example 1: On the basis of a random sample find the maximum likelihood estimator of the parameter of a Poisson distribution.

Solution: The Poisson distribution with parameter m has p.m.f.

$$f(x, m) = \frac{e^{-m} \cdot m^x}{x!} \quad (x = 0, 1, 2, ... \infty)$$

The likelihood function of the sample observations is

$$L = f(x_1, m) \cdot f(x_2, m) ... f(x_n, m)$$

$$\therefore \quad \log L = \log f(x_1, m) + \log f(x_2, m) + ... + \log f(x_n, m)$$

$$= \sum_{i=1}^{n} \log f(x_i, m)$$

$$= \sum \left[-m + x_i (\log m) - \log x_i !\right]$$

$$= -nm + (\log m) \Sigma x_i - \Sigma \log (x_i !)$$

Taking partial derivative of log L with respect to the parameter m,

$$\frac{\partial \log L}{\partial m} = -n + \frac{\Sigma x_i}{m} = -n + \frac{n\bar{x}}{m}$$

Now replacing m by m_0 and equating the result to zero,

$$\left[\frac{\partial \log L}{\partial m}\right]_{m = m_0} = -n + \frac{n\bar{x}}{m_0} = 0$$

Notes

Solving, we get $m_0 = \bar{x}$. Again,

$$\left[\frac{\partial^2 \log L}{\partial m^2}\right]_{m = m_0} = -\frac{n\bar{x}}{m_0^2} = -\frac{n\bar{x}}{\bar{x}^2} = -\frac{n}{\bar{x}} \text{ which is negative.}$$

This shows that log L is maximum at $m = m_0 = \bar{x}$. That is the m.l.e. of m is $m_0 = \bar{x}$, the sample mean.

Example 2: Find the maximum likelihood estimator of the variance σ^2 of a Normal population $N(\mu, \sigma^2)$, when the parameter μ is known. Show that this estimator is unbiased.

Solution: The p.d.f. of Normal distribution is

$$f(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}; \ (-\infty < x < +\infty)$$

The logarithm (to the base e) of the likelihood function L is

$$\log L = \sum_{i=1}^{n} \log f(x_i, \mu, \sigma^2)$$

$$= \sum\left[-\log\sigma - \frac{1}{2}\log(2x) - \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

$$= -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2x) - \frac{\sum(x_i - \mu)^2}{2\sigma^2}$$

Differentiating partially with respect to σ^2,

$$\frac{\partial}{\partial(\sigma^2)}\log L = -\frac{n}{2\sigma^2} + \frac{\sum(x_i - \mu)^2}{2(\sigma^2)^2}$$

The m.l.e. of σ^2 is obtained by solving

$$-\frac{n}{2\sigma^2} + \frac{\sum(x_i - \mu)^2}{2\sigma_0{}^4} = 0$$

$$\therefore \quad \sigma_0{}^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

It can be shown that $\left[\frac{\partial^2 \log L}{\partial(\sigma^2)^2}\right]_{\sigma^2 = \sigma_0^2} = \frac{-n}{2\sigma_0^4}$

which is negative. Thus the maximum likelihood estimator of σ^2 is

$$\sigma_0{}^2 = \sum\frac{(x_i - \mu)^2}{n}, \ (\mu \text{ known})$$

Notes

Again, since $x_1, x_2, \ldots x_n$ is a random sample and μ is the population mean, we have $E(x_i - \mu)^2 = \sigma^2$. Therefore,

$$E(\sigma_0^2) = \sum_{i=1}^{n} \frac{E(x_i - \mu)^2}{n} = \frac{\sum \sigma^2}{n} = \sigma^2$$

Thus, σ_0^2 is an unbiased estimator of σ^2.

Example 3: Find the m.l.e. of the parameters μ and σ^2 in random samples from a $N(\mu, \sigma^2)$ population, when both the parameters are unknown.

Solution: As in the preceding example,

$$\log L = -\frac{n}{2}\log \sigma^2 - n \log \sqrt{2\pi} - \frac{\sum (x_i - \mu)^2}{2\sigma^2}$$

$$\therefore \quad \left[\frac{\partial \log L}{\partial \mu}\right]_{\mu=\mu_0} = \frac{-1}{2\sigma^2} \Sigma 2(x_i - \mu_0)(-1) = 0$$

This gives $\Sigma(x_i - \mu_0) = 0$; *i.e.*, $\mu_0 = \bar{x}$, the sample mean. The m.l.e. of the parameter μ is the sample mean $\bar{x}$. (Note that this estimator is *unbiased*).

Proceeding as in Example 2 we have $\sigma_0^2 = \frac{\Sigma (x_i - \mu)^2}{n}$. But since the parameter μ is not known, it is replaced by the m.l.e. $\mu_2 = \bar{x}$. The m.l.e. of σ^2 is now

$$\sigma_0^2 = \frac{\Sigma (x_i - \bar{x})^2}{n} = S^2$$

which is the sample variance. (Note that this estimator is *biased*).

Example 4: A tossed a biased coin 50 times and got head 20 times, while B tossed it 90 times and got 40 heads. Find the maximum likelihood estimate of the probability of getting head when the coin is tossed.

Solution: Let P be the unknown probability of obtaining a head. Using binomial distribution,

Probability of 20 heads in 50 tosses = ${}^{50}C_{20}\, P^{20}(1-P)^{30}$

Probability of 40 heads in 90 tosses = ${}^{90}C_{40}\, P^{40}(1-P)^{50}$

The likelihood function is given by the product of these probabilities:

$$L = {}^{50}C_{20} \cdot {}^{90}C_{40}\, P^{30}(1-P)^{30}$$

$$\therefore \quad \log L = \log\left({}^{50}C_{20}\, {}^{90}C_{40}\right) + 60 \log P + 80 \log(1-P)$$

Hence, $$\frac{\partial \log L}{\partial P} = \frac{60}{P} - \frac{80}{1-P}$$

The maximum likelihood estimate P_0 is therefore obtained by solving

Notes

$$\frac{60}{P_0} - \frac{80}{1-P_0} = 0.$$

This gives $P_0 = \frac{60}{140} = \frac{3}{7}$. **Ans.** 3/7

(2) Method of Moments

The *Method of Moments* consists in equating the first few moments of the population with the corresponding moments of the sample *i.e.* setting

$$\mu_r' = m_r' \quad \ldots (3)$$

where $\mu_r' = E(x')$ and $m_r' = \frac{\sum x_i^r}{n}$. Since the parameters enter into the population moments, these relations when solved for the parameters give the estimates by the method of moments. Of course, this method is applicable only when the population moments exist. The method is generally applied for fitting theoretical distributions to observed data.

Example 5: Estimate the parameter *p* of the binomial distribution by the method of moments (when *n* is known).

Solution: For the binomial distribution $\mu_1' = E(x) = np$. Also $m_1' = \bar{x}$. Setting $\mu_1' = m_1'$, we have $np = \bar{x}$. Thus

$$p = \frac{\bar{x}}{n}$$

i.e. the estimated value of *p* is given by the sample mean divided by the parameter *n* (known).

Example 6: Find the estimates of μ and σ in the Normal population $N(\mu, \sigma^2)$ by the method of moments.

Solution: Equate the first two moments of the population and the sample, $\mu_1' = m_1'$ and $\mu_2' = m_2'$,

i.e. $\mu_2 = m_2$. Thus

$\mu = \bar{x}$ and $\sigma^2 = S^2$, the sample variance.

The parameters μ and σ are estimated by the sample mean $\bar{x}$ and the sample standard deviation S respectively.

29.2 Interval Estimation

In the theory of point estimation, developed earlier, any unknown parameter is estimated by a single quantity. Thus the sample mean $(\bar{x})$ is used to estimate the population mean (μ), and the sample proportion (*p*) is taken as an estimator of the population proportion (P). A single estimator of this kind, however good it may be, cannot be expected to coincide with the true value of the parameter, and may in some cases differ widely from it. In the theory of interval estimation, it is desired to find an interval which is expected to include the unknown parameter with a specified probability.

Let $x_1, x_2, \ldots x_n$ be a random sample from a population of a known mathematical form which involves an unknown parameter θ. We would try to find two functions t_1 and t_2 based on sample observations such that the probability of θ being included in the interval (t_1, t_2) has a given value, say *c*.

$$P(t_1 \le \theta \le t_2) = c \quad \ldots (4)$$

Notes

Such an interval, when it exists, is called a *Confidence Interval* for θ. The two quantities t_1 and t_2 which serve as the lower and upper limits of the interval are known as *Confidence Limits*. The probability (c) with which the confidence interval will include the true value of the parameter is known as *Confidence Coefficient* of the interval.

The significance of confidence limits is that if many independent random samples are drawn from the same population and the confidence interval is calculated from each sample, then the parameter will actually be included in the intervals in c proportion of cases in the long run. Thus the estimate of the parameter is stated as an interval with a specified degree of confidence.

The calculation of confidence limits is based on the knowledge of sampling distribution of an appropriate statistic. Suppose, we have a random sample of size n from a Normal population $N(\mu,\sigma^2)$, where the variance σ^2 is known. It is required to find 95% confidence limits for the unknown parameter μ. We know that the sample mean $(\bar{x})$ follows normal distribution with mean μ and variance σ^2/n, and so

$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

has a standard normal distribution. Since 95% of the area under the standard normal curve lies between the ordinates at $z = \pm 1.96$, we have

$$P\left[-1.96 \le \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right] = 0.95$$

i.e. in 95% of cases the following inequalities hold

$$-1.96 \le \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96$$

Separating out μ we get

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

The interval $\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ is known as the 95% confidence interval for μ, and the 95% confidence limits are

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

Again, 99% of area under the standard normal curve lies between the ordinates at $z = \pm 2.58$, and 99.73% (*i.e.* almost whole) of the area lies between $z = \pm 3$. Hence proceeding exactly in the same manner, the 99% confidence limits for μ are

$$\bar{x} \pm 2.58\frac{\sigma}{\sqrt{n}}$$

Notes

and almost sure limits for μ are:

$$\bar{x} \pm 3\frac{\sigma}{\sqrt{n}}$$

In fact, using values from the normal probability integral table (showing areas under standard normal curve), confidence limits corresponding to any specified percentage can be obtained. These are *exact* confidence limits.

In some cases, the population may not be truely a normal distribution, but the sampling distributions of statistics based on large samples are approximately normal. For example, the sample mean $(\bar{x})$ based on a large random sample drawn (with or without replacement) from any population is approximately normally distributed. Similarly, the sample proportion (*p*) calculated from a large random sample has approximately a normal distribution. It is therefore possible to utilise the properties relating to the percentage of area under the standard normal curve to find approximate confidence limits for the population mean μ and the population proportion P, provided the sample size *n* is large.

Approximate Confidence Limits (large samples) any Distribution

(1) for Mean μ:

95% confidence limits = $\bar{x} \pm 1.96(\text{S.E. of } \bar{x})$

99% confidence limits = $\bar{x} \pm 2.58(\text{S.E. of } \bar{x})$

Almost sure limits = $\bar{x} \pm 3(\text{S.E. of } \bar{x})$ ... (5)

(2) for Proportion P:

95% confidence limits = $p \pm 1.96(\text{S.E. of } p)$

99% confidence limits = $p \pm 2.58(\text{S.E. of } p)$

Almost sure limits = $p \pm 3(\text{S.E. of } p)$ ... (6)

(3) for Difference of Means $(\mu_1 - \mu_2)$:

95% confidence limits = $(\bar{x}_1 - \bar{x}_2) \pm 1.96(\text{S.E. of } \bar{x}_1 - \bar{x}_2)$

99% confidence limits = $(\bar{x}_1 - \bar{x}_2) \pm 2.58(\text{S.E. of } \bar{x}_1 - \bar{x}_2)$

Almost sure limits = $(\bar{x}_1 - \bar{x}_2 \pm 3)(\text{S.E. of } \bar{x}_1 - \bar{x}_2)$ ... (7)

(4) for Difference of Proportions $(P_1 - P_2)$:

95% confidence limits = $(p_1 - p_2) \pm 1.96(\text{S.E. of } p_1 - p_2)$

99% confidence limits = $(p_1 - p_2) \pm 2.58(\text{S.E. of } p_1 - p_2)$

Almost sure limits = $(p_1 - p_2) \pm 3(\text{S.E. of } p_1 - p_2)$ ... (8)

Notes

Notes (i) The '*probable limits*' (without any reference to the degree of confidence) may be taken to be '*almost sure limits*' in all the above cases.

(ii) The formulae for S.E. involve population parameters. If these parameters are not known, an approximate value of S.E. may be obtained by substituting the statistic for the corresponding parameter.]

Example 7: A sample of 6500 screws is taken form a large consignment and 75 are found to be defective. Estimate the percentage of defectives in the consignment and assign limits within which the percentage lies.

Solution: There are 75 defectives in a sample of size $n = 600$. Therefore, the sample proportion of defectives is

$$p = \frac{75}{600} = \frac{1}{8} = 12.5\%$$

This may be taken as an estimate of the percentage of defectives (P) in the whole consignment ('Point estimation').

The 'limits' to the percentage of defectives refer to the confidence limits, which may be given as $p \pm 3$ (S.E. of p).

$$\text{S.E. of } p = \sqrt{\frac{PQ}{n}}$$

$$= \sqrt{\frac{pq}{n}} \text{ approximately;}$$

(since the population proportion P is not known).

$$= \sqrt{\frac{\frac{1}{8}\left(1-\frac{1}{8}\right)}{600}} = \frac{1}{80}\sqrt{\frac{7}{6}} = .0135$$

Thus, the limits for P are

$$\frac{1}{8} \pm 3 \times .0135 = .125 \pm .0405$$

$$= .1655 \text{ and } .0845 = 16.55\% \text{ and } 8.45\%$$

The limits to the percentage of defectives in the consignment are 8.45% to 16.55% ('Interval estimation').

Ans. 12.5%; 8.45% to 16.55%.

Example 8: A random sample of 100 ball bearings selected from a shipment of 2000 ball bearings has an average diameter of 0.354 inch with a S.D. = .048 inch. Find 95% confidence interval for the average diameter of these 2000 ball bearings.

Solution: Theory: If a random sample of large size n is drawn *without replacement* from a *finite population* of size N, hen the 95% confidence limits for the population mean μ are $\bar{x} \pm 1.96$ (S.E. of x), where $\bar{x}$ denotes the sample mean and

$$\text{S.E. of } \bar{x} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Notes

σ denoting the Standard Deviation (S.D.) of the population. Here,

Sample size (n) = 100, Population size (N) = 2000

Sample mean $(\bar{x})$ = 0.354, Sample S.D. (S) = .048

Since σ is not known, an approximate value of S.E. is obtained on replacing the population S.D. (σ) by the sample S.D. (S).

$$\text{S.E. of } \bar{x} = \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \text{ approximately}$$

$$= \frac{.048}{\sqrt{100}}\sqrt{\frac{2000-100}{2000-1}} = .0047$$

The 95% confidence limits for the population mean μ are

$$\bar{x} \pm 1.96(\text{S.E.of}\,\bar{x}) = 0.354 \pm 1.96 \times .0047$$

$$= 0.354 \pm .0092 = 0.3632 \text{ and } 0.3448$$

Thus, the 95% confidence interval is (0.3448 to 0.3632) inch.

Example 9: A random sample of 100 articles taken from an large batch of articles contains 5 defective articles. (a) Set up 96 per cent confidence limits for the proportion of defective articles in the batch. (b) If the batch contains 2696 articles set up 95% confidence interval for the proportion of defective articles.

Solution: (a) The 96% confidence limits for the population proportion (P) are given by $p \pm 2.05$ (S.E. of p), where p is the sample proportion.

$$\text{S.E. of } p = \sqrt{\frac{PQ}{n}}$$

Since the formula involves the unknown population proportion P, an approximate value of S.E. is obtained on replacing the population proportion (P) by the sample proportion (p). Putting n = 100 and p = 5/100 = .05, (q = 1 – p = .95)

$$\text{S.E. of } p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{.05 \times .95}{100}} = .022$$

Hence, the 96% confidence limits for P are

$$p \pm 2.05 \text{ (S.E. of } p) = .05 \pm 2.05 \times .022 = .05 \pm .045$$

$$= .05 + .045 \text{ and } .05 - .045$$

$$= .095 \text{ and } .005$$

(b) The 95% confidence limits for proportion (P) are given by $p \pm 1.96$ (S.E. of p). But, when the population is of a finite size N,

$$\text{S.E. of } p = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}} \text{ (approximately)}$$

Here, n = 100, N = 2696, p = .05. Putting these values

$$\text{S.E. of } p = \sqrt{\frac{05 \times .95}{100}}\sqrt{\frac{2696-100}{2696-1}} = .022\sqrt{\frac{2596}{2696}}$$

$$= .022 \times .963 = .022 \times .98 = .0216 \text{ (approx.)}$$

Hence, the required 95% confidence limits for P are

Notes

$$p \pm 1.96 \text{ (S.E. of } p) = .05 \pm 1.96 \times .0216 = .092 \text{ and } .008$$

The 95% *confidence interval* for the proportion of defective articles is 008 to .092.

Ans. .005 and .095; .008 to .092

Example 10: 10 Life Insurance Polices in a sample of 200 taken out of 50,000 were found to be insured for less than Rs. 5,000. How many policies can be reasonably expected to be insured for less than Rs. 5,000 in the whole lot at 95%. Confidence level ?

Solution: Let us first find the confidence limits for the 'proportion' of life insurance policies insured for less than Rs. 5,000 in the whole lot. Here,

$$\text{Population size (N)} = 50{,}000,$$

$$\text{Sample size } (n) = 200$$

$$\text{Sample proportion } (p) = \frac{10}{200} = .05$$

Using,

$$\text{S.E. of } p = \sqrt{\frac{.05 \times .95}{200}} \sqrt{\frac{50{,}000 - 200}{50{,}000 - 1}}$$

$$= \sqrt{\frac{.05 \times .95 \times 49800}{200 \times 50{,}000}} \text{ (approx.)}$$

$$= .0154$$

95% confidence limits for the population proportion (P) are

$$p \pm 1.96 \text{ (S.E. of } p) = .05 \pm 1.96 \times .0154$$

$$= .05 \pm .030 = .080 \text{ and } .020$$

The means that out of the lot of 50,000, the 'proportion' of policies insured for less than Rs. 5000 lies between .020 and .080, with probability 95%. Thus the 'number' of such policies lies between 50,000 × .020 = 1000 and 50,000 × .080 = 4,000.

Ans. Between 1,000 and 4,000.

Exact Confidence Limits (any sample size) Normal Distribution

In the following cases, it is assumed that samples are drawn at random from Normal populations.

(5) for Mean μ: (s.d. known)

$$\text{95\% confidence limits less} = \bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\text{99\% confidence limits} = \bar{x} \pm 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{... (9)}$$

(6) for Difference of Means $(\mu_1 - \mu_2)$: (s.d.s known)

$$\text{95\% confidence limits} = (\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\text{99\% confidence limits} = (\bar{x}_1 - \bar{x}_2) \pm 2.58\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{... (10)}$$

Notes

Example 11: A random sample of size 10 was drawn from a normal population with an unknown mean and a variance of 44.1 (inch)^2. If the observations are (in inches): 65, 71, 80, 76, 78, 82, 68, 72, 65 and 81, obtain the 95% confidence interval for the population mean.

Solution: We are given $n = 10$, $\sigma^2 = 44.1$ and $\Sigma x_i = 738$.

$$\therefore \qquad \bar{x} = \frac{738}{10} = 73.8.$$

Since the population s.d. σ is known, using formula (9), 95% confidence limits for μ are given by

$$73.8 \pm 1.96\frac{\sqrt{44.1}}{\sqrt{10}} = 73.8 \pm 1.96 \times \sqrt{4.41}$$

$$= 73.8 \pm 1.96 \times 2.1$$

$$= 73.8 \pm 4.1 = 77.9 \text{ and } 69.7.$$

The 95% confidence interval for μ is therefore 69.7 to 77.9 inches.

(7) for Mean *m*: (s.d. unknown)

In random samples from a Normal population $N(\mu, \sigma^2)$

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n-1}}$$

follows t distribution with $(n - 1)$ degree of freedom, where S is the sample s.d. If $t_{.025}$ denotes the upper 2.5% point of t distribution with $(n - 1)$ d.f., then the 95% confidence interval for μ is obtained from

$$-t_{.025} \le \frac{\bar{x} - \mu}{S/\sqrt{n-1}} \le t_{.025}$$

Hence, for the population mean μ

95% confidence limits = $\bar{x} \pm t_{.025}\left(\frac{S}{\sqrt{n-1}}\right)$ (11 a)

99% confidence limits = $\bar{x} \pm t_{.005}\left(\frac{S}{\sqrt{n-1}}\right)$ (11 b)

(8) for Difference of Means $(\mu_1 - \mu_2)$: (common s.d. unknown)

Assuming that two independent samples are drawn from two Normal populations with means μ_1, μ_2 but a common *unknown* s.d. σ.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_1}}}$$

follows t distribution with $(n_1 + n_2 - 2)$ d.f., where

$$s^2 = \frac{(n_1 S_1^2 + n_2 S_2^2)}{(n_1 + n_2 - 2)}$$

Notes

Hence, with 95% probability the following inequalities hold

$$-t_{.025} \le \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \le t_{.025}$$

from which the 95% confidence limits for $(\mu_1 - \mu_2)$ are

$$(\bar{x}_1 - \bar{x}_2) \pm t_{.025} \cdot s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{... (12 a)}$$

Similarly the 99% confidence limits for $(\mu_1 - \mu_2)$ are

$$(\bar{x}_1 - \bar{x}_2) \pm t_{.005} \cdot s\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{... (12 b)}$$

Example 12: A random sample of 10 students of class II was selected from schools in a certain region, and their weights recorded are shown below (in. lb.): 38, 46, 45, 40, 35, 39, 44, 45, 33, 37. Find 95% confidence limits within which the mean weight of all such students in the region is expected to lie. (Given $t_{.025}$ = 2.262 for 9 d.f. and 2.228 for 10 d.f.).

Solution: From the given data, $\bar{x} = \frac{402}{10} = 40.2$. To calculate the s.d. (S), we take deviations from 40, *i.e.* $d = x - 40$.

– 2, 6, 5, 0, – 5, – 1, 4, 5, – 7, – 3

$$\sum d = 2, \ \sum d^2 = 190$$

$$\therefore \quad S^2 = \frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2 = \frac{190}{10} - \left(\frac{2}{10}\right)^2 = 18.96$$

$$S = \sqrt{18.96} = 4.35.$$

Since the population s.d. σ is *unknown* the 95% confidence limits for μ are (see formula 12 a)

$40.2 \pm 2.262 \times 4.35/\sqrt{9}$ (degrees of freedom = 9)

= 40.2 ± 3.28 = 36.92 and 43.48

The 95% confidence limits for the mean weight are (in lb.) 36.9 and 43.5.

(9) for Variance σ^2

Case I:

(*mean known*)—In random samples from $N(\mu, \sigma^2)$ population, $\sum(x_i - \mu)^2/\sigma^2$ follows chi-square distribution with *n* degrees of freedom. If $\chi^2_{.975}$ and $\chi^2_{.025}$ denote the lower and the upper 2.5% points of chi-square distribution with *n* d.f., then with probability 95% we have

$$\chi^2_{.975} \le \sum \frac{(x_i - \mu)^2}{\sigma^2} \le \chi^2_{.025}$$

From which the 95% confidence interval for σ^2 is

Notes

$$\frac{\Sigma(x_i-\mu)^2}{\chi^2_{.025}} \le \sigma^2 \le \frac{\Sigma(x_i-\mu)^2}{\chi^2_{.975}} \quad \text{... (13)}$$

Case II:

(*mean unknown*) — In this case $\frac{nS^2}{\sigma^2} = \frac{\Sigma(x_i-\bar{x})^2}{\sigma^2}$ follows chi-square distribution with $(n-1)$ degrees of freedom. Using the lower and the upper 2.5% points of chi-square distribution with $(n-1)$ d.f., we have with probability 95% the following inequalities

$$\chi^2_{.975} \le \frac{nS^2}{\sigma^2} \le \chi^2_{.025}$$

from which the 95% confidence interval for σ^2 can be given as

$$\frac{nS^2}{\chi^2_{.025}} \le \sigma^2 \le \frac{nS^2}{\chi^2_{.975}} \quad \text{... (14)}$$

Example 13: The standard deviation of a random sample of size 12 drawn from a normal population is 5.5 Calculate the 95% confidence limits for the standard deviation (σ) in the population (Given $\chi^2_{.975}$ = 3.82 and $\chi^2_{.025}$ = 21.92 for 11 degrees of freedom).

Solution: Here n = 12 and the sample s.d. (S) = 5.5. Substituting the values in formula (14), the 95% confidence interval for σ^2 is

$$\frac{12\times(5.5)^2}{21.92} \le \sigma^2 \le \frac{12\times(5.5)^2}{3.82}$$

or, $$16.56 \le \sigma^2 \le 95.03$$

i.e., $$4.1 \le \sigma \le 9.7$$

The 95% confidence limits for the population s.d. (σ) are 4.1 and 9.7.

Example 14: A sample of size 8 from a normal population yields as the ubiased estimate of population variance the value 4.4. Obtain the 99% confidence limits for the population variance σ^2 (Given $\chi^2_{.975}$ = 0.99 and $\chi^2_{.005}$ = 20.3 for 7 d.f.).

Solution: Here n = 8, and the unbiased estimate s^2 = 4.4. So, $nS^2 = (n-1)s^2$ = 7 × 4.4 = 30.8.

Hence the 99% confidence limits for σ^2 are obtained from

$$\frac{nS^2}{\chi^2_{.005}} \le \sigma^2 \le \frac{nS^2}{\chi^2_{.975}}$$

or, $$\frac{30.8}{20.3} \le \sigma^2 \le \frac{30.8}{0.99};\ \textit{i.e.,}\ 1.52 \le \sigma^2 \le 31.1$$

Notes

(10) for Variance-Ratio $\frac{\sigma_1^2}{\sigma_2^2}$: (means unknown)

If two independent random samples of sizes n_1 and n_2 are drawn from two Normal populations with unknown means μ_1, μ_2 and variances σ_1^2, σ_2^2 respectively, then $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ follows F distribution with degrees of freedom $(n_1 - 1, n_2 - 1)$. If $F_{.975}$ and $F_{.025}$ denote the lower and the upper 2.5% points of F distribution, we have with probability 95% the following inequalities

$$F_{.975} \le \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \le F_{.025} \quad \text{... (15)}$$

The 95% confidence interval for σ_1^2/σ_2^2 can be obtained from this as

$$\frac{1}{F_{.025}} \cdot \frac{s_1^2}{s_2^2} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{1}{F_{.975}} \cdot \frac{s_1^2}{s_2^2} \quad \text{... (16)}$$

where s_1^2 and s_2^2 denote unbiased estimators of σ_1^2, σ_2^2 respectively from the two samples.

Self-Assessment

1. Tick the Correct Answer

(*i*) What is the best description of a point estimate ?
- (*a*) any value from the sample used to estimate a parameter.
- (*b*) a sample statistic used to estimate a parameter.
- (*c*) the margin of error used to estimate a parameter.

(*ii*) Which best describes the lower endpoint of a confidence interval ?
- (*a*) point estimate
- (*b*) point estimate minus margin of error
- (*c*) point estimate plus margin of error

(*iii*) Which best describes the upper endpoint of a confidence interval ?
- (*a*) point estimate
- (*b*) point estimate minus margin of error
- (*c*) point estimate plus margin of error

(*iv*) Which value will be at the centre of a confidence interval
- (*a*) population parameter
- (*b*) point estimate
- (*c*) margin of error

(*v*) What is the relationship between a 95% confidence interval and a 99% confidence interval from the same sample ?
- (*a*) the 95% interval will be wider
- (*b*) the 99% interval will be wider
- (*c*) both intervals have the some width

Notes

29.3 Summary

- The object of sampling is to study the features of the population on the basis of sample observations. A carefully selection sample is expected to reveal these features, and hence we shall infer about the population from a statistical analysis of the sample. This process is known as *Statistical Inference.*
- In interval estimation, an interval within which the parameter is expected to lie is given by using two quantities based on sample values. This is known as *Confidence Interval,* and the two quantities which are used to specify the interval, are known as *Confidence Limits.*
- The *Method of Maximum Likelihood* consists in choosing as an estimator of θ that statistic, which when substituted for θ, maximises the likelihood function L. Such a statistic is called a *maximum likelihood estimator* (m.l.e.). We shall denote the m.l.e. of θ by the symbol θ_0.
- The parameters enter into the population moments, these relations when solved for the parameters give the estimates by the method of moments. Of course, this method is applicable only when the population moments exist. The method is generally applied for fitting theoretical distributions to observed data.
- In the theory of point estimation, developed earlier, any unknown parameter is estimated by a single quantity. Thus the sample mean $(\bar{x})$ is used to estimate the population mean (μ), and the sample proportion (*p*) is taken as an estimator of the population proportion (P). A single estimator of this kind, however good it may be, cannot be expected to coincide with the true value of the parameter, and may in some cases differ widely from it. In the theory of interval estimation, it is desired to find an interval which is expected to include the unknown parameter with a specified probability.
- The significance of confidence limits is that if many independent random samples are drawn from the same population and the confidence interval is calculated from each sample, then the parameter will actually be included in the intervals in *c* proportion of cases in the long run. Thus the estimate of the parameter is stated as an interval with a specified degree of confidence.

29.4 Key-Words

1. First order interaction : The interaction of two variables. Also known as a "simple interaction."
2. Fixed marginal totals : The situation in which the marginal totals in a contingency table are known before the data are collected and are not subject to sampling error.
3. Fixed model : Anova An analysis of variance model in which the levels of the independent variable are treated as fixed.

29.5 Review Questions

1. Discuss the methods of point estimation.
2. What is the difference between point estimation and interval estimation ? Is interval estimation better than point estimation ?
3. Explain the procedure of constructing a confidence interval for estimating population mean μ.
4. Explain interval estimation.
5. The central limit theorem for sample proportion can be used for estimating the population proportion. Elaborate.

Notes

Answers: Self-Assessment

1. (*i*) (*b*) (*ii*) (*b*) (*iii*) (*c*) (*iv*) (*b*) (*v*) (*b*)

29.6 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 30: Types of Hypothesis: Null and Alternative, Types of Errors in Testing Hypothesis and Level of Significance

CONTENTS

Objectives

After reading this unit students will be able to:

- Explain Null and Alternative Hypothesis.
- Know the Types of Errors in Testing Hypothesis.
- Discuss the Level of Significance.

Introduction

In unit 29 we showed how a sample could be used to develop point and interval estimates of population parameters. In this unit we continue the discussion of statistical inference by showing how hypothesis testing can be used to determine whether a statement about the value of a population parameter should or should not be rejected.

In hypothesis testing we begin by making a tentative assumption about a population parameter. This tentative assumption is called the null hypothesis and is denoted by H_0. We then define another hypothesis, called the alternative hypothesis, which is the opposite of what is stated in the null hypothesis. We denote the alternative hypothesis by H_1. The hypothesis testing procedure uses data from a sample to assess the two competing statements indicated by H_0 and H_1.

This unit shows how hypothesis tests can be conducted about a population mean and a population proportion. We begin by providing examples of approaches to formulating null and alternative hypotheses.

30.1 Null and Alternative Hypothesis

(a) Null Hypothesis: The null hypothesis asserts that there is no real difference in the sample and the population in the particular matter under consideration and that the difference found is accidental and unimportant arising out of fluctuations of sampling. The null hypothesis constitutes a challenge and the function of the experiment is to give the facts a chance to refute or fail to refute this challenge.

For example, if we want to find out whether the new vaccine has benefited the people or not, the null hypothesis, shall be set up saying that "the new vaccine has not benefited the people". The rejection of the null hypothesis indicates that the differences have statistical significance and the acceptance of the null hypothesis indicate that the differences are due to chance.

Notes

(b) **Alternative Hypothesis:** The alternative hypothesis specifies those values that the researcher believes to hold true and hopes that the sample data would lead to acceptance of this hypothesis to be true. The alternative hypothesis may embrace the whole range of values rather than single point.

As per this definition, it is very difficult to find out which is null hypothesis and which one is alternative hypothesis.

However, for statistical convenience, the hypothesis these definitions are used.

The null hypotheses are represented by the symbol H_0 and the alternative hypothesis is represented by H_1.

Developing null and alternative hypotheses

In some applications it may not be obvious how the null and alternative hypotheses should be formulated. Care must be taken to structure the hypotheses appropriately so that the conclusion from the hypothesis test provides the information the researcher or decision-maker wants. Learning to formulate hypotheses correctly will take practice. The examples in this section show a variety of forms for H_0 and H_1 depending upon the application. Guidelines for establishing the null and alternative hypotheses will he given for three types of situations in which hypothesis testing procedures are commonly used.

Statistics in Practice

Monitoring the quality of latex condoms

Many consumer products are required by law to meet specifications set out in documents known as standards. This is particularly the case when there are issues of consumer safety, such as with electrical goods, children's toys or furniture (fire resistance). In less safety-critical cases, the standards may be permissive rather than obligatory, but manufacturers will often conform to the standards, and tell consumers so, as an assurance of quality. Many standards are established internationally and are embodied in documents published by the International Standards Organization (ISO). Companies based in the UK and other EU countries usually operate according to ISO standards.

The humble latex condom is the subject of ISO standard 4074: 2002. This lays down a range of specifications, relating to materials, dimensions, packaging and performance criteria including, for obvious reasons, freedom from holes. For an outline of quality testing procedures, read the relevant pages at www.durex.com, for example. ISO 4047: 2002 makes reference to other standards documents including frequent references to ISO 4859-1, which lays down specifications for the sampling schemes that must be used to ensure quality, such as in respect of freedom from holes, For the latter characteristic, ISO 4047: 2002 specifies an acceptable quality level (AQL) of no more than 0.25 per cent defective condoms in any manufacturing batch; in other words, a probability of no more than 1 in 400 that any particular condom will be defective.

As an example of the sampling specifications, suppose ISO 4859-1 requires that, from a batch of 10000 condoms, a random sample of 200 should be taken and examined individually for freedom from holes. This is likely to be a destructive test. Suppose ISO 4859-1 then stipulates that the whole batch from which the sample was drawn can be declared satisfactory, in respect of freedom from holes, only if the sample contains no more than one defective condom. If the sample does contain more than one defective condom, the whole batch must be scrapped, or further tests of quality must be done to gather further information about the overall quality of the batch.

A statistician would refer to this decision procedure as a hypothesis test. The working hypothesis is that the batch conforms to the AQL specified in ISO 4074: 2002.

If the sampled batch contains more than one defective condom, this hypothesis is rejected. Otherwise the hypothesis is accepted. In making the decision on the basis of sample evidence, the quality controller is taking two risks. One risk is that a batch meeting the AQL requirement will be incorrectly rejected. The second risk is that a batch not meeting the AQL requirement will be incorrectly accepted. The sampling schemes laid down in ISO 4859-1 are intended to clarify and restrict the level of risk involved, and to strike a sensible balance between the two types of risk.

In this chapter you will learn about the logic of statistical hypothesis testing.

Notes

Testing research hypotheses

Consider a particular model of ear that currently attains an average fuel consumption of 7 litres of fuel per 100 kilometres of driving. A product research group develops a new fuel injection system specifically designed to decrease the fuel consumption. To evaluate the new system, several will be manufactured, installed in cars, and subjected to research-controlled driving tests. Here the product research group is looking for evidence to conclude that the new system *decreases* the mean fuel consumption. In this case, the research hypothesis is that the new fuel injection system will provide a mean litres-per-100 *km* rating below 7; that is, $\mu < 7$. As a general guideline, a research hypothesis should be stated as the *alternative hypothesis*. Hence, the appropriate null and alternative hypotheses for the study are:

$$H_0: \mu \geq 7$$

$$H_1: \mu < 7$$

If the sample results indicate that H_0 cannot be rejected, researchers cannot conclude that the new fuel injection system is better. Perhaps more research and subsequent testing should be conducted. However, if the sample results indicate that H_0 can be rejected, researchers can make the inference that $H_1: \mu < 7$ is true. With this conclusion, the researchers gain the statistical support necessary to state that the new system decreases the mean fuel consumption. Production with the new system should be considered.

In research studies such as these, the null and alternative hypotheses should he formulated so that the rejection of H_0 supports the research conclusion. The research hypothesis therefore should be expressed as the alternative hypothesis.

Testing the validity of a claim

As an illustration of testing the validity of a claim, consider the situation of a manufacturer of soft drinks who states that bottles of its products contain an average of at least 1.5 litres. A sample of bottles will be selected, and the contents will be measured to test the manufacturer's claim. In this type of hypothesis testing situation, we generally assume that the manufacturer's claim is true unless the sample evidence is contradictory Using this approach for the soft-drink example, we would state the null and alternative hypotheses as follows.

$$H_0: \mu \geq 1.5$$

$$H_1: \mu < 1.5$$

In the sample results indicate H_0 cannot be rejected, the manufacturer's claim will not be challenged. However, if the sample results indicate H_0 can he rejected, the inference will be made that $H_1: \mu < 1.5$ is true. With this conclusion, statistical evidence indicates that the manufacturer's claim is incorrect and that the soft-drink bottles are being filled with a mean less than the claimed 1.5 litres. Appropriate action against the manufacturer may be considered.

In any situation that involves testing the validity of a claim, the null hypothesis is generally based on the assumption that the claim is true. The alternative hypothesis is then formulated so that rejection of H_0 will provide statistical evidence that the stated assumption is incorrect. Action to correct the claim should be considered whenever H_0 is rejected.

Testing in decision-making situations

In testing research hypotheses or testing the validity of a claim, action is taken if H_0 is rejected. In many instances, however, action must be taken both when H_0 cannot be rejected and when H_0 can be rejected. In general, this type of situation occurs when a decision-maker must choose between two

Notes

courses of action, one associated with the null hypothesis and another associated with the alternative hypothesis. The quality-testing scenario outlined in the Statistics in Practice at the beginning of the unit is an example of this.

Suppose that, on the basis of a sample of parts from a shipment just received, a quality control inspector must decide whether to accept the shipment or to return the shipment to the supplier because it does not meet specifications. The specifications for a particular part require a mean length of two centimetres per part. If the mean length is greater or less than the two-centimeter standard, the parts will cause quality problems in the assembly operation. In this case, the null and alternative hypothesis would be formulated as follows.

$$H_0: \mu = 2$$

$$H_1: \mu \neq 2$$

If the sample results indicate H_0 cannot be rejected, the quality control inspector will have no reason to doubt that the shipment meets specifications, and the shipment will be accepted. However, if the sample results indicate H_0 should be rejected, the conclusion will be that the parts do not meet specifications. In this case, the quality control inspector will have sufficient evidence to return the shipment to the supplier. We see that for these types of situations, action is taken both when H_0 cannot be rejected and when H_0 can be rejected.

Summary of forms for null and alternative hypothesis

The hypothesis tests in this unit involve one of two population parameters: the population mean and the population proportion. Depending on the situation, hypothesis tests about a population parameter may take one of three forms; two include inequalities in the null hypothesis, the third uses only an equality in the null hypothesis. For hypothesis tests involving a population mean, we let μ_0 denote the hypothesized value and choose one of the following three forms for the hypothesis test.

$$H_0: \mu \geq \mu_0 \qquad H_0: \mu \leq \mu_0 \qquad H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0 \qquad H_1: \mu > \mu_0 \qquad H_1: \mu \neq \mu_0$$

For reasons that will be clear later, the first two forms are called one-tailed tests. The third form is called a two-tailed test.

In many situations, the choice of H_0 and H_1 is not obvious and judgment is necessary to select the proper form. However, as the preceding forms show, the equality part of the expression (either $\geq$, $\leq$ or =) *always* appears in the null hypothesis. In selecting the proper form of H_0 and H_1, keep in mind that the alternative hypothesis is often what the test is attempting to establish. Hence, asking whether the user is looking for evidence to support $\mu < \mu_0$, $\mu_1 > \mu_0$ or $\mu \neq \mu_0$ will help determine H_1. The following exercises are designed to provide practice in choosing the proper form for a hypothesis test involving a population mean.

30.2 Types of Errors in Testing Hypothesis

Ideally the hypothesis testing procedure should lead to the acceptance of the null hypothesis H_0 when it is true, and the rejection of H_0 when it is not. However, the correct decision is not always possible. Since the decision to reject or accept a hypothesis is based on sample data, there is a possibility of an incorrect decision or error. A decision-maker may commit two types of errors while testing a null hypothesis. The two types of errors that can be made in any hypothesis testing are shown in Table 1.

Notes

Table 1: Errors in Hypothesis Testing

Decision	State of Nature	
	H_0 is True	**H_0 is False**
Accept H_0	Correct decision with confidence $(1-\alpha)$	Type II error (β)
Reject H_0	Type I error (α)	Correct decision $(1-\beta)$

Type I Error: This is the *probability of rejecting the null hypothesis when it is true* and some alternative hypothesis is wrong. The probability of making a Type I error is denoted by the symbol α. It is represented by the area under the sampling distribution curve over the region of rejection.

Hypothesis testing: The process of testing a statement or belief about a population parameter by the use of information collected from a sample(s).

Type I error: The probability of rejecting a true null hypothesis.

The probability of making a Type I error, is referred to as the level of significance. The probability level of this error is decided by the decision-maker before the hypothesis test is performed and is based on his tolerance in terms of risk of rejecting the true null hypothesis. The risk of making Type I error depends on the cost and/or goodwill loss. The complement $(1-\alpha)$ of the probability of Type I error measures the probability level of not rejecting a true null hypothesis. It is also referred to as *confidence level*.

Type II Error: This is the *probability of accepting the null hypothesis when it is false* and some alternative hypothesis is true. The probability of making a Type II is denoted by the symbol β.

The probability of Type II error varies with the actual values of the population parameter being tested when null hypothesis H_0 is false. The probability of committing a Type II error depends on five factors: (i) the actual value of the population parameter, being tested, (ii) the level of significance selected, (iii) type of test (one or two tailed test) used to evaluate the null hypothesis, (iv) the sample standard deviation (also called standard error) and (v) the size of sample.

A summary of certain critical values at various significance levels for test statistic *z* is given in Table 30.2.

Level of significance: The probability of rejecting a true null hypothesis due to sampling error.

Type II error: The probability of accepting a false null hypothesis.

Table 2: Summary of Certain Critical Values for Sample Statistic *z*

Rejection Region	Level of Significance, α per cent			
	α = 0.10	**α = 0.05**	**0.01**	**α = 0.005**
One-tailed region	± 1.285	± 1.645	± 2.33	± 2.58
Two-tailed region	± 1.645	± 1.96	± 2.58	± 2.81

Power of a Statistical Test

Another way of evaluating the goodness of a statistical test is to look at the complement of Type II error, which is stated as:

$1-\beta$ = P (reject H_0 when H_1 is true)

Notes

The complement $1 - \beta$ of β, *i.e.* the probability of Type-II error, is called the *power of a statistical test* because it measures the probability of rejecting H_0 when it is true.

For example, suppose null and alternative hypotheses are stated as.

$$H_{0:}\ \mu = 80 \text{ and } H_{1:}\ \mu = 80$$

Power of a test: The ability (probability) of a test to reject the null hypothesis when it is false.

Often, when the null hypothesis is false, another alternative value of the population mean, μ is unknown. So for each of the possible values of the population mean μ, the probability of committing Type II error for several possible values of μ is required to be calculated.

Suppose a sample of size n = 50 is drawn from the given population to compute the probability of committing a Type II error for a specific alternative value of the population mean, μ. Let sample mean so obtained be $\bar{x}$ = 71 with a standard deviation, s = 21. For significance level, α = 0.05 and a two-tailed test, the table value of $z_{0.05} = \pm 1.96$. But the deserved value from sample data is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{71 - 80}{21/\sqrt{50}} = -3.03$$

Since $z_{cal} = -3.03$ value falls in the rejection region, the null hypothesis H_0 is rejected. The rejection of null hypothesis, leads to either make a correct decision or commit a Type II error. If the population mean is actually 75 instead of 80, then the probability of commiting a Type II error is determined by computing a critical region for the mean $\bar{x}_c$. This value is used as the cutoff point between the area of acceptance and the area of rejection. If for any sample mean so obtained is less than (or greater than for right-tail rejection region), $\bar{x}_c$, then the null hypothesis is rejected. Solving for the critical value of mean gives

$$z_c = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} \text{ or } \pm 1.96 = \frac{\bar{x}_c - 80}{21/\sqrt{50}}$$

$$\bar{x}_c = 80 \pm 5.82 \text{ or } 74.18 \text{ to } 85.82$$

If μ = 75, then probability of accepting the false null hypothesis H_0: μ = 80 when critical value is falling in the range $\bar{x}_c$ = 74.18 to 85.82 is calculated as follows:

$$z_1 = \frac{74.18 - 75}{21/\sqrt{50}} = -0.276$$

The area under normal curve for $z_1 = -0.276$ is 0.1064.

$$z_2 = \frac{85.82 - 75}{21\sqrt{50}} = 3.643$$

The area under normal carve for z_2 = 3.643 is 0.4995

Thus the probability of committing a Type II error (β) falls in the region:

$$\beta = P(74.18 < \bar{x}_c < 85.82) = 0.1064 + 0.4995 = 0.6059$$

The total probability 0.6059 of committing a Type II error (β) is the area to the right of $\bar{x}_c$ = 74.18 in the distribution. Hence the power of the test is $1 - \beta$ = 1 – 0.6059 = 0.3941 as shown in Figure 1 (b).

Notes

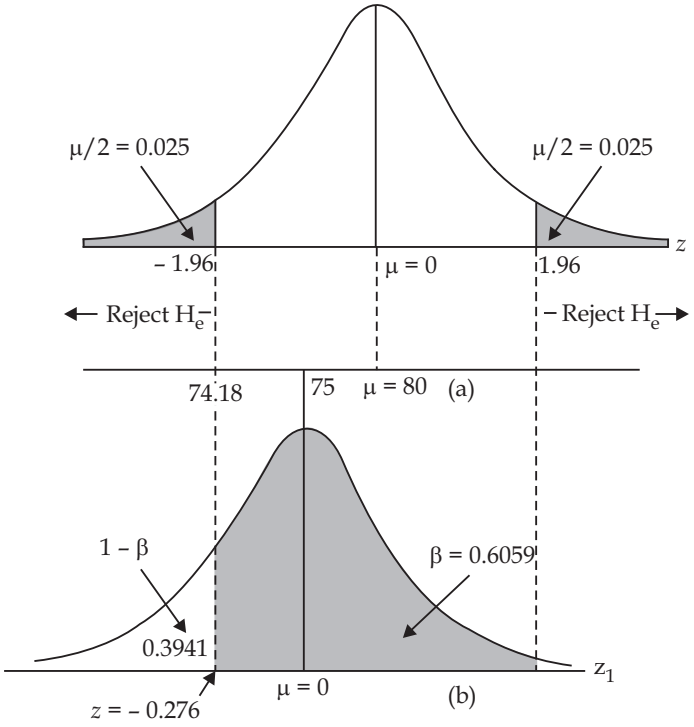


Figure 1 (a): Sampling distribution with H_0: μ = 80

Figure 1 (b): Sampling distribution with H_0: μ = 75

To keep α or β low depends on which type of error is more costly. However, if both types of errors are costly, then to keep both α and β low, then inferences can be made more reliable by reducing the variability of observations. It is preferred to have large sample size and a low α value.

Few relations between two errors α and β, the power of a test $1-\beta$, and the sample size n are stated below:

(i) If α (the sum of the two tail areas in the curve) is increased, the shaded area corresponding to β gets smaller, and vice versa.

(ii) The β value call be increased for a fixed α, by increasing the sample size n.

Special Case: Suppose hypotheses are defined as:

$$H_0: \mu = 80 \text{ and } H_1: \mu < 80$$

Given n = 50, s = 21 and $\bar{x}$ = 71. For α = 0.05 and left-tailed test, the table value $z_{0.05}$ = – 1.645. The observed z value from sample data is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{71 - 80}{21/\sqrt{50}} = -3.03$$

The critical value of the sample mean $\bar{x}_c$ for a given population mean μ = 80 is given by:

$$z_c = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} \text{ or } -1.645 = \frac{\bar{x}_c - 80}{21/\sqrt{50}}$$

Notes

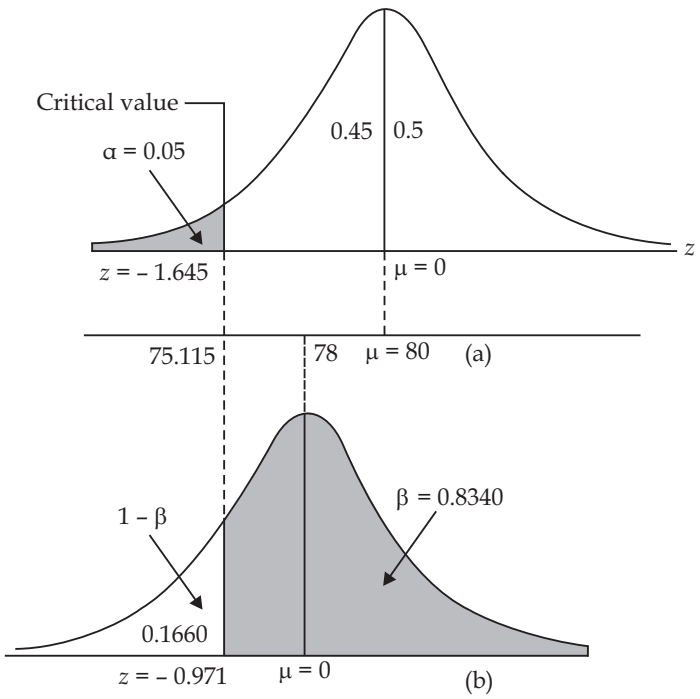


Figure 2 (a): Sampling distribution with H_0: μ = 80

Figure 2 (b): Sampling distribution with H_0: μ = 78

Figure 2 (a) shows that the distribution of values that contains critical value of mean $\bar{x}_c$ = 75.115 and below which H_0 will be rejected. Figure 2 (b) shows the distribution of values when the alternative population mean value μ = 78 is true. If H_0 is false, it is not possible to reject null hypothesis H_0 whenever sample mean is in the acceptance region, $\bar{x} \geq 75.151$. Thus critical value is computed by extending it and solved for the area to the right of $\bar{x}_c$ as follows:

$$z_1 = \frac{\bar{x}_c - \mu}{\sigma_{\bar{x}}} = \frac{75.115 - 78}{21/\sqrt{50}} = -0.971$$

This value of z yields an area of 0.3340 under the normal curve. Thus the probability = 0.3340 + 0.5000 = 0.8340 of committing a Type II error is all the area to right of $\bar{x}_c$ = 75.115.

Remark: In general, if alternative value of population mean μ is relatively more than its hypothesized value, then probability of committing a Type II error is smaller compared to the case when the alternative value is close to the hypothesized value. The probability of committing a Type II error decreases as alternative values are greater than the hypothesized mean of the population.

30.3 The Level of Significance

In Section 30.2, we introduced hypothesis testing along rather traditional lines: we defined the parts of a statistical test along with the two types of errors and their associated probabilities α and $\beta(\mu_a)$. The problem with this approach is that if other researchers want to apply the results of your study

Notes

using a different value for α then they must compute a new rejection region before reaching a decision concerning H_0 and H_a. An alternative approach to hypothesis testing follows the following steps: specify the null and alternative hypotheses, specify a value for α, collect the sample data, and determine the weight of evidence for rejecting the null hypothesis. This weight, given in terms of a probability, is called the level of significance (or ***p*-value**) of the statistical test. More formally, the level of significance is defined as follows: *the probability of obtaining a value of the test statistic that is as likely or more likely to reject H_0 as the actual observed value of the test statistic, assuming that the null hypothesis is true.* Thus, if the level of significance is a small value, then the sample data fail to support H_0 and our decision is to reject H_0. On the other hand, if the level of significance is a large value, then we fail to reject H_0. We must next decide what is a large or small value for the level of significance.

Decision Rule for Hypothesis Testing **Using the *p*-Value**

1. If the *p*-value $\leq \alpha$, then reject H_0.
2. If the *p*-value $> \alpha$, then fail to reject H_0.

We illustrate the calculation of a level of significance with several examples.

Example 1 : (a) Determine the level of significance (*p*-value) for the statistical test and reach a decision concerning the research hypothesis using α = .01.

(b) If the preset value of α is .05 instead of .01, does your decision concerning H_a change ?

Solution :

(a) The null and alternative hypotheses are

$$H_0: \mu \leq 380$$

$$H_a: \mu > 380$$

From the sample data, with *s* replacing σ, the computed value of the test statistic is

$$z = \frac{\bar{y} - 380}{\sigma / \sqrt{n}} = \frac{390 - 380}{35.2 / \sqrt{50}} = 2.01$$

The level of significance for this test (*i.e.,* the weight of evidence for rejecting H_0) is the probability of observing a value of $\bar{y}$ greater than or equal to 390 assuming that the null hypothesis is true; that is, μ = 380. This value can be computed by using the *z*-value of the test statistic, 2.01, because *p*-value = $P(\bar{y} \geq 390$, assuming μ = 380) = $P(z \geq 2.01)$

Referring to Table 30.1 in the Appendix, $P(z \geq 2.01) = 1 - P(z < 2.01) = 1 - .9778 =$ .0222. This value is shown by the shaded area in Figure 3. Because the *p*-value is greater than α (.0222 > .01). we fail to reject H_0 and conclude that the data do not support the research hypothesis.

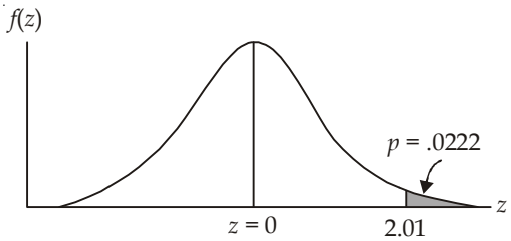


Figure 3: Level of significance for Example 1

Notes

(b) Another person examines the same data but with a preset value for α = .05. This person is willing to support a higher risk of a Type I error, and hence the decision is to reject H_0 because the *p*-value is less than α $(.0222 \le .05)$. It is important to emphasize that the value of α used in the decision rule is *preset* and not selected after calculating the *p*-value.

As we can see from Example 1, the level of significance represents the probability of observing a sample outcome more contradictory to H_0 than the observed sample result. *The smaller the value of this probability, the heavier the weight of the sample evidence against* H_0. For example, a statistical test with a level of significance of *p* = .01 shows more evidence for the rejection of H_0 than does another statistical test with *p* = .20.

Example 2 : Using a preset value of α = .05, is there sufficient evidence in the data to support the research hypothesis ?

Solution : The null and alternative hypotheses are

$$H_0 : \mu \ge 33$$

$$H_\alpha : \mu < 33$$

From the sample data, with *s* replacing σ, the computed value of the test statistic is

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{31.2 - 33}{8.4/\sqrt{35}} = -1.27$$

The level of significance for this test statistic is computed by determining which values of $\bar{y}$ are more extreme to H_0 than the observed $\bar{y}$. Because H_α specifies μ less than 33, the values of $\bar{y}$ that would be more extreme to H_0 are those values less than 31.2, the observed value. Thus,

p-value = $P(\bar{y} \le 31.2$, assuming μ = 33) = $P(z \le -1.27)$ = .1020

There is considerable evidence to support H_0. More precisely, *p*-value = .1020 > .05 = α, and hence we fail to reject H_0. Thus, we conclude that there is insufficient evidence (*p*-value = .1020) to support the research hypothesis. Note that this is exactly the same conclusion reached using the traditional approach.

For two-tailed tests, $H_\alpha : \mu \ne \mu_0$, we still determine the level of significance by computing the probability of obtaining a sample having a value of the test statistic that is more contradictory to H_0 than the observed value at the test statistic. However, for two-tailed research hypotheses, we compute this probability in terms of the magnitude of the distance from $\bar{y}$ to the null value of μ because both values of $\bar{y}$ much less than μ_0 and values of $\bar{y}$ much larger than μ_0 contradict $\mu = \mu_0$. Thus, the level of significance is written as

$$p\text{-value} = P\left(|\bar{y} - \mu_0| \ge \text{ observed } |\bar{y} - \mu_0|\right) = P\left(|z| \ge |\text{computed } z|\right)$$

$$= 2P\left(z \ge |\text{computed } z|\right)$$

To summarize, the level of significance (*p*-value) can be computed as

Notes

Case 1	Case 2	Case 3
$H_0 : \mu \le \mu_0$	$H_0 : \mu \ge \mu_0$	$H_0 : \mu = \mu_0$
$H_\alpha : \mu > \mu_0$	$H_\alpha : \mu < \mu_0$	$H_\alpha : \mu \ne \mu_0$
p-value: $P(z \ge \text{computed } z)$	$P(z \le \text{computed } z)$	$2P(z \ge \lvert\text{computed } z\rvert)$

Example 3 : Using a preset value of $\alpha = .01$, is there sufficient evidence in the data to support the research hypothesis ?

Solution : The null and alternative hypotheses are

$$H_0 : \mu = 190$$

$$H_\alpha : \mu \ne 190$$

From the sample data, with s replacing σ, the computed value of the test statistic is

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{178.2 - 190}{45.3/\sqrt{100}} = -2.60$$

The level of significance for this test statistic is computed using the formula on page 248.

$$p\text{-value} = 2P(z \ge |\text{computed } z|) = 2P(z \ge |-2.60|) = 2P(z \ge 2.60)$$

$$= 2(1 - .9953) = .0047$$

Because the p-value is very small, there is very little evidence to support H_0. More precisely, p-value = $.0047 \le .05 = \alpha$, and hence we reject H_0. Thus, there is sufficient evidence (p-value = .0047) to support the research hypothesis and conclude that the mean cholesterol level differs from 190. Note that this is exactly the same conclusion reached using the traditional approach.

There is much to he said in favor of this approach to hypothesis testing. Rather than reaching a decision directly, the statistician (or person performing the statistical test) presents the experimenter with the weight of evidence for rejecting the null hypothesis. The experimenter can then draw his or her own conclusion. Some experimenters reject a null hypothesis if $p \le .10$, whereas others require $p \le .05$ or $p \le .01$ for rejecting the null hypothesis. The experimenter is left to make the decision based on what he or she believes is enough evidence to indicate rejection of the null hypothesis.

Many professional journals have followed this approach by reporting the results of a statistical test in terms of its level of significance. Thus, we might read that a particular test was significant at the $p = .05$ level or perhaps the $p < .01$ level. By reporting results this way, the reader is left to draw his or her own conclusion.

One word of warning is needed here. The p-value of .05 has become a magic level, and many seem to feel that a particular null hypothesis should not be rejected unless the test achieves the .05 level or lower. This has resulted in part from the decision-based approach with α preset at .05. Try not to fall into this trap when reading journal articles or reporting the results of your statistical tests. After all, statistical significance at a particular level does not dictate importance or practical significance. Rather, it means that a null hypothesis can be rejected with a specified low risk of error. For example, suppose that a company is interested in determining whether the average number of miles driven per car per month for the sales force has risen above 2,600. Sample data from 400 cars show that $\bar{y}$ = 2,640 and s = 35. For these data, the z statistic for H_0: μ = 2,600 is z = 22.86 based on σ = 35; the level of significance is $p < .0000000001$. Thus, even though there has only been a 1.5% increase in the average

Notes

monthly miles driven for each car, the result is (highly) statistically significant. Is this increase of any practical significance ? Probably not. What we have done is proved *conclusively* that the mean μ has increased slightly.

The company should not just examine the size of the *p*-value. It is very important to also determine the size of the difference between the null value of the population mean μ_0 and the estimated value of the population mean $\bar{y}$. This difference is called the estimated *effect size*. In this example the estimated effect size would be $\bar{y} - \mu_0 = 2{,}640 - 2{,}600 = 40$ miles driven per month. This is the quantity that the company should consider when attempting to determine if the change in the population mean has practical significance.

Throughout the text we will conduct statistical tests from both the decision-based approach and from the level-of-significance approach to familiarize you with both avenues of thought. For either approach, remember to consider the practical significance of your finding after drawing conclusions based on the statistical test.

Self Assessment

1. Fill in the Blanks:

(*i*) The first important step in the decision making procedure is to state the ..... hypothesis.

(*ii*) When the hypothesis is true but the test rejects it, it is called ..... error.

(*iii*) When the hypothesis is false and the test accepts it, it is called ..... error.

(*iv*) The confidence with which an experimenter rejects. Or retains a null hypothesis depends upon the ..... level adopted.

(*v*) The alternative hypothesis may embrace the whole range of value rather than ..... point.

30.4 Summary

- In hypothesis testing we begin by making a tentative assumption about a population parameter. This tentative assumption is called the null hypothesis and is denoted by H_0. We then define another hypothesis, called the alternative hypothesis, which is the opposite of what is stated in the null hypothesis. We denote the alternative hypothesis by H_1. The hypothesis testing procedure uses data from a sample to assess the two competing statements indicated by H_0 and H_1.
- The null hypothesis asserts that there is no real difference in the sample and the population in the particular matter under consideration and that the difference found is accidental and unimportant arising out of fluctuations of sampling. The null hypothesis constitutes a challenge and the function of the experiment is to give the facts a chance to refute or fail to refute this challenge.
- The rejection of the null hypothesis indicates that the differences have statistical significance and the acceptance of the null hypothesis indicate that the differences are due to chance.
- The alternative hypothesis specifies those values that the researcher believes to hold true and hopes that the sample data would lead to acceptance of this hypothesis to be true. The alternative hypothesis may embrace the whole range of values rather than single point.
- The null hypotheses are represented by the symbol H_0 and the alternative hypothesis is represented by H_1.
- Care must be taken to structure the hypotheses appropriately so that the conclusion from the hypothesis test provides the information the researcher or decision-maker wants. Learning to formulate hypotheses correctly will take practice. The examples in this section show a variety of forms for H_0 and H_1 depending upon the application. Guidelines for establishing the null and alternative hypotheses will he given for three types of situations in which hypothesis testing procedures are commonly used.

Notes

- Many consumer products are required by law to meet specifications set out in documents known as standards. This is particularly the case when there are issues of consumer safety, such as with electrical goods, children's toys or furniture (fire resistance). In less safety-critical cases, the standards may be permissive rather than obligatory, but manufacturers will often conform to the standards, and tell consumers so, as an assurance of quality. Many standards are established internationally and are embodied in documents published by the International Standards Organization (ISO). Companies based in the UK and other EU countries usually operate according to ISO standards.
- In making the decision on the basis of sample evidence, the quality controller is taking two risks. One risk is that a batch meeting the AQL requirement will be incorrectly rejected. The second risk is that a batch not meeting the AQL requirement will be incorrectly accepted. The sampling schemes laid down in ISO 4859-1 are intended to clarify and restrict the level of risk involved, and to strike a sensible balance between the two types of risk.
- In research studies such as these, the null and alternative hypotheses should he formulated so that the rejection of H_0 supports the research conclusion. The research hypothesis therefore should be expressed as the alternative hypothesis.
- In any situation that involves testing the validity of a claim, the null hypothesis is generally based on the assumption that the claim is true. The alternative hypothesis is then formulated so that rejection of H_0 will provide statistical evidence that the stated assumption is incorrect. Action to correct the claim should be considered whenever H_0 is rejected.
- In testing research hypotheses or testing the validity of a claim, action is taken if H_0 is rejected. In many instances, however, action must be taken both when H_0 cannot be rejected and when H_0 can be rejected. In general, this type of situation occurs when a decision-maker must choose between two courses of action, one associated with the null hypothesis and another associated with the alternative hypothesis.
- The hypothesis tests in this unit involve one of two population parameters: the population mean and the population proportion. Depending on the situation, hypothesis tests about a population parameter may take one of three forms; two include inequalities in the null hypothesis, the third uses only an equality in the null hypothesis.
- The hypothesis testing procedure should lead to the acceptance of the null hypothesis H_0 when it is true, and the rejection of H_0 when it is not. However, the correct decision is not always possible. Since the decision to reject or accept a hypothesis is based on sample data, there is a possibility of an incorrect decision or error.
- The probability of making a Type I error, is referred to as the level of significance. The probability level of this error is decided by the decision-maker before the hypothesis test is performed and is based on his tolerance in terms of risk of rejecting the true null hypothesis. The risk of making Type I error depends on the cost and/or goodwill loss. The complement $(1-\alpha)$ of the probability of Type I error measures the probability level of not rejecting a true null hypothesis. It is also referred to as *confidence level*.
- The ability (probability) of a test to reject the null hypothesis when it is false. Often, when the null hypothesis is false, another alternative value of the population mean, μ is unknown. So for each of the possible values of the population mean μ, the probability of committing Type II error for several possible values of μ is required to be calculated.E
- However, if both types of errors are costly, then to keep both α and β low, then inferences can be made more reliable by reducing the variability of observations. It is preferred to have large sample size and a low α value.
- However, if both types of errors are costly, then to keep both α and β low, then inferences can be made more reliable by reducing the variability of observations. It is preferred to have large sample size and a low α value.

Notes

- An alternative approach to hypothesis testing follows the following steps: specify the null and alternative hypotheses, specify a value for α, collect the sample data, and determine the weight of evidence for rejecting the null hypothesis. This weight, given in terms of a probability, is called the level of significance (or ***p*-value**) of the statistical test. More formally, the level of significance is defined as follows: *the probability of obtaining a value of the test statistic that is as likely or more likely to reject H_0 as the actual observed value of the test statistic, assuming that the null hypothesis is true.* Thus, if the level of significance is a small value, then the sample data fail to support H_0 and our decision is to reject H_0. On the other hand, if the level of significance is a large value, then we fail to reject H_0. We must next decide what is a large or small value for the level of significance.

30.5 Key-Words

1. Harmonic mean : The number of elements to be averaged divided by the sum of the reciprocals of the elements.
2. Heavy tailed distribution : A distribution with a higher percentage of scores in the tails than we would expect in a normal distribution.
3. Heterogeneity of variance : A situation in which samples are drawn from populations having different variances.

30.6 Review Questions

1. What is meant by the terms hypothesis and a test of a hypothesis ?
2. What do you understand by null hypothesis and alternative hypothesis ? Explain developing null and alternative hypothesis.
3. Discuss the types of errors in testing hypothesis.
4. Define term 'level of significance'. How is it related to the probability of committing a Type I error.
5. How is power related to the probability of making a Type II error ?

Answers: Self-Assessment

1. (*i*) Null (*ii*) type I (*iii*) type II
   (*iv*) significance (*v*) single

30.7 Further Readings

Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

LOVELY PROFESSIONAL UNIVERSITY
Jalandhar-Delhi G.T. Road (NH-1)
Phagwara, Punjab (India)-144411
For Enquiry: +91-1824-300360
Fax.: +91-1824-506111
Email: odl@lpu.co.in