LOVELY PROFESSIONAL UNIVERSITY

QUANTITATIVE TECHNIQUES – I

Edited By

Dr. Dilfraz Singh

Printed by
EXCEL BOOKS PRIVATE LIMITED
A-45, Naraina, Phase-I,
New Delhi-110028
for
Lovely Professional University
Phagwara

SYLLABUS

Quantitative Techniques – I

Objectives: To familiarize the student with the theoretical foundation of qualitative business research. To familiarize the students with different statistical techniques used in business decisions.

Sr. No.	Description
1.	**Statistics:** Introduction, Importance, Scope and Limitations of Statistics
2.	**Classification, Tabulation and Presentation of Data:** Geographical, Chronological, Qualitative and Quantitative Classification, Formation of Frequency Distribution, Tabulation of Data, Types of Tables, Bar Diagrams, Pie Diagrams, Pictograms and Cartograms.
3.	**Collection of Data:** Primary and Secondary Data, Method of Collecting Data, Drafting the Questionnaire, Sources of Secondary Data.
4.	**Measures of Central Tendency:** Mean, Harmonic Mean, Geometric Mean, Median and Mode
5.	**Measures of Dispersion:** Range, Mean deviation, Quartile deviation, Standard deviation, Coefficient of variation
6.	**Correlation Analysis and Regression Analysis:** Scatter Diagram, Karl Pearson's, Rank Correlation, Regression Equations – Deviation taken from Assumed mean and Arithmetic Mean, Least square method, Graphing Regression Lines
7.	**Index Number:** Methods of Constructing index Number – Laspeyres, Paasche, Bowley's, Fisher and Marshall-Edgeworth method, Chain base Index Number
8.	**Analysis of Time Series:** Method of Semi-average, Moving average, Simple average, Ratio-to-trend method, Ratio-to-Moving average Method
9.	**Probability and Expected value:** Addition Theorem, Multiplicative Theorem
10.	**Probability Distribution:** Binomial, Poisson and Normal Distribution

CONTENT

Unit 1:	Statistics *Neha Tikoo, Lovely Professional University*	1
Unit 2:	Classification of Data *Neha Tikoo, Lovely Professional University*	20
Unit 3:	Tabulation *Neha Tikoo, Lovely Professional University*	41
Unit 4:	Presentation of Data *Neha Tikoo, Lovely Professional University*	54
Unit 5:	Collection of Data *Neha Tikoo, Lovely Professional University*	77
Unit 6:	Measures of Central Tendency *Neha Tikoo, Lovely Professional University*	90
Unit 7:	Measures of Dispersion *Neha Tikoo, Lovely Professional University*	130
Unit 8;	Correlation Analysis *Dilfraz Singh, Lovely Professional University*	166
Unit 9:	Regression Analysis *Tanima Dutta, Lovely Professional University*	185
Unit 10:	Index Number *Dilfraz Singh, Lovely Professional University*	200
Unit 11:	Analysis of Time Series *Tanima Dutta, Lovely Professional University*	224
Unit 12:	Probability and Expected Value *Dilfraz Singh, Lovely Professional University*	247
Unit 13:	Binomial Probability Distribution *Tanima Dutta, Lovely Professional University*	267
Unit 14:	Poisson Probability Distribution *Dilfraz Singh, Lovely Professional University*	284
Unit 15:	Normal Probability Distribution *Tanima Dutta, Lovely Professional University*	298

Notes

Unit 1: Statistics

CONTENTS

Objectives

Introduction

1.1 Meaning, Definition and Characteristics of Statistics

1.1.1 Statistics as a Scientific Method

1.1.2 Statistics as a Science or an Art

1.2 Importance of Statistics

1.3 Scope of Statistics

1.4 Limitations of Statistics

1.5 Summary

1.6 Keywords

1.7 Review Questions

1.8 Further Readings

Objectives

After studying this unit, you will be able to:

- Define and explain the meaning, origin and growth of statistics
- Discuss the importance of statistics
- State the characteristics of statistics
- Discuss Statistics as a Science or an Art
- Describe the limitations and scope of statistics

Introduction

Modern age is the age of science which requires that every aspect, whether it pertains to natural phenomena, politics, economics or any other field, should be expressed in an unambiguous and precise form. A phenomenon expressed in ambiguous and vague terms might be difficult to understand in proper perspective. Therefore, in order to provide an accurate and precise explanation of a phenomenon or a situation, figures are often used. The statement that prices in a country are increasing, conveys only an incomplete information about the nature of the problem. However, if the figures of prices of various years are also provided, we are in a better position to understand the nature of the problem. In addition to this, these figures can also be used to compare the extent of price changes in a country vis-a-vis the changes in prices of some other country. Using these figures, it might be possible to estimate the possible level of prices at some future date so that some policy measures can be suggested to tackle the problem. The subject which deals with such type of figures, called data, is known as Statistics.

Notes

1.1 Meaning, Definition and Characteristics of Statistics

The meaning of the word 'Statistics' is implied by the pattern of development of the subject. Since the subject originated with the collection of data and then, in later years, the techniques of analysis and interpretation were developed, the word 'statistics' has been used in both the plural and the singular sense. Statistics, in plural sense, means a set of numerical figures or data. In the singular sense, it represents a method of study and therefore, refers to statistical principles and methods developed for analysis and interpretation of data.

Statistics has been defined in different ways by different authors. These definitions can be broadly classified into two categories. In the first category are those definitions which lay emphasis on statistics as data whereas the definitions in second category emphasise statistics as a scientific method.

Statistics used in the plural sense implies a set of numerical figures collected with reference to a certain problem under investigation. It may be noted here that any set of numerical figures cannot be regarded as statistics. There are certain characteristics which must be satisfied by a given set of numerical figures in order that they may be termed as statistics. Before giving these characteristics it will be advantageous to go through the definitions of statistics in the plural sense, given by noted scholars.

1. "Statistics are numerical facts in any department of enquiry placed in relation to each other." – *A.L. Bowley*

   The main features of the above definition are:

   (i) Statistics (or Data) implies numerical facts.

   (ii) Numerical facts or figures are related to some enquiry or investigation.

   (iii) Numerical facts should be capable of being arranged in relation to each other.

   On the basis of the above features we can say that data are those numerical facts which have been expressed as a set of numerical figures related to each other and to some area of enquiry or research. We may, however, note here that all the characteristics of data are not covered by the above definition.

2. "By statistics we mean quantitative data affected to a marked extent by multiplicity of causes." – *Yule & Kendall*

   This definition covers two aspects, i.e., the data are quantitative and affected by a large number of causes.

3. "Statistics are classified facts respecting the conditions of the people in a state – especially those facts which can be stated in numbers or in tables of numbers or in any other tabular or classified arrangement." – *Webster*

On the basis of the above definitions we can now state the following characteristics of statistics as data:

1. ***Statistics are numerical facts:*** In order that any set of facts can be called as statistics or data, it must be capable of being represented numerically or quantitatively. Ordinarily, the facts can be classified into two categories: (a) Facts that are measurable and can be represented by numerical measurements. Measurement of heights of students in a college, income of persons in a locality, yield of wheat per acre in a certain district, etc., are examples of measurable facts. (b) Facts that are not measurable but we can feel the presence or absence of the characteristics. Honesty, colour of hair or eyes, beauty, intelligence, smoking habit, etc., are examples of immeasurable facts. Statistics or data can be obtained in such cases also, by counting the number of individuals in different categories. For

Notes

example, the population of a country can be divided into three categories on the basis of complexion of the people such as white, whitish or black.

2. ***Statistics are aggregate of facts:*** A single numerical figure cannot be regarded as statistics. Similarly, a set of unconnected numerical figures cannot be termed as statistics. Statistics means an aggregate or a set of numerical figures which are related to one another. The number of cars sold in a particular year cannot be regarded as statistics. On the other hand, the figures of the number of cars sold in various years of the last decade is statistics because it is an aggregate of related figures. These figures can be compared and we can know whether the sale of cars has increased, decreased or remained constant during the last decade.

   It should also be noted here that different figures are comparable only if they are expressed in same units and represent the same characteristics under different situations. In the above example, if we have the number of Ambassador cars sold in 1981 and the number of Fiat cars sold in 1982, etc., then it cannot be regarded as statistics. Similarly, the figures of, say, measurement of weight of students should be expressed in the same units in order that these figures are comparable with one another.

3. ***Statistics are affected to a marked extent by a multiplicity of factors:*** Statistical data refer to measurement of facts in a complex situation, e.g., business or economic phenomena are very complex in the sense that there are a large number of factors operating simultaneously at a given point of time. Most of these factors are even difficult to identify. We know that quantity demanded of a commodity, in a given period, depends upon its price, income of the consumer, prices of other commodities, taste and habits of the consumer. It may be mentioned here that these factors are only the main factors but not the only factors affecting the demand of a commodity. Similarly, the sale of a firm in a given period is affected by a large number of factors. Data collected under such conditions are called statistics or statistical data.

4. ***Statistics are either enumerated or estimated with reasonable standard of accuracy:*** This characteristic is related to the collection of data. Data are collected either by counting or by measurement of units or individuals. For example, the number of smokers in a village are counted while height of soldiers is measured. We may note here that if the area of investigation is large or the cost of measurement is high, the statistics may also be collected by examining only a fraction of the total area of investigation.

   When statistics are being obtained by measurement of units, it is necessary to maintain a reasonable degree or standard of accuracy in measurements. The degree of accuracy needed in an investigation depends upon its nature and objectivity on the one hand and upon time and resources on the other. For example, in weighing of gold, even milligrams may be significant where as, for weighing wheat, a few grams may not make much difference. Sometimes, a higher degree of accuracy is needed in order that the problem, to be investigated, gets highlighted by the data. Suppose the diameter of bolts produced by a machine are measured as 1.546 cms, 1.549 cms, 1.548 cms, etc. If, instead, we obtain measurements only up to two places after decimal, all the measurements would be equal and as such nothing could be inferred about the working of the machine. In addition to this, the degree of accuracy also depends upon the availability of time and resources. For any investigation, a greater degree of accuracy can be achieved by devoting more time or resources or both. As will be discussed later, in statistics, generalisations about a large group (known as population) are often made on the basis of small group (known as sample). It is possible to achieve this by maintaining a reasonable degree of accuracy of measurements. Therefore, it is not necessary to always have a high degree of accuracy but whatever degree of accuracy is once decided must be uniformly maintained throughout the investigation.

Notes

5. ***Statistics are collected in a systematic manner and for a predetermined purpose:*** In order that the results obtained from statistics are free from errors, it is necessary that these should be collected in a systematic manner. Haphazardly collected figures are not desirable as they may lead to wrong conclusions. Moreover, statistics should be collected for a well defined and specific objective, otherwise it might happen that the unnecessary statistics are collected while the necessary statistics are left out. Hence, a given set of numerical figures cannot be termed as statistics if it has been collected in a haphazard manner and without proper specification of the objective.

6. **Statistics should be capable of being placed in relation to each other:** This characteristic requires that the collected statistics should be comparable with reference to time or place or any other condition. In order that statistics are comparable it is essential that they are homogeneous and pertain to the same investigation. This can be achieved by collecting data in identical manner for different periods or for different places or for different conditions.

Hence, any set of numerical facts possessing the above mentioned characteristics can be termed as statistics or data.

The use of the word 'STATISTICS' in singular form refers to a science which provides methods of collection, analysis and interpretation of statistical data. Thus, statistics as a science is defined on the basis of its functions and different scholars have defined it in a different way. In order to know about various aspects of statistics, we now state some of these definitions.

1. "Statistics is the science of counting." – *A.L. Bowley*

2. "Statistics may rightly be called the science of averages." – *A.L. Bowley*

3. "Statistics is the science of measurement of social organism regarded as a whole in all its manifestations." – *A.L. Bowley*

4. "Statistics is the science of estimates and probabilities." – *Boddington*

All of the above definitions are incomplete in one sense or the other because each consider only one aspect of statistics. According to the first definition, statistics is the science of counting.

However, we know that if the population or group under investigation is large, we do not count but obtain estimates.

The *second* definition viz. statistics is the science of averages, covers only one aspect, i.e., measures of average but, besides this, there are other measures used to describe a given set of data.

The *third* definition limits the scope of statistics to social sciences only. Bowley himself realised this limitation and admitted that scope of statistics is not confined to this area only.

The *fourth* definition considers yet another aspect of statistics. Although, use of estimates and probabilities have become very popular in modern statistics but there are other techniques, as well, which are also very important.

The following definitions covers some more but not all aspects of statistics.

5. "The science of statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis or enumeration or collection of estimates."

– *W.I. King*

Notes

6. "Statistics or statistical method may be defined as collection, presentation, analysis and interpretation of numerical data." – *Croxton and Cowden*

   This is a simple and comprehensive definition of statistics which implies that statistics is a scientific method.

7. "Statistics is a science which deals with collection, classification and tabulation of numerical facts as the basis for the explanation, description and comparison of phenomena."

   – *Lovitt*

8. "Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry." – *Seligman*

The definitions given by Lovitt and Seligman are similar to the definition of Croxton and Cowden except that they regard statistics as a science while Croxton and Cowden has termed it as a scientific method.

With the development of the subject of statistics, the definitions of statistics given above have also become outdated. In the last few decades the discipline of drawing conclusions and making decisions under uncertainty has grown which is proving to be very helpful to decision makers, particularly in the field of business. Although, various definitions have been given which include this aspect of statistics also, we shall now give a definition of statistics, given by Spiegel, to reflect this new dimension of statistics.

9. "Statistics is concerned with scientific method for collecting, organising, summarising, presenting and analysing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis."

On the basis of the above definitions we can say that statistics, in singular sense, is a science which consists of various statistical methods that can be used for collection, classification, presentation and analysis of data relating to social, political, natural, economical, business or any other phenomena. The results of the analysis can be used further to draw valid conclusions and to make reasonable decisions in the face of uncertainty.

1.1.1 Statistics as a Scientific Method

We have seen above that, statistics as a non-experimental science can be used to study and analyse various problems of social sciences. It may, however, be pointed out that there may be situations even in natural sciences, where conducting of an experiment under hundred per cent controlled conditions is rather impossible. Statistics, under such conditions, finds its use in natural sciences, like physics, chemistry, etc.

In view of the uses of statistics in almost all the disciplines of natural as well as social sciences, it will be more appropriate to regard it as a scientific method rather than a science. Statistics as a scientific method can be divided into the following two categories:

1. ***Theoretical Statistics:*** Theoretical statistics can be further sub-divided into the following three categories:

   (a) *Descriptive Statistics:* All those methods which are used for the collection, classification, tabulation, diagrammatic presentation of data and the methods of calculating average, dispersion, correlation and regression, index numbers, etc., are included in descriptive statistics.

Notes

(b) *Inductive Statistics:* It includes all those methods which are used to make generalisations about a population on the basis of a sample. The techniques of forecasting are also included in inductive statistics.

(c) *Inferential Statistics:* It includes all those methods which are used to test certain hypotheses regarding characteristics of a population.

2. ***Applied Statistics:*** It consists of the application of statistical methods to practical problems. Design of sample surveys, techniques of quality control, decision-making in business, etc., are included in applied statistics.

1.1.2 Statistics as a Science or an Art

We have seen above that statistics is a science. Now we shall examine whether it is an art or not. We know that science is a body of systematized knowledge. How this knowledge is to be used for solving a problem is work of an art. In addition to this, art also helps in achieving certain objectives and to identify merits and demerits of methods that could be used. Since statistics possesses all these characteristics, it may be reasonable to say that it is also an art.

Thus, we conclude that since statistical methods are systematic and have general applications, therefore, statistics is a science. Further since the successful application of these methods depends, to a considerable degree, on the skill and experience of a statistician, therefore, statistics is an art also.

Did u know? R.A. Fisher is a notable contributor to the field of statistics. His book 'Statistical Methods for Research Workers', published in 1925, marks the beginning of the theory of modern statistics.

Self Assessment

Multiple Choice Questions

1. Who gave the following definitions of statistics?

   "Statistics is the science of counting".

   (a) Bowley (b) Boddington

   (c) King (d) Saligman

2. "Statistics is the science of estimates and probabilities".

   (a) Webster (b) Secrist

   (c) Boddington (d) Yule & Kendall

3. "The science of statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis or enumeration or collection of estimates".

   (a) Achenwall (b) Marshall

   (c) W.I. King (d) Croxton & Cowden

Notes

1.2 Importance of Statistics

It is perhaps difficult to imagine a field of knowledge which can do without statistics. To begin with, the State started the use of statistics and now it is being used by almost every branch of knowledge such as physics, chemistry, biology, sociology, geography, economics, business, etc. The use of statistics provides precision to various ideas and can also suggest possible ways of tackling a problem relating to any of the above subjects. The importance of statistics has been summarized by A.L. Bowley as, "A knowledge of statistics is like a knowledge of foreign language or of algebra. It may prove of use at any time under any circumstances."

We shall discuss briefly, the importance of statistics in the following major areas:

1. ***Importance to the State:*** We know that the subject of statistics originated for helping the ancient rulers in the assessment of their military and economic strength. Gradually its scope was enlarged to tackle other problems relating to political activities of the State. In modern era, the role of State has increased and various governments of the world also take care of the welfare of its people. Therefore, these governments require much greater information in the form of numerical figures for the fulfillment of welfare objectives in addition to the efficient running of their administration.

   In a democratic form of government, various political groups are also guided by the statistical analysis regarding their popularity in the masses. Thus, it can be said that it is impossible to think about the functioning of modern state in the absence of statistics.

2. ***Importance in economics:*** Statistics is an indispensable tool for a proper understanding of various economic problems. It also provides important guidelines for the formulation of various economic policies.

   Almost every economic problem is capable of being expressed in the form of numerical figures, e.g., the output of agriculture or of industry, volume of exports and imports, prices of commodities, income of the people, distribution of land holding, etc. In each case, the data are affected by a multiplicity of factors. Further, it can be shown that the other conditions prescribed for statistical data are also satisfied. Thus, we can say that the study of various economic problems is essentially the one of a statistical nature.

   Inductive method of generalisation, popularly used in economics, is also based on statistical principles. Various famous laws in economics such as, the law of diminishing marginal utility, the law of diminishing marginal returns, the theory of revealed preference, etc., are based on generalisations from observation of economic behaviour of a large number of individuals. Statistical methods are also useful in estimating a mathematical relation between various economic variables. For example, the data on prices and corresponding quantities demanded of a commodity can be used to estimate the mathematical form of the demand relationship between two variables. Further, the validity of a generalisation or relation between variables can also be tested by using statistical techniques.

   Statistical analysis of a given data can also be used for the precise understanding of an economic problem. For example, to study the problem of inequalities of income in a society, we can classify the relevant data and, if necessary, compute certain measures to bring the problem into focus. Using statistics, suitable policy measures can also be adopted for tackling this problem. Similarly, statistical methods can also be used to understand and to suggest a suitable solution for problems in other areas such as industry, agricultural, human resource development, international trade, etc.

   Realising the importance of statistics in economics, a separate branch of economics, known as econometrics has been developed during the recent years. The techniques of econometrics are based upon the principles of economics, statistics and mathematics.

Notes

3. ***Importance in national income accounting:*** The system of keeping the accounts of income and expenditure of a country is known as national income accounting. These accounts contain information on various macro-economic variables like national income, expenditure, production, savings, investments, volume of exports and imports, etc. The national income accounts of a country are very useful in having an idea about the broad features of its economy or of a particular region. The preparation of these accounts require data, regarding various variables, at the macro-level. Since such information is very difficult, if not impossible, to obtain, is often estimated by using techniques and principles of statistics.

4. ***Importance in planning:*** Planning is indispensable for achieving faster rate of growth through the best use of a nation's resources. It also requires a good deal of statistical data on various aspects of the economy. One of the aims of planning could be to achieve a specified rate of growth of the economy. Using statistical techniques, it is possible to assess the amounts of various resources available in the economy and accordingly determine whether the specified rate of growth is sustainable or not. The statistical analysis of data regarding an economy may reveal certain areas which might require special attention, e.g., a situation of growing unemployment or a situation of rising prices during past few years. Statistical techniques and principles can also guide the Government in adopting suitable policy measures to rectify such situations. In addition to this, these techniques can be used to assess various policies of the Government in the past. Thus, it is rather impossible to think of a situation where planning and evaluation of various policies can be done without the use of statistical techniques. In view of this it is sometimes said that, "Planning without statistics is a ship without rudder and compass". Hence statistics is an important tool for the quantification of various planning policies.

5. ***Importance in business:*** With the increase in size of business of a firm and with the uncertainties of business because of cut throat competition, the need for statistical information and statistical analysis of various business situations has increased tremendously. Prior to this, when the size of business used to be smaller without much complexities, a single person, usually owner or manager of the firm, used to take all decisions regarding its business. For example, he used to decide, from where the necessary raw materials and other factors of production were to be obtained, how much of output will be produced, where it will be sold, etc. This type of decision-making was usually based on experience and expectations of this single man and as such had no scientific basis.

The modern era is an era of mass production in which size and number of firms have increased enormously. The increase in the number of firms has resulted into cut throat competition among various firms and, consequently, the uncertainties in business have become greater than before. Under such circumstances, it has become almost impossible for a single man to take decisions regarding various aspects of a rather complex business. It is precisely this point from where the role of statistics started in business. Now a days no business, large or small, public or private, can prosper without the help of statistics. Statistics provides necessary techniques to a businessman for the formulation of various policies with regard to his business. In fact the process of collection and analysis of data becomes necessary right from the stage of launching a particular business. Some of the stages of business where statistical analysis has become necessary are briefly discussed below:

1. ***Decisions regarding business, its location and size:*** Before starting a business it is necessary to know whether it will be worth while to undertake this. This involves a detailed analysis of its costs and benefits which can be done by using techniques and principles of statistics. Furthermore, statistics can also provide certain guidelines which may prove to be helpful in deciding the possible location and size of the proposed business.

Notes

2. ***Planning of production:*** After a business is launched, the businessman has to plan its production so that he is able to meet the demand of its product and incurs minimum losses on account of over or under production. For this he has to estimate the pattern of demand of the product by conducting various market surveys. Based upon these surveys, he might also forecast the demand of the product at various points of time in future. In addition to this, the businessman has to conduct market surveys of various resources that will be used in the production of the given output. This may help him in the organisation of production with minimum costs.

3. ***Inventory control:*** Sometimes, depending upon the fluctuations in demand and supply conditions, it may not be possible to keep production in pace with demand of the product. There may be a situation of no demand resulting in over production and consequently the firm might have to discontinue production for some time. On the other hand, there may be a sudden rise in the demand of the product so that the firm is able to meet only a part of the total demand. Under such situations the firm may decide to have an inventory of the product for the smooth running of its business. The optimum limits of inventory, i.e., the minimum and maximum amount of stock to be kept, can be decided by the statistical analysis of the fluctuations in demand and supply of the product.

4. ***Quality control:*** Statistical techniques can also be used to control the quality of the product manufactured by a firm. This consists of the preparation of control charts by means of the specification of an average quality. A control chart shows two limits, the lower control limit and the upper control limit for variation in the quality of the product. The samples of output, being produced, are taken at regular intervals and their quality is measured. If the quality falls outside the control limits, steps are taken to rectify the manufacturing process.

5. ***Accounts writing and auditing:*** Every business firm keeps accounts of its revenue and expenditure. All activities of a firm, whether big or small, are reflected by these accounts. Whenever certain decisions are to be taken or it is desired to assess the performance of the firm or of its particular section or sections, these accounts are required to be summarised in a statistical way. This may consist of the calculation of typical measures like average production per unit of labour, average production per hour, average rate of return on investment, etc. Statistical methods may also be helpful in generalising relationships between two or more of such variables.

   Further, while auditing the accounts of a big business, it may not be possible to examine each and every transaction. Statistics provides sampling techniques to audit the accounts of a business firm. This can save a lot of time and money.

6. ***Banks and Insurance companies:*** Banks use statistical techniques to take decisions regarding the average amount of cash needed each day to meet the requirements of day to day transactions. Furthermore, various policies of investment and sanction of loans are also based on the analysis provided by statistics.

The business of insurance is based on the studies of life expectancy in various age groups. Depending upon these studies, mortality tables are constructed and accordingly the rates of premium to be charged by an insurance company are decided. All this involves the use of statistical principles and methods.

Did u know? The science of statistics received contributions from notable economists such as Augustin Cournot (1801 - 1877), Leon Walras (1834 - 1910), Vilfredo Pareto (1848 - 1923), Alfred Marshall (1842 - 1924), Edgeworth, A.L. Bowley, etc. They gave an applied form to the subject.

Notes

Self Assessment

State whether the following statements are true or false:

4. It is very difficult to imagine a field of knowledge which can do without statistics
5. Statistics is being used by almost every branch of knowledge such as physics, chemistry, biology, sociology, geography, economics, business, etc.
6. The importance of statistics has been summarized by L.A. Bowler as, "A knowledge of statistics is like a knowledge of foreign language or of algebra.
7. The subject of statistics originated for helping the ancient rulers in the assessment of their military and economic strength.
8. Statistics is a dispensable tool for a proper understanding of various economic problems.
9. The system of keeping the accounts of income and expenditure of a country is known as domestic income accounting.
10. The modern era is an era of mass production in which size and number of firms have decreased.

Task Conduct a debate on following statements and interpret them.

1. "Statistics can prove anything."
2. "Statistics can prove nothing."

Caution Statistics should not be used in the same way as a drunken man uses lamppost for support rather than for illumination.

1.3 Scope of Statistics

Statistics is used to present the numerical facts in a form that is easily understandable by human mind and to make comparisons, derive valid conclusions, etc., from these facts. R.W. Buges describes the functions of statistics in these words, "The fundamental gospel of statistics is to push back the domain of ignorance, prejudice, rule of thumb, arbitrary or premature decisions, tradition and dogmatism and to increase the domain in which decisions are made and principles are formulated on the basis of analysed quantitative facts."

The following are the main functions of statistics:

1. ***Presents facts in numerical figures:*** The first function of statistics is to present a given problem in terms of numerical figures. We know that the numerical presentation helps in having a better understanding of the nature a of problem. Facts expressed in words are not very useful because they are often vague and are likely to be understood differently by different people. For example, the statement that a large proportion of total work force of India is engaged in agriculture, is vague and uncertain. On the other hand, the statement that 70% of the total work force is engaged in agriculture is more specific and easier to grasp. Similarly, the statement that the annual rate of inflation in a country is 10% is more convincing than the statement that prices are rising.

Notes

2. ***Presents complex facts in a simplified form:*** Generally a problem to be investigated is represented by a large mass of numerical figures which are very difficult to understand and remember. Using various statistical methods, this large mass of data can be presented in a simplified form. This simplification is achieved by the summarisation of data so that broad features of the given problem are brought into focus. Various statistical techniques such as presentation of data in the form of diagrams, graphs, frequency distributions and calculation of average, dispersion, correlation, etc., make the given data intelligible and easily understandable.

3. ***Studies relationship between two or more phenomena:*** Statistics can be used to investigate whether two or more phenomena are related. For example, the relationship between income and consumption, demand and supply, etc., can be studied by measuring correlation between relevant variables. Furthermore, a given mathematical relation can also be fitted to the given data by using the technique of regression analysis.

4. ***Provides techniques for the comparison of phenomena:*** Many a times, the purpose of undertaking a statistical analysis is to compare various phenomena by computing one or more measures like mean, variance, ratios, percentages and various types of coefficients. For example, when we compute the consumer price index for a particular group of workers, then our aim could be to compare this index with that of previous year or to compare it with the consumer price index of a similar group of workers of some other city, etc. Similarly, the inequalities of income in various countries may be computed for the sake of their comparison.

5. ***Enlarges individual experiences:*** An important function of statistics is that it enlarges human experience in the solution of various problems. In the words of A.L. Bowley, "the proper function of statistics, indeed is to enlarge individual experience." Statistics is like a master key that is used to solve problems of mankind in every field. It would not be an exaggeration to say that many fields of knowledge would have remained closed to the mankind forever but for the efficient and useful techniques and methodology of the science of statistics.

6. ***Helps in the formulation of policies:*** Statistical analysis of data is the starting point in the formulation of policies in various economic, business and government activities. For example, using statistical techniques a firm can know the tastes and preferences of the consumers and decide to make its product accordingly. Similarly, the Government policies regarding taxation, prices, investments, unemployment, imports and exports, etc. are also guided by statistical studies in the relevant areas.

7. ***Helps in forecasting:*** The success of planning by the Government or of a business depends to a large extent upon the accuracy of their forecasts. Statistics provides a scientific basis for making such forecasts. Various techniques used for forecasting are time series analysis, regression analysis, etc.

8. ***Provides techniques for testing of hypothesis:*** A hypothesis is a statement about some characteristics of a population (or universe). For example, the statement that average height of students of a college is 66 inches, is a hypothesis. Here students of the college constitute the population. It is possible to test the validity of this statement by the use of statistical techniques.

9. ***Provides techniques for making decisions under uncertainty:*** Many a times we face an uncertain situation where any one of the many alternatives may be adopted. For example, a person may face a situation of rain or no rain and he wants to decide whether to take his umbrella or not. Similarly, a businessman might face a situation of uncertain investment opportunities in which he can lose or gain. He may be interested in knowing whether to

Notes

undertake a particular investment or not. The answer to such problems are provided by the statistical techniques of decision-making under uncertainty.

Did u know? Among the noteworthy Indian scholars who contributed to statistics are P.C. Mahalnobis, V.K.R.V. Rao, R.C. Desai and P.V. Sukhatme.

Self Assessment

State whether the following statements are true or false:

11. Honesty, intelligence, colour of eyes, beauty, etc. all are examples of qualitative characteristics.
12. Statistics deals only with groups and not with individuals.
13. Statistical results are surely true.
14. Statistics are liable to be misused.
15. Statistics must be used by anybody.

Task Elucidate the following statements:

1. "The successful businessman is one whose estimates closely approaches accuracy."
2. "All its best science is statistical."
3. "Statistical results are very general estimates rather than exact statements".

1.4 Limitations of Statistics

Like every other science, statistics also has its limitations. In order to have maximum advantage from the use of statistical methods, it is necessary to know their limitations. According to Newshome, "It (statistics) must be regarded as an instrument of research of great value, but having severe limitations, which are not possible to overcome and as such they need our careful attention." The science of statistics suffers from the following limitations:

1. ***Statistics deals with numerical facts only:*** Broadly speaking there are two types of facts, (a) quantitative and (b) qualitative facts.

   Quantitative facts are capable of being represented in the form of numerical figures and therefore, are also known as numerical facts. These facts can be analysed and interpreted with the help of statistical methods. Qualitative facts, on the other hand, represent only the qualitative characteristics like honesty, intelligence, colour of eyes, beauty, etc. and statistical methods cannot be used to study these types of characteristics. Sometimes, however, it is possible to make an indirect study of such characteristics through their conversion into numerical figures. For example, we may assign a number 0 for a male and 1 for a female, etc.

2. ***Statistics deals only with groups and not with individuals:*** Statistical studies are undertaken to study the characteristics of a group rather than individuals. These studies are done to compare the general behaviour of the group at different points of time or the behaviour of different groups at a particular point of time. For example, the economic performance of a country in a year is measured by its national income in that year and by

Notes

comparing national income of various years, one can know whether performance of the country is improving or not. Further, by comparing national income of different countries, one can know its relative position vis-a-vis other countries.

3. ***Statistical results are true only on the average:*** Statistical results give the behaviour of the group on the average and these may not hold for an individual of that very group. Thus, the statement that average wages of workers of a certain factory is ₹ 1,500 p.m. does not necessarily mean that each worker is getting this wage. In fact, some of the workers may be getting more while others less than or equal to ₹ 1,500. Further, when value of a variable is estimated by using some explanatory variable, the estimated value represents the value on the average for a particular value of the explanatory variable. In a similar way, all the laws of statistics are true only on the average.

4. ***Statistical results are only approximately true:*** Most of the statistical studies are based on a sample taken from the population. Under certain circumstances the estimated data are also used.

   Therefore, conclusions about a population based on such information are bound to be true only approximately. Further, if more observations are collected with a view to improve the accuracy of the results, these efforts are often offset by the errors of observation. In the words of Bowley, "When observations are extended, many sources of inaccuracy are found to be present, and it is very frequently impossible to remove them completely. Statistical results are, therefore, very general estimates rather than exact statements." Thus, whether statistical results are based on sample or census data, are bound to be true only approximately.

5. ***Statistical methods constitute only one set of methods to study a problem:*** A given problem can often be studied in many ways. Statistical methods are used to simplify the mass of data and obtain quantitative results by its analysis. However, one should not depend entirely on statistical results. These results must invariably be supplemented by the results of alternative methods of analysing the problem. It should be kept in mind that statistics is only a means and not an end. According to D. Gregory and H. Ward, "Statistics cannot run a business or a government. Nor can the study of statistics do more than provide a few suggestions or offer a few pointers as to firm's or government's future behaviour."

6. ***Statistics are liable to be misused:*** Statistical data are likely to be misused to draw any type of conclusion. If the attitude of the investigator is biased towards a particular aspect of the problem, he is likely to collect only such data which give more importance to that aspect. The conclusions drawn on the basis of such information are bound to be misleading. Suppose, for example, the attitude of the Government is biased and it wants to compute a price index which should show a smaller rise of prices than the actual one. In such a situation, the Government might use only those price quotations that are obtained from markets having lower prices.

7. ***Statistics must be used only by experts:*** Statistics, being a technical subject, is very difficult for a common man to understand. Only the experts of statistics can use it correctly and derive right conclusions from the analysis. In the words of Yule and Kendall, "Statistical methods are the most dangerous tools in the hands of inexperts." Hence, this is the most important limitation of statistics.

Notes

Notes

Origin and Growth of Statistics

Statistics, as a subject, has a very long history. The origin of STATISTICS is indicated by the word itself which seems to have been derived either from the Latin word 'STATUS' or from the Italian word 'STATISTA' or may be from the German word 'STATISTIK.' The meaning of all these words is 'political state'. Every State administration in the past collected and analysed data. The data regarding population gave an idea about the possible military strength and the data regarding material wealth of a country gave an idea about the possible source of finance to the State. Similarly, data were collected for other purposes also. On examining the historical records of various ancient countries, one might find that almost all the countries had a system of collection of data. In ancient Egypt, the data on population and material wealth of the country were collected as early as 3050 B.C., for the construction of pyramids. Census was conducted in Jidda in 2030 B.C. and the population was estimated to be 38,00,000. The first census of Rome was done as early as 435 B.C. After the 15th century the work of publishing the statistical data was also started but the first analysis of data on scientific basis was done by Captain John Graunt in the 17th century. His first work on social statistics, 'Observation on London Bills of Mortality' was published in 1662. During the same period the gamblers of western countries had started using statistics, because they wanted to know the more precise estimates of odds at the gambling table. This led to the development of the 'Theory of Probability'.

Ancient India also had the tradition of collection of statistical data. In ancient works, such as Manusmriti, Shukraniti, etc., we find evidences of collection of data for the purpose of running the affairs of the State where population, military force and other resources have been expressed in the form of figures. The fact and figures of the Chandragupta Mauraya's regime are described in 'Kautilya's Arthashastra'. Statistics were also in use during the Mughal period. The data were collected regarding population, military strength, revenue, land revenue, measurements of land, etc. During the British period too, statistics were used in various areas of activities.

Although the tradition of collection of data and its use for various purposes is very old, the development of modern statistics as a subject is of recent origin. The development of the subject took place mainly after sixteenth century. The notable mathematicians who contributed to the development of statistics are Galileo, Pascal, De-Mere, Farment and Cardeno of the 17th century. Then in later years the subject was developed by Abraham De Moivre (1667 - 1754), Marquis De Laplace (1749 - 1827), Karl Friedrich Gauss (1777 - 1855), Adolphe Quetelet (1796 - 1874), Francis Galton (1822 - 1911), etc. Karl Pearson (1857 - 1937), who is regarded as the father of modern statistics, was greatly motivated by the researches of Galton and was the first person to be appointed as Galton Professor in the University of London. William S. Gosset (1876 - 1937), a student of Karl Pearson, propounded a number of statistical formulae under the pen-name of 'Student'.

Self Assessment

Fill in the blanks:

16. Statistics is used to present the ..... facts in a form that is easily understandable by human mind and to make comparisons, derive valid conclusions, etc.
17. The first function of statistics is .....
18. Statistics presents ..... facts in a ..... form.

Notes

19. Statistics is like a ..... key that is used to solve problems of mankind in every field.

20. Businessman might face a situation of ..... investment opportunities in which he can lose or gain.

Case Study

Statistics as a Science different from Natural Sciences

Science is a body of systematized knowledge developed by generalisations of relations based on the study of cause and effect. These generalised relations are also called the laws of science. For example, there are laws in physics, chemistry, statistics, mathematics, etc. It is obvious from this that statistics is also a science like any other natural science. The basic difference between statistics and other natural sciences lies in the difference in conditions under which its experiments are conducted. Where as the experiments in natural sciences are done in laboratory, under more or less controlled conditions, the experiments in statistics are conducted under uncontrolled conditions. Consider, for example, the collection of data regarding expenditure of households in a locality. There may be a large number of factors affecting expenditure and some of these factors might be different for different households.

Due to these reasons, statistics is often termed as a non-experimental science while natural sciences are termed as experimental sciences. We may note here that social sciences like economics, business, sociology, geography, political science, etc., belong to the category of non-experimental science and thus, the laws and methods of statistics can be used to understand and analyse the problems of these sciences also.

Questions:

1. Analyze the case and interpret it.
2. Write down the case facts.
3. What do you conclude?

1.5 Summary

- The word 'Statistics' can be used in both 'the plural' and 'the singular' sense.
- In plural sense it implies a set of numerical figures, commonly known as statistical data.
- In singular sense statistics implies a scientific method used for the collection, analysis and interpretation of data.
- Any set of numerical figures cannot be regarded as statistics or data. A set of numerical figures collected for the investigation of a given problem can be regarded as data only if these are comparable and affected by a multiplicity of factors.
- As a scientific method, statistics is used in almost every subject of natural and social sciences.
- Statistics as a method can be divided into two broad categories viz. Theoretical Statistics and Applied Statistics.
- Theoretical statistics can further be divided into Descriptive, Inductive and Inferential statistics.

Notes

- Statistics is used to collect, present and analyze numerical figures on a scientific basis.
- The use of various statistical methods help in presenting complex mass of data in a simplified form so as to facilitate the process of comparison of characteristics in two or more situations.
- Statistics also provide important techniques for the study of relationship between two or more characteristics (or variable), in forecasting, testing of hypothesis, quality control, decision-making, etc.
- Statistics as a scientific method has its importance in almost all subjects of natural and social sciences.
- Statistics is an indispensable tool for the modern government to ensure efficient running of its administration in addition to fulfillment of welfare objectives.
- It is rather impossible to think of planning in the absence of statistics.
- The importance of statistics is also increasing in modern business world.
- Every business, whether big or small, uses statistics for analysing various business situations, including the feasibility of launching a new business.
- The limitations of statistics must always be kept in mind.
- Statistical methods are applicable only if data can be expressed in terms of numerical figures.
- The results of analysis are applicable to groups of individuals or units and are true only on average, etc.

1.6 Keywords

Applied Statistics: It consists of the application of statistical methods to practical problems. Design of sample surveys,

Descriptive Statistics: All those methods which are used for the collection, classification, tabulation, diagrammatic presentation of data and the methods of calculating average, dispersion, correlation and regression, index numbers, etc., are included in descriptive statistics.

Inductive Statistics: It includes all those methods which are used to make generalisations about a population on the basis of a sample. The techniques of forecasting are also included in inductive statistics.

Inferential Statistics: It includes all those methods which are used to test certain hypotheses regarding characteristics of a population.

National income accounting: The system of keeping the accounts of income and expenditure of a country is known as national income accounting.

Numerical facts: Quantitative facts are capable of being represented in the form of numerical figures and therefore, are also known as numerical facts.

Qualitative facts: These facts represent only the qualitative characteristics like honesty, intelligence, colour of eyes, beauty, etc.

Quantitative facts: The facts which are capable to be expressed in forms of quantity/amount are called.

Statistics : By statistics we mean aggregate of facts affected to a marked extent by a multiplicity of causes,numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.

Notes

1.7 Review Questions

1. Define the term statistics.
2. Distinguish between statistical methods and statistics.
3. Discuss the scope and significance of the study of statistics.
4. "Statistics are numerical statements of facts, but all facts stated numerically are not statistics". Clarify this statement and point out briefly which numerical statements of facts are statistics.
5. Discuss briefly the utility of statistics in economic analysis and business.
6. "Statistics are the straws out of which one like other economists have to make bricks". Discuss.
7. "Science without statistics bear no fruit, statistics without science have no roots". Explain the above statement.
8. "It is usually said that statistics is science and art both". Do you agree with this statement? Discuss the scope of statistics.
9. "Statistics is not a science, it is a scientific method". Discuss it critically and explain the scope of statistics.
10. Explain clearly the three meanings of the word 'Statistics' contained in the following statement:

    "You compute statistics from statistics by statistics".
11. "Economics and statistics are twin sisters". Discuss.
12. Discuss the nature and scope of statistics. What are the fields of investigation and research where statistical methods and techniques can be usefully employed?
13. "A knowledge of statistics is like the knowledge of foreign language or of algebra. It may prove of use at any time under any circumstances". Discuss the above statement.
14. "The science of statistics is a most useful servant but only of great value to those who understand its proper use."
15. Examine the above statement and discuss the limitations of statistics.
16. Explain the importance of statistics in economic analysis and planning.
17. "He who accepts statistics indiscriminately will often be duped unnecessarily. But he who distrusts statistics indiscriminately will often be ignorant unnecessarily. In fact, reliable statistics are the lamps that light our path on the road of knowledge." Comment.
18. "Statistical methods are most dangerous tools in the hands of the inexperts. Statistics is one of those sciences whose adept must exercise the self-restraint of an artist." Explain.
19. "Planning without statistics is a ship without rudder and compass." In the light of this statement, explain the importance of statistics as effective aid to National Planning in India.
20. Explain by giving reasons whether the following are data or not:

    (a) Arun is more intelligent than Avinash.

    (b) Arun got 75% marks in B.Sc. and Avinash got 70% marks in B.Com.

Notes

(c) Arun was born on August 25, 1974.

(d) The consumption function of a community is C = 1,000 + 0.8Y, therefore, the levels of consumption for different levels of income are:

Y	0	1000	2000	4000	6000	8000
C	1000	1800	2600	4200	5800	7400

21. Would you regard the following information as statistics? Explain by giving reasons.

(a) The height of a person is 160 cms.

(b) The height of Ram is 165 cms and of Shyam is 155 cms.

(c) Ram is taller than Shyam.

(d) Ram is taller than Shyam by 10 cms.

(e) The height of Ram is 165 cms and weight of Shyam is 55 kgs.

Answers: Self Assessment

1.	(a)	2.	(c)
3.	(c)	4.	True
5.	True	6.	False
7.	True	8.	False
9.	False	10.	False
11.	True	12.	True
13.	False	14.	True
15.	False	16.	numerical

17. to present a given problem in terms of numerical figures.

18.	complex, simplified	19.	master
20.	uncertain		

1.8 Further Readings

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Bhardwaj R.S., *Business Statistics*, Excel Books.

Garrett H.E. (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, Mc Graw Hill Book Company, New York.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hannagan T.J. (1982), *Mastering Statistics*, The Macmillan Press Ltd., Surrey.

Hooda R. P., *Statistics for Business and Economics*, Macmillan India, Delhi, 2008.

Notes

Jaeger R.M (1983), *Statistics: A Spectator Sport*, Sage Publications India Pvt. Ltd., New Delhi.

Lindgren B.W. (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Selvaraj R., Loganathan, C. *Quantitative Methods in Management.*

Sharma J.K., *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutta.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Online links

http://www.buseco.monash.edu.au/ebs/about/busstats.php

http://en.wikipedia.org/wiki/Mathematical_statistics

http://en.wikipedia.org/wiki/Business_statistics

http://cnx.org/content/m11061/latest/

www.imwright.org/WebEd/u02/we020304.htm

www.tl.nist.gov/div898/handbook/eda/section3/eda366j.htm

Notes

Unit 2: Classification of Data

CONTENTS

Objectives

Introduction

2.1 Classification

2.2 Types of Classification

2.3 Formation of A Frequency Distribution

2.3.1 Construction of a Discrete Frequency Distribution

2.3.2 Construction of a Continuous Frequency Distribution

2.3.3 Relative or Percentage Frequency Distribution

2.3.4 Cumulative Frequency Distribution

2.3.5 Frequency Density

2.4 Bivariate and Multivariate Frequency Distributions

2.5 Summary

2.6 Keywords

2.7 Review Questions

2.8 Further Readings

Objectives

After studying this unit, you will be able to:

- Define the term classification
- Discuss its objectives and requisites
- Explain the different types of classification
- Analyze the formation of frequency distribution
- Know about Frequency Density

Introduction

The collected data are a complex and unorganised mass of figures which is very difficult to analyse and interpret. Therefore, it becomes necessary to organise this so that it becomes easier to grasp its broad features. Further, in order to apply the tools of analysis and interpretation, it is essential that the data are arranged in a definite form. This task is accomplished by the process of classification and tabulation.

2.1 Classification

Classification is the process of arranging the available data into various homogeneous classes and subclasses according to some common characteristics or objective of investigation. In the words of L.R. Connor, "Classification is the process of arranging things (either actually or

Notes

notionally) in the groups or classes according to the unity of attributes that may subsist amongst a diversity of individuals." The chief characteristics of any classification are:

1. The collected data are arranged into homogeneous groups.
2. The basis of classification is the similarity of characteristics or features inherent in the collected data.
3. Classification of data signifies unity in diversity.
4. Classification of data may be actual or notional.
5. Classification of data may be according to certain measurable or non-measurable characteristics or according to some combination of both.

Objectives of Classification

The main objectives of any classification are:

1. To present a mass of data in a condensed form.
2. To highlight the points of similarity and dissimilarity.
3. To bring out the relationship between variables.
4. To highlight the effect of one variable by eliminating the effect of others.
5. To facilitate comparison.
6. To prepare data for tabulation and analysis.

Requisites of a Good Classification

A good classification must possess the following features:

1. ***Unambiguous:*** The classification should not lead to any ambiguity or confusion.
2. ***Exhaustive:*** A classification is said to be exhaustive if there is no item that cannot be allotted a class.
3. ***Mutually Exclusive:*** When a classification is mutually exclusive, each item of the data can be placed only in one of the classes.
4. ***Flexibility:*** A good classification should be capable of being adjusted according to the changed situations and conditions.
5. ***Stability:*** The principle of classification, once decided, should remain same throughout the analysis, otherwise it will not be possible to get meaningful results. In the absence of stability, the results of the same type of investigation at different time periods may not be comparable.
6. ***Suitability:*** The classification should be suitable to the objective(s) of investigation.
7. ***Homogeneity:*** A classification is said to be homogeneous if similar items are placed in a class.
8. ***Revealing:*** A classification is said to be revealing if it brings out essential features of the collected data. This can be done by selecting a suitable number of classes. Making few classes means over summarization while large number classes fail to reveal any pattern of behaviour of the variable.

Notes

Did u know? Different classes are said to be mutually exclusive if they are non-overlapping.

Self Assessment

Fill in the blanks:

1. The collected data are a complex and unorganised mass of figures which is very difficult to ..... and .....
2. In order to apply the tools of analysis and interpretation, it is essential that the data are arranged in a ..... form.
3. ..... is the process of arranging the available data into various homogeneous classes and sub-classes according to some common characteristics or objective of investigation.
4. A classification is said to be ..... if there is no item that cannot be allotted a class.
5. Different classes are said to be mutually exclusive if they are .....

2.2 Types of Classification

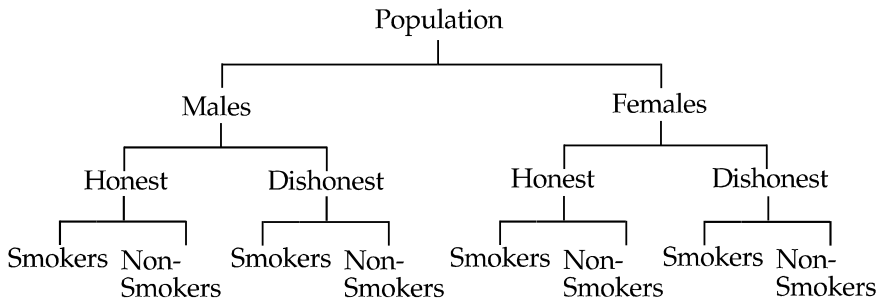
The nature of classification depends upon the purpose and objective of investigation. The following are some very common types of classification:

1. ***Geographical (or spatial) classification:*** When the data are classified according to geographical location or region, it is called a geographical classification.
2. ***Chronological classification:*** When the data are classified on the basis of its time of occurrence, it is called a chronological classification. Various time series such as; National Income figures (annual), annual output of wheat, monthly expenditure of a household, daily consumption of milk, etc., are some examples of chronological classification.
3. ***Conditional classification:*** When the data are classified according to certain conditions, other than geographical or chronological, it is called a conditional classification.
4. ***Qualitative classification or classification according to attributes:*** When the characteristics of the data are non-measurable, it is called a qualitative data. The examples of non-measurable characteristics are sex of a person, marital status, colour, honesty, intelligence, etc. These characteristics are also known as attributes. When qualitative data are given, various items can be classified into two or more groups according to a characteristic. If the data are classified only into two categories according to the presence or absence of an attribute, the classification is termed as dichotomous or twofold classification. On the other hand, if the data are classified into more than two categories according to an attribute, it is called a manifold classification. For example, classification of various students of a college according to the colour of their eyes like black, brown, grey, blue, etc. The conditional classification, given above, is also an example of a manifold classification.

   If the classification is done according to a single attribute, it is known as a one-way classification. On the other hand, the classification done according to two or more attributes is known as a two-way or multiway classification respectively. The example of a three-way classification, where population is dichotomised according to each attribute; sex, honesty and smoking habit, is given below:

Notes

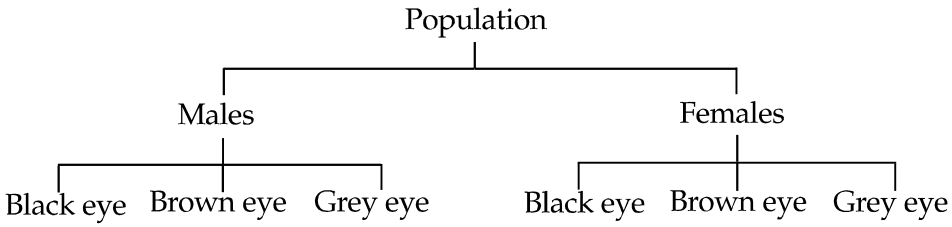
Figure 2.1: Three-way Classification



We note that there will be eight subgroups of individuals like (male, honest, smokers), (male, honest, nonsmokers), etc.

In the classification, (Figure 2.1), the population is dichotomised with respect to each of the three attributes. There may be situations where classification with respect to one attribute is dichotomous while it is manifold with respect to the other. A two way classification of this type is shown as:

Figure 2.2: Two way classification



5. ***Quantitative classification or classification according to variables:*** In case of quantitative data, the characteristic is measurable in terms of numbers and is termed as variable, e.g., weight, height, income, the number of children in a family, the number of crime cases in a city, life of an electric bulb of a company, etc. A variable can take a different value corresponding to a different item of the population or universe.

   Variables can be of two types (a) Discrete and (b) Continuous.

   (a) *Discrete Variable:* A discrete variable can assume only some specific values in a given interval. For example, the number of children in a family, the number of rooms on each floor of a multistoried building, etc.

   (b) *Continuous Variable:* A continuous variable can assume any value in a given interval. For example, monthly income of a worker can take any value, say, between ₹ 1,000 to 2,500. The income of a worker can be ₹ 1,500.25, etc.

   It must be pointed out here that, in practice, data collected on a continuous variable also look like the data of a discrete variable. This is due to the fact that measurements, done even with the finest degree of accuracy, can only be expressed in a discrete form. For example, height measured even with accuracy upto three places after decimal gives discrete values like 167.645 cms, 167.646 cms, etc. In the classification according to variables, the data are classified by the values of the variables for each item. As in the case of attributes, the classification on the basis of a single variable is termed as a one-way classification.

Notes

Similarly, there can be a two-way and multi-way classification of the data. For example, if the students of a class are classified on the basis of their marks in statistics, we get a one-way classification. However, if these students are simultaneously classified on the basis of marks in statistics and marks in economics, it becomes a two-way classification.

It should be noted here that in a two-way classification, it is possible to have simultaneous classification according to an attribute and a variable. For example, the classification of students of a class on the basis of their marks in statistics and the sex of the person.

Did u know? The life of an electric bulb is a continuous variable that can take any value from 0 to ∞.

Task Write few examples of continuous variables whose actual measurements are expressed in terms of discrete numbers.

Self Assessment

State whether the following statements are true or false:

6. When the data are classified according to geographical location or region, it is called a geographical classification.
7. When the data are classified on the basis of its time of occurrence, it is called a chronological classification.
8. When the data are classified according to certain conditions, other than geographical or chronological, it is called a unconditional classification.
9. When the characteristics of the data are non-measurable, it is called a qualitative data.
10. In a one-way classification, it is possible to have simultaneous classification according to an attribute and a variable.

2.3 Formation of a Frequency Distribution

2.3.1 Construction of a Discrete Frequency Distribution

A discrete frequency distribution may be ungrouped or grouped. In an ungrouped frequency distribution, various values of the variable are shown along with their corresponding frequencies. If this distribution fails to reveal any pattern, grouping of various observations become necessary. The resulting distribution is known as grouped frequency distribution of a discrete variable. Furthermore, a grouped frequency distribution is also constructed when the possible values that a variable can take are large.

1. ***Ungrouped Frequency Distribution of a Discrete variable:*** Suppose that a survey of 150 houses was conducted and number of rooms in each house was recorded as shown below:

5	4	4	6	3	2	2	6	6	2	6	3	3	4	5
6	3	2	2	5	3	1	4	5	1	5	1	4	3	2
5	1	5	3	2	2	4	2	2	4	4	6	3	2	4
2	3	2	4	6	3	3	2	6	4	1	4	4	5	2

Notes

4	1	4	2	1	5	1	3	3	2	5	6	1	3	1
5	3	4	3	1	1	4	1	1	2	2	1	5	2	3
6	3	5	2	2	3	3	3	3	4	5	1	6	2	1
2	1	1	6	5	2	1	1	5	6	4	2	2	3	3
3	4	3	2	1	5	2	3	1	1	4	6	4	6	2
2	4	5	6	3	6	4	1	2	4	2	2	3	4	5

Counting of frequency using Tally Marks

The method of tally marks is used to count the number of observations or the frequency of each value of the variable. Each possible value of the variable is written in a column. For every observation, a tally mark denoted by | is noted against its corresponding value. Five observations are denoted as ~~||||~~ i.e., the fifth tally mark crosses the earlier four marks and so on. The method of tally marks is used below to determine the frequencies of various values of the variable for the data given above.

Number of Rooms (X)	Tally Marks	Frequency
1	~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~	25
2	~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ||||	34
3	~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ||||	29
4	~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ |	26
5	~~||||~~ ~~||||~~ ~~||||~~ ||||	19
6	~~||||~~ ~~||||~~ ~~||||~~ ||	17
Total		150

In the above frequency distribution, the number of rooms 'X' is a discrete variable which can take integral values from 1 to 6. This distribution is also known as ungrouped frequency distribution. It should be noted here that, in case of ungrouped frequency distribution, the identity of various observations is not lost, i.e., it is possible to get back the original observations from the given frequency distribution.

2. ***Grouped Frequency Distribution of a Discrete Variable:*** Consider the data on marks obtained by 50 students in statistics. The variable 'X' denoting marks obtained is a discrete variable, let the ungrouped frequency distribution of this data be as given in the following table.

Marks	*Frequency*	*Marks*	*Frequency*	*Marks*	*Frequency*
33	1	57	1	76	1
35	2	59	1	77	2
39	1	60	2	78	1
41	2	61	1	80	1
42	1	64	1	81	1
45	1	65	3	84	1
48	2	66	2	85	2
50	1	67	1	88	1
52	1	69	2	89	1
53	1	71	1	91	1
54	1	73	2	94	2
55	2	74	2	98	1

Notes

This frequency distribution does not reveal any pattern of behaviour of the variable. In order to bring the behaviour of the variable into focus, it becomes necessary to convert this into a grouped frequency distribution.

Instead of above, if the individual marks are grouped like marks between and including 30 and 39, 40 and 49, etc. and the respective frequencies are written against them, we get a grouped frequency distribution as shown below:

Marks between and including	*Frequency*
30 - 39	4
40 - 49	6
50 - 59	8
60 - 69	12
70 - 79	9
80 - 89	7
90 - 99	4
Total	50

The above frequency distribution is more revealing than the earlier one. It is easy to understand the behaviour of marks on the basis of this distribution. It should, however, be pointed out here that the identity of observations is lost after grouping. For example, on the basis of the above distribution we can only say that 4 students have obtained marks between and including 30 - 39, etc. Thus, it is not possible to get back the original observations from a grouped frequency distribution.

2.3.2 Construction of a Continuous Frequency Distribution

As opposed to a discrete variable, a continuous variable can take any value in an interval. Measurements like height, age, income, time, etc., are some examples of a continuous variable. As mentioned earlier, when data are collected regarding these variables, it will show discreteness, which depends upon the degree of precision of the measuring instrument. Therefore, in such a situation, even if the recorded data appear to be discrete, it should be treated as continuous. Since a continuous variable can take any value in a given interval, therefore, the frequency distribution of a continuous variable is always a grouped frequency distribution.

To construct a grouped frequency distribution, the whole interval of the continuous variable, given by the difference of its largest and the smallest possible values, is divided into various mutually exclusive and exhaustive sub-intervals. These sub-intervals are termed as class intervals. Then, the frequency of each class interval is determined by counting the number of observations falling under it. The construction of such a distribution is explained below:

The figures, given below, are the 90 measurements of diameter (in mm.) of a wire.

1.86, 1.58, 1.13, 1.46, 1.53, 1.65, 1.49, 1.03, 1.10, 1.36, 1.37, 1.46, 1.44, 1.46, 1.95, 1.67, 1.59, 1.35, 1.44, 1.40, 1.50, 1.41, 1.19, 1.16, 1.27, 1.21, 1.82, 1.55, 1.52, 1.42, 1.17, 1.62, 1.42, 1.22, 1.56, 1.78, 1.98, 1.31, 1.29, 1.69, 1.32, 1.68, 1.36, 1.55, 1.54, 1.67, 1.81, 1.47, 1.30, 1.33, 1.38, 1.34, 1.40, 1.37, 1.27, 1.04, 1.87, 1.45, 1.47, 1.35, 1.24, 1.48, 1.41, 1.39, 1.38, 1.47, 1.73, 1.20, 1.77, 1.25, 1.62, 1.43, 1.51, 1.60, 1.15, 1.26, 1.76, 1.66, 1.12, 1.70, 1.57, 1.75, 1.28, 1.56, 1.42, 1.09, 1.07, 1.57, 1.92, 1.48.

The following decisions are required to be taken in the construction of any frequency distribution of a continuous variable.

Notes

1. ***Number of Class Intervals:*** Though there is no hard and fast rule regarding the number of classes to be formed, yet their number should be neither very large nor very small. If there are too many classes, the frequency distribution appears to be too fragmented to reveal the pattern of behaviour of characteristics. Fewer classes imply that the width of the class intervals will be broad and accordingly it would include a large number of observations. As will be obvious later that in any statistical analysis, the value of a class is represented by its mid-value and hence, a class interval with broader width will be representative of a large number of observations. Thus, the magnitude of loss of information due to grouping will be large when there are small number of classes. On the other hand, if the number of observations is small or the distribution of observations is irregular, i.e., not uniform, having more number of classes might result in zero or very small frequencies of some classes, thus, revealing no pattern of behaviour. Therefore, the number of classes depends upon the nature and the number of observations. If the number of observations is large or the distribution of observations is regular, one may have more number of classes. In practice, the minimum number of classes should not be less than 5 or 6 and in any case there should not be more than 20 classes.

   The approximate number of classes can also be determined by Struge's formula: $n = 1 + 3.322 \times \log 10N$, where n (rounded to the next whole number) denotes the number of classes and N denotes the total number of observations.

   Based on this formula, we have

   $n = 1 + 3.322 \times 2.000 = 7.644$ or 8, when N = 100

   $n = 1 + 3.322 \times 2.699 = 9.966$ or 10, when N = 500

   $n = 1 + 3.322 \times 4.000 = 14.288$ or 15, when N = 10,000

   $n = 1 + 3.322 \times 4.699 = 16.610$ or 17, when N = 50,000

   From the above calculations we may note that even the formation of 20 class intervals is very rarely needed.

   For the given data on the measurement of diameter, there are 90 observations. The number of classes by the Sturge's formula are

   $n = 1 + 3.322 \times \log 1090 = 7.492$ or 8

2. ***Width of a Class Interval:*** After determining the number of class intervals, one has to determine their width. The problem of determining the width of a class interval is closely related to the number of class intervals.

   As far as possible, all the class intervals should be of equal width. However, there can be situations where it may not be possible to have equal width of all the classes. Suppose that there is a frequency distribution, having all classes of equal width, in which the pattern of behaviour of the observations is not regular, i.e., there are nil or very few observations in some classes while there is concentration of observations in other classes. In such a situation, one may be compelled to have unequal class intervals in order that the frequency distribution becomes regular.

   The approximate size of a class interval can be decided by the use of the following formula:

$$\text{Class Interval} = \frac{\text{Largest observation} - \text{Smallest observation}}{\text{Number of class intervals}}$$

   or using notations, $\text{Class Interval} = \dfrac{L - S}{n}$

Notes

In the example, given above, L = 1.98 and S = 1.03 and n = 8.

∴ Approximate size of a class interval

$$= \frac{1.98 - 1.03}{8} = 0.1188 \text{ or } 0.12 \text{ (approx.)}$$

Before taking a final decision on the width of various class intervals, it is worthwhile to consider the following points:

(a) Normally a class interval should be a multiple of 5, because it is easy to grasp numbers like 5, 10, 15, ..., etc.

(b) It should be convenient to find the mid-value of a class interval.

(c) Most of the observations in a class should be uniformly distributed or concentrated around its mid-value.

(d) As far as possible, all the classes should be of equal width. A frequency distribution of equal class width is convenient to be represented diagrammatically and easy to analyse.

On the basis of above considerations, it will be more appropriate to have classes, each, with interval of 0.10 rather than 0.12. Further, the number of classes should also be revised in the light of this decision.

$$n = \frac{L - S}{\text{Class Interval}} = \frac{1.98 - 1.03}{0.10} = \frac{0.95}{0.10} = 9.5 \text{ or } 10$$

(rounded to the next whole number)

3. ***Designation of Class Limits:*** The class limits are the smallest and the largest observation in a class. These are respectively known as the lower limit and the upper limit of a class. For a frequency distribution, it is necessary to designate these class limits very unambiguously, because the mid-value of a class is obtained by using these limits. As will be obvious later, this mid-value will be used in all the computations about a frequency distribution and the accuracy of these computations will depend upon the proper specification of class limits. The class limits should be designated keeping the following points in mind:

   (a) It is not necessary to have lower limit of the first class exactly equal to the smallest observation of the data. In fact it can be less than or equal to the smallest observation. Similarly, the upper limit of the last class can be equal to or greater than the largest observation of the data.

   (b) It is convenient to have lower limit of a class either equal to zero or some multiple of 5.

   (c) The chosen class limits should be such that the observations in a class tend to concentrate around its mid-value. This will be true if the observations are uniformly distributed in a class.

The designation of class limits for various class intervals can be done in two ways: (1) Exclusive Method and (2) Inclusive Method.

1. ***Exclusive Method:*** In this method the upper limit of a class is taken to be equal to the lower limit of the following class. To keep various class intervals as mutually exclusive, the observations with magnitude greater than or equal to lower limit but less than the upper limit of a class are included in it. For example, if the lower limit of a class is 10 and its

Notes

upper limit is 20, then this class, written as 10-20, includes all the observations which are greater than or equal to 10 but less than 20. The observations with magnitude 20 will be included in the next class.

2. ***Inclusive Method:*** Here all observations with magnitude greater than or equal to the lower limit and less than or equal to the upper limit of a class are included in it.

The two types of class intervals, discussed above, are constructed for the data on the measurements of diameter of a wire as shown below:

Class Intervals	20 - 29	30 - 39	40 - 49	50 - 59	Total
Frequency	8	15	10	7	40

Mid-Value of a Class

In exclusive types of class intervals, the mid-value of a class is defined as the arithmetic mean of its lower and upper limits. However, in the case of inclusive types of class intervals, there is a gap between the upper limit of a class and the lower limit of the following class which is eliminated by determining the class boundaries. Here, the mid-value of a class is defined as the arithmetic mean of its lower and upper boundaries. To find class boundaries, we note that the given data on the measurements of diameter of a wire is expressed in terms of millimeters, approximated upto two places after decimal. This implies that a value greater than or equal to 1.095 but less than 1.10 is approximated as 1.10 and, thus, included in the class interval 1.10 - 1.19.

Similarly, an observation less than 1.095 but greater than 1.09 is approximated as 1.09 and is included in the interval 1.00-1.09. Keeping the precision of measurements in mind, various class boundaries, for the inclusive class intervals, given above, can be obtained by subtracting 0.005 from the lower limit and adding 0.005 to the upper limit of each class. These boundaries are given in the third column of the above table.

Construction of a Grouped Frequency Distribution for the Data on the Measurements of Diameter of a Wire

Taking class intervals as 1.00 - 1.10, 1.10 - 1.20, etc. and counting their respective frequencies, by the method of tally marks, we get the required frequency distribution as given below:

Class Intervals	Tally Marks	Frequency
1.00 - 1.10	IIII	4
1.10 - 1.20	~~IIII~~ II	7
1.20 - 1.30	~~IIII~~ ~~IIII~~	10
1.30 - 1.40	~~IIII~~ ~~IIII~~ IIII	14
1.40 - 1.50	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~	20
1.50 - 1.60	~~IIII~~ ~~IIII~~ III	13
1.60 - 1.70	~~IIII~~ IIII	9
1.70 - 1.80	~~IIII~~ I	6
1.80 - 1.90	IIII	4
1.90 - 2.00	III	3
Total		90

Notes

Example: Given below are the weights (in pounds) of 70 students.

1. Construct a frequency distribution when class intervals are inclusive, taking the lowest class as 60-69. Also construct class boundaries.

2. Construct a frequency distribution when class intervals are exclusive, taking the lowest class as 60-70.

61, 80, 91, 113, 100, 106, 109, 73, 88, 92, 101, 106, 107, 97, 93, 96, 102, 114, 87, 62, 74, 107, 109, 91, 72, 89, 94, 98, 112, 103, 101, 77, 92, 73, 67, 76, 84, 90, 118, 107, 108, 82, 78, 84, 77, 95, 111, 115, 104, 69, 106, 105, 63, 76, 85, 88, 96, 90, 95, 99, 83, 98, 88, 72, 75, 86, 82, 86, 93, 92.

Solution:

1. Construction of frequency distribution using inclusive class intervals.

Class Intervals	Tally Marks	Frequency	Class Boundaries
60 - 69	卌	5	59.5 - 69.5
70 - 79	卌 卌 |	11	69.5 - 79.5
80 - 89	卌 卌 ||||	14	79.5 - 89.5
90 - 99	卌 卌 卌 |||	18	89.5 - 99.5
100 - 109	卌 卌 卌 |	16	99.5 -109.5
110 - 119	卌 |	6	109.5 -119.5
Total		70	

To determine the class boundaries, we note that measured weights are approximated to the nearest pound. Therefore, a measurement less than 69.5 is approximated as 69 and included in the class interval 60 - 69. Similarly, a measurement greater than or equal to 69.5 is approximated as 70 and is included in the class interval 70 - 79. Thus, the class boundaries are obtained by subtracting 0.5 from the lower limit and adding 0.5 to the upper limit of various classes. These boundaries are shown in the last column of the above table.

2. The frequency distribution of exclusive type of class intervals can be directly written from the above table as shown below:

Class Inervals	*Frequency*
60 - 70	5
70 - 80	11
80 - 90	14
90 - 100	18
100 - 110	16
110 - 120	6
Total	70

Example: Determine the class boundaries for the following distribution of ages of 40 workers of a factory, where quoted age is the age completed on last birthday.

Class Intervals	20 - 29	30 - 39	40 - 49	50 - 59	Total
Frequency	8	15	10	7	40

Notes

Solution: The determination of class boundaries depends upon the nature of approximation. Since the quoted age is the age completed on last birthday, therefore, a number greater than 29 but less than 30 is approximated as 29. Therefore, the boundaries of this class will be 20 - 30. Similarly, the boundaries of other classes will be 30 - 40, 40 - 50 and 50 - 60, respectively.

2.3.3 Relative or Percentage Frequency Distribution

If instead of frequencies of various classes their relative or percentage frequencies are written, we get a relative or percentage frequency distribution.

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{\text{Total Frequency}}$$

$$\text{Percentage frequency of a class} = \text{Relative frequency} \times 100$$

These frequencies are shown in the following table.

Class Intervals	*Frequency*	*Relative Frequency*	*Percentage Frequency*
1.00 - 1.10	4	0.044	4.4
1.10 - 1.20	7	0.079	7.9
1.20 - 1.30	10	0.111	11.1
1.30 - 1.40	14	0.156	15.6
1.40 - 1.50	20	0.222	22.2
1.50 - 1.60	13	0.144	14.4
1.60 - 1.70	9	0.100	10.0
1.70 - 1.80	6	0.067	6.7
1.80 - 1.90	4	0.044	4.4
1.90 - 2.00	3	0.033	3.3
Total	90	1.000	100.0

2.3.4 Cumulative Frequency Distribution

The total frequency of all classes less than the upper class boundary of a given class is called the cumulative frequency of that class. "A table showing the cumulative frequencies is called a cumulative frequency distribution".

In order to answer the questions like; the measurements on diameter that are less than 1.70 or the number of measurements that are greater than 1.30, etc., a cumulative frequency distribution is constructed.

Table I

Diameters	*Cumulative Frequency*
less than 1.10	4
less than 1.20	11
less than 1.30	21
less than 1.40	35
less than 1.50	55
less than 1.60	68
less than 1.70	77
less than 1.80	83
less than 1.90	87
less than 2.00	90

Table II

Diameters	*Cumulative Frequency*
More than 1.00	90
More than 1.10	86
More than 1.20	79
More than 1.30	69
More than 1.40	55
More than 1.50	35
More than 1.60	22
More than 1.70	13
More than 1.80	7
More than 1.90	3

Notes

There are two types of cumulative frequency distributions.

Less than cumulative frequency distribution: It is obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulate is started from the lowest to the highest size.

More than cumulative frequency distribution: It is obtained by finding the cumulate total of frequencies starting from the highest to the lowest class

These frequency distributions, for the data on the measurements of diameter of a wire, are shown in Table I and Table II respectively.

Converting a cumulative frequency distribution table into a frequency distribution

The above cumulative frequency distribution table can be presented in a frequency distribution as follows**:**

Class Intervals	Frequency
1.00 - 1.10	4
1.10 - 1.20	7
1.20 - 1.30	10
1.30 - 1.40	14
1.40 - 1.50	20
1.50 - 1.60	13
1.60 - 1.70	9
1.70 - 1.80	6
1.80 - 1.90	4
1.90 - 2.00	3
Total	90

2.3.5 Frequency Density

Frequency density in a class is defined as the number of observations per unit of its width. Frequency density gives the rate of concentration of observations in a class:

$$\text{Frequency Density} = \frac{\text{Frequency of the class}}{\text{Width of the class}}$$

Table showing Frequency Density of Various Classes:

Class Intervals	Frequency	Frequency Density
1.00 - 1.10	4	40
1.10 - 1.20	7	70
1.20 - 1.30	10	100
1.30 - 1.40	14	140
1.40 - 1.50	20	200
1.50 - 1.60	13	130
1.60 - 1.70	9	90
1.70 - 1.80	6	60
1.80 - 1.90	4	40
1.90 - 2.00	3	30
Total	90	

Notes

Task Administer a test in your class. Classify the data so obtained, in the form of a frequency distribution.

Self Assessment

Multiple Choice Questions:

11. A ..... frequency distribution may be ungrouped or grouped.
    (a) Continuous Frequency Distribution
    (b) Discrete Frequency Distribution
    (c) Relative Frequency Distribution
    (d) Percentage Frequency Distribution
12. In an ..... frequency distribution, various values of the variable are shown along with their corresponding frequencies.
    (a) Grouped (b) Ungrouped
    (c) Continuous (d) Relative
13. The method of .....is used to count the number of observations or the frequency of each value of the variable.
    (a) Grouping (b) Arranging
    (c) Tally marks (d) Graphing
14. The ..... are the smallest and the largest observation in a class.
    (a) Class size (b) Class Interval
    (c) Class limits (d) True class limits
15. The ..... of a class is defined as the arithmetic mean of its lower and upper limits.
    (a) Mid value (b) Middle values
    (c) Extreme values (d) Medium Values

2.4 Bivariate and Multivariate Frequency Distributions

Bivariate Frequency Distributions

In the frequency distributions, discussed so far, the data are classified according to only one characteristic. These distributions are known as univariate frequency distributions. There may be a situation where it is necessary to classify data, simultaneously, according to two characteristics. A frequency distribution obtained by the simultaneous classification of data according to two characteristics, is known as a bivariate frequency distribution. An example of such a classification is given below, where 100 couples are classified according to the two characteristics, Age of Husband and Age of Wife. The tabular representation of the bivariate frequency distribution is known as a contingency table.

Notes

Classification according to Age of Husband and Age of Wife

Age of Husband ↓ / Age of wife →	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	Total
10 - 20	6	3	0	0	0	9
20 - 30	3	16	10	0	0	29
30 - 40	0	10	15	7	0	32
40 - 50	0	0	7	10	4	21
50 - 60	0	0	0	4	5	9
Total	9	29	32	21	9	100

It should be noted that in a bivariate classification either or both the variable can be discrete or continuous. Further, there may be a situation in which one characteristic is a variable and the other is an attribute.

Multivariate Frequency Distribution

If the classification is done, simultaneously, according to more than two characteristics, the resulting frequency distribution is known as a multivariate frequency distribution.

Example: Find the lower and upper limits of the classes when their mid-values are given as 15, 25, 35, 45, 55, 65, 75, 85 and 95.

Solution: Note that the difference between two successive mid-values is same, i.e., 10. Half of this difference is subtracted and added to the mid value of a class in order to get lower limit and the upper limit respectively. Hence, the required class intervals are 10 - 20, 20 - 30, 30 - 40, 40 - 50, 50 - 60, 60 - 70, 70 - 80, 80 - 90, 90 - 100.

Example: Find the lower and upper limits of the classes if their mid-values are 10, 20, 35, 55, 85.

Solution: Here the difference of two successive mid-values are different. In order to find the limits of the first class, half of the difference between the second and first mid-value is subtracted and added. Therefore, the first class limits are 5 - 15. The lower limit of second class is taken as equal to upper limit of the first class.

The upper limit of a class = lower limit + width,

where width = 2(Mid-value - lower limit).

$\therefore$ The upper limit of the second class = 15 + 2(20 - 15) = 25.

Thus, second class interval will be 15 - 25. Similarly, we can find the limits of third, fourth and fifth classes as 25 - 45, 45 - 65 and 65 - 105, respectively.

Notes

Notes **Distrust of Statistics**

After the study of functions and importance of statistics, one may find that statistics is indeed a very useful science. However, there are chances of its being misused by the biased or the ignorant people. The use of statistics by such people is likely to give a misleading interpretation of a given problem. This has created a feeling of distrust about statistics amongst the common people. This feeling of distrust has been expressed by different scholars as:

1. "There are three kinds of lies : lies, damned lies and statistics."
2. "Statistics can prove anything."
3. "Statistics is a rainbow of lies."
4. "Statistics are like clay of which you can make God or Devil, as you please."

All these statements reflect a feeling of distrust of statistics.

Caution Statistics is only a tool which could be used in right or wrong ways, as is obvious from the statement that, "figures won't lie but liars figure".

Task Suggest suitable measures to restore trust in statistics.

Self Assessment

State whether the following statements are true or false:

16. When data are classified according to only one characteristic, distributions are known as univariate frequency distributions.
17. A frequency distribution obtained by the simultaneous classification of data according to two characteristics, is known as a bivariate frequency distribution.
18. In a bivariate classification either or both the variable can be discrete or continuous.

Case Study

Statistical Series

The classified data when arranged in some logical order, e.g., according to the size, according to the time of occurrence or according to some other measurable or non-measurable characteristics, is known as Statistical Series.

H. Secrist defined a statistical series as, "A series, as used statistically, may be defined as things or attributes of things arranged according to some logical order." Another definition given by L. R. Connor as, "If the two variable quantities can be arranged side by side so that the measurable differences in the one correspond to the measurable differences in the other, the result is said to form a statistical series."

Contd...

Notes

A statistical series can be one of the following four types: (i) Spatial Series, (ii) Conditional Series, (iii) Time Series and (iv) Qualitative or Quantitative Series

The series formed by the geographical or spatial classification is termed as spatial series. Similarly, a series formed by the conditional classification is known as the conditional series. The examples of such series are already given under their respective classification category.

Time Series

A time series is the result of chronological classification of data. In this case, various figures are arranged with reference to the time of their occurrence. For example, the data on exports of India in various years is a time series.

Year	:	1980	1981	1982	1983	1984	1985	1986	1987	1988
Exports (in ₹ cr.)	:	6591	7242	8309	8810	9981	10427	11490	15741	20295

Qualitative or Quantitative Series

This type of series is obtained when the classification of data is done on the basis of qualitative or quantitative characteristics. Accordingly, we can have two types of series, namely, qualitative and quantitative series.

Qualitative Series:

In case of qualitative series, the number of items in each group are shown against that group. These groups are either expressed in ascending order or in descending order of the number of items in each group. The example of such a series is given below.

Distribution of Students of a College according to Sex

Sex	*Males*	Females	*Total*
No. of Students	1700	500	2200

Quantitative Series

In case of quantitative series, the number of items possessing a particular value are shown against that value.

A quantitative series can be of two types (1) Individual Series, and (2) Frequency distribution.

1. ***Individual series:*** In an individual series, the names of the individuals are written against their corresponding values. For example, the list of employees of a firm and their respective salary in a particular month.
2. ***Frequency Distribution:*** A table in which the frequencies and the associated values of a variable are written side by side, is known as a frequency distribution. According to Croxton and Cowden, "Frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude, either individually or in a group with their corresponding frequencies side by side." A frequency distribution can be discrete or continuous depending upon whether the variable is discrete or continuous.

Questions :

1. What is meant by statistical series?
2. Differentiate between quantitative and qualitative series with an example of each.
3. Explore your knowledge for different types of statistical series.

Notes

2.5 Summary

- Classification of data on the basis of one, two or more factors is termed as a one-way, two-way or multi-way classification, respectively.
- Classified data, when arranged in some logical order such as, according to size or according to time of occurrence or according to some other criterion, is known as statistical series.
- A statistical series, in which data are arranged according to magnitude of one or more characteristics, is known as a frequency distribution.
- Data classified according to the magnitude of only one characteristic is known as uni-variate frequency distribution.
- Data classified, simultaneously, according to the magnitude of two or more characteristics are known as bivariate or multivariate frequency distributions respectively.
- When a characteristics is an attribute, the data can be classified into two or more classes according to this attribute, known as dichotomous or manifold classification respectively.
- When the data are simultaneously classified according to two or more attributes, the classifications are two-way or multi-way respectively.
- It is also possible to have a two-way or multi-way classification in which one or more characteristics are variables while others are attributes.
- The process of classification is facilitated by writing the classified data in tabular form.
- Using tables, it is possible to write huge mass of data in a concise form. Further, it helps to highlight essential features of the data and make it fit for further analysis.

2.6 Keywords

Bivariate frequency distributions: Data classified, simultaneously, according to the magnitude of two characteristics are known as bivariate frequency distributions

Classification: Classification is the process of arranging things (either actually or notionally) in the groups or classes according to the unity of attributes that may subsist amongst a diversity of individuals.

Dichotomous classification: When a characteristic is an attribute, the data can be classified into two classes according to this attribute, known as dichotomous classification.

Frequency distribution: A statistical series, in which data are arranged according to magnitude of one or more characteristics, is known as a frequency distribution.

Manifold classification: When a characteristics is an attribute, the data can be classified into two or more classes according to this attribute, known as dichotomous or manifold classification respectively.

***Multivariate frequency distributions*:** Data classified, simultaneously, according to the magnitude of more than two characteristics are known as multivariate frequency distributions.

***Statistical series*:** Classified data, when arranged in some logical order such as, according to size or according to time of occurrence or according to some other criterion, is known as statistical series.

Uni-variate frequency distribution: Data classified according to the magnitude of only one characteristic is known as uni-variate frequency distribution.

Notes

2.7 Review Questions

1. What do you mean by Classification and Tabulation? Explain their importance in statistical studies.
2. What are the different factors that should be kept in mind while classifying data?
3. Distinguish between classification and tabulation. Discuss the purpose and methods of classification.
4. What are objects of classification of data? Discuss different methods of classification.
5. Discuss the purpose, methods and importance of tabulation in any statistical investigation. Mention the types of tables generally used.
6. Distinguish between an ungrouped and a grouped frequency distribution. What are the points that should be taken into consideration while determining the following:
   (a) Number of Groups
   (b) Magnitude of Class-Intervals
   (c) Class Limits.
7. Twenty students of a class appeared in an examination. Their marks out of 50 are as under:

   5, 6, 17, 17, 20, 21, 22, 22, 22, 25, 25, 26, 26, 30, 31, 31, 34, 35, 42, 48.

   Prepare a classified table by taking class intervals of 10 each, according to exclusive and inclusive methods.
8. Construct a frequency table for the following data by taking width of each class as 10. Use inclusive method of classification.

   30, 38, 43, 59, 82, 40, 45, 39, 83, 85, 72, 66, 45, 33, 53, 67, 70, 72, 52, 50, 43, 44, 60, 89, 67, 66, 78, 32, 56, 47, 65, 56, 38, 84, 64, 52, 43, 33, 31, 35, 38, 39, 40, 37, 52, 53, 60.

   If these figures represent the age of persons approximated to the nearest whole number, construct the class boundaries.
9. The number of children in 50 families of a locality are given below. Construct an appropriate discrete frequency distribution.

   2, 2, 2, 3, 2, 4, 5, 4, 6, 8, 3, 3, 1, 4, 3, 1, 3, 3, 2, 1, 3, 3, 2, 4, 3, 5, 4, 3, 3, 2, 2, 5, 2, 5, 3, 3, 3, 4, 3, 5, 4, 4, 2, 6, 3, 6, 3, 3, 7, 3.
10. Construct a frequency distribution of the marks obtained by 50 students in economics as given below:

    42, 53, 65, 63, 61, 47, 58, 60, 64, 45, 55, 57, 82, 42, 39, 51, 65, 55, 33, 70, 50, 52, 53, 45, 45, 25, 36, 59, 63, 39, 65, 30, 45, 35, 49, 15, 54, 48, 64, 26, 75, 20, 42, 40, 41, 55, 52, 46, 35, 18.

    (Take the first class interval as 10 - 20)
11. The following figures give the ages, in years, of newly married husbands and their wives. Represent the data by an appropriate frequency distribution.

Age of Husband :	24	26	27	25	28	24	27	28	25	26
Age of Wife :	17	18	19	17	10	18	18	19	18	19
Age of Husband :	25	26	27	25	27	26	25	26	26	26
Age of Wife :	17	18	19	19	20	19	17	20	17	18

Notes

12. "He who accepts statistics indiscriminately will often be duped unnecessarily. But he who distrusts statistics indiscriminately will often be ignorant unnecessarily. In fact, reliable statistics are the lamps that light our path on the road of knowledge." Comment.

Answers: Self Assessment

1.	Analyze, interpret	2.	definite
3.	Classification	4.	exhaustive
5.	Non-overlapping	6.	True
7.	True	8.	False
9.	True	10.	False
11.	(b)	12.	(b)
13.	(c)	14.	(c)
15.	(a)	16.	True
17.	True	18.	True

2.8 Further Readings

Books

Bhardwaj R. S., *Business Statistics*, Excel Books.

Balwani Nitin, *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Garrett H.E. (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, Mc Graw Hill Book Company, New York.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hannagan T.J. (1982), *Mastering Statistics*, The Macmillan Press Ltd., Surrey.

Hooda R.P., *Statistics for Business and Economics*, Macmillan India Delhi, 2008

Jaeger R.M (1983), *Statistics: A Spectator Sport*, Sage Publications India Pvt. Ltd., New Delhi.

Lindgren B.W (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Selvaraj R., Loganathan C., *Quantitative Methods in Management*.

Sharma J.K., *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutta.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Notes

Online links

http://cnx.org/content/m11061/latest

http://en.wikipedia.org/wiki/Business_statistics

http://en.wikipedia.org/wiki/Mathematical_statistics

http://www.mathsisfun.com/data

Notes

Unit 3: Tabulation

CONTENTS

Objectives

Introduction

3.1 Objectives of Tabulation

3.1.1 Difference between Classification and Tabulation

3.1.2 Main Parts of a Table

3.2 Type of Tables

3.3 Methods of Tabulation

3.4 Summary

3.5 Keywords

3.6 Review Questions

3.7 Further Readings

Objectives

After studying this unit, you will be able to:

- Define the term tabulation
- Discuss the objectives of tabulation
- Make a difference between classification and tabulation
- List the main parts of a table
- Describe various types of table
- Focus on various methods of tabulation

Introduction

Tabulation is a systematic presentation of numerical data in rows and columns. Tabulation of classified data make it more intelligible and fit for statistical analysis. According to Tuttle, " A statistical table is the logical listing of related quantitative data in vertical columns and horizontal rows of numbers, with sufficient explanatory and qualifying words, phrases and statements in the form of titles, headings and footnotes to make clear the full meaning of the data and their origin." The classified data presented in tabular form helps to bring out their essential features.

3.1 Objectives of Tabulation

The main objectives of tabulation are:

1. To simplify complex data.
2. To highlight chief characteristics of the data.
3. To clarify objective of investigation.

Notes

4. To present data in a minimum space.
5. To detect errors and omissions in the data.
6. To facilitate comparison of data.
7. To facilitate reference.
8. To identify trend and tendencies of the given data.
9. To facilitate statistical analysis.

3.1.1 Difference between Classification and Tabulation

The basic points of difference between classification and tabulation, inspite of the fact that these are closely related, are as given below:

1. Classification of data is basis for tabulation because first the data are classified and then tabulated.
2. By classification the data are divided into various groups and subgroups on the basis of their similarities and dissimilarities while tabulation is a process of arranging the classified data in rows and columns with suitable heads and subheads.

3.1.2 Main Parts of a Table

The main parts of a table are as given below:

1. ***Table Number:*** This number is helpful in the identification of a table. This is often indicated at the top of the table.
2. ***Title:*** Each table should have a title to indicate the scope, nature of contents of the table in an unambiguous and concise form.
3. ***Captions and stubs:*** A table is made up of rows and columns. Headings or subheadings used to designate columns are called captions while those used to designate rows are called stubs. A caption or a stub should be self explanatory. A provision of totals of each row or column should always be made in every table by providing an additional column or row respectively.
4. ***Main Body of the Table:*** This is the most important part of the table as it contains numerical information. The size and shape of the main body should be planned in view of the nature of figures and the objective of investigation. The arrangement of numerical data in main body is done from top to bottom in columns and from left to right in rows.
5. ***Ruling and Spacing:*** Proper ruling and spacing is very important in the construction of a table. Vertical lines are drawn to separate various columns with the exception of sides of a table. Horizontal lines are normally not drawn in the body of a table, however, the totals are always separated from the main body by horizontal lines. Further, the horizontal lines are drawn at the top and the bottom of a table.

   Spacing of various horizontal and vertical lines should be done depending on the available space. Major and minor items should be given space according to their relative importance.
6. ***Head-note:*** A head-note is often given below the title of a table to indicate the units of measurement of the data. This is often enclosed in brackets.
7. ***Foot note:*** Abbreviations, if any, used in the table or some other explanatory notes are given just below the last horizontal line in the form of footnotes.

Notes

8. ***Source-Note:*** This note is often required when secondary data are being tabulated. This note indicates the source from where the information has been obtained. Source note is also given as a footnote.

The main parts of a table can also be understood by looking at its broad structure given below:

Structure of a table

Table No : .....

Title : .....

Stub Heading	Captions		Captions			Total
	Captions	Captions	Captions	Captions	Captions	
↑ Stub Enteries ↓	M A I N B O D Y					
Total						

Foot Note:

Source:

Rules for Tabulation

The rules for tabulation of data can be divided into two broad categories: (i) Rules regarding structure of a table, explained above, and (ii) General rules.

General Rules

1. The table should be simple and compact which is not overloaded with details.
2. Tabulation should be in accordance with the objective of investigation.
3. The unit of measurements must always be indicated in the table.
4. The captions and stubs must be arranged in a systematic manner so that it is easy to grasp the table.
5. A table should be complete and self explanatory.
6. As far as possible the interpretative figures like totals, ratios and percentages must also be provided in a table.
7. The entries in a table should be accurate.
8. Table should be attractive to draw the attention of readers.

Did u know? Classification is a process of statistical analysis while tabulation is a process of presentation.

Notes

Task Give suitable example to differentiate amongst head note, footnote and source note.

Self Assessment

Fill in the Blanks:

1. ..... is a systematic presentation of numerical data in rows and columns.
2. According to....., "A statistical table is the logical listing of related quantitative data in vertical columns and horizontal rows of numbers, with sufficient explanatory and qualifying words, phrases and statements in the form of titles, headings and footnotes to make clear the full meaning of the data and their origin."
3. Classification is a process of statistical ..... while tabulation is a process of .....
4. Headings or subheadings used to designate columns are called ..... while those used to designate rows are called .....
5. A .....is often given below the title of a table to indicate the units of measurement of the data.
6. Horizontal lines are normally ..... in the body of a table.
7. The totals are always separated from the main body by ..... lines.
8. Tabulation should be in accordance with the objective of .....
9. A table should be complete and .....
10. .....indicates the source from where the information has been obtained

3.2 Type of Tables

Statistical tables can be classified into various categories depending upon the basis of their classification. Broadly speaking, the basis of classification can be any of the following:

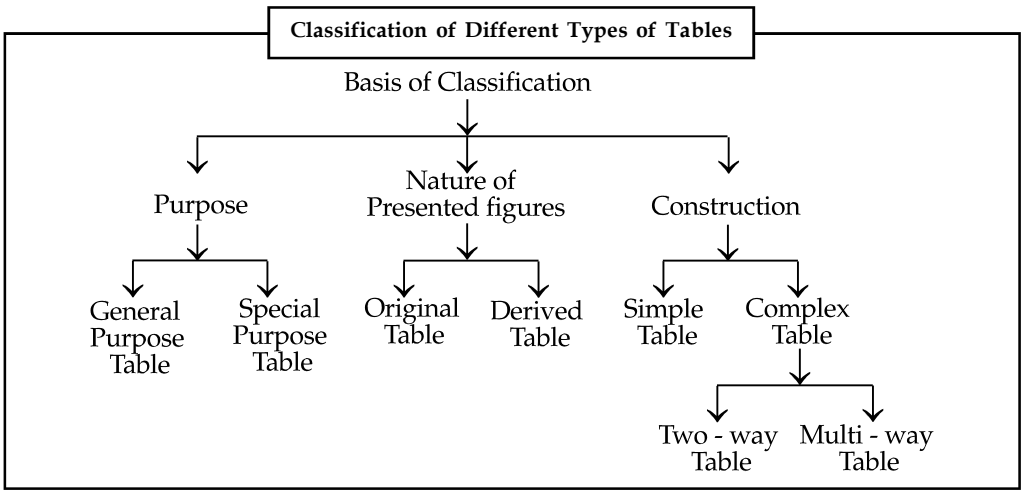
1. Purpose of investigation
2. Nature of presented figures
3. Construction

Different types of tables, thus, obtained are shown in the following chart.

1. ***Classification on the basis of purpose of investigation:*** These tables are of two types viz. (a) General purpose table and (b) Special purpose table.

   (a) *General purpose table:* A general purpose table is also called as a reference table. This table facilitates easy reference to the collected data. In the words of Croxton and Cowden, "The primary and usually the sole purpose of a reference table is to present the data in such a manner that the individual items may be readily found by a reader." A general purpose table is formed without any specific objective, but can be used for a number of specific purposes. Such a table usually contains a large mass of data and are generally given in the appendix of a report.

Notes



(b) *Special purpose table:* A special purpose table is also called a text table or a summary table or an analytical table. Such a table presents data relating to a specific problem. According to H. Secrist, "These tables are those in which are recorded, not the detailed data which have been analysed, but rather the results of analysis." Such tables are usually of smaller size than the size of reference tables and are generally found to highlight relationship between various characteristics or to facilitate their comparisons.

2. ***Classification on the basis of the nature of presented figures:*** Tables, when classified on the basis of the nature of presented figures can be (a) Primary table and (b) Derivative table.

   (a) *Primary Table:* Primary table is also known as original table and it contains data in the form in which it were originally collected.

   (b) *Derivative Table:* A table which presents figures like totals, averages, percentages, ratios, coefficients, etc., derived from original data.

   A table of time series data is an original table but a table of trend values computed from the time series data is known as a derivative table.

3. ***Classification on the basis of construction:*** Tables when classified on the basis of construction can be (a) Simple table, (b) Complex table or (c) Cross-classified table.

   (a) *Simple Table:* In this table the data are presented according to one characteristic only. This is the simplest form of a table and is also known as table of first order. The following blank table, for showing the number of workers in each shift of a company, is an example of a simple table.

Simple Table

Table No. .....

Shifts	*No. of Workers*
I	
II	
III	
Total	

Notes

(b) *Complex Table:* A complex table is used to present data according to two or more characteristics. Such a table can be two-way, three-way or multi-way, etc.

(i) Two-way table: Such a table presents data that is classified according to two characteristics. In such a table the columns of a table are further divided into sub-columns. The example of such a table is given below.

Complex Table

Table No. .....

Distribution of workers of a factory according to shifts and sex

Shifts	No. of Workers		Total
	Males	Females	
I			
II			
III			
Total			

(ii) Three-way table: When three characteristics of data are shown simultaneously, we get a three-way table as shown below.

Three way table

Table No. .....

Distribution of workers of a factory according to shifts, sex and training

Shifts	No. of Workers						Total No. of Workers
	Males		Total	Females		Total	
	Skilled	Unskilled		Skilled	Unskilled		
I							
II							
III							

(iii) Multi-way table: If each shift is further classified into three departments, say, manufacturing, packing and transportation, we shall get a four-way table, etc.

(c) *The Cross-Classified Table:* Tables that classify entries in both directions, i.e., row-wise and column-wise, are called cross-classified tables. The two ways of classification are such that each category of one classification can occur with any category of the other. The cross-classified tables can also be constructed for more than two characteristics also.

A cross-classification can also be used for analytical purpose, e.g., it is possible to make certain comparisons while keeping the effect of other factors as constant.

Example: Draw a blank table to show the population of a city according to age, sex and unemployment in various years.

Notes

Solution:

Table No. .....

Population of a city according to age, sex and unemployment in various years

Years	Age Sex	Population (in thousands)							
		Employed				Unemployed			
		below 20	20 - 60	60 & above	Total	below 20	20 - 60	60 & above	Total
2010	Males								
	Females								
	Total								
2011	Males								
	Females								
	Total								

Note: The table can be extended for the years 2012, 2013….., etc.

Example: In a sample study about coffee habit in two towns; the following information were received:

Town A: Females were 40%; total coffee drinkers were 45%; and male non-coffee drinkers were 20%.

Town B: Males were 55%; male non-coffee drinkers were 30%; and female coffee drinkers were 15%.

Represent the above data in a tabular form.

Solution:

Table No. .....

Distribution of population, according to sex and coffee habit, in two towns

Habit	Town A			Town B		
	Males	Females	Total	Males	Females	Total
Coffee Drinkers	40	5	45	25	15	40
Non - Coffee Drinkers	20	35	55	30	30	60
Total	60	40	100	55	45	100

Note: The figures are in percentage.

Example: Prepare a blank table for showing the percentage of votes polled by various political parties in India according to states, during 2010 general elections.

Notes

Solution:

Table No. .....
Percentage distribution of votes polled by political parties according to States in India during the 2010 general elections

States	*No. of Votes Polled*							
	Congress	*B.J.P.*	*Janta Dal*	*C.P.M.*	*B.S.P.*	*C.P.I.*	*Others*	*Total*
Assam								
Andhra								
Arunachal								
Bengal								
- - - - -								
- - - - -								
- - - - -								
Total								

Self Assessment

State whether the following statements are true or false:

11. A general purpose table is also called as a preference table
12. A general purpose table is formed without any specific objective, but can be used for a number of specific purposes.
13. A special purpose table is also called a text table or a summary table or an analytical table.
14. According to H. Secrist, "These tables are those in which are recorded, not the detailed data which have been analysed, but rather the results of analysis."
15. Primary table is also known as original table.
16. A table which presents figures like totals, averages, percentages, ratios, coefficients, etc., derived from original data are called derivative table.
17. Simple form of a table is also known as table of first order.
18. A complex table is used to present data according to two or more characteristics. Such a table can be two-way, three-way or multi-way, etc.

3.3 Methods of Tabulation

Tabulation of the collected data can be done in two ways: (i) By Manual Method, and (ii) By Mechanical Method.

1. ***Manual Method:*** When field of investigation is not too large and the number of characteristics are few, the work of tabulation can be done by hand.
2. ***Mechanical Method:*** This method is used when the data are very large. The use of machines save considerable amount of labour and time. With the development of high speed computers, the work of tabulation and analysis of data can be done very quickly and with greater accuracy.

Notes

Tasks

1. Prepare a blank table to represent the students of a college according to:
   (a) Percentage of marks obtained in an annual examination by taking class intervals 0 - 10, 10 - 20, ..., etc.
   (b) Sex-wise: males and females.
   (c) Faculty-wise: science, arts and commerce
2. Construct a blank table to represent two different dates and in five industries, the average wages of the four groups, males, females, eighteen years and over and under eighteen years. Suggest a suitable title.

Caution Analyze the field of investigation very carefully; only after decide for selecting manual or mechanical method of tabulation.

Notes

Statistical Investigation

The search for knowledge, done by analysing numerical facts, is known as a statistical investigation.

A statistical investigation is a process of collection and analysis of data. The relevance and accuracy of data obtained in an investigation depends directly upon the care with which it is planned. A properly planned investigation can give the best results with least cost and time. The investigation of the levels of living of the inhabitants of a particular area, the investigation of relationship between rainfall and the yield of a crop, etc., are some examples of statistical investigation.

A statistical investigation can be done by collecting numeral facts or data through the conduct of statistical surveys. The collected data are then analysed to get the results. Any process of statistical investigation can be divided into the following two broad categories:

1. Planning of a Statistical Investigation
2. Execution of a Statistical Investigation

Self Assessment

Multiple Choice Questions

19. When field of investigation is not too large and the number of characteristics are few then we use ..... method of investigation.
    (a) Manual (b) Semi manual
    (c) Mechanical (d) Automatic
20. ..... method is used when the data are very large.
    (a) Manual (b) Semi manual
    (c) Mechanical (d) Automatic

Notes

Case Study

Tabulation?

A survey of 370 students from the Commerce Faculty and 130 students from the Science Faculty revealed that 180 students were studying for only C.A. examinations, 140 for only Costing examinations and 80 for both C.A. and Costing examinations. The rest had offered part-time Management courses. Of those studying Costing only, 13 were girls and 90 boys belonged to the Commerce Faculty. Out of 80 students studying for both C.A. and Costing, 72 were from the Commerce Faculty amongst which 70 were boys. Amongst those who offered part-time Management courses, 50 boys were from the Science Faculty and 30 boys and 10 girls from the Commerce faculty. In all there were 110 boys in the Science Faculty.

Question

Present the above information in a tabular form. Find the number of students from the Science Faculty studying for part-time Management courses.

3.4 Summary

- Classification and tabulation of data are necessary to understand its broad features and to make it fit for statistical analysis.
- A table is made up of rows and columns.
- Various headings and subheadings used to designate columns and rows of a table are known as captions and stubs respectively
- A table can be of general or special purpose.
- If it represents original data, it is called a primary table otherwise it is called a derivative table.
- A table can be simple, complex, cross-classified; one, two or multi-way, etc

3.5 Keywords

Classification: Classification is a process of statistical analysis while tabulation is a process of presentation.

Complex Table: A complex table is used to present data according to two or more characteristics. Such a table can be two-way, three-way or multi-way, etc.

Cross-Classified Table: Tables that classify entries in both directions, i.e., row-wise and column-wise, are called cross-classified tables

Derivative Table: A table which presents figures like totals, averages, percentages, ratios, coefficients, etc., derived from original data.

Foot note: Abbreviations, if any, used in the table or some other explanatory notes are given just below the last horizontal line in the form of footnotes.

General purpose table: A general purpose table is also called as a reference table. This table facilitates easy reference to the collected data.

Manual Method: When field of investigation is not too large and the number of characteristics are few, the work of tabulation can be done by hand.

Notes

Mechanical Method: This method is used when the data are very large. The use of machines save considerable amount of labour and time

Primary Table: Primary table is also known as original table and it contains data in the form in which it were originally collected.

Simple Table: In this table the data are presented according to one characteristic only. This is the simplest form of a table and is also known as table of first order.

Source-Note: This note is often required when secondary data are being tabulated. This note indicates the source from where the information has been obtained. Source note is also given as a footnote.

Special purpose table: A special purpose table is also called a text table or a summary table or an analytical table. Such a table presents data relating to a specific problem.

Statistical table: A statistical table is the logical listing of related quantitative data in vertical columns and horizontal rows of numbers, with sufficient explanatory and qualifying words, phrases and statements in the form of titles, headings and footnotes to make clear the full meaning of the data and their origin.

Tabulation: Tabulation is a systematic presentation of numerical data in rows and columns

3.6 Review Questions

1. Define the term tabulation.
2. What is the difference between tabulation and classification?
3. What is the need for tabulation?
4. What are the various parts of table?
5. What is the difference between primary table and derivative table?
6. What is the difference between footnote and source note?
7. What is the difference between simple and complex table?
8. What is the difference between manual and mechanical method of tabulation?
9. Tabulate the following information:

   In a trip organized by a college, there were 80 persons each of whom paid ₹ 15.50 on an average. There were 60 students, each of whom paid ₹ 16. Members of the teaching staff were charged at a higher rate. The number of servants was 6, all males and they were not charged anything. The number of ladies was 20% of the total of which one was a lady staff member.

10. There were 850 union and 300 non union workers in a factory in 2009. Of these, 250 were females out of which 100 were non union workers. The number of union workers increased by 50 in 2010 out of which 40 were males. Of the 350 non union workers, 125 were females. In 2011, there were 1,000 workers in all and out of 400 non union workers there were only 100 females. There were only 400 male workers in the union.

    Present the above information in a tabular form.

11. A super market divided into five main sections; grocery, vegetables, medicines, textiles and novelties, recorded the following sales in 2009, 2010 and 2011:

    In 2009 the sales in groceries, vegetables, medicines and novelties were ₹ 6,25,000, ₹ 2,20,000, ₹ 1,88,000 and ₹ 94,000 respectively. Textiles accounted for 30% of the total sales during the year.

Notes

In 2010 the total sales showed 10% increase over the previous year while grocery and vegetables registered 8% and 10% increase over the corresponding previous year, medicines dropped by ₹ 13,000 and textiles increased by ₹ 53,000 over their corresponding figure of 1985.

In 2011, though total sales remained the same as in 1986, grocery fell by ₹ 22,000, vegetables by ₹ 32,000, medicines by ₹ 10,000 and novelties by ₹ 12,000. Tabulate the above data.

Answers: Self Assessment

1.	Tabulation	2.	Tuttle
3.	analysis, presentation	4.	captions, stubs
5.	head-note	6.	not drawn
7.	horizontal	8.	investigation.
9.	self explanatory	10.	Source note
11.	False	12.	True
13.	True	14.	True
15.	True	16.	True
17.	True	18.	True
19.	manual	20.	Mechanical

3.7 Further Readings

Books

Bhardwaj R. S., *Business Statistics*, Excel Books.

Balwani Nitin, *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Garrett H.E. (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, Mc Graw Hill Book Company, New York.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hannagan T.J. (1982), *Mastering Statistics*, The Macmillan Press Ltd., Surrey.

Hooda R.P., *Statistics for Business and Economics*, Macmillan India Delhi, 2008

Jaeger R.M (1983), *Statistics: A Spectator Sport*, Sage Publications India Pvt. Ltd., New Delhi.

Lindgren B.W (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Selvaraj R., Loganathan C., *Quantitative Methods in Management*.

Sharma J.K., *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Notes

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutta.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Online links

http://www.publishyourarticles.org/knowledge-hub/statistics/tabulation.html

http://www.emathzone.com/tutorials/basic-statistics/tabulation-of-data.html

http://www.jstor.org/pss/2276664

http://wiki.answers.com/Q/What_do_you_understand_by_tabulation_in_statistics_for_management_subject

http://www.mathsisfun.com/data/frequency-distribution.html

http://www.mathsisfun.com/data/relative-frequency.html

Notes

Unit 4: Presentation of Data

CONTENTS

Objectives

Introduction

4.1 Diagrammatic Presentation

4.1.1 Advantages

4.1.2 Limitations

4.1.3 General Rules for Making Diagrams

4.1.4 Choice of a Suitable Diagram

4.2 Bar Diagrams

4.3 Circular or Pie Diagrams

4.4 Pictogram and Cartogram (Map Diagram)

4.5 Summary

4.6 Keywords

4.7 Review Questions

4.8 Further Readings

Objectives

After studying this unit, you will be able to:

- Discuss utility and advantages of diagrammatic presentation
- Write the limitations of diagrammatic presentation
- State the relevance of General Rules for making diagrams
- Explain the concept of bar diagrams and focus on its types
- Tell about pie diagrams, pictograms and cartograms

Introduction

An important function of statistics is the presentation of complex mass of data in a simple way so that it becomes easier to understand. Classification and tabulation are the techniques that help in presenting the data in an intelligible form. But with increase in volume of data, it becomes more and more inconvenient to understand even after its classification and tabulation.

4.1 Diagrammatic Presentation

To understand various trends of the data at a glance and to facilitate the comparison of various situations, the data are presented in the form of diagrams and graphs.

Notes

4.1.1 Advantages

Data presented in the form of diagrams are useful as well as advantageous in many ways, as is obvious from the following:

1. ***Diagrams are attractive and impressive:*** Data presented in the form of diagrams are able to attract the attention of even a common man. It may be difficult for a common man to understand and remember the data presented in the form of figures but diagrams create a lasting impression upon his mind. Due to their attractive and impressive character, the diagrams are very frequently used by various newspapers and magazines for the explanation of certain phenomena. Diagrams are also useful in modern advertising campaign.
2. ***Diagrams simplify data:*** Diagrams are used to represent a huge mass of complex data in simplified and intelligible form which is easy to understand.
3. ***Diagrams give more information:*** In addition to the depiction of the characteristics of data, the diagrams may bring out other hidden facts and relations which are not possible to know from the classified and tabulated data.
4. ***Diagrams save time and labour:*** A lot of time is required to study the trend and significance of voluminous data. The same data, when presented in the form of diagrams, can be understood in practically no time.
5. ***Diagrams are useful in making comparisons:*** Many a times the objective of the investigation is to compare two or more situations either with respect to time or places. The task of comparison can be very conventionally done by the use of diagrams.
6. ***Diagrams have universal applicability:*** Diagrams are used in almost in every field of study like economics, business, administration, social institutions and other fields.

4.1.2 Limitations

Inspite of the above advantages of diagrams, their usefulness is somewhat limited. One has to be very careful while drawing conclusions from diagrams. Their main limitations are:

1. Diagrams give only a vague idea of the problem which may be useful for a common man but not for an expert who wishes to have an exact idea of the problem.
2. Diagrams can at best be a supplement to the tabular presentation but not an alternative to it.
3. The information given by the diagrams vis-a-vis classification and tabulation is limited.
4. The level of precision of values indicated by diagrams is very low.
5. Diagrams are helpful only when comparisons are desired. They don't lead to any further analysis of data.
6. Diagrams can portray only limited number of characteristics. Larger the number of characteristics the more difficult it is to understand them using diagrams..
7. Diagrams are liable to be misused for presenting an illusory picture of the problem.
8. Diagrams don't give a meaningful look when different measurements have wide variations.
9. Diagrams drawn on a false base lines should be analysed very carefully.

Notes

4.1.3 General Rules for Making Diagrams

A diagrammatic presentation is a simple and effective method of presenting the information contained in statistical data. The construction of a diagram is an art, which can be acquired only through practice. However, the following rules should be observed, in their construction, to make them more effective and useful.

1. ***Appropriate title and footnote:*** Every diagram must have a suitable title written at its top. The title should be able to convey the subject matter in brief and unambiguous manner. The details about the title, if necessary, should be provided below the diagram in the form of a footnote.
2. ***Attractive presentation:*** A diagram should be constructed in such a way that it has an immediate impact on the viewer. It should be neatly drawn and an appropriate proportion should be maintained between its length and breadth. The size of the diagram should neither be too big nor too small. Different aspects of the problem may be emphasised by using various shades or colours.
3. ***Accuracy:*** Diagrams should be drawn accurately by using proper scales of measurements. Accuracy should not be compromised to attractiveness.
4. ***Selection of an appropriate diagram:*** There are various types of geometrical figures and pictures which can be used to present statistical data.
5. ***Index:*** When a diagram depicts various characteristics distinguished by various shades and colours, an index explaining these should be given for clear identification and understanding.
6. ***Source-Note:*** As in case of tabular presentation, the source of data must also be indicated if the data have been acquired from some secondary source.
7. ***Simplicity:*** As far as possible, the constructed diagram should be simple so that even a layman can understand it without any difficulty

4.1.4 Choice of a Suitable Diagram

Diagrammatic presentation of data can be done in various ways. The choice of a suitable diagram is a practical problem and should be done in the light of the following considerations:

1. The nature of data
2. Purpose of the diagram
3. The calibre of the persons to whom the information is to be communicated.

The choice of a suitable diagram depends upon the nature of the given data. It may be recalled that two or three-dimensional diagrams are more appropriate if there are large variations in the magnitudes of observations. Many a times, the purpose of drawing a diagram may also give a clue to its choice. For example, if it is desired to indicate the comparison of values relating to different situations, bar diagrams will be most suitable. Further, if one wishes to indicate various components of a characteristics, sub-divided bar diagrams can be used. The relative importance of various components can be shown by using percentage sub-divided bar diagram. When the number of components become very large, i.e., more than three or four, circular diagrams are preferred because bar diagrams look more crowded. If the statistical data consists of a series of observations with different components for each observation, percentage sub-divided bar diagrams are more suitable than the circular diagrams.

Notes

Before the choice of a suitable diagram, it is very necessary to know the level of education of the person for whom the diagram is to be suitable. Further, if the data are related to different geographical areas, the cartograms may be the most appropriate drawn. For persons with little knowledge of statistics, the pictograms or cartograms may be more choice.

Caution The selection of an appropriate diagram should be carefully done keeping in view the nature of data and objective of investigation.

Self Assessment

Fill in the blanks:

1. Classification and tabulation are the techniques that help in presenting the data in an ..... form.
2. Diagrams are attractive and .....
3. Diagrams ..... data
4. Diagrams are useful in making .....
5. Diagrams are liable to be misused for presenting ..... picture of the problem.
6. A ..... is a simple and effective method of presenting the information contained in statistical data.
7. The construction of a diagram is an ..... , which can be acquired only through practice.
8. The choice of a suitable diagram depends upon the ..... of the given data.

4.2 Bar Diagrams

One-dimensional diagrams are also known as bar diagrams. In case of one-dimensional diagrams, the magnitude of the characteristics is shown by the length or height of the bar. The width of a bar is chosen arbitrarily so that the constructed diagram looks more elegant and attractive. It also depends upon the number of bars to be accommodated in the diagrams. If large number of items are to be included in the diagram, lines may also be used instead of bars. Different types of bar diagrams are:

1. Line Diagram
2. Simple Bar Diagram
3. Multiple Bar Diagram
4. Sub-divided or Component Bar Diagram
5. Percentage Sub-divided Bar Diagram
6. Deviation Bar Diagram
7. Duo-directional Bar Diagram
8. Sliding Bar Diagram
9. Pyramid Diagram
10. Broken-Scale Bar Diagram
11. Histogram

Notes

12. Frequency Polygon
13. Frequency Curve
14. 'Ogive' or Cumulative Frequency Curve

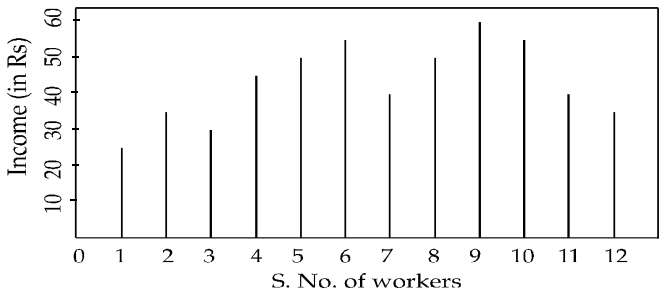
1. ***Line Diagram:*** In case of a line diagram, different values are represented by the length of the lines, drawn vertically or horizontally. The gap between the successive lines is kept uniform. The comparison of values of various items is done by the length of these lines. Although the comparison is easy, the diagram is not very attractive. This diagram is used when the number of items is relatively large.

Example: The income of 12 workers on a particular day was recorded as given below. Represent the data by a line diagram.

S. No. of workers :	1	2	3	4	5	6	7	8	9	10	11	12
Income (in ₹) :	25	35	30	45	50	55	40	50	60	55	40	35

Solution.

Line Diagram of income of 12 workers



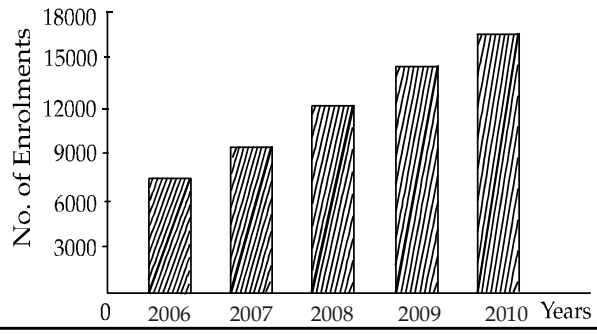
2. ***Simple Bar Diagram:*** In case of a simple bar diagram, the vertical or horizontal bars, with height proportional to the value of the item, are constructed. The width of a bar is chosen arbitrarily and is kept constant for every bar. Different bars are drawn so that the gap between the successive bars is same. Bar diagrams are particularly suitable for presenting individual series, such as time and spatial series.

Example: Represent the following data by a suitable diagram.

Years	:	2006	2007	2008	2009	2010
C. F. A Enrolments	:	7300	9400	12100	14600	16700

Solution:

Simple Bar Diagram



Notes

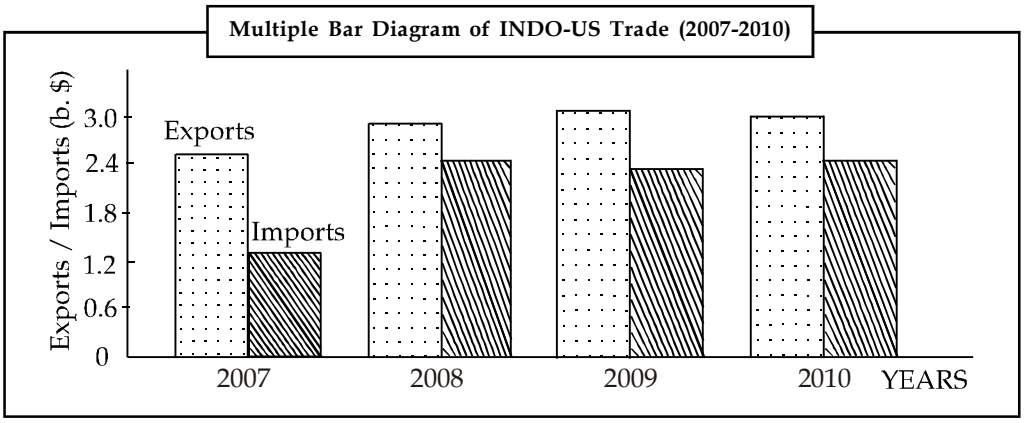
3. ***Multiple Bar Diagram:*** This type of diagram, also known as compound bar diagram, is used when comparisons are to be shown between two or more sets of data. A set of bars for a period or a related phenomena are drawn side by side without gaps while various sets of bars are separated by some arbitrarily chosen constant gap. Different bars are distinguished by different shades or colours. In order that various bars are comparable, it is necessary to draw them on the same scale.

Example: The following table gives the figures of Indo-US trade during 2007 to 2010. The figures of Indian exports and imports are in $ billion.

Year	:	2007	2008	2009	2010
Export	:	2.529	2.952	3.314	3.191
Import	:	1.460	2.484	2.463	2.486

Present the above data by a suitable diagram.

Solution:

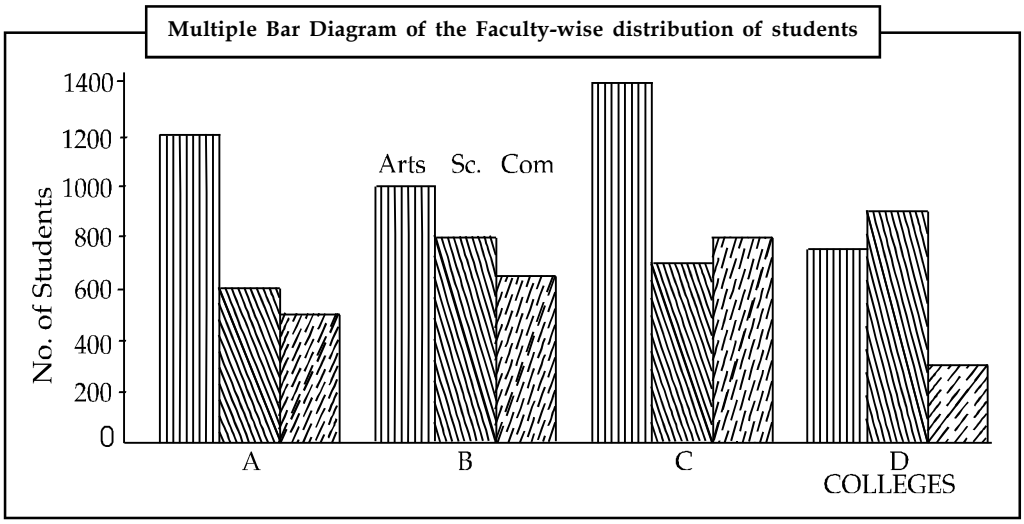


Source: U.S. Department of Commerce

4. ***Sub-divided or Component Bar Diagram:*** In case of a sub-divided bar diagram, the bar corresponding to each phenomenon is divided into various components. The portion of the bar occupied by each component denotes its share in the total. For example, the bar corresponding to the number of students in a course can be sub-divided into boys and girls. The subdivisions of different bars should always be done in the same order and they should be distinguished from each other by using different colours or shades.

   Sub-divided bar diagram is useful when it is desired to represent the comparative values of different components of a phenomenon. The main limitation of this diagram is that since various components are not drawn on a common base, they are difficult to compare. This diagram is used only if there are few components of a phenomenon.

Notes



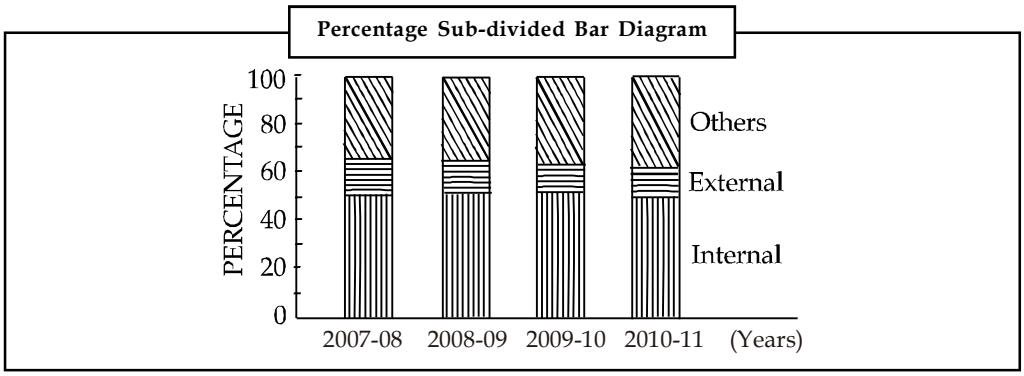
5. ***Percentage Sub-Divided Bar Diagram:*** A sub-divided diagram is used to show absolute magnitudes of various components. These magnitudes can be changed into relative by converting them as a percentage of the total. The length of each bar is taken as 100 and length of each component is denoted by its percentage value. As before the different components are distinguished from each other by different type of shades or colours.

Example: The following table gives the debt position of Central Government as the amount (₹ Crores) outstanding at the end of March 31 of each year. Represent the data by a percentage sub-divided bar diagram.

Years	:	**2007 - 08**	**2008 - 09**	**2009 - 10**	**2010 - 11**
InternalDebt	:	58537	71039	86313	98646
External Debt	:	16637	18153	20299	23223
Other Liabilities	:	38267	48422	59934	73692
Total	:	113441	137614	166546	195561

Solution: The amounts are given in absolute terms, therefore, these are to be converted into percentages in order to show them on a percentage sub-divided bar diagram.

Years	:	**2007 - 08**	**2008 - 09**	**2009 - 10**	**2010 - 11**
Internal Debt	:	51	52	52	50
External Debt	:	15	13	12	12
OtherLiabilities	:	34	35	36	38
Total	:	100	100	100	100



Notes

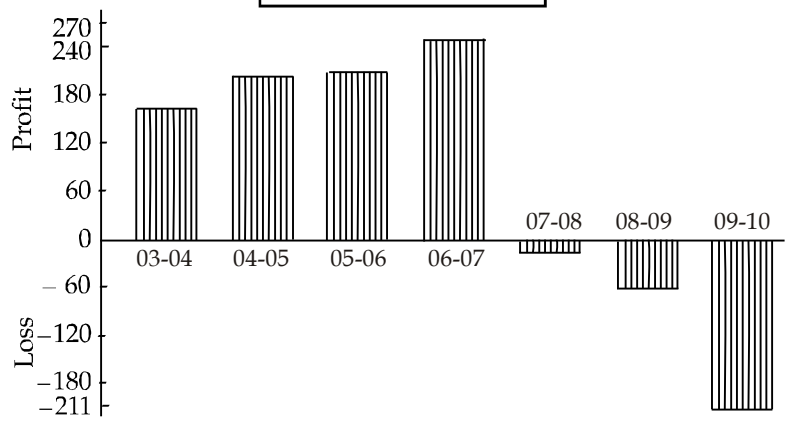
6. ***Deviation Bar Diagram:*** This diagram represents net quantities like profit and loss, positive and negative balance of trade, surplus and deficit, etc. Positive quantities are shown above X-axis and negative quantities are shown below it.

Example: Represent the following data by a suitable diagram.

Net Profit/Loss of Indian Airlines

Year	Profit/Loss(in ₹ Crores)	Year	Profit/Loss(in ₹ Crores)
2003 - 04	163 . 1	2007 - 08	(–) 15 . 2
2004 - 05	204 . 3	2008 - 09	(–) 64 . 6
2005 - 06	206 . 0	2009 - 10	(–) 211 . 0
2006 - 07	252 . 2		

Deviation Bar Diagram

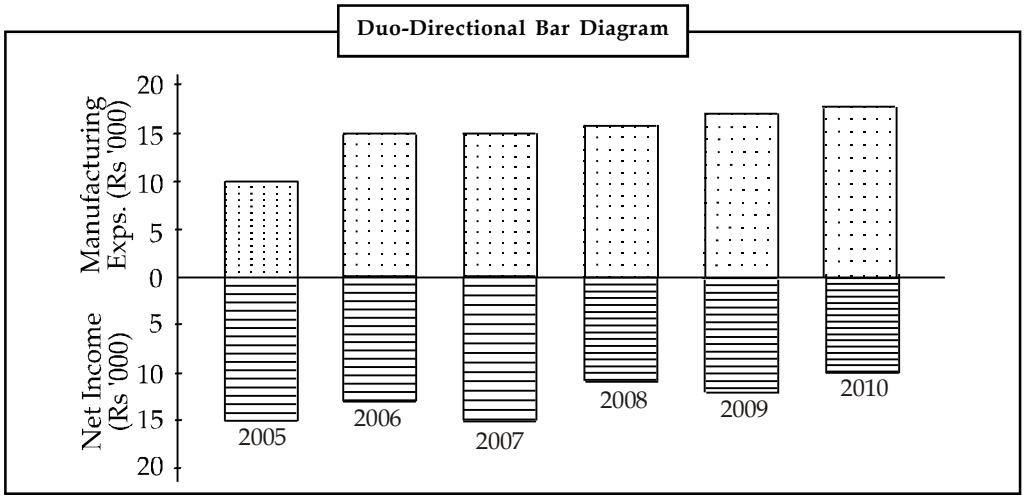


7. ***Duo-Directional Bar Diagram:*** This diagram is used to show an aggregate of two components. One of the components is shown above X-axis and the other below it. Both the components added together give total value.

Example: Represent the following data by a Duo-directional bar diagram.

Years	Total Income of a Manufacturer (in ₹ '000)	Manufacturing Expenses (in ₹ '000)	Net Income (in ₹ '000)
2005	25	10	15
2006	28	15	13
2007	30	15	15
2008	27	16	11
2009	29	17	12
2010	28	18	10

Notes

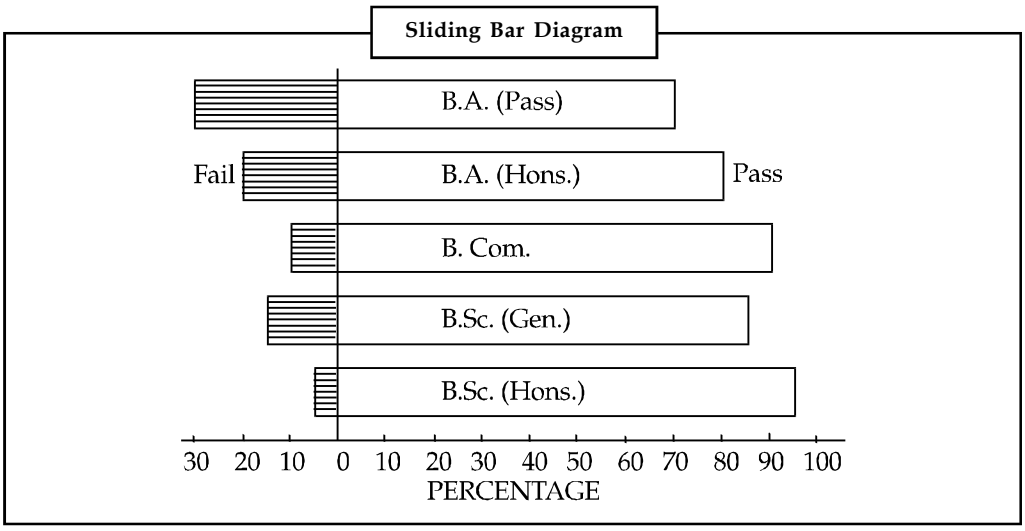


8. ***Sliding Bar Diagram:*** Sliding bar diagrams are similar to duo-directional bar diagrams. Whereas absolute values are shown by duo-directional bar diagrams, the percentage is shown using sliding bar diagrams. The length of each sliding bar is same, which represents 100%. The bars can be drawn horizontally or vertically.

Example: Represent the following data by a sliding bar diagram.

Result of a college in various courses

	Pass (Percentage)	Fail (Percentage)
1. B. A. (Pass)	70	30
2. B. A. (Hons.)	80	20
3. B. Com.	90	10
4. B. Sc. (Gen.)	85	15
5. B. Sc. (Hons.)	95	5



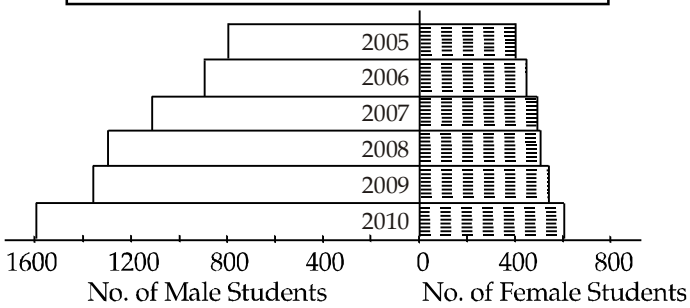
Notes

9. ***Pyramid Diagram:*** This diagram is used to represent the distribution of population according to sex, age, occupation, education, etc. The bars are drawn adjacently one above the other so as to look like a pyramid, as shown in the diagram.

Example: Given below are the figures of enrolment in a college during 2005-2010. Represent the data with the help of a suitable diagram.

Years	:	2005	2006	2007	2008	2009	2010
Male Students	:	800	850	1120	1300	1360	1600
Female Students	:	400	450	480	500	540	600
Total	:	1200	1300	1600	1800	1900	2200

Sex-wise Distribution of Students of a College

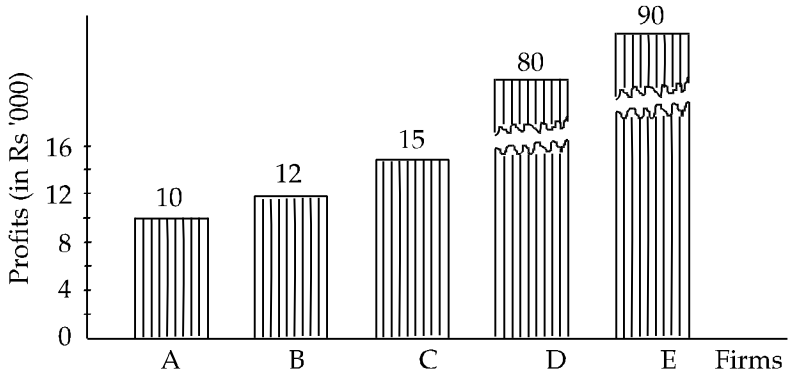


10. ***Broken-Scale Bar Diagram:*** When there are one or more figures of unusually high magnitude while the majority of the figures are of low magnitude, the diagrammatic representation is done by using a broken scale as shown in the following example.

Example: Represent the following data by a suitable diagram :

Firms	:	*A*	*B*	*C*	*D*	*E*
Profit (in Rs '000)	:	10	12	15	80	90

Bar Diagram Representing Profits of five Firms



Notes

Task "Diagrams do not add anything to the meaning of statistics but when drawn and studied intelligently they bring to view the salient characteristics of groups and series". Justify it.

Did u know? A sub-divided or percentage sub-divided bar diagram is also known as a stacked bar diagram.

11. ***Histogram:*** A histogram is a graph of a frequency distribution in which the class intervals are plotted on X- axis and their respective frequencies on Y- axis. On each class, a rectangle is erected with its height proportional to the frequency density of the class.

    (a) *Construction of a Histogram when Class Intervals are equal:* In this case the height of each rectangle is taken to be equal to the frequency of the corresponding class. The construction of such a histogram is illustrated by the following example.

Example: The frequency distribution of marks obtained by 60 students of a class in a college is given below :

Marks	30 - 34	35 - 39	40 - 44	45 - 49	50 - 54	55 - 59	60 - 64
No. of Students	3	5	12	18	14	6	2

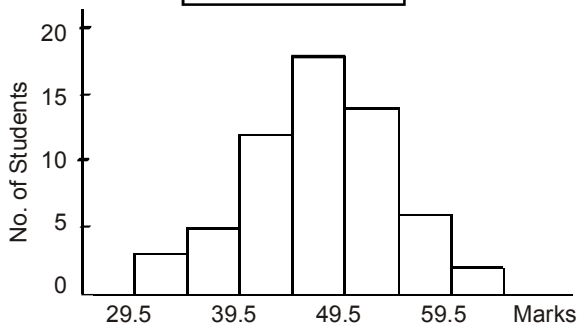
Draw a histogram for the distribution.

Solution: Since the upper limit of a class is not equal to the lower limit of its following class, the class boundaries will have to be determined. The distribution after adjustment will be as given below.

Table

Marks	No. of Students
29.5 - 34.5	3
34.5 - 39.5	5
39.5 - 44.5	12
44.5 - 49.5	18
49.5 - 54.5	14
54.5 - 59.5	6
59.5 - 64.5	2

Histogram



Notes

(b) *Construction of a Histogram when Class Intervals are not equal:* When different classes of a frequency distribution are not equal, the frequency density (frequency ÷ width) of each class is computed. The product of frequency density and the width of the class having shortest interval is taken as the height of the corresponding rectangle.

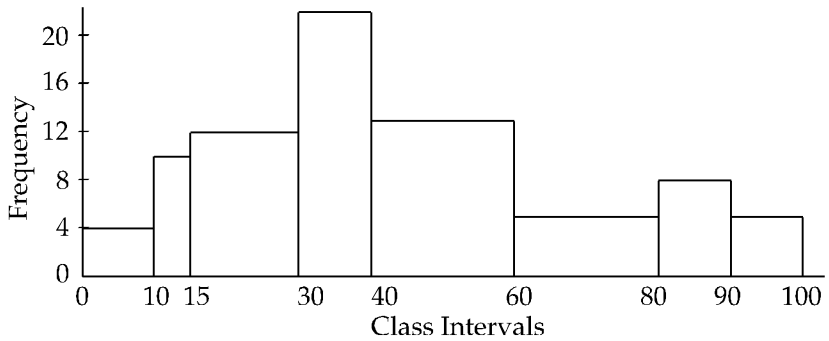
Example: Represent the following frequency distribution by a histogram.

Class Intervals	:	0-10	10-15	15-30	30-40	40-60	60-80	80-90	90-100
Frequency	:	8	10	36	44	52	20	16	10

Solution: The height of the rectangle = Frequency Density × Shortest Class Width

Class Intervals	Frequency	Frequency Density (f.d.)	Height of the rectangle = f.d.×5
0-10	8	0.8	4.0
10-15	10	2.0	10.0
15-30	36	2.4	12.0
30-40	44	4.4	22.0
40-60	52	2.6	13.0
60-80	20	1.0	5.0
80-90	16	1.6	8.0
90-100	10	1.0	5.0

Histogram



Note: If the mid points of various classes are given in place of class intervals then these must first be converted into classes.

12. ***Frequency Polygon:*** A frequency polygon is another method of representing a frequency distribution on a graph. Frequency polygons are more suitable than histograms whenever two or more frequency distributions are to be compared.

A frequency polygon is drawn by joining the mid-points of the upper widths of adjacent rectangles, of the histogram of the data, with straight lines. Two hypothetical class intervals, one in the beginning and the other in the end, are created. The ends of the polygon are extended upto base line by joining them with the mid-points of hypothetical classes. This step is necessary for making area under the polygon to be approximately equal to the area under the histogram. Frequency polygon can also be constructed without making rectangles. The points of frequency polygon are obtained by plotting mid-points of classes against the heights of various rectangles, which will be equal to the frequencies if all the classes are of equal width.

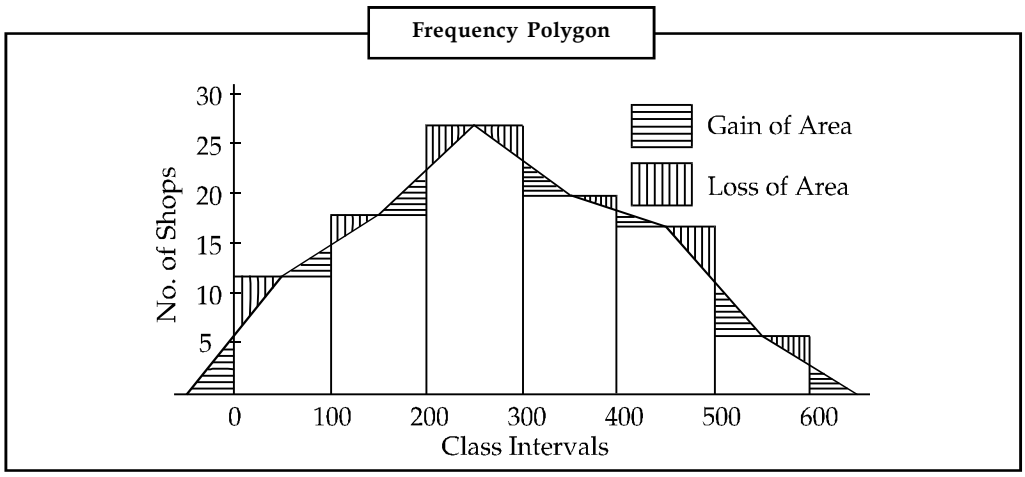
Notes

Example: The daily profits (in rupees) of 100 shops are distributed as follows :

Profit/Shop:	0-100	100-200	200-300	300-400	400-500	500-600
No. of Shops:	12	18	27	20	17	6

Construct a frequency polygon of the above distribution.

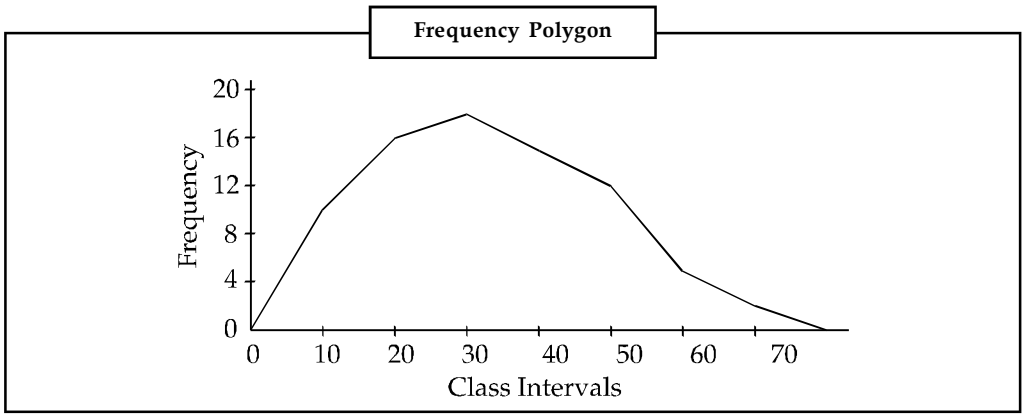
Solution:



Example: Represent the following data by a frequency polygon.

Classes:	5-15	15-25	25-35	35-45	45-55	55-65	65-75
Frequency:	10	16	18	15	12	5	2

Solution: Here the frequency polygon is drawn by plotting mid-points of class intervals against their respective frequencies.



13. ***Frequency curve:*** When the vertices of a frequency polygon are joined by a smooth curve, the resulting figure is known as a frequency curve. As the number of observations increases, there is need of having more and more classes to accommodate them and hence the width of each class will become smaller and smaller. In such a situation the variable under consideration tend to become continuous and the frequency polygon of the data tends to acquire the shape of a frequency curve. Thus, a frequency curve may be regarded as a limiting form of frequency polygon as the number of observations become large.

Notes

The construction of a frequency curve should be done very carefully by avoiding, as far as possible, the sharp and sudden turns. Smoothing should be done so that the area under the curve is approximately equal to the area under the histogram.

A frequency curve can be used for estimating the rate of increase or decrease of the frequency at a given point. It can also be used to determine the frequency of a value (or of values in an interval) of the variable. This method of determining frequencies is popularly known as interpolation method.

14. ***Cumulative Frequency Curve or Ogive:*** The curve obtained by representing a cumulative frequency distribution on a graph is known as cumulative frequency curve or ogive. Since a cumulative frequency distribution can of 'less than' or 'greater than' type and, accordingly, there are two type of ogive, 'less than ogive' and 'more than ogive'.

An ogive is used to determine certain positional averages like median, quartiles, deciles, percentiles, etc. We can also determine the percentage of cases lying between certain limits. Various frequency distributions can be compared on the basis of their ogives.

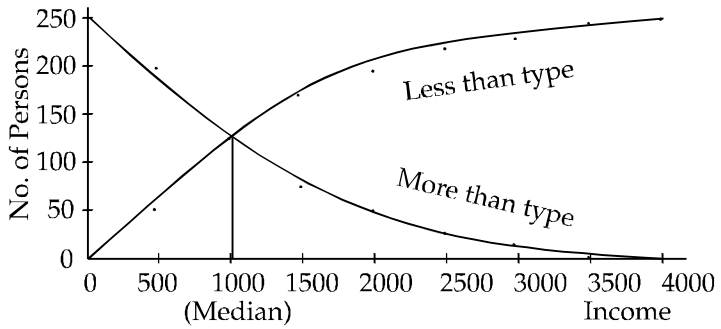
Example: Draw 'less than' and 'more than' ogives for the following distribution of monthly salary of 250 families of a certain locality.

Income Intervals	:	0-500	500-1000	1000-1500	1500-2000
No. of Families	:	50	80	40	25
Income Intervals	:	2000-2500	2500-3000	3000-3500	3500-4000
No.of Families	:	25	15	10	5

Solution: First we construct 'less than' and 'more than' type cumulative frequency distributions.

Income less than	Cumulative Frequency	Income more than	Cumulative Frequency
500	50	0	250
1000	130	500	200
1500	170	1000	120
2000	195	1500	80
2500	220	2000	55
3000	235	2500	30
3500	245	3000	15
4000	250	3500	5

Ogive



We note that the two ogives intersect at the median.

Notes

Self Assessment

State whether the following statements are true or false:

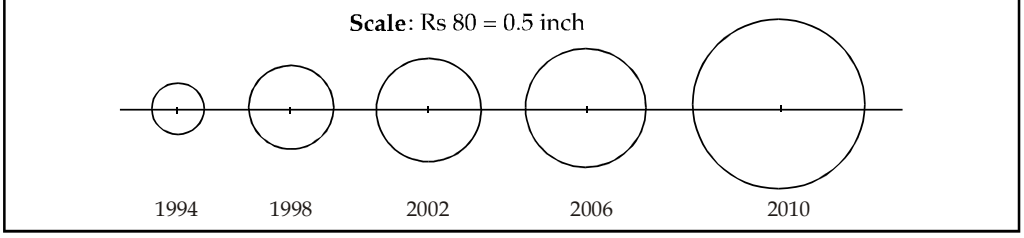
9. One-dimensional diagrams are also known as bar diagrams.
10. In case of one-dimensional diagrams, the magnitude of the characteristics is shown by the length or height of the bar.
11. Sub-divided bar diagram is useful when it is desired to represent the comparative values of different components of a phenomenon.
12. Duo-Directional Bar Diagram is used to show an aggregate of two components.
13. Pyramid Diagram is used to represent the distribution of population according to sex, age, occupation, education, etc.

4.3 Circular or Pie Diagrams

In the above example, one can also draw circles in place of squares. The radius of the circle is given by $r = \sqrt{\frac{A}{\pi}}$, where A denotes area of the circle whose value is given by the value of an item.

Year	:	1994	1998	2002	2006	2010
India©sExports (X) (in ₹ crores)	:	1823	4970	6591	9981	20295
Radius $\left(\sqrt{X+\pi}\right)$	:	24.1	39.8	45.8	56.4	80.4

Pie Diagram



In order to show proportions of various components, a circle can also be partitioned into sections in a similar manner as in component bar diagrams. Since there are 360° at the centre of a circle, these are divided in proportions to the magnitude of values of different items. The diagram, thus obtained is known as Angular Sector Diagram or more popularly as Pie Diagram. The construction of a pie diagram is explained by the following example:

Example: Show the following data of expenditure of an average working class family by a suitable diagram.

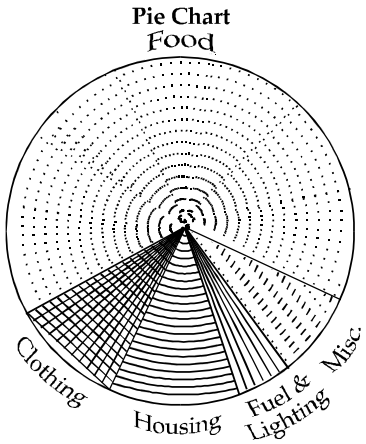
Item of Expenditure	*Percent of Total Expenditure*
(*i*) *Food*	65
(*ii*) *Clothing*	10
(*iii*) *Housing*	12
(*iv*) *Fuel and Lighting*	5
(*v*) *Miscellaneous*	8

Notes

Items of Expenditure	*Angles*
(i) *Food*	$\frac{65}{100} \times 360 = 234^\circ$
(ii) *Clothing*	$\frac{10}{100} \times 360 = 36^\circ$
(iii) *Housing*	$\frac{12}{100} \times 360 = 43.2^\circ$
(iv) *Fuel and Lighting*	$\frac{5}{100} \times 360 = 18^\circ$
(v) *Miscellaneous*	$\frac{8}{100} \times 360 = 28.8^\circ$

The angles of different sectors are calculated as shown below:

Pie Chart



Self Assessment

Multiple Choice Questions:

14. In order to show ..... of various components, a circle can also be partitioned into sections in a similar manner as in component bar diagrams.

    (a) Ratios (b) Proportions

    (c) Properties (d) Amount

15. Since there are ..... at the centre of a circle, these are divided in proportions to the magnitude of values of different items. The diagram, thus obtained is known as Pie Diagram.

    (a) 90° (b) 180°

    (c) 270° (d) 360°

16. Angular Sector Diagram is more popularly known as .....

    (a) Pie diagram (b) Bar diagram

    (c) Histogram (d) Pyramid Diagram

Notes

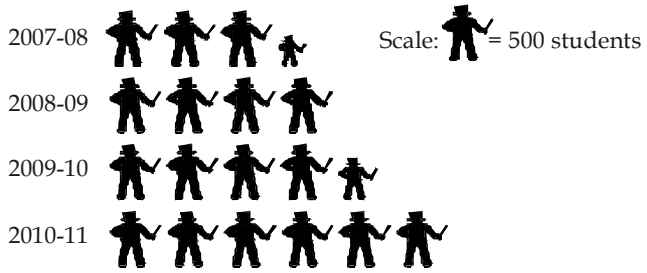
4.4 Pictogram and Cartogram (Map Diagram)

When numerical figures are represented by pictures, we get pictogram. Although such diagrams are very attractive and easier to understand, they are difficult to be drawn by everybody. The following pictogram represents the number of students in a college during the four academic sessions.

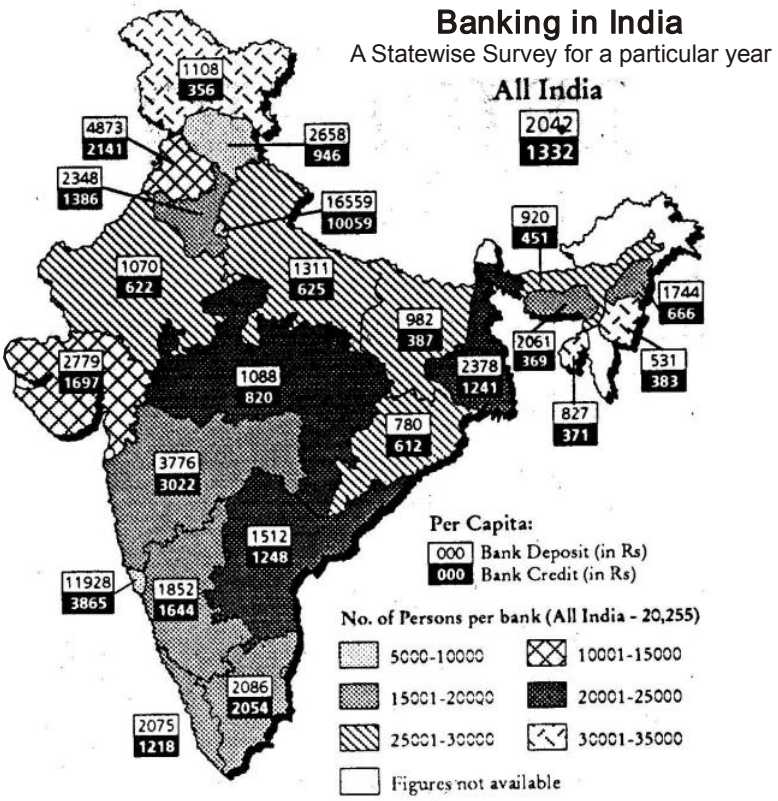
Academic Sessions	:	2007-08	2008-09	2009-10	2010-11
No. of Students	:	1625	2000	2250	3000

Pictogram

No. of Students in a College



Cartogram



Notes

Cartograms are used to represent data relating to a particular country or to a geographical area. Such a diagram can be used to represent various types of characteristics like density of population, yield of a crop, amount of rainfall, etc. The following diagram shows the per capita bank deposits, credits and the number of persons per bank in different states and union territories of India in a particular year.

Notes

Graphic Presentation

Graphic presentation is another way of pictorial presentation of the data. Graphs are commonly used for presentation of time series and frequency distributions. In situations where the diagrams as well as the graphs can be used, the later is preferred because of its advantages over the former. Graphic presentation of data, like diagrammatic presentation, also provides a quick and easier way of understanding various trends of data and to facilitate the process of comparison of two more situations. In addition to this, it can also be used as a tool of analysis. Graphic methods are sometimes used in place of mathematical computations to save time and labour, e.g., free hand curves may be fitted in place of mathematical curve to determine trend values.

Self Assessment

State whether the following statements are true or false:

17. When numerical figures are represented by pictures, we get pictogram.

18. Cartograms are used to represent data relating to a particular country or to a geographical area.

Case Study

ASSOCHAM

Associated Chamber of Commerce and Industry (ASSOCHAM) is very much concerned about the employment of youths and their pay rolls in small engineering industries, with special reference to automobile parts manufacturing, transport for hire, taxis, dealers of new and old vehicles, petrol stations and automobile repair garages. The chamber has employed you to collect the data regarding employment and pay role as on 31st March, 2000 and present it suitably through diagram so that it can be include in the final memorandum to be submitted to Minister for Industries. The data that you have collected is as follows:

	Industry	Employment on 31-3-2000	Avg. Earnings per employee per year (₹)
1.	Automobile parts manufacturers	4,34,856	56,540
2.	Transport for hire	15,26,897	26,348
3.	Taxis	11,32,560	42,685
4.	Dealers of new and used vehicles	1,09,805	13,684
5.	Retail filling stations	22,25,960	15,008
6.	Automobile repair garages	12,35,200	12,048

Present the data using a suitable diagram(s) so as to bring out the finer points.

Notes

4.5 Summary

- The diagrammatic presentation of data provides a quick and an easier way to understand the broad nature and trends of the given data.
- Diagrams are capable of being understood easily even by a common man. In addition to this, they facilitate the process of comparison of data in two or more situations.
- While using diagrams, their limitations must always be kept in mind.
- Diagrams give only a vague idea of the problem and therefore, cannot be used as a substitute for classification and tabulation.
- The diagrams can portray only a limited number of characteristics and are no longer useful when the number of characteristics become large.
- The main limitation of the diagrams being that these cannot be used as a tool of analysis.
- Various types of diagrams can be divided into five broad categories, viz. one-dimensional, two-dimensional, three-dimensional, pictograms and cartograms.
- Some important one-dimensional diagrams are line diagram, bar diagram, multiple bar diagram, component bar diagram, etc.
- Rectangular, square and circular diagrams are examples of two-dimensional diagrams.
- Cubes sphere and cylinder, etc., are three-dimensional diagrams.
- The diagrams can also be constructed by using relevant pictures or maps.

4.6 Keywords

Bar diagrams: One-dimensional diagrams are also known as bar diagrams.

Broken-Scale Bar Diagram: When there are one or more figures of unusually high magnitude while the majority of the figures are of low magnitude, the diagrammatic representation is done by using a broken scale.

Cartograms: Cartograms are used to represent data relating to a particular country or to a geographical area. Such a diagram can be used to represent various types of characteristics like density of population, yield of a crop, amount of rainfall, etc.

Deviation Bar Diagram: This diagram represents net quantities like profit and loss, positive and negative balance of trade, surplus and deficit, etc. Positive quantities are shown above X-axis and negative quantities are shown below it.

Duo-Directional Bar Diagram: This diagram is used to show an aggregate of two components. One of the components is shown above X-axis and the other below it. Both the components added together give total value.

Line Diagram: In case of a line diagram, different values are represented by the length of the lines, drawn vertically or horizontally.

Multiple Bar Diagram: This type of diagram, also known as compound bar diagram, is used when comparisons are to be shown between two or more sets of data. A set of bars for a period or a related phenomena are drawn side by side without gaps.

One-dimensional diagrams: In case of one-dimensional diagrams, the magnitude of the characteristics is shown by the length or height of the bar. The width of a bar is chosen arbitrarily so that the constructed diagram looks more elegant and attractive.

Notes

Percentage Sub-Divided Bar Diagram: A sub-divided diagram is used to show absolute magnitudes of various components. These magnitudes can be changed into relative by converting them as a percentage of the total.

Pictogram: When numerical figures are represented by pictures, we get pictogram.

Pie Diagram: In order to show proportions of various components, a circle can also be partitioned into sections in a similar manner as in component bar diagrams. Since there are 360° at the centre of a circle, these are divided in proportions to the magnitude of values of different items. The diagram, thus obtained is known as Angular Sector Diagram or more popularly as Pie Diagram.

Pyramid Diagram: This diagram is used to represent the distribution of population according to sex, age, occupation, education, etc. The bars are drawn adjacently one above the other so as to look like a pyramid.

Simple Bar Diagram: In case of a simple bar diagram, the vertical or horizontal bars, with height proportional to the value of the item, are constructed. The width of a bar is chosen arbitrarily and is kept constant for every bar.

Sliding Bar Diagram: Sliding bar diagrams are similar to duo-directional bar diagrams. Whereas absolute values are shown by duo-directional bar diagrams, the percentage is shown using sliding bar diagrams. The length of each sliding bar is same, which represents 100%. The bars can be drawn horizontally or vertically.

Sub-divided or Component Bar Diagram: In case of a sub-divided bar diagram, the bar corresponding to each phenomenon is divided into various components. The portion of the bar occupied by each component denotes its share in the total.

4.7 Review Questions

1. Describe the merits and limitations of the diagrammatic presentation of data.
2. What are different types of diagram which are used in statistics to show salient characteristics of groups and series? Illustrate your answer with examples.
3. What are the advantages of presentation of data through diagram? Give brief description of various types of diagram.
4. Explain clearly the necessity and importance of diagrams in statistics. What precautions should be taken in drawing a good diagram?
5. Describe, with suitable examples, the following type of diagrams:
   (a) Bar Diagram
   (b) Multiple Bar Diagram
   (c) Pie Diagram
   (d) Pictogram
6. Describe, in brief, different types of two-dimensional diagrams.
7. Discuss the usefulness of diagrammatic representation of facts and explain how would you construct circular diagrams?

Notes

8. Represent the following data by a line diagram:

S.No.	:	1	2	3	4	5	6	7	8	9	10
Weekly Income (₹)	:	240	270	315	318	330	345	354	360	375	390

9. Represent the following data by multiple bar diagram:

Name of the Bank	*Profits (in Rs Crores)*	
	2009-10	2010-11
State Bank of India	107.0	175.1
Canara Bank	76.0	156.6
Punjab National Bank	43.7	112.4
Bank of Baroda	35.1	95.1
Bank of India	22.5	56.6

10. The following table gives the support price of Rabi-crops during 2009-2010 and 2010-2011. Represent the given data by a suitable diagram.

Name of the Crop	*Support price (in Rs / quintal)*	
	2009-10	2010-11
Wheat	250	305
Barley	210	260
Gram	500	600
Rapeseed / Mustad	670	760
Safflower	640	720

11. The following table shows the yield of wheat in five countries during 2010. Represent the data by a suitable diagram.

Country	:	*China*	*EEC*	*USA*	*India*	*Canada*
Yield of Wheat (in Kg. / hect.)	:	3194	5118	2656	2125	2272

12. Represent the increase or decrease of production of the year 2010 through bar diagram:

    (a) Increase in production of cotton textile industry = 60%

    (b) Increase in the production of iron and steel industry = 50%

    (c) Decrease in sugar production = 40%

    (d) Decrease in cement production = 30%

13. Show the following data by means of a pie diagram:

    Area-wise exports from India during 2010-11.

West Europe	:	31.38%
Asia	:	31.04%
America	:	16.49%
East Europe	:	18.24%
Africa and others	:	2.85%

Notes

Answers: Self Assessment

1.	intelligible	2.	impressive
3.	simplify	4.	comparisons
5.	an illusory	6.	diagrammatic presentation
7.	art	8.	nature
9.	True	10.	True
11.	True	12.	True
13.	True	14.	(b)
15.	(d)	16.	(a)
17.	True	18.	True

4.8 Further Readings

Books

Bhardwaj R. S., *Business Statistics*, Excel Books.

Balwani Nitin, *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Garrett H.E. (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, Mc Graw Hill Book Company, New York.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hannagan T.J. (1982), *Mastering Statistics*, The Macmillan Press Ltd., Surrey.

Hooda R.P., *Statistics for Business and Economics*, Macmillan India Delhi, 2008

Jaeger R.M (1983), *Statistics: A Spectator Sport*, Sage Publications India Pvt. Ltd., New Delhi.

Lindgren B.W (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Selvaraj R., Loganathan C., *Quantitative Methods in Management*.

Sharma J.K., *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutta.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Notes

Online links

http://www.staff.vu.edu.au/mcaonline/units/statistics/presentation.html

http://www.slideshare.net/edithosb/graphic-presentation-of-data

http://nos.org/318courseE/L-6%20PRESENTATION%20OF%20STATISTICAL%20DATA.pdf

http://www.stars.rdg.ac.uk/data.html

http://www.mathsisfun.com/data/bar-graphs.html

http://www.mathsisfun.com/data/pie-charts.html

http://www.mathsisfun.com/data/pictographs.html

http://www.mathsisfun.com/data/histograms.html

Notes

Unit 5: Collection of Data

CONTENTS

Objectives

Introduction

5.1 Collection of Data

5.2 Method of Collecting Data

5.2.1 Drafting a Questionnaire or a Schedule

5.3 Sources of Secondary Data

5.3.1 Secondary Data

5.4 Summary

5.5 Keywords

5.6 Review Questions

5.7 Further Readings

Objectives

After studying this unit, you will be able to:

- Tell about the collection of data
- Describe the methods of collecting primary data
- State the merits and demerits of different methods of collecting primary data
- Explain the concept of drafting a Questionnaire or a Schedule
- Know about collection of secondary data

Introduction

The collection and analysis of data constitute the main stages of execution of any statistical investigation. However, between these two stages there is always an intermediate stage, known as the editing of data. The process of editing refines the collected data by checking inconsistencies, inaccuracies, illegible writings and other types of deficiencies or errors present in the collected information.

5.1 Collection of Data

The procedure for collection of data depends upon various considerations such as objective, scope, nature of investigation, etc. Availability of resources like money, time, manpower, etc., also affect the choice of a procedure. Data may be collected either from a primary or from a secondary source. Data from a primary source are collected, for the first time, keeping in view the objective of investigation. Secondary data, on the other hand, are available from certain publications or reports. Such data are already collected by some other agency in the past for some other purpose. Thus, the primary data collected with a specified objective of investigation, are likely to be more reliable as compared to secondary data. The use of secondary data, whenever

Notes

necessary, must be done very carefully. The cost of collection of primary data, however, are much higher.

Did u know? The system of data collection was described in Tuzuk-i-Bab4ri and Ain-i- Akabari. During Akbar's period, his revenue minister, Raja Todarmal, made a well organised survey of land for the collection of land revenue.

Self Assessment

Fill in the blanks:

1. The collection and analysis of data constitute the main stages of ..... of any statistical investigation.
2. Between collection and analysis of data, there is always an intermediate stage, known as the .....
3. The process of ..... refines the collected data by checking inconsistencies, inaccuracies, illegible writings and other types of deficiencies or errors present in the collected information.
4. The procedure for collection of data depends upon various ..... such as objective, scope, nature of investigation, etc.
5. Data may be collected either from a ..... or from a ..... source.
6. Primary data collected with a specified objective of investigation, are likely to be ..... reliable as compared to secondary data.
7. The cost of collection of primary data, however, are .....

5.2 Method of Collecting Data

For collection of primary data, the investigator may chose any one or a combination of the following methods:

1. Direct Personal observation
2. Indirect Oral Interview
3. Information through Local Agencies or Correspondents
4. Information through Questionnaires filled by Respondents
5. Information through Schedules filled by Investigators

1. ***Direct Personal Observation:*** Under this method, the investigator collects data by having direct contact with the units of investigation. The accuracy of collected data depends, to a large extent, upon the training and attitude of the investigator and the supporting attitude of the respondents.

   This method is suitable for an intensive type of investigation where (i) the scope of investigation is narrow, (ii) the process of investigation is so complex that it requires personal attention of the investigator, (iii) the investigation is confidential and (iv) more emphasis is to be given to the accuracy of the data.

Notes

Merits

(a) Original data are collected.

(b) Collected data are more accurate and reliable.

(c) The investigator can modify or put indirect questions in order to extract satisfactory information.

(d) The collected data are often homogeneous and comparable.

(e) Some additional information may also get collected, along with the regular information, which may prove to be helpful in future investigations.

(f) Misinterpretations or misgivings, if any, on the part of the respondents can be avoided by the investigators.

Demerits

(a) This method is expensive and time consuming, particularly when the field of investigation is large.

(b) It is not possible to properly train a large team of investigators.

(c) The bias or prejudice of investigators can affect the accuracy of data to a large extent.

(d) Data are collected as per the convenience and willingness of the respondents.

2. ***Indirect Oral Interview:*** This method is used when the area of investigation is very large or the respondents are reluctant to part with the information due to various reasons. Under this method, the investigator collects data from a third party or witness or head of an institution, etc., who is supposed to be in touch with the respondents. When the field of investigation is very large, the information about a large number of respondents can indirectly be obtained from one person who may be head (or pradhan) of that community.

Merits

(a) This method is suitable when the area of investigation is large or when the respondents are reluctant to part with the information.

(b) It is economical in terms of time, money and manpower.

(c) Since the information is collected from the persons who are well aware of the situation, it is likely to be unbiased and reliable.

(d) This method is particularly suitable for the collection of confidential information. For example, a person may not like to reveal his habit of drinking, smoking, gambling, etc., which may be revealed by others.

Demerits

(a) In the absence of direct contact between investigator and the respondent, it may happen that many important points remain unnoticed.

(b) As compared with direct personal observation, the degree of accuracy of the data is likely to be lower.

(c) The persons, providing the information, may be prejudiced or biased.

(d) Since the interest of the person, providing the information, is not at stake, the collected information is often vague and unreliable.

(e) The information collected from different persons may not be homogeneous and comparable.

Notes

3. ***Information Through Local Agencies or Correspondents:*** Under this method, local agents or correspondents are appointed in different parts of the area under investigation. These agents send the desired information at regular intervals of time. This method is often adopted by newspapers.

Merits

(a) This method is useful in situations where the area of investigation is very large and periodic information is to be collected from the distant places.

(b) It is economical in terms of time, money and labour.

Demerits

(a) The collected information lacks originality.

(b) The information is likely to be affected by the bias of the corespondents.

(c) It is not possible to obtain results with high degree of accuracy.

(d) The information supplied by different correspondents often lacks homogeneity and hence, not comparable.

4. ***Information Through Questionnaires Filled by Respondents:*** The information, in this method, is collected through the filling of questionnaires by the respondents. A questionnaire consists of a list of questions pertaining to the investigation. Blank spaces are left for writing answers. The questionnaire is sent to the respondents along with a covering letter for soliciting their cooperation by giving the correct information and returning the filled questionnaires well in time. With a view to get accurate information, the respondents may also be acquainted with the objective(s) of the investigation along with the assurance that the supplied information will be kept confidential and shall not, in any way, be misused against them. To get better response, self addressed and stamped envelope should also be sent to the respondents, in case the information is to be obtained by post. The basic assumption underlying this method is that the respondents are educated and have no difficulty in filling and sending the questionnaires. This method is adopted by research workers and other official and non-official agencies.

Merits

(a) This method is useful for the collection of information from an extensive area of investigation.

(b) This method is economical as it requires less time, money and labour.

(c) The collected information is original and more reliable.

(d) It is free from the bias of the investigator.

Demerits

(a) Very often, there is problem of 'non-response' as the respondents are not willing to provide answers to certain questions.

(b) The respondents may provide wrong information if the questions are not properly understood.

(c) It is not possible to collect information if the respondents are not educated.

(d) Since it is not possible to ask supplementary questions, the method is not flexible.

(e) The results of an investigation are likely to be misleading if the attitude of the respondents is biased.

Notes

(f) The process is time consuming, particularly when the information is to be obtained by post.

5. ***Information Through Schedules Filled by Investigators:*** The information obtained through mailing the questionnaires to the respondents is generally incomplete and unrepresentative. To avoid this problem, the work of filling of a questionnaire, termed as schedule here, can be done by the investigator through personal contact with the respondent. In order to get reliable information, the investigator should be tactful, well trained, unbiased and hard working.

Merits

(a) This method is suitable for an extensive area of investigation.

(b) Since the investigator has a direct contact with the respondents, it is possible to get accurate and reliable information.

(c) By asking cross questions it is possible to test the truth of the supplied information.

(d) The problem of non-response is minimised.

(e) It is possible to get the information even if the respondents are not educated.

Demerits

(a) This method is very expensive and time consuming.

(b) The collected information is affected, to a large extent, by the bias of the investigator.

(c) If the investigators are negligent or not properly trained, the results of investigations are likely to be misleading.

5.2.1 Drafting a Questionnaire or a Schedule

A questionnaire or a schedule is a list of questions relating to the problem under investigation. There is no basic difference between a questionnaire and a schedule. A questionnaire is filled by the respondent, while a schedule is filled by the investigator.

The quality of information collected through the filling of a questionnaire depends, to a large extent, upon the drafting of its questions. Hence, it is extremely important that the questions be designed or drafted very carefully and in a tactful manner. Although, there are no hard and fast rules to be followed, but the following general points must always be kept in mind to draft a good questionnaire or schedule.

1. The questions should be simple, unambiguous and precise.

2. The number of questions should be as small as possible. Only those questions which have a direct relevance to the problem be included.

3. The question should be framed in such a manner that their answers are specific and precise.

   The following four type of questions are generally framed in any questionnaire :

   (a) *Simple alternative questions:* Such questions are answered by yes/no or right/wrong, etc.

   (b) *Multiple choice questions:* In such questions, the possible answers are printed in the questionnaire and the respondent is supposed to tick any one of them.

Notes

Example: What is your occupation?

1. Agriculture
2. Industry
3. Trade
4. Service
5. Any other

(c) *Specific information questions:* Such questions are used to extract specific information like; how many members are there in your family, what is your monthly income, etc.

(d) *Open questions:* These types of questions are to be answered by the respondent in his own words. The questions should be such that it is possible to answer them in few words. Care should be taken to avoid answers that use the words like probably, fairly good, etc.

4. The questions should be capable of being easily answered by the respondents. The questions that rely too much on the memory of the respondent should be avoided.
5. The questions affecting the pride and sentiments of the respondents should be avoided. Similarly the questions pertaining to private affairs of the respondents should never be asked.
6. The questions relating to mathematical computations should be avoided.
7. The questions should follow a logical sequence so that a natural and spontaneous reply to each question is induced.
8. The questions should be directly related to the objective(s) of investigation.
9. Certain corroborative questions must also be asked to verify the accuracy of the supplied information.
10. Necessary instructions for filling the questionnaire should also be given in simple and precise form.
11. Enough space should be provided for answers. The questionnaire should look as attractive as possible.
12. Before actually using a questionnaire, a test check must always be done by obtaining answers from some respondents. The questionnaire should be modified, if necessary, in the light of these answers.

Format of a Questionnaire for assessing preferences of the users of two-wheelers.

QUESTIONNAIRE

Note: Please put tick mark ( ) in the relevant box of each question.

Section A

1. Name : _____
2. Address : _____
3. Age : _____
4. Sex : Male Female

Contd...

Notes

5. Marital Status : Married Unmarried Others
6. Number of children, if married:

   None One Two More than two
7. Highest educational qualification:

   Under Matric Matric Senior Secondary

   Graduate Post Graduate Higher
8. Occupation:

   Service Manufacture Trade

   Agriculture Others
9. Annual Income (in ₹):

   Less than 20,000 20,000 - 30,000 30,000 - 40,000

   40,000 - 50,000 50,000 - 60,000 60,000 and above

Section B

1. What is the brand name of your scooter?

   Bajaj L.M.L. Vespa Kinetic Honda

   Vijay Super Any other
2. How did you know this brand?

   Through advertisement in a newspaper Through advertisement in radio

   Through advertisement in television Through some friend

   Through some other agency
3. You purchased this scooter because:

   It gives more mileage It requires less maintenance

   It is more powerful It is more stable

   It has a better after-sales service

   (You can put tick mark in more than one box.)
4. You purchased this scooter from:

   Company showroom Authorised dealer

   Local dealer Others
5. What was the mode of payment?

   On cash down payment On Instalment basis

   On Hire-Purchase basis Others
6. How old is your scooter?

   Less than 1 year 1 - 3 years 3 - 5 years 5 - 10 years 10 or more years
7. Are you satisfied with your scooter? yes no

Notes

8. The maintenance expenditure, per month, of your scooter:

   Less than ₹ 100 ₹ 100–200 ₹ 200 or more

9. For major repairs, you go to:

   The authorised service station The local service station

   Any other workshop

10. If given a chance to purchase a new scooter, would you like to purchase the same scooter?
    yes no

Signatures

Task You are the sales promotion officer of Delta Cosmetics Co. Ltd. Your company is about to market a new product. Design a suitable questionnaire to conduct a consumer survey before the product is launched. State various types of persons that may be approached for replying to the questionnaire.

Did u know? Indirect oral interview is generally adopted by the police department for the collection of information regarding bad elements of a locality.

Self Assessment

State whether the following statements are true or false:

8. The accuracy of collected data depends, to a large extent, upon the training and attitude of the investigator and the supporting attitude of the respondents.
9. The collected data are often non homogeneous and comparable.
10. A questionnaire consists of a list of questions pertaining to the investigation.
11. The information obtained through mailing the questionnaires to the respondents is generally incomplete and unrepresentative.
12. There is no basic difference between a questionnaire and a schedule.
13. The quantity of information collected through the filling of a questionnaire depends, to a large extent, upon the drafting of its questions.
14. Simple alternative questions are answered by yes/no or right/wrong, etc.
15. Closed questions are to be answered by the respondent in his own words.

5.3 Sources of Secondary Data

The data, collected and used by some other person or agency for an investigation in the past, when used for the investigation of a current problem, is known as secondary data.

5.3.1 Secondary Data

Secondary data may be available in published or unpublished form. In published form, the data are available in magazines, research papers, newspapers, government publications, international publications, etc.

Notes

Caution Before using any secondary data, the user must satisfy himself regarding the following points:

1. Are data suitable for the current investigation?
2. Are data adequate for the current investigation?
3. Are data reliable?

Task Amongst various methods of rounding, which one is best and why?

Notes

Editing of Data

The next logical step, after the collection of data, is the process of refining it for proper utilisation. This process is known as editing. Editing of data includes the identification and dropping of the unwanted information. In addition to this, steps are taken to complete the left out information. Another aim of editing is to find out and rectify possible errors or irregularities during the collection of data. Thus, the process of editing implies the scrutiny of data in various ways. The process of editing of primary data can be divided into the following six stages:

1. Deciphering
2. Scrutiny for completeness
3. Scrutiny for uniformity
4. Scrutiny for consistency
5. Scrutiny for accuracy
6. Coding of data

Self Assessement

Multiple Choice Questions:

16. The data, collected and used by some other person or agency for an investigation in the past, when used for the investigation of a current problem, is known as ..... data.

    (a) Primary (b) Secondary

    (c) Simple (d) Complex

17. ..... data may be available in published or unpublished form

    (a) Secondary (b) Primary

    (c) Simple (d) Complex

Notes

18. In ..... form, the data are available in magazines, research papers, newspapers, government publications, international publications, etc.

(a) Published (b) Unpublished

(c) Reduced (d) Enlarged

Case Study

Chand Contractors

Chand Contractors supplies contract labor to various industrial units for carrying out their various production activities in and around Bhilai. Mr.R.B. Tripathi is the chief consultant and is responsible to manage the continuous supply of contract labor on weekly basis. The daily wages of contract labor varies from ₹ 25 to 95 per day depending on the skill, experience and the nature of work in the industry utilising the services of contract laborers. The daily wages and number of workers data have been compiled by Shri Tripathi for estimating the number of workers demanded and their average wages.

Daily Wages (₹)	No. of Workers	Daily Wages (₹)	No. off workers
20-25	21	60-65	36
25-30	29	65-70	45
30-35	19	70-75	27
35-40	39	75-80	48
40-45	43	80-85	21
45-50	94	85-90	12
50-55	73	90-95	5
55-60	68		

1. Draw a suitable diagram of the data to enable the chief executive of Chand Contractors to understand the relations between wages and number of workers.
2. Find out the number of workers getting wages lower than 57 and more than 77 using Ogive graphs.

5.4 Summary

- The stage of collection of data follows the stage of planning of a statistical investigation.
- The process for the collection of data depends upon the objective, scope, nature of investigation, etc.
- The data may be collected either from a primary or from a secondary source.
- Data from a primary source are original and correspond to the objective of investigation.
- The secondary data are often available in published form, collected originally by some other agency with a similar or different objective.
- The primary data are more reliable than the secondary data which, however, are more economical.
- There are several methods of collection of primary data.
- The choice of a particular method depends, apart from objective, scope and nature of investigation, on the availability of resources, literacy level of the respondents, etc.

Notes

- Secondary data should be used very carefully, only after examining that these are suitable, adequate and reliable for the purpose of investigation under consideration.

5.5 Keywords

Direct Personal Observation: Under this method, the investigator collects data by having direct contact with the units of investigation.

Editing of data: The collection and analysis of data constitute the main stages of execution of any statistical investigation. However, between these two stages there is always an intermediate stage, known as the editing of data.

Indirect Oral Interview: This method is used when the area of investigation is very large or the respondents are reluctant to part with the information due to various reasons. Under this method, the investigator collects data from a third party or witness or head of an institution, etc., who is supposed to be in touch with the respondents.

Multiple choice questions: In such questions, the possible answers are printed in the questionnaire and the respondent is supposed to tick any one of them.

Open questions: These types of questions are to be answered by the respondent in his own words.

Questionnaire/Schedule: questionnaire/schedule is a list of questions relating to the problem under investigation.

Secondary data: The data, collected and used by some other person or agency for an investigation in the past, when used for the investigation of a current problem, is known as secondary data.

Specific information questions: Such questions are used to extract specific information like; how many members are there in your family, what is your monthly income, etc.

5.6 Review Questions

1. What are various methods of collecting statistical data? Which of these is more reliable and why?
2. Discuss the comparative merits of various methods of collecting primary data. Which method would you recommend for the following investigations:
   (a) A family budget enquiry of teachers of a university.
   (b) Survey of economic conditions of workers in cottage and small scale industries of a town.
3. "In collection of statistical data, common sense is the chief requisite and experience is the chief teacher". Discuss this statement.
4. What do you understand by secondary data? State their chief sources and point out dangers involved in their use. What precaution must be taken while using such data for further investigation?
5. "Statistics especially other people's statistics are full of pitfalls for the user unless used with caution". Explain the meaning of this statement and mention various merits and demerits of using secondary data.
6. What are the requisites of a good questionnaire? Explain the procedure for collection of data through mailing of questionnaire.

Notes

7. Distinguish between a questionnaire and a schedule. What precautions should be taken in drafting a questionnaire?

8. Select the correct answer from the following:

   In order to collect primary data:

   (a) It is necessary for the investigator to have personal contact with the respondents.

   (b) It is not necessary for the investigator to have personal contact with the respondents.

   (c) The investigator has to mail questionnaires.

   (d) The respondents must be educated.

9. Fill in the blanks:

   (a) Secondary data are also ..... data for some other investigation.

   (b) Published data are termed as ..... data.

   (c) Primary data are ..... reliable than secondary data.

   (d) The distinction between primary and secondary data is of ..... only.

10. Explain the need for editing of data. What is the difference between editing of primary and secondary data?

11. Round the following figures to 2 digits:

    (a) 25,136 (b) 1.28 (c) 0.057 (d) 0.0085

12. Compute and interpret a useful ratio or percentage from the following:

    (i) A company has issued 5,00,000 shares and its earning is ₹ 15,00,000.

    (ii) Out of 500 workers of a firm only 150 are matriculates.

    (iii) The cost of petrol of a company, having 15 cars was ₹ 80,000 last year.

    (iv) The number of fatal road accidents during last month in a city was 15.

13. Collect data on any company's financial performance from different sources like internet, newspapers, magazines, etc. Are there any differences between the same? What inferences do you draw on the objectives of that particular type of media when they are presenting data.

14. Try and collect as much data as possible from different sources about the health levels of the people residing in your area. What problems come in while collecting this data?

15. Mount Shivalik Distilleries is a progressive manufacturer of 'Wasp' brand export quality rum. It follows the modern practices of presentation of data in various board meetings. The data collected by its finance director over a period of 3 years pertaining to its operations is shown below. The Finance Director desires that the data should be presented diagrammatically. Would you please help him in presenting the data.

	Industry	Employment on 31-3-2011	Avg. Earnings per employee per year (Rs)
1.	Automobile parts manufacturers	4,34,856	56,540
2.	Transport for hire	15,26,897	26,348
3.	Taxies	11,32,560	42,685
4.	Dealers of new and used vehicles	1,09,805	13,684
5.	Retail filling stations	22,25,960	15,008
6.	Automobile repair garages	12,35,200	12,048

Present the above data using a suitable diagram(s) so as to bring out the finer points.

Notes

16. Ansal Builders is engaged in the construction of a multistory building. It has recently conducted a cost audit. The manager (cost accounting) has collected the figures of total cost and its major constituents. The information collected as percentage of expenditure is shown below. Represent the data with the help of a suitable diagram.

Item	Expenditure %
Wages	25
Bricks	15
Cement	20
Steel	15
Wood	10
Supervison and Misc	15

Answers: Self Assessment

1.	execution	2.	editing of data
3.	editing	4.	considerations
5.	primary, secondary	6.	more
7.	much higher	8.	True
9.	False	10.	True
11.	True	12.	True
13.	False	14.	True
15.	False	16.	(b)
17.	(a)	18.	(a)

5.7 Further Readings

Books

Balwani Nitin, *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Bhardwaj R S., *Business Statistics*, Excel Books.

Garrett H.E (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, McGraw Hill Book Company, New York.

Selvaraj R, Loganathan, C *Quantitative Methods in Management.*

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Online links

http://en.wikipedia.org/wiki/Data_collection

http://www.balancedscorecard.org/Portals/0/PDF/datacoll.pdf

http://www.mathsisfun.com/data/univariate-bivariate.html

http://msdn.microsoft.com/en-us/library/bb677179.aspx

Notes

Unit 6: Measures of Central Tendency

CONTENTS

Objectives

Introduction

6.1 Average

6.1.1 Functions of an Average

6.1.2 Characteristics of a Good Average

6.1.3 Various Measures of Average

6.2 Arithmetic Mean

6.2.1 Calculation of Simple Arithmetic Mean

6.2.2 Weighted Arithmetic Mean

6.2.3 Properties of Arithmetic Mean

6.2.4 Merits and Demerits of Arithmetic Mean

6.3 Median

6.3.1 Determination of Median

6.3.2 Properties of Median

6.3.3 Merits, Demerits and Uses of Median

6.4 Other Partition or Positional Measures

6.4.1 Quartiles

6.4.2 Deciles

6.4.3 Percentiles

6.5 Mode

6.5.1 Determination of Mode

6.5.2 Merits and Demerits of Mode

6.5.3 Relation between Mean, Median and Mode

6.6 Geometric Mean

6.6.1 Calculation of Geometric Mean

6.6.2 Weighted Geometric Mean

6.6.3 Geometric Mean of the Combined Group

6.6.4 Average Rate of Growth of Population

6.6.5 Suitability of Geometric Mean for Averaging Ratios

6.6.6 Properties of Geometric Mean

6.6.7 Merits, Demerits and Uses of Geometric Mean

Contd...

Notes

6.7 Harmonic Mean
6.7.1 Calculation of Harmonic Mean
6.7.2 Weighted Harmonic Mean
6.7.3 Merits and Demerits of Harmonic Mean
6.8 Summary
6.9 Keywords
6.10 Review Questions
6.11 Further Readings

Objectives

After studying this unit, you will be able to:

- Define the term average and state its functions and characteristics.
- Write the uses, merits and demerits of mean,median and mode
- Tell about Mathematical Averages-AM,GM and HM
- Establish the relationship amongst mean, median and mode
- Establish the relationship amongst AM, GM and HM

Introduction

Summarization of the data is a necessary function of any statistical analysis. As a first step in this direction, the huge mass of unwieldy data is summarized in the form of tables and frequency distributions. In order to bring the characteristics of the data into sharp focus, these tables and frequency distributions need to be summarized further. A measure of central tendency or an average is very essential and an important summary measure in any statistical analysis.

6.1 Average

The average of a distribution has been defined in various ways. Some of the important definitions are:

"An average is an attempt to find one single figure to describe the whole of figures".

– *Clark and Sekkade*

"Average is a value which is typical or representative of a set of data".

– *Murray R. Spiegal*

"An average is a single value within the range of the data that is used to represent all the values in the series. Since an average is somewhere within the range of data it is sometimes called a measure of central value".

– *Croxton and Cowden*

"A measure of central tendency is a typical value around which other figures congregate".

– *Sipson and Kafka*

Notes

6.1.1 Functions of an Average

1. ***To present huge mass of data in a summarised form:*** It is very difficult for human mind to grasp a large body of numerical figures. A measure of average is used to summarise such data into a single figure which makes it easier to understand and remember.
2. ***To facilitate comparison:*** Different sets of data can be compared by comparing their averages. For example, the level of wages of workers in two factories can be compared by mean (or average) wages of workers in each of them.
3. ***To help in decision making:*** Most of the decisions to be taken in research, planning, etc., are based on the average value of certain variables. For example, if the average monthly sales of a company are falling, the sales manager may have to take certain decisions to improve it.

6.1.2 Characteristics of a Good Average

A good measure of average must posses the following characteristics:

1. It should be rigidly defined, preferably by an algebraic formula, so that different persons obtain the same value for a given set of data.
2. It should be easy to compute.
3. It should be easy to understand.
4. It should be based on all the observations.
5. It should be capable of further algebraic treatment.
6. It should not be unduly affected by extreme observations.
7. It should not be much affected by the fluctuations of sampling.

6.1.3 Various Measures of Average

Various measures of average can be classified into the following three categories:

1. ***Mathematical Averages***
   - (a) Arithmetic Mean or Mean
   - (b) Geometric Mean
   - (c) Harmonic Mean
   - (d) Quadratic Mean
2. ***Positional Averages***
   - (a) Median
   - (b) Mode
3. ***Commercial Average***
   - (a) Moving Average
   - (b) Progressive Average
   - (c) Composite Average

Out of above mentioned, we will discuss here only mathematical averages and positional averages.

Notes

Did u know? An average is a single value which can be taken as representative of the whole distribution.

Self Assessment

Fill in the blanks:

1. ..... of the data is a necessary function of any statistical analysis.
2. The huge mass of unwieldy data is summarized in the form of .....and .....
3. A ..... or an average is very essential and an important summary measure in any statistical analysis.
4. An ..... is a single value which can be taken as representative of the whole distribution.
5. A measure of central tendency is a typical value around which other figures.....
6. Different sets of data can be compared by comparing their .....
7. AM and GM comes under .....averages.

6.2 Arithmetic Mean

Before the discussion of arithmetic mean, we shall introduce certain notations. It will be assumed that there are n observations whose values are denoted by $X_1, X_2, \ldots\ldots X_n$ respectively. The sum of these observations $X_1 + X_2 + \ldots\ldots + X_n$ will be denoted in abbreviated form as, where $\sum$ (called sigma) denotes summation sign. The subscript of X, i.e., 'i' is a positive integer, which indicates the serial number of the observation. Since there are n observations, variation in i will be from 1 to n. This is indicated by writing it below and above $\sum$, as written earlier. When there is no ambiguity in range of summation, this indication can be skipped and we may simply write $X_1 + X_2 + \ldots\ldots + X_n = \sum X_i$.

Arithmetic Mean is defined as the sum of observations divided by the number of observations. It can be computed in two ways: (i) Simple arithmetic mean and (ii) weighted arithmetic mean. In case of simple arithmetic mean, equal importance is given to all the observations while in weighted arithmetic mean, the importance given to various observations is not same.

6.2.1 Calculation of Simple Arithmetic Mean

When Individual Observations are given

Let there be n observations $X_1, X_2 \ldots\ldots X_n$. Their arithmetic mean can be calculated either by direct method or by short cut method. The arithmetic mean of these observations will be denoted by.

Direct Method

Under this method, $\overline{X}$ is obtained by dividing sum of observations by number of observations, i.e.,

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

Notes

Short-cut Method

This method is used when the magnitude of individual observations is large. The use of short-cut method is helpful in the simplification of calculation work.

Let A be any assumed mean. We subtract A from every observation. The difference between an observation and A, i.e., $X_i - A$ is called the deviation of ith observation from A and is denoted by d_i. Thus, we can write ; $d_1 = X_1 - A$, $d_2 = X_2 - A$, ... $d_n = X_n - A$. On adding these deviations and dividing by n we get

$$\frac{\sum d_i}{n} = \frac{\sum (X_i - A)}{n} = \frac{\sum X_i - nA}{n} = \frac{\sum X_i}{n} - A$$

$$\text{or } \bar{d} = \bar{X} - A \qquad \left(\text{Where } \bar{d} = \frac{\sum d_i}{n}\right)$$

$$\text{On rearranging, we get } \bar{X} = A + \bar{d} = A + \frac{\sum d_i}{n}$$

This result can be used for the calculation of $\bar{X}$

Remarks: Theoretically we can select any value as assumed mean. However, for the purpose of simplification of calculation work, the selected value should be as nearer to the value of $\bar{X}$ as possible.

Example: The following figures relate to monthly output of cloth of a factory in a given year:

Months	:	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*
Output (in '000 metres)	:	*80*	*88*	*92*	*84*	*96*	*92*	*96*	*100*	*92*	*94*	*98*	*86*

Calculate the average monthly output.

Solution:

1. *Using Direct Method*

$$\bar{X} = \frac{80 + 88 + 92 + 84 + 96 + 92 + 96 + 100 + 92 + 94 + 98 + 86}{12}$$

= 91.5 ('000 mtrs)

2. *Using Short-cut Method*

Let A = 90.

X_i	80	88	92	84	96	92	96	100	92	94	98	86	*Total*
$d_i = X_i - A$	– 10	– 2	2	– 6	6	2	6	10	2	4	8	– 4	$\sum d_i = 18$

$$\therefore \quad \bar{X} = 90 + \frac{18}{12} = 90 + 1.5 = 91.5 \text{ thousand mtrs}$$

Notes

When Data are in the form of an Ungrouped Frequency Distribution

Let there be n values X_1, X_2, ... X_n out of which X_1 has occurred f_1 times, X_2 has occurred f_2 times, ... X_n has occurred f_n times. Let N be the total frequency, i.e., N = $\sum_{i=1}^{n} f_i$ f_1. Alternatively, this can be written as follows:

Values	X_1	X_2	- - -	X_n	*Total Frequency*
Frequency	f_1	f_2	- - -	f_n	N

Direct Method

The arithmetic mean of these observations using direct method is given by

$$X = \frac{\underbrace{X_1 + X_1 + ... + X_1}_{f_1\ times} + \underbrace{X_2 + ... + ... + X_2 + ...}_{f_2\ times} + \underbrace{... + X_n + ... + X_n}_{f_n\ times}}{f_1 + f_2 + ... + f_n}$$

Since $X_1 + X_1 + ... + X_1$ added f_1 times can also be written f_1X_1. Similarly, by writing other observation in same manner, we have

$$\overline{X} = \frac{f_1X_1 + f_2X_2 + ... + f_nX_n}{f_1 + f_2 + ... + f_n} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i X_i}{N}$$

Short-cut Method

As before, we take the deviations of observations from an arbitrary value A. The deviation of ith observation from A is $d_i = X_i - A$.

Multiplying both sides by f_i we have $f_i\, d_i = f_i\, (X_i - A)$

Taking sum over all the observations

$$\Sigma f_i d_i = \Sigma f_i(X_i - A) = \Sigma f_i X_i - A\Sigma f_i = \Sigma f_i X_i - A.N$$

Dividing both sides by N we have

$$\frac{\Sigma f_i d_i}{N} = \frac{\Sigma f_i X_i}{N} - A = \overline{X} - A \text{ or } \overline{X} = A + \frac{\Sigma f_i d_i}{N} = A + \overline{d}\ .$$

Example: The following is the frequency distribution of age of 670 students of a school. Compute the arithmetic mean of the data.

X (in years)	5	6	7	8	9	10	11	12	13	14
Frequency	25	45	90	165	112	96	81	26	18	12

Notes

Solution:

Direct Method

The computations are shown in the following table:

X	5	6	7	8	9	10	11	12	13	14	Total
f	25	45	90	165	112	96	81	26	18	12	$\sum f = 670$
fX	125	270	630	1320	1008	960	891	312	234	168	$\sum fX = 5918$

$$\overline{X} = \frac{\sum fX}{\sum f} = \frac{5918}{670} = 8.83 \text{ years.}$$

Short-cut Method

The method of computations are shown in the following table:

X	5	6	7	8	9	10	11	12	13	14	*Total*
f	25	45	90	165	112	96	81	26	18	12	670
$d = X - 8$	−3	−2	−1	0	1	2	3	4	5	6	
fd	−75	−90	−90	0	112	192	243	104	90	72	558

$$\therefore \quad \overline{X} = A + \frac{\sum fd}{N} = 8 + \frac{558}{670} = 8 + 0.83 = 8.83 \text{ years.}$$

When Data are in the Form of a Grouped Frequency Distribution

In a grouped frequency distribution, there are classes along with their respective frequencies. Let l_i be the lower limit and u_i be the upper limit of ith class. Further, let the number of classes be n, so that i = 1, 2,.....n. Also let f_i be the frequency of i th class. This distribution can written in tabular form, as shown.

Notes Here u_1 may or may not be equal to l_2, i.e., the upper limit of a class may or may not be equal to the lower limit of its following class.

It may be recalled here that, in a grouped frequency distribution, we only know the number of observations in a particular class interval and not their individual magnitudes. Therefore, to calculate mean, we have to make a fundamental assumption that the observations in a class are uniformly distributed. Under this assumption, the mid-value of a class will be equal to the mean of observations in that class and hence can be taken as their representative. Therefore, if X_i is the mid-value of i th class with frequency f_i, the above assumption implies that there are f_i observations each with magnitude X_i (i = 1 to n).

Remarks: The accuracy of arithmetic mean calculated for a grouped frequency distribution depends upon the validity of the fundamental assumption. This assumption is rarely met in practice. Therefore, we can only get an approximate value of the arithmetic mean of a grouped frequency distribution.

Notes

Example: Calculate arithmetic mean of the following distribution:

Class Intervals	:	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80
Frequency	:	3	8	12	15	18	16	11	5

Solution: Here only short-cut method will be used to calculate arithmetic mean but it can also be calculated by the use of direct-method.

Class Intervals	Mid Values (X)	Frequency (f)	$d = X - 35$	fd
0-10	5	3	−30	−90
10-20	15	8	−20	−160
20-30	25	12	−10	−120
30-40	35	15	0	0
40-50	45	18	10	180
50-60	55	16	20	320
60-70	65	11	30	330
70-80	75	5	40	200
Total		88		660

$$\therefore \qquad \bar{X} = A + \frac{\sum fd}{N} = 35 + \frac{660}{88} = 42.5$$

6.2.2 Weighted Arithmetic Mean

In the computation of simple arithmetic mean, equal importance is given to all the items. But this may not be so in all situations. If all the items are not of equal importance, then simple arithmetic mean will not be a good representative of the given data. Hence, weighing of different items becomes necessary. The weights are assigned to different items depending upon their importance, i.e., more important items are assigned more weight. For example, to calculate mean wage of the workers of a factory, it would be wrong to compute simple arithmetic mean if there are a few workers (say managers) with very high wages while majority of the workers are at low level of wages. The simple arithmetic mean, in such a situation, will give a higher value that cannot be regarded as representative wage for the group. In order that the mean wage gives a realistic picture of the distribution, the wages of managers should be given less importance in its computation. The mean calculated in this manner is called weighted arithmetic mean. The computation of weighted arithmetic is useful in many situations where different items are of unequal importance, e.g., the construction index numbers, computation of standardised death and birth rates, etc.

Formulae for Weighted Arithmetic Mean

Let X_1, X_2 ....., X_n be n values with their respective weights w_1, w_2 ....., w_n. Their weighted arithmetic mean denoted as $\bar{X}_w$ is given by,

1. $$\bar{X}_w = \frac{\sum w_i X_i}{\sum w_i}$$ (Using direct method),

Notes

2. $\bar{X}_w = A + \frac{\sum w_i d_i}{\sum w_i}$ (where di = Xi - A) (Using short-cut method),

3. $\bar{X}_w = A + \frac{\sum w_i u_i}{\sum w_i} \times h$ (where $u_i = \frac{X_i - A}{h}$) (Using step-deviation method)

Remarks: If $\bar{X}$ denotes simple mean and $\bar{X}_w$ denotes the weighted mean of the same data, then

1. $\bar{X} = \bar{X}_w$, when equal weights are assigned to all the items.
2. $\bar{X} > \bar{X}_w$, when items of small magnitude are assigned greater weights and items of large magnitude are assigned lesser weights.
3. $\bar{X} < \bar{X}_w$, when items of small magnitude are assigned lesser weights and items of large magnitude are assigned greater weights.

6.2.3 Properties of Arithmetic Mean

Arithmetic mean of a given data possess the following properties:

1. The sum of deviations of the observations from their arithmetic mean is always zero.

   According to this property, the arithmetic mean serves as a point of balance or a centre of gravity of the distribution; since sum of positive deviations (i.e., deviations of observations which are greater than $\bar{X}$) is equal to the sum of negative deviations (i.e., deviations of observations which are less than $\bar{X}$).

2. The sum of squares of deviations of observations is minimum when taken from their arithmetic mean. Because of this, the mean is sometimes termed as 'least square' measure of central tendency.

3. Arithmetic mean is capable of being treated algebrically.

   This property of arithmetic mean highlights the relationship between $\bar{X}$, $\sum f_i X_i$ and N. According to this property, if any two of the three values are known, the third can be easily computed.

4. If $\bar{X}_1$ and N_1 are the mean and number of observations of a series and $\bar{X}_2$ and N_2 are the corresponding magnitudes of another series, then the mean $\bar{X}$ of the combined series of $N_1 + N_2$ observations is given by $\bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$.

5. If a constant B is added (subtracted) from every observation, the mean of these observations also gets added (subtracted) by it.

6. If every observation is multiplied (divided) by a constant b, the mean of these observations also gets multiplied (divided) by it.

7. If some observations of a series are replaced by some other observations, then the mean of original observations will change by the average change in magnitude of the changed observations.

Notes

Example: Find out the missing item (x) of the following frequency distribution whose arithmetic mean is 11.37.

X :	5	7	(x)	11	13	16	20
f :	2	4	29	54	11	8	4

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{(5\times 2)+(7\times 4)+29x+(11\times 54)+(13\times 11)+(16\times 8)+(20\times 4)}{112}$$

$$11.37 = \frac{10+28+29x+594+143+128+80}{112} \text{ or } 11.37 \times 112 = 983 + 29x$$

$$\therefore \quad x = \frac{290.44}{29} = 10.015 = 10 \text{ (approximately)}$$

Example: The arithmetic mean of 50 items of a series was calculated by a student as 20. However, it was later discovered that an item 25 was misread as 35. Find the correct value of mean.

Solution.

N = 50 and $\bar{X} = 20$ $\therefore$ $SX_i = 50 \times 20 = 1000$

Thus $SX_{i\ (corrected)} = 1000 + 25 - 35 = 990$ and $\bar{X}_{(corrected)} = \frac{990}{50} = 19.8$

Alternatively, using property 7:

$$\bar{X}_{new} = \bar{X}_{old} + \text{ average change in magnitude } = 20 - \frac{10}{50} = 20 - 0.2 = 19.8$$

6.2.4 Merits and Demerits of Arithmetic Mean

Merits

Out of all averages arithmetic mean is the most popular average in statistics because of its merits given below:

1. Arithmetic mean is rigidly defined by an algebraic formula.
2. Calculation of arithmetic mean requires simple knowledge of addition, multiplication and division of numbers and hence, is easy to calculate. It is also simple to understand the meaning of arithmetic mean, e.g., the value per item or per unit, etc.
3. Calculation of arithmetic mean is based on all the observations and hence, it can be regarded as representative of the given data.
4. It is capable of being treated mathematically and hence, is widely used in statistical analysis.
5. Arithmetic mean can be computed even if the detailed distribution is not known but sum of observations and number of observations are known.
6. It is least affected by the fluctuations of sampling.
7. It represents the centre of gravity of the distribution because it balances the magnitudes of observations which are greater and less than it.
8. It provides a good basis for the comparison of two or more distributions.

Notes

Demerits

Although, arithmetic mean satisfies most of the properties of an ideal average, it has certain drawbacks and should be used with care. Some demerits of arithmetic mean are:

1. It can neither be determined by inspection nor by graphical location.
2. Arithmetic mean cannot be computed for a qualitative data; like data on intelligence, honesty, smoking habit, etc.
3. It is too much affected by extreme observations and hence, it does not adequately represent data consisting of some extreme observations.
4. The value of mean obtained for a data may not be an observation of the data and as such it is called a fictitious average.
5. Arithmetic mean cannot be computed when class intervals have open ends. To compute mean, some assumption regarding the width of class intervals is to be made.
6. In the absence of a complete distribution of observations the arithmetic mean may lead to fallacious conclusions. For example, there may be two entirely different distributions with same value of arithmetic mean.
7. Simple arithmetic mean gives greater importance to larger values and lesser importance to smaller values.

Self Assessment

State whether the following statements are true or false:

8. Arithmetic Mean is defined as the sum of squares of observations divided by the number of observations.
9. In case of simple arithmetic mean, equal importance is not given to all the observations.
10. In weighted arithmetic mean, the importance given to various observations is same.
11. In an individual frequency distribution, we only know the number of observations in a particular class interval and not their individual magnitudes.
12. The accuracy of arithmetic mean calculated for a grouped frequency distribution does not depends upon the validity of the fundamental assumption.

6.3 Median

Median of distribution is that value of the variate which divides it into two equal parts. In terms of frequency curve, the ordinate drawn at median divides the area under the curve into two equal parts. Median is a positional average because its value depends upon the position of an item and not on its magnitude.

6.3.1 Determination of Median

When Individual Observations are given

The following steps are involved in the determination of median:

1. The given observations are arranged in either ascending or descending order of magnitude.

Notes

2. Given that there are n observations, the median is given by:

   (a) The size of $\left(\frac{n+1}{2}\right)$th observations, when n is odd.

   (b) The mean of the sizes of $\frac{n}{2}$th and $\left(\frac{n+1}{2}\right)$th observations, when n is even.

Example: Find median of the following observations:

20, 15, 25, 28, 18, 16, 30.

Solution:

Writing the observations in ascending order, we get 15, 16, 18, 20, 25, 28, 30.

Since n = 7, i.e., odd, the median is the size of $\left(\frac{7+1}{2}\right)^{\text{th}}$ i.e., 4th observation.

Hence, median, denoted by Md = 20.

Note: The same value of Md will be obtained by arranging the observations in descending order of magnitude.

Example: Find median of the data : 245, 230, 265, 236, 220, 250.

Solution:

Arranging these observations in ascending order of magnitude, we get

220, 230, 236, 245, 250, 265. Here n = 6, i.e., even.

$\therefore$ Median will be arithmetic mean of the size of $\frac{6}{2}$ th, i.e., 3rd and $\left(\frac{6}{2}+1\right)^{\text{th}}$, i.e., 4th observations.

Hence $M_d = \frac{236+245}{2} = 240.5$

When ungrouped frequency distribution is given

In this case, the data are already arranged in the order of magnitude. Here, cumulative frequency is computed and the median is determined in a manner similar to that of individual observations.

Example: Locate median of the following frequency distribution:

Variable (X)	:	10	11	12	13	14	15	16
Frequency (f)	:	8	15	25	20	12	10	5

Solution:

X	:	10	11	12	13	14	15	16
f	:	8	15	25	20	12	10	5
$c.f.$	:	8	23	48	68	80	90	95

Here N = 95, which is odd. Thus, median is size of $\left[\frac{95+1}{2}\right]^{th}$ i.e., 48th observation. From the table 48th observation is 12, $\therefore$ $M_d = 12$.

Notes

Alternative Method

$\frac{N}{2} = \frac{95}{2} = 47.5.$ Looking at the frequency distribution we note that there are 48 observations which are less than or equal to 12 and there are 72 (i.e., 95 – 23) observations which are greater than or equal to 12. Hence, median is 12.

Locate median of the following frequency distribution:

X	:	0	1	2	3	4	5	6	7
f	:	7	14	18	36	51	54	52	20

Solution:

X	0	1	2	3	4	5	6	7
f	7	14	18	36	51	54	52	20
$c.f.$	7	21	39	75	126	180	232	252

Here N = 252, i.e., even.

$$\text{Now } \frac{N}{2} = \frac{252}{2} = 126 \text{ and } \frac{N}{2} + 1 = 127.$$

$\therefore$ Median is the mean of the size of 126th and 127th observation. From the table we note that 126th observation is 4 and 127th observation is 5.

$$\therefore \; M_d = \frac{4+5}{2} = 4.5$$

Alternative Method

Looking at the frequency distribution we note that there are 126 observations which are less than or equal to 4 and there are 252 – 75 = 177 observations which are greater than or equal to 4.

Similarly, observation 5 also satisfies this criterion. Therefore, median $= \frac{4+5}{2} = 4.5.$

When grouped frequency distribution is given (Interpolation formula)

The determination of median, in this case, will be explained with the help of the following example.

Example: The following table shows the daily sales of 230 footpath sellers of Chandni Chowk:

Sales (in ₹)	:	0-500	500-1000	1000-1500	1500-2000
No. of Sellers	:	12	18	35	42

Sales (in ₹)	:	2000-2500	2500-3000	3000-3500	3500-4000
No. of Sellers	:	50	45	20	8

Locate the median of the above data using

(i) only the less than type ogive, and

(ii) both, the less than and the greater than type ogives.

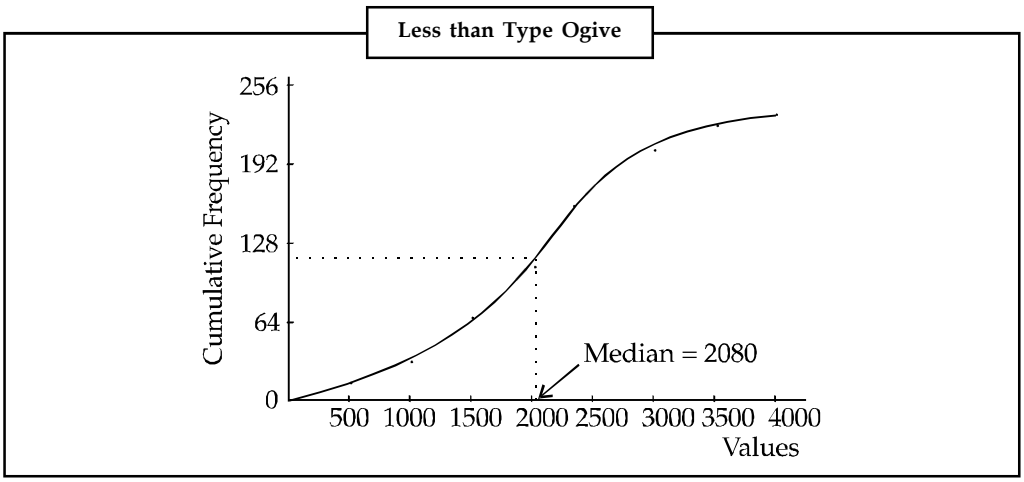
Notes

Solution:

To draw ogives, we need to have a cumulative frequency distribution.

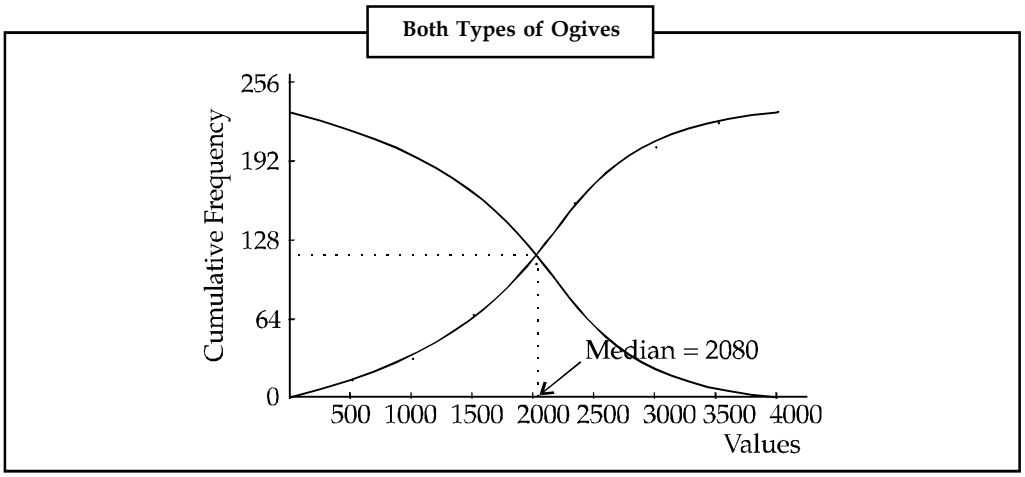
Class Intervals	*Frequency*	*Less than c.f.*	*More than c.f.*
0 - 500	12	12	230
500 - 1000	18	30	218
1000 - 1500	35	65	200
1500 - 2000	42	107	165
2000 - 2500	50	157	123
2500 - 3000	45	202	73
3000 - 3500	20	222	28
3500 - 4000	8	230	8

1. Using the less than type ogive



The value $\frac{N}{2}$ = 115 is marked on the vertical axis and a horizontal line is drawn from this point to meet the ogive at point S. Drop a perpendicular from S. The point at which this meets X- axis is the median.

2. Using both types of ogives



Notes

A perpendicular is dropped from the point of intersection of the two ogives. The point at which it intersects the X-axis gives median. It is obvious from above figures that median = 2080.

6.3.2 Properties of Median

1. It is a positional average.
2. It can be shown that the sum of absolute deviations is minimum when taken from median. This property implies that median is centrally located.

6.3.3 Merits, Demerits and Uses of Median

Merits

1. It is easy to understand and easy to calculate, especially in series of individual observations and ungrouped frequency distributions. In such cases it can even be located by inspection.
2. Median can be determined even when class intervals have open ends or not of equal width.
3. It is not much affected by extreme observations. It is also independent of range or dispersion of the data.
4. Median can also be located graphically.
5. It is centrally located measure of average since the sum of absolute deviation is minimum when taken from median.
6. It is the only suitable average when data are qualitative and it is possible to rank various items according to qualitative characteristics.
7. Median conveys the idea of a typical observation.

Demerits

1. In case of individual observations, the process of location of median requires their arrangement in the order of magnitude which may be a cumbersome task, particularly when the number of observations is very large.
2. It, being a positional average, is not capable of being treated algebraically.
3. In case of individual observations, when the number of observations is even, the median is estimated by taking mean of the two middle-most observations, which is not an actual observation of the given data.
4. It is not based on the magnitudes of all the observations. There may be a situation where different sets of observations give same value of median. For example, the following two different sets of observations, have median equal to 30.

   Set I : 10, 20, 30, 40, 50 and Set II : 15, 25, 30, 60, 90.
5. In comparison to arithmetic mean, it is much affected by the fluctuations of sampling.
6. The formula for the computation of median, in case of grouped frequency distribution, is based on the assumption that the observations in the median class are uniformly distributed. This assumption is rarely met in practice.
7. Since it is not possible to define weighted median like weighted arithmetic mean, this average is not suitable when different items are of unequal importance.

Uses

Notes

1. It is an appropriate measure of central tendency when the characteristics are not measurable but different items are capable of being ranked.
2. Median is used to convey the idea of a typical observation of the given data.
3. Median is the most suitable measure of central tendency when the frequency distribution is skewed. For example, income distribution of the people is generally positively skewed and median is the most suitable measure of average in this case.
4. Median is often computed when quick estimates of average are desired.
5. When the given data has class intervals with open ends, median is preferred as a measure of central tendency since it is not possible to calculate mean in this case.

Self Assessment

Multiple Choice Questions:

13. ..... of distribution is that value of the variate which divides it into two equal parts.

    (a) Mean (b) Median

    (c) Mode (d) Standard deviation

14. Median is a ..... average because its value depends upon the position of an item and not on its magnitude.

    (a) Positional (b) Arithmetic

    (c) Mathematical (d) Commercial

15. Median is often computed when ..... of average are desired

    (a) Estimate (b) Slow estimate

    (c) Quick estimate (d) Negligible estimate

6.4 Other Partition or Positional Measures

Median of a distribution divides it into two equal parts. It is also possible to divide it into more than two equal parts. The values that divide a distribution into more than two equal parts are commonly known as partition values or fractiles. Some important partition values are discussed in the following sections.

6.4.1 Quartiles

The values of a variable that divide a distribution into four equal parts are called quartiles. Since three values are needed to divide a distribution into four parts, there are three quartiles, viz. Q_1, Q_2 and Q_3, known as the first, second and the third quartile respectively.

For a discrete distribution, the first quartile (Q_1) is defined as that value of the variate such that at least 25% of the observations are less than or equal to it and at least 75% of the observations are greater than or equal to it.

For a continuous or grouped frequency distribution, Q_1 is that value of the variate such that the area under the histogram to the left of the ordinate at Q_1 is 25% and the area to its right is 75%.

Notes

The formula for the computation of Q_1 can be written by making suitable changes in the formula of median.

After locating the first quartile class, the formula for Q_1 can be written as follows:

$$Q_1 = L_{Q1} + \frac{\frac{n}{4} - C}{f_{Q_1}} \times h$$

Here, L_{Q_1} is lower limit of the first quartile class, h is its width, f_{Q_1} is its frequency and C is cumulative frequency of classes preceding the first quartile class.

By definition, the second quartile is median of the distribution. The third quartile (Q_3) of a distribution can also be defined in a similar manner.

For a discrete distribution, Q_3 is that value of the variate such that at least 75% of the observations are less than or equal to it and at least 25% of the observations are greater than or equal to it.

For a grouped frequency distribution, Q_3 is that value of the variate such that area under the histogram to the left of the ordinate at Q_3 is 75% and the area to its right is 25%. The formula for computation of Q_3 can be written as $Q_3 = L_{Q_3} + \frac{\left(\frac{3N}{4} - C\right)}{f_{Q_3}} \times h$, where the symbols have their usual meaning.

6.4.2 Deciles

Deciles divide a distribution into 10 equal parts and there are, in all, 9 deciles denoted as D_1, D_2, ... D_9 respectively.

For a discrete distribution, the i th decile D_i is that value of the variate such that at least (10i)% of the observation are less than or equal to it and at least (100 - 10i)% of the observations are greater than or equal to it (i = 1, 2, ... 9).

For a continuous or grouped frequency distribution, D_i is that value of the variate such that the area under the histogram to the left of the ordinate at D_i is (10i)% and the area to its right is (100 - 10i)%. The formula for the i th decile can be written as

$$D_i = L_{D_i} + \frac{\left(\frac{iN}{10} - C\right)}{f_{D_i}} \times h \quad (i = 1, 2, ... 9)$$

6.4.3 Percentiles

Percentiles divide a distribution into 100 equal parts and there are, in all, 99 percentiles denoted as P_1, P_2, ... P_{25}, ... P_{40}, ... P_{60}, ... P_{99} respectively.

For a discrete distribution, the kth percentile P_k is that value of the variate such that at least k% of the observations are less than or equal to it and at least (100 – k)% of the observations are greater than or equal to it.

Notes

For a grouped frequency distribution, P_k is that value of the variate such that the area under the histogram to the left of the ordinate at P_k is k% and the area to its right is (100 – k)% . The formula for the kth percentile can be written as

$$P_k = L_{P_k} + \frac{\left(\frac{kN}{100} - C\right)}{f_{P_k}} \times h, \text{ (k = 1, 2, ... 99)}$$

Example: Locate Median, Q_1, Q_3, D_4, D_7, P_{15}, P_{60} and P_{90} from the following data:

Daily Profit (in Rs) :	75	76	77	78	79	80	81	82	83	84	85
No. of Shops :	15	20	32	35	33	22	20	10	8	3	2

Solution:

First we calculate the cumulative frequencies, as in the following table:

Daily Profit (in Rs)	75	76	77	78	79	80	81	82	83	84	85
No. of Shops (f)	15	20	32	35	33	22	20	10	8	3	2
Less than c.f.	15	35	67	102	135	157	177	187	195	198	200

1. ***Determination of Median:*** Here $\frac{N}{2}$ = 100. From the cumulative frequency column, we note that there are 102 (greater than 50% of the total) observations that are less than or equal to 78 and there are 133 observations that are greater than or equal to 78. Therefore, M_d = ₹ 78.

2. ***Determination of Q_1 and Q_3:*** First we determine $\frac{N}{4}$ which is equal to 50. From the cumulative frequency column, we note that there are 67 (which is greater than 25% of the total) observations that are less than or equal to 77 and there are 165 (which is greater than 75% of the total) observations that are greater than or equal to 77. Therefore, Q_1 = ₹ 77. Similarly, Q_3 = ₹ 80.

3. ***Determination of D_4 and D_7:*** From the cumulative frequency column, we note that there are 102 (greater than 40% of the total) observations that are less than or equal to 78 and there are 133 (greater than 60% of the total) observations that are greater than or equal to 78. Therefore, D_4 = ₹ 78. Similarly, D_7 = ₹ 80.

4. ***Determination of P_{15}, P_{60} and P_{90}:*** From the cumulative frequency column, we note that there are 35 (greater than 15% of the total) observations that are less than or equal to 76 and there are 185 (greater than 85% of the total) observations that are greater than or equal to 76. Therefore, P_{15} = ₹ 76. Similarly, P_{60} = ₹ 79 and P_{90} = ₹ 82.

Example: The following incomplete table gives the number of students in different age groups of a town. If the median of the distribution is 11 years, find out the missing frequencies.

Age Group :	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30	*Total*
No. of Students :	15	125	?	66	?	4	300

Solution:

Let x be the frequency of age group 10 - 15. Then the frequency of the age group 20 - 25 will be 300 – (15 + 125 + x + 66 + 4) = 90 – x.

Notes

Making a cumulative frequency table we have

Age Groups	*No. of Students*	*c.f. (less than)*
0-5	15	15
5-10	125	140
10-15	x	$140+x$
15-20	66	$206+x$
20-25	$90-x$	296
25-30	4	300

Here $\frac{N}{2} = \frac{300}{2} = 150$. Since median is given as 11, the median class is 10 - 15.

Hence, 11 $= 10 + \frac{150-140}{x} \times 5$ or x = 50.

Also, frequency of the age group 20-25 is 90 – 50 = 40.

Self Assessment

State whether the following statements are true or false:

16. The values of a variable that divide a distribution into four equal parts are called quartiles.
17. Q_1, Q_2 and Q_3, known as the first, second and the third quartile respectively.
18. Deciles divide a distribution into 10 equal parts.
19. There are, in all, 9 deciles denoted as D_1, D_2, ... D_9 respectively.
20. Percentiles divide a distribution into 100 equal parts.
21. There are, in all, 99 percentiles denoted as P_1, P_2, ... P_{25}, ... P_{40}, ... P_{60}, ... P_{99} respectively.

6.5 Mode

Mode is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed. In the words of Croxton and Cowden, " The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded the most typical of a series of values." Further, according to A.M. Tuttle, "Mode is the value which has the greatest frequency density in its immediate neighbourhood."

If the frequency distribution is regular, then mode is determined by the value corresponding to maximum frequency. There may be a situation where concentration of observations around a value having maximum frequency is less than the concentration of observations around some other value. In such a situation, mode cannot be determined by the use of maximum frequency criterion. Further, there may be concentration of observations around more than one value of the variable and, accordingly, the distribution is said to be bi-model or multi-model depending upon whether it is around two or more than two values.

The concept of mode, as a measure of central tendency, is preferable to mean and median when it is desired to know the most typical value, e.g., the most common size of shoes, the most common size of a ready-made garment, the most common size of income, the most common size of pocket expenditure of a college student, the most common size of a family in a locality, the most common duration of cure of viral-fever, the most popular candidate in an election, etc.

Notes

6.5.1 Determination of Mode

When data are either in the form of individual observations or in the form of ungrouped frequency distribution

Given individual observations, these are first transformed into an ungrouped frequency distribution. The mode of an ungrouped frequency distribution can be determined in two ways, as given below:

1. By inspection or
2. By method of Grouping

1. ***By inspection:*** When a frequency distribution is fairly regular, then mode is often determined by inspection. It is that value of the variate for which frequency is maximum. By a fairly regular frequency distribution we mean that as the values of the variable increase the corresponding frequencies of these values first increase in a gradual manner and reach a peak at certain value and, finally, start declining gradually in, approximately, the same manner as in case of increase.

Example: Compute mode of the following data:

3, 4, 5, 10, 15, 3, 6, 7, 9, 12, 10, 16, 18,

20, 10, 9, 8, 19, 11, 14, 10, 13, 17, 9, 11

Solution.

Writing this in the form of a frequency distribution, we get

Values :	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequency :	2	1	1	1	1	1	3	4	2	1	1	1	1	1	1	1	1	1

$\therefore$ Mode = 10

Remarks:

1. If the frequency of each possible value of the variable is same, there is no mode.
2. If there are two values having maximum frequency, the distribution is said to be bi-modal.

Example: Determine the mode of the following distribution

X	:	10	11	12	13	14	15	16	17	18	19
f	:	8	15	20	100	98	95	90	75	50	30

Solution:

This distribution is not regular because there is sudden increase in frequency from 20 to 100. Therefore, mode cannot be located by inspection and hence the method of grouping is used. Various steps involved in this method are as follows:

1. Prepare a table consisting of 6 columns in addition to a column for various values of X.
2. In the first column, write the frequencies against various values of X as given in the question.
3. In second column, the sum of frequencies, starting from the top and grouped in twos, are written.

Notes

4. In third column, the sum of frequencies, starting from the second and grouped in twos, are written.
5. In fourth column, the sum of frequencies, starting from the top and grouped in threes are written.
6. In fifth column, the sum of frequencies, starting from the second and grouped in threes are written.
7. In the sixth column, the sum of frequencies, starting from the third and grouped in threes are written.

The highest frequency total in each of the six columns is identified and analysed to determine mode. We apply this method for determining mode of the above example.

X	*f* (1)	(2)	(3)	(4)	(5)	(6)
10	8					
11	15	23				
12	20		35	43		
13	100	120			135	
14	98		198			218
15	95	193		293		
16	90		185		283	
17	75	165				260
18	50		125	215		
19	30	80			155	

Analysis Table

Columns		*V*	*A*	*R*	*I*	*A*	*B*	*L*	*E*	
	10	11	12	13	14	15	16	17	18	19
1				1						
2					1	1				
3				1	1					
4				1	1	1				
5					1	1	1			
6						1	1	1		
Total	0	0	0	3	4	4	2	1	0	0

Since the value 14 and 15 are both repeated maximum number of times in the analysis table, therefore, mode is ill defined. Mode in this case can be approximately located by the use of the following formula, which will be discussed later, in this unit.

Mode = 3 Median – 2 mean

Calculation of Median and Mean

X	10	11	12	13	14	15	16	17	18	19	*Total*
f	8	15	20	100	98	95	90	75	50	30	581
c.f.	8	23	43	143	241	336	426	501	551	581	
fX	80	165	240	1300	1372	1425	1440	1275	900	570	8767

Median = Size of $\left(\frac{581+1}{2}\right)^{th}$, i.e., 291st observation = 15. Mean = $\frac{8767}{581}$ = 15.09

$\therefore$ Mode = 3 × 15 – 2 × 15.09 = 45 – 30.18 = 14.82

Notes

Remarks: If the most repeated values, in the above analysis table, were not adjacent, the distribution would have been bi-modal, i.e., having two modes.

Example: The monthly profits (in ₹) of 100 shops are distributed as follows:

Profit per Shop :	0 - 100	100 - 200	200 - 300	300 - 400	400 - 500	500 - 600
No. of Shops :	12	18	27	20	17	6

Determine the 'modal value' of the distribution graphically and verify the result by calculation.

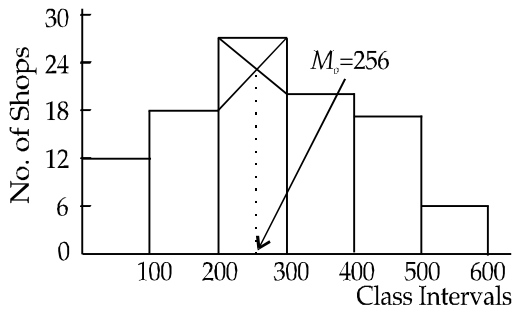
Solution:

Since the distribution is regular, the modal class would be a class having the highest frequency. The modal class, of the given distribution, is 200 - 300.

Graphical Location of Mode

To locate mode we draw a histogram of the given frequency distribution. The mode is located as shown in Figure.

Figure 6.1: Graphical Location of Mode



From the figure, mode = ₹ 256.

Determination of Mode by interpolation formula

Since the modal class is 200 – 300, $L_m = 200$, $\Delta_1 = 27 - 18 = 9$, $\Delta_2 = 27 - 20 = 7$ and h = 100.

$$\therefore \quad M_o = 200 + \frac{9}{9+7} \times 100 = ₹\ 256.25$$

6.5.2 Merits and Demerits of Mode

Merits

1. It is easy to understand and easy to calculate. In many cases it can be located just by inspection.
2. It can be located in situations where the variable is not measurable but categorisation or ranking of observations is possible.
3. Like mean or median, it is not affected by extreme observations. It can be calculated even if these extreme observations are not known.
4. It can be determined even if the distribution has open end classes.
5. It can be located even when the class intervals are of unequal width provided that the width of modal and that of its preceding and following classes are equal.
6. It is a value around which there is more concentration of observations and hence the best representative of the data.

Notes

Demerits

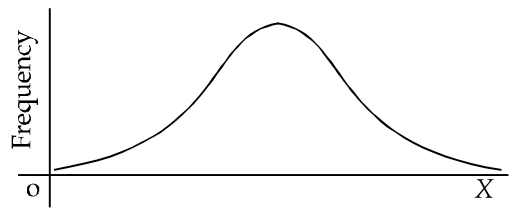
1. It is not based on all the observations.
2. It is not capable of further mathematical treatment.
3. In certain cases mode is not rigidly defined and hence, the important requisite of a good measure of central tendency is not satisfied.
4. It is much affected by the fluctuations of sampling.
5. It is not easy to calculate unless the number of observations is sufficiently large and reveal a marked tendency of concentration around a particular value.
6. It is not suitable when different items of the data are of unequal importance.
7. It is an unstable average because, mode of a distribution, depends upon the choice of width of class intervals.

6.5.3 Relation between Mean, Median and Mode

The relationship between the above measures of central tendency will be interpreted in terms of a continuous frequency curve.

If the number of observations of a frequency distribution are increased gradually, then accordingly, we need to have more number of classes, for approximately the same range of values of the variable, and simultaneously, the width of the corresponding classes would decrease. Consequently, the histogram of the frequency distribution will get transformed into a smooth frequency curve, as shown in Figure 6.2.

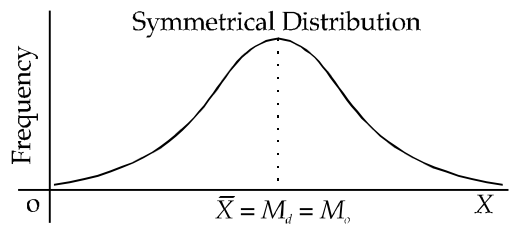
Figure 6.2: Frequency Curve



For a given distribution, the mean is the value of the variable which is the point of balance or centre of gravity of the distribution. The median is the value such that half of the observations are below it and remaining half are above it. In terms of the frequency curve, the total area under the curve is divided into two equal parts by the ordinate at median. Mode of a distribution is a value around which there is maximum concentration of observations and is given by the point at which peak of the curve occurs.

For a symmetrical distribution, all the three measures of central tendency are equal i.e. = M_d = M_o, as shown in Figure 6.3.

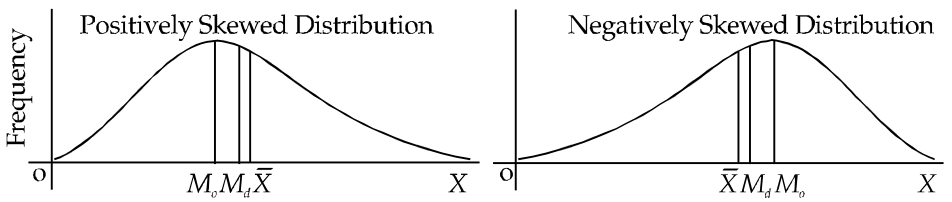
Figure 6.3: Symmetrical Distribution



Notes

Imagine a situation in which the above symmetrical distribution is made asymmetrical or positively (or negatively) skewed (by adding some observations of very high (or very low) magnitudes, so that the right hand (or the left hand) tail of the frequency curve gets elongated. Consequently, the three measures will depart from each other. Since mean takes into account the magnitudes of observations, it would be highly affected. Further, since the total number of observations will also increase, the median would also be affected but to a lesser extent than mean. Finally, there would be no change in the position of mode. More specifically, we shall have $M_o < M_d < \overline{X}$, when skewness is positive and $\overline{X} < M_d < M_o$, when skewness is negative, as shown in Figure 6.4.

Figure 6.4: Positively and Negatively Skewed Distribution



Empirical Relation between Mean, Median and Mode

Empirically, it has been observed that for a moderately skewed distribution, the difference between mean and mode is approximately three times the difference between mean and median, i.e., $\overline{X} - M_o = 3\left(\overline{X} - M_d\right)$

This relation can be used to estimate the value of one of the measures when the values of the other two are known.

Example:

(a) The mean and median of a moderately skewed distribution are 42.2 and 41.9 respectively. Find mode of the distribution.

(b) For a moderately skewed distribution, the median price of men's shoes is ₹ 380 and modal price is ₹ 350. Calculate mean price of shoes.

Solution:

(a) Here, mode will be determined by the use of empirical formula.

$$\overline{X} - M_o = 3\left(\overline{X} - M_d\right) \quad or \quad M_o = 3M_d - 2\overline{X}$$

It is given that $\overline{X}$ = 42.2 and M_d = 41.9

$\therefore$ M_o = 3 × 41.9 - 2 × 42.2 = 125.7 – 84.4 = 41.3

(b) Using the empirical relation, we can write $\overline{X} = \dfrac{3M_d - M_o}{2}$

It is given that M_d = ₹ 380 and M_o = ₹ 350

$\therefore$ $\overline{X} = \dfrac{3 \times 380 - 350}{2}$ = ₹ 395

Notes

Did u know?

3median = mode + 2 mean

mode = 3median – 2 mean

2mean = 3median – mode.

Self Assessment

Fill in the blanks:

22. .....is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed.
23. The concept of mode, as a measure of central tendency, is preferable to mean and median when it is desired to know the .....
24. For a moderately skewed distribution, the difference between mean and mode is approximately ..... the difference between mean and median

6.6 Geometric Mean

The geometric mean of a series of n positive observations is defined as the nth root of their product.

6.6.1 Calculation of Geometric Mean

Individual Series

If there are n observations, X_1, X_2, ... X_n, such that $X_i > 0$ for each i, their geometric mean (GM) is defined as $(X_1, X_2 .....X_n)\frac{1}{n} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$, where the symbol P is used to denote the product of observations.

To evaluate GM, we have to use logarithms. Taking log of both sides we have

$$\log (GM) = \frac{1}{n}\log(X_1.X_2.....X_n)$$

$$= \frac{1}{n}\left[\log X_1 + \log X_2 + \cdots\cdots + \log X_n\right] = \frac{\sum \log X_i}{n}$$

Taking antilog of both sides, we have

$$GM = \text{antilog}\left[\frac{\sum \log X_i}{n}\right].$$

This result shows that the GM of a set of observations is the antilog of the arithmetic mean of their logarithms.

Notes

Example: Calculate geometric mean of the following data:

1, 7, 29, 92, 115 and 375

Solution:

Calculation of Geometric Mean

X	1	7	29	92	115	375	$\sum \log X$
log X	0.0000	0.8451	1.4624	1.9638	2.0607	2.5740	8.9060

$$GM = \text{antilog}\left[\frac{\sum \log X}{n}\right] = \text{antilog}\left[\frac{8.9060}{6}\right] = 30.50$$

GM = antilog = antilog = 30.50

Ungrouped Frequency Distribution

If the data consists of observations X_1, X_2, ... X_n with respective frequencies f_1, f_2, ... f_n, where $\sum_{i=1}^{n} f_i = N$, the geometric mean is given by:

$$\text{GM} = \left[\underbrace{X_1 \,.\, X_1 \ldots\ldots X_1}_{f_1 \text{ times}} \underbrace{X_2 \ldots\ldots X_2}_{f_2 \text{ times}} \ldots\ldots \underbrace{X_n \ldots\ldots X_n}_{f_n \text{ times}}\right]^{\frac{1}{N}} = \left[X_1^{f_1} . X_2^{f_2} \ldots\ldots X_n^{f_n}\right]^{\frac{1}{N}}$$

Taking log of both sides, we have

$$\log (\text{GM}) = \frac{1}{N}\left[\log X_1^{f_1} + \log X_2^{f_2} + \cdots\cdots + \log X_n^{f_n}\right]$$

$$= \frac{1}{N}\left[f_1 \log X_1 + f_2 \log X_2 + \cdots\cdots + f_n \log X_n\right] = \frac{\sum_{i=1}^{n} f_i \log X_i}{N}$$

or GM = antilog $\left(\frac{1}{N}\sum_{i=1}^{n} f_i \log X_i\right)$, which is again equal to the antilog of the arithmetic mean of the logarithm of observations.

Example: Calculate geometric mean of the following distribution:

X	:	5	10	15	20	25	30
f	:	13	18	50	40	10	6

Notes

Solution.

Calculation of GM

X	*f*	*log X*	*f log X*
5	13	0.6990	9.0870
10	18	1.0000	18.0000
15	50	1.1761	58.8050
20	40	1.3010	52.0400
25	10	1.3979	13.9790
30	6	1.4771	8.8626
Total	137		160.7736

$$\therefore \text{ GM} = \text{antilog}\left[\frac{160.7736}{137}\right] = \text{antilog } 1.1735 = 14.91$$

Continuous Frequency Distribution

In case of a continuous frequency distribution, the class intervals are given. Let X_1, X_2, ...X_n denote the mid-values of the first, second ..... nth class interval respectively with corresponding frequencies f_1, f_2, ... f_n, such that Sf_i = N. The formula for calculation of GM is same as the formula used for an ungrouped frequency distribution

$$\text{i.e., GM} = \text{antilog}\left[\frac{\sum f_i \log X_i}{N}\right]$$

Example: Calculate geometric mean of the following distribution:

Class Intervals	:	5 - 15	15 - 25	25 - 35	35 - 45	45 - 55
Frequencies	:	10	22	25	20	8

Solution:

Calculation of GM

Class	*f*	Mid-Value (X)	logX	*f* logX
5-15	10	10	1.0000	10.0000
15-25	22	20	1.3010	28.6227
25-35	25	30	1.4771	36.9280
35-45	20	40	1.6020	32.0412
45-55	8	50	1.6990	13.5918
Total	85			121.1837

$$\text{GM} = \text{antilog}\ \frac{121.1837}{85} = \text{antilog } 1.4257 = 26.65$$

6.6.2 Weighted Geometric Mean

If various observations, X_1, X_2, ...X_n, are not of equal importance in the data, weighted geometric mean is calculated. Weighted GM of the observations X_1, X_2, ...X_n with respective weights as w_1, w_2 .....w_n is given by :

$$\text{GM} = \text{antilog}\left[\frac{\sum w_i \log X_i}{\sum w_i}\right]$$, i.e., weighted geometric mean of observations is equal to the antilog of weighted arithmetic mean of their logarithms.

Notes

Example: Calculate weighted geometric mean of the following data:

Variable (X)	:	5	8	44	160	500
Weights (w)	:	10	9	3	2	1

How does it differ from simple geometric mean?

Solution:

Calculation of weighted and simple GM

X	*Weights* (w)	$\log X$	$w\log X$
5	10	0.6990	6.9900
8	9	0.9031	8.1278
44	3	1.6435	4.9304
160	2	2.2041	4.4082
500	1	2.6990	2.6990
Total	25	8.1487	27.1554

$$\text{Weighted GM} = \text{antilog}\ \frac{27.1554}{25} = \text{antilog } 1.0862 = 12.20$$

$$\text{Simple GM} = \text{antilog}\ \frac{8.1487}{5}\ (n = 5) = \text{antilog } 1.6297 = 42.63$$

Note that the simple GM is greater than the weighted GM because the given system of weights assigns more importance to values having smaller magnitude.

Did u know? Simple GM is greater than the weighted GM because the given system of weights assigns more importance to values having smaller magnitude.

6.6.3 Geometric Mean of the Combined Group

If $G_1, G_2, \ldots\ldots G_k$ are the geometric means of k groups having $n_1, n_2, \ldots\ldots n_k$ observations respectively, the geometric mean G of the combined group consisting of $n_1 + n_2 + \ldots\ldots + n_k$ observations is given by

$$G = \text{antilog}\left[\frac{n_1 \log G_1 + n_2 \log G_2 + \cdots\cdots + n_k \log G_k}{n_1 + n_2 + \cdots\cdots + n_k}\right] \text{antilog}\left[\frac{\sum n_i \log G_i}{\sum n_i}\right]$$

Example: If the geometric means of two groups consisting of 10 and 25 observations are 90.4 and 125.5 respectively, find the geometric mean of all the 35 observations combined into a single group.

Solution.

$$\text{Combined GM} = \text{antilog}\left[\frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}\right]$$

Here $n_1 = 10$, $G_1 = 90.4$ and $n_2 = 25$, $G_2 = 125.5$

Notes

$$\therefore \quad GM = \text{antilog}\left[\frac{10\log 90.4 + 25\log 125.5}{35}\right]$$

$$= \text{antilog}\left[\frac{10\times 1.9562 + 25\times 2.0986}{35}\right] = \text{antilog } 2.0579 = 114.27$$

To determine the average rate of change of price for the entire period when the rate of change of prices for different periods are given

Let P_0 be the price of a commodity in the beginning of the first year. If it increases by k_1 % in the first year, the price at the end of 1st year (or beginning of second year) is given by

$P_1 = P_0 + P_0\frac{k_1}{100} = P_0\left(1+\frac{k_1}{100}\right) = P_0(1 + r_1)$, where $r_1 = \frac{k_1}{100}$ denotes the rate of increase per rupee in first year. Similarly, if the price changes by k_2% in second year, the price at the end of second year is given by

$$P_2 = P_1 + P_1\frac{k_2}{100} = P_1\left(1+\frac{k_2}{100}\right) = P_1(1 + r_2)$$

Replacing the value of P_1 as $P_0(1 + r_1)$ we can write

$$P_2 = P_0(1 + r_1)(1 + r_2)$$

Proceeding in this way, if $100r_i$% is the rate of change of price in the i th year, the price at the end of nth period, P_n, is given by

$$P_n = P_0(1 + r_1)(1 + r_2) \ldots\ldots (1 + r_n) \qquad \ldots. (1)$$

Further, let 100r % per year be the average rate of increase of price that gives the price P_n at the end of n years. Therefore, we can write

$$P_n = P_0(1 + r)(1 + r) \ldots\ldots (1 + r) = P_0(1 + r)^n \qquad \ldots. (2)$$

Equating (1) and (2), we can write

$$(1 + r)^n = (1 + r_1)(1 + r_2) \ldots\ldots (1 + r_n)$$

or $$(1 + r) = \left[(1+r_1)(1+r_2)\ldots\ldots(1+r_n)\right]^{\frac{1}{n}} \qquad \ldots. (3)$$

This shows that (1 + r) is geometric mean of $(1 + r_1)$, $(1 + r_2)$, ... and $(1 + r_n)$.

From (3), we get

$$r = \left[(1+r_1)(1+r_2)\ldots\ldots(1+r_n)\right]^{\frac{1}{n}} - 1 \qquad \ldots. (4)$$

Note: Here r denotes the per unit rate of change. This rate is termed as the rate of increase or the rate of growth if positive and the rate of decrease or the rate of decay if negative.

6.6.4 Average Rate of Growth of Population

The average rate of growth of price, denoted by r in the above section, can also be interpreted as the average rate of growth of population. If P_0 denotes the population in the beginning of the

Notes

period and P_n the population after n years, using Equation (2), we can write the expression for the average rate of change of population per annum as $r = \left(\frac{P_n}{P_0}\right)^{\frac{1}{n}} - 1$.

Similarly, Equation (4), given above, can be used to find the average rate of growth of population when its rates of growth in various years are given.

Remarks: The formulae of price and population changes, considered above, can also be extended to various other situations like growth of money, capital, output, etc.

Example: The population of a country increased from 2,00,000 to 2,40,000 within a period of 10 years. Find the average rate of growth of population per year.

Solution:

Let r be the average rate of growth of population per year for the period of 10 years. Let P_0 be initial and P_{10} be the final population for this period.

We are given P_0 = 2,00,000 and P_{10} = 2,40,000.

$$\therefore \qquad r = \left(\frac{P_{10}}{P_0}\right)^{\frac{1}{10}} - 1 = \left(\frac{2,40,000}{2,00,000}\right)^{\frac{1}{10}} - 1$$

Now
$$\left(\frac{24}{20}\right)^{\frac{1}{10}} = \text{antilog}\left[\frac{1}{10}(\log 24 - \log 20)\right]$$

$$= \text{anti}\log\left[\frac{1}{10}(1.3802 - 1.3010)\right] = \text{anti}\log(0.0079) = 1.018$$

Thus, r = 1.018 - 1 = 0.018.

Hence, the percentage rate of growth = 0.018 × 100 = 1.8% p. a.

6.6.5 Suitability of Geometric Mean for Averaging Ratios

It will be shown here that the geometric mean is more appropriate than arithmetic mean while averaging ratios.

Let there be two values of each of the variables x and y, as given below:

x	y	$Ratio\left(\frac{x}{y}\right)$	$Ratio\left(\frac{y}{x}\right)$
40	60	2/3	3/2
20	80	1/4	4

Now AM of (x/y) ratios = $\frac{\frac{2}{3}+\frac{1}{4}}{2} = \frac{11}{24}$ and the AM of (y/x) ratios = $\frac{\frac{3}{2}+4}{2} = \frac{11}{4}$. We note that their product is not equal to unity.

However, the product of their respective geometric means, i.e., $\frac{1}{\sqrt{6}}$ and $\sqrt{6}$, is equal to unity.

Since it is desirable that a method of average should be independent of the way in which a ratio is expressed, it seems reasonable to regard geometric mean as more appropriate than arithmetic mean while averaging ratios.

Notes

6.6.6 Properties of Geometric Mean

1. As in case of arithmetic mean, the sum of deviations of logarithms of values from the log GM is equal to zero.

   This property implies that the product of the ratios of GM to each observation, that is less than it, is equal to the product the ratios of each observation to GM that is greater than it. For example, if the observations are 5, 25, 125 and 625, their GM = 55.9. The above property implies that

$$\frac{55.9}{5} \times \frac{55.9}{25} = \frac{125}{55.9} \times \frac{625}{55.9}$$

2. Similar to the arithmetic mean, where the sum of observations remains unaltered if each observation is replaced by their AM, the product of observations remains unaltered if each observation is replaced by their GM.

6.6.7 Merits, Demerits and Uses of Geometric Mean

Merits

1. It is a rigidly defined average.
2. It is based on all the observations.
3. It is capable of mathematical treatment. If any two out of the three values, i.e., (i) product of observations, (ii) GM of observations and (iii) number of observations, are known, the third can be calculated.
4. In contrast to AM, it is less affected by extreme observations.
5. It gives more weights to smaller observations and vice-versa.

Demerits

1. It is not very easy to calculate and hence not very popular.
2. Like AM, it may be a value which does not exist in the set of given observations.

Uses

1. It is most suitable for averaging ratios and exponential rates of changes.
2. It is used in the construction of index numbers.
3. It is often used to study certain social or economic phenomena.

Task Calculate AM, GM and HM of first five multiples of 3.Which is greatest? Which is smallest?

Caution Geometric mean cannot be calculated if any observation is zero or negative.

Notes

Self Assessment

State whether the following statements are true or false:

25. The geometric mean of a series of n positive observations is defined as the nth root of their product.

26. In case of arithmetic mean, the sum of deviations of logarithms of values from the log GM is equal to zero.

6.7 Harmonic Mean

The harmonic mean of n observations, none of which is zero, is defined as the reciprocal of the arithmetic mean of their reciprocals.

6.7.1 Calculation of Harmonic Mean

Individual Series

If there are n observations X_1, X_2, ... X_n, their harmonic mean is defined as

$$HM = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots\cdots + \frac{1}{X_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

Example: Obtain harmonic mean of 15, 18, 23, 25 and 30.

Solution:

$$HM = \frac{5}{\frac{1}{15} + \frac{1}{18} + \frac{1}{23} + \frac{1}{25} + \frac{1}{30}} = \frac{5}{0.239} = 20.92$$

Ungrouped Frequency Distribution

For ungrouped data, i.e., each X_1, X_2, ... X_n, occur with respective frequency f_1, f_2 ..... f_n, where $\Sigma f_i = N$ is total frequency, the arithmetic mean of the reciprocals of observations is given by $\frac{1}{N}\sum_{i=1}^{n}\frac{f_i}{X_i}$.

Thus, $$HM = \frac{N}{\sum \frac{f_i}{X_i}}$$

Example: Calculate harmonic mean of the following data:

X	:	10	11	12	13	14
f	:	5	8	10	9	6

Notes

Solution:

Calculation of Harmonic Mean

X	10	11	12	13	14	*Total*
Frequency (f)	5	8	10	9	6	38
$f \times \frac{1}{X}$	0.5000	0.7273	0.8333	0.6923	0.4286	3.1815

$$\therefore \quad HM = \frac{38}{3.1815} = 11.94$$

Continuous Frequency Distribution

In case of a continuous frequency distribution, the class intervals are given. The mid-values of the first, second ..... nth classes are denoted by X_1, X_2, ... X_n. The formula for the harmonic mean is same, as given in (b) above.

Example: Find the harmonic mean of the following distribution :

Class Intervals :	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80
Frequency :	5	8	11	21	35	30	22	18

Solution.

Calculation of Harmonic Mean

Class Intervals	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	*Total*
Frequency (f)	5	8	11	21	35	30	22	18	150
Mid-Values (X)	5	15	25	35	45	55	65	75	
$\frac{f}{X}$	1.0000	0.5333	0.4400	0.6000	0.7778	0.5455	0.3385	0.2400	4.4751

$$\therefore \text{ HM} = \frac{150}{4.4751} = 33.52$$

6.7.2 Weighted Harmonic Mean

If X_1, X_2, ... X_n are n observations with weights w_1,w_2, ... w_n respectively, their weighted harmonic mean is defined as follows :

$$\text{HM} = \frac{\sum w_i}{\sum \frac{w_i}{X_i}}$$

Example: A train travels 50 kms at a speed of 40 kms/hour, 60 kms at a speed of 50 kms/hour and 40 kms at a speed of 60 kms/hour. Calculate the weighted harmonic mean of the speed of the train taking distances travelled as weights. Verify that this harmonic mean represents an appropriate average of the speed of train.

Notes

Solution:

$$\text{HM} = \frac{\sum w_i}{\sum \frac{w_i}{X_i}} = \frac{150}{\frac{50}{40}+\frac{60}{50}+\frac{40}{60}} = \frac{150}{1.25+1.20+0.67} \quad \text{.... (1)}$$

= 48.13 kms/hour

Verification: Average speed $= \frac{\text{Total distance travelled}}{\text{Total time taken}}$

We note that the numerator of Equation (1) gives the total distance travelled by train. Further, its denominator represents total time taken by the train in travelling 150 kms, since $\frac{50}{40}$ is time taken by the train in travelling 50 kms at a speed of 40 kms/hour. Similarly $\frac{60}{50}$ and $\frac{40}{60}$ are time taken by the train in travelling 60 kms and 40 kms at the speeds of 50 kms./hour and 60 kms/hour respectively. Hence, weighted harmonic mean is most appropriate average in this case.

Example: Ram goes from his house to office on a cycle at a speed of 12 kms/hour and returns at a speed of 14 kms/hour. Find his average speed.

Solution:

Since the distances of travel at various speeds are equal, the average speed of Ram will be given by the simple harmonic mean of the given speeds.

$$\text{Average speed} = \frac{2}{\frac{1}{12}+\frac{1}{14}} = \frac{2}{0.1547} = 12.92 \text{ kms/hour}$$

6.7.3 Merits and Demerits of Harmonic Mean

Merits

1. It is a rigidly defined average.
2. It is based on all the observations.
3. It gives less weight to large items and vice-versa.
4. It is capable of further mathematical treatment.
5. It is suitable in computing average rate under certain conditions.

Demerits

1. It is not easy to compute and is difficult to understand.
2. It may not be an actual item of the given observations.
3. It cannot be calculated if one or more observations are equal to zero.
4. It may not be representative of the data if small observations are given correspondingly small weights.

Notes

Did u know? If all the observations of a variable are same, all the three measures of central tendency coincide, i.e., AM = GM = HM. Otherwise, we have AM > GM > HM.

Notes

Quadratic Mean

Quadratic mean is the square root of the arithmetic mean of squares of observations.

If X_1, X_2 ..... X_n are n observations, their quadratic mean is given by

$$QM = \sqrt{\frac{X_1^2 + X_2^2 + \cdots\cdots + X_n^2}{n}} = \sqrt{\frac{\sum X_i^2}{n}}$$

Similarly, the QM of observations X_1, X_2 ..... X_n with their respective frequencies as f_1, f_2 ..... f_n is given by QM $= \sqrt{\frac{\sum f_i X_i^2}{N}}$, where N = $\sum f_i$.

Moving Average

This is a special type of average used to eliminate periodic fluctuations from the time series data.

Progressive Average

A progressive average is a cumulative average which is computed by taking all the available figures in each succeeding years. The average for different periods are obtained as shown below:

$$X_1, \frac{X_1 + X_2}{2}, \frac{X_1 + X_2 + X_3}{3}, \cdots\cdots \text{etc.}$$

This average is often used in the early years of a business.

Composite Average

A composite average is an average of various other averages. If for example, $\overline{X}_1, \overline{X}_2, \ldots\ldots \overline{X}_k$ are the arithmetic means of k series, their composite average $= \frac{\overline{X}_1 + \overline{X}_2 + \ldots\ldots + \overline{X}_k}{k}$.

Self Assessment

Fill in the blanks:

27. The harmonic mean of n observations, none of which is..... , is defined as the reciprocal of the arithmetic mean of their reciprocals.

28. If all the observations of a variable are same, all the three measures of central tendency coincide, i.e., AM = GM = HM. Otherwise, we have .....

Notes

6.8 Summary

- Summarization of the data is a necessary function of any statistical analysis.
- Average is a value which is typical or representative of a set of data.
- Arithmetic Mean is defined as the sum of observations divided by the number of observations.
- In order that the mean wage gives a realistic picture of the distribution, the wages of managers should be given less importance in its computation. The mean calculated in this manner is called weighted arithmetic mean.
- If a constant B is added (subtracted) from every observation, the mean of these observations also gets added (subtracted) by it.
- If every observation is multiplied (divided) by a constant b, the mean of these observations also gets multiplied (divided) by it.
- If some observations of a series are replaced by some other observations, then the mean of original observations will change by the average change in magnitude of the changed observations.
- Arithmetic mean is rigidly defined by an algebraic formula.
- Median of distribution is that value of the variate which divides it into two equal parts.
- Median conveys the idea of a typical observation
- The values of a variable that divide a distribution into four equal parts are called quartiles
- Deciles divide a distribution into 10 equal parts and there are, in all, 9 deciles denoted as D_1, D_2, ... D_9 respectively.
- Percentiles divide a distribution into 100 equal parts and there are, in all, 99 percentiles denoted as P_1, P_2, ... P_{25}, ... P_{40}, ... P_{60}, ... P_{99} respectively.
- Mode is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed.
- For a moderately skewed distribution, the difference between mean and mode is approximately three times the difference between mean and median, i.e.,
- The geometric mean of a series of n positive observations is defined as the nth root of their product.
- The harmonic mean of n observations, none of which is zero, is defined as the reciprocal of the arithmetic mean of their reciprocals.

6.9 Keywords

Average: An average is a single value within the range of the data that is used to represent all the values in the series.

Arithmetic Mean: Arithmetic Mean is defined as the sum of observations divided by the number of observations.

Deciles: Deciles divide a distribution into 10 equal parts and there are, in all, 9 deciles denoted as D_1, D_2, ... D_9 respectively.

Notes

Geometric Mean: The geometric mean of a series of n positive observations is defined as the nth root of their product.

Harmonic Mean: The harmonic mean of n observations, none of which is zero, is defined as the reciprocal of the arithmetic mean of their reciprocals

Measure of Central Tendency: A measure of central tendency is a typical value around which other figures congregate.

Measure of Central Value: Since an average is somewhere within the range of data it is sometimes called a measure of central value.

Median: Median of distribution is that value of the variate which divides it into two equal parts.

Mode: Mode is that value of the variate which occurs maximum number of times in a distribution and around which other items are densely distributed.

Partition Values: The values that divide a distribution into more than two equal parts are commonly known as partition values or fractiles.

Percentiles: Percentiles divide a distribution into 100 equal parts and there are, in all, 99 percentiles denoted as P_1, P_2, ... P_{25}, ... P_{40}, ... P_{60}, ... P_{99} respectively.

Quartiles: The values of a variable that divide a distribution into four equal parts are called quartiles.

Weighted Arithmetic Mean: Weights are assigned to different items depending upon their importance, i.e., more important items are assigned more weight.

Weighted Geometric Mean: Weighted geometric mean of observations is equal to the antilog of weighted arithmetic mean of their logarithms.

6.8 Review Questions

1. What are the functions of an average? Discuss the relative merits and demerits of various types of statistical averages.
2. Give the essential requisites of a measure of 'Central Tendency'. Under what circumstances would a geometric mean or a harmonic mean be more appropriate than arithmetic mean?
3. Compute arithmetic mean of the following series:

Marks :	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
No. of Students :	12	18	27	20	17	6

4. Calculate arithmetic mean of the following data:

Mid - Values :	10	12	14	16	18	20
Frequency :	3	7	12	18	10	5

5. Find out the missing frequency in the following distribution with mean equal to 30.

Class :	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
Frequency :	5	6	10	?	13

6. A distribution consists of three components each with total frequency of 200, 250 and 300 and with means of 25, 10 and 15 respectively. Find out the mean of the combined distribution.
7. The mean of a certain number of items is 20. If an observation 25 is added to the data, the mean becomes 21. Find the number of items in the original data.

Notes

8. The mean age of a combined group of men and women is 30 years. If the mean age of the men's group is 32 years and that for the women's group is 27 years, find the percentage of men and women in the combined group.

9. Locate median of the following data:

   65, 85, 55, 75, 96, 76, 65, 60, 40, 85, 80, 125, 115, 40

10. In a class of 16 students, the following are the marks obtained by them in statistics. Find out the lower quartile, upper quartile, seventh decile and thirty fifth percentile.

S.No.	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Marks	:	5	12	17	23	28	31	37	41	42	49	54	58	65	68	17	77

11. Find out median from the following:

No. of Workers	:	1-5	6-10	11-15	16-20	21-25
No. of factories	:	3	8	13	11	5

12. Find median from the following distribution:

X	:	1	2	3	4	5-9	10-14	15-19	20-25
f	:	5	10	16	20	30	15	8	6

13. Determine the mode of the following data:

    58, 60, 31, 62, 48, 37, 78, 43, 65, 48

14. Find geometric mean from the following daily income (in ₹) of 10 families:

    85, 70, 15, 75, 500, 8, 45, 250, 40 and 36.

15. Calculate geometric mean of the following distribution:

Marks (less than)	:	10	20	30	40	50
No. of Students	:	12	27	72	93	100

16. The value of a machine depreciates at a constant rate from the cost price of ₹ 1,000 to the scrap value of ₹ 100 in ten years. Find the annual rate of depreciation and the value of the machine at the end of one, two, three years.

17. The price of a commodity increased by 12% in 1986, by 30% in 1987 and by 15% in 1988. Calculate the average increase of price per year.

18. The population of a city was 30 lakh in 1981 which increased to 45 lakh in 1991. Determine the rate of growth of population per annum. If the same growth continues, what will be the population of the city in 1995.

19. The value of a machine depreciated by 30% in 1st year, 13% in 2nd year and by 5% in each of the following three years. Determine the average rate of depreciation for the entire period.

20. The geometric means of three groups consisting of 15, 20 and 23 observations are 14.5, 30.2 and 28.8 respectively. Find geometric mean of the combined group.

21. A sum of money was invested for 3 years. The rates of interest in the first, second and third year were 10%, 12% and 14% respectively. Determine the average rate of interest per annum.

22. The weighted geometric mean of four numbers 8, 25, 17 and 30 is 15.3. If the weights of first three numbers are 5, 3 and 4 respectively, find the weight of the fourth number.

Notes

23. The annual rates of growth of output of a factory in five years are 5.0, 6.5, 4.5, 8.5 and 7.5 percent respectively. What is the compound rate of growth of output per annum for the period?

24. Premier Automobiles Ltd. does statistical analysis for an automobile racing team. Here are the fuel consumption figures in Kilometer per litre for the team's cars in the recent races.

4.77	6.11	6.11	5.05	5.99	4.91	5.27	6.01
5.75	4.89	6.05	5.22	6.02	5.24	6.11	5.02

(a) Calculate the median fuel consumption.

(b) Calculate the mean fuel consumption.

(c) Group the given data into equally sized classes. What is the fuel consumption value of the modal classes.

(d) Which of the three measures of central tendency is best for Allison to use when she orders fuel? Explain.

Answers: Self Assessment

1.	Summarization	2.	tables, frequency distributions
3.	measure of central tendency	4.	average
5.	congregate	6.	averages
7.	Mathematical	8.	False
9.	False	10.	False
11.	False	12.	False
13.	(b)	14.	(a)
15.	(c)	16.	True
17.	True	18.	True
19.	True	20.	True
21.	True	22.	Mode
23.	most typical value	24.	three times
25.	True	26.	True
27.	zero	28.	AM > GM > HM.

6.9 Further Readings

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Bhardwaj R.S., *Business Statistics*, Excel Books.

Garrett H.E. (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Notes

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, Mc Graw Hill Book Company, New York.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hannagan T.J. (1982), *Mastering Statistics*, The Macmillan Press Ltd., Surrey.

Hooda R. P., *Statistics for Business and Economics*, Macmillan India, Delhi, 2008.

Jaeger R.M (1983), *Statistics: A Spectator Sport*, Sage Publications India Pvt. Ltd., New Delhi.

Lindgren B.W. (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Selvaraj R., Loganathan, C. *Quantitative Methods in Management.*

Sharma J.K., *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutta.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Online links

http://www.mathsisfun.com/data/histograms.html

http://www.mathsisfun.com/data/mean-machine.html

http://www.mathsisfun.com/mode.html

http://www.mathsisfun.com/median.html

http://www.mathsisfun.com/data/central-measures.html

http://stattrek.com/lesson3/centraltendency.aspx

http://www.icai.org/resource_file/16796CENTRAL-TEN.pdf

http://www.icoachmath.com/math_dictionary/measures_of_central_tendency.html

Notes

Unit 7: Measures of Dispersion

CONTENTS

Objectives

Introduction

7.1 Definitions

7.2 Objectives of Measuring Dispersion

7.3 Characteristics of a Good Measure of Dispersion

7.4 Measures of Dispersion

7.5 Range

7.5.1 Merits and Demerits of Range

7.5.2 Uses of Range

7.6 Interquartile Range

7.6.1 Interpercentile Range

7.6.2 Quartile Deviation or Semi-Interquartile Range

7.6.3 Merits and Demerits of Quartile Deviation

7.7 Mean Deviation or Average Deviation

7.7.1 Calculation of Mean Deviation

7.7.2 Merits and Demerits of Mean Deviation

7.8 Standard Deviation

7.8.1 Calculation of Standard Deviation

7.8.2 Coefficient of Variation

7.8.3 Properties of Standard Deviation

7.8.4 Merits, Demerits and Uses of Standard Deviation

7.8.5 Skewness

7.8.6 Graphical Measure of Dispersion

7.8.7 Empirical relation among various measures of dispersions

7.9 Summary

7.10 Keywords

7.11 Review Questions

7.12 Further Readings

Objectives

After studying this unit, you will be able to:

- Define the term dispersion and range
- Discuss the objectives and characteristics of a good measure of dispersion

Notes

- State the merits and demerits of mean deviation
- Explain the concept of standard deviation
- Describe coefficient of variation

Introduction

A measure of central tendency summarizes the distribution of a variable into a single figure which can be regarded as its representative. This measure alone, however, is not sufficient to describe a distribution because there may be a situation where two or more different distributions have the same central value. Conversely, it is possible that the pattern of distribution in two or more situations is same but the values of their central tendency are different. Hence, it is necessary to define some additional summary measures to adequately represent the characteristics of a distribution. One such measure is known as the measure of dispersion or the measure of variation.

7.1 Definitions

The concept of dispersion is related to the extent of scatter or variability in observations. The variability, in an observation, is often measured as its deviation from a central value. A suitable average of all such deviations is called the measure of dispersion.

Some important definitions of dispersion are given below:

"Dispersion is the measure of variation of the items." – *A.L. Bowley*

"Dispersion is the measure of extent to which individual items vary." – *L.R. Connor*

"The measure of the scatteredness of the mass of figures in a series about an average is called the measure of variation or dispersion." – *Simpson and Kafka*

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

Did u know?

1. Measures of central tendency are known as the averages of first order.
2. Measures of dispersion are known as the averages of second order.

Self Assessment

Fill in the blanks:

1. A measure of central tendency summarizes the distribution of a variable into a single figure which can be regarded as its .....
2. It is possible that the pattern of distribution in two or more situations is ..... but the values of their central tendency are .....
3. The concept of .....is related to the extent of scatter or variability in observations.
4. Measures of central tendency are known as the .....
5. Dispersion are also known as the .....

Notes

7.2 Objectives of Measuring Dispersion

The main objectives of measuring dispersion of a distribution are:

1. ***To test reliability of an average:*** A measure of dispersion can be used to test the reliability of an average. A low value of dispersion implies that there is greater degree of homogeneity among various items and, consequently, their average can be taken as more reliable or representative of the distribution.
2. ***To compare the extent of variability in two or more distributions:*** The extent of variability in two or more distributions can be compared by computing their respective dispersions. A distribution having lower value of dispersion is said to be more uniform or consistent.
3. ***To facilitate the computations of other statistical measures:*** Measures of dispersions are used in computations of various important statistical measures like correlation, regression, test statistics, confidence intervals, control limits, etc.
4. ***To serve as the basis for control of variations:*** The main objective of computing a measure of dispersion is to know whether the given observations are uniform or not. This knowledge may be utilised in many ways. In the words of Spurr and Bonini, "In matters of health, variations in body temperature, pulse beat and blood pressure are basic guides to diagnosis. Prescribed treatment is designed to control their variations. In industrial production, efficient operation requires control of quality variations, the causes of which are sought through inspection and quality control programs". The extent of inequalities of income and wealth in any society may help in the selection of an appropriate policy to control their variations.

Self Assessment

State whether the following statements are true or false:

6. A measure of dispersion can be used to test the reliability of an average.
7. A high value of dispersion implies that there is greater degree of homogeneity among various items.
8. The extent of variability in two or more distributions can be compared by computing their respective dispersions.
9. A distribution having lower value of dispersion is said to be more uniform or consistent.
10. Measures of dispersions are used in computations of various important statistical measures like correlation, regression, test statistics, confidence intervals, control limits, etc.

7.3 Characteristics of a Good Measure of Dispersion

Like the characteristics of a measure of central tendency, a good measure of dispersion should posses the following characteristics:

1. It should be easy to calculate.
2. It should be easy to understand.
3. It should be rigidly defined.
4. It should be based on all the observations.
5. It should be capable of further mathematical treatment.

Notes

6. It should not be unduly affected by extreme observations.
7. It should not be much affected by the fluctuations of sampling.

Self Assessment

Fill in the blanks:

11. A good measure of dispersion should be .....defined.
12. A good measure of dispersion should be ..... all the observations.
13. A good measure of dispersion should not be unduly affected by .....
14. A good measure of dispersion should not be much affected by the ..... of sampling.

7.4 Measures of Dispersion

Various measures of dispersion can be classified into two broad categories:

1. The measures which express the spread of observations in terms of distance between the values of selected observations. These are also termed as distance measures, e.g., range, interquartile range, interpercentile range, etc.
2. The measures which express the spread of observations in terms of the average of deviations of observations from some central value. These are also termed as the averages of second order, e.g., mean deviation, standard deviation, etc.

   The following are some important measures of dispersion

   (a) Range

   (b) Inter-Quartile Range

   (c) Mean Deviation

   (d) Standard Deviation

Self Assessment

State whether the following statements are true or false:

15. The measures which express the spread of observations in terms of distance between the values of selected observations are termed as Kilometres measures.
16. The measures which express the spread of observations in terms of the average of deviations of observations from some central value. These are also termed as the averages of second order, e.g., mean deviation, standard deviation, etc.
17. An absolute measure of dispersion is expressed in terms of the units of measurement of the variable.
18. A relative measure of dispersion, popularly known as coefficient of dispersion, is expressed as a pure number, independent of the units of measurement of the variable.

Notes

7.5 Range

The range of a distribution is the difference between its two extreme observations, i.e., the difference between the largest and smallest observations. Symbolically, R = L – S where R denotes range, L and S denote largest and smallest observations, respectively. R is the absolute measure of range. A relative measure of range, also termed as the coefficient of range, is defined as:

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

Example: Find range and coefficient of range for each of the following data:

1. Weekly wages of 10 workers of a factory are:

   310, 350, 420, 105, 115, 290, 245, 450, 300, 375.

2. The distribution of marks obtained by 100 students:

Marks	:	0-10	10-20	20-30	30-40	40-50
No. of Students	:	6	14	21	20	18
Marks	:	50-60	60-70	70-80	80-90	90-100
No. of Students	:	10	5	3	2	1

3. The age distribution of 60 school going children:

Age (in years)	:	5-7	8-10	11-13	14-16	17-19
Frequency	:	20	18	10	8	4

Solution:

1. Range = 450 – 105 = ₹ 345

   Coefficient of Range = $\frac{450-105}{450+105} = 0.62.$

2. Range = 100 – 0 = 100 marks

   Coefficient of Range = $\frac{100-0}{100+0} = 1.$

3. Range = 19 – 5 = 12 Years

   Coefficient of Range = $\frac{19-5}{19+5} = 0.583$

7.5.1 Merits and Demerits of Range

Merits

1. It is easy to understand and easy to calculate.
2. It gives a quick measure of variability.

Notes

Demerits

1. It is not based on all the observations.
2. It is very much affected by extreme observations.
3. It only gives rough idea of spread of observations.
4. It does not give any idea about the pattern of the distribution. There can be two distribution with the same range but different patterns of distribution.
5. It is very much affected by fluctuations of sampling.
6. It is not capable of being treated mathematically.
7. It cannot be calculated for a distribution with open ends.

7.5.2 Uses of Range

Inspite of many serious demerits, it is useful in the following situations:

1. It is used in the preparation of control charts for controlling the quality of manufactured items.
2. It is also used in the study of fluctuations of, say, price of a commodity, temperature of a patient, amount of rainfall in a given period, etc.

Self Assessment

Fill in the blanks:

19. The ..... of a distribution is the difference between its two extreme observations.
20. Symbolically, R = L – S where R denotes range, L and S denote ..... observations, respectively.
21. A relative measure of range, also termed as the .....

7.6 Interquartile Range

Interquartile Range is an absolute measure of dispersion given by the difference between third quartile (Q_3) and first quartile (Q_1)

Symbolically, Interquartile range = $Q_3 - Q_1$.

7.6.1 Interpercentile Range

The difficulty of extreme observations can also be tackled by the use of interpercentile range or simply percentile range.

Symbolically, percentile range = $P_{(100 - i)} - P_i$ (i < 50).

This measure excludes i% of the observations at each end of the distribution and is a range of the middle (100 – 2i)% of the observations.

Normally, a percentile range corresponding to i = 10, i.e., $P_{90} - P_{10}$ is used. Since $Q_1 = P_{25}$ and $Q_3 = P_{75}$, therefore, interquartile range is also a percentile range.

Notes

Example: Determine the interquartile range and percentile range of the following distribution:

Class Intervals	:	11 – 13	13 – 15	15 – 17	17 – 19	19 – 21	21 – 23	23 – 25
Frequency	:	8	10	15	20	12	11	4

Solution:

Class Intervals	*Frequency*	*Less than c.f.*
11 – 13	8	8
13 – 15	10	18
15 – 17	15	33
17 – 19	20	53
19 – 21	12	65
21 – 23	11	76
23 – 25	4	80

1. Calculation of Interquartile Range

   Calculation of Q_1

   Since $\frac{N}{4} = \frac{80}{4} = 20,$ the first quartile class is 15 - 17

   $\therefore\ l_{Q_1} = 15, f_{Q_1} = 15$, h = 2 and C = 18

   Hence, $Q_1 = 15 + \frac{20-18}{15} \times 2 = 15.27$

   Calculation of Q_3

   Since $\frac{3N}{4} = \frac{3 \times 80}{4} = 60,$ the third quartile class is 19 - 21

   $\therefore\ l_{Q_3} = 19, f_{Q_3} = 12$, h = 2 and C = 53

   Hence, $Q_3 = 19 + \frac{60-53}{12} \times 2 = 20.17$

   Thus, the interquartile range = 20.17 – 15.27 = 4.90

2. Calculation of Percentile Range

   Calculation of P_{10}

   Since, $\frac{10N}{100} = \frac{10 \times 80}{100} = 8,$ P_{10} lies in the class interval 11 – 13

   $\therefore\ l_{P_{10}} = 11, f_{P_{10}} = 8$, h = 2 and C = 0

   Hence, $P_{10} = 11 + \frac{8-0}{8} \times 2 = 13$

Notes

Calculation of P_{90}

Since $\frac{90N}{100} = \frac{90 \times 80}{100} = 72$, P_{90} lies in the class interval 21 - 23

$\therefore$ $l_{P_{90}} = 21, f_{P_{90}} = 11$, h = 2 and C = 65

Hence, $P_{90} = 21 + \frac{72-65}{11} \times 2 = 22.27$

Thus, the percentile range = $P_{90} - P_{10}$ = 22.27 – 13.0 = 9.27.

7.6.2 Quartile Deviation or Semi-Interquartile Range

Half of the interquartile range is called the quartile deviation or semi-interquartile range. Symbolically, $Q.D. = \frac{Q_3 - Q_1}{2}$.

The value of Q.D. gives the average magnitude by which the two quartiles deviate from median.

If the distribution is approximately symmetrical, then $M_d \pm$ Q.D. will include about 50% of the observations and, thus, we can write $Q_1 = M_d$ – Q.D. and $Q_3 = M_d$ + Q.D.

Further, a low value of Q.D. indicates a high concentration of central 50% observations and vice-versa.

Quartile deviation is an absolute measure of dispersion. The corresponding relative measure is known as coefficient of quartile deviation defined as

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

Analogous to quartile deviation and the coefficient of quartile deviation we can also define a percentile deviation and coefficient of percentile deviation as $\frac{P_{100-i} - P_i}{2}$ and $\frac{P_{100-i} - P_i}{P_{100-i} + P_i}$, respectively.

Example: Find the quartile deviation, percentile deviation and their coefficients from the following data:

Age (in years)	:	15	16	17	18	19	20	21
No. of Students	:	4	6	10	15	12	9	4

Solution:

Table for the calculation of Q.D. and P.D.

Age (X)	*No.of Students* (f)	*Less than c.f.*
15	4	4
16	6	10
17	10	20
18	15	35
19	12	47
20	9	56
21	4	60

Notes

We have, $\frac{N}{4} = \frac{60}{4} = 15$ $\therefore$ $Q_1 = 17$ (by inspection)

$\frac{3N}{4} = \frac{3 \times 60}{4} = 45$ $\therefore$ $Q_3 = 19$ "

$\frac{10N}{100} = \frac{10 \times 60}{100} = 6$ $\therefore$ $P_{10} = 16$ "

$\frac{90N}{100} = \frac{90 \times 60}{100} = 54$ $\therefore$ $P_{90} = 20$ "

Thus, Q.D. = $\frac{19-17}{2}$ = 1 year and P.D. = $\frac{20-16}{2}$ = 2 years

Also, Coefficient of Q.D. = $\frac{19-17}{19+17}$ = 0.056

and Coefficient of P.D. = $\frac{20-16}{20+16}$ = 0.11

7.6.3 Merits and Demerits of Quartile Deviation

Merits

1. It is rigidly defined.
2. It is easy to understand and easy to compute.
3. It is not affected by extreme observations and hence a suitable measure of dispersion when a distribution is highly skewed.
4. It can be calculated even for a distribution with open ends.

Demerits

1. Since it is not based on all the observations, hence, not a reliable measure of dispersion.
2. It is very much affected by the fluctuations of sampling.
3. It is not capable of being treated mathematically.

Self Assessment

State whether the following statements are true or false:

22. Interquartile Range is an absolute measure of dispersion given by the difference between second quartile (Q_3) and first quartile (Q_1).
23. Symbolically, Interquartile range = $Q_3 - Q_1/2$
24. 60% of the interquartile range is called the quartile deviation or semi-interquartile range.

Notes

7.7 Mean Deviation or Average Deviation

Mean deviation is a measure of dispersion based on all the observations. It is defined as the arithmetic mean of the absolute deviations of observations from a central value like mean, median or mode. Here the dispersion in each observation is measured by its deviation from a central value. This deviation will be positive for an observation greater than the central value and negative for less than it.

7.7.1 Calculation of Mean Deviation

The following are the formulae for the computation of mean deviation (M.D.) of an individual series of observations $X_1, X_2, \ldots\ldots X_n$:

1. M.D. from $\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|$

2. M.D. from $M_d = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - M_d\right|$

3. M.D. from $M_o = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - M_o\right|$

In case of an ungrouped frequency distribution, the observations $X_1, X_2, \ldots\ldots X_n$ occur with respective frequencies $f_1, f_2, \ldots\ldots f_n$ such that $\sum_{i=1}^{n} f_i = N$. The corresponding formulae for M.D. can be written as:

1. M.D. from $\overline{X} = \frac{1}{N}\sum_{i=1}^{n} f_i\left|X_i - \overline{X}\right|$

2. M.D. from $M_d = \frac{1}{N}\sum_{i=1}^{n} f_i\left|X_i - M_d\right|$

3. M.D. from $M_o = \frac{1}{N}\sum_{i=1}^{n} f_i\left|X_i - M_o\right|$

The above formulae are also applicable to a grouped frequency distribution where the symbols $X_1, X_2, \ldots\ldots X_n$ will denote the mid-values of the first, second ..... nth classes respectively.

Remarks: We state without proof that the mean deviation is minimum when deviations are taken from median.

Coefficient of Mean Deviation

The above formulae for mean deviation give an absolute measure of dispersion. The formulae for relative measure, termed as the coefficient of mean deviation, are given below:

4. Coefficient of M.D. from $\overline{X} = \frac{M.D.\ from\ \overline{X}}{\overline{X}}$

Notes

5. Coefficient of M.D. from $M_d = \dfrac{M.D.\ from\ M_d}{M_d}$

6. Coefficient of M.D. from $M_o = \dfrac{M.D.\ from\ M_o}{M_o}$

Example: Calculate mean deviation from mean and median for the following data of heights (in inches) of 10 persons.

60, 62, 70, 69, 63, 65, 60, 68, 63, 64

Also calculate their respective coefficients.

Solution:

Calculation of M.D. from $\overline{X}$

$$\overline{X} = \frac{60+62+70+69+63+65+60+68+63+64}{10} = 64.4 \text{ inches}$$

Sum of observations greater than $\overline{X}$ = 70 + 69 + 65 + 68 = 272.

Sum of observations less than $\overline{X}$ = 60 + 62 + 63 + 60 + 63 + 64 = 372.

Also, k_2= 4 and k_1 = 6

$\therefore$ M.D. from $\overline{X} = \dfrac{1}{10}\left[272 - 372 - (4-6)64.4\right]$ = 2.88 inches

Also, coefficient of M.D. from $\overline{X} = \dfrac{2.88}{64.4}$ = 0.045

Calculation of M.D. from M_d

Arranging the observations in order of magnitude, we have

60, 60, 62, 63, 63, 64, 65, 68, 69, 70

The median of the above observations is = $\dfrac{63+64}{2}$ = 63.5 inches.

Sum of observations greater than M_d = 64 + 65 + 68 + 69 + 70 = 336

Sum of observations less than M_d = 60 + 60 + 62 + 63 + 63 = 308

Also, k_2= 5 and k_1 = 5.

$\therefore$ M.D. from M_d = $\dfrac{\left[336 - 308 - (5-5)63.5\right]}{10}$ = 2.8 inches

Also, the coefficient of M.D. from M_d = $\dfrac{2.8}{63.5}$ = 0.044.

Example: Calculate mean deviation from median and its coefficient from the given data:

X	:	0	1	2	3	4	5	6	7	8	9
f	:	15	45	91	162	110	95	82	26	13	2

Solution:

Notes

Calculation of Mean Deviation

X	f	*Less than c. f.*	$\lvert X-4\rvert$	$f\lvert X-4\rvert$
0	15	15	4	60
1	45	60	3	135
2	91	151	2	182
3	162	313	1	162
4	110	423	0	0
5	95	518	1	95
6	82	600	2	164
7	26	626	3	78
8	13	639	4	52
9	2	641	5	10
Total				938

Since $\frac{N}{2} = \frac{641}{2} = 320.5$ $\therefore$ $M_d = 4$ (by inspection)

Thus, M.D. = $\frac{938}{641} = 1.46$ and the coefficient of M.D. = $\frac{1.46}{4} = 0.365$

Example: Calculate mean deviation from median for the following data :

Class Intervals :	20 – 25	25 – 30	30 – 40	40 – 45	45 – 50	50 – 55	55 – 60	60 – 70	70 – 80
Frequency :	6	12	17	30	10	10	8	5	2

Also calculate the coefficient of Mean Deviation.

Solution:

Calculation of Median and Mean Deviation

Class Intervals	*Frequency*(f)	*c.f.*	*Mid Values*	*fX*
20 – 25	6	6	22.5	135.0
25 – 30	12	18	27.5	330.0
30 – 40	17	35	35.0	595.0
40 – 45	30	65	42.5	1275.0
45 – 50	10	75	47.5	475.0
50 – 55	10	85	52.5	525.0
55 – 60	8	93	57.5	460.0
60 – 70	5	98	65.0	325.0
70 – 80	2	100	75.0	150.0

Since $\frac{N}{2} = 50$, the median class is 40 - 45.

$\therefore$ $L_m = 40$, $f_m = 30$, h = 5 and C = 35

Hence, $M_d = 40 + \frac{50-35}{30} \times 5 = 42.5$

Notes

Calculation of M.D.

Sum of observations which are greater than M_d

= 475 + 525 + 460 + 325 + 150 = 1,935

Sum of observations which are less than M_d

= 135 + 330 + 595 = 1060

No. of observations which are greater than M_d , i.e., k_2

= 10 +10 +8 +5 + 2 = 35

No. of observations which are less than M_d , i.e., k_1

= 6 +12 +17 = 35

$\therefore$ M.D. = $\frac{1935-1060}{100}$ = 8.75 and the coefficient of M.D. = $\frac{8.75}{42.5}$ = 0.206

7.7.2 Merits and Demerits of Mean Deviation

Merits

1. It is easy to understand and easy to compute.
2. It is based on all the observations.
3. It is less affected by extreme observations vis-a-vis range or standard deviation (to be discussed in the next section).
4. It is not much affected by fluctuations of sampling.

Demerits

1. It is not capable of further mathematical treatment. Since mean deviation is the arithmetic mean of absolute values of deviations, it is not very convenient to be algebraically manipulated.
2. This necessitates a search for a measure of dispersion which is capable of being subjected to further mathematical treatment.
3. It is not well defined measure of dispersion since deviations can be taken from any measure of central tendency.

Uses of M.D.

The mean deviation is a very useful measure of dispersion when sample size is small and no elaborate analysis of data is needed. Since standard deviation gives more importance to extreme observations the use of mean deviation is preferred in statistical analysis of certain economic, business and social phenomena.

Task Calculate the Mean Deviation from mean as well as from median of first ten prime numbers.

Self Assessment

Notes

Multiple Choice Questions:

25. An important requirement of a measure of dispersion is that it should be based on ..... the observations.

    (a) Some of (b) Few of

    (c) All (d) None of

26. ..... is a measure of dispersion based on all the observations.

    (a) Mean (b) Mean deviation

    (c) Quartiles (d) Standard deviation

27. Mean deviation is defined as the ..... of the absolute deviations of observations from a central value like mean, median or mode.

    (a) Mean (b) Arithmetic mean

    (c) Geometric mean (d) Harmonic mean

28. Mean deviation will be ..... for an observation greater than the central value.

    (a) Zero (b) Positive

    (c) Negative (d) undetermined

7.8 Standard Deviation

From the mathematical point of view, the practice of ignoring minus sign of the deviations, while computing mean deviation, is very inconvenient and this makes the formula, for mean deviation, unsuitable for further mathematical treatment. Further, if the signs are taken into account, the sum of deviations taken from their arithmetic mean is zero. This would mean that there is no dispersion in the observations. However, the fact remains that various observations are different from each other. In order to escape this problem, the squares of the deviations from arithmetic mean are taken and the positive square root of the arithmetic mean of sum of squares of these deviations is taken as a measure of dispersion. This measure of dispersion is known as standard deviation or root-mean square deviation. Square of standard deviation is known as variance. The concept of standard deviation was introduced by Karl Pearson in 1893.

The standard deviation is denoted by Greek letter 'σ' which is called 'small sigma' or simply sigma.

In terms of symbols

$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$, for n individual observations, X_1, X_2 ..... X_n, and

$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\left(X_i - \overline{X}\right)^2}$, for a grouped or ungrouped frequency distribution, where an observation X_i occurs with frequency f_i for i = 1,2, ... n and $\sum_{i=1}^{n} f_i = N$.

It should be noted here that the units of σ are same as the units of X.

Notes

7.8.1 Calculation of Standard Deviation

There are two methods of calculating standard deviation: (i) Direct Method (ii) Short-cut Method

Direct Method

1. ***Individual Series:*** If there are n observations $X_1, X_2, \ldots\ldots X_n$, various steps in the calculation of standard deviation are:

   (a) Find $\overline{X} = \frac{\sum X_i}{n}$.

   (b) Obtain deviations $\left(X_i - \overline{X}\right)$ for each i = 1, 2, ... n.

   (c) Square these deviations and add to obtain $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

   (d) Compute variance, i.e., $\sigma^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}$.

   (e) Obtain s as the positive square root of σ^2.

The above method is appropriate when $\overline{X}$ is a whole number. If $\overline{X}$ is not a whole number, the standard deviation is conveniently computed by using the transformed form of the above formula, given below.

$$\sigma^2 = \frac{1}{n}\Sigma\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\Sigma\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)$$

$$\frac{1}{n}\Sigma\left(X_i^2 - \overline{X}X_i\right)^2 - \frac{\overline{X}}{n}\Sigma\left(X_i - \overline{X}\right) = \frac{1}{n}\Sigma X_i^2 - \overline{X}\frac{\Sigma X_i}{n}$$

(The 2nd term is sum of deviations from $\overline{X}$, which is equal to zero.)

$$= \frac{1}{n}\sum X_i^2 - \overline{X}^2 = \frac{1}{n}\sum X_i^2 - \left(\frac{\sum X_i}{n}\right)^2 \text{ or}$$

= Mean of squares – Square of mean.

Example: Calculate variance and standard deviation of the weights of ten persons.

Weights (in kgs) : 45, 49, 55, 50, 41, 44, 60, 58, 53, 55

Solution:

Calculation of standard deviation

Let $u = X - \overline{X}$

											Total
Weights (X)	45	49	55	50	41	44	60	58	53	55	510
u	–6	–2	4	–1	–10	–7	9	7	2	4	0
u^2	36	4	16	1	100	49	81	49	4	16	356

Notes

From the above table

$\overline{X} = \frac{510}{10}$ = 51 kgs and $\sigma^2 = \frac{356}{10}$ =35.6 kgs^2

$\therefore$ σ = 5.97 kgs

2. ***Ungrouped or Grouped Frequency Distributions:*** Let the observations X_1, X_2 ..... X_n appear with respective frequencies f_1, f_2 ..... f_n, where $\sum f_i = N$. As before, if the distribution is grouped, then X_1, X_2 ..... X_n will denote the mid-values of the first, second ..... n^{th} class intervals respectively. The formulae for the calculation of variance and standard deviation can be written as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(X_i - \overline{X}\right)^2 \text{ and } \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i\left(X_i - \overline{X}\right)^2} \text{ respectively.}$$

Here also, we can show that

Variance = Mean of squares – Square of the mean

Therefore, we can write

$$\sigma^2 = \frac{\sum f_i X_i^2}{N} - \left(\frac{\sum f_i X_i}{N}\right)^2 \text{ and } \sigma = \sqrt{\frac{\sum f_i X_i^2}{N} - \left(\frac{\sum f_i X_i}{N}\right)^2}$$

Example: Calculate standard deviation of the following data :

X	:	10	11	12	13	14	15	16	17	18
f	:	2	7	10	12	15	11	10	6	3

Solution.

Calculation of Standard Deviation

Let $u = X - \overline{X}$

X	f	fX	u	u^2	fu^2	fX^2
10	2	20	–4	16	32	200
11	7	77	–3	9	63	847
12	10	120	–2	4	40	1440
13	12	156	–1	1	12	2028
14	15	210	0	0	0	2940
15	11	165	1	1	11	2475
16	10	160	2	4	40	2560
17	6	102	3	9	54	1734
18	3	54	4	16	48	972
Total	76	1064			300	15196

$$\overline{X} = \frac{1064}{76} = 14, \sigma^2 = \frac{300}{76} = 3.95 \text{ and } s.d.\ \sigma = \sqrt{3.95} = 1.99$$

Notes

Alternative Method

From the last column of the above table, we have

Sum of squares = 15196

$\therefore$ Mean of squares = $\frac{15196}{76}$ = 199.95

Thus, σ^2 = Mean of squares – Square of the mean = 199.96 – $(14)^2$ = 3.96

Short-cut Method

Before discussing this method, we shall examine an important property of the variance (or standard deviation), given below:

The variance of a distribution is independent of the change of origin but not of change of scale.

Change of Origin

If from each of the observations, X_1, X_2 ..... X_n, a fixed number, say A, is subtracted, the resulting values are $X_1 - A$, $X_2 - A$ ..... $X_n - A$.

We denote $X_i - A$ by d_i, where i = 1, 2 ..... n the values d_1, d_2 ..... d_n are said to be measured from A. In order to understand this, we consider the following figure.

X_i Values	:	0	1	2	3	4	5	6	7	8
d_i (=X_i – 3) Values	:	–3	–2	–1	0	1	2	3	4	5

In the above, the origin of X_i values is the point at which $X_i = 0$. When we make the transformation $d_i = X_i - 3$, the origin of d_i values shift at the value 3 because $d_i = 0$ when $X_i = 3$.

The first part of the property says that the variance of X_i values is equal to the variance of the d_i values, i.e., $\sigma_X^2 = \sigma_d^2$.

Change of Scale

To make change of scale every observation is divided (or multiplied) by a suitable constant. For example, if X_i denotes inches, then $Y_i = X_i / 12$ will denote feet or if X_i denotes rupees, then $Y_i = 100$ $X_i = \frac{X_i}{0.01}$ will denote paise, etc.

We can also have simultaneous change of origin and scale, by making the transformation $u_i = \frac{X_i - A}{h}$, where A refers to change of origin and h refers to change of scale.

According to second part of the property $\sigma_X^2 \neq \sigma_Y^2$ or $\sigma_X^2 \neq \sigma_u^2$.

The Relation between σ_X^2 and σ_u^2

Consider $\sigma_X^2 = \frac{1}{N}\sum f_i\left(X_i - \overline{X}\right)^2$ .... (1)

Notes

Let $u_i = \dfrac{X_i - A}{h}$, $\therefore$ $X_i = A + hu_i$ .... (2)

Also $\Sigma f_i X_i = \Sigma f_i(A + hu_i) = AN + h\Sigma f_i u_i$

Dividing both sides by N, we have

$$\frac{\sum f_i X_i}{N} = A + h \cdot \frac{\sum f_i u_i}{N} \text{ or } \overline{X} = A + h\overline{u} \quad \text{.... (3)}$$

Substituting the values of X_i and $\overline{X}$ in equation (1), we have

$$\sigma_X^2 = \frac{1}{N}\sum f_i\left(A + hu_i - A - h\overline{u}\right)^2 = h^2\left[\frac{1}{N}\sum f_i\left(u_i - \overline{u}\right)^2\right] = h^2\sigma_u^2 \quad \text{.... (4)}$$

The result shows that variance is independent of change of origin but not of change of scale. Using this we can write down a short-cut formula for variance of X.

$$\sigma_X^2 = h^2\left[\frac{\sum f_i u_i^2}{N} - \left(\frac{\sum f_i u_i}{N}\right)^2\right] \quad \text{.... (5)}$$

Further, when only change of origin is made

$$\sigma_X^2 = \left[\frac{\sum f_i d_i^2}{N} - \left(\frac{\sum f_i d_i}{N}\right)^2\right], \text{ where } d_i = X_i - A$$

Example: Calculate standard deviation of the following series:

Weekly wages	*No. of workers*	*Weekly wages*	*No. of workers*
100-105	200	130-135	410
105-110	210	135-140	320
110-115	230	140-145	280
115-120	320	145-150	210
120-125	350	150-155	160
125-130	520	155-160	90

Solution:

Calculation of S.D. by using d_i (= X_i – A)

Class Intervals	*No. of Workers (f)*	*Mid values (X)*	*d=X - 127.5*	*fd*	*fd²*
100-105	200	102.5	-25	-5000	125000
105-110	210	107.5	-20	-4200	84000
110-115	230	112.5	-15	-3450	51750
115-120	320	117.5	-10	-3200	32000
120-125	350	122.5	-5	-1750	8750
125-130	520	127.5	0	0	0
130-135	410	132.5	5	2050	10250
135-140	320	137.5	10	3200	32000
140-145	280	142.5	15	4200	63000
145-150	210	147.5	20	4200	84000
150-155	160	152.5	25	4000	100000
155-160	90	157.5	30	2700	81000
Total	3300			2750	671750

Notes

$$\sigma_X^2 = \frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2 = \frac{671750}{3300} - \left(\frac{2750}{3300}\right)^2 = 202.87$$

$$\therefore \quad \sigma_X^2 = \sqrt{202.87} = ₹\,14.24$$

7.8.2 Coefficient of Variation

The standard deviation is an absolute measure of dispersion and is expressed in the same units as the units of variable X. A relative measure of dispersion, based on standard deviation is known as coefficient of standard deviation and is given by $\frac{\sigma}{\overline{X}} \times 100$.

This measure introduced by Karl Pearson, is used to compare the variability or homogeneity or stability or uniformity or consistency of two or more sets of data. The data having a higher value of the coefficient of variation is said to be more dispersed or less uniform, etc.

Example: Calculate standard deviation and its coefficient of variation from the following data:

Measurements	:	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25
Frequency	:	4	1	10	3	2

Solution:

Calculation of $\overline{X}$ and σ

Class Intervals	*Frequency (f)*	*Mid-values (X)*	*u*	*fu*	*fu*^{2}
0-5	4	2.5	-2	-8	16
5-10	1	7.5	-1	-1	1
10-15	10	12.5	0	0	0
15-20	3	17.5	1	3	3
20-25	2	22.5	2	4	8
Total	20			-2	28

Here , $u = \frac{X - 12.5}{5}$

Now $\overline{X} = 12.5 - \frac{5 \times 2}{20} = 12$ and $\sigma = 5\sqrt{\frac{28}{20} - \left(\frac{2}{20}\right)^2} = 5.89$

Thus, the coefficient of variation (CV) $= \frac{5.89}{12} \times 100 = 49\%$

Example: The mean and standard deviation of 200 items are found to be 60 and 20 respectively. If at the time of calculations, two items were wrongly recorded as 3 and 67 instead of 13 and 17, find the correct mean and standard deviation. What is the correct value of the coefficient of variation?

Notes

Solution:

It is given that $\overline{X} = 60$, $\sigma = 20$ and n = 200

The sum of observations $= \sum X_i = n\overline{X} = 200 \times 60 = 12{,}000$

To find the sum of squares of observations, we use the relation

$$\sum X_i^2 = n\left(\sigma^2 + \overline{X}^2\right)$$

From this we can write $\sum X_i^2 = 200(400 + 3600) = 8{,}00{,}000$

Further the corrected sum of observations (ΣX_i) = uncorrected sum of observations – sum of wrongly recorded observations + sum of correct observations = 12,000 – (3 + 67) + (13 +17) = 11,960.

$$\therefore \quad \text{Corrected } \overline{X} = \frac{11960}{200} = 59.8$$

Similarly, the corrected sum of squares (ΣX_i^2) = uncorrected sum of squares - sum of squares of wrongly recorded observations + sum of squares of correct observations

$= 8{,}00{,}000 - (3^2 + 67^2) + (13^2 + 17^2) = 7{,}95{,}960$

Hence, corrected $\sigma^2 = \frac{795960}{200} - (59.8)^2 = 403.76$ or corrected $\sigma = 20.09$

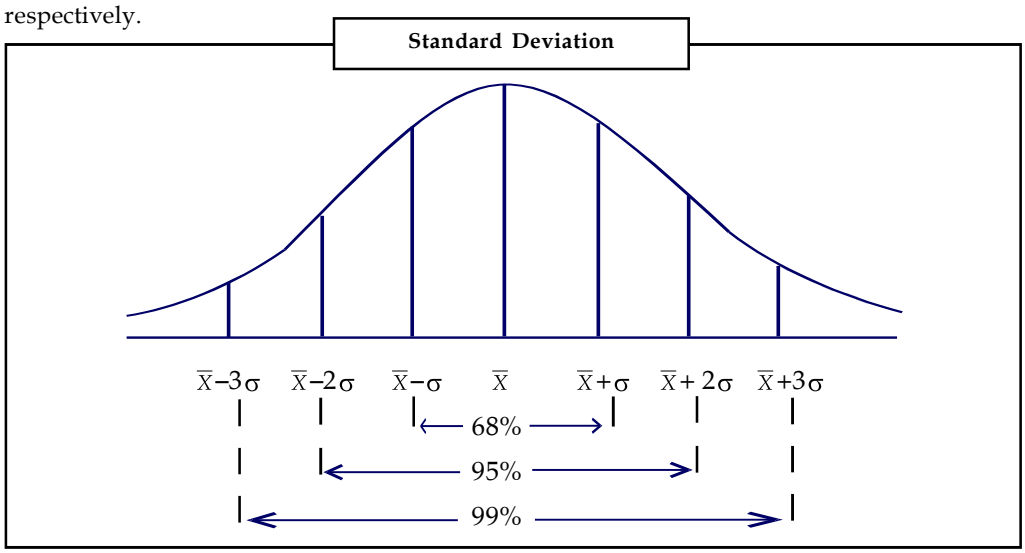
Also, $\text{CV} = \frac{20.09}{59.8} \times 100 = 33.60$

Example: Find the missing information from the following:

	Group I	Group II	Group III	Combined
Number of observations	50	?	90	200
Standard deviation	6	7	?	7.746
Mean	113	?	115	116

Solution:

Let n_1, n_2, n_3 and n denote the number of observations, $\overline{X}_1, \overline{X}_2, \overline{X}_3$ and $\overline{X}$ be the means and σ_1, σ_2, σ_3 and σ be the standard deviations of the first, second, third and combined group respectively.



Notes

From the given information we can easily determine the number of observations in group II, i.e, $n_2 = n - n_1 - n_3 = 200 - 50 - 90 = 60$.

Further the relation between means is given by

$$\overline{X} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3}{n_1 + n_2 + n_3}$$

$$\therefore \qquad \overline{X}_2 = \frac{(n_1 + n_2 + n_3)\overline{X} - n_1\overline{X}_1 - n_3\overline{X}_3}{n_2} = \frac{200 \times 116 - 50 \times 113 - 90 \times 115}{60} = 120$$

To determine s_3, consider the following relationship between variances:

$$n\sigma^2 = n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_3(\sigma_3^2 + d_3^2)$$

or $$\sigma_3^2 = \frac{n\sigma^2 - n_1\left(\sigma_1^2 + d_1^2\right) - n_2\left(\sigma_2^2 + d_2^2\right)}{n_3} - d_3^2$$

Here $d_1 = 113 - 116 = -3$, $d_2 = 120 - 116 = 4$, $d_3 = 115 - 116 = -1$

$$\therefore \qquad \sigma_3^2 = \frac{200(7.746)^2 - 50(36+9) - 60(49+16)}{90} - 1 = 64. \text{ Thus, } \sigma_3 = 8.$$

7.8.3 Properties of Standard Deviation

1. Standard deviation of a given set of observations is not greater than any other root mean square deviation, i.e. $\sqrt{\frac{1}{n}\sum\left(X_i - \overline{X}\right)^2} \le \sqrt{\frac{1}{n}\sum\left(X_i - A\right)^2}$.
2. Standard Deviation of a given set of observations is not less than mean deviation from mean, i.e., Standard Deviation $\ge$ Mean Deviation from mean.
3. In an approximately normal distribution, $\overline{X} \pm \sigma$ covers about 68% of the distribution, $\overline{X} \pm 2\sigma$ covers about 95% of the distribution and $\overline{X} \pm 3\sigma$ covers about 99%, i.e., almost whole of the distribution. This is an Empirical Rule that is based on the observations of several bell shaped symmetrical distributions. This rule is helpful in determining whether the deviation of a particular value from its mean is unusual or not. The deviations of more than 2*s* are regarded as unusual and warrant some remedial action. Furthermore, all observations with deviations of more than 3*s* from their mean are regarded as not belonging to the given data set.

7.8.4 Merits, Demerits and Uses of Standard Deviation

Merits

1. It is a rigidly defined measure of dispersion.
2. It is based on all the observations.
3. It is capable of being treated mathematically. For example, if standard deviations of a number of groups are known, their combined standard deviation can be computed.
4. It is not very much affected by the fluctuations of sampling and, therefore, is widely used in sampling theory and test of significance.

Notes

Demerits

1. As compared to the quartile deviation and range, etc., it is difficult to understand and difficult to calculate.
2. It gives more importance to extreme observations.
3. Since it depends upon the units of measurement of the observations, it cannot be used to compare the dispersions of the distributions expressed in different units.

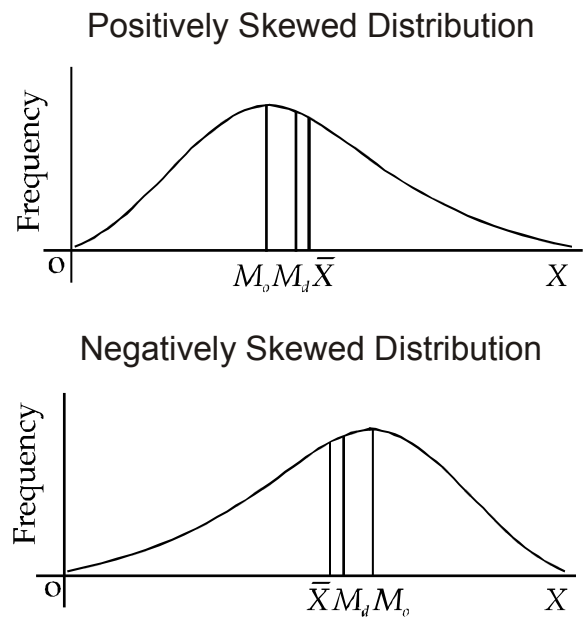
Uses of Standard Deviation

1. Standard deviation can be used to compare the dispersions of two or more distributions when their units of measurements and arithmetic means are same.
2. It is used to test the reliability of mean. It may be pointed out here that the mean of a distribution with lower standard deviation is said to be more reliable.

7.8.5 Skewness

Skewness of a distribution refers to its asymmetry. The symmetry of a distribution implies that for a given deviation from a central value, there are equal number of observations on either side of it. If the distribution is asymmetrical or skewed, its frequency curve would have a prolonged tail either towards its left or towards its right hand side. Thus, the skewness of a distribution is defined as the departure from symmetry. We may note here that there may be a situation where two or more frequency distributions are same with regard to mean and variance but not so with regard to skewness.

Figure 7.1: Positively verses Negatively Skewed Distribution



Notes

In a symmetrical distribution, mean, median and mode are equal and the ordinate at mean divides the frequency curve into two parts such that one part is the mirror image of the other, positive skewness results if some observations of high magnitude are added to a symmetrical distribution so that the right hand tail of the frequency curve gets elongated. In such a situation, we have Mode < Median < Mean. Similarly, negative skewness results when some observations of low magnitude are added to the distribution so that left hand tail of the frequency curve gets elongated and we have Mode > Median > Mean. As shown in Figure 7.1).

Measures of Skewness

A measure of skewness gives the extent and direction of skewness of a distribution. As in case of dispersion, we can define the absolute and the relative measures of skewness. Various measures of skewness can be divided into three broad categories : Measures of Skewness based on (i) $\overline{X}$, M_d and M_o, (ii) quartiles or percentiles and (iii) moments.

1. ***Measure of Skewness based on $\overline{X}$, M_d and M_o:*** This measure was suggested by Karl Pearson. According to this method, the difference between $\overline{X}$ and M_o can be taken as an absolute measure of skewness in a distribution, i.e., absolute measure of skewness = $\overline{X} - M_o$.

   Alternatively, when mode is ill defined and the distribution is moderately skewed, the above measure can also be approximately expressed as $3(\overline{X} - M_d)$.

   A relative measure, known as Karl Pearson's Coefficient of Skewness, is given by

$$S_k = \frac{\overline{X} - M_o}{\sigma} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} \text{ or } S_k = \frac{3(\overline{X} - M_d)}{\sigma}$$

   We note that if $S_k > 0$, the distribution is positively skewed,

   if $S_k < 0$, the distribution is negatively skewed and

   if $S_k = 0$, the distribution is symmetrical.

2. ***Measure of Skewness based on Quartiles or Percentiles***

   (a) *Using Quartiles*

   This measure, suggested by Bowley, is based upon the fact that Q_1 and Q_3 are equidistant from median of a symmetrical distribution, i.e., $Q_3 - M_d = M_d - Q_1$. Therefore, $(Q_3 - M_d) - (M_d - Q_1)$ can be taken as an absolute measure of skewness.

   A relative measure, known as Bowley's Coefficient of Skewness, is defined as

$$S_Q = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = \frac{Q_3 - 2M_d + Q_1}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1}$$

   The value of S_Q will lie between – 1 and + 1.

   It may be noted here that S_k and S_Q are not comparable, though, in the absence of skewness, both of them are equal to zero.

   (b) *Using Percentiles*

   Bowley's measure of skewness leaves 25% observations on each extreme of the distribution and hence is based only on the middle 50% of the observations. As an improvement to this, Kelly suggested a measure based on the middle 80% of the observations.

Notes

Kelly's absolute measure of Skewness = $(P_{90} - P_{50}) - (P_{50} - P_{10})$ and

Kelly's Coefficient of Skewness $S_P = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})} = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$

(We note that $P_{50} = M_d$).

3. ***Measure of Skewness based on Moments***

This measure is based on the property that all odd ordered moments of a symmetrical distribution are zero. Therefore, a suitable α-coefficient can be taken as a relative measure of skewness.

Since $\alpha_1 = 0$ and $\alpha_2 = 1$ for every distribution, these do not provide any information about the nature of a distribution. The third α-coefficient, i.e., α_3 can be taken as a measure of the coefficient of skewness. The skewness will be positive, negative or zero (i.e. symmetrical distribution) depending upon whether $\alpha_3 > 0$, < 0 or $= 0$. Thus, the coefficient of Skewness based on moments is given as

$$S_M = \alpha_3 = \frac{\mu_3}{\sigma^3} = \pm\sqrt{\beta_1} = \gamma_1$$

Alternatively, the skewness is expressed in terms of β_1. Since β_1 is always a non-negative number, the sign of skewness is given by the sign of μ_3.

Example: Calculate the Karl Pearson's coefficient of skewness from the following data:

Size	:	1	2	3	4	5	6	7
Frequency	:	10	18	30	25	12	3	2

Solution:

To calculate Karl Pearson's coefficient of skewness, we first find $\overline{X}$, M_o and s from the given distribution.

Size (X)	Frequency (f)	$d = X - 4$	fd	fd^2
1	10	−3	−30	90
2	18	−2	−36	72
3	30	−1	−30	30
4	25	0	0	0
5	12	1	12	12
6	3	2	6	12
7	2	3	6	18
Total	100		−72	234

$$\overline{X} = A + \frac{\sum fd}{N} = 4 + \frac{-72}{100} = 3.28$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{234}{100} - \left(\frac{-72}{100}\right)^2} = 1.35$$

Also, M_o (by inspection) = 3.00

Notes

$$\therefore \qquad S_k = \frac{\overline{X} - M_o}{\sigma} = \frac{3.28 - 3.00}{1.35} = 0.207$$

Since S_k is positive and small, the distribution is moderately positively skewed.

Example: Calculate Karl Pearson's coefficient of skewness from the following data :

Weights (lbs.)	*No. of Students*	*Weights (lbs.)*	*No. of Students*
90 - 100	4	140 - 150	23
100 - 110	10	150 - 160	16
110 - 120	17	160 - 170	5
120 - 130	22	170 - 180	3
130 - 140	30		

Solution.

Calculation of $\overline{X}$, σ and M_o

Class Intervals	Frequency (f)	Mid – Points (X)	$u = \frac{X-135}{10}$	fu	fu^2
90-100	4	95	–4	–16	64
100-110	10	105	–3	–30	90
110-120	17	115	–2	–34	68
120-130	22	125	–1	–22	22
130-140	30	135	0	0	0
140-150	23	145	1	23	23
150-160	16	155	2	32	64
160-170	5	165	3	15	45
170-180	3	175	4	12	48
Total	130			–20	424

1. $$\overline{X} = A + h\frac{\sum fu}{N} = 135 + 10 \times \frac{-20}{130} = 133.46$$

2. $$\sigma = h \times \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 10 \times \sqrt{\frac{424}{130} - \left(\frac{-20}{130}\right)^2} = 18.0$$

3. $$M_o = L_m + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h$$

By inspection, the modal class is 130 - 140.

$\therefore$ $L_m = 130$, $\Delta_1 = 30 - 22 = 8$, $\Delta_2 = 30 - 23 = 7$ and $h = 10$

Thus, $M_o = 130 + \frac{8}{15} \times 10 = 135.33$

Hence, $S_k = \frac{\overline{X} - M_o}{\sigma} = \frac{133.46 - 135.33}{18.0} = -0.10$, *i.e.*, the distribution is moderately negatively skewed.

Notes

Example: Calculate Karl Pearson's coefficient of skewness from the following data :

Class Intervals	:	40-60	30-40	20-30	15-20	10-15	5-10	3-5	0-3
Frequency	:	25	15	12	8	6	4	3	2

Solution:

Since mode is ill defined, skewness will be computed by the use of the median.

Calculation of $\overline{X}$, σ and M_d

Class Intervals	Freq. (f)	M.V. (X)	$d = X - 25$	fd	fd^2	Less than (c.f.)
0-3	2	1.5	−23.5	−47.0	1104.5	2
3-5	3	4.0	−21.0	−63.0	1323.0	5
5-10	4	7.5	−17.5	−70.0	1225.0	9
10-15	6	12.5	−12.5	−75.0	937.5	15
15-20	8	17.5	−7.5	−60.0	450.0	23
20-30	12	25.0	0.0	0.0	0.0	35
30-40	15	35.0	10.0	150.0	1500.0	50
40-60	25	50.0	25.0	625.0	15625.0	75
Total	75			460.0	22165.0	

1. $\overline{X} = 25 + \frac{460}{75} = 31.13$

2. $\sigma = \sqrt{\frac{22165}{75} - \left(\frac{460}{75}\right)^2} = 16.06$

3. Since $\frac{N}{2} = \frac{75}{2} = 37.5$, median class is 30 - 40.

Thus, $L_m = 30, C = 35, f_m = 15, h = 10$

$\therefore \quad M_d = 30 + \frac{37.5 - 35}{15} \times 10 = 31.67$

Also $\quad S_k = \frac{3(\overline{X} - M_d)}{\sigma} = \frac{3(31.13 - 31.67)}{16.06} = -0.10$

Example: Calculate Bowley's coefficient of skewness from the following data:

Class Intervals	:	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	:	7	10	20	13	17	10	14	9

Notes

Solution:

Calculation of M_d , Q_1 and Q_3

Class Intervals	Frequency (f)	Less than (c.f.)
0-5	7	7
5-10	10	17
10-15	20	37
15-20	13	50
20-25	17	67
25-30	10	77
30-35	14	91
35-40	9	100
Total	100	

1. Since $\frac{N}{2} = 50$, the median class is 15 - 20.

   Thus, L_m =15, f_m = 13, C =37, h = 5, hence $M_d = 15 + \frac{50-37}{13} \times 5 = 20$

2. Since $\frac{N}{4} = 25$, the first quartile class is 10 - 15.

   Thus, $L_{Q_1} = 10, f_{Q_1} = 20$, C = 17, h = 5, hence $Q_1 = 10 + \frac{25-17}{20} \times 5 = 12$

3. Since $\frac{3N}{4} = 75$, the third quartile class is 25 - 30.

   Thus, $L_{Q_3} = 25, f_{Q_3} = 10$, C = 67, h = 5, hence $Q_3 = 25 + \frac{75-67}{10} \times 5 = 29$

   $\therefore$ Bowley's Coefficient of Skewness $S_Q = \frac{29 - 2 \times 20 + 12}{29 - 12} = \frac{1}{17}$ = 0.06

   Thus, the distribution is approximately symmetrical.

Example: In a frequency distribution the coefficient of skewness based upon quartiles is 0.6. If the sum of upper and lower quartiles is 100 and median is 38, find the values of upper and lower quartiles.

Solution:

It is given that $Q_3 + Q_1 = 100$, $M_d = 38$ and $S_Q = 0.6$

Substituting these values in Bowley's formula, we get

$$0.6 = \frac{100 - 2 \times 38}{Q_3 - Q_1} \quad \Rightarrow \quad Q_3 - Q_1 = 40$$

Adding the equations $Q_3 + Q_1 = 100$ and $Q_3 - Q_1 = 40$, we get

$2Q_3 = 140$ or $Q_3 = 70$

Also $Q_1 = 30$ ($\because Q_1 + Q_3 = 100$).

Notes

Example: Calculate the Kelly's coefficient of skewness from the following data :

Wages (₹)	No.of Workers	Wages (₹)	No.of Workers
800-900	10	1200-1300	160
900-1000	33	1300-1400	80
1000-1100	47	1400-1500	60
1100-1200	110		

Solution:

Calculation of P_{10}, P_{50} and P_{90}

Class Intervals	No.of Workers (f)	Less than (c.f.)
800-900	10	10
900-1000	33	43
1000-1100	47	90
1100-1200	110	200
1200-1300	160	360
1300-1400	80	440
1400-1500	60	500
Total	500	

1. Since $\frac{10}{100}N = \frac{10 \times 500}{100} = 50$, P_{10} lies in the interval 1000 - 1100.

   Thus, $L_{P_{10}} = 1000, C = 43, f_{P_{10}} = 47, h = 100$

   Hence, $P_{10} = 1000 + \frac{50-43}{47} \times 100 =$ ₹ 1014.89

2. Since $\frac{50}{100}N = 250$, P_{50} lies in the interval 1200 - 1300.

   Thus, $L_{P_{50}} = 1200, C = 200, f_{P_{50}} = 160, h = 100$

   Hence, $P_{50} = 1200 + \frac{250-200}{160} \times 100 =$ ₹ 12.31.25

(iii) Since $\frac{90}{100}N = 450$, P_{90} lies in the class 1400 - 1500.

   Thus, $L_{P_{90}} = 1400, C = 440, f_{P_{90}} = 60, h = 100$

   Hence, $P_{90} = 1400 + \frac{450-440}{60} \times 100 =$ ₹ 14.16.67

$$\therefore S_P = \frac{1416.67 + 1014.89 - 2 \times 1231.25}{1416.67 - 1014.89} = \frac{-30.94}{401.78} = -0.08$$

Example: Compute the moment measure of skewness from the following distribution.

Marks obtained :	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
No. of Students :	8	14	22	26	15	10	5

Notes

Solution:

Calculation of Skewness

Class Intervals	Freq. (f)	M.V. (X)	$u = \frac{X-35}{10}$	fu	fu^2	fu^3
0-10	8	5	−3	−24	72	−216
10-20	14	15	−2	−28	56	−112
20-30	22	25	−1	−22	22	−22
30-40	26	35	0	0	0	0
40-50	15	45	1	15	15	15
50-60	10	55	2	20	40	80
60-70	5	65	3	15	45	135
Total	100			−24	250	−120

$$\mu_1' = h\frac{\sum fu}{N} = \frac{10\times(-24)}{100} = -2.4,\ \mu_2' = h^2\frac{\sum fu^2}{N} = \frac{100\times 250}{100} = 250 \text{ and}$$

$$\mu_3' = h^3\frac{\sum fu^3}{N} = \frac{1000\times(-120)}{100} = -1200.$$

Thus, $\mu_2 = 250 - (-2.4)^2 = 244.24$ and $\mu_3 = -1200 + 3\times 250\times 2.4 + 2(-2.4)^3 = 572.35$

$$\text{Hence, } \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(572.35)^2}{(244.24)^3} = 0.02248$$

Since the value of β_1 is small and μ_3 is positive, therefore, the distribution is moderately positively skewed.

Since β_1 is a coefficient, its value can directly be obtained from moments of u, *i.e.,* the moments without adjustment by the scale factor h. Let us denote various moments of u as follows:

$$\delta_1' = \frac{\sum fu}{N} = \frac{-24}{100} = -0.24,\ \delta_2' = \frac{\sum fu^2}{N} = \frac{250}{100} = 2.50,\ \delta_3' = \frac{\sum fu^3}{N} = \frac{-120}{100} = -1.2$$

$$\therefore\ \delta_2 = \delta_2' - \delta_1'^2 = 2.50 - (-0.24)^2 = 2.4424$$

Note: At least 4 places after decimal should be taken to get the correct results.

$$\delta_3 = \delta_3' - 3\delta_2'\delta_1' + 2\delta_1'^3 = -1.2 - 3\times 2.5\times(-0.24) + 2\times(-0.24)^3 = 0.5724$$

$$\therefore\ \beta_1 = \frac{\delta_3^2}{\delta_2^3} = \frac{(0.5724)^2}{(2.4424)^3} = 0.02249$$

It may also be pointed out that the central moments can also be obtained from δ_2, δ_3, etc., by suitable multiplication of the scale factor.

Example: If the first three moments of an empirical frequency distribution about the value 2 are 1, 16 and − 40. Examine the skewness of the distribution.

Notes

Solution: We are given raw moments; $\mu_1' = 1, \mu_2' = 16$ and $\mu_3' = -40$, which should be converted into central moments.

Now $\mu_2 = \mu_2' - \mu_1'^2 = 16 - 1 = 15$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -40 - 3 \times 16 + 2 = -86$$

$$\therefore \quad \gamma_1 = -\sqrt{\beta_1} = -\frac{86}{(15)^{3/2}} = -1.48$$

7.8.6 Empirical Relation among Various Measures of Dispersions

Although much depends upon the nature of a frequency distribution, it has been observed that for a symmetrical or moderately skewed distribution, the following approximate results hold true.

QD $\approx$ 0.8453 (approximately $\frac{5}{6}$) × MD

QD $\approx$ 0.6745 (approximately $\frac{2}{3}$) × SD

MD $\approx$ 0.7979 (approximately $\frac{4}{5}$) × SD

or we can say that 6 SD $\approx$ 9 QD $\approx$ 7.5 MD

Also Range $\approx$ 6 SD

Caution Standard deviation is independent of change of origin but not of change of scale. This implies that if a constant is added (or subtracted) to all the observations, the value of standard deviation remains unaffected. On the other hand, if all the observations are multiplied or divided by a constant, the standard deviation also gets multiplied (or divided) by this constant.

Notes

Choice of a Suitable Measure of Dispersion

The choice of a suitable measure depends upon: (i) The nature of available data, (ii) the objective of measuring dispersion and (iii) the characteristics of the measure of dispersion.

The nature of available data may restrict the choice of a measure of dispersion. For example, if the distribution has class intervals with open ends, one can only calculate quartile deviation, percentile deviation, etc. On the other hand, if the objective is to know the extent of variations in the values of a variable in a given time or situation, the calculation of range may be more appropriate, e.g., maximum and minimum rainfall in a season, maximum and minimum temperature on a particular day, etc. Similarly, Lorenz Curves are generally used to compare the extent of inequalities of income or wealth in two or more situations. Further, if one is interested in obtaining the magnitude of variation in observations on the average from a central value, mean deviation and standard deviation are used. In statistical analysis, the use of standard deviation is preferred to mean deviation because of its several merits over the later. In the words of M.M. Blair, "These two measures

Contd...

Notes

(Mean Deviation and Standard Deviation) are to the statistician what the axe and cross-cut saw are to the woodsman – the basic tools for working upon his raw materials". Since the mean deviation is inconvenient to handle mathematically, the standard deviation is often used as a measure of dispersion in the absence of any special reason.

Self Assessment

State whether the following statements are true or false:

29. The squares root of the deviations from arithmetic mean are taken and the positive square of the arithmetic mean of sum of squares of these deviations is taken as a measure of dispersion. This measure of dispersion is known as standard deviation or root-mean square deviation.
30. The concept of standard deviation was introduced by Karl Pearson in 1897.
31. The standard deviation is an absolute measure of dispersion and is expressed in the same units as the units of variable X.
32. A relative measure of dispersion, based on standard deviation is known as coefficient of standard deviation.

Case Study

Philips India Ltd.

Philips India Ltd., manufactures the famous Philips tube-lights of 40 watts. The company has developed a new variety of Flouroscent 24 watt tube lights for specific applications in control equipments used in defence components. Before it is commercially launched the manager R&D desires to ensure its reliability and quality. The test results conducted on 400 such tube lights are shown below.

Life Time (hours)	No. of Tubes
14	300-400
46	400-500
58	500-600
76	600-700
68	700-800
62	800-900
48	900-1000
22	1000-1100
6	1100-1200

Compute the coefficient of skewness.

7.9 Summary

- Range = L – S, L = largest observation and S = smallest observation.
- Coefficient of Range = $\frac{L-S}{L+S}$

Notes

- Quartile Deviation or Semi-Interquartile Range QD = $\frac{Q_3 - Q_1}{2}$
- Coefficient of QD = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$
- Mean Deviation from $\overline{X} = \frac{1}{N}\sum f_i \left| X_i - \overline{X} \right|$

  Mean Deviation from $M_d = \frac{1}{N}\sum f_i \left| X_i - M_d \right|$

  Mean Deviation from $M_o = \frac{1}{N}\sum f_i \left| X_i - M_o \right|$
- Coefficient of $MD = \frac{M.D.}{\overline{X}}$ or $\frac{M.D.}{M_d}$ or $\frac{M.D.}{M_o}$
- Standard Deviation $\sigma = \sqrt{Mean\ of\ squares - Square\ of\ the\ Mean}$

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

$$= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}, \text{ where } d = X - A.$$

$$= h \times \sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2}, \text{ where } u = \frac{X - A}{h}$$

- Coefficient of s.d. = $\frac{\sigma}{\overline{X}}$
- Coefficient of Variation = $\frac{\sigma}{\overline{X}} \times 100$
- Standard Deviation of the combined series

$$\sigma = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right) + \cdots + n_k\left(\sigma_k^2 + d_k^2\right)}{n_1 + n_2 + \cdots + n_k}},$$

  where $d_i = \overline{X}_i - \overline{X}$ for i = 1,2, ... k

7.10 Keywords

Averages of second order : The measures which express the spread of observations in terms of the average of deviations of observations from some central value are termed as the averages of second order, e.g., mean deviation, standard deviation, etc.

Coefficient of standard deviation: A relative measure of dispersion, based on standard deviation is known as coefficient of standard deviation.

Dispersion: Dispersion is the measure of extent to which individual items vary.

Notes

Distance measures: The measures which express the spread of observations in terms of distance between the values of selected observations. These are also termed as distance measures, e.g., range, interquartile range, interpercentile range, etc.

Interquartile Range: Interquartile Range is an absolute measure of dispersion given by the difference between third quartile (Q_3) and first quartile (Q_1)

Symbolically, Interquartile range = $Q_3 - Q_1$.

Measure of central tendency: A measure of central tendency summarizes the distribution of a variable into a single figure which can be regarded as its representative.

Measure of variation: The measure of the scatteredness of the mass of figures in a series about an average is called the measure of variation.

Quartile deviation or semi-interquartile range: Half of the interquartile range is called the quartile deviation or semi-interquartile range.

Range: The range of a distribution is the difference between its two extreme observations, i.e., the difference between the largest and smallest observations. Symbolically, R = L – S where R denotes range, L and S denote largest and smallest observations.

Standard deviation or root-mean square deviation: The squares of the deviations from arithmetic mean are taken and the positive square root of the arithmetic mean of sum of squares of these deviations is taken as a measure of dispersion. This measure of dispersion is known as standard deviation or root-mean square deviation

Variance: Square of standard deviation is known as variance.

7.11 Review Questions

1. "Frequency distribution may either differ in numerical size of their averages though not necessarily in their formation or they may have the same values of their averages yet differ in their respective formation". Explain and illustrate how the measures of dispersion afford a supplement to the information about frequency distribution furnished by averages.
2. "Indeed the averages and measures of variation together cover most of the need of practical statistician but their interpretation and use in combination require a good knowledge of statistical theory". — Tippet

   Discuss this statement with the help of arithmetic mean and standard deviation.
3. " Measures of dispersion and central tendency are complementary to each other in highlighting the characteristics of a frequency distribution". Explain this statement with suitable examples.
4. Explain briefly the meaning of (i) Range (ii) Quartile Deviation.
5. Distinguish between an absolute measure and relative measure of dispersion. What are the advantages of using the latter?
6. Explain how the standard deviation is a better measure as compared to other measures of dispersion? Mention its defects, if any.
7. What do you understand by mean deviation? Explain its merits and demerits.
8. Explain mean deviation, quartile deviation and standard deviation. Discuss the circumstances in which they may be used.

Notes

9. What do you understand by coefficient of variation? Discuss the relative advantages of coefficient of variation and standard deviation as a measure of variability.

10. Calculate range and its coefficient from the following data:

    (a) 159, 167, 139, 119, 117, 168, 133, 135, 147, 160

    (b)

Weights (lbs.) :	115-125	125-135	135-145	145-155	155-165	165-175
Frequency :	4	5	6	3	1	1

11. Find out quartile deviation and its coefficient from the following data:

Class :	0-4	5-9	10-14	15-19	20-24	25-29
Frequency :	15	26	12	5	4	3

12. Find out the range of income of (a) middle 50% of workers, (b) middle 80% of the workers and hence the coefficients of quartile deviation and percentile deviation from the following data :

Wages less than :	40	50	60	70	80	90	100
No. of workers :	5	8	15	20	30	33	35

13. The following data denote the weights of 9 students of certain class. Calculate mean deviation from median and its coefficient.

S.No. :	1	2	3	4	5	6	7	8	9
Weight :	40	42	45	47	50	51	54	55	57

14. Calculate mean deviation from median for the following data :

Wages per week :	50-59	60-69	70-79	80-89	90-99	100-109	110-119
No. of workers :	15	40	50	60	45	90	15

15. Calculate the coefficient of mean deviation from mean and median from the following data :

Marks :	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Students :	2	6	12	18	25	20	10	7

16. Calculate mean deviation from mode of the following data:

    (a) 7, 4, 6, 4, 4, 5, 2, 4, 1, 7, 7, 6, 2, 3, 4, 2

    (b)

Size :	0-10	10-20	20-30	30-40	40-50	50-60
Frequency :	6	20	44	26	3	1

17. Calculate the standard deviation of the following series:

Marks :	0-10	10-20	20-30	30-40	40-50
Frequency :	10	8	15	8	4

18. Calculate the standard deviation from the following data:

Age less than (in years) :	10	20	30	40	50	60	70	80
No. of Persons :	15	30	53	75	100	110	115	125

Notes

19. Find out standard deviation from the following data:

Midvalue	:	30	35	40	45	50	55	60	65	70	75	80
Frequency	:	1	2	4	7	9	13	17	12	7	6	3

20. The mean of 150 observations is 35 and their standard deviation is 4. Find sum and sum of squares of all the observations.

21. The mean and standard deviation of two distributions having 100 and 150 observations are 50, 5 and 40, 6 respectively. Find the mean and standard deviation of all the 250 observations taken together.

22. The mean and standard deviation of 100 items are found to be 40 and 10. If, at the time of calculations, two items were wrongly taken as 30 and 70 instead of 3 and 27, find the correct mean and standard deviation.

23. The sum and the sum of squares of a set of observations are 75 and 435 respectively. Find the number of observations if their standard deviation is 2.

24. The sum and the sum of squares of 50 observations from the value 20 are -10 and 452 respectively. Find standard deviation and coefficient of variation.

25. The mean and standard deviation of marks obtained by 40 students of a class in statistics are 55 and 8 respectively. If there are only 5 girls in the class and their respective marks are 40, 55, 63, 75 and 87, find mean and standard deviation of the marks obtained by boys.

26. There are 60 male and 40 female workers in a factory. The standard deviations of their wages (per hour) were calculated as ₹ 8 and ₹ 11 respectively. The mean wages of the two groups were found to be equal. Compute the combined standard deviation of the wages of all the workers.

Answers: Self Assessment

1.	representative	2.	same,different
3.	dispersion	4.	averages of first order
5.	averages of second order	6.	True
7.	False	8.	True
9.	True	10.	True
11.	rigidly	12.	based on
13.	extreme observations	14.	fluctuations
15.	False	16.	True
17.	True	18.	True
19.	range	20.	largest and smallest
21.	coefficient of range	22.	False
23.	False	24.	False
25.	(c)	26.	(b)
27.	(b)	28.	(b)
29.	False	30.	False
31.	True	32.	True

7.9 Further Readings

Notes

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Bhardwaj R.S., *Business Statistics*, Excel Books.

Garrett H.E. (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, Mc Graw Hill Book Company, New York.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hannagan T.J. (1982), *Mastering Statistics*, The Macmillan Press Ltd., Surrey.

Hooda R. P., *Statistics for Business and Economics*, Macmillan India, Delhi, 2008.

Jaeger R.M (1983), *Statistics: A Spectator Sport*, Sage Publications India Pvt. Ltd., New Delhi.

Lindgren B.W. (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Selvaraj R., Loganathan, C. *Quantitative Methods in Management.*

Sharma J.K., *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutta.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Online links

http://en.wikipedia.org/wiki/Statistical_dispersion

http://www.quickmba.com/stats/dispersion/

http://www.mathsisfun.com/data/quartiles.html

http://www.investopedia.com/terms/s/standarddeviation.asp

http://en.wikipedia.org/wiki/Coefficient_of_variation

Notes

Unit 8: Correlation Analysis

CONTENTS

Objectives

Introduction

8.1 Correlation

8.1.1 Definitions of Correlation

8.1.2 Scope of Correlation Analysis

8.1.3 Properties of Coefficient of Correlation

8.1.4 Scatter Diagram

8.1.5 Karl Pearson's Coefficient of Linear Correlation

8.1.6 Merits and Limitations of Coefficient of Correlation

8.2 Spearman's Rank Correlation

8.2.1 Case of Tied Ranks

8.2.2 Limits of Rank Correlation

8.3 Summary

8.4 Keywords

8.5 Review Questions

8.6 Further Readings

Objectives

After studying this unit, you will be able to:

- Differentiate between univariate distribution and Bivariate Distribution
- Categorize study of relationship between two or more variables
- State the definition and scope of Correlation Analysis
- Discuss the properties, merits and demerits of coefficient of correlation
- Explain spearman's rank correlation
- Analyse the case of tied ranks and focus on limits of rank correlation

Introduction

So far we have considered distributions relating to a single characteristics. Such distributions are known as Univariate Distribution. When various units under consideration are observed simultaneously, with regard to two characteristics, we get a Bivariate Distribution. For example, the simultaneous study of the heights and weights of students of a college. For such data also, we can compute mean, variance, skewness, etc., for each individual characteristics. In addition to this, in the study of a bivariate distribution, we are also interested in knowing whether there exists some relationship between two characteristics or in other words, how far the two variables, corresponding to two characteristics, tend to move together in same or opposite directions i.e. how far they are associated.

Notes

The knowledge of this type of relationship is useful for predicting the value of one variable given the value of the other. It also helps in understanding and analysis of various economic and business problems. It should be noted here that statistical relations are different from the exact mathematical relations. Given a statistical relation Y = a + bX, between two variables X and Y, we can only get a value of Y that we expect on the average for a given value of X. The study of relationship between two or more variables can be divided into two broad categories:

1. To determine whether there exists some sort of association between the variables. If so, what is the degree of association or the magnitude of correlation between the two.
2. To determine the most suitable form of the relationship between the variables given that they are correlated.

The first category relates to the study of 'Correlation' which will be discussed in this unit and the second relates to the study of 'Regression', to be discussed in next unit.

8.1 Correlation

Various experts have defined correlation in their own words and their definitions, broadly speaking, imply that correlation is the degree of association between two or more variables.

8.1.1 Definitions of Correlation

Some important definitions of correlation are given below:

"If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated."

– *L.R. Connor*

"Correlation is an analysis of covariation between two or more variables."

– *A.M. Tuttle*

"When the relationship is 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."

– *Croxton and Cowden*

"Correlation analysis attempts to determine the 'degree of relationship' between variables".

– *Ya Lun Chou*

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

8.1.2 Scope of Correlation Analysis

If there is a correlation between two variables, it may be due to any of the following situations:

1. ***One of the variable may be affecting the other:*** A correlation coefficient calculated from the data on quantity demanded and corresponding price of tea would only reveal that the degree of association between them is very high. It will not give us any idea about whether price is affecting demand of tea or vice-versa. In order to know this, we need to have some additional information apart from the study of correlation. For example, if on the basis of some additional information, we say that the price of tea affects its demand, then price will be the cause and quantity will be the effect. The causal variable is also termed as independent variable while the other variable is termed as dependent variable.

Notes

2. ***The two variables may act upon each other:*** Cause and effect relation exists in this case also but it may be very difficult to find out which of the two variables is independent. For example, if we have data on price of wheat and its cost of production, the correlation between them may be very high because higher price of wheat may attract farmers to produce more wheat and more production of wheat may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of wheat. For the purpose of determining a relationship between the two variables in such situations, we can take any one of them as independent variable.

3. ***The two variables may be acted upon by the outside influences:*** In this case we might get a high value of correlation between the two variables, however, apparently no cause and effect type relation seems to exist between them. For example, the demands of the two commodities, say X and Y, may be positively correlated because the incomes of the consumers are rising. Coefficient of correlation obtained in such a situation is called a spurious or nonsense correlation.

4. ***A high value of the correlation coefficient may be obtained due to sheer coincidence (or pure chance):*** This is another situation of spurious correlation. Given the data on any two variables, one may obtain a high value of correlation coefficient when in fact they do not have any relationship. For example, a high value of correlation coefficient may be obtained between the size of shoe and the income of persons of a locality.

8.1.3 Properties of Coefficient of Correlation

1. ***The coefficient of correlation is independent of the change of origin and scale of measurements.***

   This property is very useful in the simplification of computations of correlation. On the basis of this property, we can write a short-cut formula for the computation of r_{XY} :

$$r_{XY} = \frac{n\sum u_i v_i - \left(\sum u_i\right)\left(\sum v_i\right)}{\sqrt{n\sum u_i^2 - \left(\sum u_i\right)^2}\sqrt{n\sum v_i^2 - \left(\sum v_i\right)^2}}$$

2. ***If X and Y are independent they are uncorrelated, but the converse is not true:*** This property points our attention to the fact that r_{XY} is only a measure of the degree of linear association between X and Y. If the association is non-linear, the computed value of r_{XY} is no longer a measure of the degree of association between the two variables.

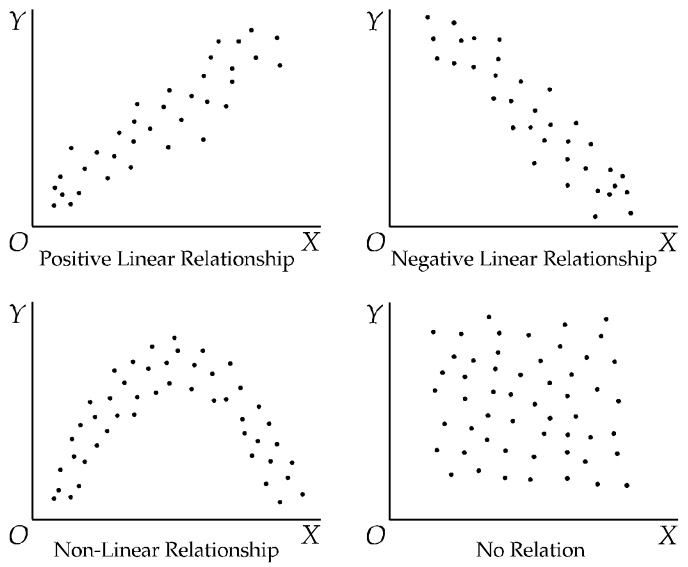
8.1.4 Scatter Diagram

Let the bivariate data be denoted by (X_i, Y_i), where i = 1, 2 ..... n. In order to have some idea about the extent of association between variables X and Y, each pair (X_i, Y_i), i = 1, 2.....n, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

Each pair of values (X_i, Y_i) is denoted by a point on the graph. The set of such points (also known as dots of the diagram) may cluster around a straight line or a curve or may not show any tendency of association. Various possible situations are shown with the help of following diagrams:

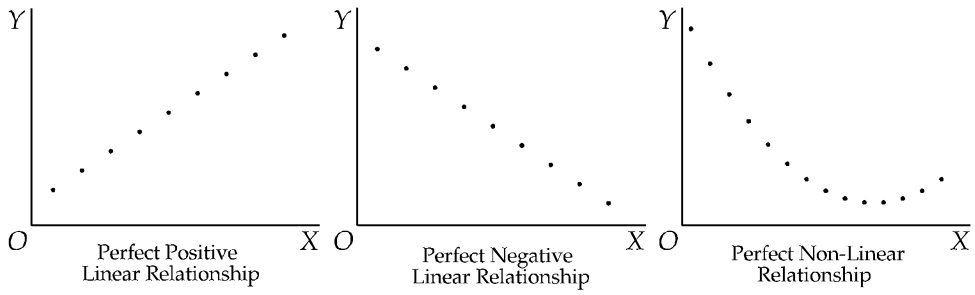
Notes

Figure 8.1: Scatter Diagram



If all the points or dots lie exactly on a straight line or a curve, the association between the variables is said to be perfect. This is shown below:

Figure 8.2: Association between the Variables



A scatter diagram of the data helps in having a visual idea about the nature of association between two variables. If the points cluster along a straight line, the association between variables is linear. Further, if the points cluster along a curve, the corresponding association is non-linear or curvilinear. Finally, if the points neither cluster along a straight line nor along a curve, there is absence of any association between the variables.

It is also obvious from the above figure that when low (high) values of X are associated with low (high) value of Y, the association between them is said to be positive. Contrary to this, when low (high) values of X are associated with high (low) values of Y, the association between them is said to be negative.

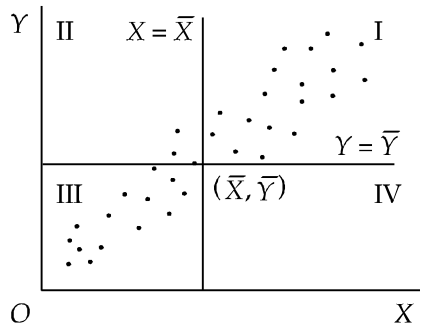
This unit deals only with linear association between the two variables X and Y. We shall measure the degree of linear association by the Karl Pearson's formula for the coefficient of linear correlation.

Notes

8.1.5 Karl Pearson's Coefficient of Linear Correlation

Let us assume, again, that we have data on two variables X and Y denoted by the pairs (X_i, Y_i), i = 1,2, ... n. Further, let the scatter diagram of the data be as shown in figure 8.3.

Figure 8.3: Scatter Diagram of the Data



Let $\bar{X}$ and $\bar{Y}$ be the arithmetic means of X and Y respectively. Draw two lines $X = \bar{X}$ and $Y = \bar{Y}$ on the scatter diagram. These two lines, intersect at the point $(\bar{X}, \bar{Y})$ and are mutually perpendicular, divide the whole diagram into four parts, termed as I, II, III and IV quadrants, as shown.

As mentioned earlier, the correlation between X and Y will be positive if low (high) values of X are associated with low (high) values of Y. In terms of the above figure, we can say that when values of X that are greater (less) than $\bar{X}$ are generally associated with values of Y that are greater (less) than $\bar{Y}$, the correlation between X and Y will be positive. This implies that there will be a general tendency of points to concentrate in I and III quadrants. Similarly, when correlation between X and Y is negative, the point of the scatter diagram will have a general tendency to concentrate in II and IV quadrants.

Further, if we consider deviations of values from their means, i.e., $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$, we note that:

1. Both $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ will be positive for all points in quadrant I.

2. $(X_i - \bar{X})$ will be negative and $(Y_i - \bar{Y})$ will be positive for all points in quadrant II.

3. Both $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$ will be negative for all points in quadrant III.

4. $(X_i - \bar{X})$ will be positive and $(Y_i - \bar{Y})$ will be negative for all points in quadrant IV.

It is obvious from the above that the product of deviations, i.e., $(X_i - \bar{X})(Y_i - \bar{Y})$ will be positive for points in quadrants I and III and negative for points in quadrants II and IV.

Since, for positive correlation, the points will tend to concentrate more in I and III quadrants than in II and IV, the sum of positive products of deviations will outweigh the sum of negative products of deviations. Thus, $\sum (X_i - \bar{X})(Y_i - \bar{Y})$ will be positive for all the n observations.

Similarly, when correlation is negative, the points will tend to concentrate more in II and IV quadrants than in I and III. Thus, the sum of negative products of deviations will outweigh the

Notes

sum of positive products and hence $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ will be negative for all the n observations.

Further, if there is no correlation, the sum of positive products of deviations will be equal to the sum of negative products of deviations such that $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ will be equal to zero.

On the basis of the above, we can consider $\sum(X_i - \bar{X})(Y_i - \bar{Y})$ as an absolute measure of correlation. This measure, like other absolute measures of dispersion, skewness, etc., will depend upon (i) the number of observations and (ii) the units of measurements of the variables.

In order to avoid its dependence on the number of observations, we take its average, i.e., $\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$. This term is called covariance in statistics and is denoted as Cov(X,Y).

To eliminate the effect of units of measurement of the variables, the covariance term is divided by the product of the standard deviation of X and the standard deviation of Y. The resulting expression is known as the Karl Pearson's coefficient of linear correlation or the product moment correlation coefficient or simply the coefficient of correlation, between X and Y.

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad \text{.... (1)}$$

or

$$r_{XY} = \frac{\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum(Y_i - \bar{Y})^2}} \qquad \text{.... (2)}$$

Cancelling $\frac{1}{n}$ from the numerator and the denominator, we get

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} \qquad \text{.... (3)}$$

Consider $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum(X_i - \bar{X})Y_i - \bar{Y}\sum(X_i - \bar{X})$

$$= \sum X_i Y_i - \bar{X}\sum Y_i \text{ (second term is zero)}$$

$$= \sum X_i Y_i - n\bar{X}\bar{Y} \quad \left(\sum Y_i = n\bar{Y}\right)$$

Similarly we can write $\sum(X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$

and $\sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$

Substituting these values in equation (3), we have

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left[\sum X_i^2 - n\bar{X}^2\right]}\sqrt{\left[\sum Y_i^2 - n\bar{Y}^2\right]}} \qquad \text{.... (4)}$$

Notes

$$r_{XY} = \frac{\sum X_i Y_i - n \cdot \frac{\sum X_i}{n} \times \frac{\sum Y_i}{n}}{\sqrt{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2}\sqrt{\sum Y_i^2 - n\left(\frac{\sum Y_i}{n}\right)^2}}$$

$$= \frac{\sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}}{\sqrt{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}\sqrt{\sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}}} \quad .... (5)$$

On multiplication of numerator and denominator by n, we can write

$$r_{XY} = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{n\sum X_i^2 - \left(\sum X_i\right)^2}\sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}} \quad .... (6)$$

Further, if we assume $x_i = X_i - \overline{X}$ and $y_i = Y_i - \overline{Y}$, equation (2), given above, can be written as

$$r_{XY} = \frac{\frac{1}{n}\sum x_i y_i}{\sqrt{\frac{1}{n}\sum x_i^2}\sqrt{\frac{1}{n}\sum y_i^2}} \quad .... (7)$$

$$or \quad r_{XY} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \quad .... (8)$$

$$or \quad r_{XY} = \frac{1}{n}\frac{\sum x_i y_i}{\sigma_x \sigma_y} \quad .... (9)$$

Equations (5) or (6) are often used for the calculation of correlation from raw data, while the use of the remaining equations depends upon the forms in which the data are available. For example, if standard deviations of X and Y are given, equation (9) may be appropriate.

Example: Calculate the Karl Pearson's coefficient of correlation from the following pairs of values:

Values of X :	12	9	8	10	11	13	7
Values of Y :	14	8	6	9	11	12	3

Solution:

The formula for Karl Pearson's coefficient of correlation is

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

The values of different terms, given in the formula, are calculated from the following table:

Notes

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
12	14	168	144	196
9	8	72	81	64
8	6	48	64	36
10	9	90	100	81
11	11	121	121	121
13	12	156	169	144
7	3	21	49	9
70	63	676	728	651

Here n = 7 (no. of pairs of observations)

$$r_{XY} = \frac{7 \times 676 - 70 \times 63}{\sqrt{7 \times 728 - (70)^2}\sqrt{7 \times 651 - (63)^2}} = 0.949$$

Example: Calculate the Karl Pearson's coefficient of correlation between X and Y from the following data:

No. of pairs of observations n = 8, $\sum (X_i - \bar{X})^2 = 184$, $\sum (Y_i - \bar{Y})^2 = 148$,

$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 164$, $\bar{X} = 11$ and $\bar{Y} = 10$

Solution:

Using the formula, $r_{XY} = \dfrac{\sum (X - \bar{X}_i)(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2}\sqrt{\sum (Y_i - \bar{Y})^2}}$, we get

$$r_{XY} = \frac{164}{\sqrt{184}\sqrt{148}} = 0.99$$

Example: Calculate the coefficient of correlation between age group and rate of mortality from the following data:

Age group	:	0 - 20	20 - 40	40 - 60	60 - 80	80 - 100
Rate of Mortality	:	350	280	540	760	900

Solution:

Since class intervals are given for age, their mid-values shall be used for the calculation of r.

Table for calculation of r

Age group	*M.V.* (X)	*Rate of Mort.* (Y)	$u_i = \frac{X_i - 50}{20}$	$v_i = \frac{Y_i - 540}{10}$	u_iv_i	u_i^2	v_i^2
0 - 20	10	350	- 2	- 19	38	4	361
20 - 40	30	280	- 1	- 26	26	1	676
40 - 60	50	540	0	0	0	0	0
60 - 80	70	760	1	22	22	1	484
80 - 100	90	900	2	36	72	4	1296
Total			0	13	158	10	2817

Notes

Here n = 5. Using the formula (10) for correlation, we get

$$r_{XY} = \frac{5 \times 158 - 0 \times 13}{\sqrt{5 \times 10 - 0^2}\sqrt{5 \times 2817 - 13^2}} = 0.95$$

Example: From the following table, find the missing values and calculate the coefficient of correlation by Karl Pearson's method:

X :	6	2	10	4	?
Y :	9	11	?	8	7

Arithmetic means of X and Y series are 6 and 8 respectively.

Solution:

The missing value in X-series = 5 × 6 – (6 + 2 + 10 + 4) = 30 – 22 = 8

The missing value in Y-series = 5 × 8 – (9 + 11 + 8 + 7) = 40 – 35 = 5

Table for Calculation of r

X	Y	$X-\bar{X}$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
6	9	0	1	0	0	1
2	11	−4	3	−12	16	9
10	5	4	−3	−12	16	9
4	8	−2	0	0	4	0
8	7	2	−1	−2	4	1
Total				−26	40	20

Using formula of correlation, we get $r = \frac{-26}{\sqrt{40}\sqrt{20}} = -0.92$

Example: Calculate Karl Pearson's coefficient of correlation for the following series :

Price (in ₹) :	10	11	12	13	14	15	16	17	18	19
Demand (in kgs) :	420	410	400	310	280	260	240	210	210	200

Solution.

Table for calculation of r

Price (X)	*Demand* (Y)	$u = X - 14$	$v = \frac{Y-310}{10}$	uv	u^2	v^2
10	420	−4	11	−44	16	121
11	410	−3	10	−30	9	100
12	400	−2	9	−18	4	81
13	310	−1	0	0	1	0
14	280	0	−3	0	0	9
15	260	1	−5	−5	1	25
16	240	2	−7	−14	4	49
17	210	3	−10	−30	9	100
18	210	4	−10	−40	16	100
19	200	5	−11	−55	25	121
Total		5	−16	−236	85	706

Notes

$$r = \frac{-10 \times 236 + 5 \times 16}{\sqrt{10 \times 85 - 25}\sqrt{10 \times 706 - 256}} = -0.96$$

Example:

A computer while calculating the correlation coefficient between two variables, X and Y, obtained the following results :

$n = 25, \Sigma X = 125, \Sigma X^2 = 650, \Sigma Y = 100, \Sigma Y^2 = 460, \Sigma XY = 508.$

It was, however, discovered later at the time of checking that it had copied down two pairs of observations as $\begin{array}{c|c} X & Y \\ \hline 6 & 14 \\ 8 & 6 \end{array}$ in place of the correct pairs $\begin{array}{c|c} X & Y \\ \hline 8 & 12 \\ 6 & 8 \end{array}$. Obtain the correct value of r.

Solution:

First we have to correct the values of $\Sigma X, \Sigma X^2$..... etc.

Corrected ΣX = 125 – (6 + 8) + (8 + 6) = 125

Corrected ΣX^2 = 650 – (36 + 64) + (64 + 36) = 650

Corrected ΣY = 100 – (14 + 6) + (12 + 8) = 100

Corrected ΣY^2 = 460 – (196 + 36) + (144 + 64) = 436

Corrected ΣXY = 508 - (84 + 48) + (96 + 48) = 520

$$r = \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\sqrt{25 \times 436 - (100)^2}} = 0.67$$

8.1.6 Merits and Limitations of Coefficient of Correlation

The only merit of Karl Pearson's coefficient of correlation is that it is the most popular method for expressing the degree and direction of linear association between the two variables in terms of a pure number, independent of units of the variables. This measure, however, suffers from certain limitations, given below:

1. Coefficient of correlation r does not give any idea about the existence of cause and effect relationship between the variables. It is possible that a high value of r is obtained although none of them seem to be directly affecting the other. Hence, any interpretation of r should be done very carefully.
2. It is only a measure of the degree of linear relationship between two variables. If the relationship is not linear, the calculation of r does not have any meaning.
3. Its value is unduly affected by extreme items.
4. As compared with other methods, the computations of r are cumbersome and time consuming.

Notes

Caution If the data are not uniformly spread in the relevant quadrants the value of r may give a misleading interpretation of the degree of relationship between the two variables. For example, if there are some values having concentration around a point in first quadrant and there is similar type of concentration in third quadrant, the value of r will be very high although there may be no linear relation between the variables.

Did u know? The coefficient of correlation lies between –1 and + 1.

Self Assessment

Fill in the blanks:

1. Distributions relating to a single characteristics are known as .....
2. When various units under consideration are observed simultaneously, with regard to two characteristics, we get a .....
3. Study of 'Correlation' is meant to determine whether there exists some sort of ..... between the variables.
4. ..... is the degree of association between two or more variables.
5. Correlation is an analysis of ..... between two or more variables.
6. Correlation analysis attempts to determine the ..... between variables.
7. ..... is a numerical measure of the degree of association between two or more variables.
8. A correlation coefficient calculated from the data on quantity demanded and corresponding price of tea would only reveal that the degree of association between them is very .....
9. The coefficient of correlation lies between ..... and .....
10. A .....of the data helps in having a visual idea about the nature of association between two variables.

8.2 Spearman's Rank Correlation

This is a crude method of computing correlation between two characteristics. In this method, various items are assigned ranks according to the two characteristics and a correlation is computed between these ranks. This method is often used in the following circumstances:

1. When the quantitative measurements of the characteristics are not possible, e.g., the results of a beauty contest where various individuals can only be ranked.
2. Even when the characteristics is measurable, it is desirable to avoid such measurements due to shortage of time, money, complexities of calculations due to large data, etc.
3. When the given data consist of some extreme observations, the value of Karl Pearson's coefficient is likely to be unduly affected. In such a situation the computation of the rank correlation is preferred because it will give less importance to the extreme observations.
4. It is used as a measure of the degree of association in situations where the nature of population, from which data are collected, is not known.

Notes

The coefficient of correlation obtained on the basis of ranks is called 'Spearman's Rank Correlation' or simply the 'Rank Correlation'. This correlation is denoted by ρ (rho) .

Let X_i be the rank of i th individual according to the characteristics X and Y_i be its rank according to the characteristics Y. If there are n individuals, there would be n pairs of ranks (X_i, Y_i), i = 1, 2, ... n. We assume here that there are no ties, i.e., no two or more individuals are tied to a particular rank. Thus, X_i's and Y_i's are simply integers from 1 to n, appearing in any order.

The means of X and Y, i.e., $\overline{X} = \overline{Y} = \frac{1+2+\cdots\cdots n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$. Also,

$$\sigma_X^2 = \sigma_Y^2 = \frac{1}{n}[1^2+2^2+\cdots+n^2] - \frac{(n+1)^2}{4} = \frac{1}{n}\left[\frac{n(n+1)(2n+1)}{6}\right] - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}$$

Let d_i be the difference in ranks of the i th individual, i.e.,

$$d_i = X_i - Y_i = \left(X_i - \overline{X}\right) - \left(Y_i - \overline{Y}\right) \qquad \left(\because \overline{X} = \overline{Y}\right)$$

Squaring both sides and taking sum over all the observations, we get

$$\sum d_i^2 = \sum\left[\left(X_i - \overline{X}\right) - \left(Y_i - \overline{Y}\right)\right]^2$$

$$= \sum\left(X_i - \overline{X}\right)^2 + \sum\left(Y_i - \overline{Y}\right)^2 - 2\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

Dividing both sides by n, we get

$$\frac{1}{n}\sum d_i^2 = \frac{1}{n}\sum\left(X_i - \overline{X}\right)^2 + \frac{1}{n}\sum\left(Y_i - \overline{Y}\right)^2 - \frac{2}{n}\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

$$= \sigma_X^2 + \sigma_Y^2 - 2Cov(X,Y) = 2\sigma_X^2 - 2Cov(X,Y) \qquad \left(\because \sigma_X^2 = \sigma_Y^2\right)$$

$$= 2\sigma_X^2 - 2\rho\sigma_X\sigma_Y = 2\sigma_X^2 - 2\rho\sigma_X^2 = 2\sigma_X^2(1-\rho) \qquad \left(\because \rho = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}\right)$$

From this, we can write $1-\rho = \frac{1}{n}\times\frac{\sum d_i^2}{2\sigma_X^2}$

$$\text{or } \rho = 1 - \frac{1}{n}\times\frac{\sum d_i^2}{2\sigma_X^2} = 1 - \frac{1}{n}\times\frac{\sum d_i^2}{2}\times\frac{12}{n^2-1} = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)}$$

Note: This formula is not applicable in case of a bivariate frequency distribution.

8.2.1 Case of Tied Ranks

In case of a tie, i.e., when two or more individuals have the same rank, each individual is assigned a rank equal to the mean of the ranks that would have been assigned to them in the event of there being slight differences in their values. To understand this, let us consider the series 20, 21, 21, 24, 25, 25, 25, 26, 27, 28. Here the value 21 is repeated two times and the value 25 is repeated three times. When we rank these values, rank 1 is given to 20. The values 21 and 21 could have been assigned ranks 2 and 3 if these were slightly different from each other. Thus, each value will be assigned a rank equal to mean of 2 and 3, i.e., 2.5. Further, the value 24 will be assigned a rank equal to 4 and each of the values 25 will be assigned a rank equal to 6, the mean of 5, 6 and 7 and so on.

Notes

Since the Spearman's formula is based upon the assumption of different ranks to different individuals, therefore, its correction becomes necessary in case of tied ranks. It should be noted that the means of the ranks will remain unaffected. Further, the changes in the variances are usually small and are neglected. However, it is necessary to correct the term Sd_i^2 and accordingly the correction factor $\frac{m(m^2-1)}{12}$, where m denotes the number of observations tied to a particular rank, is added to it for every tie. We note that there will be two correction factors, i.e., $\frac{2(4-1)}{12}$ and $\frac{3(9-1)}{12}$ in the above example.

8.2.2 Limits of Rank Correlation

A positive rank correlation implies that a high (low) rank of an individual according to one characteristic is accompanied by its high (low) rank according to the other. Similarly, a negative rank correlation implies that a high (low) rank of an individual according to one characteristic is accompanied by its low (high) rank according to the other. When $\rho = +1$, there is said to be perfect consistency in the assignment of ranks, i.e., every individual is assigned the same rank with regard to both the characteristics. Thus, we have $\Sigma d_i^2 = 0$ and hence, $\rho = 1$.

Similarly, when $\rho = -1$, an individual that has been assigned 1st rank according to one characteristic must be assigned nth rank according to the other and an individual that has been assigned 2nd rank according to one characteristic must be assigned (n – 1)th rank according to the other, etc. Thus, the sum of ranks, assigned to every individual, is equal to (n + 1), i.e., $X_i + Y_i = n + 1$ or $Y_i = (n + 1) - X_i$, for all i = 1, 2, ... n.

Further, $d_i = X_i - Y_i = X_i - (n + 1) + X_i = 2X_i - (n + 1)$

Squaring both sides, we have

$$d_i^2 = \left[2X_i - (n+1)\right]^2 = 4X_i^2 + (n+1)^2 - 4(n+1)X_i$$

Taking sum over all the observations, we have

$$\sum d_i^2 = 4\sum X_i^2 + n(n+1)^2 - 4(n+1)\sum X_i = \frac{4n(n+1)(2n+1)}{6} + n(n+1)^2 - \frac{4n(n+1)^2}{2}$$

$$= n(n+1)\left[\frac{2}{3}(2n+1) + (n+1) - 2(n+1)\right] = \frac{n(n+1)(n-1)}{3} = \frac{n(n^2-1)}{3}$$

Substituting this value in the formula for rank correlation we have

$$\rho = 1 - \frac{6n(n^2-1)}{3} \times \frac{1}{n(n^2-1)} = -1$$

Hence, the Spearman's coefficient of correlation lies between - 1 and + 1.

Example: The following table gives the marks obtained by 10 students in commerce and statistics. Calculate the rank correlation.

Marks in Statistics :	35	90	70	40	95	45	60	85	80	50
Marks in Commerce :	45	70	65	30	90	40	50	75	85	60

Notes

Solution:

Calculation Table

Marks in Statistics	Marks in Commerce	Rank of Marks in Statistics X	Rank of Marks in Commerce Y	$d_i = X_i - \overline{Y}_i$	d_i^2
35	45	1	3	–2	4
90	70	9	7	2	4
70	65	6	6	0	0
40	30	2	1	1	1
95	90	10	10	0	0
45	40	3	2	1	1
60	50	5	4	1	1
85	75	8	8	0	0
80	85	7	9	–2	4
50	60	4	5	–1	1

From the above table, we have $\sum d_i^2 = 16.$

$\therefore$ Rank Correlation $\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 16}{10\times 99} = 0.903$

Example: Calculate the Spearman's rank correlation coefficient between the series A and B given below:

Series A	:	57	59	62	63	64	65	55	58	57
Series B	:	113	117	126	126	130	129	111	116	112

Solution:

Calculation Table

Series A	Series B	(X) Rank of Series A	(Y) Rank of Series B	$d_i = X_i - Y_i$	d_i^2
57	113	2.5	3	–0.5	0.25
59	117	5	5	0	0
62	126	6	6.5	–0.5	0.25
63	126	7	6.5	0.5	0.25
64	130	8	9	–1	1.00
65	129	9	8	1	1.00
55	111	1	1	0	0
58	116	4	4	0	0
57	112	2.5	2	0.5	0.25

From the above table, we get $\sum d_i^2 = 3.00$

Since there are two ties and two observations are tied in each case, therefore, the correction factor will be $\frac{m\left(m^2-1\right)}{12} + \frac{m\left(m^2-1\right)}{12} = \frac{m\left(m^2-1\right)}{6} = \frac{2(4-1)}{6} = 1$

Thus, corrected $\sum d_i^2 = 3.00 + 1 = 4.00$

Now, $\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 4}{9\times 80} = 0.967$

Notes

Notes

Probable error of r

It is an old measure to test the significance of a particular value of r without the knowledge of test of hypothesis.

According to Horace Secrist "The probable error of correlation coefficient is an amount which if added to and subtracted from the mean correlation coefficient, gives limits within which the chances are even that a coefficient of correlation from a series selected at random will fall."

Since standard error of r, i.e., $S.E._r = \frac{1-r^2}{\sqrt{n}}$, $\therefore P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$

Did u know?

1. Probable error of r, denoted by P.E.(r) is 0.6745 times its standard error.
2. The value 0.6745 is obtained from the fact that in a normal distribution covers 50% of the total distribution.

Task Calculate Karl Pearson's coefficient of correlation of first 5 coprime numbers. Also find the rank correlation for the same data. Observe the difference, if any.

Self Assessment

Multiple Choice Questions:

11. Spearman's rank correlation is a ..... method of computing correlation between two characteristics.
    (a) Simple (b) Complex
    (c) Rude (d) Crude
12. In Spearman's rank correlation, various items are assigned ..... according to the two characteristics and a correlation is computed between these ranks.
    (a) Ranks (b) Tanks
    (c) Numbers (d) Standard
13. When the given data consist of some ..... observations, the value of karl pearson's coefficient is likely to be unduly affected.
    (a) Special (b) Extreme
    (c) General (d) Usual
14. The coefficient of correlation obtained on the basis of ranks is called .....
    (a) Spearman's rank correlation (b) Correlation
    (c) Regression (d) Karl Pearson correlation

Notes

15. A .....rank correlation implies that a high (low) rank of an individual according to one characteristic is accompanied by its high (low) rank according to the other.

(a) Positive (b) Negative

(c) Neutral (d) Zero

Case Study

Rank Correlation

A teacher ranked seven of his pupils according to their academic achievement. The order of achievement from high to low, together with family income of each pupil is given as follows:

Ray (₹ 9,200), Bhatnagar (₹ 4,500), Kaul (₹ 6,500), Das (₹ 8,000), Aggarwal (₹ 27,000), Banerjee (₹ 17,500) and Kannan (₹ 16,700). Compute rank correlation between academic achievement and family income.

8.3 Summary

Summary of Formulae

- $$r = \frac{n\sum X_iY_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{n\sum X_i^2 - \left(\sum X_i\right)^2}\sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}}$$ (without deviations)
- $$r = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum\left(X_i - \bar{X}\right)^2}\sqrt{\sum\left(Y_i - \bar{Y}\right)^2}}$$ (when deviations are taken from means)
- $$r = \frac{n\sum u_iv_i - \left(\sum u_i\right)\left(\sum v_i\right)}{\sqrt{n\sum u_i^2 - \left(\sum u_i\right)^2}\sqrt{n\sum v_i^2 - \left(\sum v_i\right)^2}}$$, where $u_i = \frac{X_i - A}{h}$ and $v_i = \frac{Y_i - B}{k}$
- $$r = \frac{N\sum\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f_j'v_j\right)}{\sqrt{N\sum f_iu_i^2 - \left(\sum f_iu_i\right)^2}\sqrt{N\sum f_j'v_j^2 - \left(\sum f_j'v_j\right)^2}}$$, where $N = \sum\sum f_{ij}$
- Standard Error of $r = \frac{1 - r^2}{\sqrt{n}}$
- Probable Error of r = 0.6745 × SE_r
- $$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$ (when ranks are not repeated)
- $$\rho = 1 - \frac{6\left[\sum d_i^2 + \frac{m\left(m^2 - 1\right)}{12} + \frac{m\left(m^2 - 1\right)}{12} + \cdots\cdots\right]}{n\left(n^2 - 1\right)}$$ (repeated ranks)

Notes

- $r_C = \pm\sqrt{\pm\left(\frac{2C - D}{D}\right)}$ (concurrent deviation formula)

8.4 Keywords

Bivariate Distribution: When various units under consideration are observed simultaneously, with regard to two characteristics, we get a Bivariate Distribution

Correlation: When the relationship is 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.

Correlation analysis: Correlation analysis attempts to determine the 'degree of relationship' between variables.

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

Dots of the diagram: Each pair of values (Xi, Yi) is denoted by a point on the graph. The set of such points (also known as dots of the diagram.

Scatter Diagram: Let the bivariate data be denoted by (Xi, Yi), where i = 1, 2 ..... n. In order to have some idea about the extent of association between variables X and Y, each pair (Xi, Yi), i = 1, 2.....n, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

Spearman's Rank Correlation: This is a crude method of computing correlation between two characteristics. In this method, various items are assigned ranks according to the two characteristics and a correlation is computed between these ranks.

Univariate Distribution: Distributions relating to a single characteristics are known as univariate Distribution.

8.5 Review Questions

1. Define correlation between two variables. Distinguish between positive and negative correlation. Illustrate by using diagrams.
2. Write down an expression for the Karl Pearson's coefficient of linear correlation. Why is it termed as the coefficient of linear correlation? Explain.
3. "If two variables are independent the correlation between them is zero, but the converse is not always true". Explain the meaning of this statement.
4. Distinguish between the Spearman's coefficient of rank correlation and Karl Pearson's coefficient of correlation. Explain the situations under which Spearman's coefficient of rank correlation can assume a maximum and a minimum value. Under what conditions will Spearman's formula and Karl Pearson's formula give equal results?
5. Write short notes on scatter diagram.
6. Compute Karl Pearson's coefficient of correlation from the following data:

X	:	8	11	15	10	12	16
Y	:	6	9	11	7	9	12

Notes

7. Calculate Karl Pearson's coefficient of correlation between the marks obtained by 10 students in economics and statistics.

Roll No.	:	1	2	3	4	5	6	7	8	9	10
Marks in eco.	:	23	27	28	29	30	31	33	35	36	39
Marks in stat.	:	18	22	23	24	25	26	28	29	30	32

8. Find Karl Pearson's coefficient of correlation from the following data and interpret its value.

Wages (Rs)	:	100	101	103	102	100	99	97	98	96	95
Cost of Living (Rs)	:	98	99	99	97	95	92	95	94	90	91

9. Find the coefficient of correlation between X and Y. Assume 69 and 112 as working origins for X and Y respectively.

X	:	78	89	96	69	59	79	68	61
Y	:	125	137	156	112	107	136	123	108

10. (a) Calculate the coefficient of correlation from the following data and interpret the result.

$$\sum XY = 8425, \bar{X} = 28.5, \bar{Y} = 28.0, \sigma_X = 10.5, \sigma_Y = 5.6, \text{and } n = 10$$

(b) Draw a scatter diagram of the following data and indicate whether the correlation between the variables is positive or negative.

Height (inches)	:	62	72	70	60	67	70	64	65	60	70
Weight (lbs.)	:	50	65	63	52	56	60	59	58	54	65

11. The coefficient of rank correlation of the marks obtained by 10 students in biology and chemistry was found to be 0.8. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 2 instead of 5. Find the correct value of coefficient of correlation.

12. Rank correlation coefficient for a certain number of pairs of observations was found to be 0.75. If the sum of squares of the differences between the corresponding ranks is 91, find the number of pairs.

13. Coefficient of rank correlation and the sum of squares of differences in corresponding ranks are 0.9021 and 28 respectively. Determine the number of pairs of observations.

Answers: Self Assessment

1.	Univariate Distribution	2.	Bivariate Distribution
3.	association	4.	Correlation
5.	covariation	6.	'degree of relationship'
7.	Correlation Coefficient	8.	high
9.	–1,+1	10.	scatter diagram
11.	(d)	12.	(a)
13.	(b)	14.	(a)
15.	(a)		

Notes

8.6 Further Readings

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Bhardwaj R.S., *Business Statistics*, Excel Books.

Garrett H.E. (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, Mc Graw Hill Book Company, New York.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hannagan T.J. (1982), *Mastering Statistics*, The Macmillan Press Ltd., Surrey.

Hooda R. P., *Statistics for Business and Economics*, Macmillan India, Delhi, 2008.

Jaeger R.M (1983), *Statistics: A Spectator Sport*, Sage Publications India Pvt. Ltd., New Delhi.

Lindgren B.W. (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Selvaraj R., Loganathan, C. *Quantitative Methods in Management.*

Sharma J.K., *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutta.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Online links

http://www.mathsisfun.com/data/correlation.html

http://www.mathsisfun.com/data/scatter-xy-plots.html

http://www.experiment-resources.com/correlation-and-regression.html

Notes

Unit 9: Regression Analysis

CONTENTS

Objectives

Introduction

9.1 Two Lines of Regression

9.1.1 Line of Regression of Y on X

9.1.2 Line of Regression of X on Y

9.1.3 Correlation Coefficient and the two Regression Coefficients

9.1.4 Regression Coefficient in a Bivariate Frequency Distribution

9.2 Least Square Methods

9.2.1 Fitting of Linear Trend

9.2.2 Fitting of Parabolic Trend

9.2.3 Fitting of Exponential Trend

9.3 Summary

9.4 Keywords

9.5 Review Questions

9.6 Further Readings

Objectives

After studying this unit, you will be able to:

- Define the term regression equation
- State the relevance of regression equation in statistics
- Discuss two lines of regression
- Analyze correlation coefficient and the two regression coefficients
- Explain various methods of least square and mention its merits and demerits

Introduction

If the coefficient of correlation calculated for bivariate data (X_i, Y_i), i = 1,2, ... n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as *regression equation* in statistics. Since the coefficient of correlation is measure of the degree of linear association of the variables, we shall discuss only linear regression equation. This does not, however, imply the non-existence of non-linear regression equations.

The regression equations are useful for predicting the value of dependent variable for given value of the independent variable. As pointed out earlier, the nature of a regression equation is different from the nature of a mathematical equation, e.g., if Y = 10 + 2X is a mathematical

Notes

equation then it implies that Y is exactly equal to 20 when X = 5. However, if Y = 10 + 2X is a regression equation, then Y = 20 is an average value of Y when X = 5.

Did u know? The term regression was first introduced by Sir Francis Galton in 1877.

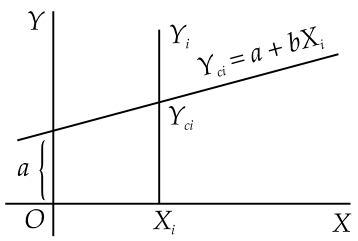
9.1 Two Lines of Regression

For a bivariate data (X_i, Y_i), i = 1,2, ... n, we can have either X or Y as independent variable. If X is independent variable then we can estimate the average values of Y for a given value of X. The relation used for such estimation is called regression of Y on X. If on the other hand Y is used for estimating the average values of X, the relation will be called regression of X on Y. For a bivariate data, there will always be two lines of regression. It will be shown later that these two lines are different, i.e., one cannot be derived from the other by mere transfer of terms, because the derivation of each line is dependent on a different set of assumptions.

9.1.1 Line of Regression of Y on X

The general form of the line of regression of Y on X is $Y_{Ci} = a + bX_i$, where Y_{Ci} denotes the average or predicted or calculated value of Y for a given value of X = X_i. This line has two constants, a and b. The constant a is defined as the average value of Y when X = 0. Geometrically, it is the intercept of the line on Y-axis. Further, the constant b, gives the average rate of change of Y per unit change in X, is known as the regression coefficient.

Figure 9.1



The above line is known if the values of a and b are known. These values are estimated from the observed data (X_i, Y_i), i = 1,2, ... n.

Note: It is important to distinguish between Y_{Ci} and Y_i. Where as Y_i is the observed value, Y_{Ci} is a value calculated from the regression equation.

Deviation taken from Actual Mean as well as from assumed mean

Using the regression $Y_{Ci} = a + bX_i$, we can obtain Y_{C1}, Y_{C2}, ... Y_{Cn} corresponding to the X values X_1, X_2, ... X_n respectively. The difference between the observed and calculated value for a particular value of X say X_i is called error in estimation of the ith observation on the assumption of a particular line of regression. There will be similar type of errors for all the n observations. We denote by $e_i = Y_i - Y_{Ci}$ (i = 1,2,.....n), the error in estimation of the ith observation. As is obvious from figure, e_i will be positive if the observed point lies above the line and will be negative if the observed point lies below the line. Therefore, in order to obtain a figure of total error, e_i's are squared and added. Let S denote the sum of squares of these errors,

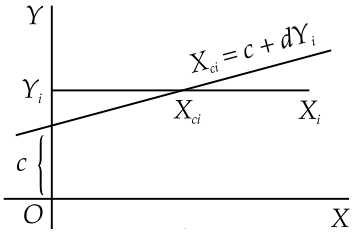
i.e., $S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - Y_{Ci})^2$.

Notes

9.1.2 Line of Regression of X on Y

The general form of the line of regression of X on Y is $X_{Ci} = c + dY_i$, where X_{Ci} denotes the predicted or calculated or estimated value of X for a given value of $Y = Y_i$ and c and d are constants. d is known as the regression coefficient of regression of X on Y.

Figure 9.2



In this case, we have to calculate the value of c and d so that

$$S' = \Sigma(X_i - X_{Ci})^2 \text{ is minimised.}$$

As in the previous section, the normal equations for the estimation of c and d are

$$\Sigma X_i = nc + d\,\Sigma Y_i \qquad \text{.... (13)}$$

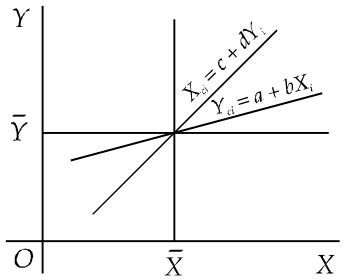
and $$\Sigma X_i Y_i = c\,\Sigma Y_i + d\,\Sigma Y_i^2 \qquad \text{.... (14)}$$

Dividing both sides of equation (13) by n, we have $\overline{X} = c + d\overline{Y}$.

Graphing Regression Lines

This shows that the line of regression also passes through the point $(\overline{X}, \overline{Y})$. Since both the lines of regression passes through the point $(\overline{X}, \overline{Y})$, therefore $(\overline{X}, \overline{Y})$ is their point of intersection as shown in Figure 9.3.

Figure 9.3



We can write $c = \overline{X} - d\overline{Y}$ .... (15)

As before, the various expressions for d can be directly written, as given below.

$$d = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{\sum Y_i^2 - n\overline{Y}^2} \qquad \text{.... (16)}$$

or $$d = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (Y_i - \overline{Y})^2} \qquad \text{.... (17)}$$

Notes

or $$d = \frac{\sum x_i y_i}{\sum y_i^2} \quad \text{.... (18)}$$

$$= \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum (Y_i - \bar{Y})^2} = \frac{Cov(X,Y)}{\sigma_Y^2} \quad \text{.... (19)}$$

Also $$d = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum Y_i^2 - \left(\sum Y_i\right)^2} \quad \text{.... (20)}$$

This expression is useful for calculating the value of d. Another short-cut formula for the calculation of d is given by

$$d = \frac{h}{k}\left[\frac{n\sum u_i v_i - \left(\sum u_i\right)\left(\sum v_i\right)}{n\sum v_i^2 - \left(\sum v_i\right)^2}\right] \quad \text{.... (21)}$$

where $u_i = \frac{X_i - A}{h}$ and $v_i = \frac{Y_i - B}{k}$

Consider equation (19)

$$d = \frac{Cov(X,Y)}{\sigma_Y^2} = \frac{r\sigma_X\sigma_Y}{\sigma_Y^2} = r \cdot \frac{\sigma_X}{\sigma_Y} \quad \text{.... (22)}$$

Substituting the value of c from equation (15) into line of regression of X on Y we have

$$X_{Ci} = \bar{X} - d\bar{Y} + dY_i \quad \text{or} \quad \left(X_{Ci} - \bar{X}\right) = d\left(Y_i - \bar{Y}\right) \quad \text{.... (23)}$$

or $$\left(X_{Ci} - \bar{X}\right) = r \cdot \frac{\sigma_X}{\sigma_Y}\left(Y_i - \bar{Y}\right) \quad \text{.... (24)}$$

This shows that the line of regression also passes through the point $(\bar{X}, \bar{Y})$. Since both the lines of regression passes through the point $(\bar{X}, \bar{Y})$, therefore $(\bar{X}, \bar{Y})$ is their point of intersection as shown in Figure 9.3.

9.1.3 Correlation Coefficient and the two Regression Coefficients

Since $b = r \cdot \frac{\sigma_Y}{\sigma_X}$ and $d = r \cdot \frac{\sigma_X}{\sigma_Y}$ we have

$b.d = r\frac{\sigma_Y}{\sigma_X} \cdot r\frac{\sigma_X}{\sigma_Y} = r^2$ or $r = \sqrt{b.d}$ This shows that correlation coefficient is the geometric mean of the two regression coefficients.

Example: From the data given below, find:

1. The two regression equations.
2. The coefficient of correlation between marks in economics and statistics.

Notes

3. The most likely marks in statistics when marks in economics are 30.

Marks in Eco. : 25 28 35 32 31 36 29 38 34 32

Marks in Stat. : 43 46 49 41 36 32 31 30 33 39

Solution:

Calculation table

Marks in Eco. (X)	Marks in Stat. (Y)	$u = X - 31$	$v = Y - 41$	uv	u^2	v^2
25	43	– 6	2	– 12	36	4
28	46	– 3	5	– 15	9	25
35	49	4	8	32	16	64
32	41	1	0	0	1	0
31	36	0	– 5	0	0	25
36	32	5	– 9	– 45	25	81
29	31	– 2	– 10	20	4	100
38	30	7	– 11	– 77	49	121
34	33	3	– 8	– 24	9	64
32	39	1	– 2	– 2	1	4
Total		10	– 30	– 123	150	488

From the table, we have

$$\overline{X} = 31 + \frac{10}{10} = 32 \text{ and } \overline{Y} = 41 - \frac{30}{10} = 38.$$

1. ***The lines of regression***

(a) *Regression of Y on X*

$$b = \frac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{n\sum u^2 - \left(\sum u\right)^2} = \frac{-1230 + 300}{1500 - 100} = -0.66$$

$$a = \overline{Y} - b\overline{X} \qquad = 38 + 0.66 \times 32 = 59.26$$

∴ Regression equation is

YC = 59.26 - 0.66X

(b) Regression of X on Y

$$d = \frac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{n\sum v^2 - \left(\sum v\right)^2} = \frac{-1230 + 300}{4880 - 900} = -0.23$$

$$c = \overline{X} - d\overline{Y} = 32 + 0.23 \times 38 = 40.88$$

∴ Regression equation is

XC = 40.88 - 0.23Y

Notes

2. ***Coefficient of correlation***

$$r = \sqrt{b \cdot d} = -\sqrt{-0.66 \times -0.23} = -0.39$$

Note that r, b and d are of same sign.

Since we have to estimate marks in statistics denoted by Y, therefore, regression of Y on X will be used. The most likely marks in statistics when marks in economics are 30, is given by

$$Y_C = 59.26 - 0.66 \times 30 = 39.33$$

Example: For a bivariate data, you are given the following information:

$\Sigma(X - 58) = 46$ $\quad$ $\Sigma(X - 58)^2 = 3086$

$\Sigma(Y - 58) = 9$ $\quad$ $\Sigma(Y - 58)^2 = 483$

$\Sigma(X - 58)(Y - 58) = 1095.$

Number of pairs of observations = 7. You are required to determine (i) the two regression equations and (ii) the coefficient of correlation between X and Y.

Solution:

Let u = X – 58 and v = Y – 58. In terms of our notations, we are given Su = 46, Su2 = 3086, Sv = 9, Sv2 = 483, Suv = 1095 and n = 7.

$$\text{Now } \overline{X} = 58 + \frac{46}{7} = 64.7 \quad \text{and} \quad \overline{Y} = 58 + \frac{9}{7} = 59.29$$

1. For regression equation of Y on X, we have

$$b = \frac{7 \times 1095 - 46 \times 9}{7 \times 3086 - (46)^2} = 0.37$$

and $a = \overline{Y} - b\overline{X} = 59.29 - 0.37 \times 64.57 = 35.40$

$\therefore$ The line of regression of Y on X is given by

$$Y_C = 35.40 + 0.37X$$

2. For regression equation of X on Y, we have

$$d = \frac{7 \times 1095 - 46 \times 9}{7 \times 483 - (9)^2} = 2.20$$

and $c = \overline{X} - d\overline{Y} = 64.57 - 2.2 \times 59.29 = -65.87$

$\therefore$ The line of regression of X on Y is given by

$$X_C = -65.87 + 2.2Y$$

3. The coefficient of correlation

$$r = \sqrt{b \cdot d} = \sqrt{0.37 \times 2.2} = 0.90$$

Notes

9.1.4 Regression Coefficient in a Bivariate Frequency Distribution

As in case of calculation of correlation coefficient, we can directly write the formula for the two regression coefficients for a bivariate frequency distribution as given below :

$$b = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f'_jY_j\right)}{N\sum f_iX_i^2 - \left(\sum f_iX_i\right)^2}$$

or, if we define $u_i = \dfrac{X_i - A}{h}$ and $v_j = \dfrac{Y_j - B}{k}$

$$b = \frac{k}{h}\left[\frac{N\sum\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f'_jv_j\right)}{N\sum f_iu_i^2 - \left(\sum f_iu_i\right)^2}\right]$$

Similarly, $$d = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f'_jY_j\right)}{N\sum f'_jY_j^2 - \left(\sum f'_jY_j\right)^2}$$

or $$d = \frac{h}{k}\left[\frac{N\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f'_jv_j\right)}{N\sum f'_jv_j^2 - \left(\sum f'_jv_j\right)^2}\right]$$

Caution A different line of regression means a different pair of constants a and b.

Self Assessment

Fill in the blanks:

1. Study of ..... meant to determine the most suitable form of the relationship between the variables given that they are correlated.
2. If the coefficient of correlation calculated for bivariate data (X_i, Y_i), i = 1,2, ... n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as ..... in statistics.
3. Coefficient of correlation is measure of the degree of .....of the variables.
4. The regression equations are useful for predicting the value of ..... variable for given value of the independent variable.
5. The nature of a regression equation is ..... the nature of a mathematical equation.

Multiple Choice Questions:

6. The term regression was first introduced by Sir Francis Galton in .....

   (a) 1857 (b) 1871

   (c) 1877 (d) 1987

Notes

7. If X is independent variable then we can estimate the average values of Y for a given value of X. The relation used for such estimation is called regression of .....

   (a) X on X (b) Y on Y

   (c) X on Y (d) Y on X

8. If Y is used for estimating the average values of X, the relation will be called regression of .....

   (a) X on X (b) Y on Y

   (c) X on Y (d) Y on X

9. For a bivariate data, there will always be .....of regression.

   (a) Single line (b) Two lines

   (c) Three lines (d) Four lines

10. Derivation of each line is dependent on a different set of .....

   (a) Functions (b) Assumptions

   (c) Symbols (d) Presumptions

9.2 Least Square Methods

This is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, are minimised. We shall use this method in the fitting of following trends:

1. Linear Trend
2. Parabolic Trend
3. Exponential Trend

9.2.1 Fitting of Linear Trend

Given the data (Y_t, t) for n periods, where t denotes time period such as year, month, day, etc., we have to find the values of the two constants, a and b, of the linear trend equation $Y_t = a + bt$.

Using the least square method, the normal equation for obtaining the values of a and b are:

$$\Sigma Y_t = na + b\Sigma t \quad \text{and}$$
$$\Sigma tY_t = a\Sigma t + b\Sigma t^2$$

Let X = t – A, such that $\Sigma X = 0$, where A denotes the year of origin.

The above equations can also be written as

$$\Sigma Y = na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

(Dropping the subscript t for convenience).

Since $\Sigma X = 0$, we can write $a = \frac{\sum Y}{n}$ and $b = \frac{\sum XY}{\sum X^2}$

Notes

9.2.2 Fitting of Parabolic Trend

The mathematical form of a parabolic trend is given by $Y_t = a + bt + ct^2$ or $Y = a + bt + ct^2$ (dropping the subscript for convenience). Here a, b and c are constants to be determined from the given data.

Using the method of least squares, the normal equations for the simultaneous solution of a, b, and c are:

$$\Sigma Y = na + b\Sigma t + c\Sigma t^2$$
$$\Sigma tY = a\Sigma t + b\Sigma t^2 + c\Sigma t^3$$
$$\Sigma t^2Y = a\Sigma t^2 + b\Sigma t^3 + c\Sigma t^4$$

By selecting a suitable year of origin, i.e., define X = t - origin such that SX = 0, the computation work can be considerably simplified. Also note that if SX = 0, then SX3 will also be equal to zero. Thus, the above equations can be rewritten as:

$$\Sigma Y = na + c\Sigma X^2 \quad ..(i)$$

$$\Sigma XY = b\Sigma X^2 \quad ...(ii)$$

$$\Sigma X^2Y = a\Sigma X^2 + c\Sigma X^4 \quad ...(iii)$$

From equation (ii), we get $b = \dfrac{\sum XY}{\sum X^2}$ .... (iv)

Further, from equation (i), we get $a = \dfrac{\sum Y - c\sum X^2}{n}$ .... (v)

And from equation (iii), we get $c = \dfrac{n\sum X^2Y - \left(\sum X^2\right)\left(\sum Y\right)}{n\sum X^4 - \left(\sum X^2\right)^2}$ .... (vi)

Thus, equations (iv), (v) and (vi) can be used to determine the values of the constants a, b and c.

9.2.3 Fitting of Exponential Trend

The general form of an exponential trend is Y = a.bt, where a and b are constants to be determined from the observed data.

Taking logarithms of both sides, we have *log*Y = *log* a + t *log* b.

This is a linear equation in log Y and t and can be fitted in a similar way as done in case of linear trend. Let A = log a and B = log b, then the above equation can be written as log Y = A + Bt.

The normal equations, based on the principle of least squares are:

$$\Sigma \log Y = nA + B\Sigma t$$

and $$\Sigma t\log Y = A\Sigma t + B\Sigma t^2.$$

Notes

By selecting a suitable origin, i.e., defining X = t - origin, such that SX = 0, the computation work can be simplified. The values of A and B are given by $A = \frac{\sum \log Y}{n}$ and $B = \frac{\sum X \log Y}{\sum X^2}$ respectively. Thus, the fitted trend equation can be written as log Y = A + BX

or Y = Antilog [A + BX] = Antilog [log a + X log b]

= Antilog [log $a.b^X$] = $a.b^X$.

Notes

Merits and Demerits of Least Squares Method

Merits

1. Given the mathematical form of the trend to be fitted, the least squares method is an objective method.
2. Unlike the moving average method, it is possible to compute trend values for all the periods and predict the value for a period lying outside the observed data.
3. The results of the method of least squares are most satisfactory because the fitted trend satisfies the two important properties, i.e., (i) $\Sigma(Y_o - Y_t) = 0$ and (ii) $\Sigma(Y_o - Y_t)^2$ is minimum. Here Yo denotes the observed value and Yt denotes the calculated trend value.

   The first property implies that the position of fitted trend equation is such that the sum of deviations of observations above and below this is equal to zero. The second property implies that the sum of squares of deviations of observations, about the trend equation, are minimum.

Demerits

1. As compared with the moving average method, it is a cumbersome method.
2. It is not flexible like the moving average method. If some observations are added, then the entire calculations are to be done once again.
3. It can predict or estimate values only in the immediate future or past.
4. The computation of trend values, on the basis of this method, doesn't take into account the other components of a time series and hence not reliable.
5. Since the choice of a particular trend is arbitrary, the method is not, strictly, objective.
6. This method cannot be used to fit growth curves, the pattern followed by the most of the economic and business time series.

Task Study various methods used for fitting of trends and analyze them.

Self Assessment

State whether the following statements are true or false:

11. Least square root method is one of the most popular methods of fitting a mathematical trend.

Notes

12. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, are minimized.
13. The procedure for calculation of the two constants is highly different for even and odd number of observations.
14. The general form of linear trend equation is $Y_t = a + bt$.
15. The mathematical form of a parabolic trend is given by $Yt = a + bt - ct^2$ or $Y = a + bt - ct^2$
16. The general form of an exponential trend is Y = a.bt
17. Given the mathematical form of the trend to be fitted, the least squares method is an descriptive method.
18. The results of the method of least squares are most satisfactory.
19. Least square method can be used to fit growth curves.
20. Least square method estimate values only in the immediate future or past.

Case Study

Average Growth Rate

The weights (in lbs.) of a newly born calf are taken at weekly intervals. Below are the observations for ten weeks.

Age (X weeks) :	1	2	3	4	5	6	7	8	9	10
Weight (Y lbs.) :	52.5	58.7	65.0	70.2	75.4	81.1	87.2	95.5	102.2	108.4

Let Y = a + bu, where u = 2X - 11. Use normal equations to estimate a and b. Use these values to obtain the line of best fit of Y on X and write down the average rate of growth of weight of the calf per week.

9.3 Summary

Regression of Y on X

- ***Regression coefficient:*** $b = \dfrac{\sum XY - n\bar{X}\,\bar{Y}}{\sum X^2 - n\bar{X}^2} = \dfrac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2}$

  Also b $= \dfrac{Cov\left(X,Y\right)}{\sigma_X^2} = r \cdot \dfrac{\sigma_Y}{\sigma_X}$
- ***Change of scale and origin:***

  $$\text{If } u = \frac{X-A}{h} \text{ and } v = \frac{Y-B}{h}, \text{ then } b = \frac{k}{h}\left[\frac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{n\sum u^2 - \left(\sum u\right)^2}\right]$$
- ***Constant term:*** $a = \bar{Y} - b\bar{X}$
- ***Alternative form of regression equation:***

  $$Y_C - \bar{Y} = \left(X - \bar{X}\right) \quad or \quad Y_C - \bar{Y} = r \cdot \frac{\sigma_Y}{\sigma_X}\left(X - \bar{X}\right)$$

Notes

- Regression coefficient in bivariate frequency distribution

$$b = \frac{k}{h}\left[\frac{N\sum\sum f_{ij}u_i v_j - \left(\sum f_i u_i\right)\left(\sum f_j' v_j\right)}{N\sum f_i u_i^2 - \left(\sum f_i u_i\right)^2}\right]$$

- Standard Error of the estimate

$s_{Y.X} = \sigma_Y\sqrt{1-r^2}$ for large n (i.e., n > 30)

$$= \sqrt{\frac{\sum\left(Y_i - \bar{Y}\right)^2\left(1-r^2\right)}{n-2}}$$ for small n

Regression of X on Y

- Regression Coefficient $d = \dfrac{\sum XY - n\bar{X}\,\bar{Y}}{\sum Y^2 - n\bar{Y}^2} = \dfrac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum Y^2 - \left(\sum Y\right)^2}$

$$= \frac{Cov\left(X,Y\right)}{\sigma_Y^2} = r\cdot\frac{\sigma_X}{\sigma_Y}$$

- Change of scale and origin

If $u = \dfrac{X-A}{h}$ and $v = \dfrac{Y-B}{h}$, then $d = \dfrac{h}{k}\left[\dfrac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{n\sum v^2 - \left(\sum v\right)^2}\right]$

- Constant term $c = \bar{X} - d\bar{Y}$
- Alternative form of regression equation

$$X_C - \bar{X} = d\left(Y - \bar{Y}\right) \quad or \quad X_C - \bar{X} = r.\frac{\sigma_X}{\sigma_Y}\left(Y - \bar{Y}\right)$$

- Regression coefficient in a bivariate frequency distribution

$$d = \frac{h}{k}\left[\frac{N\sum\sum f_{ij}u_i v_j - \left(\sum f_i u_i\right)\left(\sum f_j' v_j\right)}{N\sum f_j' v_j^2 - \left(\sum f_j' v_j\right)^2}\right]$$

- Standard error of the estimate

$s_{X.Y} = \sigma_X\sqrt{1-r^2}$ for large n (i.e., n > 30)

$$= \sqrt{\frac{\sum\left(X_i - \bar{X}\right)^2\left(1-r^2\right)}{n-2}}$$ for small n

- Relation of r with b and d

$$b\times d = r\cdot\frac{\sigma_Y}{\sigma_X}\cdot r\cdot\frac{\sigma_X}{\sigma_Y} = r^2$$

or $r = \sqrt{b\times d}$

Notes

9.4 Keywords

Exponential trend: The general form of an exponential trend is $Y = a.b^t$, where a and b are constants.

Least square methods: This is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, are minimized.

Line of Regression Y on X: The general form of the line of regression of Y on X is $Y_{Ci} = a + bX_i$, where Y_{Ci} denotes the average or predicted or calculated value of Y for a given value of $X = X_i$. This line has two constants, a and b.

Line of Regression of X on Y: The general form of the line of regression of X on Y is $X_{Ci} = c + dY_i$, where X_{Ci} denotes the predicted or calculated or estimated value of X for a given value of $Y = Y_i$ and c and d are constants. d is known as the regression coefficient of regression of X on Y.

***Linear Trend*:** The linear trend equation is given by relation $Y_t = a + bt$. where t denotes time period such as year, month, day, etc., and a, b are the constants.

Parabolic Trend: The mathematical form of a parabolic trend is given by $Y_t = a + bt + ct^2$ or $Y = a + bt + ct^2$ Where a, b and c are constants.

Regression equation: If the coefficient of correlation calculated for bivariate data (X_i, Y_i), i = 1,2, ... n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as regression equation in statistics.

9.5 Review Questions

1. Distinguish between correlation and regression. Discuss least square method of fitting regression.
2. What do you understand by linear regression? Why there are two lines of regression? Under what condition(s) can there be only one line?
3. Define the regression of Y on X and of X on Y for a bivariate data (X_i, Y_i), i = 1, 2, ... n. What would be the values of the coefficient of correlation if the two regression lines (a) intersect at right angle and (b) coincide?
4. (a) Show that the proportion of variations explained by a regression equation is r^2.

   (b) What is the relation between Total Sum of Squares (TSS), Explained Sum of Squares (ESS) and Residual Sum of squares (RSS)? Use this relationship to prove that the coefficient of correlation has a value between –1 and +1.
5. Write a note on the standard error of the estimate.
6. "The regression line gives only a 'best estimate' of the quantity in question. We may assess the degree of uncertainty in this estimate by calculating its standard error ". Explain.
7. Show that the coefficient of correlation is the geometric mean of the two regression coefficients.
8. What is the method of least squares? Show that the two lines of regression obtained by this method are irreversible except when r = ± 1. Explain.

Notes

9. Show that, in principle, there are always two lines of regression for a bivariate data. Prove that the coefficient of correlation between two variables is either + 1 or –1 when the two lines are identical and is zero when they are perpendicular.

10. Fit a linear regression of Y on X to the following data:

X	:	1	2	3	4	5	6	7	8
Y	:	65	80	45	86	178	205	200	250

11. The following table gives the data relating to purchases and sales. Obtain the two regression equations by the method of least squares and estimate the likely sales when purchases equal 100.

Purchases	:	62	72	98	76	81	56	76	92	88	49
Sales	:	112	124	131	117	132	96	120	136	97	85

12. The following table gives the marks of ten students in economics (X) and statistics (Y). Compute the appropriate regression equation to estimate the marks in statistics of a student who scored 65 marks in economics.

X	:	54	50	63	65	50	65	54	55	61	60
Y	:	65	58	78	72	62	72	60	63	66	70

13. In a partially destroyed record the following data are available:

    The two regression lines are 5X + 3Y = 290 and 3X + 2Y = 180. The variance of X = 16.

    Find (a) Mean values of X and Y

    (b) Standard deviation of Y

    (c) Coefficient of correlation between X and Y.

14. The two regression lines obtained by a student were as given below:

    3X - 4Y = 5

    8X + 16Y = 15

    Do you agree with him? Explain with reasons.

15. Obtain the lines of regression of Y on X and X on Y for the data given below:

$$\sum X = 50, \sum Y = 60, \sum XY = 350, n = 10, \sigma_X^2 = 4 \text{ and } \sigma_Y^2 = 9$$

Answers: Self Assessment

1.	'Regression'	2.	regression equation
3.	linear association	4.	dependent
5.	different from	6.	(c)
7.	(d)	8.	(c)
9.	(b)	10.	(b)
11.	False	12.	True
13.	False	14.	True
15.	False	16.	True
17.	False	18.	True
19.	False	20.	True

9.6 Further Readings

Notes

Books

Bhardwaj R S., *Business Statistics*, Excel Books.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hooda R.P., *Statistics for Business and Economics*, Macmillan India Delhi, 2008.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Sharma J.K, *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Online links

http://www.experiment-resources.com/correlation-and-regression.html

http://www.managers-net.com/regression.html

http://www.investopedia.com/terms/r/regression.asp#axzz1VvD6cSeJ

http://en.wikipedia.org/wiki/Regression_analysis

http://www.law.uchicago.edu/files/files/20.Sykes_.Regression.pdf

Notes

Unit 10: Index Number

CONTENTS

Objectives

Introduction

10.1 Definitions and Characteristics of Index Numbers

10.2 Uses of Index Numbers

10.3 Construction of Index Numbers

10.4 Notations and Terminology

10.5 Price Index Numbers

10.6 Quantity Index Numbers

10.7 Value Index Number

10.8 Comparison of Laspeyres's and Paasche's Index Numbers

10.9 Relation between Weighted Aggregative and Weighted Arithmetic Average of Price Relatives Index Numbers

10.9.1 Change in the Cost of Living due to Change in Price of an Item

10.10 Chain Base Index Numbers

10.10.1 Chained Index Numbers

10.10.2 Conversion of Chain Base Index Number into Fixed Base Index Number and vice-versa

10.11 Summary

10.12 Keywords

10.13 Review Questions

10.14 Further Readings

Objectives

After studying this unit, you will be able to:

- Define the term index number
- Discuss the features and uses of an index number
- Familiar with notations and terminologies used in index numbers
- Establish relation between weighted aggregative and weighted arithmetic average of price relatives index numbers
- Make comparison of Laspeyres's and Paasche's Index Numbers
- Tell about chain base index numbers

Notes

Introduction

An index number is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations. Suppose that we want to compare the average price level of different items of food in 1992 with what it was in 1990. Let the different items of food be wheat, rice, milk, eggs, ghee, sugar, pulses, etc. If the prices of all these items change in the same ratio and in the same direction; assume that prices of all the items have increased by 10% in 1992 as compared with their prices in 1990; then there will be no difficulty in finding out the average change in price level for the group as a whole. Obviously, the average price level of all the items taken as a group will also be 10% higher in 1992 as compared with prices of 1990. However, in real situations, neither the prices of all the items change in the same ratio nor in the same direction, i.e., the prices of some commodities may change to a greater extent as compared to prices of other commodities. Moreover, the price of some commodities may rise while that of others may fall. For such situations, the index numbers are very useful device for measuring the average change in prices or any other characteristics like quantity, value, etc., for the group as a whole.

Another important feature of the index number is that it is often used to average a characteristics expressed in different units for different items of a group. For example, the price of wheat may be quoted as ₹/kg., price of milk as ₹/litre, price of eggs as ₹/dozen, etc. To arrive at a single figure that expresses the average change in price for the whole group, various prices have to be combined and averaged in a suitable way. This single figure is known as price index and can be used to determine the extent and direction of average change in the prices for the group. In a similar way we can construct quantity index numbers, value index numbers, etc. It should be noted here that.

Did u know? Index numbers are specialized type of averages that are used to measure the changes in a characteristics which is not capable of being directly measured. For example, it is not possible to measure business activity in a direct way, however, relative changes in a business activity can be determined by the direct measurement of changes in some factors that affect it. Similarly, it is not possible to measure, directly, the price level of a group of items, but changes in price level can be measured by using price index numbers.

10.1 Definitions and Characteristics of Index Numbers

Some important definitions of index numbers are given below:

"An index number is a device for comparing the general level of magnitude of a group of distinct, but related, variables in two or more situations." – *Karmel and Polasek*

"An index number is a special type of average that provides a measurement of relative changes from time to time or from place to place." – *Wessell, Wilett and Simone*

"Index number shows by its variation the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice." – *Edgeworth*

"An index number is a single ratio (usually in percentage) which measures the combined (i.e., averaged) change of several variables between two different times, places or situations."

– *Tuttle*

Notes

On the basis of the above definitions, the following characteristics of index numbers are worth mentioning:

1. ***Index numbers are specialised averages:*** As we know that an average of data is its representative summary figure. In a similar way, an index number is also an average, often a weighted average, computed for a group. It is called a specialised average because the figures, that are averaged, are not necessarily expressed in homogeneous units.

2. ***Index numbers measure the changes for a group which are not capable of being directly measured: The examples of such magnitudes are:*** Price level of a group of items, level of business activity in a market, level of industrial or agricultural output in an economy, etc.

3. ***Index numbers are expressed in terms of percentages:*** The changes in magnitude of a group are expressed in terms of percentages which are independent of the units of measurement. This facilitates the comparison of two or more index numbers in different situations.

10.2 Uses of Index Numbers

The main uses of index numbers are:

1. ***To measure and compare changes:*** The basic purpose of the construction of an index number is to measure the level of activity of phenomena like price level, cost of living, level of agricultural production, level of business activity, etc. It is because of this reason that sometimes index numbers are termed as barometers of economic activity. It may be mentioned here that a barometer is an instrument which is used to measure atmospheric pressure in physics.

   The level of an activity can be expressed in terms of index numbers at different points of time or for different places at a particular point of time. These index numbers can be easily compared to determine the trend of the level of an activity over a period of time or with reference to different places.

2. ***To help in providing guidelines for framing suitable policies:*** Index numbers are indispensable tools for the management of any government or non-government organisation. For example, the increase in cost of living index is helpful in deciding the amount of additional dearness allowance that should be paid to the workers to compensate them for the rise in prices. In addition to this, index numbers can be used in planning and formulation of various government and business policies.

3. ***Price index numbers are used in deflating:*** This is a very important use of price index numbers. These index numbers can be used to adjust monetary figures of various periods for changes in prices. For example, the figure of national income of a country is computed on the basis of the prices of the year in question. Such figures, for various years often known as national income at current prices, do not reveal the real change in the level of production of goods and services. In order to know the real change in national income, these figures must be adjusted for price changes in various years. Such adjustments are possible only by the use of price index numbers and the process of adjustment, in a situation of rising prices, is known as deflating.

4. ***To measure purchasing power of money:*** We know that there is inverse relation between the purchasing power of money and the general price level measured in terms of a price index number. Thus, reciprocal of the relevant price index can be taken as a measure of the purchasing power of money.

Notes

10.3 Construction of Index Numbers

To illustrate the construction of an index number, we reconsider various items of food mentioned earlier. Let the prices of different items in the two years, 1990 and 1992, be as given below:

Item	Price in 2009 (in ₹/unit)	Price in 2011 (in ₹/unit)
1. Wheat	300/quintal	360/quintal
2. Rice	12/kg.	15/kg.
3. Milk	7/litre	8/litre
4. Eggs	11/dozen	12/dozen
5. Ghee	80/kg.	88/kg.
6. Sugar	9/kg.	10/kg.
7. Pulses	14/kg.	16/kg.

The comparison of price of an item, say wheat, in 1992 with its price in 1990 can be done in two ways, explained below:

1. By taking the difference of prices in the two years, i.e., 360 – 300 = 60, one can say that the price of wheat has gone up by ₹ 60/quintal in 1992 as compared with its price in 1990.

2. By taking the ratio of the two prices, i.e., $\frac{360}{300} = 1.20$, one can say that if the price of wheat in 1990 is taken to be 1, then it has become 1.20 in 1992. A more convenient way of comparing the two prices is to express the price ratio in terms of percentage, i.e., $\frac{360}{300} \times 100 = 120$, known as Price Relative of the item. In our example, price relative of wheat is 120 which can be interpreted as the price of wheat in 1992 when its price in 1990 is taken as 100. Further, the figure 120 indicates that price of wheat has gone up by 120 – 100 = 20% in 1992 as compared with its price in 1990.

The first way of expressing the price change is inconvenient because the change in price depends upon the units in which it is quoted. This problem is taken care of in the second method, where price change is expressed in terms of percentage. An additional advantage of this method is that various price changes, expressed in percentage, are comparable. Further, it is very easy to grasp the 20% increase in price rather than the increase expressed as ₹ 60/quintal.

For the construction of index number, we have to obtain the average price change for the group in 1992, usually termed as the Current Year, as compared with the price of 1990, usually called the Base Year. This comparison can be done in two ways:

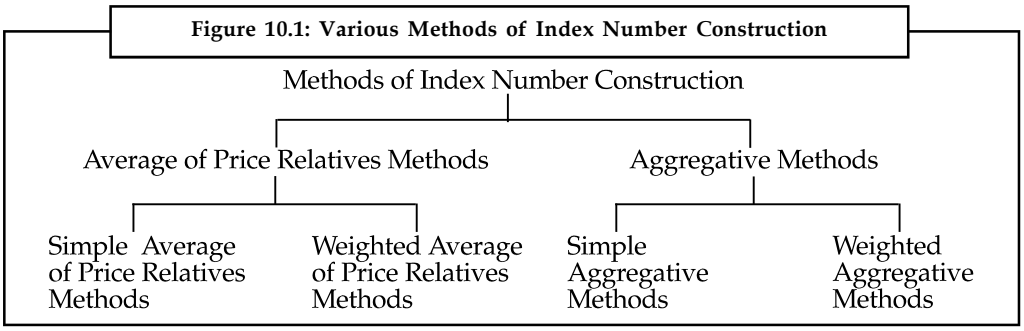
1. By taking suitable average of price relatives of different items. The methods of index number construction based on this procedure are termed as Average of Price Relative Methods.

2. By taking ratio of the averages of the prices of different items in each year. These methods are popularly known as Aggregative Methods.

Since the average in each of the above methods can be simple or weighted, these can further be divided as simple or weighted. Various methods of index number construction can be classified as shown below:

Notes

Figure 10.1: Various Methods of Index Number Construction



In addition to this, a particular method would depend upon the type of average used. Although, geometric mean is more suitable for averaging ratios, arithmetic mean is often preferred because of its simplicity with regard to computations and interpretation.

10.4 Notations and Terminology

Before writing various formulae of index numbers, it is necessary to introduce certain notations and terminology for convenience.

Base Year: The year from which comparisons are made is called the base year. It is commonly denoted by writing '0' as a subscript of the variable.

Current Year: The year under consideration for which the comparisons are to be computed is called the current year. It is commonly denoted by writing '1' as a subscript of the variable.

Let there be n items in a group which are numbered from 1 to n. Let p_0^i denote the price of the ith item in base year and p_1^i denote its price in current year, where i = 1, 2, ... n. In a similar way q_0^i and q_1^i will denote the quantities of the ith item in base and current years respectively.

Using these notations, we can write an expression for price relative of the ith item as $P_i = \frac{p_{1i}}{p_{0i}} \times 100$, and quantity relative of the i th item as $Q_i = \frac{q_{1i}}{q_{0i}} \times 100$

Further, P_{01} will be used to denote the price index number of period '1' as compared with the prices of period '0'. Similarly, Q_{01} and V_{01} would denote the quantity and the value index numbers respectively of period '1' as compared with period '0'.

10.5 Price Index Numbers

1. ***Simple Average of Price Relatives:***
   (a) When arithmetic mean of price relatives is used

   The index number formula is given by $P_{01} = \frac{\sum P_i}{n}$ or $P_{01} = \frac{\sum \frac{p_{1i}}{p_{0i}} \times 100}{n}$ Omitting the subscript i, the above formula can also be written as $P_{01} = \frac{\sum \frac{p_1}{p_0} \times 100}{n}$

   (b) When geometric mean of price relatives is used
   The index number formula is given by

$$P_{01} = \left(P_1 \times P_2 \times \cdots\cdots P_n\right)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} P_i\right)^{\frac{1}{n}} = \text{Antilog}\left[\frac{\sum \log P_i}{n}\right]$$

   (Π is used to denote the product of terms.)

Notes

Example: Given below are the prices of 5 items in 2005 and 2010. Compute the simple price index number of 1990 taking 2005 as base year. Use (a) arithmetic mean and (b) geometric mean.

Item	Price in 2005 (Rs/unit)	Price in 2010 (Rs/unit)
1	15	20
2	8	7
3	200	300
4	60	110
5	100	130

Solution:

Calculation Table

Item	Price in 2005 (P_{0i})	Price in 2010 (P_{0i})	Price Relative $P_i = \frac{p_{1i}}{p_{0i}} \times 100$	log P_i
1	15	20	133.33	2.1249
2	8	7	87.50	1.9420
3	200	300	150.00	2.1761
4	60	110	183.33	2.2632
5	100	130	130.00	2.1139
Total			684.16	10.6201

$\therefore$ Index number, using A.M., is $P_{01} = \frac{684.16}{5} = 136.83$

and Index number, using G.M., is $P_{01} = \text{Antilog}\left[\frac{10.6201}{5}\right] = 133.06$

2. ***Weighted Average of Price Relatives:*** In the method of simple average of price relatives, all the items are assumed to be of equal importance in the group. However, in most of the real life situations, different items of a group have different degree of importance. In order to take this into account, weighing of different items, in proportion to their degree of importance, becomes necessary.

   Let w_i be the weight assigned to the i th item (i = 1, 2, ... n). Thus, the index number, given by the weighted arithmetic mean of price relatives, is $P_{01} = \frac{\sum P_i w_i}{\sum w_i}$.

   Similarly, the index number, given by the weighted geometric mean of price relatives can be written as follows:

$$P_{01} = \left[P_1^{w_1}.P_2^{w_2}.....P_n^{w_n}\right]^{\frac{1}{\sum w_i}} = \left[\prod P_i^{w_i}\right]^{\frac{1}{\sum w_i}}$$

or $P_{01} = \text{Antilog}\left[\frac{\sum w_i \log P_i}{\sum w_i}\right]$

Notes

Nature of weights

While taking weighted average of price relatives, the values are often taken as weights. These weights can be the values of base year quantities valued at base year prices, i.e., $p_{0i}q_{0i}$, or the values of current year quantities valued at current year prices, i.e., $p_{1i}q_{1i}$, or the values of current year quantities valued at base year prices, i.e., $p_{0i}q_{1i}$, etc., or any other value.

Example: Construct an index number for 2002 taking 2010 as base for the following data, by using

(a) weighted arithmetic mean of price relatives and

(b) weighted geometric mean of price relatives.

Commodities	Prices in 2002	Prices in 2010	Weights
A	60	100	30
B	20	20	20
C	40	60	24
D	100	120	30
E	120	80	10

Solution:

Calculation Table

Comm - odities	Prices in 2002(p_0)	Prices in 2010 (p_1)	P.R.(P) $= \frac{p_1}{p_0} \times 100$	Wts (w)	Pw	log P	w log P
A	60	100	166.67	30	5000.1	2.2219	66.657
B	20	20	100.00	20	2000.0	2.0000	40.000
C	40	60	150.00	24	3600.0	2.1761	52.226
D	100	120	120.00	30	3600.0	2.0792	62.376
E	120	80	66.67	10	666.7	1.8239	18.239
Total				114	14866.8		239.48

$\therefore$ Index number using A.M. is $P_{01} = \frac{14866.8}{114} = 130.41$

and index number using G.M. is $P_{01} = \text{Antilog}\left[\frac{239.498}{114}\right] = 126.15$

3. ***Simple Aggregative Method:*** In this method, the simple arithmetic mean of the prices of all the items of the group for the current as well as for the base year are computed separately. The ratio of current year average to base year average multiplied by 100 gives the required index number.

Using notations, the arithmetic mean of prices of n items in current year is given by $\frac{\sum p_{0i}}{n}$

$$\therefore \text{ Simple aggregative price index } P_{01} = \frac{\frac{\sum p_{1i}}{n}}{\frac{\sum p_{0i}}{n}} \times 100 = \frac{\sum p_{1i}}{\sum p_{0i}} \times 100$$

Notes

Omitting the subscript i, the above index number can also be written as:

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

Example: The following table gives the prices of six items in the years 2010 and 2011. Use simple aggregative method to find index of 2011 with 2010 as base.

Item	Price in 2010 (₹)	Price in 2011 (₹)
A	40	50
B	60	60
C	20	30
D	50	70
E	80	90
F	100	100

Solution:

Let p_0 be the price in 2010 and p_1 be the price in 2011. Thus, we have

$Sp_0 = 350$ and $Sp_1 = 400$

$$\therefore \quad P_{01} = \frac{400}{350} \times 100 = 114.29$$

4. ***Weighted Aggregative Method:*** This index number is defined as the ratio of the weighted arithmetic means of current to base year prices multiplied by 100.

   Using the notations, defined earlier, the weighted arithmetic mean of current year prices can be written as = $\frac{\sum p_{1i} w_i}{\sum w_i}$

   Similarly, the weighted arithmetic mean of base year prices $= \frac{\sum p_{0i} w_i}{\sum w_i}$

   $$\therefore \text{ Price Index Number, } P_{01} = \frac{\frac{\sum p_{1i} w_i}{\sum w_i}}{\frac{\sum p_{0i} w_i}{\sum w_i}} \times 100 = \frac{\sum p_{1i} w_i}{\sum p_{0i} w_i} \times 100$$

   Omitting the subscript, we can also write $P_{01} = \frac{\sum p_1 w}{\sum p_0 w} \times 100$

Nature of Weights

In case of weighted aggregative price index numbers, quantities are often taken as weights. These quantities can be the quantities purchased in base year or in current year or an average of base year and current year quantities or any other quantities. Depending upon the choice of weights, some of the popular formulae for weighted index numbers can be written as follows:

Notes

1. ***Laspeyres's Index:*** Laspeyres' price index number uses base year quantities as weights. Thus, we can write:

$$P_{01}^{La} = \frac{\sum p_{1i}q_{0i}}{\sum p_{0i}q_{0i}} \times 100 \quad \text{or} \quad P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

2. ***Paasche's Index:*** This index number uses current year quantities as weights. Thus, we can write

$$P_{01}^{Pa} = \frac{\sum p_{1i}q_{1i}}{\sum p_{0i}q_{1i}} \times 100 \quad \text{or} \quad P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

3. ***Fisher's Ideal Index:*** As will be discussed later that the Laspeyres's Index has an upward bias and the Paasche's Index has a downward bias. In view of this, Fisher suggested that an ideal index should be the geometric mean of Laspeyres' and Paasche's indices. Thus, the Fisher's formula can be written as follows:

$$P_{01}^{F} = \sqrt{P_{01}^{La} \times P_{01}^{Pa}} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100}$$

If we write L = $\frac{\sum p_1 q_0}{\sum p_0 q_0}$ and P = $\frac{\sum p_1 q_1}{\sum p_0 q_1}$, the Fisher's Ideal Index can also be written as $P_{01} = \sqrt{L \times P} \times 100$

4. ***Dorbish and Bowley's Index:*** This index number is constructed by taking the arithmetic mean of the Laspeyres's and Paasche's indices.

$$P_{01}^{DB} = \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 + \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100\right]$$

$$= \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100 = \frac{1}{2}\left[L + P\right] \times 100$$

5. ***Marshall and Edgeworth's Index:*** This index number uses arithmetic mean of base and current year quantities.

$$P_{01}^{ME} = \frac{\sum p_1\left(\frac{q_0 + q_1}{2}\right)}{\sum p_0\left(\frac{q_0 + q_1}{2}\right)} \times 100 = \frac{\sum p_1\left(q_0 + q_1\right)}{\sum p_0\left(q_0 + q_1\right)} \times 100 = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

Notes

Example: Calculate the weighted aggregative price index for 2010 from the following data:

Item	*Price in* 2000	*Price in* 2010	*Weights*
A	8	9.5	5
B	12	12.5	1
C	6.5	9	3
D	4	4.5	6
E	6	7	4
F	2	4	3

Solution:

Calculation Table

Item	*Price in* 2000 (p_0)	*Price in* 2010 (p_1)	*Weights* (w)	p_0w	p_1w
A	8	9.5	5	40.0	47.5
B	12	12.5	1	12.0	12.5
C	6.5	9	3	19.5	27.0
D	4	4.5	6	24.0	27.0
E	6	7	4	24.0	28.0
F	2	4	3	6.0	12.0
Total				125.5	154.0

$\therefore$ Price Index (2000 = 100) $P_{01} = \frac{154.0}{125.5} \times 100 = 122.71$

Note: The term within bracket, i.e., 2000= 100, indicates that base year is 2000.

Example: For the data given in the following table, compute

1. Laspeyres's Price Index
2. Paasche's Price Index
3. Fisher's Ideal Index
4. Dorbish and Bowley's Price Index
5. Marshall and Edgeworth's Price Index

Commodity	p_0	q_0	p_1	q_1
A	10	30	12	50
B	8	15	10	25
C	6	20	6	30
D	4	10	6	20

Notes

Solution:

Calculation Table

Comm-odity	p_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1	q_0q_1	$\sqrt{q_0q_1}$	$p_0\sqrt{q_0q_1}$	$p_1\sqrt{q_0q_1}$
A	10	30	12	50	300	360	500	600	1500	38.73	387.3	464.8
B	8	15	10	25	120	150	200	250	375	19.36	154.9	193.6
C	6	20	6	30	120	120	180	180	600	24.49	146.9	146.9
D	4	10	6	20	40	60	80	120	200	14.14	56.6	84.8
					580	690	960	1150			745.7	890.1

The calculation of various price index numbers are done as given below:

1. $P_{01}^{La} = \frac{690}{580} = 118.97$

2. $P_{01}^{Pa} = \frac{1150}{960} \times 100 = 119.79$

3. $P_{01}^{Fi} = \sqrt{\frac{690}{580} \times \frac{1150}{960}} \times 100 = 119.38$

4. $P_{01}^{DB} = \frac{1}{2}\left[\frac{690}{580} + \frac{1150}{960}\right] \times 100 = 119.4$

5. $P_{01}^{ME} = \frac{690 + 1150}{580 + 960} \times 100 = 119.48$

10.6 Quantity Index Numbers

A quantity index number measures the change in quantities in current year as compared with a base year. The formulae for quantity index numbers can be directly written from price index numbers simply by interchanging the role of price and quantity. Similar to a price relative, we can define a quantity relative as $Q = \frac{q_1}{q_0} \times 100$

Various formulae for quantity index numbers are as given below :

1. Simple aggregative index $Q_{01} = \frac{\sum q_1}{\sum q_0} \times 100$

2. Simple average of quantity relatives

   (a) Taking A.M. $Q_{01} = \frac{\frac{\sum q_1}{\sum q_0} \times 100}{n} = \frac{\sum Q}{n}$

   (b) Taking G.M. $Q_{01} = Antilog\left[\frac{\sum \log Q}{n}\right]$

Notes

3. Weighted aggregative index

(a) $Q_{01}^{La} = \dfrac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$ (base year prices are taken as weights)

(b) $Q_{01}^{Pa} = \dfrac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$ (current year prices are taken as weights)

(c) $Q_{01}^{Fi} = \sqrt{\dfrac{\sum q_1 p_0}{\sum q_0 p_0} \times \dfrac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100$

Other aggregative formulae can also be written in a similar way.

4. Weighted average of quantity relatives

(a) Taking A.M. $Q_{01} = \dfrac{\sum Qw}{\sum w}$

(b) Taking G.M. $Q_{01} = Antilog\left[\dfrac{\sum w \log Q}{\sum w}\right]$

Like weighted average of price relatives, values are taken as weights.

Example: Using Fisher's formula, the quantity index number from the following data:

Article	2009		2011	
	Price (Rs)	Value (Rs)	Price (Rs)	Value (Rs)
A	5	50	4	48
B	8	48	7	49
C	6	18	5	20

Solution:

Article	2009			2011			p_0q_1	p_1q_0
	p_0	V_0	$q_0 = \frac{V_0}{p_0}$	p_1	V_1	$q_1 = \frac{V_1}{p_1}$		
A	5	50	10	4	48	12	60	40
B	8	48	6	7	49	7	56	42
C	6	18	3	5	20	4	24	15
Total	$\sum p_0q_0 = 116$			$\sum p_1q_1 = 117$			140	97

$$Q_{01}^{Fi} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \times 100 = \sqrt{\frac{140}{116} \times \frac{117}{97}} = 120.65$$

Notes

10.7 Value Index Number

A value index number gives the change in value in current period as compared with base period. The value index, denoted by V01, is given by the formula

$$V_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0} \times 100 .$$

Tasks

1. "For the construction of index numbers, the best method on theoretical grounds is not the best from practical point of view, so, out of a long list of methods no method is really ideal ". Comment.
2. Study conceptual differences between price index number and quantity index numbers.

Self Assessment

Fill in the blanks:

1. An ..... is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations.
2. The index numbers are very useful device for measuring the average change in prices or any other characteristics like....., ...etc.
3. Index number is often used to average a ..... expressed in different units for different items of a group.
4. Price of eggs is expressed as .....
5. Price index can be used to determine the ..... and ..... of average change in the prices for the group.
6. The year from which comparisons are made is called the .....year.
7. The year under consideration for which the comparisons are to be computed is called the ..... year.
8. While taking weighted average of price relatives, the ..... are often taken as weights.
9. Laspeyres's Index has an ..... bias.
10. Paasche's Index has a ..... bias.

10.8 Comparison of Laspeyres's and Paasche's Index Numbers

Out of various formulae discussed so far, the Laspeyres's and Paasche's formulae are generally preferred for the construction of index numbers. The main reason for this is that the values of these index numbers have a simple interpretation. For example, in case of Laspeyres's index, the base year quantities are used as weights and Sp_1q_0 gives the cost of base year bundle of goods valued at current year prices. Similarly, Sp_0q_0 gives the cost of base year bundle valued at base year prices. Therefore, the ratio $\frac{\sum p_1 q_0}{\sum p_0 q_0}$ gives the change in cost of purchasing the bundle q_0.

Notes

In a similar manner the Paasche's price index can be interpreted as the change in cost of purchasing the bundle q_1. Out of these two, the Laspeyres's index is preferred because weights do not change over different periods and hence the index numbers of various periods remain comparable. Furthermore, Laspeyres's index requires less calculation work than the one with changing weights in every period. The main disadvantage of Laspeyres's formula is that with passage of time the relative importance of various items may change and the base year weights may become outdated. Paasche's index, on the other hand, uses current year weights which truly reflect the relative importance of the items. The main difficulty, in this case, is that index numbers of various periods are not comparable because of changing weights. Moreover, it may be too expensive and difficult to obtain these weights.

When both the index number formulae are applied to the same data, they will in general give different values. However, "if prices of all the commodities change in the same ratio, then the Laspeyres's index is equal to Paasche's index, for then the two weighing systems become irrelevant; or, if quantities of all the commodities change in same ratio, the two index numbers will again be equal, for then the two weighing systems are same relatively." (Karmel & Polasek)

In order to show this, let p_{1i} be the price of i th commodity in current year and p_{0i} be its price in base year. If prices of all the commodities increase by 5%, then we can write $\frac{p_{1i}}{p_{0i}} = \frac{105}{100}$ or $p_{1i} = 1.05 \times p_{0i}$, for all values of i. To generalise, we assume that $p_{1i} = a.p_{0i}$ (or $p_1 = a.p_0$, on dropping the subscript i), where a is constant.

We can write the Laspeyres's index as $P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

Substituting $p_1 = ap_0$, we have

$$P_{01}^{La} = \frac{\sum \alpha p_0 q_0}{\sum p_0 q_0} \times 100 = \alpha \frac{\sum p_0 q_0}{\sum p_0 q_0} \times 100 = 100\alpha \qquad \text{.... (1)}$$

Similarly, the Paasche's index is given by

$$P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{\sum \alpha p_0 q_1}{\sum p_0 q_1} \times 100 = \alpha \frac{\sum p_0 q_1}{\sum p_0 q_1} = 100\alpha \qquad \text{.... (2)}$$

Hence, $P_{01}^{La} = P_{01}^{Pa}$

Further, when quantities of all the commodities change in same proportion, we can write, $q_1 = b.q_0$, for all commodities. Here b is a constant.

Thus, we can write the Paasche's index as

$$P_{01}^{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{\beta \sum p_1 q_0}{\beta \sum p_0 q_0} \times 100 = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Hence, $P_{01}^{Pa} = P_{01}^{La}$

Notes

Self Assessment

Multiple Choice Questions:

11. The Laspeyres's and Paasche's formulae are generally preferred for the construction of .....

    (a) Index numbers (b) Frequency table

    (c) Pie charts (d) Bar graphs

12. Index numbers are expressed in terms of .

    (a) Constant (b) Decimals

    (c) Percentages (d) Decimals

13. Laspeyres's index requires ..... calculation work than the one with changing weights in every period.

    (a) Simple (b) Complex

    (c) Less (d) More

14. In practical situations, neither ..... nor ..... change in the same proportion, the two index numbers are in general different from each other.

    (a) Prices, quantities (b) Size, Volume

    (c) Price, Value (d) Ratio, Quantity

10.9 Relation between Weighted Aggregative and Weighted Arithmetic Average of Price Relatives Index Numbers

It will be shown here that, basically, the two types of index numbers, weighted aggregative and weighted arithmetic average of price relatives, are same and that one type of index number can be obtained from the other by suitable selection of weights. Since the weighted aggregative index numbers are easy to calculate and have simple interpretation, they are preferred to weighted arithmetic average of price relatives indices.

Consider the Laspeyre's index $P_{01}^{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

Rewriting $\Sigma p_1 q_0$ as $\sum \frac{p_1}{p_0} \cdot p_0 q_0$, we have $P_{01}^{La} = \frac{\sum \frac{p_1}{p_0} \cdot p_0 q_0}{\sum p_0 q_0} \times 100 = \frac{\sum Pw}{\sum w}$ (1)

Here $P = \frac{p_1}{p_0} \times 100$ and w = p0q0

In a similar way, the other aggregative type of index numbers can also be converted into average type index numbers.

Notes

Further, it can be shown that an arithmetic average type of index number can be converted into an aggregative type by a suitable selection of weights.

Consider $P_{01} = \dfrac{\sum \dfrac{p_1}{p_0} \times w}{\sum w} \times 100$

Let w = p_0q_1, then the above equation can be written as

$$P_{01} = \frac{\sum \frac{p_1}{p_0} \times p_0 q_1}{\sum p_0 q_1} \times 100 = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = P_{01}^{Pa}$$

From the practical point of view, the weighted aggregative methods are preferred to weighted average of price relatives because the former are easy to compute and have simple interpretation. However, the weighted average of price relatives may be useful when we are interested in knowing the extent of homogeneity in price movements of a certain sub-group or the whole group of commodities. Further, to determine the relative importance of an item, it is necessary to write the index number formula in weighted average of price relative form.

From equation (1), we can write $P_{01}^{La} = \sum \left(\dfrac{p_1}{p_0} \cdot \dfrac{p_0 q_0}{\sum p_0 q_0} \right) \times 100$

Let $w = \dfrac{p_0 q_0}{\sum p_0 q_0}$ then Sw = 1. Thus, $\dfrac{p_0 q_0}{\sum p_0 q_0}$ for an item gives its relative importance (or weight) in the group.

10.9.1 Change in the Cost of Living due to Change in Price of an Item

Let q_1, q_2, ... q_n be the fixed quantities of the n commodities consumed by a group of consumers irrespective of price changes; and p_{01}, p_{02}, ... p_{0n} and p_{11}, p_{12}, ... p_{1n} be their prices in base and current years respectively. Their cost of living index, measured by the change in expenditure of purchasing a given bundle of commodities, can be written as $P_{01} = \dfrac{\sum p_{1i} q_i}{\sum p_{0i} q_i} \times 100$

Let the price of ith commodity changes by 100a%. Thus, the new price, denoted as p'_{01}, can be written as $(1 + a)p_{1i}$ and the changed index number would be

$$P'_{01} = \frac{\left(\sum p_{1i} q_i + \alpha p_{1i} q_i\right)}{\sum p_{0i} q_i} \times 100 = P_{01} + \frac{\alpha p_{1i} q_i}{\sum p_{0i} q_i} \times 100$$

Hence, the absolute change in the cost of living is given by

$P'_{01} - P_{01} = \dfrac{\alpha p_{1i} q_i}{\sum p_{0i} q_i} \times 100$ and the proportionate change is given by

$$\frac{P'_{01} - P_{01}}{P_{01}} = \frac{\alpha p_{1i} q_i}{\sum p_{0i} q_i} \times 100 \times \frac{\sum p_{0i} q_i}{\sum p_{1i} q_i} \times \frac{1}{100} = \frac{\alpha p_{1i} q_i}{\sum p_{1i} q_i}$$

Notes

We note that $\frac{p_{1i}q_i}{\sum p_{1i}q_i}$ is the proportion of expenditure on the ith commodity before the change of price.

Alternatively, the above equation can be written as:

$$\begin{pmatrix} \text{Proportionate Change in} \\ \text{price of the commodity} \end{pmatrix} \times \begin{pmatrix} \text{Proportion of expenditure} \\ \text{on the commodity} \end{pmatrix} = \begin{pmatrix} \text{Proportionate Change in} \\ \text{the cost of living} \end{pmatrix}$$

It may be pointed out here that the above result assumes that the consumption of the commodity remains unchanged as a result of change in its price.

Did u know? Consumer price index number was formerly known as cost of living index.

Self Assessment

Fill in the blanks:

15. Weighted aggregative and weighted arithmetic average of price relatives, are .....
16. One type of index number can be obtained from the other by ..... of weights.
17. The weighted aggregative index numbers are ..... to calculate and have ..... interpretation.

10.10 Chain Base Index Numbers

So far, we have considered index numbers where comparisons of various periods were done with reference to a particular period, termed as base period. Such type of index number series is known as fixed base series. There are several examples of fixed base series like the series of index numbers of industrial production, of agricultural production, of wholesale prices, etc. The main problem with a fixed base series arises when the base year becomes too distant from the current year. In such a situation, it may happen that commodities which used to be very important in the base year are no longer so in current year. Furthermore, certain new commodities might be in use while some old commodities are dropped in current year. In short, this implies that the relative importance of various items is likely to change and, therefore, the comparison of a particular year with a remote base year may appear to be meaningless. A way out to this problem is to construct Chain Base Index Numbers, where current year is compared with its preceding year.

Similar to price relatives, here we define link relatives. A link relative of a commodity in a particular year is equal to the ratio of this year's price to last year's price multiplied by hundred.

Using symbols, the link relative of i th commodity in period t is written as $L_{ti} = \frac{p_{ti}}{p_{t-1i}} \times 100$.

When there are n commodities, the chain base index for period t is given by a suitable average of their link relatives. For example, taking simple arithmetic mean of link relatives we can write

the chain base index as $P_t^{CB} = \frac{\sum L_t}{n} \times 100 = \frac{\sum \frac{p_t}{p_{t-1}} \times 100}{n}$ .... (1)

We may note here that a chain base index is equal to link relative of a commodity when there is only one commodity.

Notes

10.10.1 Chained Index Numbers

The chain base index numbers, obtained above, are as such of not much use because these have been computed with reference to a different base period and hence not comparable with each other. To avoid this difficulty, these are required to be chained to a common base period. The process of chaining is based upon the concept of circular test. The expression for chained index for period 't' with '0' as base period, denoted as P_{0t}^{Ch}, can be written as

$$P_{0t}^{Ch} = \frac{P_1^{CB}}{100} \times \frac{P_2^{CB}}{100} \times L \times \frac{P_t^{CB}}{100} \times 100$$

or $$P_{0t}^{Ch} = \frac{P_t^{CB} \times P_{0(t-1)}^{Ch}}{100} \left(Q \quad P_{0(t-1)}^{Ch} = \frac{P_1^{CB}}{100} \times L \times \frac{P_{t-1}^{CB}}{100} \times 100 \right) \quad .... (2)$$

10.10.2 Conversion of Chain Base Index Number into Fixed Base Index Number and vice-versa

We can write $P_t^{CB} = \frac{P_{0t}^{FB}}{P_{0t-1}^{FB}} \times 100$

$$\text{i.e.,} \begin{pmatrix} \text{Chain Base Index} \\ \text{Number of current} \\ \text{year} \end{pmatrix} = \frac{\text{Fixed Base Index of current year}}{\text{Fixed Base Index of previous year}} \times 100 \quad .... (3)$$

$$\begin{pmatrix} \text{Fixed Base Index} \\ \text{of Current year} \end{pmatrix} = \frac{\begin{pmatrix} \text{Chain Base Index} \\ \text{of Current year} \end{pmatrix} \times \begin{pmatrix} \text{Fixed Base Index} \\ \text{of previous year} \end{pmatrix}}{100} \quad .... (4)$$

Example: From the following data, construct chain base index numbers:

Items	Years				
	2006	2007	2008	2009	2010
	Prices in ₹				
A	5	8	10	12	15
B	3	6	8	10	12
C	2	3	5	7	10.5

Solution:

Calculation of Chain Base Index Numbers

LR* → Items B	2006	2007	2008	2009	2010
A	100	$\frac{8}{5} \times 100 = 160$	$\frac{10}{8} \times 100 = 125$	$\frac{12}{10} \times 100 = 120$	$\frac{15}{12} \times 100 = 125$
B	100	$\frac{6}{3} \times 100 = 200$	$\frac{8}{6} \times 100 = 133.3$	$\frac{10}{8} \times 100 = 125$	$\frac{12}{10} \times 100 = 120$
C	100	$\frac{3}{2} \times 100 = 150$	$\frac{5}{3} \times 100 = 166.7$	$\frac{7}{5} \times 100 = 140$	$\frac{10.5}{7} \times 100 = 150$
Total	300	510	425.0	385	395
CBI	$\frac{300}{3} = 100$	$\frac{510}{3} = 170$	$\frac{425}{3} = 141.7$	$\frac{385}{3} = 128.3$	$\frac{395}{3} = 131.7$

* LR = Link Relatives

Notes

Example: From the chain base index numbers given below, prepare fixed base index numbers; (a) when 1975 is not the base year and (b) when 1975 is taken as base year.

Years	:	2007	2008	2009	2010	2011
Index	:	80	110	120	90	140

Solution:

$$\text{Fixed Base Index of current year} = \frac{\begin{pmatrix}\text{Chain Base Index}\\ \text{of current year}\end{pmatrix} \times \begin{pmatrix}\text{Fixed Base Index}\\ \text{of previous year}\end{pmatrix}}{100}$$

Calculation of Fixed Base Index Numbers

Years	C.B.I.	F.B.I. $1975 \neq 100$	F.B.I. 1975=100
2007	80	80	100
2008	110	$\frac{110\times80}{100}=88$	$\frac{110\times100}{100}=110$
2009	120	$\frac{120\times88}{100}=105.6$	$\frac{120\times110}{100}=132$
2010	90	$\frac{90\times105.6}{100}=95$	$\frac{90\times132}{100}=118.8$
2011	140	$\frac{140\times95}{100}=133$	$\frac{140\times118.8}{100}=166.32$

Notes

Some additional weighted index numbers

1. ***Walsh's Index:*** Geometric mean of base and current year quantities are used as weights in this index number.

$$P_{01}^{Wa} = \frac{\sum p_1\sqrt{q_0q_1}}{\sum p_0\sqrt{q_0q_1}} \times 100$$

2. ***Kelly's Fixed Weights Aggregative Index:*** The weights, in this index number, are quantities which may not necessarily relate to base or current year. The weights, once decided, remain fixed for all periods. The main advantage of this index over Laspeyres's index is that weights do not change with change of base year. Using symbols, the Kelly's Index can be written as

$$P_{01}^{Ke} = \frac{\sum p_1q}{\sum p_0q} \times 100$$

Caution An index number can be computed by using a number of formulae and different formulae will give different results. Unless a proper method is used, the results are likely to be inaccurate and misleading.

Notes

Self Assessment

State whether the following statements are true or false

18. Index numbers where comparisons of various periods were done with reference to a particular period, termed as base period.

19. There is no problem with a fixed base series even when the base year becomes too distant from the current year.

20. When there is a single commodity, the chained index will be equal to the fixed base index.

Case Study

Cost of Living Index

An enquiry into the budgets of middle class families of certain city revealed that, on an average, the percentage expenses on the different groups were as follows:

Food 45, Rent 15, Clothing 12, Fuel and Light 8, Miscellaneous 20.

The group index numbers for the current year as compared to a fixed base were 410, 150, 343, 248 and 285 respectively. Calculate the cost of living index for the current year.

Mr. X was getting ₹ 240 in the base year and ₹ 430 in current year. State, how much he ought to have received as extra allowance to maintain his former standard of living.

10.11 Summary

- An index number is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations.
- In real situations, neither the prices of all the items change in the same ratio nor in the same direction, i.e., the prices of some commodities may change to a greater extent as compared to prices of other commodities.
- The index numbers are very useful device for measuring the average change in prices or any other characteristics like quantity, value, etc., for the group as a whole.
- Index numbers are specialized type of averages that are used to measure the changes in a characteristics which is not capable of being directly measured.
- The changes in magnitude of a group are expressed in terms of percentages which are independent of the units of measurement. This facilitates the comparison of two or more index numbers in different situations.
- Index numbers are indispensable tools for the management of any government or non-government organizations.
- There is inverse relation between the purchasing power of money and the general price level measured in terms of a price index number.
- The reciprocal of the relevant price index can be taken as a measure of the purchasing power of money.
- The year from which comparisons are made is called the base year. It is commonly denoted by writing '0' as a subscript of the variable.

Notes

- While taking weighted average of price relatives, the values are often taken as weights. These weights can be the values of base year quantities valued at base year prices.
- In case of weighted aggregative price index numbers, quantities are often taken as weights. These quantities can be the quantities purchased in base year or in current year or an average of base year and current year quantities or any other quantities.
- A quantity index number measures the change in quantities in current year as compared with a base year.
- Index numbers where comparisons of various periods were done with reference to a particular period, termed as base period. Such type of index number series is known as fixed base series.

10.12 Keywords

Barometers of economic activity: Sometimes index numbers are termed as barometers of economic activity.

Base Year: The year from which comparisons are made is called the base year. It is commonly denoted by writing '0' as a subscript of the variable.

Current Year: The year under consideration for which the comparisons are to be computed is called the current year. It is commonly denoted by writing '1' as a subscript of the variable.

Dorbish and Bowley's Index: This index number is constructed by taking the arithmetic mean of the Laspeyres's and Paasche's indices.

Fisher's Index: Fisher suggested that an ideal index should be the geometric mean of Laspeyres' and Paasche's indices.

Index number: An index number is a statistical measure used to compare the average level of magnitude of a group of distinct but related variables in two or more situations.

Kelly's Fixed Weights Aggregative Index: The weights, in this index number, are quantities which may not necessarily relate to base or current year. The weights, once decided, remain fixed

Laspeyres's Index: Laspeyres' price index number uses base year quantities as weights

Marshall and Edgeworth's Index: This index number uses arithmetic mean of base and current year quantities.

Paasche's Index: This index number uses current year quantities as weights.

Quantity Index Number: A quantity index number measures the change in quantities in current year as compared with a base year.

Simple Aggregative Method: In this method, the simple arithmetic mean of the prices of all the items of the group for the current as well as for the base year are computed separately. The ratio of current year average to base year average multiplied by 100 gives the required index number

Value index Number: A value index number gives the change in value in current period as compared with base period. The value index is denoted by V01. for all periods

Walsh's Index: Geometric mean of base and current year quantities are used as weights in this index number.

Weighted Aggregative Method: This index number is defined as the ratio of the weighted arithmetic means of current to base year prices multiplied by 100.

Notes

10.13 Review Questions

1. What are index numbers? Discuss their uses?
2. Examine various steps in the construction of an index number.
3. "Index numbers are barometers of economic activity". Explain the meaning of this statement.
4. "An index number is a specialised type of average". Explain.
5. Distinguish between average type and aggregative type of index numbers. Discuss the nature of weights used in each case.
6. Explain the role of weighing in the construction of an index number. What are commonly proposed weighing schemes ?
7. Distinguish between simple and weighted index numbers. Explain 'weighted aggregative' and 'weighted average of relatives' methods for the construction of index numbers.
8. Explain the role of weighing in the construction of an index number of prices. What are the commonly proposed weighing schemes?
8. "In the construction of index numbers the advantages of geometric mean are greater than those of arithmetic mean ". Discuss.
9. If we wish to calculate an average of price relatives of commodities, each of which is regarded as being of equal importance, which average would you use: Arithmetic Mean, Geometric Mean or Harmonic Mean? Explain, why?
10. Show that the Laspeyres's index has an upward bias and the Paasche's index has a downward bias. Under what conditions the two index numbers will be equal?
11. Show that Laspeyres price index number can be written as a weighted average of price relatives. What are the weights?
12. "Index numbers are used to measure the changes in some magnitude that is not capable of being directly observed ". Explain this statement and point out the uses and limitations of index numbers.
13. Distinguish between fixed base and chain base methods of index number construction. What are the advantages and disadvantages of the two methods?
14. Explain the meaning of circular test. Discuss the use of this test in the construction of chain base index numbers.
15. What are time reversal and factor reversal tests? Under what conditions these tests are satisfied by Laspeyres's and Paasche's index numbers ?
16. Calculate price index of 2011 with 2007 as base from the following data by:
    (a) Simple aggregative method
    (b) Laspeyres's method
    (c) Paasche's method
    (d) Fisher's ideal index method

Notes

(e) Dorbish and Bowley's method and

(f) Marshall and Edgeworth's method.

Commodity	2007		2011	
	Price	*Quantity*	*Price*	*Quantity*
A	16	50	24	45
B	18	30	24	25
C	20	5	15	8
D	10	6	12	6
E	10	10	14	12

Answers: Self Assessment

1.	Index Number	2.	quantity, value
3.	characteristics	4.	₹/dozen
5.	extent, direction	6.	base
7.	current	8.	values
9.	upward	10.	downward
11.	(a)	12.	(c)
13.	(c)	14.	(a)
15.	same	16.	suitable selection
17.	easy, simple	18.	True
19.	False	20.	True

10.14 Further Readings

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi

Bhardwaj R S., *Business Statistics*, Excel Books

Lindgren B.W (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Selvaraj R, Loganathan, *C Quantitative Methods in Management*

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutia.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Notes

Online links

http://www.netcomuk.co.uk/~jenolive/sindexes.html

http://www.mathsisfun.com/data/index.html

www.imf.org/external/np/sta/tegppi/ch15.pdf

Notes

Unit 11: Analysis of Time Series

CONTENTS

Objectives

Introduction

11.1 Time Series

11.1.1 Objectives of Time Series Analysis

11.1.2 Components of a Time Series

11.1.3 Analysis of Time Series

11.1.4 Method of Averages

11.2 Seasonal Variations

11.2.1 Methods of Measuring Seasonal Variations

11.3 Summary

11.4 Keywords

11.5 Review Questions

11.6 Further Readings

Objectives

After studying this unit, you will be able to:

- Define the term time series
- Discuss the objective and components of time series
- Make analysis of time series
- Brief about seasonal variations
- Explain various methods of measuring seasonal variations and state their merits and demerits

Introduction

The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.

Did u know? The successive intervals are usually equal time intervals, e.g., it can be 10 years, a year, a quarter, a month, a week, a day, an hour, etc.

Notes

11.1 Time Series

A series of observations, on a variable, recorded after successive intervals of time is called a time series.

It should be noted here that the time series data are bivariate data in which one of the variables is time. This variable will be denoted by t. The symbol Y_t will be used to denote the observed value, at point of time t, of the other variable. If the data pertains to n periods, it can be written as (t, Y_t), t = 1, 2, ... n.

11.1.1 Objectives of Time Series Analysis

The analysis of time series implies its decomposition into various factors that affect the value of its variable in a given period. It is a quantitative and objective evaluation of the effects of various factors on the activity under consideration.

There are two main objectives of the analysis of any time series data:

1. To study the past behaviour of data.
2. To make forecasts for future.

The study of past behaviour is essential because it provides us the knowledge of the effects of various forces. This can facilitate the process of anticipation of future course of events and, thus, forecasting the value of the variable as well as planning for future.

11.1.2 Components of a Time Series

An observed value of a time series, Y_t, is the net effect of many types of influences such as changes in population, techniques of production, seasons, level of business activity, tastes and habits, incidence of fire floods, etc. It may be noted here that different types of variables may be affected by different types of factors, e.g., factors affecting the agricultural output may be entirely different from the factors affecting industrial output. However, for the purpose of time series analysis, various factors are classified into the following three general categories applicable to any type of variable.

1. Secular Trend or simply Trend
2. Periodic or Oscillatory Variations
   (a) Seasonal Variations
   (b) Cyclical Variations
3. Random or Irregular Variations

1. ***Secular Trend:*** Secular trend or simply trend is the general tendency of the data to increase or decrease or stagnate over a long period of time. Most of the business and economic time series would reveal a tendency to increase or to decrease over a number of years. For example, data regarding industrial production, agricultural production, population, bank deposits, deficit financing, etc., show that, in general, these magnitudes have been rising over a fairly long period. As opposed to this, a time series may also reveal a declining trend, e.g., in the case of substitution of one commodity by another, the demand of the substituted commodity would reveal a declining trend such as the demand for cotton clothes, demand for coarse grains like bajra, jowar, etc. With the improved medical facilities, the death rate is likely to show a declining trend, etc. The change in trend, in either case, is attributable to the fundamental forces such as changes in population, technology, composition of production, etc.

Notes

Objectives of Measuring Trend

There are four main objectives of measuring trend of a time series data:

(a) To study past growth or decline of the series. On ignoring the short-term fluctuations, trend describes the basic growth or decline tendency of the data.

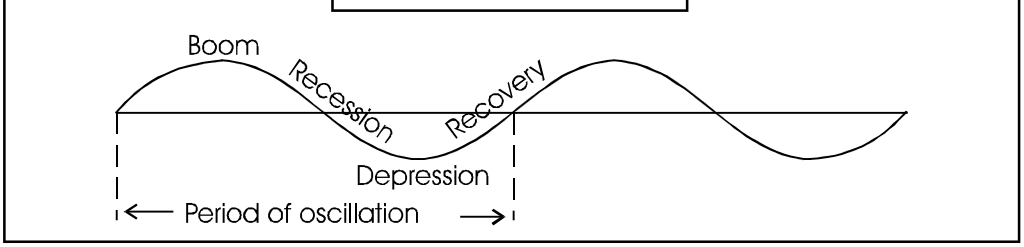
(b) Assuming that the same behaviour would continue in future also, the trend curve can be projected into future for forecasting.

(c) In order to analyse the influence of other factors, the trend may first be measured and then eliminated from the observed values.

(d) Trend values of two or more time series can be used for their comparison.

2. ***Periodic Variations:*** These variations, also known as oscillatory movements, repeat themselves after a regular interval of time. This time interval is known as the period of oscillation. These oscillations are shown in the following figure:

Figure 11.1: Periodic Variations



The oscillatory movements are termed as Seasonal Variations if their period of oscillation is equal to one year, and as Cyclical Variations if the period is greater than one year.

A time series, where the time interval between successive observations is less than or equal to one year, may have the effects of both the seasonal and cyclical variations. However, the seasonal variations are absent if the time interval between successive observations is greater than one year.

Although the periodic variations are more or less regular, they may not necessarily be uniformly periodic, i.e., the pattern of their variations in different periods may or may not be identical in respect of time period and size of periodic variations. For example, if a cycle is completed in five years then its following cycle may take greater or less than five years for its completion.

(a) *Causes of Seasonal variations:* The main causes of seasonal variations are:

(i) Climatic Conditions

(ii) Customs and Traditions

(i) Climatic Conditions: The changes in climatic conditions affect the value of time series variable and the resulting changes are known as seasonal variations. For example, the sale of woolen garments is generally at its peak in the month of November because of the beginning of winter season. Similarly, timely rainfall may increase agricultural output, prices of agricultural commodities are lowest during their harvesting season, etc., reflect the effect of climatic conditions on the value of time series variable.

Notes

(ii) Customs and Traditions: The customs and traditions of the people also give rise to the seasonal variations in time series. For example, the sale of garments and ornaments may be highest during the marriage season, sale of sweets during Diwali, etc., are variations that are the results of customs and traditions of the people.

It should be noted here that both of the causes, mentioned above, occur regularly and are often repeated after a gap of less than or equal to one year.

Objectives of Measuring Seasonal Variations

The main objectives of measuring seasonal variations are:

(i) To analyse the past seasonal variations.

(ii) To predict the value of a seasonal variation which could be helpful in short-term planning.

(iii) To eliminate the effect of seasonal variations from the data.

(b) *Causes of Cyclical Variations:* Cyclical variations are revealed by most of the economic and business time series and, therefore, are also termed as trade (or business) cycles. Any trade cycle has four phases which are respectively known as boom, recession, depression and recovery phases. These phases are shown in Figure 11.1. Various phases repeat themselves regularly one after another in the given sequence. The time interval between two identical phases is known as the period of cyclical variations. The period is always greater than one year. Normally, the period of cyclical variations lies between 3 to 10 years.

Objectives of Measuring Cyclical Variations

The main objectives of measuring cyclical variations are:

(i) To analyse the behaviour of cyclical variations in the past.

(ii) To predict the effect of cyclical variations so as to provide guidelines for future business policies.

3. ***Random or Irregular Variations:*** As the name suggests, these variations do not reveal any regular pattern of movements. These variations are caused by random factors such as strikes, floods, fire, war, famines, etc. Random variations is that component of a time series which cannot be explained in terms of any of the components discussed so far. This component is obtained as a residue after the elimination of trend, seasonal and cyclical components and hence is often termed as residual component.

Random variations are usually short-term variations but sometimes their effect may be so intense that the value of trend may get permanently affected.

11.1.3 Analysis of Time Series

As mentioned earlier, the purpose of analysis of a time series is to decompose Y_t into various components. However, before doing this, we have to make certain assumptions regarding the manner in which these components have combined themselves to give the value Y_t. Very often it is assumed that Y_t is given by either the summation or the multiplication of various components, and accordingly we shall assume two type of models, i.e., additive model or multiplicative model.

Notes

1. ***Additive Model:*** This model is based on the assumption that the value of the variable of a time series, at a point of time t, is the sum of the four components. Using symbols, we can write

   $Y_t = T_t + S_t + C_t + R_t$, where T_t, S_t, C_t and R_t are the values of trend, seasonal, cyclical and random components respectively, at a point of time t.

   This model assumes that all the four components of time series act independently of one another. This assumption implies that one component has no effect on the other(s) irrespective of their magnitudes.

2. ***Multiplicative Model:*** This model assumes that Y_t is given by the multiplication of various components. Symbolically, we can write

   $$Y_t = T_t \times S_t \times C_t \times R_t$$

   This model implies that although the four components may be due to different causes, these are, strictly speaking, not independent of each other. For example, the seasonal component may be some percentage of trend. Similarly, we can have other components expressed in terms of certain percentage.

   There is, in fact, very little agreement amongst the experts about the validity of the models assumed above. It is not very certain that the components combine themselves in the manner mentioned in the two models. Consequently, various mixed type of models have also been suggested, such as

   $$Y_t = T_t.S_t.C_t + R_t$$

   or $Y_t = T_t.C_t + S_t.R_t$ or $Y_t = T_t + C_t.S_t.R_t$, etc.

   Out of all the models, given above, the additive and the multiplicative models are often used. The two models, when applied to the same data, would give different answers. Though, the additive model may be appropriate in some of the situations, yet it is the multiplicative model which characterises the majority of the time series in economic and business fields.

11.1.4 Method of Averages

1. ***Method of Selected Points:*** In this method, two points, considered to be the most representative or normal, are joined by a straight line to get secular trend. This, again, is a subjective method since different persons may have different opinions regarding the representative points. Further, only linear trend can be determined by this method.

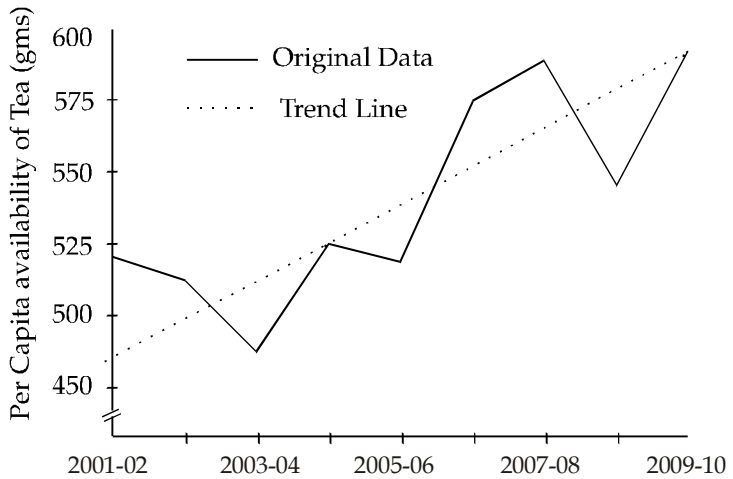
Example: Determine the trend of the following time series data by the method of selected points:

Years	:	2001 – 02	2002 – 03	2003 – 04	2004 – 05	2005 – 06
Per Capita availability of Tea (gms)	:	521	511	462	525	518
Years	:	2006 – 07	2007 – 08	2008 – 09	2009 – 10	
Per Capita availability of Tea (gms)	:	575	589	546	593	

Solution:

Notes

Figure



In the above figure, we have taken the years 2004-2005 and 2009-2010 as the normal years. The corresponding points are joined by a straight line to get the trend of the observed values.

2. ***Method of Semi-averages:*** The given time series data are divided into two equal parts and the arithmetic mean of the values of each part is computed. The computed means are termed as semi-averages. Each semi-average is paired with the centre of time period of its part. The two pairs are then plotted on a graph paper and the points are joined by a straight line to get the trend. It should be pointed out here that in case of odd number of observations, the two equal parts are obtained by dropping the middle most observation.

   Merits

   (a) It is simple method of measuring trend.

   (b) It is an objective method because anyone applying this to a given data would get identical trend values.

   Demerits

   (a) This method can give only a linear trend of the data irrespective of whether it exists or not.

   (b) This is only a crude method of measuring trend, since we do not know whether the effects of other components is completely eliminated or not.

3. ***Method of Moving Average:*** This method is based on the principle that the total effect of periodic variations at different points of time in its cycle gets completely neutralised, i.e., $\Sigma S_t = 0$ in one year and $\Sigma C_t = 0$ in the period of cyclical variations.

   In the method of moving average, successive arithmetic averages are computed from overlapping groups of successive values of a time series. Each group includes all the observations in a given time interval, termed as the period of moving average. The next group is obtained by replacing the oldest value by the next value in the series. The averages of such groups are known as the moving averages.

Notes

The moving average of a group is always shown at the centre of its period. The process of computing moving averages smoothens out the fluctuations in the time series data. It can be shown that if the trend is linear and the oscillatory variations are regular, the moving average with period equal to the period of oscillatory variations would completely eliminate them. Further, the effect of random variations would get minimised because the average of a number of observations must lie between the smallest and the largest observation. It should be noted here that the larger is the period of moving average the more would be the reduction in the effect of random component but more information is lost at the two ends of data.

When the trend is non-linear, the moving averages would give biased rather than the actual trend values.

Let Y_1, Y_2, ... Y_n be the n values of a time series for successive time periods 1, 2, ... n respectively. The calculation of 3-period and 4-period moving averages are shown in the following tables:

Time Period	*Values of Y*	*3 - period M.A.*
1	Y_1	...
2	Y_2	$\frac{Y_1+Y_2+Y_3}{3}$
3	Y_3	$\frac{Y_2+Y_3+Y_4}{3}$
4	Y_4	$\frac{Y_3+Y_4+Y_5}{3}$
5	Y_5	⋮
⋮	⋮	⋮
n	Y_n	...

Time Period	*Values of Y*	*4 - period M.A.*	*Centered Values*
1	Y_1	...	...
2	Y_2	...	...
3	Y_3	$\frac{Y_1+Y_2+Y_3+Y_4}{4}=A_1$	$\frac{A_1+A_2}{2}$
4	Y_4	$\frac{Y_2+Y_3+Y_4+Y_5}{4}=A_2$	$\frac{A_2+A_3}{2}$
5	Y_5	$\frac{Y_3+Y_4+Y_5+Y_6}{4}=A_3$	⋮
⋮	⋮	⋮	⋮
n	Y_n	...	...

It should be noted that, in case of 3-period moving average, it is not possible to get the moving averages for the first and the last periods. Similarly, the larger is the period of moving average the more information will be lost at the ends of a time series.

When the period of moving average is even, the computed average will correspond to the middle of the two middle most periods. These values should be centered by taking arithmetic mean of the two successive averages. The computation of moving average in such a case is also illustrated in the above table.

Example: Determine the trend values of the following data by using 3-year moving average. Also find short-term fluctuations for various years, assuming additive model. Plot the original and the trend values on the same graph.

Year	:	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Production ('000' tonnes)	:	26	27	28	30	29	27	30	31	32	31

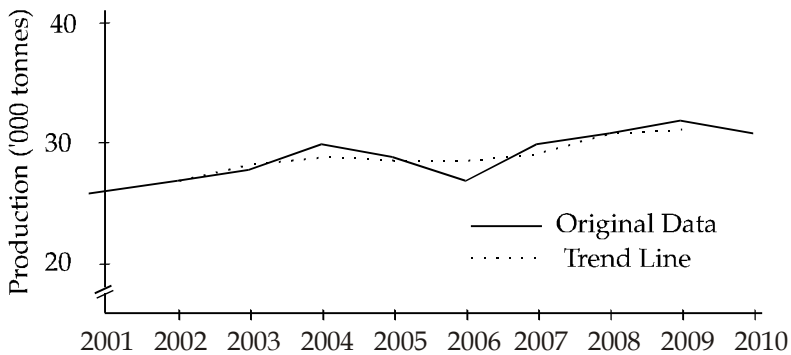
Notes

Solution:

Calculation of Trend and Short-term fluctuations

Years	Production (Y)	3-Year Moving Total	3-Year M.A. or Trend values (T)	Short-term fluctuations (Y – T)
2001	26	…	…	…
2002	27	81	27.00	0.00
2003	28	85	28.33	–0.33
2004	30	87	29.00	1.00
2005	29	86	28.67	0.33
2006	27	86	28.67	–1.67
2007	30	88	29.33	0.67
2008	31	93	31.00	0.00
2009	32	94	31.33	0.67
2010	31	…	…	…

Graphical presentation of Y and T values



Merits

1. This method is easy to understand and easy to use because there are no mathematical complexities involved.
2. It is an objective method.
3. It is a flexible method in the sense that if a few more observations are added, the entire calculations are not changed.
4. When the period of oscillatory movements is equal to the period of moving average, these movements are completely eliminated.
5. By the indirect use of this method, it also possible to isolate seasonal, cyclical and random components.

Demerits

1. It is not possible to calculate trend values for all the items of the series. Some information is always lost at its ends.
2. This method can determine accurate values of trend only if the oscillatory and random fluctuations are uniform in terms of period and amplitude and the trend is, at least, approximately linear. However, these conditions are rarely met in practice. When the trend is not linear, the moving averages will not give correct values of trend.

Notes

3. The selection of period of moving average is a difficult task and a great deal of care is needed to determine it.
4. Like arithmetic mean, the moving averages are too much affected by extreme values.
5. The trend values obtained by moving averages may not follow any mathematical pattern and thus, cannot be used for forecasting which perhaps is the main task of any time series analysis.

Task Assuming a four-yearly cycle, find the trend values for the following data by the method of moving average.

Year	:	1997	1998	1999	2000	2001	2002	2003
Sales (in ₹ '000)	:	74	100	97	87	90	115	126
Year	:	2004	2005	2006	2007	2008	2009	2010
Sales (in ₹ '000)	:	108	100	125	118	113	122	126

Caution Method of moving average should be used only if the trend is linear, When the trend is not linear, we often use weighted rather than simple moving average.

Self Assessment

Fill in the blanks:

1. The successive intervals are usually ..... time intervals.
2. ..... can be 10 years, a year, a quarter, a month, a week, a day, an hour, etc.
3. The data on the population of India is a time series data where time interval between two successive figures is .....
4. Figures of national income, agricultural and industrial production, etc., are available on ..... basis.
5. A series of observations, on a variable, recorded after successive intervals of time is called a .....
6. The time series data are ..... data in which one of the variables is time.
7. Factors affecting the agricultural output may be ..... from the factors affecting industrial output.
8. ..... is the general tendency of the data to increase or decrease or stagnate over a long period of time.
9. Periodic Variations are also known as .....which repeat themselves after a regular interval of time.
10. Any trade cycle has four phases which are respectively known as boom, recession, ... and ..... phases.
11. The time interval between two identical phases is known as the period of .....
12. Normally, the period of cyclical variations lies between ..... years.
13. ..... or ..... Variations variations do not reveal any regular pattern of movements.

Notes

11.2 Seasonal Variations

If the time series data are in terms of annual figures, the seasonal variations are absent. These variations are likely to be present in data recorded on quarterly or monthly or weekly or daily or hourly basis. As discussed earlier, the seasonal variations are of periodic nature with period equal to one year. These variations reflect the annual repetitive pattern of the economic or business activity of any society. The main objectives of measuring seasonal variations are :

1. To understand their pattern.
2. To use them for short-term forecasting or planning.
3. To compare the pattern of seasonal variations of two or more time series in a given period or of the same series in different periods.
4. To eliminate the seasonal variations from the data. This process is known as deseasonalisation of data.

11.2.1 Methods of Measuring Seasonal Variations

The measurement of seasonal variation is done by isolating them from other components of a time series. There are four methods commonly used for the measurement of seasonal variations. These method are :

1. Method of Simple Averages
2. Ratio to Trend Method
3. Ratio to Moving Average Method
4. Method of Link Relatives

Note: In the discussion of the above methods, we shall often assume a multiplicative model. However, with suitable modifications, these methods are also applicable to the problems based on additive model.

Method of Simple Averages

This method is used when the time series variable consists of only the seasonal and random components. The effect of taking average of data corresponding to the same period (say 1st quarter of each year) is to eliminate the effect of random component and thus, the resulting averages consist of only seasonal component. These averages are then converted into seasonal indices, as explained in the following examples.

Example: Assuming that trend and cyclical variations are absent, compute the seasonal index for each month of the following data of sales (in ₹ '000) of a company.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2008	46	45	44	46	45	47	46	43	40	40	41	45
2009	45	44	43	46	46	45	47	42	43	42	43	44
2010	42	41	40	44	45	45	46	43	41	40	42	45

Notes

Solution:

Calculation Table

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2008	46	45	44	46	45	47	46	43	40	40	41	45
2009	45	44	43	46	46	45	47	42	43	42	43	44
2010	42	41	40	44	45	45	46	43	41	40	42	45
Total	133	130	127	136	136	137	139	128	124	122	126	134
A_i	44.3	43.3	42.3	45.3	45.3	45.7	46.3	42.7	41.3	40.7	42.0	44.7
S.I.	101.4	99.1	96.8	103.7	103.7	104.6	105.9	97.7	94.5	93.1	96.1	102.3

In the above table, Ai denotes the average and S.I. the seasonal index for a particular month of various years. To calculate the seasonal index, we compute grand average G, given by $G = \frac{\Sigma A_i}{12} = \frac{524}{12} = 43.7$. Then the seasonal index for a particular month is given by. $S.I. = \frac{A_i}{G} \times 100$

Further, $\Sigma S.I. = 1198.9 \neq 1200$. Thus, we have to adjust these values such that their total is 1200. This can be done by multiplying each figure by $\frac{1200}{1198.9}$. The resulting figures are the adjusted seasonal indices, as given below:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
101.5	99.2	96.9	103.8	103.8	104.7	106.0	97.8	94.6	93.2	96.2	102.3

Remarks: The totals equal to 1200, in case of monthly indices and 400, in case of quarterly indices, indicate that the ups and downs in the time series, due to seasons, neutralise themselves within that year. It is because of this that the annual data are free from seasonal component.

Example: Compute the seasonal index from the following data by the method of simple averages.

Year	Quarter	Y	Year	Quarter	Y	Year	Quarter	Y
2005	I	106	2007	I	90	2009	I	80
	II	124		II	112		II	104
	III	104		III	101		III	95
	IV	90		IV	85		IV	83
2006	I	84	2008	I	76	2010	I	104
	II	114		II	94		II	112
	III	107		III	91		III	102
	IV	88		IV	76		IV	84

Solution.

Notes

Calculation of Seasonal Indices

Years	1st Qr	2nd Qr	3rd Qr	4th Qr
2005	106	124	104	90
2006	84	114	107	88
2007	90	112	101	85
2008	76	94	91	76
2009	80	104	95	83
2010	104	112	102	84
Total	540	660	600	506
A_i	90	110	100	84.33
$\frac{A_i}{G}\times 100$	93.67	114.49	104.07	87.77

We have $G = \frac{\sum A_i}{4} = \frac{384.33}{4} = 96.08$. Further, since the sum of terms in the last row of the table is 400, no adjustment is needed. These terms are the seasonal indices of respective quarters.

Merits and Demerits

This is a simple method of measuring seasonal variations which is based on the unrealistic assumption that the trend and cyclical variations are absent from the data. However, we shall see later that this method being a part of the other methods of measuring seasonal variations, is very useful.

Ratio to Trend Method

This method is used when cyclical variations are absent from the data, i.e., the time series variable Y consists of trend, seasonal and random components.

Using symbols, we can write Y = T.S.R

Various steps in the computation of seasonal indices are:

(i) Obtain the trend values for each month or quarter, etc., by the method of least squares.

(ii) Divide the original values by the corresponding trend values. This would eliminate trend values from the data. To get figures in percentages, the quotients are multiplied by 100.

Thus, we have $\frac{Y}{T}\times 100 = \frac{T.S.R}{T}\times 100 = S.R.100$

(iii) Finally, the random component is eliminated by the method of simple averages.

Notes

Example: Assuming that the trend is linear, calculate seasonal indices by the ratio to moving average method from the following data:

Quarterly output of coal in 4 years (in thousand tonnes)

Year	*I*	*II*	*III*	*IV*
2007	65	58	56	61
2008	68	63	63	67
2009	70	59	56	52
2010	60	55	51	58

Solution:

By adding the values of all the quarters of a year, we can obtain annual output for each of the four years. Fit a linear trend to the data and obtain trend values for each quarter.

Year	*Output*	$X = 2(t - 1983.5)$	*XY*	X^2
2007	240	– 3	– 720	9
2008	261	– 1	– 261	1
2009	237	1	237	1
2010	224	3	672	9
Total	962	0	– 72	20

From the above table, we get $a = \frac{962}{4} = 240.5$ and $b = \frac{-72}{20} = -3.6$

Thus, the trend line is Y = 240.5 - 3.6X, Origin: 1st January 2009, unit of X: 6 months.

The quarterly trend equation is given by $Y = \frac{240.5}{4} - \frac{3.6}{8}X$ or Y = 60.13 – 0.45X, Origin : 1st January 2009, unit of X : 1 quarter (i.e., 3 months).

Shifting origin to 15th Feb. 2009, we get

Y = 60.13 - 0.45(X + $\frac{1}{2}$) = 59.9 – 0.45X, origin I-quarter, unit of X = 1 quarter.

The table of quarterly values is given by

Year	*I*	*II*	*III*	*IV*
2007	63.50	63.05	62.60	62.15
2008	61.70	61.25	60.80	60.35
2009	59.90	59.45	59.00	58.55
2010	58.10	57.65	57.20	56.75

The table of Ratio to Trend Values, i.e., $\frac{Y}{T} \times 100$

Notes

Years	I	II	III	IV
2007	102.36	91.99	89.46	98.15
2008	110.21	102.86	103.62	111.02
2009	116.86	99.24	94.92	88.81
2010	103.27	95.40	89.16	102.20
Total	432.70	389.49	377.16	400.18
Average	108.18	97.37	94.29	100.05
S.I.	108.20	97.40	94.32	100.08

Note: Grand Average, $G = \frac{399.89}{4} = 99.97$

Example: Find seasonal variations by the ratio to trend method, from the following data:

Year	I-Qr	II-Qr	III-Qr	IV-Qr
2006	30	40	36	34
2007	34	52	50	44
2008	40	58	54	48
2009	54	76	68	62
2010	80	92	86	82

Solution:

First we fit a linear trend to the annual totals.

Years	Annual Totals (Y)	X	XY	X^2
2006	140	– 2	– 280	4
2007	180	– 1	– 180	1
2008	200	0	0	0
2009	260	1	260	1
2010	340	2	680	4
Total	1120	0	480	10

Now $a = \frac{1120}{5} = 224$ and $b = \frac{480}{10} = 48.$

$\therefore$ The trend equation is Y = 224 + 48X, origin : 1st July 2008, unit of X = 1 year.

The quarterly trend equation is $Y = \frac{224}{4} + \frac{48}{16}X = 56 + 3X$, origin: 1st July 2008, unit of X = 1 quarter.

Shifting the origin to III quarter of 2008, we get

$$Y = 56 + 3(X + \frac{1}{2}) = 57.5 + 3X$$

Notes

Table of Quarterly Trend Values

Year	I	II	III	IV
2006	27.5	30.5	33.5	36.5
2007	39.5	42.5	45.5	48.5
2008	51.5	54.5	57.5	60.5
2009	63.5	66.5	69.5	72.5
2010	75.5	78.5	81.5	84.5

Ratio to Trend Values

Year	I	II	III	IV
2006	109.1	131.1	107.5	93.2
2007	86.1	122.4	109.9	90.7
2008	77.7	106.4	93.9	79.3
2009	85.0	114.3	97.8	85.5
2010	106.0	117.2	105.5	97.0
Total	463.9	591.4	514.6	445.7
A_i	92.78	118.28	102.92	89.14
S.I.	92.10	117.35	102.11	88.44

Note that the Grand Average $G = \frac{403.12}{4} = 100.78$. Also check that the sum of indices is 400.

Remarks: If instead of multiplicative model we have an additive model, then Y = T + S + R or S + R = Y – T. Thus, the trend values are to be subtracted from the Y values. Random component is then eliminated by the method of simple averages.

Merits and Demerits

It is an objective method of measuring seasonal variations. However, it is very complicated and doesn't work if cyclical variations are present.

Ratio to Moving Average Method

The ratio to moving average is the most commonly used method of measuring seasonal variations. This method assumes the presence of all the four components of a time series. Various steps in the computation of seasonal indices are as follows:

1. Compute the moving averages with period equal to the period of seasonal variations. This would eliminate the seasonal component and minimise the effect of random component. The resulting moving averages would consist of trend, cyclical and random components.
2. The original values, for each quarter (or month) are divided by the respective moving average figures and the ratio is expressed as a percentage, i.e., , where R′ and R′′ denote the changed random components.
3. Finally, the random component R′′ is eliminated by the method of simple averages.

Notes

Example: Given the following quarterly sale figures, in thousand of rupees, for the year 2007-2010, find the specific seasonal indices by the method of moving averages.

Year	I	II	III	IV
2007	34	33	34	37
2008	37	35	37	39
2009	39	37	38	40
2010	42	41	42	44

Solution:

Calculation of Ratio to Moving Averages

Year / Quarter	Sales	4 – Period Moving Total	Centred Total	4 Period M	$\frac{Y}{M} \times 100$
2007 I	34		…	…	…
II	33		…	…	…
III	34	138	279	34.9	97.4
IV	37	141	284	35.5	104.2
2008 I	37	143	289	36.1	102.5
II	35	146	294	36.8	95.1
III	37	148	298	37.3	99.2
IV	39	150	302	37.8	103.2
2009 I	39	152	305	38.1	102.4
II	37	153	307	38.4	96.4
III	38	154	311	38.9	97.7
IV	40	157	318	39.8	100.5
2010 I	42	161	326	40.8	102.9
II	41	165	334	41.8	98.1
III	42	169	…	…	…
IV	44		…	…	…

Calculation of Seasonal Indices

Year	I	II	III	IV
2007	–	–	97.4	104.2
2008	102.5	95.1	99.2	103.2
2009	102.4	96.4	97.7	100.5
2010	102.9	98.1	–	–
Total	307.8	289.6	294.3	307.9
A_i	102.6	96.5	98.1	102.6
S.I.	102.7	96.5	98.1	102.7

Note that the Grand Average $G = \frac{399.8}{4} = 99.95$. Also check that the sum of indices is 400.

Notes

Merits and Demerits

This method assumes that all the four components of a time series are present and, therefore, widely used for measuring seasonal variations. However, the seasonal variations are not completely eliminated if the cycles of these variations are not of regular nature. Further, some information is always lost at the ends of the time series.

Link Relatives Method

This method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern. As discussed in earlier unit, the link relatives are percentages of the current period (quarter or month) as compared with previous period. With the computation of link relatives and their average, the effect of cyclical and random component is minimised. Further, the trend gets eliminated in the process of adjustment of chained relatives. The following steps are involved in the computation of seasonal indices by this method:

1. Compute the link relative (L.R.) of each period by dividing the figure of that period with the figure of previous period. For example, link relative of 3rd quarter $= \frac{\text{figure of 3rd quarter}}{\text{figure of 2nd quarter}} \times 100$

2. Obtain the average of link relatives of a given quarter (or month) of various years. A.M. or M_d can be used for this purpose. Theoretically, the later is preferable because the former gives undue importance to extreme items.

3. These averages are converted into chained relatives by assuming the chained relative of the first quarter (or month) equal to 100. The chained relative (C.R.) for the current period (quarter or month) $= \frac{\text{C.R. of the previous period} \times \text{L.R. of the current period}}{100}$

4. Compute the C.R. of first quarter (or month) on the basis of the last quarter (or month). This is given by $\frac{\text{C.R. of last quarter (or month)} \times \text{average L.R. of 1st quarter (or month)}}{100}$

   This value, in general, be different from 100 due to long term trend in the data. The chained relatives, obtained above, are to be adjusted for the effect of this trend. The adjustment factor is $d = \frac{1}{4}[\text{New } C.R. \text{ for 1st quarter} - 100]$ for quarterly data and $d = \frac{1}{12}[\text{New } C.R. \text{ for 1st month} - 100]$ for monthly data.

   On the assumption that the trend is linear, d, 2d, 3d, etc., is respectively subtracted from the 2nd, 3rd, 4th, etc., quarter (or month).

5. Express the adjusted chained relatives as a percentage of their average to obtain seasonal indices.

6. Make sure that the sum of these indices is 400 for quarterly data and 1200 for monthly data.

Notes

Example: Determine the seasonal indices from the following data by the method of link relatives:

Year	1st Qr	2nd Qr	3rd Qr	4th Qr
2006	26	19	15	10
2007	36	29	23	22
2008	40	25	20	15
2009	46	26	20	18
2010	42	28	24	21

Solution.

Calculation Table

Year	I	II	III	IV
2006	–	73.1	78.9	66.7
2007	360.0	80.5	79.3	95.7
2008	181.8	62.5	80.0	75.0
2009	306.7	56.5	76.9	90.0
2010	233.3	66.7	85.7	87.5
Total	1081.8	339.3	400.8	414.9
Mean	270.5	67.9	80.2	83.0
C.R.	100.0	67.9	54.5	45.2
C.R.(adjusted)	100.0	62.3	43.3	28.4
S.I.	170.9	106.5	74.0	48.6

The chained relative (C.R.) of the 1st quarter on the basis of C.R. of the 4th quarter $= \frac{270.5 \times 45.2}{100} = 122.3$.

The trend adjustment factor $d = \frac{1}{4}(122.3 - 100) = 5.6$

Thus, the adjusted C.R. of 1st quarter = 100

and for 2nd = 67.9 – 5.6 = 62.3

for 3rd = 54.5 - 2 × 5.6 = 43.3

for 4th = 45.2 - 3 × 5.6 = 28.4

The grand average of adjusted C.R., $G = \frac{100 + 62.3 + 43.3 + 28.4}{4} = 58.5$

The seasonal index of a quarter $= \frac{\text{Adjusted C.R.} \times 100}{G}$

Merits and Demerits

This method is less complicated than the ratio to moving average and the ratio to trend methods. However, this method is based upon the assumption of a linear trend which may not always hold true.

Notes

Notes

Deseasonalisation of Data

The deseasonalisation of data implies the removal of the effect of seasonal variations from the time series variable. If Y consists of the sum of various components, then for its deseasonalisation, we subtract seasonal variations from it. Similarly, in case of multiplicative model, the deseasonalisation is done by taking the ratio of Y value to the corresponding seasonal index. A clue to this is provided by the fact that the sum of seasonal indices is equal to zero for an additive model while their sum is 400 or 1200 for a multiplicative model.

It may be pointed out here that the deseasonalisation of a data is done under the assumption that the pattern of seasonal variations, computed on the basis of past data, is similar to the pattern of seasonal variations in the year of deseasonalisation.

Did u know? Ratio to moving average method is most general and, therefore, most popular method of measuring seasonal variations.

Self Assessment

Multiple Choice Questions:

14. If the time series data are in terms of annual figures, the seasonal variations are.....
    (a) Present (b) Absent
    (c) In fixed ratio (d) transitory
15. The seasonal variations are of ..... nature with period equal to one year.
    (a) Linear (b) Cyclic
    (c) Periodic (d) Varying
16. The measurement of seasonal variation is done by ..... them from other components of a time series.
    (a) Separating (b) Dissociating
    (c) Isolating (d) Filtering

State whether the following statements are true or false:

17. Method of Simple Averages is used when the time series variable consists of only the seasonal and random components.
18. Ratio to Trend Method is used when cyclical variations are absent from the data.
19. Link Relatives Method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern.
20. Seasonalisation is the process to eliminate the seasonal variations from the data.

Notes

Case Study

Production Estimate

Suppose that there is a series of quarterly production figures (in thousand tonnes) in an industry for the years 2004 to 2010 and the equation of linear trend fitted to the annual data is Xt = 107.2 + 2.93t , where t = year 2007 and Xt the annual production in time period t. Use this equation to estimate the annual production for the year 2005 and 2011.

Suppose now the quarterly indices of seasonal variations are: January-March 125, April-June 105, July-September 87, October-December 88. (The multiplicative model for the time series is assumed.) Use these indices to estimate the production during the first quarter of 1987.

11.3 Summary

- A series of observations, on a variable, recorded after successive intervals of time is called a time series.
- The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.
- The analysis of time series implies its decomposition into various factors that affect the value of its variable in a given period.
- It is a quantitative and objective evaluation of the effects of various factors on the activity under consideration.
- Secular trend or simply trend is the general tendency of the data to increase or decrease or stagnate over a long period of time.
- Trend values of two or more time series can be used for their comparison
- Oscillatory movements, repeat themselves after a regular interval of time. This time interval is known as the period of oscillation.
- The main objective of measuring seasonal variations is to eliminate the effect of seasonal variations from the data.
- Random variations are usually short-term variations but sometimes their effect may be so intense that the value of trend may get permanently affected.
- The main objectives of measuring seasonal variations is to understand their pattern.
- The measurement of seasonal variation is done by isolating them from other components of a time series.
- There are four methods commonly used for the measurement of seasonal variations. These methods are:
    - Method of Simple Averages
    - Ratio to Trend Method
    - Ratio to Moving Average Method
    - Method of Link Relatives

Notes

11.4 Keywords

Additive Model: This model is based on the assumption that the value of the variable of a time series, at a point of time t, is the sum of the four components. Using symbols, we can write $Y_t = T_t + S_t + C_t + R_t$, where T_t, S_t, C_t and R_t are the values of trend, seasonal, cyclical and random components respectively, at a point of time t.

Cyclical Variations: The oscillatory movements are termed as Cyclical Variations if their period of oscillation is greater than one year.

Link Relatives Method: This method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern.

Multiplicative Model: This model assumes that Yt is given by the multiplication of various components. Symbolically, we can write Yt = Tt × St × Ct × Rt.

Periodic Variations: These variations, also known as oscillatory movements, repeat themselves after a regular interval of time. This time interval is known as the period of oscillation.

Random or Irregular Variations: As the name suggests, these variations do not reveal any regular pattern of movements. These variations are caused by random factors such as strikes, floods, fire, war, famines, etc.

Seasonal Variations: The oscillatory movements are termed as Seasonal Variations if their period of oscillation is equal to one year.

Secular Trend: Secular trend or simply trend is the general tendency of the data to increase or decrease or stagnate over a long period of time.

Time Series: A series of observations, on a variable, recorded after successive intervals of time is called a time series.

11.5 Review Questions

1. Explain the meaning and objectives of time series analysis. Describe briefly the methods of measurement of trend.
2. What is a time series? What are its main components? How would you study the seasonal variations in any time series?
3. Distinguish between secular trend and periodic variations. How would you measure trend in a time series data by the method of least squares? Explain your answer with an example.
4. Explain the method of moving average for the determination of trend in a time series data. What are its merits and demerits?
5. Discuss the underlying assumptions of additive and multiplicative models in a time series analysis. Which of these is more popular in practice and why?
6. Distinguish between the ratio to trend and the ratio to moving average methods of measuring seasonal variations. Which method is more general and why?
7. "All periodic variations are not necessarily seasonal". Discuss the above statement with a suitable example.
8. Determine secular trend by the method of semi-averages from the following data on the production of sugarcane (in million tonnes). Plot the observed and the trend values on a graph.

Notes

Years	:	2003	2004	2005	2006	2007	2008	2009	2010
Production	:	152	129	154	186	189	174	170	186

9. Draw a trend line by the method of semi-averages from the following data. The figures of sales are in ₹ '000.

Years	:	2004	2005	2006	2007	2008	2009	2010
Sales	:	85	102	80	89	92	87	94

10. Fit a linear trend to the following data, on average monthly output, with origin at mid-point of the year 2006. Convert this into a monthly trend equation. Estimate the average output for June and August, 2006.

Year	:	2002	2003	2004	2005	2006	2007	2008	2009	2010
Output	:	6.3	7.4	9.3	7.4	8.3	10.6	9.0	8.7	7.9

11. Fit a straight line trend to the following data on average monthly domestic demand (in million barrels) for motor fuel:

Year	:	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Demand	:	61	66	72	76	82	90	96	100	103	110	114

12. Fit a straight line trend by the method of least squares for the following data and obtain the trend values in the form of average monthly production for each year. Also estimate the average monthly production for 2011 and the production for individual months, May and July of 2011.

Year	:	2004	2005	2006	2007	2008	2009	2010
Production	:	70	80	82	73	84	89	82

Answers: Self Assessment

1.	equal	2.	Successive intervals
3.	10 years	4.	yearly
5.	time series	6.	bivariate
7.	entirely different	8.	Secular trend
9.	Oscillatory movements	10.	depression, recovery
11.	cyclical variations.	12.	3 to 10
13.	Random, Irregular	14.	(b)
15.	(c)	16.	(c)
17.	True	18.	True
19.	True	20.	False

11.6 Further Readings

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Bhardwaj R.S., *Business Statistics*, Excel Books.

Notes

Garrett H.E. (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, Mc Graw Hill Book Company, New York.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hannagan T.J. (1982), *Mastering Statistics*, The Macmillan Press Ltd., Surrey.

Hooda R. P., *Statistics for Business and Economics*, Macmillan India, Delhi, 2008.

Jaeger R.M (1983), *Statistics: A Spectator Sport*, Sage Publications India Pvt. Ltd., New Delhi.

Lindgren B.W. (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Selvaraj R., Loganathan, C. *Quantitative Methods in Management.*

Sharma J.K., *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutta.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Online links

http://en.wikipedia.org/wiki/Time_series

http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm

http://www.abs.gov.au/websitedbs/d3310114.nsf/4a256353001af3ed4b2562bb00121564/b81ecff00cd36415ca256ce10017de2f!OpenDocument

http://www.statsoft.com/textbook/time-series-analysis/

http://www.iasri.res.in/ebook/EBADAT/5Modeling%20and%20Forecasting%20Techniques%20in%20Agriculture/2-time_series_analysis_22-02-07_revised.pdf

Notes

Unit 12: Probability and Expected Value

CONTENTS

Objectives

Introduction

12.1 Definitions

12.2 Theorems on Expectation

12.2.1 Expected Monetary Value (EMV)

12.2.2 Expectation of the Sum or Product of two Random Variables

12.2.3 Expectation of a Function of Random Variables

12.3 Counting Techniques

12.3.1 Fundamental Principle of Counting

12.3.2 Permutation

12.3.3 Combination

12.3.4 Ordered Partitions

12.3.5 Statistical or Empirical Definition of Probability

12.3.6 Axiomatic or Modern Approach to Probability

12.3.7 Theorems on Probability

12.4 Summary

12.5 Keywords

12.6 Review Questions

12.7 Further Readings

Objectives

After studying this unit, you will be able to:

- Define the term random experiment and probability
- Discuss various associated terms while defining probability
- State Addition theorem of probability
- Explain multiplicative theorem of probability
- Describe various theorem of expectation

Introduction

The theory of probability is a study of Statistical or Random Experiments. It is the backbone of Statistical Inference and Decision Theory that are essential tools of the analysis of most of the modern business and economic problems.

Notes

Often, in our day-to-day life, we hear sentences like 'it may rain today', 'Mr X has fifty-fifty chances of passing the examination', 'India may win the forthcoming cricket match against Sri lanka', 'the chances of making profits by investing in shares of company A are very bright', etc. Each of the given sentences involves an element of uncertainty.

A phenomenon or an experiment which can result into more than one possible outcome, is called a random phenomenon or random experiment or statistical experiment. Although, we may be aware of all the possible outcomes of a random experiment, it is not possible to predetermine the outcome associated with a particular experimentation or trial.

Consider, for example, the toss of a coin. The result of a toss can be a head or a tail, therefore, it is a random experiment. Here we know that either a head or a tail would occur as a result of the toss, however, it is not possible to predetermine the outcome. With the use of probability theory, it is possible to assign a quantitative measure, to express the extent of uncertainty, associated with the occurrence of each possible outcome of a random experiment.

Did u know? The concept of probability originated from the analysis of the games of chance in the 17th century.

12.1 Definitions

Classical Definition: This definition, also known as the mathematical definition of probability, was given by J. Bernoulli. With the use of this definition, the probabilities associated with the occurrence of various events are determined by specifying the conditions of a random experiment.

Did u know? The classical definition of probability is also known as 'a priori' definition of probability.

Definition

If n is the number of equally likely, mutually exclusive and exhaustive outcomes of a random experiment out of which m outcomes are favourable to the occurrence of an event A, then the probability that A occurs, denoted by P(A), is given by:

$$P(A) = \frac{\text{Number of outcomes favourable to A}}{\text{Number of exhaustive outcomes}} = \frac{m}{n}$$

Various terms used in the above definition are explained below:

1. ***Equally likely outcomes:*** The outcomes of random experiment are said to be equally likely or equally probable if the occurrence of none of them is expected in preference to others. For example, if an unbiased coin is tossed, the two possible outcomes, a head or a tail are equally likely.
2. ***Mutually exclusive outcomes:*** Two or more outcomes of an experiment are said to be mutually exclusive if the occurrence of one of them precludes the occurrence of all others in the same trial i.e. they cannot occur jointly. For example, the two possible outcomes of toss of a coin are mutually exclusive. Similarly, the occurrences of the numbers 1, 2, 3, 4, 5, 6 in the roll of a six faced die are mutually exclusive.
3. ***Exhaustive outcomes:*** It is the totality of all possible outcomes of a random experiment. The number of exhaustive outcomes in the roll of a die are six. Similarly, there are 52 exhaustive outcomes in the experiment of drawing a card from a pack of 52 cards.

Notes

4. ***Event:*** The occurrence or non-occurrence of a phenomenon is called an event. For example, in the toss of two coins, there are four exhaustive outcomes, viz. (H, H), (H, T), (T, H), (T, T). The events associated with this experiment can be defined in a number of ways. For example, (i) the event of occurrence of head on both the coins, (ii) the event of occurrence of head on at least one of the two coins, (iii) the event of non-occurrence of head on the two coins, etc.

Caution An event can be simple or composite depending upon whether it corresponds to a single outcome of the experiment or not. In the example, given above, the event defined by (i) is simple, while those defined by (ii) and (iii) are composite events.

Example: Find the probability of obtaining an odd number in the roll of an unbiased die.

Solution:

The number of equally likely, mutually exclusive and exhaustive outcomes, i.e., n = 6. There are three odd numbers out of the numbers 1, 2, 3, 4, 5 and 6. Therefore, m = 3.

Thus, probability of occurrence of an odd number

Example: What is the chance of drawing a face card in a draw from a pack of 52 well-shuffled cards?

Solution:

Total possible outcomes n = 52.

Since the pack is well-shuffled, these outcomes are equally likely. Further, since only one card is to be drawn, the outcomes are mutually exclusive.

There are 12 face cards, $\therefore$ m = 12.

Thus, probability of drawing a face card $= \frac{12}{52} = \frac{3}{13}$

Task What is the probability that a leap year selected at random will contain 53 Mondays?

Self Assessment

State whether the following statements are true or false:

1. The concept of probability originated from the analysis of the games of chance in the 17th century.
2. The theory of probability is a study of Statistical or Random Experiments.
3. It is the backbone of Statistical Inference and Decision Theory that are essential tools of the analysis of most of the modern business and economic problems.
4. A phenomenon or an experiment which can result into more than one possible outcome, is called a random phenomenon or random experiment or statistical experiment.
5. The result of a toss can be a head or a tail. thus, it is a non random experiment

Notes

6. Mathematical definition of probability, was given by J. Bernoulli.
7. Classical definition is also known as 'a priori' definition of probability.
8. Two or more outcomes of an experiment are said to be mutually exclusive if the occurrence of one of them precludes the occurrence of all others in the same trial i.e. they cannot occur jointly.

12.2 Theorems on Expectation

Theorem 1:

Expected value of a constant is the constant itself, i.e., E(b) = b, where b is a constant.

Theorem 2:

E(aX) = aE(X), where X is a random variable and a is constant.

12.2.1 Expected Monetary Value (EMV)

When a random variable is expressed in monetary units, its expected value is often termed as expected monetary value and symbolised by EMV.

Example: If it rains, an umbrella salesman earns ₹ 100 per day. If it is fair, he loses ₹ 15 per day. What is his expectation if the probability of rain is 0.3?

Solution:

Here the random variable X takes only two values, X1 = 100 with probability 0.3 and X2 = – 15 with probability 0.7.

Thus, the expectation of the umbrella salesman

$$= 100 \times 0.3 - 15 \times 0.7 = 19.5$$

The above result implies that his average earning in the long run would be ₹ 19.5 per day.

12.2.2 Expectation of the Sum or Product of two Random Variables

Theorem 1:

If X and Y are two random variables, then E(X + Y) = E(X) + E(Y).

Theorem 2:

If X and Y are two independent random variables, then

$$E(X.Y) = E(X).E(Y)$$

12.2.3 Expectation of a Function of Random Variables

Let $\phi(X,Y)$ be a function of two random variables X and Y. Then we can write

$$E\left[\phi(X,Y)\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}\phi\left(X_i,Y_j\right)p_{ij}$$

Notes

Expression for Covariance

As a particular case, assume that $\phi(X_i, Y_j) = (X_i - \mu_X)(Y_j - \mu_Y)$, where $E(X) = \mu_X$ and $E(Y) = \mu_Y$

Thus, $E\left[(X - \mu_X)(Y - \mu_Y)\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_i - \mu_X)(Y_j - \mu_Y)p_{ij}$

The above expression, which is the mean of the product of deviations of values from their respective means, is known as the Covariance of X and Y denoted as Cov(X, Y) or σ_{XY}. Thus, we can write $Cov(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$

An alternative expression of Cov(X, Y)

$$\begin{aligned} Cov(X,Y) &= E\left[\{X - E(X)\}\{Y - E(Y)\}\right] \\ &= E\left[X.\{Y - E(Y)\} - E(X).\{Y - E(Y)\}\right] \\ &= E\left[X.Y - X.E(Y)\right] = E(X.Y) - E(X).E(Y) \end{aligned}$$

Note that E[{Y – E(Y)}] = 0, the sum of deviations of values from their arithmetic mean.

Mean and Variance of a Linear Combination

Let $Z = \phi(X, Y) = aX + bY$ be a linear combination of the two random variables X and Y, then using the theorem of addition of expectation, we can write

$$\mu_Z = E(Z) = E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y$$

Further, the variance of Z is given by

$$\begin{aligned} \sigma_Z^2 &= E\left[Z - E(Z)\right]^2 = E\left[aX + bY - a\mu_X - b\mu_Y\right]^2 = E\left[a(X - \mu_X) + b(Y - \mu_Y)\right]^2 \\ &= a^2E(X - \mu_X)^2 + b^2E(Y - \mu_Y)^2 + 2abE(X - \mu_X)(Y - \mu_Y) \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY} \end{aligned}$$

Example: A random variable X has the following probability distribution:

X	:	-2	-1	0	1	2
Probability	:	$\frac{1}{6}$	p	$\frac{1}{4}$	p	$\frac{1}{6}$

1. Find the value of p.
2. Calculate E(X + 2) and E(2X² + 3X + 5).

Solution:

Since the total probability under a probability distribution is equal to unity, the value of p should be such that $\frac{1}{6} + p + \frac{1}{4} + p + \frac{1}{6} = 1$.

Notes

This condition gives $p = \frac{5}{24}$

Further

$$E(X) = -2.\frac{1}{6} - 1.\frac{5}{24} + 0.\frac{1}{4} + 1.\frac{5}{24} + 2.\frac{1}{6} = 0$$

$$E(X^2) = 4.\frac{1}{6} + 1.\frac{5}{24} + 0.\frac{1}{4} + 1.\frac{5}{24} + 4.\frac{1}{6} = \frac{7}{4}$$

$$E(X+2) = E(X) + 2 = 0 + 2 = 2$$

And $E(2X^2 + 3X + 5) = 2E(X^2) + 3E(X) + 5 = 2.\frac{7}{4} + 0 + 5 = 8.5$

Self Assessment

Fill in the blanks:

9. Expected value of a constant is the .....
10. When a random variable is expressed in monetary units, its expected value is often termed as .....
11. If X and Y are two random variables, then E(X + Y) = ...

12.3 Counting Techniques

Counting techniques or combinatorial methods are often helpful in the enumeration of total number of outcomes of a random experiment and the number of cases favourable to the occurrence of an event.

12.3.1 Fundamental Principle of Counting

If the first operation can be performed in any one of the m ways and then a second operation can be performed in any one of the n ways, then both can be performed together in any one of the m × n ways.

This rule can be generalised. If first operation can be performed in any one of the n_1 ways, second operation in any one of the n_2 ways, ... kth operation in any one of the n_k ways, then together these can be performed in any one of the $n_1 \times n_2 \times ..... \times n_k$ ways.

12.3.2 Permutation

A permutation is an arrangement of a given set of objects in a definite order. Thus composition and order both are important in a permutation.

1. ***Permutations of n objects:*** The total number of permutations of n distinct objects is n!. Using symbols, we can write = n!, (where n denotes the permutations of n objects, all taken together).

   Let us assume there are n persons to be seated on n chairs. The first chair can be occupied by any one of the n persons and hence, there are n ways in which it can be occupied. Similarly, the second chair can be occupied in n – 1 ways and so on. Using the fundamental principle of counting, the total number of ways in which n chairs can be occupied by n persons or the permutations of n objects taking all at a time is given by:

   ^nP_n = n(n – 1)(n – 2) ..... 3.2.1 = n!

Notes

2. ***Permutations of n objects taking r at a time:*** In terms of the example, considered above, now we have n persons to be seated on r chairs,

   where $r \le n$.

   Thus, ${}^{n}P_{r}$ = n(n – 1)(n – 2) ..... [n – (r – 1)] = n(n – 1)(n – 2) ..... (n – r + 1).

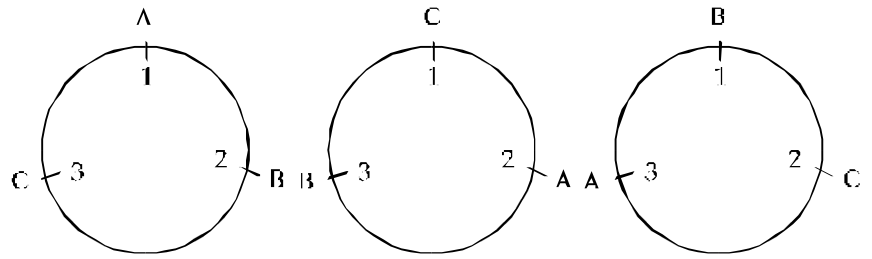
   On multiplication and division of the R.H.S. by (n - r)!, we get

$$ {}^{n}P_{r} = \frac{n(n-1)(n-2)\ ....\ (n-r+1)(n-r)!}{(n-r)!} = \frac{n!}{(n-r)!} $$

3. ***Permutations of n objects taking r at a time when any object may be repeated any number of times:*** Here, each of the r places can be filled in n ways. Therefore, total number of permutations is nr.

4. ***Permutations of n objects in a circular order:*** Suppose that there are three persons A, B and C, to be seated on the three chairs 1, 2 and 3, in a circular order. Then, the following three arrangements are identical:

Figure 12.1: Permutations of n objects in a circular order



   Similarly, if n objects are seated in a circle, there will be n identical arrangements of the above type. Thus, in order to obtain distinct permutation of n objects in circular order we divide ${}^{n}P_{n}$ by n, where ${}^{n}P_{n}$ denotes number of permutations in a row. Hence, the number of permutations in a circular order $\frac{n!}{n} = (n-1)!$

5. ***Permutations with restrictions:*** If out of n objects n_1 are alike of one kind, n_2 are alike of another kind, ... n_k are alike, the number of permutations are $\frac{n!}{n_1!n_2!\ ....\ n_k!}$

   Since permutation of ni objects, which are alike, is only one (i = 1, 2, ... k). Therefore, n! is to be divided by $n_1!$, $n_2!$ .... $n_k!$, to get the required permutations.

Example: What is the total number of ways of simultaneous throwing of (i) 3 coins, (ii) 2 dice and (iii) 2 coins and a die?

Solution:

1. Each coin can be thrown in any one of the two ways, i.e, a head or a tail, therefore, the number of ways of simultaneous throwing of 3 coins = 2^3 = 8.

2. Similarly, the total number of ways of simultaneous throwing of two dice is equal to 6^2 = 36 and

Notes

3. the total number of ways of simultaneous throwing of 2 coins and a die is equal to $2^2 \times 6 = 24$.

Example:

1. In how many ways can the letters of the word EDUCATION be arranged?
2. In how many ways can the letters of the word STATISTICS be arranged?
3. In how many ways can 20 students be allotted to 4 tutorial groups of 4, 5, 5 and 6 students respectively?
4. In how many ways 10 members of a committee can be seated at a round table if (a) they can sit anywhere (b) president and secretary must not sit next to each other?

Solution:

1. The given word EDUCATION has 9 letters. Therefore, number of permutations of 9 letters is 9! = 3,62,880.
2. The word STATISTICS has 10 letters in which there are 3S's, 3T's, 2I's, 1A and 1C. Thus, the required number of permutations = 50,400.
3. Required number of permutations = 9,77,72,87,522
4. (a) Number of permutations when they can sit anywhere = (10-1)!= 9! = 3,62,880.

   (b) We first find the number of permutations when president and secretary must sit together. For this we consider president and secretary as one person. Thus, the number of permutations of 9 persons at round table = 8! = 40,320.

$\therefore$ The number of permutations when president and secretary must not sit together = 3,62,880 – 40,320 = 3,22,560.

12.3.3 Combination

When no attention is given to the order of arrangement of the selected objects, we get a combination. We know that the number of permutations of n objects taking r at a time is ${}^nP_r$. Since r objects can be arranged in r! ways, therefore, there are r! permutations corresponding to one combination. Thus, the number of combinations of n objects taking r at a time, denoted by ${}^nC_r$, can be obtained by dividing ${}^nP_r$ by r!, i.e., ${}^nC_r = \frac{{}^nP_r}{r!} = \frac{n!}{r!(n-r)!}$

Note:

1. Since ${}^nC_r = {}^nC_{n-r}$, therefore, ${}^nC_r$ is also equal to the combinations of n objects taking (n - r) at a time.
2. The total number of combinations of n distinct objects taking 1, 2, ... n respectively, at a time is ${}^nC_1 + {}^nC_2 + \ldots\ldots + {}^nC_n = 2^n - 1$.

Notes

Tasks A committee of 8 teachers is to be formed out of 6 science, 8 arts teachers and a physical instructor. In how many ways the committee can be formed if

1. Any teacher can be included in the committee.
2. There should be 3 science and 4 arts teachers on the committee such that (i) any science teacher and any arts teacher can be included, (ii) one particular science teacher must be on the committee, (iii) three particular arts teachers must not be on the committee?

12.3.4 Ordered Partitions

1. ***Ordered Partitions (distinguishable objects)***

   (a) The total number of ways of putting n distinct objects into r compartments which are marked as 1, 2, ... r is equal to r^n.

   Since first object can be put in any of the r compartments in r ways, second can be put in any of the r compartments in r ways and so on.

   (b) The number of ways in which n objects can be put into r compartments such that the first compartment contains n_1 objects, second contains n_2 objects and so on the rth compartment contains n_r objects, where $n_1 + n_2 + ... + n_r = n$, is given by

   $$\frac{n!}{n_1!n_2!\ .....\ n_r!}.$$

   To illustrate this, let r = 3. Then n_1 objects in the first compartment can be put in ${}^nC_{n_1}$ ways. Out of the remaining $n - n_1$ objects, n_2 objects can be put in the second compartment in ${}^{n-n_1}C_{n_2}$ ways. Finally the remaining $n - n_1 - n_2 = n_3$ objects can be put in the third compartment in one way. Thus, the required number of ways is

   $${}^nC_{n_1} \times {}^{n-n_1}C_{n_2} = \frac{n!}{n_1!n_2!n_3!}.$$

2. ***Ordered Partitions (identical objects)***

   (a) The total number of ways of putting n identical objects into r compartments marked as 1, 2, ... r, is ${}^{n+r-1}C_{r-1}$, where each compartment may have none or any number of objects.

   We can think of n objects being placed in a row and partitioned by the (r - 1) vertical lines into r compartments. This is equivalent to permutations of (n + r - 1) objects out of which n are of one type and (r - 1) of another type. The required number of permutations are $\frac{(n+r-1)!}{n!(r-1)!}$, which is equal to ${}^{(n+r-1)}C_n$ or ${}^{(n+r-1)}C_{(r-1)}$.

   (b) The total number of ways of putting n identical objects into r compartments is ${}^{(n-r)+(r-1)}C_{(r-1)}$ or ${}^{(n-1)}C_{(r-1)}$, where each compartment must have at least one object.

   In order that each compartment must have at least one object, we first put one object in each of the r compartments. Then the remaining (n - r) objects can be placed as in (a) above.

Notes

(c) The formula, given in (b) above, can be generalised. If each compartment is supposed to have at least k objects, the total number of ways is ${}^{(n-kr)+(r-1)}C_{(r-1)}$, where k = 0, 1, 2, ... etc. such that $k < \frac{n}{r}$.

Example: 4 couples occupy eight seats in a row at random. What is the probability that all the ladies are sitting next to each other?

Solution:

Eight persons can be seated in a row in 8! ways.

We can treat 4 ladies as one person. Then, five persons can be seated in a row in 5! ways. Further, 4 ladies can be seated among themselves in 4! ways.

$\therefore$ The required probability $= \frac{5!4!}{8!} = \frac{1}{14}$

Example: 12 persons are seated at random (1) in a row, (2) in a ring. Find the probabilities that three particular persons are sitting together.

Solution:

1. The required probability $= \frac{10!3!}{12!} = \frac{1}{22}$

2. The required probability $= \frac{9!3!}{11!} = \frac{3}{55}$

Example: 5 red and 2 black balls, each of different sizes, are randomly laid down in a row. Find the probability that

1. the two end balls are black,
2. there are three red balls between two black balls and
3. the two black balls are placed side by side.

Solution:

The seven balls can be placed in a row in 7! ways.

1. The black can be placed at the ends in 2! ways and, in-between them, 5 red balls can be placed in 5! ways.

   $\therefore$ The required probability $= \frac{2!5!}{7!} = \frac{1}{21}$

2. We can treat BRRRB as one ball. Therefore, this ball along with the remaining two balls can be arranged in 3! ways. The sequence BRRRB can be arranged in 2! 3! ways and the three red balls of the sequence can be obtained from 5 balls in ways.

   $\therefore$ The required probability $= \frac{3!2!3!}{7!} \times {}^5C_3 = \frac{1}{7}$

Notes

3. The 2 black balls can be treated as one and, therefore, this ball along with 5 red balls can be arranged in 6! ways. Further, 2 black ball can be arranged in 2! ways.

   $\therefore$ The required probability $= \frac{6!2!}{7!} = \frac{2}{7}$

12.3.5 Statistical or Empirical Definition of Probability

The scope of the classical definition was found to be very limited as it failed to determine the probabilities of certain events in the following circumstances:

1. When n, the exhaustive outcomes of a random experiment is infinite.
2. When actual value of n is not known.
3. When various outcomes of a random experiment are not equally likely.

In addition to the above this definition doesn't lead to any mathematical treatment of probability.

In view of the above shortcomings of the classical definition, an attempt was made to establish a correspondence between relative frequency and the probability of an event when the total number of trials become sufficiently large.

Definition (R. Von Mises)

If an experiment is repeated n times, under essentially the identical conditions and, if, out of these trials, an event A occurs m times, then the probability that A occurs is given by P(A) = $\lim_{n\to\infty} \frac{m}{n}$, provided the limit exists.

This definition of probability is also termed as the empirical definition because the probability of an event is obtained by actual experimentation.

Although, it is seldom possible to obtain the limit of the relative frequency, the ratio $\frac{m}{n}$ can be regarded as a good approximation of the probability of an event for large values of n.

This definition also suffers from the following shortcomings:

1. The conditions of the experiment may not remain identical, particularly when the number of trials is sufficiently large.
2. The relative frequency, $\frac{m}{n}$, may not attain a unique value no matter how large is the total number of trials.
3. It may not be possible to repeat an experiment a large number of times.
4. Like the classical definition, this definition doesn't lead to any mathematical treatment of probability.

12.3.6 Axiomatic or Modern Approach to Probability

This approach was introduced by the Russian mathematician, A. Kolmogorov in 1930s. In his book, 'Foundations of Probability' published in 1933, he introduced probability as a function of the outcomes of an experiment, under certain restrictions. These restrictions are known as Postulates or Axioms of probability theory. Before discussing the above approach to probability, we shall explain certain concepts that are necessary for its understanding.

Notes

Sample Space

It is the set of all possible outcomes of a random experiment. Each element of the set is called a sample point or a simple event or an elementary event. The sample space of a random experiment is denoted by S and its element are denoted by e_i, where i = 1, 2, ... n. Thus, a sample space having n elements can be written as:

$$S = \{e_1, e_2, ..., e_n\}.$$

If a random experiment consists of rolling a six faced die, the corresponding sample space consists of 6 elementary events. Thus, S = {1, 2, 3, 4, 5, 6}.

Similarly, in the toss of a coin S = {H, T}.

The elements of S can either be single elements or ordered pairs. For example, if two coins are tossed, each element of the sample space would consist of the set of ordered pairs, as shown below :

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

Finite and Infinite Sample Space

A sample space consisting of finite number of elements is called a finite sample space, while if the number of elements is infinite, it is called an infinite sample space. The sample spaces discussed so far are examples of finite sample spaces. As an example of infinite sample space, consider repeated toss of a coin till a head appears. Various elements of the sample space would be:

$$S = \{(H), (T, H), (T, T, H), ... \}.$$

Discrete and Continuous Sample Space

A discrete sample space consists of finite or countably infinite number of elements. The sample spaces, discussed so far, are some examples of discrete sample spaces. Contrary to this, a continuous sample space consists of an uncountable number of elements. This type of sample space is obtained when the result of an experiment is a measurement on continuous scale like measurements of weight, height, area, volume, time, etc.

Event

An event is any subset of a sample space. In the experiment of roll of a die, the sample space is S = {1, 2, 3, 4, 5, 6}. It is possible to define various events on this sample space, as shown below:

Let A be the event that an odd number appears on the die. Then A = {1, 3, 5} is a subset of S. Further, let B be the event of getting a number greater than 4. Then B = {5, 6} is another subset of S. Similarly, if C denotes an event of getting a number 3 on the die, then C = {3}.

It should be noted here that the events A and B are composite while C is a simple or elementary event.

Occurrence of an Event

An event is said to have occurred whenever the outcome of the experiment is an element of its set. For example, if we throw a die and obtain 5, then both the events A and B, defined above, are said to have occurred.

Notes

It should be noted here that the sample space is certain to occur since the outcome of the experiment must always be one of its elements.

Definition of Probability (Modern Approach)

Let S be a sample space of an experiment and A be any event of this sample space. The probability of A, denoted by P(A), is defined as a real value set function which associates a real value corresponding to a subset A of the sample space S. In order that P(A) denotes a probability function, the following rules, popularly known as axioms or postulates of probability, must be satisfied.

Axiom I : For any event A in sample space S, we have $0 \le P(A) \le 1$.

Axiom II : P(S) = 1.

Axiom III : If $A_1, A_2, \ldots\ldots A_k$ are k mutually exclusive events (i.e., $A_i \bigcap_{i \neq j} A_j = \phi$, where ϕ denotes a null set) of the sample space S, then

$$P(A_1 \cup A_2 \ \ldots\ldots \ \cup A_k) = \sum_{i=1}^{k} P(A_i)$$

The first axiom implies that the probability of an event is a non-negative number less than or equal to unity. The second axiom implies that the probability of an event that is certain to occur must be equal to unity. Axiom III gives a basic rule of addition of probabilities when events are mutually exclusive.

The above axioms provide a set of basic rules that can be used to find the probability of any event of a sample space.

Probability of an Event

Let there be a sample space consisting of n elements, i.e., $S = \{e_1, e_2, \ldots\ldots e_n\}$. Since the elementary events $e_1, e_2, \ldots\ldots e_n$ are mutually exclusive, we have, according to axiom III, $P(S) = \sum_{i=1}^{n} P(e_i)$.

Similarly, if $A = \{e_1, e_2, \ldots\ldots e_m\}$ is any subset of S consisting of m elements, where $m \le n$, then $P(A) = \sum_{i=1}^{m} P(e_i)$. Thus, the probability of a sample space or an event is equal to the sum of probabilities of its elementary events.

It is obvious from the above that the probability of an event can be determined if the probabilities of elementary events, belonging to it, are known.

The Assignment of Probabilities to various Elementary Events

The assignment of probabilities to various elementary events of a sample space can be done in any one of the following three ways:

1. ***Using Classical Definition:*** We know that various elementary events of a random experiment, under the classical definition, are equally likely and, therefore, can be assigned equal probabilities. Thus, if there are n elementary events in the sample space of an

Notes

experiment and in view of the fact that $P(S)=\sum_{i=1}^{n} P(e_i)\ 1$ (from axiom II), we can assign a probability equal to $\frac{1}{n}$ to every elementary event or, using symbols, we can write $P(e_i)=\frac{1}{n}$ for i = 1, 2, ... n.

Further, if there are m elementary events in an event A, we have, $P(A)=\frac{1}{n}+\frac{1}{n}+\ .....\ +\frac{1}{n}\ \ (m \text{ times})=\frac{m}{n}\ =\frac{n(A),\ i.e.,\ number\ of\ elements\ in\ A}{n(S),\ i.e.,\ number\ of\ elements\ in\ S}$

We note that the above expression is similar to the formula obtained under classical definition.

2. ***Using Statistical Definition:*** Using this definition, the assignment of probabilities to various elementary events of a sample space can be done by repeating an experiment a large number of times or by using the past records.

3. ***Subjective Assignment:*** The assignment of probabilities on the basis of the statistical and the classical definitions is objective. Contrary to this, it is also possible to have subjective assignment of probabilities. Under the subjective assignment, the probabilities to various elementary events are assigned on the basis of the expectations or the degree of belief of the statistician. These probabilities, also known as personal probabilities, are very useful in the analysis of various business and economic problems where it is neither possible to repeat the experiment nor the outcomes are equally likely.

   It is obvious from the above that the Modern Definition of probability is a general one which includes the classical and the statistical definitions as its particular cases. Besides this, it provides a set of mathematical rules that are useful for further mathematical treatment of the subject of probability.

12.3.7 Theorems on Probability

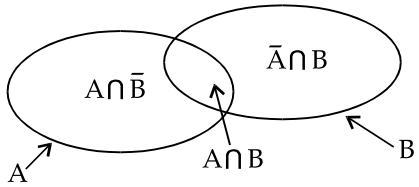
Theorem 1:

$P(\phi)=0$, where ϕ is a null set.

Theorem 2:

$P(\overline{A})=1-P(A)$, where $\overline{A}$ is compliment of A.

Figure 12.2: Venn Diagram



Notes

Theorem 3:

For any two events A and B in a sample space S

$$P\left(\overline{A}\cap B\right)=P\left(B\right)-P\left(A\cap B\right)$$

Theorem 4: (Addition of Probabilities)

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

Example: What is the probability of drawing a black card or a king from a well-shuffled pack of playing cards?

Solution:

There are 52 cards in a pack, $\therefore$ n(S) = 52.

Let A be the event that the drawn card is black and B be the event that it is a king. We have to find.

$$P\left(A\cup B\right)$$

Since there are 26 black cards, 4 kings and two black kings in a pack, we have n(A) = 26, n(B) = 4 and $n\left(A\cap \mathrm{B}\right)=2$ Thus, $P\left(A\cup B\right)=\dfrac{26+4-2}{52}=\dfrac{7}{13}$

Alternative Method

The given information can be written in the form of the following table:

	B	$\overline{B}$	$Total$
A	2	24	26
$\overline{A}$	2	24	26
$Total$	4	48	52

From the above, we can write

$$P\left(A\cup B\right)=1-P\left(\overline{A}\cap\overline{B}\right)=1-\frac{24}{52}=\frac{7}{13}$$

Theorem 5: (Multiplication or Compound Probability Theorem)

A compound event is the result of the simultaneous occurrence of two or more events. For convenience, we assume that there are two events, however, the results can be easily generalised. The probability of the compound event would depend upon whether the events are independent or not. Thus, we shall discuss two theorems: (a) Conditional Probability Theorem, and (b) Multiplicative Theorem for Independent Events.

1. ***Conditional Probability Theorem:*** For any two events A and B in a sample space S, the probability of their simultaneous occurrence, is given by

$$P\left(A\cap B\right)=P\left(A\right)P\left(B/A\right)$$

or equivalently $$=P\left(B\right)P\left(A/B\right)$$

Here, P(B/A) is the conditional probability of B given that A has already occurred. Similar interpretation can be given to the term P(A/B).

Notes

2. ***Multiplicative Theorem for Independent Events:*** If A and B are independent, the probability of their simultaneous occurrence is given by

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

Example: Two unbiased dice are tossed. Let w denote the number on the first die and r denote the number on the second die. Let A be the event that $w + r \leq 4$ and B be the event that $w + r \leq 3$. Are A and B independent?

Solution:

The sample space of this experiment consists of 36 elements, i.e., n(S) = 36. Also, A = {(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)} and B = {(1, 1), (1, 2), (2, 1)}.

From the above, we can write

$$P(A) = \frac{6}{36} = \frac{1}{6},\ P(B) = \frac{3}{36} = \frac{1}{12}$$

Also $(A \cap B) = \{(1,1),(1,2),(2,1)\}$ $\therefore$ $P(A \cap B) = \frac{3}{36} = \frac{1}{12}$

Since $P(A \cap B) \neq P(A)P(B)$, A and B are not independent.

Notes

Attributes

An attribute is a qualitative characteristics. One can only feel the presence or absence of this characteristic while observing individuals or items under consideration. For example, honesty, marriage, colour, beauty, etc., are attributes. It is, however, possible to classify various individuals or items into two or more categories according to an attribute.

Self Assessment

Multiple Choice Questions:

12. ..... techniques are often helpful in the enumeration of total number of outcomes of a random experiment and the number of cases favourable to the occurrence of an event.

    (a) Counting (b) Discounting

    (c) Estimation (d) Investigation

13. If the first operation can be performed in any one of the m ways and then a second operation can be performed in any one of the n ways, then both can be performed together in any one of the ..... ways.

    (a) m+n (b) m-n

    (c) m × n (d) none

Notes

14. A ..... is an arrangement of a given set of objects in a definite order.
    (a) Permutation (b) Combination
    (c) Arrangement (d) System
15. When no attention is given to the order of arrangement of the selected objects, we get a .....
    (a) Permutation (b) Combination
    (c) Arrangement (d) System
16. A ..... is the result of the simultaneous occurrence of two or more events.
    (a) Event (b) Simple event
    (c) Complex event (d) compound event

Case Study

Independent Events

Suppose a company hires both MBAs and non-MBAs for the same kind of managerial task. After a period of employment some of each category are promoted and some are not. Table below gives the proportion of company's managers among the said classes:

Promotional Status	Academic Qualification		Total
	MBA (A)	Non - MBA ($\overline{A}$)	
Promoted (B)	0.42	0.18	0.60
Not Promoted($\overline{B}$)	0.28	0.12	0.40
Total	0.70	0.30	1.00

Question

Calculate P(A/B) and P(B/A) and find out whether A and B are independent events.

12.4 Summary

1. (a) The number of permutations of n objects taking n at a time are n!
   (b) The number of permutations of n objects taking r at a time, are $^{n}P_r = \frac{n!}{(n-r)!}$
   (c) The number of permutations of n objects in a circular order are (n - 1)!
   (d) The number of permutations of n objects out of which n_1 are alike, n_2 are alike, ... n_k are alike, are $\frac{n!}{n_1!n_2!\ldots n_k!}$
   (e) The number of combinations of n objects taking r at a time are $^{n}C_r = \frac{n!}{r!(n-r)!}$
2. (a) The probability of occurrence of at least one of the two events A and B is given by: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1 - P(\overline{A} \cap \overline{B})$.
   (b) The probability of occurrence of exactly one of the events A or B is given by: $P(A \cap \overline{B}) + P(\overline{A} \cap B)$ or $P(A \cup B) - P(A \cap B)$

Notes

3. (a) The probability of simultaneous occurrence of the two events A and B is given by: $P(A \cap B) = P(A).P(B/A)$ or $= P(B).P(A/B)$

   (b) If A and B are independent $P(A \cap B) = P(A).P(B)$.

12.5 Keywords

Combination: When no attention is given to the order of arrangement of the selected objects, we get a combination.

Counting techniques or combinatorial methods: These are often helpful in the enumeration of total number of outcomes of a random experiment and the number of cases favourable to the occurrence of an event.

Equally likely outcomes: The outcomes of random experiment are said to be equally likely or equally probable if the occurrence of none of them is expected in preference to others.

Expected Monetary Value: When a random variable is expressed in monetary units, its expected value is often termed as expected monetary value and symbolized by EMV.

Expected Value: Expected value of a constant is the constant itself, i.e., E(b) = b, where b is a constant.

Mutually exclusive outcomes: Two or more outcomes of an experiment are said to be mutually exclusive if the occurrence of one of them precludes the occurrence of all others in the same trial i.e. they cannot occur jointly.

Permutation: A permutation is an arrangement of a given set of objects in a definite order. Thus composition and order both are important in a permutation

Priori' definition of probability: If n is the number of equally likely, mutually exclusive and exhaustive outcomes of a random experiment out of which m outcomes are favourable to the occurrence of an event A, then the probability that A occurs, denoted by P(A).

Random phenomenon: A phenomenon or an experiment which can result into more than one possible outcome, is called a random phenomenon or random experiment or statistical experiment.

12.6 Review Questions

1. Define the term 'probability' by (a) The Classical Approach, (b) The Statistical Approach. What are the main limitations of these approaches?
2. Discuss the axiomatic approach to probability. In what way it is an improvement over classical and statistical approaches?
3. Explain the meaning of conditional probability. State and prove the multiplication rule of probability of two events when (a) they are not independent, (b) they are independent.
4. Explain the concept of independence and mutually exclusiveness of two events A and B. If A and B are
5. Explain the meaning of a statistical experiment and corresponding sample space. Write down the sample space of an experiment of simultaneous toss of two coins and a die.
6. What is the probability of getting exactly two heads in three throws of an unbiased coin?
7. What is the probability of getting a sum of 2 or 8 or 12 in single throw of two unbiased dice?

Notes

8. Two cards are drawn at random from a pack of 52 cards. What is the probability that the first is a king and second is a queen?
9. What is the probability of successive drawing of an ace, a king, a queen and a jack from a pack of 52 well shuffled cards? The drawn cards are not replaced.
10. 5 unbiased coins with faces marked as 2 and 3 are tossed. Find the probability of getting a sum of 12.
11. If 15 chocolates are distributed at random among 5 children, what is the probability that a particular child receives 8 chocolates?
12. A and B stand in a ring with 10 other persons. If arrangement of 12 persons is at random, find the chance that there are exactly three persons between A and B.
13. Two different digits are chosen at random from the set 1, 2, 3, 4, 5, 6, 7, 8. Find the probability that sum of two digits exceeds 13.
14. 5-letter words are formed from the letters of the word ORDINATES. What is the probability that the word so formed consists of 2 vowels and 3 consonants?
15. Maximum number of different committees are formed out of 100 teachers, including principal, of a college such that each committee consists of the same number of members. What is the probability that principal is a member of any committee?
16. If n persons are seated around a round table, find the probability that in no two ways a man has the same neighbours.
17. 6 teachers, of whom 2 are from science, 2 from arts and 2 from commerce, are seated in a row. What is the probability that the teachers of the same discipline are sitting together?
18. A problem in economics is given to 3 students whose chances of solving it are and respectively. What is the probability that the problem will be solved?
19. What is the chance that a non-leap year selected at random will contain 53 Sundays?
20. Two men M_1 and M_2 and three women W_1, W_2 and W_3, in a big industrial firm, are trying for promotion to a single post which falls vacant. Those of the same sex have equal probabilities of getting promotion but each man is twice as likely to get the promotion as any women.
    (a) Find the probability that a woman gets the promotion.
    (b) If M_2 and W_2 are husband and wife, find the probability that one of them gets the promotion.

Answers: Self Assessment

1.	True	2.	True
3.	True	4.	True
5.	False	6.	True
7.	True	8.	True
9.	constant itself	10.	expected monetary value
11.	E(X) + E(Y).	12.	(a)
13.	(c)	14.	(a)
15.	(b)	16.	(d)

Notes

12.7 Further Readings

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002, Excel Books, New Delhi.

Bhardwaj R S., *Business Statistics*, Excel Books.

Selvaraj R, Loganathan, *C Quantitative Methods in Management.*

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Online links

http://en.wikipedia.org/wiki/Probability

http://www.cut-the-knot.org/probability.shtml

http://www.probability.net/

http://www.tutors4you.com/probabilitytutorial.htm

Notes

Unit 13: Binomial Probability Distribution

CONTENTS

Objectives
Introduction
13.1 Concept of Probablity Distribution
13.1.1 Probability Distribution of a Random Variable
13.1.2 Discrete and Continuous Probability Distributions
13.2 The Binomial Probability Distribution
13.2.1 Probability Function or Probability Mass Function
13.2.2 Summary Measures of Binomial Distribution
13.3 Fitting of Binomial Distribution
13.3.1 Features of Binomial Distribution
13.3.2 Uses of Binomial Distribution
13.4 Summary
13.5 Keywords
13.6 Review Questions
13.7 Further Readings

Objectives

After studying this unit, you will be able to:

- Brief about theoretical probability distribution
- Categorize theoretical probability distribution
- Define the term binomial probability distribution
- Explain the various features of binomial probability distribution
- Discuss the uses and measures of binomial probability distribution

Introduction

The study of a population can be done either by constructing an observed (or empirical) frequency distribution, often based on a sample from it, or by using a theoretical distribution. We have already studied the construction of an observed frequency distribution and its various summary measures. Now we shall learn a more scientific way to study a population through the use of theoretical probability distribution of a random variable. It may be mentioned that a theoretical probability distribution gives us a law according to which different values of the random variable are distributed with specified probabilities. It is possible to formulate such laws either on the basis of given conditions (a priori considerations) or on the basis of the results (a posteriori inferences) of an experiment.

If a random variable satisfies the conditions of a theoretical probability distribution, then this distribution can be fitted to the observed data.

Notes

The knowledge of the theoretical probability distribution is of great use in the understanding and analysis of a large number of business and economic situations. For example, with the use of probability distribution, it is possible to test a hypothesis about a population, to take decision in the face of uncertainty, to make forecast, etc.

Theoretical probability distributions can be divided into two broad categories, viz. discrete and continuous probability distributions, depending upon whether the random variable is discrete or continuous. Although, there are a large number of distributions in each category, we shall discuss only some of them having important business and economic applications.

Did u know? In order to discuss the applications of probability to practical situations, it is necessary to associate some numerical characteristics with each possible outcome of the random experiment. This numerical characteristic is termed as random variable.

13.1 Concept of Probablity Distribution

A probability distribution is a rule that assigns a probability to every possible outcome of an experiment.

In order to discuss the applications of probability to practical situations, it is necessary to associate some numerical characteristics with each possible outcome of the random experiment. This numerical characteristic is termed as random variable. Or we can say, An event whose numerical value is determined by the outcome of an experiment is called a variate or often a random variable.

Example: Three coins are tossed simultaneously. Write down the sample space of the random experiment. What are the possible values of the random variable X, if it denotes the number of heads obtained?

Solution:

The sample space of the experiment can be written as:

S = {(H,H,H), (H,H,T), (H,T,H), (T,H,H), (H,T,T), (T,H,T), (T,T,H), (T,T,T)}

We note that the first element of the sample space denotes 3 heads, therefore, the corresponding value of the random variable will be 3. Similarly, the value of the random variable corresponding to each of the second, third and fourth element will be 2 and it will be 1 for each of the fifth, sixth and seventh element and 0 for the last element. Thus, the random variable X, defined above can take four possible values, i.e., 0, 1, 2 and 3.

It may be pointed out here that it is possible to define another random variable on the above sample space.

13.1.1 Probability Distribution of a Random Variable

Given any random variable, corresponding to a sample space, it is possible to associate probabilities to each of its possible values. For example, in the toss of 3 coins, assuming that they are unbiased, the probabilities of various values of the random variable X, defined in example above, can be written as:

$$P(X=0)=\frac{1}{8}, P(X=1)=\frac{3}{8}, P(X=2)=\frac{3}{8} \text{ and } P(X=3)=\frac{1}{8}.$$

Notes

The set of all possible values of the random variable X along with their respective probabilities is termed as Probability Distribution of X. The probability distribution of X, defined in example above, can be written in a tabular form as given below:

X	:	0	1	2	3	Total
$p(X)$	:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

Note that the total probability is equal to unity.

In general, the set of n possible values of a random variable X, i.e., $\{X_1, X_2, \ldots\ldots X_n\}$ along with their respective probabilities $p(X_1), p(X_2), \ldots\ldots p(X_n)$, where $\sum_{i=1}^{n} p(X_i) = 1$, is called a probability distribution of X. The expression p(X) is called the probability function of X.

13.1.2 Discrete and Continuous Probability Distributions

Like any other variable, a random variable X can be discrete or continuous. If X can take only finite or countably infinite set of values, it is termed as a discrete random variable. On the other hand, if X can take an uncountable set of infinite values, it is called a continuous random variable.

The random variable defined in previous example is a discrete random variable. However, if X denotes the measurement of heights of persons or the time interval of arrival of a specified number of calls at a telephone desk, etc., it would be termed as a continuous random variable.

The distribution of a discrete random variable is called the Discrete Probability Distribution and the corresponding probability function p(X) is called a Probability Mass Function. In order that any discrete function p(X) may serve as probability function of a discrete random variable X, the following conditions must be satisfied:

1. $p(X_i) \geq 0 \; \forall \; i = 1, 2, \ldots\ldots n$ and
2. $\sum_{i=1}^{n} p(X_i) = 1$

In a similar way, the distribution of a continuous random variable is called a Continuous Probability Distribution and the corresponding probability function p(X) is termed as the Probability Density Function. The conditions for any function of a continuous variable to serve as a probability density function are:

1. $p(X) \geq 0 \; \forall$ real values of X, and
2. $\int_{-\infty}^{\infty} p(X)\,dX = 1$

Remarks:

1. When X is a continuous random variable, there are an infinite number of points in the sample space and thus, the probability that X takes a particular value is always defined to be zero even though the event is not regarded as impossible. Hence, we always measure the probability of a continuous random variable lying in an interval.
2. The concept of a probability distribution is not new. In fact it is another way of representing a frequency distribution. Using statistical definition, we can treat the relative frequencies of various values of the random variable as the probabilities.

Notes

Example: Two unbiased die are thrown. Let the random variable X denote the sum of points obtained. Construct the probability distribution of X.

Solution:

The possible values of the random variable are:

2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

The probabilities of various values of X are shown in the following table:

Probability Distribution of X

X	2	3	4	5	6	7	8	9	10	11	12	Total
p(X)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Example: Three marbles are drawn at random from a bag containing 4 red and 2 white marbles. If the random variable X denotes the number of red marbles drawn, construct the probability distribution of X.

Solution:

The given random variable can take 3 possible values, i.e., 1, 2 and 3. Thus, we can compute the probabilities of various values of the random variable as given below:

$$P(X = 1, \text{ i.e., 1R and 2 W marbles are drawn}) = \frac{{}^4C_1 \times {}^2C_2}{{}^6C_3} = \frac{4}{20}$$

$$P(X = 2, \text{ i.e., 2R and 1W marbles are drawn}) = \frac{{}^4C_2 \times {}^2C_1}{{}^6C_3} = \frac{12}{20}$$

$$P(X = 3, \text{ i.e., 3R marbles are drawn}) = \frac{{}^4C_3}{{}^6C_3} = \frac{4}{20}$$

Note: In the event of white balls being greater than 2, the possible values of the random variable would have been 0, 1, 2 and 3.

13.2 The Binomial Probability Distribution

If the assumptions of the Bernoulli process are satisfied and if the probability of a success on one trial is p, then the probability distribution of the number of successes, r, in n trials, is a binomial distribution, and is given by the formula:

$$P = {}_nC_{n-r}\, p^r (1-p)^{n-r}$$

Performing computations using the above equation can be tedious if the number of trials is large.

Binomial distribution is a theoretical probability distribution which was given by James Bernoulli. This distribution is applicable to situations with the following characteristics:

1. An experiment consists of a finite number of repeated trials.
2. Each trial has only two possible, mutually exclusive, outcomes which are termed as a 'success' or a 'failure'.

Notes

3. The probability of a success, denoted by p, is known and remains constant from trial to trial. The probability of a failure, denoted by q, is equal to 1 – p.
4. The sequence of trials under the above assumptions is also termed as Bernoulli Trials

Caution Different trials are independent, i.e., outcome of any trial or sequence of trials has no effect on the outcome of the subsequent trials.

13.2.1 Probability Function or Probability Mass Function

Let n be the total number of repeated trials, p be the probability of a success in a trial and q be the probability of its failure so that q = 1 – p.

Let r be a random variable which denotes the number of successes in n trials. The possible values of r are 0, 1, 2, ... n. We are interested in finding the probability of r successes out of n trials, i.e., P(r).

To find this probability, we assume that the first r trials are successes and remaining n – r trials are failures. Since different trials are assumed to be independent, the probability of this sequence is

$$\underbrace{p.p.\ \ldots.\ p}_{r\ times}\ \underbrace{q.q.\ \ldots.\ q}_{(n-r)\ times} \text{ i.e. prqn-r.}$$

Since out of n trials any r trials can be success, the number of sequences showing any r trials as success and remaining (n – r) trials as failure is ${}^nC_r$, where the probability of r successes in each trial is p^rq^{n-r}. Hence, the required probability is , $P(r) = {}^nC_r p^r q^{n-r}$ where r = 0, 1, 2, ... n.

Writing this distribution in a tabular form, we have:

r	0	1	2	……	n	*Total*
$P(r)$	${}^nC_0p^0q^n$	${}^nC_1p\ q^{n-1}$	${}^nC_2p^2q^{n-2}$	……	${}^nC_np^nq^0$	1

It should be noted here that the probabilities obtained for various values of r are the terms in the binomial expansion of (q + p)n and thus, the distribution is termed as Binomial Distribution. $P(r) = {}^nC_r p^r q^{n-r}$ is termed as the probability function or probability mass function (p.m.f.) of the distribution.

13.2.2 Summary Measures of Binomial Distribution

1. ***Mean:*** The mean of a binomial variate r, denoted by μ , is equal to E(r), i.e.,

$$\mu = E(r) = \sum_{r=0}^{n} rP(r) = \sum_{r=1}^{n} r.{}^nC_r p^r q^{n-r} \quad \text{(note that the term for r = 0 is 0)}$$

$$= \sum_{r=1}^{n} \frac{r.n!}{r!(n-r)!} \cdot p^r q^{n-r} = \sum_{r=1}^{n} \frac{n.(n-1)!}{(r-1)!(n-r)!} \cdot p^r q^{n-r}$$

$$= np\sum_{r=1}^{n} \frac{(n-1)!}{(r-1)!(n-r)!} \cdot p^{r-1} q^{n-r} = np(q+p)^{n-1} = np \quad [\because q+p=1]$$

2. ***Variance:*** The variance of r, denoted by σ^2 , is given by

$$\sigma^2 = E\left[r - E(r)\right]^2 = E\left[r - np\right]^2 = E\left[r^2 - 2npr + n^2p^2\right]$$

Notes

$$= E\left(r^2\right) - 2npE\left(r\right) + n^2p^2 = E\left(r^2\right) - 2n^2p^2 + n^2p^2$$

$$= E\left(r^2\right) - n^2p^2 \qquad \text{.... (1)}$$

Thus, to find we first determine $E(r^2)$.

Now, $E\left(r^2\right) = \sum_{r=1}^{n} r^2 . {}^nC_r p^r q^{n-r} = \left[r\left(r-1\right)+r\right] {}^nC_r p^r q^{n-r}$

$$= \sum_{r=2}^{n} r\left(r-1\right) {}^nC_r p^r q^{n-r} + \sum_{r=1}^{n} r . {}^nC_r p^r q^{n-r} = \sum_{r=2}^{n} \frac{r\left(r-1\right)n!}{r!\left(n-r\right)!} \,.\, p^r q^{n-r} + np$$

$$= \sum_{r=2}^{n} \frac{n!}{\left(r-2\right)!\left(n-r\right)!} \,.\, p^r q^{n-r} + np = \sum_{r=2}^{n} \frac{n\left(n-1\right).\left(n-2\right)!}{\left(r-2\right)!\left(n-r\right)!} \,.\, p^r q^{n-r} + np$$

$$= n\left(n-1\right)p^2 \sum_{r=2}^{n} \frac{\left(n-2\right)!}{\left(r-2\right)!\left(n-r\right)!} \,.\, p^{r-2} q^{n-r} + np$$

$$= n\left(n-1\right)p^2\left(q+p\right)^{n-2} + np = n\left(n-1\right)p^2 + np$$

Substituting this value in equation (1), we get

$$\sigma^2 = n\left(n-1\right)p^2 + np - n^2p^2 = np\left(1-p\right) = npq$$

Or the standard deviation $= \sqrt{npq}$

Remarks: $\sigma^2 = npq = mean \times q$, which shows that since 0 < q <1.

3. The values of μ_3, μ_4, β_1 and β_2

Proceeding as above, we can obtain

$$\mu_3 = E\left(r-np\right)^3 = npq\left(q-p\right)$$

$$\mu_4 = E\left(r-np\right)^4 = 3n^2p^2q^2 + npq\left(1-6pq\right)$$

Also $\qquad \beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{n^2p^2q^2\left(q-p\right)^2}{n^3p^3q^3} = \dfrac{\left(q-p\right)^2}{npq}$

The above result shows that the distribution is symmetrical when

p = q = $\frac{1}{2}$, negatively skewed if q < p, and positively skewed if q > p

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3n^2p^2q^2 + npq\left(1-6pq\right)}{n^2p^2q^2} = 3 + \frac{\left(1-6pq\right)}{npq}$$

The above result shows that the distribution is leptokurtic if 6pq < 1, platykurtic if 6pq > 1 and mesokurtic if 6pq = 1.

4. ***Mode:*** Mode is that value of the random variable for which probability is maximum.

If r is mode of a binomial distribution, we have

$P(r-1) \le P(r) \ge P(r+1)$

Consider the inequality $P(r) \ge P(r+1)$

Notes

or ${}^nC_r p^r q^{n-r} \geq {}^nC_{r+1} p^{r+1} q^{n-r-1}$

or $\dfrac{n!}{r!(n-r)!} p^r q^{n-r} \geq \dfrac{n!}{(r+1)!(n-r-1)!} p^{r+1} q^{n-r-1}$

or $\dfrac{1}{(n-r)}.q \geq \dfrac{1}{(r+1)}.p$ or $qr + q \geq np - pr$

Solving the above inequality for r, we get

$$r \geq (n+1)p - 1 \qquad \text{.... (1)}$$

Similarly, on solving the inequality P(r - 1) ≤ P(r) for r, we can get

$$r \leq (n+1)p \qquad \text{.... (2)}$$

Combining inequalities (1) and (2), we get

$$(n+1)p - 1 \leq r \leq (n+1)p$$

Case I. When (n + 1)p is not an integer

When (n + 1)p is not an integer, then (n + 1)p - 1 is also not an integer. Therefore, mode will be an integer between (n + 1)p - 1 and (n + 1)p or mode will be an integral part of (n + 1)p.

Case II. When (n + 1)p is an integer

When (n + 1)p is an integer, the distribution will be bimodal and the two modal values would be (n + 1)p - 1 and (n + 1)p.

Example: An unbiased die is tossed three times. Find the probability of obtaining (1) no six, (2) one six, (3) at least one six, (4) two sixes and (5) three sixes.

Solution.

The three tosses of a die can be taken as three repeated trials which are independent. Let the occurrence of six be termed as a success. Therefore, r will denote the number of six obtained. Further, n = 3 and $p = \dfrac{1}{6}$.

1. Probability of obtaining no six, i.e.,

$$P(r=0) = {}^3C_0 p^0 q^3 = 1.\left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^3 = \frac{125}{216}$$

2. $P(r=1) = {}^3C_1 p^1 q^2 = 3.\left(\dfrac{1}{6}\right)\left(\dfrac{5}{6}\right)^2 = \dfrac{25}{72}$

3. Probability of getting at least one six = 1 – P(r = 0) $= 1 - \dfrac{125}{216} = \dfrac{91}{216}$

4. $P(r=2) = {}^3C_2 p^2 q^1 = 3.\left(\dfrac{1}{6}\right)^2 \left(\dfrac{5}{6}\right) = \dfrac{5}{72}$

5. $P(r=3) = {}^3C_3 p^3 q^0 = 3.\left(\dfrac{1}{6}\right)^3 = \dfrac{1}{216}$

Notes

Example: Assuming that it is true that 2 in 10 industrial accidents are due to fatigue, find the probability that:

1. Exactly 2 of 8 industrial accidents will be due to fatigue.
2. At least 2 of the 8 industrial accidents will be due to fatigue.

Solution:

Eight industrial accidents can be regarded as Bernoulli trials each with probability of success $p = \frac{2}{10} = \frac{1}{5}$. The random variable r denotes the number of accidents due to fatigue.

1. $P(r=2) = {}^8C_2\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^6 = 0.294$

2. We have to find P(r ≥ 2). We can write
   P(r ≥ 2) = 1 – P(0) – P(1), thus, we first find P(0) and P(1).

   We have $P(0) = {}^8C_0\left(\frac{1}{5}\right)^0\left(\frac{4}{5}\right)^8 = 0.168$

   and $P(1) = {}^8C_1\left(\frac{1}{5}\right)^1\left(\frac{4}{5}\right)^7 = 0.336$

   ∴ P(r ≥ 2) = 1- 0.168 - 0.336 = 0.496

Example: The proportion of male and female students in a class is found to be 1 : 2. What is the probability that out of 4 students selected at random with replacement, 2 or more will be females?

Solution:

Let the selection of a female student be termed as a success. Since the selection of a student is made with replacement, the selection of 4 students can be taken as 4 repeated trials each with probability of success $p = \frac{2}{3}$.

Thus, P(r ≥ 2) = P(r = 2) + P(r = 3) +P(r = 4)

$$= {}^4C_2\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^2 + {}^4C_3\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right) + {}^4C_4\left(\frac{2}{3}\right)^4 = \frac{8}{9}$$

Note that P(r ≥ 2) can alternatively be found as 1 – P(0) – P(1)

Task The probability of a bomb hitting a target is 1/5. Two bombs are enough to destroy a bridge. If six bombs are aimed at the bridge, find the probability that the bridge is destroyed.

Example: An insurance salesman sells policies to 5 men all of identical age and good health. According to the actuarial tables, the probability that a man of this particular age will be alive 30 years hence is 2/3. Find the probability that 30 years hence (1) at least 1 man will be alive, (2) at least 3 men will be alive.

Notes

Solution:

Let the event that a man will be alive 30 years hence be termed as a success. Therefore, n = 5 and $p = \frac{2}{3}$

1. $P(r \geq 1) = 1 - P(r = 0)\ 1 - {}^5C_0\left(\frac{2}{3}\right)^0\left(\frac{1}{3}\right)^5 = \frac{242}{243}$

2. $P(r \geq 3) = P(r = 3) + P(r = 4) + P(r = 5)$

$$= {}^5C_3\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^2 + {}^5C_4\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^1 + {}^5C_5\left(\frac{2}{3}\right)^5 = \frac{64}{81}$$

Example: Ten percent of items produced on a machine are usually found to be defective. What is the probability that in a random sample of 12 items (1) none, (2) one, (3) two, (4) at the most two, (5) at least two items are found to be defective?

Solution:

Let the event that an item is found to be defective be termed as a success. Thus, we are given n = 12 and p = 0.1.

1. $P(r = 0) = {}^{12}C_0(0.1)^0(0.9)^{12} = 0.2824$

2. $P(r = 1) = {}^{12}C_1(0.1)^1(0.9)^{11} = 0.3766$

3. $P(r = 2) = {}^{12}C_2(0.1)^2(0.9)^{10} = 0.2301$

4. $P(r \leq 2) = P(r = 0) + P(r = 1) + P(r = 2)$

   $= 0.2824 + 0.3766 + 0.2301 = 0.8891$

5. $P(r \geq 2) = 1 - P(0) - P(1) = 1 - 0.2824 - 0.3766 = 0.3410$

Example: In a large group of students 80% have a recommended statistics book. Three students are selected at random. Find the probability distribution of the number of students having the book. Also compute the mean and variance of the distribution.

Solution:

Let the event that 'a student selected at random has the book' be termed as a success. Since the group of students is large, 3 trials, i.e., the selection of 3 students, can be regarded as independent with probability of a success p = 0.8. Thus, the conditions of the given experiment satisfies the conditions of binomial distribution.

The probability mass function $P(r) = {}^3C_r(0.8)^r(0.2)^{3-r}$

where r = 0, 1, 2 and 3

The mean is np = 3 $\times$ 0.8 = 2.4 and Variance is npq = 2.4 $\times$ 0.2 = 0.48

Example:

1. The mean and variance of a discrete random variable X are 6 and 2 respectively. Assuming X to be a binomial variate, find $P(5 \leq X \leq 7)$.

Notes

2. In a binomial distribution consisting of 5 independent trials, the probability of 1 and 2 successes are 0.4096 and 0.2048 respectively. Calculate the mean, variance and mode of the distribution.

Solution:

1. It is given that np = 6 and npq = 2

$\therefore\ q = \frac{npq}{np} = \frac{2}{6} = \frac{1}{3}$ so that $p = 1 - \frac{1}{3} = \frac{2}{3}$ and $n = 6 \times \frac{3}{2} = 9$

Now P(5 ≤ X ≤ 7) = P(X = 5) + P(X = 6) +P(X = 7)

$$= {}^9C_5\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^4 + {}^9C_6\left(\frac{2}{3}\right)^6\left(\frac{1}{3}\right)^3 + {}^9C_7\left(\frac{2}{3}\right)^7\left(\frac{1}{3}\right)^2$$

$$= \frac{2^5}{3^9}\left[{}^9C_5 + {}^9C_6 \times 2 + {}^9C_7 \times 4\right] = \frac{2^5}{3^9} \times 438$$

2. Let p be the probability of a success. It is given that

${}^5C_1 p(1-p)^4 = 0.4096$ and ${}^5C_2 p^2(1-p)^3 = 0.2048$

Using these conditions, we can write

$\frac{5p(1-p)^4}{10p^2(1-p)^3} = \frac{0.4096}{0.2048} = 2$ or $\frac{(1-p)}{p} = 4$. This gives $p = \frac{1}{5}$

Thus, mean is $np = 5 \times \frac{1}{5} = 1$ and $npq = 1 \times \frac{4}{5} = 0.8$

Since (n +1)p, i.e $6 \times \frac{1}{5}$ is not an integer, mode is its integral part, i.e., = 1.

Example: 5 unbiased coins are tossed simultaneously and the occurrence of a head is termed as a success. Write down various probabilities for the occurrence of 0, 1, 2, 3, 4, 5 successes. Find mean, variance and mode of the distribution.

Solution.

Here n = 5 and $p = q = \frac{1}{2}$.

The probability mass function is $P(r) = {}^5C_r\left(\frac{1}{2}\right)^5$, r = 0, 1, 2, 3, 4, 5.

The probabilities of various values of r are tabulated below:

r	0	1	2	3	4	5	*Total*
$P(r)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$	1

Mean = $np = 5 \times \frac{1}{2} = 2.5$ and variance $= 2.5 \times \frac{1}{2} = 1.25$

Since (n +1)p = $6 \times \frac{1}{2} = 3$ is an integer, the distribution is bimodal and the two modes are 2 and 3.

Notes

Notes

Pascal Distribution

In binomial distribution, we derived the probability mass function of the number of successes in n (fixed) Bernoulli trials. We can also derive the probability mass function of the number of Bernoulli trials needed to get r (fixed) successes. This distribution is known as Pascal distribution. Here r and p become parameters while n becomes a random variable.

We may note that r successes can be obtained in r or more trials i.e. possible values of the random variable are r, (r + 1), (r + 2), ... etc. Further, if n trials are required to get r successes, the nth trial must be a success. Thus, we can write the probability mass function of Pascal distribution as follows:

$$P(n) = \begin{pmatrix} \text{Probability of } (r-1) \text{ successes} \\ \text{out of } (n-1) \text{ trials} \end{pmatrix} \times \begin{pmatrix} \text{Probability of a success} \\ \text{in nth trial} \end{pmatrix}$$

$$= {}^{n-1}C_{r-1} p^{r-1} q^{n-r} \times p = {}^{n-1}C_{r-1} p^{r} q^{n-r}$$

where n = r, (r + 1), (r + 2), ... etc.

It can be shown that the mean and variance of Pascal distribution are $\frac{r}{p}$ and $\frac{rq}{p^2}$ respectively.

This distribution is also known as Negative Binomial Distribution because various values of P(n) are given by the terms of the binomial expansion of $p^r(1 - q)^{-r}$.

Self Assessment

State whether the following statements are true or false:

1. The study of a population can be done either by constructing an observed (or empirical) frequency distribution, often based on a sample from it, or by using a theoretical distribution.
2. It is not possible to formulate various laws either on the basis of given conditions (a priori considerations) or on the basis of the results (a posteriori inferences) of an experiment.
3. If a random variable satisfies the conditions of a theoretical probability distribution, then this distribution can be fitted to the observed data.
4. The knowledge of the theoretical probability distribution is of no use in the understanding and analysis of a large number of business and economic situations.
5. It is possible to test a hypothesis about a population, to take decision in the face of uncertainty, to make forecast, etc.
6. Theoretical probability distributions can be divided into two broad categories, viz. discrete and continuous probability distributions.
7. Binomial distribution is a theoretical probability distribution which was given by James Bernoulli.
8. In Binomial distribution, an experiment consists of a finite number of repeated trials.
9. Each trial has only two possible, mutually exclusive, outcomes which are termed as a 'success' or a 'failure'.

Notes

10. Theoretical probability distribution gives us a law according to which different values of the random variable are distributed with non-specified probabilities.

13.3 Fitting of Binomial Distribution

The fitting of a distribution to given data implies the determination of expected (or theoretical) frequencies for different values of the random variable on the basis of this data.

The purpose of fitting a distribution is to examine whether the observed frequency distribution can be regarded as a sample from a population with a known probability distribution.

To fit a binomial distribution to the given data, we find its mean. Given the value of n, we can compute the value of p and, using n and p, the probabilities of various values of the random variable. These probabilities are multiplied by total frequency to give the required expected frequencies. In certain cases, the value of p may be determined by the given conditions of the experiment.

Example: The following data give the number of seeds germinating (X) out of 10 on damp filter for 80 sets of seed. Fit a binomial distribution to the data.

X :	0	1	2	3	4	5	6	7	8	9	10
f :	6	20	28	12	8	6	0	0	0	0	0

Solution:

Here the random variable X denotes the number of seeds germinating out of a set of 10 seeds. The total number of trials n = 10.

The mean of the given data $\bar{X} = \frac{0\times 6+1\times 20+2\times 28+3\times 12+4\times 8+5\times 6}{80} = \frac{174}{80} = 2.175$

Since mean of a binomial distribution is np, $\therefore$ np = 2.175. Thus, we get $p = \frac{2.175}{10} = 0.22$ (approx.)

Further, q = 1 - 0.22 = 0.78.

Using these values, we can compute $P(X) = {}^{10}C_X (0.22)^X (0.78)^{10-X}$ and then expected frequency [= N × P(X)] for X = 0, 1, 2, ... 10. The calculated probabilities and the respective expected frequencies are shown in the following table:

X	$P(X)$	$N\times P(X)$	*Approximated Frequency*	X	$P(X)$	$N\times P(X)$	*Approximated Frequency*
0	0.0834	6.67	6	6	0.0088	0.71	1
1	0.2351	18.81	19	7	0.0014	0.11	0
2	0.2984	23.87	24	8	0.0001	0.01	0
3	0.2244	17.96	18	9	0.0000	0.00	0
4	0.1108	8.86	9	10	0.0000	0.00	0
5	0.0375	3.00	3	*Total*	1.0000		80

13.3.1 Features of Binomial Distribution

1. It is a discrete probability distribution.
2. It depends upon two parameters n and p. It may be pointed out that a distribution is known if the values of its parameters are known.

Notes

3. The total number of possible values of the random variable are n + 1. The successive binomial coefficients are ${}^nC_0$, ${}^nC_1$, ${}^nC_2$, ... ${}^nC_n$ Further, since ${}^nC_r = {}^nC_{n-r}$, these coefficients are symmetric.

   The values of these coefficients, for various values of n, can be obtained directly by using Pascal's triangle.

Pascal's Triangle

n	*Binomial Coefficients*	*Sum of Coefficients* (2^n)
1	1 1	$2^1 = 2$
2	1 2 1	$2^2 = 4$
3	1 3 3 1	$2^3 = 8$
4	1 4 6 4 1	$2^4 = 16$
5	1 5 10 10 5 1	$2^5 = 32$

   We can note that it is very easy to write this triangle. In the first row, both the coefficients will be unity because ${}^1C_0 = {}^1C_1$. To write the second row, we write 1 in the beginning and the end and the value of the middle coefficients is obtained by adding the coefficients of the first row. Other rows of the Pascal's triangle can be written in a similar way.

4. (a) The shape and location of binomial distribution changes as the value of p changes for a given value of n. It can be shown that for a given value of n, if p is increased gradually in the interval (0, 0.5), the distribution changes from a positively skewed to a symmetrical shape. When p = 0.5, the distribution is perfectly symmetrical. Further, for larger values of p the distribution tends to become more and more negatively skewed.

   (b) For a given value of p, which is neither too small nor too large, the distribution becomes more and more symmetrical as n becomes larger and larger.

13.3.2 Uses of Binomial Distribution

Binomial distribution is often used in various decision making situations in business. Acceptance sampling plan, a technique of quality control, is based on this distribution. With the use of sampling plan, it is possible to accept or reject a lot of items either at the stage of its manufacture or at the stage of its purchase.

Notes

Decision Analysis

Decision making is needed whenever an individual or an organization (private or public) is faced with a situation of selecting an optimal (or best in view of certain objectives) course of action from among several available alternatives. For example, an individual may have to decide whether to build a house or to purchase a flat or live in a rented accommodation; whether to join a service or to start own business; which company's car should be purchased, etc. Similarly, a business firm may have to decide the type of technique to be used in production, what is the most appropriate method of advertising its product, etc.

The decision analysis provides certain criteria for the selection of a course of action such that the objective of the decision maker is satisfied. The course of action selected on the basis of such criteria is termed as the optimal course of action.

Notes

Self Assessment

Fill in the blanks:

11. The ..... to given data implies the determination of expected (or theoretical) frequencies for different values of the random variable on the basis of data.
12. The purpose of fitting a distribution is to examine whether the observed frequency distribution can be regarded as a ..... from a population with a known probability distribution.
13. To fit a binomial distribution to the given data, we find its .....
14. A distribution is known if the values of its ..... are known.

Multiple Choice Questions:

15. Binomial distribution is often used in various ..... situations in business.
    (a) Analytic (b) Decision tree
    (c) Decision free (d) Decision-making
16. Acceptance sampling plan, a technique of quality control, is based on .....
    (a) Sampling (b) Binomial distribution
    (c) Poisson distribution (d) Normal distribution
17. The values of different coefficients, for different values of n, can be obtained directly by using .....
    (a) Triangle (b) Hero triangle
    (c) Pascal Triangle (d) None

Case Study

Efficiency of Machine

From past experience it is known that a machine is set up correctly on 90% of occasions. If the machine is set up correctly then 95% of good parts are expected but if the machine is not set up correctly then the probability of a good part is only 30%.

On a particular day the machine is set up and the first component produced and found to be good. What is the probability that the machine is set up correctly?

13.4 Summary

- The study of a population can be done either by constructing an observed (or empirical) frequency distribution, often based on a sample from it, or by using a theoretical distribution.
- A theoretical probability distribution gives us a law according to which different values of the random variable are distributed with specified probabilities.
- It is possible to formulate such laws either on the basis of given conditions (a priori considerations) or on the basis of the results (a posteriori inferences) of an experiment.

Notes

- If a random variable satisfies the conditions of a theoretical probability distribution, then this distribution can be fitted to the observed data.
- The knowledge of the theoretical probability distribution is of great use in the understanding and analysis of a large number of business and economic situations.
- Binomial distribution is a theoretical probability distribution which was given by James Bernoulli.
- An experiment consists of a finite number of repeated trials.
- Each trial has only two possible, mutually exclusive, outcomes which are termed as a 'success' or a 'failure'.
- The probability of a success, denoted by p, is known and remains constant from trial to trial. The probability of a failure, denoted by q, is equal to 1 – p.
- Different trials are independent, i.e., outcome of any trial or sequence of trials has no effect on the outcome of the subsequent trials.
- The sequence of trials under the various above stated assumptions is also termed as Bernoulli Trials.
- The purpose of fitting a distribution is to examine whether the observed frequency distribution can be regarded as a sample from a population with a known probability distribution.
- To fit a binomial distribution to the given data, we find its mean.
- For a given value of p, which is neither too small nor too large, the distribution becomes more and more symmetrical as n becomes larger and larger.
- Binomial distribution is often used in various decision making situations in business.
- Acceptance sampling plan, a technique of quality control, is based on this distribution.

13.5 Keywords

Binomial distribution: Binomial distribution is a theoretical probability distribution which was given by James Bernoulli.

Experiment: An experiment consists of a finite number of repeated trials

Fitting of a binomial distribution: The fitting of a distribution to given data implies the determination of expected (or theoretical) frequencies for different values of the random variable on the basis of this data.

Posteriori inferences: These are the basis of results.

Priori considerations: These are the basis of given conditions.

Theoretical probability distribution: A theoretical probability distribution gives us a law according to which different values of the random variable are distributed with specified probabilities.

13.6 Review Questions

1. What do you understand by a theoretical probability distribution? How it is useful in business decision-making?

Notes

2. Define a binomial distribution. State the conditions under which binomial probability model is appropriate.
3. What are the parameters of a binomial distribution? Obtain expressions for mean and variance of the binomial variate in terms of these parameters.
4. Assume that the probability that a bomb dropped from an aeroplane will strike a target is 1/5. If six bombs are dropped, find the probability that (i) exactly two will strike the target, (ii) at least two will strike the target.
5. An unbiased coin is tossed 5 times. Find the probability of getting (i) two heads, (ii) at least two heads.
6. An experiment succeeds twice as many times as it fails. Find the probability that in 6 trials there will be (i) no successes, (ii) at least 5 successes, (iii) at the most 5 successes.
7. In an army battalion 60% of the soldiers are known to be married and remaining unmarried. If p(r) denotes the probability of getting r married soldiers from 5 soldiers, calculate p(0), p(1), p(2), p(3), p(4) and p(5). If there are 500 rows each consisting of 5 soldiers, approximately how many rows are expected to contain (i) all married soldiers, (ii) all unmarried soldiers?
8. A company has appointed 10 new secretaries out of which 7 are trained. If a particular executive is to get three secretaries, selected at random, what is the chance that at least one of them will be untrained?
9. The overall pass rate in a university examination is 70%. Four candidates take up such examination. What is the probability that (i) at least one of them will pass (ii) at the most 3 will pass (iii) all of them will pass, the examination?
10. 20% of bolts produced by a machine are defective. Deduce the probability distribution of the number of defectives in a sample of 5 bolts.
11. 25% employees of a firm are females. If 8 employees are chosen at random, find the probability that (i) 5 of them are males (ii) more than 4 are males (iii) less than 3 are females.
12. Suppose that the probability is that a car stolen in Delhi will be recovered. Find the probability that at least one out of 20 cars stolen in the city on a particular day will be recovered.
13. A sales man makes a sale on the average to 40 percent of the customer he contacts. If 4 customers are contacted today, what is the probability that he makes sales to exactly two? What assumption is required for your answer?
14. In a binomial distribution consisting of 5 independent trials, the probabilities of 1 and 2 successes are 0.4096 and 0.2048 respectively. Determine the distribution and write down the probability of at least three successes.
15. In a binomial distribution with 6 independent trials, the probabilities of 3 and 4 successes are found to be 0.1933 and 0.0644 respectively. Find the parameters 'p' and 'q' of the distribution.
16. The probability that a teacher will given an unannounced test during any class meeting is 1/5. If a student is absent twice. What is the probability that he will miss at least one test?
17. Components are placed into bins containing 100. After inspection of a large number of bins the average number of defective parts was found to be 10 with a standard deviation of 3.

Notes

Assuming that the same production conditions continue, except that bins containing 300 were used:

(a) What would be the average number of defective components per larger bin?

(b) What would be the standard deviation of the number of defectives per larger bin?

(c) How many components must each bin hold so that the standard deviation of the number of defective components is equal to 1% of the total number of components in the bin?

Answers: Self Assessment

1.	True	2.	False
3.	True	4.	False
5.	True	6.	True
7.	True	8.	True
9.	True	10.	False
11.	fitting of a distribution	12.	sample
13.	mean	14.	parameters
15.	(d)	16.	(b)
17.	(c)		

13.7 Further Readings

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Bhardwaj R S., *Business Statistics*, Excel Books.

Garrett H.E (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Selvaraj R, Loganathan, *C Quantitative Methods in Management.*

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Online links

http://en.wikipedia.org/wiki/Binomial_distribution

http://stattrek.com/lesson2/binomial.aspx

http://www.stat.yale.edu/Courses/1997-98/101/binom.htm

http://www.investopedia.com/terms/b/binomialdistribution.asp

http://www.mathsrevision.net/alevel/pages.php?page=72

Notes

Unit 14: Poisson Probability Distribution

CONTENTS

Objectives

Introduction

14.1 Poisson Distribution

14.1.1 Probability Mass Function

14.1.2 Summary Measures of Poisson Distribution

14.1.3 Poisson Approximation to Binomial

14.1.4 Fitting of a Poisson Distribution

14.2 Features and Uses of Poisson Distribution

14.3 Summary

14.4 Keywords

14.5 Review Questions

14.6 Further Readings

Objectives

After studying this unit, you will be able to:

- Define the term Poisson distribution
- Discuss Poisson process in brief
- Describe probability mass function
- Focus on various measures of poisson distribution
- Analyze Poisson approximation to binomial
- Tell the uses of Poisson Probability Distribution

Introduction

Poisson distribution was a limiting case of binomial distribution, when the number of trials n tends to become very large and the probability of success in a trial p tends to become very small such that their product np remains a constant

Did u know? Poisson distribution was derived by a noted mathematician, Simon D. Poisson, in 1837.

Notes

14.1 Poisson Distribution

This distribution is used as a model to describe the probability distribution of a random variable defined over a unit of time, length or space. For example, the number of telephone calls received per hour at a telephone exchange, the number of accidents in a city per week, the number of defects per meter of cloth, the number of insurance claims per year, the number breakdowns of machines at a factory per day, the number of arrivals of customers at a shop per hour, the number of typing errors per page etc.

Notes

Poisson Process

Let us assume that on an average 3 telephone calls are received per 10 minutes at a telephone exchange desk and we want to find the probability of receiving a telephone call in the next 10 minutes. In an effort to apply binomial distribution, we can divide the interval of 10 minutes into 10 intervals of 1 minute each so that the probability of receiving a telephone call (i.e., a success) in each minute (i.e., trial) becomes 3/10 (note that p = m/n, where m denotes mean). Thus, there are 10 trials which are independent, each with probability of success = 3/10. However, the main difficulty with this formulation is that, strictly speaking, these trials are not Bernoulli trials. One essential requirement of such trials, that each trial must result into one of the two possible outcomes, is violated here. In the above example, a trial, i.e. an interval of one minute, may result into 0, 1, 2, ... successes depending upon whether the exchange desk receives none, one, two, ... telephone calls respectively.

One possible way out is to divide the time interval of 10 minutes into a large number of small intervals so that the probability of receiving two or more telephone calls in an interval becomes almost zero. This is illustrated by the following table which shows that the probabilities of receiving two calls decreases sharply as the number of intervals are increased, keeping the average number of calls, 3 calls in 10 minutes in our example, as constant.

Using symbols, we may note that as n increases then p automatically declines in such a way that the mean m (= np) is always equal to a constant. Such a process is termed as a Poisson Process. The chief characteristics of Poisson process can be summarised as given below:

1. The number of occurrences in an interval is independent of the number of occurrences in another interval.
2. The expected number of occurrences in an interval is constant.
3. It is possible to identify a small interval so that the occurrence of more than one event, in any interval of this size, becomes extremely unlikely.

14.1.1 Probability Mass Function

The probability mass function (p.m.f.) of Poisson distribution can be derived as a limit of p.m.f. of binomial distribution when $n \to \infty$ such that m (= np) remains constant. Thus, we can write

$$P(r) = \lim_{n \to \infty} {}^nC_r \left(\frac{m}{n}\right)^r \left(1 - \frac{m}{n}\right)^{n-r} = \lim_{n \to \infty} \frac{n!}{r!(n-r)!} \left(\frac{m}{n}\right)^r \left(1 - \frac{m}{n}\right)^{n-r}$$

Notes

$$= \frac{m^r}{r!} \cdot \lim_{n \to \infty} \left[n(n-1)(n-2) \ .... \ (n-r+1) . \frac{1}{n^r} . \left(1 - \frac{m}{n}\right)^{n-r} \right]$$

$$= \frac{m^r}{r!} \cdot \lim_{n \to \infty} \left[\frac{\frac{n}{n}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right) \ .... \ \left(1-\frac{(r-1)}{n}\right)\left(1-\frac{m}{n}\right)^n}{\left(1-\frac{m}{n}\right)^r} \right]$$

$$= \frac{m^r}{r!} \lim_{n \to \infty} \left(1 - \frac{m}{n}\right)^n$$ since each of the remaining terms will tend to unity as $n \to \infty$

$$\frac{m^r . e^{-m}}{r!} \text{since} \lim_{n \to \infty} \left(1 - \frac{m}{n}\right)^n = \lim_{n \to \infty} \left\{ \left(1 - \frac{m}{n}\right)^{\frac{n}{m}} \right\}^m = e^{-m}$$

Thus, the probability mass function of Poisson distribution is

$$P(r) = \frac{e^{-m} . m^r}{r!}, \text{ where } r = 0, 1, 2, \ ..... \infty$$

Here e is a constant with value = 2.71828... . Note that Poisson distribution is a discrete probability distribution with single parameter m.

Total probability $= \sum_{r=0}^{\infty} \frac{e^{-m} . m^r}{r!} = e^{-m} \left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \frac{m^3}{3!} + \ ....\right)$

$$= e^{-m} . e^m = 1$$

14.1.2 Summary Measures of Poisson Distribution

1. ***Mean:*** The mean of a Poisson variate r is defined as

$$E(r) = \sum_{r=0}^{\infty} r . \frac{e^{-m} . m^r}{r!} = e^{-m} \sum_{r=1}^{\infty} \frac{m^r}{(r-1)!} = e^{-m} \left[m + m^2 + \frac{m^3}{2!} + \frac{m^4}{3!} + \ ....\right]$$

$$= me^{-m} \left[1 + m \ + \frac{m^2}{2!} + \frac{m^3}{3!} + \ ....\right] = me^{-m} e^m = m$$

2. ***Variance:*** The variance of a Poisson variate is defined as

   $Var(r) = E(r - m)^2 = E(r^2) - m^2$

   Now $E(r^2) = \sum_{r=0}^{\infty} r^2 P(r) = \sum_{r=0}^{\infty} [r(r-1) + r] P(r)$

Notes

$$= \sum_{r=2}^{\infty}\left[r(r-1)\right]\frac{e^{-m}.m^r}{r!} + m = e^{-m}\sum_{r=2}^{\infty}\frac{m^r}{(r-2)!} + m$$

$$= m + e^{-m}\left(m^2 + m^3 + \frac{m^4}{2!} + \frac{m^5}{3!} + \ ....\right)$$

$$= m + m^2 e^{-m}\left(1 + m \ + \frac{m^2}{2!} + \frac{m^3}{3!} + \ ....\right) = m + m^2$$

Thus, Var(r) = m + m^2 - m^2 = m.

Also standard deviation $\sigma = \sqrt{m}$

3. ***The values of*** β_1***:*** It can be shown that μ_3 = m and μ_4 = m + $3m^2$.

$$\therefore \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}$$

Since m is a positive quantity, therefore, β_1 is always positive and hence the Poisson distribution is always positively skewed. We note that $\to 0$ as m $\to \infty$, therefore the distribution tends to become more and more symmetrical for large values of m.

Further, This result shows that the distribution becomes normal for large values of m.

4. ***Mode:*** As in binomial distribution, a Poisson variate r will be mode if

$$P(r-1) \le P(r) \ge P(r+1)$$

The inequality $P(r-1) \le P(r)$ can be written as

$$\frac{e^{-m}.m^{r-1}}{(r-1)!} \le \frac{e^{-m}.m^r}{r!} \quad \Rightarrow \quad 1 \le \frac{m}{r} \quad \Rightarrow \quad r \le m \qquad .... (1)$$

Similarly, the inequality $P(r) \ge P(r+1)$ can be shown to imply that

$$r \ge m - 1 \qquad .... (2)$$

Combining (1) and (2), we can write $m - 1 \le r \le m$.

Case I. When m is not an integer

The integral part of m will be mode.

Case II. When m is an integer

The distribution is bimodal with values m and m – 1.

Example: The average number of customer arrivals per minute at a super bazaar is 2. Find the probability that during one particular minute (1) exactly 3 customers will arrive, (2) at the most two customers will arrive, (3) at least one customer will arrive.

Notes

Solution:

It is given that m = 2. Let the number of arrivals per minute be denoted by the random variable r. The required probability is given by

1. $P(r=3)=\frac{e^{-2}.2^3}{3!}=\frac{0.13534\times 8}{6}=0.18045$

2. $P(r\le 2)=\sum_{r=0}^{2}\frac{e^{-2}.2^r}{r!}=e^{-2}\left[1+2+\frac{4}{2}\right]=0.13534\times 5=0.6767.$

3. $P(r\ge 1)=1-P(r=0)=1-\frac{e^{-2}.2^0}{0!}=1-0.13534=0.86464.$

Example: An executive makes, on an average, 5 telephone calls per hour at a cost which may be taken as ₹ 2 per call. Determine the probability that in any hour the telephone calls' cost (i) exceeds ₹ 6, (ii) remains less than ₹ 10.

Solution:

The number of telephone calls per hour is a random variable with mean = 5. The required probability is given by

1. $P(r>3)=1-P(r\le 3)=1-\sum_{r=0}^{3}\frac{e^{-5}.5^r}{r!}$

$$=1-e^{-5}\left[1+5+\frac{25}{2}+\frac{125}{6}\right]=1-0.00678\times\frac{236}{6}=0.7349.$$

2. $P(r\le 4)=\sum_{r=0}^{4}\frac{e^{-5}.5^r}{r!}=e^{-5}\left[1+5+\frac{25}{2}+\frac{125}{6}+\frac{625}{24}\right]=0.00678\times\frac{1569}{24}=0.44324.$

Example: A company makes electric toys. The probability that an electric toy is defective is 0.01. What is the probability that a shipment of 300 toys will contain exactly 5 defectives?

Solution:

Since n is large and p is small, Poisson distribution is applicable. The random variable is the number of defective toys with mean m = np = 300 × 0.01 = 3. The required probability is given by

$$P(r=5)=\frac{e^{-3}.3^5}{5!}=\frac{0.04979\times 243}{120}=0.10082.$$

Example: In a town, on an average 10 accidents occur in a span of 50 days. Assuming that the number of accidents per day follow Poisson distribution, find the probability that there will be three or more accidents in a day.

Notes

Solution:

The random variable denotes the number accidents per day. Thus, we have $m = \frac{10}{50} = 0.2$. The required probability is given by

$$P(r \geq 3) = 1 - P(r \leq 2) = 1 - e^{-0.2}\left[1 + 0.2 + \frac{(0.2)^2}{2!}\right] = 1 - 0.8187 \times 1.22 = 0.00119.$$

Example: A car hire firm has two cars which it hire out every day. The number of demands for a car on each day is distributed as a Poisson variate with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused. [$e^{-1.5} = 0.2231$]

Solution:

When both car are not used, r = 0

$\therefore\ P(r = 0) = e^{-1.5} = 0.2231$. Hence the proportion of days on which neither car is used is 22.31%.

Further, some demand is refused when more than 2 cars are demanded, i.e., r > 2

$$\therefore\ P(r > 2) = 1 - P(r \leq 2) = 1 - \sum_{r=0}^{2} \frac{e^{-1.5}(1.5)^r}{r!} = 1 - 0.2231\left[1 + 1.5 + \frac{(1.5)^2}{2!}\right] = 0.1913.$$

Hence the proportion of days is 19.13%.

Example: A firm produces articles of which 0.1 percent are usually defective. It packs them in cases each containing 500 articles. If a wholesaler purchases 100 such cases, how many cases are expected to be free of defective items and how many are expected to contain one defective item?

Solution:

The Poisson variate is number of defective items with mean

$$m = \frac{1}{1000} \times 500 = 0.5.$$

Probability that a case is free of defective items

$P(r = 0) = e^{-0.5} = 0.6065$. Hence the number of cases having no defective items = 0.6065 × 100 = 60.65

Similarly, $P(r = 1) = e^{-0.5} \times 0.5 = 0.6065 \times 0.5 = 0.3033$. Hence the number of cases having one defective item are 30.33.

Example: A manager accepts the work submitted by his typist only when there is no mistake in the work. The typist has to type on an average 20 letters per day of about 200 words each. Find the chance of her making a mistake (1) if less than 1% of the letters submitted by her are rejected; (2) if on 90% of days all the work submitted by her is accepted. [As the probability of making a mistake is small, you may use Poisson distribution. Take e = 2.72].

Notes

Solution:

Let p be the probability of making a mistake in typing a word.

1. Let the random variable r denote the number of mistakes per letter. Since 20 letters are typed, r will follow Poisson distribution with mean = 20 × p.

   Since less than 1% of the letters are rejected, it implies that the probability of making at least one mistake is less than 0.01, i.e.,

   $P(r \geq 1) \leq 0.01$ or $1 - P(r = 0) \leq 0.01$

   $\Rightarrow \quad 1 - e^{-20p} \leq 0.01$ or $e^{-20p} \geq 0.99$

   Taking log of both sides

   $-20p.\log 2.72 \geq \log 0.99$

   $-(20 \times 0.4346)p \geq \bar{1}.9956$

   $-8.692p \geq -0.0044$ or $p \leq \dfrac{0.0044}{8.692} = 0.00051.$

2. In this case r is a Poisson variate which denotes the number of mistakes per day. Since the typist has to type 20 × 200 = 4000 words per day, the mean number of mistakes = 4000 p.

   It is given that there is no mistake on 90% of the days, i.e.,

   $P(r = 0) = 0.90$ or $e^{-4000p} = 0.90$

   Taking log of both sides, we have

   $-4000p \log 2.72 = \log 0.90$ or $-4000 \times 0.4346p = \bar{1}.9542 = -0.0458$

   $\therefore \quad p = \dfrac{0.0458}{4000 \times 0.4346} = 0.000026.$

Lot Acceptance using Poisson Distribution

Example: Videocon company purchases heaters from Amar Electronics. Recently a shipment of 1000 heaters arrived out of which 60 were tested. The shipment will be accepted if not more than two heaters are defective. What is the probability that the shipment will be accepted? From past experience, it is known that 5% of the heaters made by Amar Electronics are defective.

Solution:

Mean number of defective items in a sample of 60 = 60 × $\dfrac{5}{100}$ = 3

$$P(r \leq 2) = \sum_{r=0}^{2} \frac{e^{-3}.3^r}{r!}$$

$$= e^{-3}\left[1 + 3 + \frac{3^2}{2!}\right] = e^{-3}.8.5 = 0.4232$$

Notes

14.1.3 Poisson Approximation to Binomial

When n, the number of trials become large, the computation of probabilities by using the binomial probability mass function becomes a cumbersome task. Usually, when $n \geq 20$ and $p \leq 0.05$, Poisson distribution can be used as an approximation to binomial with parameter m = np.

Example: Find the probability of 4 successes in 30 trials by using (1) binomial distribution and (2) Poisson distribution. The probability of success in each trial is given to be 0.02.

Solution:

1. Here n = 30 and p = 0.02

$$\therefore\ P(r=4) = {}^{30}C_4 (0.02)^4 (0.98)^{26} = 27405 \times 0.00000016 \times 0.59 = 0.00259.$$

2. Here m = np = 30 × 0.02 = 0.6

$$\therefore\ P(r=4) = \frac{e^{-0.6}(0.6)^4}{4!} = \frac{0.5488 \times 0.1296}{24} = 0.00296.$$

14.1.4 Fitting of a Poisson Distribution

To fit a Poisson distribution to a given frequency distribution, we first compute its mean m. Then the probabilities of various values of the random variable r are computed by using the probability mass function $P(r) = \frac{e^{-m}.m^r}{r!}$. These probabilities are then multiplied by N, the total frequency, to get expected frequencies.

Example: The following mistakes per page were observed in a book:

No. of mistakes per page	:	0	1	2	3
Frequency	:	211	90	19	5

Fit a Poisson distribution to find the theoretical frequencies.

Solution:

The mean of the given frequency distribution is

$$m = \frac{0 \times 211 + 1 \times 90 + 2 \times 19 + 3 \times 5}{211 + 90 + 19 + 5} = \frac{143}{325} = 0.44$$

Calculation of theoretical (or expected) frequencies

We can write $P(r) = \frac{e^{-0.44}(0.44)^r}{r!}$. Substituting r = 0, 1, 2 and 3, we get the probabilities for various values of r, as shown in the following table.

Notes

r	$P(r)$	$N \times P(r)$	Expected Frequencies Approximated to the nearest integer
0	0.6440	209.30	210
1	0.2834	92.10	92
2	0.0623	20.25	20
3	0.0091	2.96	3
Total			325

Did u know? Poisson distribution serves as a reasonably good approximation to binomial distribution when $n \geq 20$ and $p \geq 0.05$.

Task A manufacturer of pins knows that on an average 5% of his product is defective. He sells pins in boxes of 100 and guarantees that not more than 4 pins will be defective. What is the probability that the box will meet the guaranteed quality?

Self Assessment

State whether the following statements are true or false:

1. Poisson distribution was derived by a noted mathematician, Simon D. Poisson, in 1857.
2. Poisson distribution is a limiting case of binomial distribution, when the number of trials n tends to become very large and the probability of success in a trial p tends to become very small such that their product np remains a constant.
3. Poisson distribution is used as a model to describe the probability distribution of a random variable defined over a unit of time, length or space.
4. The number of telephone calls received per hour at a telephone exchange, the number of accidents in a city per week, the number of defects per meter of cloth, the number of insurance claims per year, the number breakdowns of machines at a factory per day, the number of arrivals of customers at a shop per hour, the number of typing errors per page, etc., all are examples of poisson distribution.
5. As n increases then p automatically increases in such a way that the mean m (= np) is always equal to a constant .
6. The number of occurrences in an interval is dependent of the number of occurrences in another interval.
7. The expected number of occurrences in an interval is constant.
8. It is possible to identify a small interval so that the occurrence of more than one event, in any interval of this size, becomes extremely unlikely.
9. The probability mass function (p.m.f.) of Poisson distribution can be derived as a limit of p.m.f. of binomial distribution when $n \to \infty$ such that m (= np) remains constant.

Notes

14.2 Features and Uses of Poisson Distribution

Features

1. It is discrete probability distribution.
2. It has only one parameter m.
3. The range of the random variable is $0 \le r < \infty$.
4. The Poisson distribution is a positively skewed distribution. The skewness decreases as m increases.

Uses

This distribution is applicable to situations where the number of trials is large and the probability of a success in a trial is very small.

Caution When n, the number of trials become large, the computation of probabilities by using the binomial probability mass function becomes a non cumbersome task.

Self Assessment

Fill in the blanks:

11. Poisson distribution is..... probability distribution.
12. Poisson distribution has .....
13. In Poisson distribution, range of the random variable is .....
14. The Poisson distribution is a .....distribution.
15. Poisson distribution is applicable to situations where the number of trials is ..... and the probability of a success in a trial is .....

Case Study

Poisson Distribution

Customers arrive randomly at a service point at an average rate of 30 per hour. Assuming a Poisson distribution calculate the probability that:

1. no customer arrives in any particular minute.
2. exactly one customer arrives in any particular minute.
3. two or more customers arrive in any particular minute.
4. three or fewer customers arrive in any particular minute.

Notes

14.3 Summary

- This Poisson distribution was derived by a noted mathematician, Simon D. Poisson, in 1837.
- This distribution was derived as a limiting case of binomial distribution,
- When the number of trials n tends to become very large and the probability of success in a trial p tends to become very small such that their product np remains a constant.
- This distribution is used as a model to describe the probability distribution of a random variable defined over a unit of time, length or space.
- The number of telephone calls received per hour at a telephone exchange, the number of accidents in a city per week, the number of defects per meter of cloth, the number of insurance claims per year, the number breakdowns of machines at a factory per day, the number of arrivals of customers at a shop per hour, the number of typing errors per page etc. all are examples of poisson distribution
- To fit a Poisson distribution to a given frequency distribution, we first compute its mean m.
- The range of the random variable is $0 \leq r < \infty$.
- The Poisson distribution is a positively skewed distribution. The skewness decreases as m increases.
- This distribution is applicable to situations where the number of trials is large and the probability of a success in a trial is very small.
- It serves as a reasonably good approximation to binomial distribution when $n \geq 20$ and $p \leq 0.05$.

14.4 Keywords

Poisson Approximation to Binomial: Poisson distribution can be used as an approximation to binomial with parameter m = np.

Poisson Distribution: This is a a limiting case of binomial distribution, when the number of trials n tends to become very large and the probability of success in a trial p tends to become very small such that their product np remains a constant. This distribution is used as a model to describe the probability distribution of a random variable defined over a unit of time, length or space.

Poisson Process: Using symbols, we may note that as n increases then p automatically declines in such a way that the mean m (= np) is always equal to a constant. Such a process is termed as a Poisson Process.

Probability Mass Function: The probability mass function (p.m.f.) of Poisson distribution can be derived as a limit of p.m.f. of binomial distribution when $n \to \infty$ such that m (= np) remains constant.

Notes

14.5 Review Questions

1. What is a 'Poisson Process'? Obtain probability mass function of Poisson variate as a limiting form of the probability mass function of binomial variate.
2. Obtain mean and standard deviation of a Poisson random variate. Discuss some business and economic situations where Poisson probability model is appropriate.
3. How will you use Poisson distribution as an approximation to binomial? Explain with the help of an example.
4. State clearly the assumptions under which a binomial distribution tends to Poisson distribution.
5. A manufacturer, who produces medicine bottles, finds that 0.1% of the bottles are defective. The bottles are packed in boxes containing 500 bottles. A drug manufacturer buys 100 boxes from the producer of the bottles. Use Poisson distribution to find the number of boxes containing (i) no defective bottles (ii) at least two defective bottles.
6. A factory turning out lenses, supplies them in packets of 1,000. The packet is considered by the purchaser to be unacceptable if it contains 50 or more defective lenses. If a purchaser selects 30 lenses at random from a packet and adopts the criterion of rejecting the packet if it contains 3 or more defectives, what is the probability that the packet (i) will be accepted, (ii) will not be accepted?
7. 800 employees of a company are covered under the medical group insurance scheme. Under the terms of coverage, 40 employees are identified as belonging to 'High Risk' category. If 50 employees are selected at random, what is the probability that (i) none of them is in the high risk category, (ii) at the most two are in the high risk category? (You may use Poisson approximation to Binomial).
8. The following table gives the number of days in 50 day-period during which automobile accidents occurred in a certain part of the city. Fit a Poisson distribution to the data.

No. of accidents	:	0	1	2	3	4
No. of days	:	19	18	8	4	1

9. Comment on the following statements:
   (a) The mean of a Poisson variate is 4 and standard deviation is $\sqrt{3}$.
   (b) The second raw moment of a Poisson distribution is 2. The probability $P(X = 0) = e^{-1}$.
   (c) If for a Poisson variate X, $P(X = 1) = P(X = 2)$, then $E(X) = 2$.
   (d) If for a Poisson variate X, $P(X = 0) = P(X = 1)$, then $P(X > 0) = e^{-1}$.
10. A firm buys springs in very large quantities and from past records it is known that 0.2% are defective. The inspection department sample the springs in batches of 500. It is required to set a standard for the inspectors so that if more than the standard number of defectives is found in a batch the consignment can be rejected with at least 90% confidence that the supply is truly defective.

    How many defectives per batch should be set as the standard?

Notes

Answers: Self Assessment

1.	False	2.	True
3.	True	4.	True
5.	False	6.	False
7.	True	8.	True
9.	True	10.	False
11.	discrete	12.	only one parameter m
13.	$0 \le r < \infty$	14.	positively skewed
15.	large, very small		

14.6 Further Readings

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Bhardwaj R.S., *Business Statistics*, Excel Books.

Garrett H.E. (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Guilford J.P. (1965), *Fundamental Statistics in Psychology and Education*, Mc Graw Hill Book Company, New York.

Gupta S.P., *Statistical Method*, Sultan Chand and Sons, New Delhi, 2008.

Hannagan T.J. (1982), *Mastering Statistics*, The Macmillan Press Ltd., Surrey.

Hooda R. P., *Statistics for Business and Economics*, Macmillan India, Delhi, 2008.

Jaeger R.M (1983), *Statistics: A Spectator Sport*, Sage Publications India Pvt. Ltd., New Delhi.

Lindgren B.W. (1975), *Basic Ideas of Statistics*, Macmillan Publishing Co. Inc., New York.

Richard I. Levin, *Statistics for Management*, Pearson Education Asia, New Delhi, 2002.

Selvaraj R., Loganathan, C. *Quantitative Methods in Management.*

Sharma J.K., *Business Statistics*, Pearson Education Asia, New Delhi, 2009.

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Walker H.M. and J. Lev, (1965), *Elementary Statistical Methods*, Oxford & IBH Publishing Co., Calcutta.

Wine R.L. (1976), *Beginning Statistics*, Winthrop Publishers Inc., Massachusetts.

Online links

http://en.wikipedia.org/wiki/Poisson_distribution

http://hyperphysics.phy-astr.gsu.edu/hbase/math/poifcn.html

Notes

http://itl.nist.gov/div898/handbook/eda/section3/eda366j.htm

http://www.childrensmercy.org/stats/definitions/poisson.htm

http://en.wikipedia.org/wiki/Poisson_distribution

http://www.intmath.com/counting-probability/13-poisson-probability-distribution.php

Notes

Unit 15: Normal Probability Distribution

CONTENTS

Objectives

Introduction

15.1 The Conditions of Normality

15.2 Probability Density Function

15.3 Properties of Normal Probability Curve

15.4 Probability of Normal Variate in an Interval

15.5 Normal Approximation to Binominal Distribution

15.6 Normal Approximation to Poisson Distribution

15.7 Summary

15.8 Keywords

15.9 Review Questions

15.10 Further Readings

Objectives

After studying this unit, you will be able to:

- Tell about normal probability distribution
- Discuss various conditions of normality
- State the relevance of Probability Density Function
- Explain shape and properties of normal distribution curve
- Focus on Fitting a normal curve

Introduction

The normal probability distribution occupies a place of central importance in Modern Statistical Theory. This distribution was first observed as the normal law of errors by the statisticians of the eighteenth century. They found that each observation X involves an error term which is affected by a large number of small but independent chance factors. This implies that an observed value of X is the sum of its true value and the net effect of a large number of independent errors which may be positive or negative each with equal probability. The observed distribution of such a random variable was found to be in close conformity with a continuous curve, which was termed as the normal curve of errors or simply the normal curve.

Did u know? Since Gauss used this curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies, it is also called as Gaussian curve.

Notes

15.1 The Conditions of Normality

In order that the distribution of a random variable X is normal, the factors affecting its observations must satisfy the following conditions:

1. ***A large number of chance factors:*** The factors, affecting the observations of a random variable, should be numerous and equally probable so that the occurrence or non-occurrence of any one of them is not predictable.
2. ***Condition of homogeneity:*** The factors must be similar over the relevant population although, their incidence may vary from observation to observation.
3. ***Condition of independence:*** The factors, affecting observations, must act independently of each other.
4. ***Condition of symmetry:*** Various factors operate in such a way that the deviations of observations above and below mean are balanced with regard to their magnitude as well as their number.

15.2 Probability Density Function

If X is a continuous random variable, distributed normally with mean *m* and standard deviation *s*, then its p.d.f. is given by $p(X) = \frac{1}{\sigma\sqrt{2\pi}}.e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$ Where $-\infty < X < \infty$.

Here *p* and e are absolute constants with values 3.14159.... and 2.71828.... respectively.

It may be noted here that this distribution is completely known if the values of mean *m* and standard deviation σ are known. Thus, the distribution has two parameters, viz. mean and standard deviation.

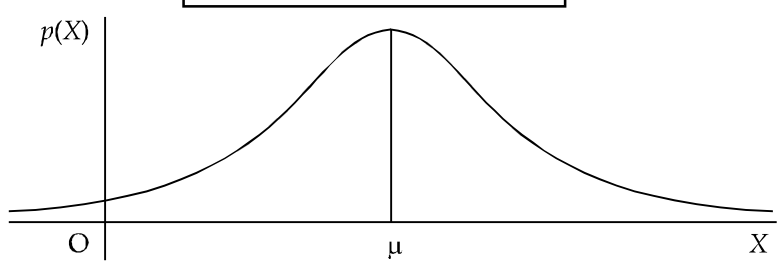
Notes

Shape of Normal Probability Curve

For given values of the parameters, *m* and *s*, the shape of the curve corresponding to normal probability density function p(X) is as shown in Figure below.

It should be noted here that although we seldom encounter variables that have a range from - ∞ to ∞, as shown by the normal curve, nevertheless the curves generated by the relative frequency histograms of various variables closely resembles the shape of normal curve.

Normal Probability Curve



Notes

15.3 Properties of Normal Probability Curve

A normal probability curve or normal curve has the following properties:

1. It is a bell shaped symmetrical curve about the ordinate at X = μ. The ordinate is maximum at X = μ.
2. It is unimodal curve and its tails extend infinitely in both directions, i.e., the curve is asymptotic to X axis in both directions.
3. All the three measures of central tendency coincide, i.e.,

   mean = median = mode

4. The total area under the curve gives the total probability of the random variable taking values between $-\infty$ to ∞. Mathematically, it can be shown that

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} p(X)dX = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} dX = 1.$$

5. Since median = μ, the ordinate at X = μ divides the area under the normal curve into two equal parts, i.e.,

$$\int_{-\infty}^{\mu} p(X)dX = \int_{\mu}^{\infty} p(X)dX = 0.5$$

6. The value of p(X) is always non-negative for all values of X, i.e., the whole curve lies above X axis.
7. The points of inflexion (the point at which curvature changes) of the curve are at X = $\mu \pm \sigma$.
8. The quartiles are equidistant from median, i.e., $M_d - Q_1 = Q_3 - M_d$, by virtue of symmetry. Also $Q_1 = \mu - 0.6745\sigma$, $Q_3 = \mu + 0.6745\sigma$, quartile deviation = 0.6745σ and mean deviation = 0.8*s*, approximately.
9. Since the distribution is symmetrical, all odd ordered central moments are zero.
10. The successive even ordered central moments are related according to the following recurrence formula

    $\mu_{2n} = (2n - 1)\sigma^2 \mu_{2n-2}$ for n = 1, 2, 3, ...

11. The value of moment coefficient of skewness β_1 is zero.
12. The coefficient of kurtosis $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3.$

    Note that the above expression makes use of property 10.

13. Additive or reproductive property

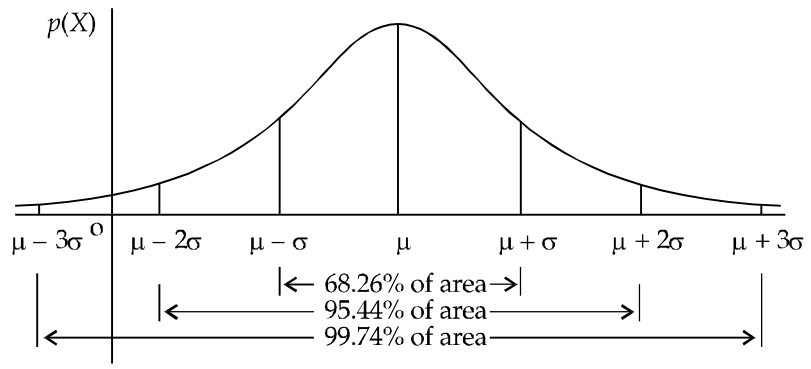
    If $X_1, X_2, \ldots\ldots X_n$ are n independent normal variates with $\mu_1, \mu_2, \ldots\ldots \mu_n$ means and variances $\sigma_1^2, \sigma_2^2, \ldots\ldots \sigma_n^2$, respectively, then their linear combination $a_1X_1 + a_2X_2 + \ldots\ldots + a_nX_n$ is also a normal variate with mean $\sum_{i=1}^{n} a_i\mu_i$ and variance $\sum_{i=1}^{n} a_i^2\sigma_i^2$.

Notes

In particular, if $a_1 = a_2 = ... = a_n = 1$, we have $\sum X_i$ is a normal variate with mean $\sum \mu_i$ and variance $\sum \sigma_i^2$. Thus the sum of independent normal variates is also a normal variate.

14. ***Area property:*** The area under the normal curve is distributed by its standard deviation in the following manner:

Figure 15.1: Area under the Normal Curve



(a) The area between the ordinates at $\mu - \sigma$ and $\mu + \sigma$ is 0.6826. This implies that for a normal distribution about 68% of the observations will lie between $\mu - \sigma$ and $\mu + \sigma$.

(b) The area between the ordinates at $\mu - 2\sigma$ and $\mu + 2\sigma$ is 0.9544. This implies that for a normal distribution about 95% of the observations will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.

(c) The area between the ordinates at $\mu - 3\sigma$ and $\mu + 3\sigma$ is 0.9974. This implies that for a normal distribution about 99% of the observations will lie between $\mu - 3\sigma$ $m - 3s$ and $\mu + 3\sigma$. This result shows that, practically, the range of the distribution is $6s$ although, theoretically, the range is from $-\infty$ to ∞.

Did u know? All the three measures of central tendency coincide, i.e., mean = median = mode

Notes

Kurtosis

This is another measure of the shape of a frequency curve. While skewness refers to the extent of lack of symmetry, kurtosis refers to the extent to which a frequency curve is peaked. Kurtosis is a Greek word which means bulginess. In statistics, the word is used for a measure of the degree of peakedness of a frequency curve.

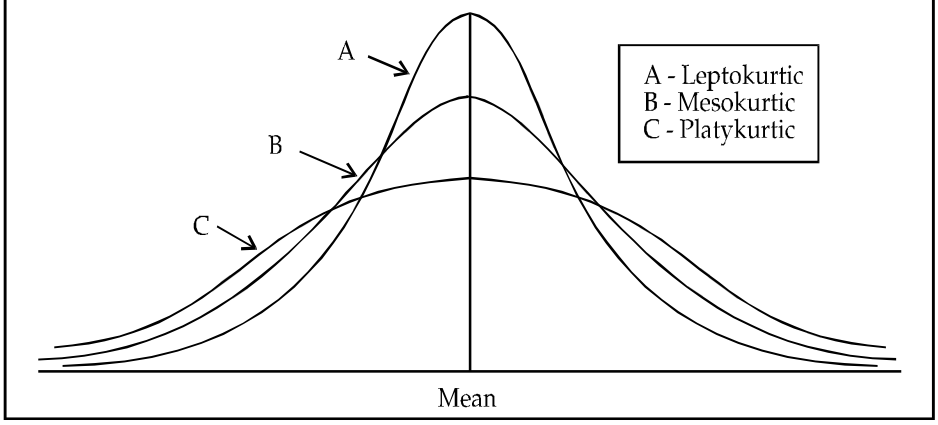
Karl Pearson, in 1905, introduced three types of curves depending upon the shape of their peaks. These three shapes are known as Mesokurtic, Leptokurtic and Platykurtic. A mesokurtic shaped curve is neither too much peaked nor too much flattened. This in fact is the frequency curve of a normal distribution. A curve that is more peaked than a normal curve is known as leptokurtic while a relatively flat topped curve is known as platykurtic.

Contd...

Notes

The three types of curves are shown in the following figure.

Mesokurtic, Leptokurtic and Platykurtic Curves



15.4 Probability of Normal Variate in an Interval

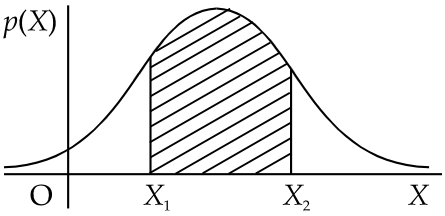
Let X be a normal variate distributed with mean μ and standard deviation σ, also written in abbreviated form as $X \sim N(\mu, \sigma)$ The probability of X lying in the interval (X_1, X_2) is given by

$$P(X_1 \le X \le X_2) = \int_{X_1}^{X_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} dX$$

In terms of figure, this probability is equal to the area under the normal curve between the ordinates at $X = X_1$ and $X = X_2$ respectively.

Note: It may be recalled that the probability that a continuous random variable takes a particular value is defined to be zero even though the event is not impossible.

Figure 15.2



It is obvious from the above that, to find $P(X_1 \le X \le X_2)$, we have to evaluate an integral which might be cumbersome and time consuming task. Fortunately, an alternative procedure is available for performing this task. To devise this procedure, we define a new variable $z = \frac{X-\mu}{\sigma}$.

We note that $E(z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\left[E(X) - \mu\right] = 0$

and $Var(z) = Var\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(X-\mu) = \frac{1}{\sigma^2} Var(X) = 1.$

Further, from the reproductive property, it follows that the distribution of z is also normal.

Notes

Thus, we conclude that if X is a normal variate with mean μ and standard deviation σ, then $z = \frac{X-\mu}{\sigma}$ is a normal variate with mean zero and standard deviation unity. Since the parameters of the distribution of z are fixed, it is a known distribution and is termed as standard normal distribution (s.n.d.). Further, z is termed as a standard normal variate (s.n.v.).

It is obvious from the above that the distribution of any normal variate X can always be transformed into the distribution of standard normal variate z. This fact can be utilised to evaluate the integral given above.

We can write $P(X_1 \le X \le X_2) = P\left[\left(\frac{X_1-\mu}{\sigma}\right) \le \left(\frac{X-\mu}{\sigma}\right) \le \left(\frac{X_2-\mu}{\sigma}\right)\right]$

$= P(z_1 \le z \le z_2)$, where $z_1 = \frac{X_1-\mu}{\sigma}$ and $z_2 = \frac{X_2-\mu}{\sigma}$.

In terms of figure, this probability is equal to the area under the standard normal curve between the ordinates at $z = z_1$ and $z = z_2$. Since the distribution of z is fixed, the probabilities of z lying in various intervals are tabulated. These tables can be used to write down the desired probability.

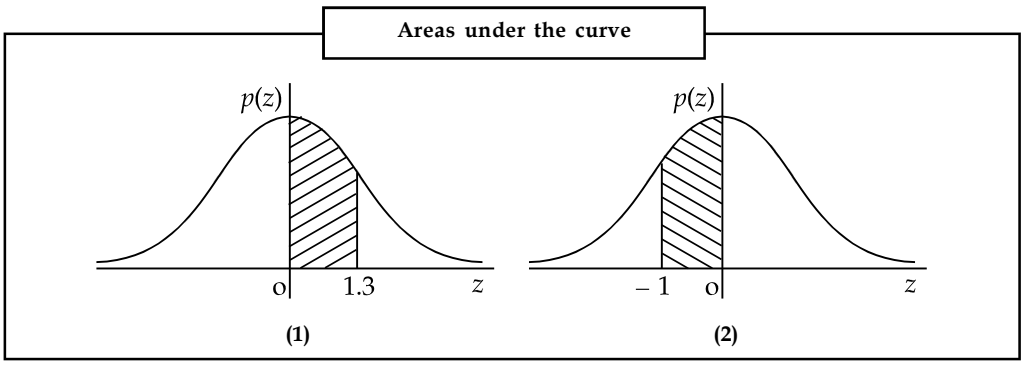
Example: Using the table of areas under the standard normal curve, find the following probabilities :

1. $P(0 \le z \le 1.3)$
2. $P(-1 \le z \le 0)$
3. $P(-1 \le z \le 2)$
4. $P(z \ge 1.54)$
5. $P(|z| > 2)$
6. $P(|z| < 2)$

Solution:

The required probability, in each question, is indicated by the shaded are of the corresponding figure.

1. From the table, we can write $P(0 \le z \le 1.3) = 0.4032$.



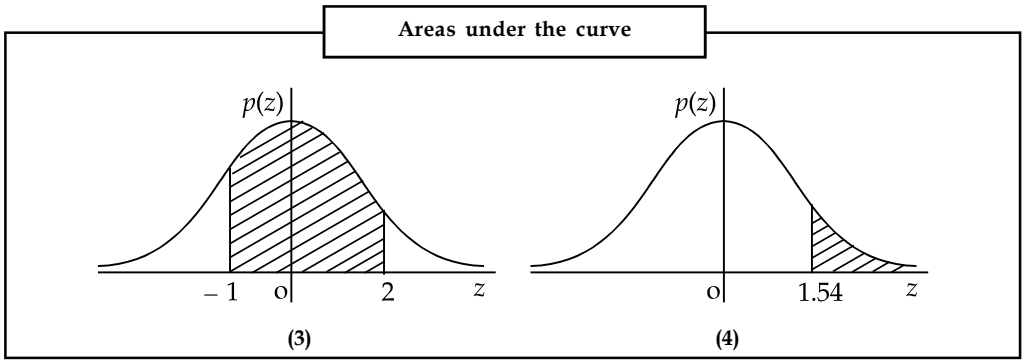
2. We can write $P(-1 \le z \le 0) = P(0 \le z \le 1)$, because the distribution is symmetrical. From the table, we can write $P(-1 \le z \le 0) = P(0 \le z \le 1) = 0.3413$.

3. We can write

   $P(-1 \le z \le 2) = P(-1 \le z \le 0) + P(0 \le z \le 2)$

   $= P(0 \le z \le 1) + P(0 \le z \le 2) = 0.3413 + 0.4772 = 0.8185$.

Notes

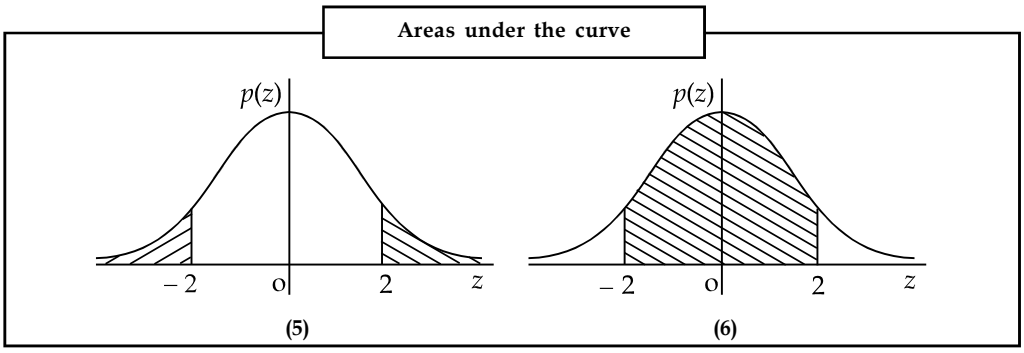


4. We can write

   $P(z \geq 1.54) = 0.5000 - P(0 \leq z \leq 1.54) = 0.5000 - 0.4382 = 0.0618.$

5. $P(|z| > 2) = P(z > 2) + P(z < -2) = 2P(z > 2) = 2[0.5000 - P(0 \leq z \leq 2)]$

   $= 1 - 2P(0 \leq z \leq 2) = 1 - 2 \times 0.4772 = 0.0456.$



6. $P(|z| < 2) = P(-2 \leq z \leq 0) + P(0 \leq z \leq 2) = 2P(0 \leq z \leq 2) = 2 \times 0.4772 = 0.9544.$

Example: Determine the value or values of z in each of the following situations:

1. Area between 0 and z is 0.4495.
2. Area between $-\infty$ to z is 0.1401.
3. Area between $-\infty$ to z is 0.6103.
4. Area between – 1.65 and z is 0.0173.
5. Area between – 0.5 and z is 0.5376.

Solution:

1. On locating the value of z corresponding to an entry of area 0.4495 in the table of areas under the normal curve, we have z = 1.64. We note that the same situation may correspond to a negative value of z. Thus, z can be 1.64 or – 1.64.
2. Since the area between $-\infty$ to z < 0.5, z will be negative. Further, the area between z and 0 = 0.5000 – 0.1401 = 0.3599. On locating the value of z corresponding to this entry in the table, we get z = – 1.08.
3. Since the area between $-\infty$ to z > 0.5000, z will be positive. Further, the area between 0 to z = 0.6103 – 0.5000 = 0.1103. On locating the value of z corresponding to this entry in the table, we get z = 0.28.

Notes

4. Since the area between – 1.65 and z < the area between – 1.65 and 0 (which, from table, is 0.4505), z is negative. Further z can be to the right or to the left of the value –1.65. Thus, when z lies to the right of – 1.65, its value, corresponds to an area (0.4505 – 0.0173) = 0.4332, is given by z = –1.5 (from table). Further, when z lies to the left of – 1.65, its value, corresponds to an area (0.4505 + 0.0173) = 0.4678, is given by z = – 1.85 (from table).
5. Since the area between – 0.5 to z > area between – 0.5 to 0 (which, from table, is 0.1915), z is positive. The value of z, located corresponding to an area (0.5376– 0.1915) = 0.3461, is given by 1.02.

Task If X is a random variate which is distributed normally with mean 60 and standard deviation 5, find the probabilities of the following events:

(i) $60 \le X \le 70$, (ii) $50 \le X \le 65$, (iii) $X > 45$, (iv) $X \le 50$.

Self Assessment

State whether the following statements are true or false:

1. The normal probability distribution occupies a place of central importance in Ancient Statistical Theory.
2. This distribution was first observed as the normal law of errors by the statisticians of the eighteenth century.
3. An observed value of X is the sum of its true value and the net effect of a large number of independent errors which may be positive or negative each with equal probability.
4. The observed distribution of a random variable was found to be in close conformity with a continuous curve, which was termed as the normal curve of errors or simply the normal curve.
5. Various factors operate in such a way that the deviations of observations above and below mean are balanced with regard to their magnitude as well as their number.
6. p and e are absolute constants with values 3.14159.... and 2.71828.... respectively.
7. Normal probability distribution is a bell shaped symmetrical curve about the ordinate at X.
8. Normal probability distribution is unimodal curve and its tails extend infinitely in both directions.
9. In NPD, All the three measures of central tendency coincide, i.e.,

   mean + median + mode=0
10. The total area under the curve gives the total probability of the random variable taking values between - ∞ to ∞.
11. In Normal distribution, since the distribution is symmetrical, all odd ordered central moments are zero.
12. The value of p(X) is always non-negative for all values of X, i.e., the whole curve lies above X axis.
13. The quartiles are equidistant from median, i.e., $Md + Q_1 = Q_3 - Md$, by virtue of symmetry.
14. The value of moment coefficient of skewness is never zero.

Notes

15.5 Normal Approximation to Binominal Distribution

Normal distribution can be used as an approximation to binomial distribution when n is large and neither p nor q is very small. If X denotes the number of successes with probability p of a success in each of the n trials, then X will be distributed approximately normally with mean np and standard deviation $\sqrt{npq}$.

$$\text{Further, } z = \frac{X - np}{\sqrt{npq}} \sim N(0,1).$$

It may be noted here that as X varies from 0 to n, the standard normal variate z would vary from – ∞ to ∞ because

$$\text{when } X = 0, \lim_{n \to \infty}\left(\frac{-np}{\sqrt{npq}}\right) = \lim_{n \to \infty}\left(-\sqrt{\frac{np}{q}}\right) = -\infty$$

$$\text{and when } X = n, \lim_{n \to \infty}\left(\frac{n - np}{\sqrt{npq}}\right) = \lim_{n \to \infty}\left(\frac{nq}{\sqrt{npq}}\right) = \lim_{n \to \infty}\left(\sqrt{\frac{nq}{p}}\right) = \infty$$

Correction for Continuity

Since the number of successes is a discrete variable, to use normal approximation, we have make corrections for continuity. For example,

$P(X_1 \le X \le X_2)$ is to be corrected as $P\left(X_1 - \frac{1}{2} \le X \le X_2 + \frac{1}{2}\right)$, while using normal approximation to binomial since the gap between successive values of a binomial variate is unity. Similarly, $P(X_1 < X < X_2)$ is to be corrected as $P\left(X_1 + \frac{1}{2} \le X \le X_2 - \frac{1}{2}\right)$, since $X_1 < X$ does not include X_1 and $X < X_2$ does not include X_2.

Note: The normal approximation to binomial probability mass function is good when $n \ge 50$ and neither p nor q is less than 0.1.

Example: An unbiased die is tossed 600 times. Use normal approximation to binomial to find the probability obtaining

1. more than 125 aces,
2. number of aces between 80 and 110,
3. exactly 150 aces.

Solution:

Let X denote the number of successes, i.e., the number of aces.

$$\therefore \mu = np = 600 \times \frac{1}{6} = 100 \text{ and } \sigma = \sqrt{npq} = \sqrt{600 \times \frac{1}{6} \times \frac{5}{6}} = 9.1$$

1. To make correction for continuity, we can write

$$P(X > 125) = P(X > 125 + 0.5)$$

Notes

Thus, $P(X \geq 125.5) = P\left(z \geq \frac{125.5-100}{9.1}\right) = P(z \geq 2.80)$

$$= 0.5000 - P(0 \leq z \leq 2.80) = 0.5000 - 0.4974 = 0.0026.$$

2. In a similar way, the probability of the number of aces between 80 and 110 is given by

$$P(79.5 \leq X \leq 110.5) = P\left(\frac{79.5-100}{9.1} \leq z \leq \frac{110.5-100}{9.1}\right)$$

$$= P(-2.25 \leq z \leq 1.15) = P(0 \leq z \leq 2.25) + P(0 \leq z \leq 1.15)$$

$$= 0.4878 + 0.3749 = 0.8627$$

3. $P(X = 120) = P(119.5 \leq X \leq 120.5) = P\left(\frac{19.5}{9.1} \leq z \leq \frac{20.5}{9.1}\right)$

$$= P(2.14 \leq z \leq 2.25) = P(0 \leq z \leq 2.25) - P(0 \leq z \leq 2.14)$$

$$= 0.4878 - 0.4838 = 0.0040$$

Self Assessment

Fill in the blanks:

15. Normal distribution can be used as an approximation to binomial distribution when n is large and ..... p ..... q is very small.
16. In Normal distribution, the standard normal variate z would vary from .....to .....

15.6 Normal Approximation to Poisson Distribution

Normal distribution can also be used to approximate a Poisson distribution when its parameter $m \geq 10$. If X is a Poisson variate with mean m, then, for $m \geq 10$, the distribution of X can be taken as approximately normal with mean m and standard deviation $\sqrt{m}$ so that $z = \frac{X - m}{\sqrt{m}}$ is a standard normal variate.

Example: A random variable X follows Poisson distribution with parameter 25. Use normal approximation to Poisson distribution to find the probability that X is greater than or equal to 30.

Solution:

$P(X \geq 30) = P(X \geq 29.5)$ (after making correction for continuity).

$$= P\left(z \geq \frac{29.5-25}{5}\right) = P(z \geq 0.9)$$

$$= 0.5000 - P(0 \leq z \leq 0.9) = 0.5000 - 0.3159 = 0.1841$$

Fitting a Normal Curve

A normal curve is fitted to the observed data with the following objectives:

1. To provide a visual device to judge whether it is a good fit or not.
2. Use to estimate the characteristics of the population.

Notes

The fitting of a normal curve can be done by

(a) The Method of Ordinates or

(b) The Method of Areas.

(a) *Method of Ordinates:* In this method, the ordinate f(X) of the normal curve, for various values of the random variate X are obtained by using the table of ordinates for a standard normal variate.

We can write $f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} = \frac{1}{\sigma}\phi(z)$

where $z = \frac{X-\mu}{\sigma}$ and $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.

The expected frequency corresponding to a particular value of X is given by $y = N.f(X) = \frac{N}{\sigma}\phi(z)$ and therefore, the expected frequency of a class = $y \times h$, where h is the class interval.

Example: Fit a normal curve to the following data :

Class Intervals :	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80	*Total*
Frequency :	2	11	24	33	20	8	2	100

Solution:

First we compute mean and standard deviation of the given data.

Class Intervals	*Mid - values* (*X*)	*Frequency* (*f*)	$d = \frac{X - 45}{10}$	*fd*	fd^2
10 - 20	15	2	- 3	- 6	18
20 - 30	25	11	- 2	- 22	44
30 - 40	35	24	- 1	- 24	24
40 - 50	45	33	0	0	0
50 - 60	55	20	1	20	20
60 - 70	65	8	2	16	32
70 - 80	75	2	3	6	18
Total		100		- 10	156

Note: If the class intervals are not continuous, they should first be made so.

$$\therefore \ \mu = 45 - 10 \times \frac{10}{100} = 44$$

and $\sigma = 10\sqrt{\frac{156}{100} - \left(\frac{10}{100}\right)^2} = 10\sqrt{1.55} = 12.4$

Notes

Table for the fitting of Normal Curve

Class Intervals	Mid-values (X)	$z = \frac{X-\mu}{\sigma}$	$\phi(z)$ (from table)	$y = \frac{N}{\sigma}\phi(z)$	f_e *
10-20	15	– 2.34	0.0258	0.2081	2
20-30	25	– 1.53	0.1238	0.9984	10
30-40	35	– 0.73	0.3056	2.4645	25
40-50	45	0.08	0.3977	3.2073	32
50-60	55	0.89	0.2685	2.1653	22
60-70	65	1.69	0.0957	0.7718	8
70-80	75	2.50	0.0175	0.1411	1

(b) *Method of Areas:* Under this method, the probabilities or the areas of the random variable lying in various intervals are determined. These probabilities are then multiplied by N to get the expected frequencies. This procedure is explained below for the data of the above example.

Class Intervals	Lower Limit (X)	$z = \frac{X-44}{12.4}$	Area from 0 to z	Area under the class	f_e*
10-20	10	– 2.74	0.4969	0.0231	2
20-30	20	– 1.94	0.4738	0.1030	10
30-40	30	– 1.13	0.3708	0.2453	25
40-50	40	– 0.32	0.1255	0.3099	31
50-60	50	0.48	0.1844	0.2171	22
60-70	60	1.29	0.4015	0.0806	8
70-80	70	2.10	0.4821	0.0160	2
	80	2.90	0.4981		

*Expected frequency approximated to the nearest integer.

Self Assessment

Multiple Choice Questions:

17. A normal curve is fitted to the observed data to provide a .....to judge whether it is a good fit or not

    (a) Device (b) Visual device

    (c) Equipment (d) Apparatus

18. A ..... is fitted to estimate the characteristics of the population.

    (a) Statistical curve (b) Binomial curve

    (c) Normal curve (d) None

19. The fitting of a normal curve can be done by the .....

    (a) Method of Ordinates (b) Method of Averages

    (c) the Method of Abscissa (d) None

20. Under Method of Areas, the ..... or the areas of the random variable lying in various intervals are determined.

    (a) Chances (b) Mean

    (c) Probabilities (d) Coefficient

Notes

Case Study

Rattle Publishing Company Limited

Rattle Publishing Company Limited are planning to introduce a new ABC text-book. The company's marketing department estimates that the prior distribution for likely sales is normal with a mean of 10,000 books. In addition it has determined that there is a probability of one half that the likely sales will lie between 8,000 and 12,000 books.

The text-book will sell for ₹ 10 per copy but the publishing company pays the author 10% of revenue in royalties and the fixed costs of printing and marketing the book are calculated to be ₹ 25,000. Using current printing facilities, the variable production costs are ₹ 4 per book. However, the Rattle Publishing Company has the option of hiring a special machine for ₹ 14,000 which will reduce the variable production costs to ₹ 2.50 per book.

Required:

1. Show that the standard deviation of likely sales is approximately $\sigma = 3{,}000$.
2. Using $\sigma = 3{,}000$ determine the probability that the company will at least break even if:
   (a) existing printing facilities are used,
   (b) the special machine is hired
3. By comparing expected profits, decide whether or not the publishing company should hire the special machine.
4. By using the normal distribution it can be shown that the following probability distribution may be applied to book sales:

Sales ('000)	0-5	5-8	8-10	10-12	12-15	15-20
Probability	0.05	0.20	0.25	0.25	0.20	0.05

By assuming that the actual sales can only take the midpoints of these classes, determine the expected value of perfect information and interpret its value.

15.7 Summary

- The normal probability distribution occupies a place of central importance in Modern Statistical Theory. This distribution was first observed as the normal law of errors by the statisticians of the eighteenth century.
- Random variables observed in many phenomena related to economics, business and other social as well as physical sciences are often found to be distributed normally.
- Here π and e are absolute constants with values 3.14159.... and 2.71828.... respectively.
- This distribution is completely known if the values of mean μ and standard deviation σ are known. Thus, the distribution has two parameters, viz. mean and standard deviation.
- We seldom encounter variables that have a range from $-\infty$ to ∞, nevertheless the curves generated by the relative frequency histograms of various variables closely resembles the shape of normal curve.

Notes

- It is a bell shaped symmetrical curve about the ordinate at X.
- The ordinate is maximum at X.
- It is unimodal curve and its tails extend infinitely in both directions.
- The curve is asymptotic to X-axis in both directions.
- The value of p(X) is always non-negative for all values of X, i.e., the whole curve lies above X-axis
- Since the distribution is symmetrical, all odd ordered central moments are zero.
- The area between the ordinates at $\mu - \sigma$ and $\mu + \sigma$ is 0.6826. This implies that for a normal distribution about 68% of the observations will lie between $\mu - \sigma$ and $\mu + \sigma$.
- The area between the ordinates at $\mu - 2\sigma$ and $\mu + 2\sigma$ is 0.9544. This implies that for a normal distribution about 95% of the observations will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- The area between the ordinates at $\mu - 3\sigma$ and $\mu + 3\sigma$ is 0.9974. This implies that for a normal distribution about 99% of the observations will lie between $\mu - 3\sigma$ and $\mu + 3\sigma$. This result shows that, practically, the range of the distribution is 6σ although, theoretically, the range is from $-\infty$ to ∞.
- Normal distribution can be used as an approximation to binomial distribution when n is large and neither p nor q is very small.

15.8 Keywords

Condition of homogeneity: The factors must be similar over the relevant population although, their incidence may vary from observation to observation.

Condition of independence: The factors, affecting observations, must act independently of each other.

Condition of symmetry: Various factors operate in such a way that the deviations of observations above and below mean are balanced with regard to their magnitude as well as their number.

Fitting a Normal Curve: A normal curve is fitted to the observed data with the objectives (1) To provide a visual device to judge whether it is a good fit or not. (2) Use to estimate the characteristics of the population.

Method of Areas: Under this method, the probabilities or the areas of the random variable lying in various intervals are determined. These probabilities are then multiplied by N to get the expected frequencies.

Method of Ordinates: In this method, the ordinate f(X) of the normal curve, for various values of the random variate X are obtained by using the table of ordinates for a standard normal variate.

Normal Approximation to Poisson Distribution: Normal distribution can also be used to approximate a Poisson distribution when its parameter $m \geq 10$.

Normal Probability Distribution: The normal probability distribution occupies a place of central importance in Modern Statistical Theory. This distribution was first observed as the normal law of errors by the statisticians of the eighteenth century.

Notes

15.9 Review Questions

1. Under what conditions will a random variable follow a normal distribution? State some important features of a normal probability curve.
2. What is a standard normal distribution? Discuss the importance of normal distribution in statistical theory.
3. State clearly the assumptions under which a binomial distribution tends normal distribution.
4. Find the probability that the value of an item drawn at random from a normal distribution with mean 20 and standard deviation 10 will be between (i) 10 and 15, (ii) –5 and 10 and (iii) 15 and 25.
5. In a particular examination an examinee can get marks ranging from 0 to 100. Last year, 1,00,000 students took this examination. The marks obtained by them followed a normal distribution. What is the probability that the marks obtained by a student selected at random would be exactly 63?
6. A collection of human skulls is divided into three classes according to the value of a 'length breadth index' x. Skulls with $x < 75$ are classified as 'long', those with $75 < x < 80$ as 'medium' and those with $x > 80$ as 'short'. The percentage of skulls in the three classes in this collection are respectively 58, 38 and 4. Find, approximately, the mean and standard deviation of x on the assumption that it is normally distributed.
7. In a large group of men, it is found that 5% are under 60 inches and 40% are between 60 and 65 inches in height. Assuming the distribution to be exactly normal, find the mean and standard deviation of the height. The values of z for area equal to 0.45 and 0.05 between 0 to z are 1.645 and 0.125 respectively.
8. Packets of a certain washing powder are filled with an automatic machine with an average weight of 5 kg. and a standard deviation of 50 gm. If the weights of packets are normally distributed, find the percentage of packets having weight above 5.10 kg.
9. For a normal distribution with mean 3 and variance 16, find the value of y such that the probability of the variate lying in the interval (3, y) is 0.4772.
10. The mean income of people working in an industrial city is approximated by a normal distribution with a mean of ₹ 24,000 and a standard deviation of ₹ 3,000. What percentage of the people in this city have income exceeding ₹ 28,500? In a random sample of 50 employed persons of this city, about how many can be expected to have income less than ₹ 19,500?
11. A batch of 5,000 electric lamps have a mean life of 1,000 hours and a standard deviation of 75 hours. Assume a Normal Distribution.
    (a) How many lamps will fail before 900 hours?
    (b) How many lamps will fail between 950 and 1,000 hours?
    (c) What proportion of lamps will fail before 925 hours?
    (d) Given the same mean life, what would the standard deviation have to be to ensure that not more than 20% of lamps fail before 916 hours?

Notes

12. A firm buys springs in very large quantities and from past records it is known that 0.2% are defective. The inspection department sample the springs in batches of 500. It is required to set a standard for the inspectors so that if more than the standard number of defectives is found in a batch the consignment can be rejected with at least 90% confidence that the supply is truly defective.

    How many defectives per batch should be set as the standard?

13. Assume that your working hours X are distributed normally with m = 5 and s = 2. What is the probability of your working 9 hours or more than 9 hours?

14. An assembly line contains 2,000 components each one of which has a limited life. Records show that the life of the components is normally distributed with a mean of 900 hours and a standard deviation of 80 hours.

    (a) What proportion of components will fail before 1,000 hours?

    (b) What proportion will fail before 750 hours?

    (c) What proportion of components fail between 850 and 880 hours?

    (d) Given that the standard deviation will remain at 80 hours what would the average life have to be to ensure that not more than 10% of components fail before 900 hours?

Answers: Self Assessment

1.	False	2.	True
3.	True	4.	True
5.	True	6.	True
7.	True	8.	True
9.	False	10.	True
11.	True	12.	True
13.	False	14.	False
15.	neither, nor	16.	$-\infty$, $+\infty$
17.	(b)	18.	(c)
19.	(a)	20.	(c)

15.10 Further Readings

Books

Balwani Nitin *Quantitative Techniques*, First Edition: 2002. Excel Books, New Delhi.

Bhardwaj R S., *Business Statistics*, Excel Books.

Garrett H.E (1956), *Elementary Statistics*, Longmans, Green & Co., New York.

Selvaraj R, Loganathan, *C Quantitative Methods in Management.*

Stockton and Clark, *Introduction to Business and Economic Statistics*, D.B. Taraporevala Sons and Co. Private Limited, Bombay.

Notes

Online links

http://www.mathsisfun.com/data/standard-normal-distribution.html

http://www.mathsisfun.com/data/standard-normal-distribution-table.html

http://stattrek.com/lesson2/normal.aspx

http://stat.wvu.edu/srs/modules/Normal/normal.html