Dr. C.V. Raman University
Kargi Road, Kota, BILASPUR, (C. G.),
Ph. : +07753-253801, +07753-253872
E-mail : info@cvru.ac.in | Website : www.cvru.ac.in

Business Statistics

Institute of **Open** and **Distance Education**

Faculty of Management

Business Statistics

3BBA5

DR. C.V. RAMAN UNIVERSITY

Chhattisgarh, Bilaspur

A STATUTORY UNIVERSITY UNDER SECTION 2(F) OF THE UGC ACT

3BBA5

Business Statistics

3BBA5, Business Statistics

Edition: March 2024

Compiled, reviewed and edited by Subject Expert team of University

1. Dr. Niket Shukla

(Professor, Dr. C. V. Raman University)

2. Dr. Priyank Mishra

(Associate Professor, Dr. C. V. Raman University)

3. Dr. Anshul Shrivastava

(Assistant Professor, Dr. C. V. Raman University)

Published by:

Dr. C.V. Raman University

Kargi Road, Kota, Bilaspur, (C. G.),

Ph. +07753-253801,07753-253872

E-mail: info@cvru.ac.in

Website: www.cvru.ac.in

TABLE OF CONTENTS

NOTES

1 INTRODUCTION OF STATISTICS

Chapter Includes :

- INTRODUCTION
- KINDS OR BRANCHES STATISTICS
- CHARACTERISTICS OF STATISTICS
- FUNCTIONS OR USES OF STATISTICS
- IMPORTANCE OF STATISTICAL METHODS
- LIMITATIONS OF STATISTICS
- CLASSIFICATION OF DATA
- BASES OF CLASSIFICATION
- COLLECTION OF STATISTICAL DATA
- CLASSIFICATION OF DATA COLLECTION
- METHODS OF COLLECTING PRIMARY DATA
- TEXTUAL PRESENTATION
- GRAPHICAL PRESENTATION OF DATA

Learning Objective :

After going through this chapter, you should be able to:

- Understand concept of statistics
- Learn kinds, functions and characteristics of statistics
- Understand data collection methods
- Explain classification of Data.
- Know presentation of Data.

NOTES

INTRODUCTION

Statistics is an old discipline, as old as the human activity. Its utility has bee increasing as the ages goes by. In the olden days it was used in the administrativ departments of the states and the scope was limited. Earlier it was used b governments to keep record of birth, death, population etc., for administrativ purpose. John Graunt was the first man to make a systematic study of birth an death statistics and the calculation of expectation of life at different age in th 17th century which led to the idea of Life Insurance.

The word 'Statistics' seems to have been derived from the Latin word 'status' o Italian word 'statista' or the German word 'Statistik' each of which means political state. Fields like agriculture, economics, sociology, business manage ment etc., are now using Statistical Method for different purposes.

In the plural sense it means a set of numerical figures called 'data' obtained b counting, or, measurement. In the singular sense it means collection, classifica tion, presentation, analysis, comparison and meaningful interpretation of 'rav data'.

Statistical data help us to understand the economic problems, e.g., balance o trade, disparities of income and wealth, national income accounts, supply anc demand curves, living and whole sale price index numbers, production, con sumption, etc., formulate economic theories and test old hypothesis. It also helps in planning and forecasting.

The success of modern business firms depends on the proper analysis of statistical data. Before expansion and diversification of the existing business or setting up a new venture, the top executives must analyse all facts like raw material prices, consumer-preferences, sales records, demand of products, labour conditions, taxes, etc., statistically. It helps to determine the location and size of business, introduce new products or drop an existing product and in fixing product price and administration. It has also wide application in Operations Research.

Meaning and Definition of Statistics

- "Microsoft reported 80% growth in the revenue during the 3rd quarter",
- "Population growth in the country is 2%"

Above statements are statistical conclusions. These statements are very convenient for the reader or listener to understand the net effect. These statements also help to make policies in the respective areas. To prepare these numerical statements, we need to be familiar with those methods and techniques which are used in data collection presentation, organization and analysis and interpretations. The study of these techniques and method is the science of Statistics.

Definition of Statistics

Statistics has been defined differently by different writers.

According to Webster "Statistics are the classified facts representing the conditions of the people in a state especially those facts which can be stated in numbers or any tabular or classified arrangement."

NOTES

According to Bowley statistics are "Numerical statements of facts in any department of enquiry placed in relation to each other."

According to Yule and Kendall, Statistics means quantitative data affected to a marked extent by multiplicity of causes.

More broad definition of statistics was given by **Horace Secrist,** According to him, statistics means aggregate of facts affected to marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.

This definition points out some essential characteristic that must possess numerical facts so that they may be called statistics. These characteristics are:

1. They are enumerated or estimated according to a reasonable standard of accuracy
2. They are affected by multiplicity of factors
3. They must be numerically expressed
4. They must be aggregate of facts

KINDS OR BRANCHES STATISTICS

Statistics may be divided into two main branches:

(1) Descriptive Statistics

(2) Inferential Statistics

(1) Descriptive Statistics:

In descriptive statistics, it deals with collection of data, its presentation in various forms, such as tables, graphs and diagrams and findings averages and other measures which would describe the data.

For Example: Industrial statistics, population statistics, trade statistics etc. Such as businessman make to use descriptive statistics in presenting their annual reports, final accounts, bank statements.

(2) Inferential Statistics:

In inferential statistics, it deals with techniques used for analysis of data, making the estimates and drawing conclusions from limited information taken on sample basis and testing the reliability of the estimates.

For Example: Suppose we want to have an idea about the percentage of illiterates in our country. We take a sample from the population and find the proportion of illiterates in the sample. This sample proportion with the help of probability enables us to make some inferences about the population proportion. This study belongs to inferential statistics.

CHARACTERISTICS OF STATISTICS

Some of its important characteristics are given below:

- Statistics are aggregates of facts.

NOTES

- Statistics are numerically expressed.
- Statistics are affected to a marked extent by multiplicity of causes.
- Statistics are enumerated or estimated according to a reasonable standard of accuracy.
- Statistics are collected for a predetermine purpose.
- Statistics are collected in a systemic manner.
- Statistics must be comparable to each other.

FUNCTIONS OR USES OF STATISTICS:

1) Statistics helps in providing a better understanding and exact description of a phenomenon of nature.
2) Statistical helps in proper and efficient planning of a statistical inquiry in any field of study.
3) Statistical helps in collecting an appropriate quantitative data.
4) Statistics helps in presenting complex data in a suitable tabular, diagrammatic and graphic form for an easy and clear comprehension of the data.
5) Statistics helps in understanding the nature and pattern of variability of a phenomenon through quantitative observations.
6) Statistics helps in drawing valid inference, along with a measure of their reliability about the population parameters from the sample data.

ROLE OF STATISTICS IN DIFFERENT FIELDS

Statistics plays a vital role in every fields of human activity. Statistics has important role in determining the existing position of per capita income, unemployment, population growth rate, housing, schooling medical facilities etc...in a country. Now statistics holds a central position in almost every field like Industry, Commerce, Trade, Physics, Chemistry, Economics, Mathematics, Biology, Botany, Psychology, Astronomy etc, so application of statistics is very wide. Now we discuss some important fields in which statistics is commonly applied.

1) **Business:** Statistics play an important role in business. A successful businessman must be very quick and accurate in decision making. He knows that what his customers wants, he should therefore, know what to produce and sell and in what quantities. Statistics helps businessman to plan production according to the taste of the costumers, the quality of the products can also be checked more efficiently by using statistical methods. So all the activities of the businessman based on statistical information. He can make correct decision about the location of business, marketing of the products, financial resources etc...
2) **In Economics:** Statistics play an important role in economics. Economics largely depends upon statistics. National income accounts are multipurpose

NOTES

indicators for the economists and administrators. Statistical methods are used for preparation of these accounts. In economics research statistical methods are used for collecting and analysis the data and testing hypothesis. The relationship between supply and demands is studies by statistical methods, the imports and exports, the inflation rate, the per capita income are the problems which require good knowledge of statistics.

3) **In Mathematics:** Statistical plays a central role in almost all natural and social sciences. The methods of natural sciences are most reliable but conclusions draw from them are only probable, because they are based on incomplete evidence. Statistical helps in describing these measurements more precisely. Statistics is branch of applied mathematics. The large number of statistical methods like probability averages, dispersions, estimation etc... is used in mathematics and different techniques of pure mathematics like integration, differentiation and algebra are used in statistics.

4) **In Banking:** Statistics play an important role in banking. The banks make use of statistics for a number of purposes. The banks work on the principle that all the people who deposit their money with the banks do not withdraw it at the same time. The bank earns profits out of these deposits by lending to others on interest. The bankers use statistical approaches based on probability to estimate the numbers of depositors and their claims for a certain day.

5) **In State Management (Administration):** Statistics is essential for a country. Different policies of the government are based on statistics. Statistical data are now widely used in taking all administrative decisions. Suppose if the government wants to revise the pay scales of employees in view of an increase in the living cost, statistical methods will be used to determine the rise in the cost of living. Preparation of federal and provincial government budgets mainly depends upon statistics because it helps in estimating the expected expenditures and revenue from different sources. So statistics are the eyes of administration of the state.

6) **In Accounting and Auditing:** Accounting is impossible without exactness. But for decision making purpose, so much precision is not essential the decision may be taken on the basis of approximation, know as statistics. The correction of the values of current asserts is made on the basis of the purchasing power of money or the current value of it. In auditing sampling techniques are commonly used. An auditor determines the sample size of the book to be audited on the basis of error.

7) **In Natural and Social Sciences:** Statistics plays a vital role in almost all the natural and social sciences. Statistical methods are commonly used for analyzing the experiments results, testing their significance in Biology, Physics, Chemistry, Mathematics, Meteorology, Research chambers of commerce, Sociology, Business, Public Administration, Communication and Information Technology etc...

NOTES

8) **In Astronomy:** Astronomy is one of the oldest branches of statistical study; it deals with the measurement of distance, sizes, masses and densities of heavenly bodies by means of observations. During these measurements errors are unavoidable so most probable measurements are founded by using statistical methods. Example: This distance of moon from the earth is measured. Since old days the astronomers have been statistical methods like method of least squares for finding the movements of stars.

IMPORTANCE OF STATISTICAL METHODS:

Statistical methods are used not only in the social, economic and political fields but in every field of science and knowledge. Statistical analysis has become more significant in global relations and in the age of fast developing information technology.

According to Prof. Bowley, ***"The proper function of statistics is to enlarge individual experiences".***

Following are some of the **important functions of Statistics:**

a) To provide numerical facts.

b) To simplify complex facts.

c) To enlarge human knowledge and experience.

d) Helps in formulation of policies.

e) To provide comparison.

f) To establish mutual relations.

g) Helps in forecasting.

h) Test the accuracy of scientific theories.

i) To study extensively and intensively.

The use of statistics has become almost essential in order to clearly understand and solve a problem. Statistics proves to be much **useful in unfamiliar fields of application and complex situations** such as:-

a) Planning

b) Administration

c) Economics

d) Trade & Commerce

e) Production management

f) Quality control

g) Helpful in inspection

h) Insurance business

i) Railways & transport Co

a) Banking Institutions

b) Speculation and Gambling

c) Underwriters and Investors

d) Politicians & social workers.

LIMITATIONS OF STATISTICS:

The important limitations of statistics are:

1) Statistics laws are true on average. Statistics are aggregates of facts. So single observation is not a statistics, it deals with groups and aggregates only.

2) Statistical methods are best applicable on quantitative data.

3) Statistical cannot be applied to heterogeneous data.

4) It sufficient care is not exercised in collecting, analyzing and interpretation the data, statistical results might be misleading.

5) Only a person who has an expert knowledge of statistics can handle statistical data efficiently.

6) Some errors are possible in statistical decisions. Particularly the inferential statistics involves certain errors. We do not know whether an error has been committed or not.

CLASSIFICATION OF DATA:

The process of arranging data into homogenous group or classes according to some common characteristics present in the data is called classification.

For Example: The process of sorting letters in a post office, the letters are classified according to the cities and further arranged according to streets.

Connor defined classification as: "the process of arranging things in groups or classes according to their resemblances and affinities and gives expression to the unity of attributes that may subsist amongst a diversity of individuals".

The raw data, collected in real situations and arranged haphazardly, do not give a clear picture.

Thus to locate similarities and reduce mental strain we resort to classification. Classification condenses the data by dropping out unnecessary details. It facilitates comparison between different sets of data clearly showing the different points of agreement and disagreement. It enables us to study the relationship between several characteristics and make further statistical treatment like tabulation, etc.

During population census, people in the country are classified according to sex (males/females), marital status (married/unmarried), place of residence (rural/urban), Age (0-5 years, 6-10 years, 11-15 years, etc.), profession (agriculture, production, commerce, transport, doctor, others), residence in states (West Bengal, Bihar, Mumbai, Delhi, etc.), etc.

NOTES

Main objectives of Classification:-

(i) To make the data easy and precise

(ii) To facilitate comparison

(iii) Classified facts expose the cause-effect relationship.

(iv) To arrange the data in proper and systematic way

(v) The data can be presented in a proper tabular form only.

Essentials of an Ideal Classification:-

(i) Classification should be so exhaustive and complete that every individual unit is included in one or the other class.

(ii) Classification should be suitable according to the objectives of investigation.

(iii) There should be stability in the basis of classification so that comparison can be made.

(iv) The facts should be arranged in proper and systematic way.

(v) Data should be classified according to homogeneity.

(vi) It should be arithmetically accurate.

Primary Rules of Classification

In quantitative classification, we classify data by assigning arbitrary limits called class-limits. The group between any two class-limits is termed as class or class-interval. The primary rules of classification are given below:

1) There should not be any ambiguity in the definition of classes. It will eliminate all doubts while including a particular item in a class.

2) All the classes should preferably have equal width or length. Only in some special cases, we use classes of unequal width.

3) The class-limits (integral or fractional) should be selected in such a way that no value of the item in the raw data coincides with the value of the limit.

4) The number of classes should preferably be between 10 and 20, i.e., neither too large nor too small.

5) The classes should be exhaustive, i.e., each value of the raw data should be included in them.

6) The classes should be mutually exclusive and non-overlapping, i.e., each item of the raw data should fit only in one class.

7) The classification must be suitable for the object of inquiry.

8) The classification should be flexible and items included in each class must be homogeneous.

9) Width of class-interval is determined by first fixing the no. of class-intervals and then dividing the total range by that number.

BASES OF CLASSIFICATION:

There are four important bases of classification:

- Qualitative Base
- Quantitative Base
- Temporal Base
- Spatial

1) Qualitative classification: It is done according to attributes or non-measurable characteristics; like social status, sex, nationality, occupation, etc. For example, the population of the whole country can be classified into four categories as married, unmarried, widowed and divorced. When only one attribute, e.g., sex, is used for classification, it is called simple classification. When more than one attributes, e.g., deafness, sex and religion, are used for classification, it is called manifold classification.

2) Quantitative classification: It is done according to numerical size like weights in kg or heights in cm. Here we classify the data by assigning arbitrary limits known as class-limits. The quantitative phenomenon under study is called a variable. For example, the population of the whole country may be classified according to different variables like age, income, wage, price, etc. Hence this classification is often called 'classification by variables'.

   Variable: A variable in statistics means any measurable characteristic or quantity which can assume a range of numerical values within certain limits, e.g., income, height, age, weight, wage, price, etc. A variable can be classified as either discrete or continuous.

   a) Discrete variable: A variable which can take up only exact values and not any fractional values is called a 'discrete' variable. Number of workmen in a factory, members of family, students in a class, number of births in a certain year, number of telephone calls in a month, etc., are examples of discrete-variable.

   b) Continuous variable: A variable which can take up any numerical value (integral/fractional) within a certain range is called a 'continuous' variable. Height, weight, rainfall, time, temperature, etc., are examples of continuous variables. Age of students in a school is a continuous variable as it can be measured to the nearest fraction of time, i.e., years, months, days, etc.

3) Temporal classification: It is done according to time, e.g., index numbers arranged over a period of time, population of a country for several decades, exports and imports of India for different five year plans, etc.

4) Spatial classification: It is done with respect to space or places, e.g., production of cereals in quintals in various states, population of a country according to states, etc.

NOTES

Check Your Progress:

1. What is statistics?
2. Name different kinds of statistics?

NOTES

TYPES OF CLASSIFICATION:

One -way Classification:

If we classify observed data keeping in view single characteristic, this type of classification is known as one-way classification. For Example: The population of world may be classified by religion as Muslim, Christians etc.

Two -way Classification:

If we consider two characteristics at a time in order to classify the observed data then we are doing two way classifications. For Example: The population of world may be classified by Religion and Sex.

Multi -way Classification:

We may consider more than two characteristics at a time to classify given data or observed data. In this way we deal in multi-way classification. For Example: The population of world may be classified by Religion, Sex and Literacy.

COLLECTION OF STATISTICAL DATA

Data means information. The first step in any enquiry (investigation) is collection of data. The data may be collected for the whole population or for a sample only. It is mostly collected on sample basis. Collection of data is very difficult job. The enumerator or investigator is the well trained person who collects the statistical data. The respondents (information) are the persons whom the information is collected.

It should be remembered that data collection is the foundation stone of statistical investigation, on which the entire structure of investigation is constructed. Therefore, data should be collected with maximum efficiency, ability and accuracy. Because if there is any deficiency in this process; the conclusion drawn will be fallacious and unreliable. Collection of data means collection of numerical information related to the subject matter coming under the purview of the investigation.

Collection of data is the basic activity of statistical science. It means collection of facts and figures relating to particular phenomenon under the study of any problem whether it is in business economics, social or natural sciences.

Such material can be obtained directly from the individual units, called primary sources or from the material published earlier elsewhere known as the secondary sources.

Data collected expressly for a specific purpose are called 'Primary data' e.g., data collected by a particular person or organization from the primary source for his own use, collection of data about the population by censuses and surveys, etc. Data collected and published by one organization and subsequently used by other organizations are called 'Secondary data'.

The various sources of collection for secondary data are: newspapers and periodicals; publications of trade associations; research papers published by university departments, U.G.C. or research bureaus; official publications of central, state and the local and foreign governments, etc.

NOTES

The collection expenses of primary data are more than secondary data. Secondary data should be used with care. The various methods of collection of primary data are: (i) Direct personal investigation (interview/observation); (ii) Indirect oral investigation; (iii) Data from local agents and correspondents; (iv) Mailed questionnaires; (v) Questionnaires to be filled in by enumerators; (vi) Results of experiments, etc. Data collected in this manner are called 'raw data'. These are generally voluminous and have to be arranged properly before use.

Statistical Data: A sequence of observation, made on a set of objects included in the sample drawn from population is known as statistical data.

Ungrouped Data: Data which have been arranged in a systematic order are called raw data or ungrouped data.

Grouped Data: Data presented in the form of frequency distribution is called grouped data.

CLASSIFICATION OF DATA COLLECTION -

By now you have known that data could be classified in the following three ways:

a) Quantitative' and Qualitative Data.

b) Sample and Census Data.

c) Primary and Secondary data.

a) **Quantitative and Qualitative data:** Quantitative data are those set of information which are quantifiable and can be expressed in some standard units like rupees, kilograms, litres, etc. For example, pocket money of students of a class and income of their parents can be expressed in so many rupees; production or import of wheat can be expressed in so many kilograms or lakh quintals; consumption of petrol and diesel in India as so many lakh litres in one year and so on.

   Qualitative data, on the other hand, are not quantifiable, that is, cannot be expressed in standard units of measurement like rupees, kilograms, litres, etc. This is because they are 'features', 'qualities' or 'characteristics' like eye color, skin complexion, honesty, good or bad, etc. These are also referred to as attributes. In this case, however, it is possible to count the number of individuals (or items) possessing a particular attribute.

b) **Sample and Census Data:** The data can be collected either by census method or sample method. Information collected through sample inquiry is called sample data and the one collected through census inquiry is called census data. Population census data are collected every ten years in India.

c) **Primary and Secondary Data:** Primary data are collected by the investigator through field survey. Such data are in raw form and must be refined before use. On the other hand, secondary data are extracted from the existing published or unpublished sources, that from the data already collected by others. Collection of data is the first basic step towards the statistical analysis of any problem. The collected data are suitably transformed and analysed to draw conclusions about the population. These conclusions may be either or both of the following:

NOTES

- To estimate one or more parameters of a population or the nature of the population itself. This forms the subject matter of the theory of estimation.
- To test a hypothesis. A hypothesis is a statement regarding the parameters or the nature of population.

Primary Data means those data which are originally collected by an investigator or agency for the first time for any statistical investigation. The primary data are the first hand information collected, compiled and published by organization for some purpose. They are most original data in character and have not undergone any sort of statistical treatment. Example: Population census reports are primary data because these are collected, complied and published by the population census organization.

According to Secrist, "By primary data we meant those which are original, that is, those in which little are no grouping has been made, the instance being recorded or itemized as encountered. They are essentially raw materials."

Secondary Data are those which are already collected by other persons or agencies and the investigator just uses them. The secondary data are the second hand information which are already collected by someone (organization) for some purpose and are available for the present study. The secondary data are not pure in character and have undergone some treatment at least once. Example: Economics survey of England is secondary data because these are collected by more than one organization like Bureau of statistics, Board of Revenue, the Banks etc.

According to M.M. Blair, "Secondary data are those already in existence and which have been collected for some other purpose than the answering of the question at hand."

METHODS OF COLLECTING PRIMARY DATA:

The methods commonly used for the collection of primary data are as follows:

a) Direct Personal Investigation,

b) Indirect Oral Investigation

c) Data collected though Correspondents or Local Sources

d) Investigation through Schedules and Questionnaires filled in by informants

e) Information through Questionnaires filled in by the enumerators

f) Investigation through Registration Method

Primary data are collected by the following methods:

- **Personal Investigation:** This is a very general method of collecting primary data. Here the investigator directly contacts the informants, solicits their cooperation and enumerates the data. The information is collected by direct personal interviews. The novelty of this method is its simplicity. It is neither difficult for the enumerator nor the informants. Because both are present at the spot of data collection. This method provides most accurate information as the investigator

NOTES

collects them personally. But as the investigator alone is involved in the process, his personal bias may influence the accuracy of the data. So it is necessary that the investigator should be honest, unbiased and experienced. In such cases the data collected may be fairly accurate. However, the method is quite costly and time-consuming. So the method should be used when the scope of enquiry is small. This method of collecting data is only applicable in case of small research projects.

Merits- The main advantages of direct personal investigation are as follows:

a) **High Level of Accuracy:** As the data are collected by the investigator himself, they are bound to be more accurate, reliable and standardized.

b) **Flexibility:** This method is flexible because the investigator makes the necessary adjustments in the questions, while collecting data. Moreover, he can explain in a better way the meaning, objective and spirit of the questions.

c) **Homogeneity:** As the data are collected by one person, the quality of homogeneity and uniformity exist in data which make data easily comparable.

d) **Cross-checking:** In this method an investor visits the spot himself and thus he can check the correctness of information by careful observation and intelligent cross-questioning.

e) **Collection of other information:** The investigator may collect other information also which may be used easily in other investigations.

Demerits: The main demerits of this method are as follows:

a) **Limited Area:** This method is suitable only for intensive study in limited area. In other words, this method is not suitable if the field of investigation is too wide in terms of the number of persons to be contacted and area to be covered.

b) **Effect of personal bias:** Every investigator has its own approach, assumptions and prejudices, which may affect the findings of investigation.

c) **More costly:** This method is more costly and requires more time.

d) **Doubt in reliability:** On account of limited area of investigation it is just possible that data used may not represent properly the whole universe and findings may become fallacious. The success of such investigation largely depends upon the ability, efficiency and tactful skill of investigator and in case of absence of such qualities, the result obtained may not be fully reliable.

- **Through Investigation:** Trained investigators are employed to collect the data. These investigators contact the individuals and fill in questionnaire after asking the required information. Most of the organizing implied this method.
- **Collection through Questionnaire:** The researchers get the data from local representation or agents that are based upon their own experience. This method is quick but gives only rough estimate.

NOTES

- **Through Telephone:** The researchers get information through telephone this method is quick and give accurate information.

SOURCES OF COLLECTING SECONDARY DATA

The secondary data are collected by the following sources:

- Official: e.g. The publications of the Statistical Division, Ministry of Finance, the Federal Bureaus of Statistics, Ministries of Food, Agriculture, Industry, Labor etc...
- Semi-Official; e.g. State Bank, Railway Board, Central Cotton Committee, Boards of Economic Enquiry etc.
- Publication of Trade Associations, Chambers of Commerce etc.
- Technical and Trade Journals and Newspapers.
- Research Organizations such as Universities and other institutions.

DIFFERENCE BETWEEN PRIMARY AND SECONDARY DATA

The difference between primary and secondary data is only a change of hand. The primary data are the first hand data information which is directly collected form one source. They are most original data in character and have not undergone any sort of statistical treatment while the secondary data are obtained from some other sources or agencies. They are not pure in character and have undergone some treatment at least once.

For Example: Suppose we interested to find the average age of MS students. We collect the age's data by two methods; either by directly collecting from each student himself personally or getting their ages from the university record. The data collected by the direct personal investigation is called primary data and the data obtained from the university record is called secondary data.

	Primary Data	Secondary Data
Basic Nature	Primary data are original and are collected for the first time.	Data which are collected earlier by someone else, and which are now in published or unpublished state
Collecting Agency	These data are collected by the investigator himself	Secondary data were collected earlier by some other person.
Post collection alterations	These data do not need alteration as they are according to the requirement of the investigation	These have to be analyzed and necessary changes have to be made to make them useful as per the requirements of investigation.
Time & Money	More time, energy and money has to be spent in collection of these data.	Comparatively less time and money is to be spent.

PRESENTATION OF STATISTICAL DATA

Statistical data can be presented in three different ways: (a) Textual presentation, (b) Tabular presentation, and (c) Graphical presentation.

NOTES

TEXTUAL PRESENTATION:

This is a descriptive form. Following is an example of such a presentation of data about deaths from industrial diseases in Great Britain in 1935-39 and 1940-44.

Example: Numerical data with regard to industrial diseases and deaths there form in Great Britain during the years 1935-39 and 1940-44 are given in a descriptive form:

"During the quinquennium 1935-39, there were in Great Britain 1, 775 cases of industrial diseases made up of 677 cases of lead poisoning, 111 of other poisoning, 144 of anthrax, and 843 of gassing. The number of deaths reported was 20 p.c. of the cases for all the four diseases taken together, that for lead poisoning was 135, for other poisoning 25 and that for anthrax was 30.

During the next quinquennium, 1940-44, the total number of cases reported was 2, 807. But lead poisoning cases reported fell by 351 and anthrax cases by 35. Other poisoning cases increased by 784 between the two periods. The number of deaths reported decreased by 45 for lead poisoning, but decreased only by 2 for anthrax from the pre-war to the post-war quinquennium. In the later period, 52 deaths were reported for poisoning other than lead poisoning. The total number of deaths reported in 1940-44 including those from gassing was 64 greater than in 1935-39".

The disadvantages of textual presentation are:

(i) it is too lengthy;

(ii) there is repetition of words;

(iii comparisons cannot be made easily;

(iv) it is difficult to get an idea and take appropriate reaction.

TABULAR PRESENTATION OR TABULATION

Tabulation may be defined as the systematic presentation of numerical data in rows or/and columns according to certain characteristics. It expresses the data in concise and attractive form which can be easily understood and used to compare numerical figures. Before drafting a table, you should be sure what you want to show and who will be the reader.

The descriptive form of previous example has been condensed below in the form of a Table.

NOTES

Death from Industrial Diseases in Great Britain

Sl. No.	Diseases	1935–39			1940–44		
		Number of cases	Number of deaths	% of deaths	Number of cases	Number of deaths	% of deaths
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	Lead poisoning	677	135	19.94	326	90	27.60
2	Anthrax	144	30	20.83	109	28	25.69
3	Gassing	843	165	19.57	1. 477	249	16.86
4	Other poisoning	111	25	22.52	895	52	5.81
	Total	1. 775	355	20.00	2. 807	419	14.93

The advantages of a tabular presentation over the textual presentation are:

(i) it is concise;

(ii) there is no repetition of explanatory matter;

(iii) comparisons can be made easily;

(iv) the important features can be highlighted; and

(v) errors in the data can be detected.

An ideal statistical table should contain the following items:

a) **Table number:** A number must be allotted to the table for identification, particularly when there are many tables in a study.

b) **Title:** The title should explain what is contained in the table. It should be clear, brief and set in bold type on top of the table. It should also indicate the time and place to which the data refer.

c) **Date:** The date of preparation of the table should be given.

d) **Stubs or Row designations:** Each row of the table should be given a brief heading. Such designations of rows are called "stubs", or, "stub items" and the entire column is called "stub column".

e) **Column headings or Captions:** Column designation is given on top of each column to explain to what the figures in the column refer. It should be clear and precise. This is called a "caption", or, "heading". Columns should be numbered if there are four or more columns.

f) **Body of the table:** The data should be arranged in such a way that any figure can be located easily. Various types of numerical variables should be arranged in an ascending order, i.e., from left to right in rows and from top to bottom in columns. Column and row totals should be given.

g) **Unit of measurement:** If the unit of measurement is uniform throughout the table, it is stated at the top right-hand corner of the table along with the title. If different rows and columns contain figures in different units, the units may be stated along with "stubs", or, "captions". Very large figures may be rounded up but the method of rounding should be explained.

NOTES

h) **Source:** At the bottom of the table a note should be added indicating the primary and secondary sources from which data have been collected.

i) **Footnotes and references:** If any item has not been explained properly, a separate explanatory note should be added at the bottom of the table.

A table should be logical, well-balanced in length and breadth and the comparable columns should be placed side by side. Light/heavy/thick or double rulings may be used to distinguish sub columns, main columns and totals. For large data more than one table may be used.

Example: Draw up a blank table to show the number of employees in a large commercial firm, classified according to (i) Sex: Male and Female; (ii) Three age-groups: below 30, 30 and above but below 45, 45 and above; and (iii) Four income-groups: below Rs. 400, Rs. 400-750, Rs. 750-1, 000, above Rs. 1, 000.

Number of employees in a large commercial firm classified by sex, three-age groups and four income-groups

Age-groups	Below 30 (Nos)			30–45 (Nos)			45 and above (Nos)			Total (Nos)		
Sex / Income groups	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
1. Below Rs. 400												
2. Rs. 400 to Rs. 750												
3. Rs. 750 to Rs. 1000												
4. Above Rs. 1000												
Grand total												

Example: *(a)* Industrial finance in India showed great variation in respect of sources during the first, second and third plans. There were two main sources, viz., internal and external. The former had two sources-depreciation and free reserves and surplus. The latter had three sources-capital issues, borrowing and 'other sources'.

During the first plan internal and external sources accounted for 62% and 38% of the total, and in this depreciation, fresh capital and 'other sources' formed 29%, 7% and 10.6% respectively.

During the second plan internal sources decreased by 17.3% and external sources increased by 17.3% as compared to the first plan, and depreciation was 24.5%. The external finance during the same period consisted of fresh capital 10.9% and borrowings 28.9%.

Compared to the second plan, during the third plan external finance decreased by 4.4% and borrowings and other sources were 29.4% and 14.9%. During the third plan, internal finance increased by 4.4% and free reserves and surplus formed 18.6%.

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

NOTES

Industrial Finance Pattern of India (in Percentage)

Plans	Sources						
	Internal			External			
	Depre-ciation	Free reserves and surplus	Total	Capital issues	Borrow-ings	Other sources	Total
First	29	33	62	7	20.4	10.6	38
Second	24.5	20.2	44.7	10.9	28.9	15.5	55.3
Third	30.5	18.6	49.1	6.6	29.4	14.9	50.9

(b) A survey of 370 students from commerce faculty and 130 students from 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 for Costing only 13 were girls and 90 boys belonged to commerce faculty. Out of 80 studying both C.A. and Costing; 72 were from commerce faculty amongst which 70 were boys. Amongst those who offered part-time management courses; 50 boys were from science faculty and 30 boys and 10 girls from commerce faculty. In all there were 110 boys in science faculty. Out of 180 studying C.A., only, 150 boys and 8 girls were from commerce faculty and 6 girls from science faculty.

Present the above information in a tabular form. Find the number of students from science faculty studying for part-time management courses.

Number of Students Studying in Different Faculties and Courses

Faculty / Courses	Commerce (Nos)			Science (Nos)			Total (Nos)		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
Part-time Management	30	10	40	50	10	60	80	20	100
C.A.	150	8	158	16	6	22	166	14	180
Costing	90	10	100	37	3	40	127	13	140
C.A. and Costing	70	2	72	7	1	8	77	3	80
Total	340	30	370	110	20	130	450	50	500

Objectives of Tabulation

The main objectives of tabulation are stated below:

(i) to carry out investigation;

(ii) to do comparison;

(iii) to locate omissions and errors in the data;

(iv) to use space economically;

(v) to study the trend;

(vi) to simplify data;

(vii) to use it as future reference.

NOTES

Sorting

Sorting of data is the last process of tabulation. It is a time-consuming process when the data is too large. After classification the data may be sorted using either of the following methods:

(i) Manual method: Here the sorting is done by hand by giving tally marks for the number of times each event has occurred. Next the total tally marks are counted. The method is simple and suitable for limited data.

(ii) Mechanical and electrical method: To reduce the sorting time mechanical devices may be used. This is described as mechanical tabulation. For electrical tabulation data should be codified first and then punched on card. For each data a separate card is used. The punched cards are checked by a machine called 'verifier'. Next the cards are sorted out into different groups as desired by a machine called 'sorter'. Finally, the tabulation is done by using a tabulator. The same card may be sorted out more than once for completing tables under different titles.

(iii) Tabulation using electronic computer: It is convenient to use electronic computer for sorting when (a) data are very large; (b) data have to be sorted for future use and (c) the requirements of the table are changing. Such tabulation is less time-consuming and more accurate than the manual method.

GRAPHICAL PRESENTATION OF DATA

A graph refers to the plotting of different valves of the variables on a graph paper which gives the movement or a change in the variable over a period of time. Diagrams can present the data in an attractive style but still there is a method more reliable than this. Diagrams are often used for publicity purposes but are not of much use in statistical analysis. Hence graphic presentation is more effective and result oriented.

Diagrams can present the data in an attractive style but still there is a method more reliable than this. Diagrams are often used for publicity purposes but are not of much use in statistical analysis. Hence graphic presentation is more effective and meaningful.

According to A. L. Boddington, "The wandering of a line is more powerful in its effect on the mind than a tabulated statement; it shows what is happening and what is likely to take place, just as quickly as the eye is capable of working."

Advantages of Graphs

The presentation of statistics in the form of graphs facilitates many processes in economics. The main uses of graphs are as under:

- Attractive and Effective presentation of Data: The statistics can be presented in attractive and effective way by graphs. A fact that an ordinary man cannot understand easily, could understand in a better way by graphs. Therefore, it is said that a picture is worth of a thousand words.

NOTES

- **Simple and Understandable Presentation of Data:** Graphs help to present complex data in a simple and understandable way. Therefore, graphs help to remove the complex nature of statistics.
- **Useful in Comparison:** Graphs also help to compare the statistics. IF investment made in two different ventures is presented through graphs, then it becomes easy to understand the difference between the two.
- **Useful for Interpretation:** Graphs also help to interpret the conclusion. It saves time as well as labour.
- **Remembrance for long period:** Graphs help to remember the facts for a long time and they cannot be forgotten.
- **Helpful in Predictions:** Through graphs, tendencies that could occur in near future can be predicted in a better way.
- **Universal utility:** In modern era, graphs can be used in all spheres such as trade, economics, government departments, advertisement, etc.
- **Information as well as Entertainment:** Graphs help us in entertainment as well as for providing information. By graphs there occurs no hindrance in the deep analysis of every information.
- **Helpful in Transmission of Information:** Graphs help in the process of transmission as well as information of facts.
- **No Need for training:** when facts are presented through graphs there is any need for special training for these interpretations.

Rules for the construction of Graph

The following are the main rules to construct a graph:

- Every graph must have a suitable title which should clearly convey the main idea, the graph intends to portray.
- The graph must suit to the size of the paper.
- The scale of the graph should be in even numbers or in multiples.
- Footnotes should be given at the bottom to illustrate the main points about the graph.
- Graph should be as simple as possible.
- In order to show many items in a graph, index for identification should be given.
- A graph should be neat and clean. It should be appealing to the eyes.
- Every graph should be given with a table to ensure whether the data has been presented accurately or not.
- The test of a good graph depends on the case with which the observer can interpret it. Thus economy in cost and energy should be exercised in drawing the graph.

NOTES

Limitations

Following are the main drawbacks/ limitations of graphs.

Limited Application: Graphic representation is useful for a common man but for an expert, its utility is limited.

Lack of Accuracy: Graphs do not measure the magnitude of the data. They only depict the fluctuations in them.

Subjective: Graphs are subjective in character. Their interpretation varies from person to person.

Misleading Conclusions: The person who has no knowledge can draw misleading conclusions from graphs.

Simplicity: Graph should be as simple as possible.

Index: In order to show many items in a graph, index for identification should be given.

Scale for a graph:

The scale indicates the unit of a variable that a fixed length of axis would represent. Scale may be different for both the axes. It should be taken in such a way so as to accommodate whole of the data on a given graph paper in a lucid and attractive style. Sometimes data to be presented does not have low values but with large terms. We have to use the graph so as it may present the given data for comparison even.

Types of Graphs:

There are two types of graphs.

- Time series Graphs
- Frequency Distribution Graphs.

Time series graphs may be of one variable, two variables or more variables graph. Frequency distribution graphs present (a) histograms (b) Frequency Polygons (c) Frequency Curves and (d) Ogives.

Histogram, Frequency Polygon and Frequency Curve

Histogram is a very common type of graph for displaying classified data. It is a set of rectangles erected vertically. It has the following features:

a) It is a rectangular diagram.

b) Since the rectangles are drawn with specified width and height, histogram is a two dimensional diagram. The width of a rectangle equals to the class interval and height

$$= \frac{\text{Class frequency x Width of the shortest class interval in the data}}{\text{Width of the class interval}}$$

c) The area of each rectangle is proportional to the frequency of the respective class.

Check Your Progress:

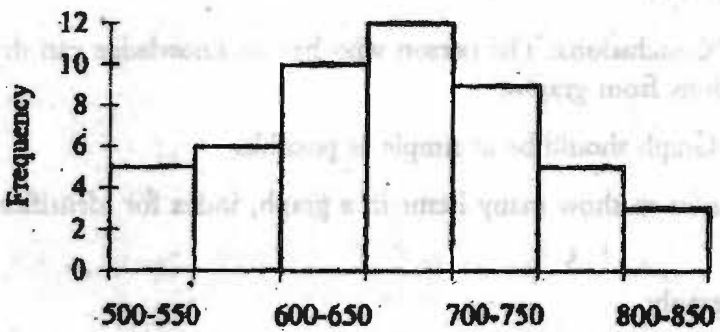
3. What is data classification?
4. What are methods of primary data collection?

NOTES

Construction of Histogram

To plot a histogram of the frequency distribution given in Table 3% on a graph paper, we mark off class intervals like 500 - 550,550 - 600, etc. on the horizontal axis. Similarly, we mark off frequencies on the vertical axis. Since all the classes are of equal width, the height of each rectangle is taken to be equal to the frequency of the respective class. The histogram is shown in Fig.

Histogram



Advantages of histogram are:

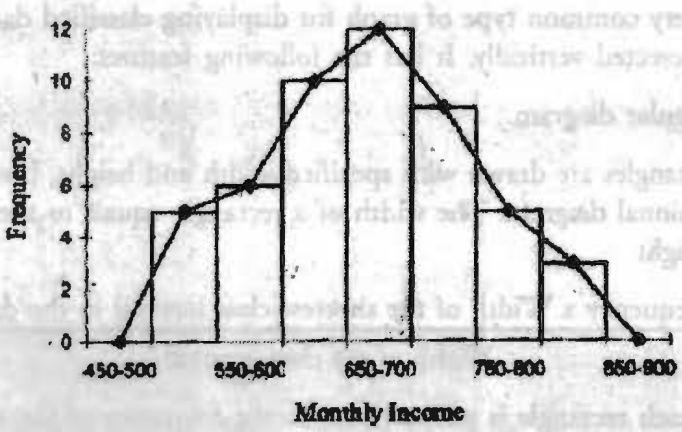
1) The width of various rectangles show the nature of classes in the distribution, i.e., whether of equal width or not.

2) Area of a rectangle shows the proportion of the class frequency in the total.

Frequency Polygon

Frequency Polygon has been derived from the word "polygon" which means many sides. In statistics, it means a graph of a frequency distribution. A frequency polygon is obtained from a histogram by joining the mid-points of the top of various rectangles with the help of straight lines, as shown in Fig. 3.4. In order that total area under the polygon remains equal to the area under histogram, two arbitrary classes, each with zero frequency, are added on both ends, as shown below.

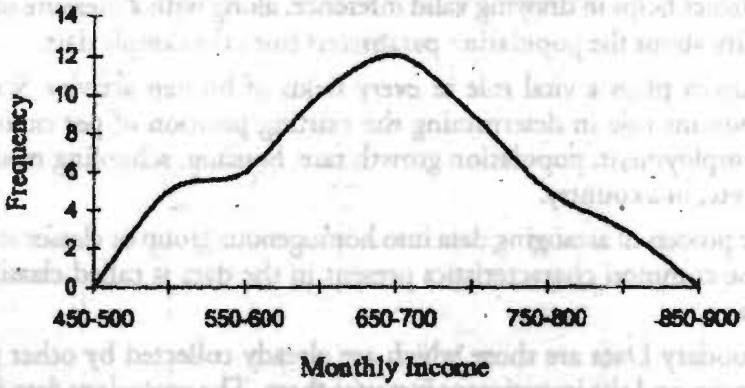
Frequency Polygon



Frequency Curve

If the points, obtained in the case of frequency polygon are joined with the help of a smooth curve, we get a frequency curve as shown in Fig. 3.5.

Frequency Curve



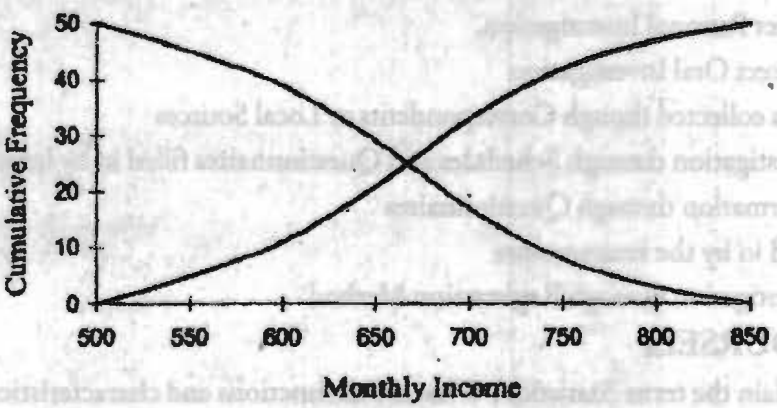
Cumulative Frequency Curve - Ogives

The graph of a cumulative frequency distribution is known as cumulative frequency curve or ogive. Since a cumulative frequency distribution can be of 'less than' or 'greater than' type, accordingly, we can have 'less than' or 'greater than' type of ogives.

Ogivcs can be used to locate, graphically, certain partition values. We can also determine the percentage of observations lying between given limits. The ogives for the cumulative frequency distributions given in Tables 3.12 and 3.13 are drawn in Fig. 3.6.

Note that to draw a less than type ogive, we add a class interval of 'less than 500' with frequency equal to zero. Similarly, we add a class interval of 'more than 900' with frequency zero for the construction of a greater than type ogive. 'Less than' and 'More than' type Ogives

'Less than' and 'More than' type Ogives



SUMMARY

- Statistics are the classified facts representing the conditions of the people in a state especially those facts which can be stated in numbers or any tabular or classified arrangement.

NOTES

- Statistics helps in providing a better understanding and exact description of a phenomenon of nature.
- Statistics helps in presenting complex data in a suitable tabular, diagrammatic and graphic form for an easy and clear comprehension of the data.
- Statistics helps in understanding the nature and pattern of variability of a phenomenon through quantitative observations.
- Statistics helps in drawing valid inference, along with a measure of their reliability about the population parameters from the sample data.
- Statistics plays a vital role in every fields of human activity. Statistics has important role in determining the existing position of per capita income, unemployment, population growth rate, housing, schooling medical facilities etc, in a country.
- The process of arranging data into homogenous group or classes according to some common characteristics present in the data is called classification of data.
- Secondary Data are those which are already collected by other persons or agencies and the investigator just uses them. The secondary data are the second hand information which are already collected by someone (organization) for some purpose and are available for the present study.
- A frequency distribution is a tabular arrangement of data into classes according to the size or magnitude along with corresponding class frequencies (the number of values fall in each class).

ANSWERS TO CHECK YOUR PROGRESS

1. "Statistics are the classified facts representing the conditions of the people in a state especially those facts which can be stated in numbers or any tabular or classified arrangement."
2. Statistics may be divided into two main branches:
   (1) Descriptive Statistics
   (2) Inferential Statistics
3. The process of arranging data into homogenous group or classes according to some common characteristics present in the data is called classification.
4. The methods commonly used for the collection of primary data are as follows:
   - Direct Personal Investigation,
   - Indirect Oral Investigation
   - Data collected though Correspondents or Local Sources
   - Investigation through Schedules and Questionnaires filled in by informants
   - Information through Questionnaires
   - filled in by the enumerators
   - Investigation through Registration Method

TEST YOURSELF

1) Explain the term 'Statistics'. What are its functions and characteristics?
2) Explain kinds or branches of Statistics.
3) What is the importance of Statistics in different fields?
4) What do you mean by Primary and Secondary data?
5) What are the methods of collecting Primary and Secondary Data?

2 FREQUENCY DISTRIBUTION AND CENTRAL TENDENCY

Chapter Includes :

- INTRODUCTION
- FREQUENCY DISTRIBUTION
- CENTRAL TENDENCY
- MEASURES OF CENTRAL TENDENCIES
- MEAN
- ARITHMETIC MEAN
- GEOMETRIC MEAN
- HARMONIC MEAN
- CONCEPT OF MODE
- MEDIAN
- QUARTILES
- DECILES
- PERCENTILES

Learning Objective :

After going through this chapter, you should be able to:

- Discuss the term Frequency Distribution.
- Understand the concept of central tendency.
- Define mean, mode and median
- Describe quartiles, deciles and percentiles
- Explain weighted harmonic mean and geometric mean

NOTES

INTRODUCTION

A modern student of statistics is mainly interested in the study of variability and uncertainty. We live in a changing world. Changes are taking place in every sphere of life. A man of statistics does not show much interest in those things which are constant. The total area of the earth may not be very important to a research minded person but the area under different crops, area covered by forests, area covered by residential and commercial buildings are figures of great importance because these figures keep on changing from time to time and from place to place. Very large number of experts is engaged in the study of changing phenomenon. Experts working in different countries of the world keep a watch on forces which are responsible for bringing changes in the fields of human interest. The agricultural, industrial and mineral production and their transportation from one part to the other parts of the world are the matters of great interest to the economists, statisticians, and other experts. The changes in human population, the changes in standard living, and changes in literacy rate and the changes in price attract the experts to make detailed studies about them and then correlate these changes with the human life. Thus variability or variation is something connected with human life and its study is very important for mankind.

Data can be summarized numerically also. Here we use summary measures like measures of central tendency (such as Arithmetic, Geometric and Harmonic Means, Mode and Median); and measures of dispersion (such as Range, Quartile Deviation, Mean Deviation, and Standard Deviation); measures of association in bivariate analysis (such as Correlation and Regression), Index Numbers, etc are discussed in next chapters. In this chapter we plan to discuss measures of central tendency. The classified and tabulated data need to be summarized to a few summary measures, which can reveal their basic features. One of the essential and important summary measures in statistics is average.

"A single value which can represent the whole set of data is called an average". If the average tends to lie or indicating the center of the distribution is called measure of central tendency or sometimes they locate the general position of the data, so they are also called measure of location.

According to Clark and Schakade, "Average is an attempt to find one single figure to describe whole group of figures."

According to Croxtn and Cowden, "An average is a single value within the range of the data that is used to represent all of the values in the series."

FREQUENCY DISTRIBUTION

A frequency distribution is a tabular arrangement of data into classes according to the size or magnitude along with corresponding class frequencies (the number of values fall in each class).

Ungrouped Data or Raw Data: Data which have not been arranged in a systemic order is called ungrouped or raw data.

Grouped Data: Data presented in the form of frequency distribution is called grouped data.

Array:

The numerical raw data is arranged in ascending or descending order is called an array. It is an arrangement of given raw data in ascending or descending order. In the ascending order, the observations are arranged in increasing order of magnitude. For example, numbers 3,5,7,8,9,10 are arranged in ascending order. In descending order, it is the reverse. For example, the numbers 10,9,8,7,6,5,3 are in descending order.

Class Limits:

The variant values of the classes or groups are called the class limits. The smaller value of the class is called lower class limit and larger value of the class is called upper class limit. Class limits are also called inclusive classes.

For Example: Let us take the class 10 - 19, the smaller value 10 is lower class limit and larger value 19 is called upper class limit.

Class Boundaries:

The true values, which describe the actual class limits of a class, are called class boundaries. The smaller true value is called the lower class boundary and the larger true value is called the upper class boundary of the class. It is important to note that the upper class boundary of a class coincides with the lower class boundary of the next class. Class boundaries are also known as exclusive classes.

For Example:

Weights in Kg	No of Students
60 – 65	8
65 – 70	12
70 – 75	5
	25

A student whose weights are between 60kg and 64.5kg would be included in the 60 - 65 class. A student whose weight is 65kg would be included in next class 65 - 70.

Open-end Classes:

A class has either no lower class limit or no upper class limit in a frequency table is called an open-end class. We do not like to use open-end classes in practice, because they create problems in calculation.

For Example:

Weights (Pounds)	No of Persons
Below – 110	6
110 – 120	12
120 – 130	20
130 – 140	10
140 – Above	2

NOTES

Class Mark or Mid Point:

The class marks or mid point is the mean of lower and upper class limits or boundaries. So it divides the class into two equal parts. It is obtained by dividing the sum of lower and upper class limit or class boundaries of a class by 2. For Example: The class mark or mid-point of the class 60 - 69 is 60+69/2 = 64.5

Size of Class Interval:

The difference between the upper and lower class boundaries (not between class limits) of a class or the difference between two successive mid points is called size of class interval.

Frequency Distribution of Discrete Data:

Discrete data is generated by counting; each and every observation is exact. When an observation is repeated, it is counted the number for which the observation is repeated is called frequency of that observation. The class limits in discrete data are true class limit; there are no class boundaries in discrete data.

Example:

The following are the number of female employees in different branches of commercial banks. Make a frequency distribution.

2, 4, 6, 1, 3, 5, 3, 7, 8, 6, 4, 7, 4, 4, 2, 1, 3, 6, 4, 2, 5, 7, 9, 1, 2, 10, 1, 8, 9, 2, 3, 1, 2, 3, 4, 4, 4, 6, 6, 5, 5, 4, 5, 8, 5, 4, 3, 3, 2, 5, 0, 5, 9, 9, 8, 10, 0, 4, 10, 10, 1, 1, 2, 2, 1, 8, 6, 9, 10

Solution:

The involved variable is "the number of female employees" which is a discrete variable. The largest and smallest values of the given data are 10 and 0 respectively.

Number of Employees (Classes) x	Tally Marks	Branches (Frequency) f
0	II	2
1	𝍸 III	8
2	𝍸 IIII	9
3	𝍸 II	7
4	𝍸 𝍸 I	11
5	𝍸 III	8
6	𝍸 I	6
7	III	3
8	𝍸	5
9	𝍸	5
10	𝍸	5

Continuous or Grouped Frequency Distribution

Numbers like 1, 2, 3, 4, 5, 20, 40, etc. are discrete numbers and are used where no value between the two consecutive numbers is possible. As in the case of the number

NOTES

of children, it will be impossible as well as funny to say that a particular family has 2.083 or 2.1 or 2.75 number of children. The family can have either 2 or 3 children and not a fraction in between. In this Sub-section we propose to illustrate the construction of continuous or grouped frequency distribution from the raw data of Table on monthly income of the 50 families.

To construct a grouped frequency distribution, the range of the given data, i.e., the difference of the highest and the lowest observations is divided into various mutually exclusive and exhaustive sub-intervals, also shown as class-intervals. The frequency of each class interval is then counted and written against it.

Frequency Distribution of Monthly Income of Families

Monthly Income (Rs.)	*Tally Sheet*	*No. of Families (Frequency)*
500 - 550	𝍸	5
550 - 600	𝍸 I	6
600 - 650	𝍸 𝍸	10
650 - 700	𝍸 𝍸 II	12
700 - 750	𝍸 IIII	9
750 - 800	𝍸	5
800 - 850	III	3
Total		50

In this Table we have completed an exercise where the variable "income of the family" has been grouped in order to reduce it to a manageable form called grouped data or Continuous Frequency Distribution. However, prior to the construction of any grouped frequency distribution, it is very important to find answers to the following questions:

1) What should be the number of class intervals?

2) What should be the width of each class interval?

3) How will the class limits be designated?

1) What should be the 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 too small nor too large. If the number of classes is too small, i.e., width of each class is large, there is likelihood of greater loss of information due to grouping. On the other hand, if the number class is very large, the distribution may appear to be too fragmented and may not reveal any pattern of behaviour of the variable. Based on experience, it has been observed that 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.

Usually the formula to determine the number of classes is given by

Number of classes = $1 + 3322 \times \log_{,,} N$,

where N is the total number of observations.

NOTES

In our example of raw data on incomes of 50 families, the number of classes can be calculated as under:

Number of classes = 1 + 3.322 x log10 50 = 1 + 3.322 x 1.6990

= 1 + 5.644 = 6.644 = 7.

2) **What should be the width of each class interval?**

As far as possible, all the class intervals should be of equal width. However, when a frequency distribution, based on equal class intervals, does not reveal a regular pattern of behaviour of observations, it might become necessary to re-group the observations into class intervals of unequal width. By a regular pattern of behavior we mean that there are no classes, with possible exclusion of extreme classes, where there are nil or very few observations while there is concentration of observations in their adjoining classes.

The approximate width of a class can be determined by the following formula:

$$\text{Width of a Class} = \frac{\text{Largest Observation - Smallest Observation}}{\text{Number of Class Intervals}}$$

However, the final decision, regarding width of class intervals, should also take into account the following points.

i) As far as possible, the width should be a multiple of 5, because it is easy to grasp numbers like 5, 10, 15, ... etc.

ii) It should be convenient to find the mid-value of a class.

iii) The observations in a class should be uniformly distributed.

3) **How will the class limits be designated?**

The smallest and the largest observations of a class interval are known as class limits. These are also termed as the lower and upper limits of a class, respectively. Since the mid-value of a class, which is used to compute mean, standard deviation, etc., is obtained from the class limits, it is necessary to define these limits in an unambiguous manner. The following points should be kept in mind while defining class limits:

a) It is not necessary that the lower limit of the first class be 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 may be greater than or equal to the largest observation of the data.

b) It is convenient to have the lower limit of a class either equal to zero or some multiple of 5 or 10.

c) The chosen class limits should be such that the observations in a class are uniformly distributed.

The class limits can be defined in either of the following methods:

NOTES

i) Exclusive Method, and ii) Inclusive Method

i) **Exclusive Method:** In this method, the upper limit of a class is taken to be equal to the lower limit of the following class. In order to keep various class intervals as mutually exclusive, it is decided that 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, the class 500 - 550 shall include all observations with magnitude greater than or equal to 500 but less than 550. An observation with magnitude equal to 550 will be included in the next class, i.e., the class 550 - 600.

The major benefit of exclusive class intervals is that it ensures continuity of data because the upper limit of one class is the lower limit of the next class. In our example on monthly income (Previous Table), there are 5 families whose income lies between Rs. 500 to Rs. 550, i.e., Rs. 500 to 549 and 6 families whose income lies between Rs. 550 to Rs. 600, i.e., Rs. 550 to 599, and so on. Based on this presumption we can rewrite this frequency distribution in the form of given table also.

Exclusive Class Intervals

Monthly Income (Rs.)	*Number of Families* (Frequency)
500 but less than 550	5
550 but less than 600	6
600 but less than 650	10
650 but less than 700	12
700 but less than 750	9
750 but less than 800	5
800 but less than 850	3
Total	50

ii) **Inclusive Method:** In this method, all the observations with magnitude greater than or equal to the lower limit but less than or equal to the upper limit of a class is included in it. Now observe given table. Income of Rs. 549 is included in the class 500 to 549 so that an income of Rs. 550 automatically goes to the next class of 550 to 599. Since the upper limit of one class is not equal to the lower limit of the following class, this saves us from the confusion whether Rs. 550 goes to (500 to 549) or (550 to 599) class.

Inclusive Class Intervals

Monthly Income (Rs.)	*Number of Families* (Frequency)
500 - 549	5
550 - 599	6
600 - 649	10
650 - 699	12
700 - 749	9
750 - 799	5
800 - 849	3
Total	50

NOTES

The choice between exclusive and inclusive methods depends upon whether we are dealing with continuous variable like income, heights, weights, etc. or a discrete variable like number of children in a family. For a continuous variable it is desirable to construct frequency distribution by the exclusive method because, as we have seen earlier, it ensures continuity. For a discrete variable like number of children in a family or number of students getting first division, the frequency distributions should be constructed by using inclusive type of class intervals.

Class Boundaries of Inclusive Class Intervals

Monthly Income (Rs.)	*Number of Families* (Frequency)
499.5 - 549.5	5
549.5 - 599.5	6
599.5 - 649.5	10
649.5 - 699.5	12
699.5 - 749.5	9
749.5 - 799.5	5
799.5 - 849.5	3
Total	50

Mid-Value of a Class

In exclusive type of class intervals, the mid-value or class mark of a class is defined as the arithmetic mean of its lower and upper limits. However, in case of inclusive class intervals, there is a gap between the upper limit of a class and the lower limit of the following class. This gap is eliminated by adding half of the gap to the upper limit and subtracting half of the gap from the lower limit. The new class limits, thus obtained, are known as class boundaries.

Various Forms of Frequency Distributions

Here we propose to introduce the meaning of the following frequency distributions:

a) Open End Frequency Distribution

b) Frequency Distribution with Unequal Class Width

c) Cumulative Frequency Distribution

d) Relative Frequency Distribution

a) Open End Frequency Distribution

Open-end frequency distribution is one which has at least one of its ends open. Either the lower limit of the first class or upper limit of the last class or both are not specified. The words "below" or "less than" and "above" or "more than" are used. In the former the value extends to - ? and in the latter to + ?. Example of such a frequency distribution is given in following table.

NOTES

Class	Frequency
Below 25	1
25 - 30	3
30 - 40	5
40 - 50	2
50 and above	1
Total	12

b) A Frequency Distribution with Unequal Class Width

The classes of a frequency distribution may or may not be of equal width. A frequency distribution with unequal class width is reproduced in following table. Here, the width of lst, 2nd and 5th classes is 5, while that of 3rd is 10 and that of 4th is 15.

Class	Frequency
20 - 25	1
25 - 30	3
30 - 40	5
40 - 55	2
55 - 60	1
Total	12

c) Cumulative Frequency Distribution

Suppose that, with reference to data given in Table 3.6, we ask the following questions:

i) How many families have their monthly income less than or equal to Rs. 700?

ii) How many families have their monthly income greater than or equal to Rs. 600?

The answers to the above questions can be easily obtained by forming an appropriate cumulative frequency distribution. To answer that question, we need to form a "less than type" cumulative frequency distribution while a "great & than type" cumulative frequency distribution is required for answering the second question. These distributions are given in Tables 1 and 2 respectively.

NOTES

Table 1

"Less-than type" Cumulative Frequency Distribution

Monthly Income (Rs.)	Frequencies		
	Simple		Cumulative
Less than 550	5		5
Less than 600	6	5+6	11
Less than 650	10	5+6+10	21
Less than 700	12	5+6+10+12	33
Less than 750	9	5+6+10+12+9	42
Less than 800	5	5+6+10+12+9+5	47
Less than 850	3	5+6+10+12+9+5+3	50

Table 2

"More-than type" Cumulative Frequency Distribution

Monthly Income (Rs.)	Frequencies		
	Simple		Cumulative
More than 500	5	3+5+9+12+10+6+5	50
More than 550	6	3+5+9+12+10+6	45
More than 600	10	3+5+9+12+10	39
More than 650	12	3+5+9+12	29
More than 700	9	3+5+9	17
More than 750	5	3+5	8
More than 800	3		3

d) Relative Frequency Distribution

So far we have expressed the frequency of a value or that of a class as the number of times an observation is repeated. We can also express these frequencies as a fraction or a percentage of the total number of observations. Such frequencies are known as the relative frequencies.

Relative Frequency Distribution of Monthly Income of 50 Families

Class	Frequency	Relative Frequency	
		As a fraction	As a percentage
500 - 549	5	5 ÷ 50 = 0.10	0.10 × 100 = 10
550 - 599	6	6 ÷ 50 = 0.12	0.12 × 100 = 12
600 - 649	10	10 ÷ 50 = 0.20	0.20 × 100 = 20
650 - 699	12	12 ÷ 50 = 0.24	0.24 × 100 = 24
700 - 749	9	9 ÷ 50 = 0.18	0.18 × 100 = 18
750 - 799	5	5 ÷ 50 = 0.10	0.10 × 100 = 10
800 - 849	3	3 ÷ 50 = 0.06	0.06 × 100 = 6
Total	50	1	100

From the above table it is clear that sum of the relative frequencies should be either 1 (in case of fraction) or 100 (in case of percentage).

CENTRAL TENDENCY

NOTES

One of the main characteristics of numerical data is central tendency. It is found that the observations tend to cluster around a point. The point, around which the observation concentrate (or cluster) lies in general, is the central part of the data. This point is called the central point or the central value of the data.

MEASURES OF CENTRAL TENDENCIES

The term central tendency refers to the "middle" value or perhaps a typical value of the data, and is measured using the mean, median, or mode. Each of these measures is calculated differently, and the one that is best to use depends upon the situation.

According to Simpson and Kafka, "A measure of central tendency is a typical value around which other figures congregate." Thus in brief, we have, "A measure of central tendency is a simple value within the range of the entire mass of data that is used to represent the whole data."

Desirable Qualities / Characteristics of a Good Average (or Measure of Central Tendency):

An average possesses all or most of the following qualities (characteristics) are considered a good average:

- It should be easy to calculate and simple to understand.
- It should be clearly defined by a mathematical formula.
- It should not be affected by extreme values.
- It should be based on all the observations.
- It should be capable of further mathematical treatment.
- It should have sample stability.

MEAN

The mean is the most commonly-used measure of central tendency. When we talk about an "average", we usually are referring to the mean. The mean is simply the sum of the values divided by the total number of items in the set. The result is referred to as the arithmetic mean. Sometimes it is useful to give more weighting to certain data points, in which case the result is called the weighted arithmetic mean.

Check your progress:

1) What is frequency distribution?
2) Define Central Tendency?

Types of Averages

Mathematical averages are:

- Arithmetic Mean
- Geometric Mean
- Harmonic Mean
- Median
- Mode

NOTES

ARITHMETIC MEAN

It is the most commonly used average or measure of the central tendency applicable only in case of quantitative data. Arithmetic mean is also simply called "mean". Arithmetic mean is defined as:

"Arithmetic mean is quotient of sum of the given values and number of the given values".

The arithmetic mean can be computed for both ungroup data (raw data: a data without any statistical treatment) and grouped data (a data arranged in tabular form containing different groups).

If X is the involved variable, then arithmetic mean of X is abbreviated as A.M. of X and denoted by . The arithmetic mean of X can be computed by any of the following methods.

Method's Name	Nature of Data	
	Ungrouped Data	Grouped Data
Direct Method	$A.M \text{ of } X = \bar{X} = \frac{\Sigma x}{n}$	$A.M \text{ of } X = \bar{X} = \frac{\Sigma fx}{\Sigma f}$
Indirect or Short-Cut Method	$A.M \text{ of } X = \bar{X} = A + \frac{\Sigma D}{n}$	$A.M \text{ of } X = \bar{X} = A + \frac{\Sigma fD}{\Sigma f}$
Method of Step-Deviation	$A.M \text{ of } X = \bar{X} = A + \frac{\Sigma u}{n} \times c$	$A.M \text{ of } X = \bar{X} = A + \frac{\Sigma fu}{\Sigma f} \times h$

Where:

Indicates values of the variable X.

Indicates number of values of X.

Indicates frequency of different groups.

Indicates assumed mean

Indicates deviation from i.e,

Step-deviation and

Indicates common divisor

Indicates size of class or class interval in case of grouped data

Summation or addition

Example (1):

The one-sided train fare of five selected BS students is recorded as follows: 10, 5, 15, 8 and 12. Calculate arithmetic mean of the following data.

Solution:

Let train fare is indicated by x, then

x
10
5
15
8
12
$\Sigma x = 50$

Arithmetic mean of

$$X = \overline{X} = \frac{\Sigma x}{n}$$

We decide to use above-mentioned formula. Form the given data, we have ?x = 50 and n = 5. Placing these two quantities in above formula, we get the arithmetic mean for given data.

$$\overline{X} = \frac{50}{5} = 10$$

Example (2):

Given the following frequency distribution of first year students of a particular college. Calculate arithmetic mean of the following data.

Age (Years)	13	14	15	16	17
Number of Students	2	5	13	7	3

Solution:

The given distribution belongs to a grouped data and the variable involved is ages of first year students. While the number of students Represent frequencies.

Ages (Years) x	Number of Students f	fx
13	2	26
14	5	70
15	13	195
16	7	112
17	3	51
Total	$\Sigma f = 30$	$\Sigma fx = 454$

Now we will find the Arithmetic Mean as

NOTES

= 15.13 years

Example (3):

The following data shows distance covered by 100 persons to perform their routine jobs. Calculate arithmetic mean of the following data.

Distance (Km)	0 – 10	10 – 20	20 – 30	30 – 40
Number of Persons	10	20	40	30

Solution:

The given distribution belongs to a grouped data and the variable involved is ages of "distance covered". While the "number of persons" Represent frequencies.

Distance (Km)	Number of Persons f	Mid Points x	fx
0 – 10	10	5	50
10 – 20	20	15	300
20 – 30	40	25	1000
30 – 40	30	35	1050
Total	$\Sigma f = 100$		$\Sigma fx = 2400$

Now we will find the Arithmetic Mean as

$$\overline{X} = \frac{\Sigma fx}{\Sigma f} = \frac{2400}{100} = 24$$

= 24 Km.

Example (4):

The following data shows distance covered by 100 persons to perform their routine jobs.

Distance (Km)	0 – 10	10 – 20	20 – 30	30 – 40
Number of Persons	10	20	40	30

Calculate Arithmetic Mean by Step-Deviation Method; also explain why it is better than direct method in this particular case.

NOTES

Solution:

The given distribution belongs to a grouped data and the variable involved is ages of "distance covered". While the "number of persons" Represent frequencies.

Distance Covered in (Km)	Number of Persons f	Mid Points x	$u=\left(\frac{x-5}{10}\right)$	fu
0 – 10	10	5	–1	–10
10 – 20	20	15	0	0
20 – 30	40	25	+1	40
30 – 40	30	35	+2	60
Total	$\Sigma f = 100$			$\Sigma fu = 90$

Now we will find the Arithmetic Mean as

$$\bar{X} = A + \frac{\Sigma fu}{\Sigma f} \times h$$

Where

$A = 15$

$\Sigma fu = 90$

$\Sigma f = 100$

$h = 10$

$$\bar{X} = 15 + \frac{90}{100} \times 10 = 24$$

$= 24$ Km

Explanation:

Here from the mid points ( ) it is very much clear that each mid point is multiple of 5 and there is also a gap of 10 from mid point to mid point i.e. class size or interval ( ). Keeping in view this, we should prefer to take method of Step-Deviation instead of Direct Method.

Example 5:

The following frequency distribution shows the marks obtained by 50 students in statistics at a certain college. Find the arithmetic mean using (1) Direct Method (2) Short-Cut Method (3) Step-Deviation.

Marks	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89
Frequency	1	5	12	15	9	6	2

NOTES

Solution:

			Direct Method	Short-Cut Method		Step-Deviation Method	
Marks	f	x	fx	$D = x - A$	fD	$u = \frac{x-A}{h}$	fu
20 – 29	1	24.5	24.5	–30	–30	–3	–3
30 – 39	5	34.5	172.5	–20	–100	–2	–10
40 – 49	12	44.5	534.5	–10	–120	–1	–12
50 – 59	15	54.5	817.5	0	0	0	0
60 – 69	9	64.5	580.5	10	90	1	9
70 – 79	6	74.5	447.5	20	120	2	12
80 – 89	2	84.5	169.5	30	60	3	6
Total	50		2745		20		2

(1) Direct Method:

$$\bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{2745}{50} = 54.9$$

= 55 Marks

(2) Short-Cut Method:

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

$$A = 54.5$$

$$= 54.5 + \frac{20}{50} = 54.5 + 0.4 = 54.9$$

= 55 Marks

(3) Step-Deviation Method:

$$\bar{X} = A + \frac{\Sigma fu}{\Sigma f} \times h$$

$$A = 54.5$$

$$h = 10$$

$$= 54.5 + \frac{2}{50} \times 10$$

$$= 54.5 + 0.4 = 54.9$$

= 55 Marks

Weighted Arithmetic Mean

In calculation of arithmetic mean, the importance of all the items was considered to be equal. However, there may be situations in which all the items under consid-

erations are not equal importance. For example, we want to find average number of marks per subject who appeared in different subjects like Mathematics, Statistics, Physics and Biology. These subjects do not have equal importance. If we find arithmetic mean by giving Mean.

Thus, *arithmetic mean computed by considering relative importance of each items is called weighted arithmetic mean.* To give due importance to each item under consideration, we assign number called weight to each item in proportion to its relative importance.

Weighted Arithmetic Mean is computed by using following formula:

$$\overline{X}_w = \frac{\sum wx}{\sum w}$$

Where:

$\overline{X}_w$ = Stands for weighted arithmetic mean.

x = Stands for values of the items and

w = Stands for weight of the item

Example:

A student obtained 40, 50, 60, 80, and 45 marks in the subjects of Math, Statistics, Physics, Chemistry and Biology respectively. Assuming weights 5, 2, 4, 3, and 1 respectively for the above mentioned subjects. Find Weighted Arithmetic Mean per subject.

Solution:

Subjects	Marks Obtained x	Weight w	wx
Math	40	5	200
Statistics	50	2	100
Physics	60	4	240
Chemistry	80	3	240
Biology	45	1	45
Total		$\sum w = 15$	$\sum wx = 825$

Now we will find weighted arithmetic mean as:

$$\overline{X}_w = \frac{\sum wx}{\sum w} = \frac{825}{15} = 55$$

= 55 marks/subject.

Merits and Demerits of Arithmetic Mean

Merits:

- It is rigidly defined.

NOTES

- It is easy to calculate and simple to follow.
- It is based on all the observations.
- It is determined for almost every kind of data.
- It is finite and not indefinite.
- It is readily put to algebraic treatment.
- It is least affected by fluctuations of sampling.

Demerits:

- The arithmetic mean is highly affected by extreme values.
- It cannot average the ratios and percentages properly.
- It is not an appropriate average for highly skewed distributions.
- It cannot be computed accurately if any item is missing.
- The mean sometimes does not coincide with any of the observed value

GEOMETRIC MEAN

It is another measure of central tendency based on mathematical footing like arithmetic mean. Geometric mean can be defined in the following terms:

"Geometric mean is the nth positive root of the product of "n" positive given values"

Hence, geometric mean for a value X containing n values such as $x_1, x_2, x_3, \ldots, x_n$ is denoted by $G.M$ of X and given as under (For Ungrouped Data)

$$G.M \text{ of } X = \overline{X} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \cdots \cdot x_n}$$

If we have a series of n positive values with repeated values such as $x_1, x_2, x_3, \ldots, x_k$ are repeated times respectively then geometric mean will becomes (For Grouped Data)

$$G.M \text{ of } X = \overline{X} = \sqrt[n]{x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \cdot \cdots \cdot x_k^{f_k}}$$

Where

$$n = f_1 + f_2 + f_3 + \cdots + f_k$$

Example:

Find the Geometric Mean of the values 10, 5, 15, 8, 12

Solution:

Here

$$x_1 = 10 \; x_2 = 5 \; x_3 = 15 \; x_4 = 8 \; x_5 = 12$$

NOTES

And n = 5

$$G.M \text{ of } X = \overline{X} = \sqrt[5]{10 \times 5 \times 15 \times 8 \times 12}$$

$$\overline{X} = \sqrt[5]{72000} = (72000)^{\frac{1}{5}} = 9.36$$

Example:

Find the Geometric Mean of the following Data:

X	13	14	15	16	17
f	2	5	13	7	3

Solution:

We may write it as given below:

Here $x_1 = 13\ x_2 = 14\ x_3 = 15\ x_4 = 16\ x_5 = 17$

$f_1 = 2\ f_2 = 5\ f_3 = 13\ f_4 = 7\ f_5 = 3$

$$n = \Sigma f = f_1 + f_2 + f_3 + f_4 + f_5 = 2 + 5 + 13 + 7 + 3 = 30$$

Using the formula of geometric mean for grouped data, geometric mean in this case will become:

$$G.M \text{ of } X = \overline{X} = \sqrt[n]{x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \cdot x_4^{f_4} \cdot x_5^{f_5}}$$

$$\overline{X} = \sqrt[30]{(13)^2 \cdot (14)^5 \cdot (15)^{13} \cdot (16)^7 \cdot (17)^3}$$

$$\overline{X} = \sqrt[30]{2.33292 \times 10^{35}} = (2.33292 \times 10^{35})^{\frac{1}{30}}$$

$$\overline{X} = 15.0984 \approx 15.10$$

The method explained above for the calculation of geometric mean is useful when the numbers of values in given data are small in number and the facility of electronic calculator is available. When a set of data contains large number of values then we need an alternative way for computing geometric mean. The modified or alternative way of computing geometric mean is given as under:

For Ungrouped Data	For Grouped Data
$G.M \text{ of } X = \overline{X} = Anti\log\left(\frac{\Sigma \log x}{n}\right)$	$G.M \text{ of } X = \overline{X} = Anti\log\left(\frac{\Sigma f \log x}{\Sigma f}\right)$

NOTES

Example:

Find the Geometric Mean of the values 10, 5, 15, 8, 12

x	$\log x$
10	1.0000
5	0.6990
15	1.1761
8	0.9031
12	1.0792
Total	$\Sigma \log x = 4.8573$

$$G.M \text{ of } X = \overline{X} = Anti\log\left(\frac{\Sigma \log x}{n}\right)$$

$$\overline{X} = Anti\log\left(\frac{4.8573}{5}\right)$$

$$\overline{X} = Anti\log(0.9715)$$

$$\overline{X} = 9.36$$

Example:

Find the Geometric Mean for the following distribution of students' marks:

Marks	0 – 30	30 – 50	50 – 80	80 – 100
No. of Students	20	30	40	10

Solution:

Marks	**No. of Students** f	**Mid Points** x	$f \log x$
0 – 30	20	15	$20 \log 15 = 23.5218$
30 – 50	30	40	$30 \log 40 = 48.0168$
50 – 80	40	65	$40 \log 65 = 72.5165$
80 – 100	10	90	$10 \log 90 = 19.5424$
Total	$\Sigma f = 100$		$\Sigma f \log x = 163.6425$

NOTES

$$\overline{X} = Anti\log\left(\frac{163.6425}{100}\right)$$

$$\overline{X} = Anti\log(1.6364)$$

$$\overline{X} = 43.29$$

Properties of Geometric Mean:

The main properties of geometric mean are:

- The geometric mean is less than arithmetic mean, G.M<A.M
- The product of the items remains unchanged if each item is replaced by the geometric mean.
- The geometric mean of the ratio of corresponding observations in two series is equal to the ratios their geometric means.
- The geometric mean of the products of corresponding items in two series is equal to the product of their geometric mean.

Merits:

- It is rigidly defined and its value is a precise figure.
- It is based on all observations.
- It is capable of further algebraic treatment.
- It is not much affected by fluctuation of sampling.
- It is not affected by extreme values.

Demerits:

- It cannot be calculated if any of the observation is zero or negative.
- Its calculation is rather difficult.
- It is not easy to understand.
- It may not coincide with any of the observations.

HARMONIC MEAN

Harmonic mean is another measure of central tendency and also based on mathematic footing like arithmetic mean and geometric mean. Like arithmetic mean and geometric mean, harmonic mean is also useful for quantitative data. Harmonic mean is defined in following terms:

Harmonic mean is quotient of "number of the given values" and "sum of the reciprocals of the given values".

Harmonic mean in mathematical terms is defined as follows:

For Ungrouped Data	For Grouped Data
$H.M \text{ of } X = \bar{X} = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$	$H.M \text{ of } X = \bar{X} = \frac{\Sigma f}{\Sigma\left(\frac{f}{x}\right)}$

Example:

Calculate the harmonic mean of the numbers: 13.5, 14.5, 14.8, 15.2 and 16.1

Solution:

The harmonic mean is calculated as below:

x	$\frac{1}{x}$
13.2	0.0758
14.2	0.0704
14.8	0.0676
15.2	0.0658
16.1	0.0621
Total	$\Sigma\left(\frac{1}{x}\right) = 0.3417$

$$H.M \text{ of } X = \bar{X} = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$$

$$H.M \text{ of } X = \bar{X} = \frac{5}{0.3417} = 14.63$$

Example:

Given the following frequency distribution of first year students of a particular college. Calculate the Harmonic Mean.

Age (Years)	13	14	15	16	17
Number of Students	2	5	13	7	3

Solution:

The given distribution belongs to a grouped data and the variable involved is ages of first year students. While the number of students Represent frequencies.

Ages (Years) x	Number of Students f	$\frac{1}{x}$
13	2	0.1538
14	5	0.3571
15	13	0.8667
16	7	0.4375
17	3	0.1765
Total	$\Sigma f = 30$	$\Sigma\left(\frac{1}{x}\right) = 1.9916$

Now we will find the Harmonic Mean as

$$\overline{X} = \frac{\Sigma f}{\Sigma\left(\frac{f}{x}\right)} = \frac{30}{1.9916} = 15.0631 \approx 15$$

= 15 years

Example:

Calculate the harmonic mean for the given below:

Marks	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
f	2	3	11	20	32	25	7

Solution:

The necessary calculations are given below:

Marks	x	f	$\frac{f}{x}$
30 – 39	34.5	2	0.0580
40 – 49	44.5	3	0.0674
50 – 59	54.5	11	0.2018
60 – 69	64.5	20	0.3101
70 – 79	74.5	32	0.4295
80 – 89	84.5	25	0.2959
90 – 99	94.5	7	0.0741
Total		$\Sigma f = 100$	$\Sigma\left(\frac{f}{x}\right) = 1.4368$

Now we will find the Harmonic Mean as

$$\overline{X} = \frac{\Sigma f}{\Sigma\left(\frac{f}{x}\right)} = \frac{100}{1.4368} = 69.60$$

NOTES

Merits:

- It is based on all observations.
- It not much affected by the fluctuation of sampling.
- It is capable of algebraic treatment.
- It is an appropriate average for averaging ratios and rates.
- It does not give much weight to the large items.

Demerits:

- Its calculation is difficult.
- It gives high weight-age to the small items.
- It cannot be calculated if any one of the items is zero.
- It is usually a value which does not exist in the given data.

CONCEPT OF MODE:

Mode is the value which occur the greatest number of times in the data. When each value occur the same numbers of times in the data, there is no mode. If two or more values occur the same numbers of time, then there are two or more modes and distribution is said to be multi-mode. If the data having only one mode the distribution is said to be uni-model and data having two modes, the distribution is said to be bi-model.

Mode from Ungrouped Data:

Mode is calculated from ungrouped data by inspecting the given data. We pick out that value which occur the greatest numbers of times in the data.

Mode from Grouped Data:

When frequency distribution with equal class interval sizes, the class which has maximum frequency is called model class.

$$Mode = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

Or

$$Mode = l + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

Where

l = Lower class boundary of the model class

f_m = Frequency of the model class (maximum frequency)

f_1 = Frequency preceding the model class frequency

f_2 = Frequency following the model class frequency

h = Class interval size of the model class

Mode from Discrete Data:

When the data follows discrete set of values, the mode may be found by inspection Mode is the value of X corresponding to the maximum frequency.

Example:

Find the mode of the values 5, 7, 2, 9, 7, 10, 8, 5, 7

Solution:

Mode is 7 because it occur the greatest number of times in the data.

Example:

The weights of 50 college students are given in the following table. Find the mode of the distribution.

Weights (Kg)	60 – 64	65 – 69	70 – 74	75 – 79	80 – 84
No of Students	5	9	16	12	8

Solution:

Weights (Kg)	No of Students f	Class Boundary
60 – 64	5	59.5 – 64.5
65 – 69	9	64.5 – 69.5
70 – 74	16	69.5 – 74.5
75 – 79	12	74.5 – 79.5
80 – 84	8	79.5 – 84.5

$$Mode = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

$$= 69.5 + \frac{16 - 9}{(16 - 9) + (16 - 12)} \times 5$$

$$= 69.5 + \frac{7}{7 + 4} \times 5 = 69.5 + 3.18$$

$$Mode = 72.68$$

NOTES

Check your progress:

3) What is arithmetic mean?

4) Define Mean?

NOTES

Example:

The following frequency distribution shows the numbers of children in each family in a locality. Find the mode.

No of Children	0	1	2	3	4	5	6
No of Families	6	30	42	55	25	18	5

Solution:

The data follows discrete set of values

So, Mode = 3 (corresponding to the maximum frequency)

Advantages:

- It is easy to understand and simple to calculate.
- It is not affected by extreme large or small values.
- It can be located only by inspection in ungrouped data and discrete frequency distribution.
- It can be useful for qualitative data.
- It can be computed in open-end frequency table.
- It can be located graphically.

Disadvantages:

- It is not well defined.
- It is not based on all the values.
- It is stable for large values and it will not be well defined if the data consists of small number of values.
- It is not capable of further mathematical treatment.
- Sometimes, the data having one or more than one mode and sometimes the data having no mode at all.

MEDIAN

Median is the most middle value in the arrayed data. It means that when the data are arranged, median is the middle value if the number of values is odd and the mean of the two middle values, if the numbers of values is even. A value which divides the arrayed set of data in two equal parts is called median, the values greater than the median is equal to the values smaller than the median. It is also known as a positional average. It is denoted by $\tilde{X}$ read as X- tilde.

Median from Ungrouped Data:

$$\text{Median} = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

NOTES

Example:

Find the median of the values 4, 1, 8, 13, 11

Solution:

Arrange the data 1, 4, 8, 11, 13

Arrange the data 1, 4, 8, 11, 13

$$\text{Median} = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

$$\text{Median} = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \left(\frac{6}{2}\right) = 3^{rd} \text{ item}$$

Median = 8

Example:

Find the median of the values 5, 7, 10, 20, 16, 12

Solution:

Arrange the data 5, 7, 10, 12, 16, 20

$$\text{Median} = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

$$\text{Median} = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \left(\frac{7}{2}\right) = 3.5^{th} \text{ item}$$

$$\text{Median} = \frac{10+12}{2} = 11$$

Median from grouped Data:

The median for grouped data, we find the cumulative frequencies and then calculated the median number $\frac{n}{2}$. The median lies in the group (class) which corresponds to the cumulative frequency in which $\frac{n}{2}$ lies. We use following formula to find the median.

$$Median = l + \frac{h}{f}\left(\frac{n}{2} - c\right)$$

NOTES

Where l = Lower class boundary of the model class

f = Frequency of the median class

$n = \sum f$ = Number of values or total frequency

c = Cumulative frequency of the class preceding the median class

h = Class interval size of the model class

Example:

Calculate median from the following data.

Group	60 – 64	65 – 69	70 – 74	75 – 79	80 – 84	85 – 89
Frequency	1	5	9	12	7	2

Solution:

Group	f	Class Boundary	Cumulative Frequency
60 – 64	1	59.5 – 64.5	1
65 – 69	5	64.5 – 69.5	6
70 – 74	9	69.5 – 74.5	15
75 – 79	12	74.5 – 79.5	27
80 – 84	7	79.5 – 84.5	34
85 – 89	2	84.5 – 89.5	36

$$Median = l + \frac{h}{f}\left(\frac{n}{2} - c\right) \qquad \because \left(\frac{n}{2}\right)^{th} \text{ item} = \frac{36}{2} = 18^{th} \text{ item}$$

$$= 74.5 + \frac{5}{12}(18 - 15) = 74.5 + \frac{5}{12}(3)$$

$$Median = 74.5 + 1.25 = 75.75$$

Median from Discrete Data:

When the data follows the discrete set of values grouped by size, we use the formula $\left(\frac{n+1}{2}\right)^{th}$ item for finding median. First we form a cumulative frequency distribution and median is that value which corresponds to the cumulative frequency in which $\left(\frac{n+1}{2}\right)^{th}$ item lies.

Example:

The following frequency distribution is classified according to the number of leaves on different branches. Calculate the median number of leaves per branch.

No of Leaves	1	2	3	4	5	6	7
No of Branches	2	11	15	20	25	18	10

Solution:

No of Leaves X	No of Branches f	Cumulative Frequency $C.F$
1	2	2
2	11	13
3	15	28
4	20	48
5	25	73
6	18	91
7	10	101
Total	101	

$$\text{Median} = \text{Size of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

$$= \frac{101+1}{2} = \frac{102}{2} = 51^{st} \text{ item}$$

Median = 5 because 51^{st} item lies corresponding to 5

QUARTILES

There are three quartiles called, first quartile, second quartile and third quartile. There quartiles divides the set of observations into four equal parts. The second quartile is equal to the median. The first quartile is also called lower quartile and is denoted by . The third quartile is also called upper quartile and is denoted by . The lower quartile is a point which has 25% observations less than it and 75% observations are above it. The upper quartile is a point with 75% observations below it and 25% observations above it.

Quartile for Individual Observations (Ungrouped Data):

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)^{th} \text{ item}$$

$$Q_2 = \text{Value of } 2\left(\frac{n+1}{4}\right)^{th} \text{ item} = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = Median$$

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

NOTES

Quartile for a Frequency Distribution (Discrete Data):

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)^{th} \text{ item} \qquad (n = \Sigma f)$$

$$Q_2 = \text{Value of } 2\left(\frac{n+1}{4}\right)^{th} \text{ item} = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = Median$$

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

Quartile for Grouped Frequency Distribution:

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right) \qquad (n = \Sigma f)$$

$$Q_2 = l + \frac{h}{f}\left(\frac{2n}{4} - c\right) = l + \frac{h}{f}\left(\frac{n}{2} - c\right) = Median$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

Example:

The wheat production (in Kg) of 20 acres is given as: 1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750, and 1885. Find the quartile deviation and coefficient of quartile deviation.

Solution:

After arranging the observations in ascending order, we get 1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)th \text{ item}$$

$$= \text{Value of } \left(\frac{20+1}{4}\right)th \text{ item}$$

$$= \text{Value of } (5.25)th \text{ item}$$

$$= 5th \text{ item} + 0.25(6th \text{ item} - 5th \text{ item}) = 1240 + 0.25(1320 - 1240)$$

$$Q_1 = 1240 + 20 = 1260$$

$$Q_3 = \text{Value of } \frac{3(n+1)}{4}th \text{ item}$$

NOTES

$= \text{Value of } \frac{3(20+1)}{4} th \text{ item}$

$= \text{Value of } (15.75) th \text{ item}$

$= 15th \text{ item} + 0.75(16th \text{ item} - 15th \text{ item}) = 1750 + 0.75(1755 - 1750)$

$Q_3 = 1750 + 3.75 = 1753.75$

Example:

Calculate the quartile deviation and coefficient of quartile deviation from the data given below:

Maximum Load (short-tons)	Number of Cables
9.3 – 9.7	2
9.8 – 10.2	5
10.3 – 10.7	12
10.8 – 11.2	17
11.3 – 11.7	14
11.8 – 12.2	6
12.3 – 12.7	3
12.8 – 13.2	1

Solution:

The necessary calculations are given below:

Maximum Load (short-tons)	Number of Cables (f)	Class Boundaries	Cumulative Frequencies
9.3 – 9.7	2	9.25 – 9.75	2
9.8 – 10.2	5	9.75 – 10.25	2 + 3 = 7
10.3 – 10.7	12	10.25 – 10.75	7 + 12 = 19
10.8 – 11.2	17	10.75 – 11.25	19 + 17 = 36
11.3 – 11.7	14	11.25 – 11.75	36 + 14 = 50
11.8 – 12.2	6	11.75 – 12.25	50 + 6 = 56
12.3 – 12.7	3	12.25 – 12.75	56 + 3 = 59
12.8 – 13.2	1	12.75 – 13.25	59 + 1 = 60

$Q_1 = \text{Value of } \left(\frac{n}{4}\right) th \text{ item} = \text{Value of } \left(\frac{60}{4}\right) th \text{ item} = 15th \text{ item}$

Q_1 lies in the class $10.25 - 10.75$

$\therefore \; Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$

NOTES

Where $l = 10.25,\ h = 0.5,\ f = 12,\ n/4 = 15$ and

$$\therefore\ Q_1 = 10.25 + \frac{0.5}{12}(15 - 7) = 10.25 + 0.33 = 10.58$$

Q_3 = Value of $\left(\frac{3n}{4}\right)th$ item = Value of $\left(\frac{3\times 60}{4}\right)th$ item=45th item

Q_3 lies in the class $11.25 - 11.75$

$$\therefore\ Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

Where $l = 11.25,\ h = 0.5,\ f = 14,\ 3n/4 = 45$ and

$$\therefore\ Q_1 = 11.25 + \frac{0.5}{14}(45 - 36) = 11.25 + 0.32 = 11.57$$

DECILES:

The deciles are the partition values which divides the set of observations into ten equal parts. There are nine deciles namely $D_1, D_2, D_3, \ldots, D_9$. The first decile is D_1 is a point which has 10% of the observations below it.

Deciles for Individual Observations (Ungrouped Data):

$$D_1 = \text{Value of } \left(\frac{n+1}{10}\right)^{th} \text{item}$$

$$D_2 = \text{Value of } 2\left(\frac{n+1}{10}\right)^{th} \text{item}$$

$$D_3 = \text{Value of } 3\left(\frac{n+1}{10}\right)^{th} \text{item}$$

$$\vdots$$

$$D_9 = \text{Value of } 9\left(\frac{n+1}{10}\right)^{th} \text{item}$$

Quartile for a Frequency Distribution (Discrete Data):

NOTES

$$D_1 = \text{Value of } \left(\frac{n+1}{10}\right)^{th} \text{ item} \qquad (n = \Sigma f)$$

$$D_2 = \text{Value of } 2\left(\frac{n+1}{10}\right)^{th} \text{ item}$$

$$D_3 = \text{Value of } 3\left(\frac{n+1}{10}\right)^{th} \text{ item}$$

$$\vdots$$

$$D_9 = \text{Value of } 9\left(\frac{n+1}{10}\right)^{th} \text{ item}$$

Quartile for Grouped Frequency Distribution:

$$D_1 = l + \frac{h}{f}\left(\frac{n}{10} - c\right) \qquad (n = \Sigma f)$$

$$D_2 = l + \frac{h}{f}\left(\frac{2n}{10} - c\right)$$

$$D_3 = l + \frac{h}{f}\left(\frac{3n}{10} - c\right)$$

$$\vdots$$

$$D_9 = l + \frac{h}{f}\left(\frac{9n}{10} - c\right)$$

PERCENTILES

The percentiles are the points which divide the set of observations into one hundred equal parts. These points are denoted by

$P_1, P_2, P_3, \ldots, P_{99}$, and are called the first, second, third, ..., ninety ninth percentiles. The percentiles are calculated for very large number of observations like workers in factories and the population in provinces or countries. The percentiles are usually calculated for grouped data. The first percentile denoted and calculated as

PERCENTILES

The percentiles are the points which divide the set of observations into one hundred equal parts. These points are denoted by

$P_1, P_2, P_3, \ldots, P_{99}$, and are called the first, second, third, ..., ninety ninth percentiles. The percentiles are calculated for very large number of observations like workers in factories and the population in provinces or countries. The percentiles are usually calculated for grouped data. The first percentile denoted and calculated as

Check your progress:

5) What is dispersion of data?
6) What do you mean by Standard Deviation?

$$P_1 = \text{Value of } \left(\frac{n}{100}\right)^{th} \text{ item}$$

NOTES

We find the group in which the $\left(\frac{n}{100}\right)^{th}$ item lies and then

P_1 is interpolated from the formula.

$$P_1 = l + \frac{h}{f}\left(\frac{n}{100} - c\right) \qquad (n = \Sigma f)$$

$$P_2 = l + \frac{h}{f}\left(\frac{2n}{100} - c\right)$$

$$P_3 = l + \frac{h}{f}\left(\frac{3n}{100} - c\right)$$

$$\vdots$$

$$P_{99} = l + \frac{h}{f}\left(\frac{99n}{100} - c\right)$$

SUMMARY

- A single value which can represent the whole set of data is called an average.
- The term central tendency refers to the "middle" value or perhaps a typical value of the data, and is measured using the mean, median, or mode.
- The mean is the most commonly-used measure of central tendency. When we talk about an "average", we usually are referring to the mean. The mean is simply the sum of the values divided by the total number of items in the set.
- The arithmetic mean can be computed for both ungroup data (raw data: a data without any statistical treatment) and grouped data (a data arranged in tabular form containing different groups).
- Geometric mean is the nth positive root of the product of "n" positive given values.
- Harmonic mean is another measure of central tendency and also based on mathematic footing like arithmetic mean and geometric mean.
- Harmonic mean is quotient of "number of the given values" and "sum of the reciprocals of the given values.
- Mode is the value which occur the greatest number of times in the data. When each value occur the same numbers of times in the data, there is

no mode. If two or more values occur the same numbers of time, then there are two or more modes and distribution is said to be multi-mode. If the data having only one mode the distribution is said to be uni-model and data having two modes, the distribution is said to be bi-model.

- Median is the most middle value in the arrayed data. It means that when the data are arranged, median is the middle value if the number of values is odd and the mean of the two middle values, if the numbers of values is even.
- The quartiles are the points which divide the set of observations into four equal parts.
- The deciles are the points which divide the set of observations into ten equal parts.
- The percentiles are the points which divide the set of observations into one hundred equal parts.

ANSWERS TO CHECK YOUR PROGRESS

1. A frequency distribution is a tabular arrangement of data into classes according to the size or magnitude along with corresponding class frequencies (the number of values fall in each class).
2. The term central tendency refers to the "middle" value or perhaps a typical value of the data, and is measured using the mean, median, or mode.
3. "Arithmetic mean is quotient of sum of the given values and number of the given values".
4. The mean is the most commonly-used measure of central tendency. When we talk about an "average", we usually are referring to the mean. The mean is simply the sum of the values divided by the total number of items in the set.
5. The degree to which numerical data tend to spread about an average value is called the dispersion or variation of the data.
6. The standard deviation plays a dominating role for the study of variation in the data. It is a very widely used measure of dispersion.

TEST YOURSELF

1) What do you mean by Grouped Frequency Distribution?
2) Explain measures of Central Tendencies.
3) What do you mean by Arithmetic Mean? Explain its merits and demerits.
4) Explain Geometric Mean. What are the properties of Geometric Mean?
5) Write a short note on Harmonic Mean.
6) Explain the concept of Mode. What are its advantages and Disadvantages?
7) What do you mean by Median?

NOTES

8) Explain following terms in brief:

i) Percentiles

ii) Deciles

iii) Quartiles

9) Calculate Arithmetic Mean by Step-Deviation Method from the following data shows the marks obtained by 100 students in Mathematics at a certain college.

Marks	20 - 40	40- 60	60 - 80	80 - 100
Number of students	10	30	40	20

10) Find the Geometric Mean for the following distribution of students' marks:

Marks	0 - 20	20 - 50	50 - 70	70 - 100
No. of Students	10	30	50	10

11) Find the mode of the following distribution.

Weights (Kg)	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74
No of Students	10	12	16	12	8

12) Calculate median from the following data.

Group	40 – 44	45 – 49	50 – 54	55 – 59	60 – 64	65 – 69
Frequency	2	6	10	12	8	2

13) Following are the marks of 10 students of a college. Find the range and the coefficient of range. Marks: 400, 450, 520, 380, 485, 495, 575, 440, 410, 460

14) Calculate the mean deviation from mean and its coefficients from the following data.

Size	4 - 5	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10
Frequency	4	6	20	60	75	22

3 MEASURE OF VARIATION OR DISPERSION, SKEWNESS AND KURTOSIS

Chapter Includes :

- INTRODUCTION
- DISPERSION
- RANGE AND COEFFICIENT OF RANGE
- QUARTILE DEVIATION
- THE MEAN DEVIATION
- STANDARD DEVIATION
- THE VARIANCE
- SKEWNESS
- KARL PEARSON'S MEASURES OF SKEWNESS
- BOWLEY'S MEASURE OF SKEWNESS
- KELLY'S MEASURE OF SKEWNESS
- KURTOSIS

Learning Objective :

After going through this chapter, you should be able to:

- Understand Dispersion, Range and Coefficient of Range
- Explain mean deviation
- Describe standard deviation
- Learn variance
- Describe skewness

NOTES

INTRODUCTION

In this chapter you will learn various techniques to distinguish between various shapes of a frequency distribution. This chapter will make you familiar with the concept of dispersion, skewness and kurtosis. The need to study these concepts arises from the fact that the measures of central tendency fail to describe a distribution completely. It is possible to have frequency distributions which differ widely in their nature and composition and yet may have same central tendency.

DISPERSION:

The word dispersion has a technical meaning in statistics. The average measures the center of the data. It is one aspect observations. Another feature of the observations is as to how the observations are spread about the center. The observation may be close to the center or they may be spread away from the center. If the observation are close to the center (usually the arithmetic mean or median), we say that dispersion, scatter or variation is small. If the observations are spread away from the center, we say dispersion is large. Suppose we have three groups of students who have obtained the following marks in a test. The arithmetic means of the three groups are also given below:

Group A: 46, 48, 50, 52, 54 $\quad \overline{X}_A = 50$

Group B: 30, 40, 50, 60, 70 $\quad \overline{X}_B = 50$

Group C: 40, 50, 60, 70, 80 $\quad \overline{X}_C = 60$

In a group A and B arithmetic means are equal i.e. $\overline{X}_A = \overline{X}_B = 50$. But in group A the observations are concentrated on the center. All students of group A have almost the same level of performance. We say that there is consistence in the observations in group A. In group B the mean is 50 but the observations are not closed to the center. One observation is as small as 30 and one observation is as large as 70. Thus there is greater dispersion in group B. In group C the mean is 60 but the spread of the observations with respect to the center 60 is the same as the spread of the observations in group B with respect to their own center which is 50. Thus in group B and C the means are different but their dispersion is the same. In group A and C the means are different and their dispersions are also different. Dispersion is an important feature of the observations and it is measured with the help of the measures of dispersion, scatter or variation. The word variability is also used for this idea of dispersion.

The study of dispersion is very important in statistical data. If in a certain factory there is consistence in the wages of workers, the workers will be satisfied. But if some workers have high wages and some have low wages, there will be conflict among the low paid workers and they might go on strikes and arrange demonstrations. If in a certain country some people are very poor and some are very high rich, we say there is economic disparity. It means that dispersion is large.

The idea of dispersion is important in the study of wages of workers, prices of commodities, standard of living of different people, distribution of wealth, distribution of land among framers and various other fields of life. Some brief definitions of dispersion are:

i. The degree to which numerical data tend to spread about an average value is called the dispersion or variation of the data.

ii. Dispersion or variation may be defined as a statistics signifying the extent of the scatteredness of items around a measure of central tendency.

iii. Dispersion or variation is the measurement of the scatter of the size of the items of a series about the average.

Measures of Dispersion:

For the study of dispersion, we need some measures which show whether the dispersion is small or large. There are two types of measure of dispersion which are:

- Absolute Measure of Dispersion
- Relative Measure of Dispersion

Absolute Measures of Dispersion:

These measures give us an idea about the amount of dispersion in a set of observations. They give the answers in the same units as the units of the original observations. When the observations are in kilograms, the absolute measure is also in kilograms. If we have two sets of observations, we cannot always use the absolute measures to compare their dispersion. We shall explain later as to when the absolute measures can be used for comparison of dispersion in two or more than two sets of data. The absolute measures which are commonly used are:

1. The Range
2. The Quartile Deviation
3. The Mean Deviation
4. The Standard deviation and Variance

Relative Measure of Dispersion:

These measures are calculated for the comparison of dispersion in two or more than two sets of observations. These measures are free of the units in which the original data is measured. If the original data is in dollar or kilometers, we do not use these units with relative measure of dispersion. These measures are a sort of ratio and are called coefficients. Each absolute measure of dispersion can be converted into its relative measure. Thus the relative measures of dispersion are:

1. Coefficient of Range or Coefficient of Dispersion.
2. Coefficient of Quartile Deviation or Quartile Coefficient of Dispersion.
3. Coefficient of Mean Deviation or Mean Deviation of Dispersion.
4. Coefficient of Standard Deviation or Standard Coefficient of Dispersion.
5. Coefficient of Variation (a special case of Standard Coefficient of Dispersion)

NOTES

RANGE AND COEFFICIENT OF RANGE:

THE RANGE:

Range is defined as the difference between the maximum and the minimum observation of the given data. If x_m denotes the maximum observation and denotes the minimum observation then the range is defined as:

$$\text{Range} = x_m - x_0$$

In case of grouped data, the range is the difference between the upper boundary of the highest class and the lower boundary of the lowest class. It is also calculated by using the difference between the mid points of the highest class and the lowest class. It is the simplest measure of dispersion. It gives a general idea about the total spread of the observations. It does not enjoy any prominent place in statistical theory. But it has its application and utility in quality control methods which are used to maintain the quality of the products produced in factories. The quality of products is to be kept within certain range of values.

The range is based on the two extreme observations. It gives no weight to the central values of the data. It is a poor measure of dispersion and does not give a good picture of the overall spread of the observations with respect to the center of the observations. Let us consider three groups of the data which have the same range:

Group A: 30, 40, 40, 40, 40, 40, 50

Group B: 30, 30, 30, 40, 50, 50, 50

Group C: 30, 35, 40, 40, 40, 45, 50

In all the three groups the range is 50 - 30 = 20. In group A there is concentration of observations in the center. In group B the observations are friendly with the extreme corner and in group C the observations are almost equally distributed in the interval from 30 to 50. The range fails to explain these differences in the three groups of data. This defect in range cannot be removed even if we calculate the coefficient of range which is a relative measure of dispersion. If we calculate the range of a sample, we cannot draw any inferences about the range of the population.

Coefficient of Range:

It is relative measure of dispersion and is based on the value of range. It is also called coefficient of dispersion. It is defined as:

Coefficient of Range

$$= \frac{x_m - x_0}{x_m + x_0}$$

The range $x_m - x_0$ is standardized by the total $x_m + x_0$

Let us take two sets of observations. Set A contains marks of five students in Mathematics out of 25 marks and group B contains marks of the same student in English out of 100 marks.

NOTES

Set A: 10, 15, 18, 20, 20

Set B: 30, 35, 40, 45, 50

The values of range and coefficient of range are calculated as:

	Range	Coefficient of Range
Set A: (Mathematics)	$20-10=10$	$\frac{20-10}{20+10}=0.33$
Set B: (English)	$50-30=20$	$\frac{50-30}{50+30}$ 0.25

In set A the range is 10 and in set B the range is 20. Apparently it seems as if there is greater dispersion in set B. But this is not true. The range of 20 in set B is for large observations and the range of 10 in set A is for small observations. Thus 20 and 10 cannot be compared directly. Their base is not the same. Marks in Mathematics are out of 25 and marks of English are out of 100. Thus, it makes no sense to compare 10 with 20. When we convert these two values into coefficient of range, we see that coefficient of range for set A is greater than that of set B. Thus there is greater dispersion or variation in set A. The marks of students in English are more stable than their marks in Mathematics.

Example:

Following are the wages of 8 workers of a factory. Find the range and the coefficient of range.

Wages in (Rs.) 1400, 1450, 1520, 1380, 1485, 1495, 1575, 1440

Solution:

Here Largest value $= x_m = 1575$ and Smallest Value $= x_0 = 1380$

$$\text{Range} = x_m - x_0 = 1575 - 1380 = 195$$

Coefficient of Range

$$= \frac{x_m - x_0}{x_m + x_0} = \frac{1575-1380}{1575+1380} = \frac{195}{2955} = 0.66$$

Example:

The following distribution gives the numbers of houses and the number of persons per house.

Number of Persons	1	2	3	4	5	6	7	8	9	10
Number of Houses	26	113	120	95	60	42	21	14	5	4

Calculate the range and coefficient of range.

NOTES

Solution:

Here Largest value $= x_m = 10$ and Smallest Value $= x_0 = 1$ Range $= x_m - x_0 = 10 - 1 = 9$

Coefficient of Range

$$= \frac{x_m - x_0}{x_m + x_0} = \frac{10-1}{10+1} = \frac{9}{11} = 0.818$$

Example:

Find the range of the weight of the students of a university.

Weights (Kg)	60 – 62	63 – 65	66 – 68	69 – 71	72 – 74
Number of Students	5	18	42	27	8

Calculate the range and coefficient of range.

Solution:

Weights (Kg)	Class Boundaries	Mid Value	No. of Students
60 – 62	59.5 – 62.5	61	5
63 – 65	62.5 – 65.5	64	18
66 – 68	65.5 – 68.5	67	42
69 – 71	68.5 – 71.5	70	27
72 – 74	71.5 – 74.5	73	8

Method 1:

Here, x_m = Upper class boundary of the highest class $= 74.5$

x_0 = Lower class boundary of the lowest class $= 59.5$

Range $= x_m - x_0 = 74.5 - 59.5 = 15$ Kilogram

Coefficient of Range

$$= \frac{x_m - x_0}{x_m + x_0} = \frac{74.5 - 59.5}{74.5 + 59.5} = \frac{15}{134} = 0.1119$$

Method 2:

Here x_m = Mid value of the highest class $= 73$

x_0 = Mid value of the lowest class $= 61$

Range $= x_m - x_0 = 73 - 61 = 12$ Kilogram

$$\text{Coefficient of Range} = \frac{x_m - x_0}{x_m + x_0} = \frac{73 - 61}{73 + 61} = \frac{12}{134} = 0.0895$$

QUARTILE DEVIATION

It is based on the lower quartile Q_1 and the upper quartile Q_3. The difference $Q_3 - Q_1$ is called the inter quartile range. The difference $Q_3 - Q_1$ divided by 2 is called semi-inter-quartile range or the quartile deviation. Thus,

Quartile Deviation (Q.D)

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

The quartile deviation is a slightly better measure of absolute dispersion than the range. But it ignores the observation on the tails. If we take difference samples from a population and calculate their quartile deviations, their values are quite likely to be sufficiently different. This is called sampling fluctuation. It is not a popular measure of dispersion. The quartile deviation calculated from the sample data does not help us to draw any conclusion (inference) about the quartile deviation in the population.

Coefficient of Quartile Deviation:

A relative measure of dispersion based on the quartile deviation is called the coefficient of quartile deviation. It is defined as

Coefficient of Quartile Deviation

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

It is pure number free of any units of measurement. It can be used for comparing the dispersion in two or more than two sets of data.

Example:

The wheat production (in Kg) of 20 acres is given as: 1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750, and 1885. Find the quartile deviation and coefficient of quartile deviation.

NOTES

Solution:

After arranging the observations in ascending order, we get

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)th \text{ item}$$

$$= \text{Value of } \left(\frac{20+1}{4}\right)th \text{ item}$$

$$= \text{Value of } (5.25)th \text{ item}$$

$$= 5th \text{ item} + 0.25(6th \text{ item} - 5th \text{ item}) = 1240 + 0.25(1320 - 1240)$$

$$Q_1 = 1240 + 20 = 1260$$

$$Q_3 = \text{Value of } \frac{3(n+1)}{4}th \text{ item}$$

$$= \text{Value of } \frac{3(20+1)}{4}th \text{ item}$$

$$= \text{Value of } (15.75)th \text{ item}$$

$$= 15th \text{ item} + 0.75(16th \text{ item} - 15th \text{ item}) = 1750 + 0.75(1755 - 1750)$$

$$Q_3 = 1750 + 3.75 = 1753.75$$

Quartile Deviation (Q.D)

$$= \frac{Q_3 - Q_1}{2} = \frac{1753.75 - 1260}{2} = \frac{492.75}{2} = 246.875$$

Coefficient of Quartile Deviation

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{1753.75 - 1260}{1753.75 + 1260} = 0.164$$

Example:

Calculate the quartile deviation and coefficient of quartile deviation from the data given below:

Maximum Load (short-tons)	Number of Cables
9.3 – 9.7	2
9.8 – 10.2	5
10.3 – 10.7	12
10.8 – 11.2	17
11.3 – 11.7	14
11.8 – 12.2	6
12.3 – 12.7	3
12.8 – 13.2	1

Solution:

The necessary calculations are given below:

Maximum Load (short-tons)	Number of Cables f	Class Boundaries	Cumulative Frequencies
9.3 – 9.7	2	9.25 – 9.75	2
9.8 – 10.2	5	9.75 – 10.25	2 + 3 = 7
10.3 – 10.7	12	10.25 – 10.75	7 + 12 = 19
10.8 – 11.2	17	10.75 – 11.25	19 + 17 = 36
11.3 – 11.7	14	11.25 – 11.75	36 + 14 = 50
11.8 – 12.2	6	11.75 – 12.25	50 + 6 = 56
12.3 – 12.7	3	12.25 – 12.75	56 + 3 = 59
12.8 – 13.2	1	12.75 – 13.25	59 + 1 = 60

Q_1 = Value of $\left(\frac{n}{4}\right)th$ item = Value of $\left(\frac{60}{4}\right)th$ item = $15th$ item

Q_1 lies in the class 10.25 – 10.75

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - c\right)$$

Where $l = 10.25$, $h = 0.5$, $f = 12$, $n/4 = 15$ and $c = 7$

$$\therefore \quad Q_1 = 10.25 + \frac{0.5}{12}(15 - 7) = 10.25 + 0.33 = 10.58$$

Q_3 = Value of $\left(\frac{3n}{4}\right)th$ item = Value of $\left(\frac{3 \times 60}{4}\right)th$ item = $45th$ item

Q_3 lies in the class 11.25 – 11.75

$$\therefore \quad Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - c\right)$$

NOTES

Where $l = 11.25$, $h = 0.5$, $f = 14, 3n/4 = 45$ and $c = 36$

$$\therefore\ Q_1 = 11.25 + \frac{0.5}{14}(45 - 36) = 11.25 + 0.32 = 11.57$$

Quartile Deviation (Q.D)

$$= \frac{Q_3 - Q_1}{2} = \frac{11.57 - 10.58}{2} = \frac{0.99}{2} = 0.495$$

Coefficient of Quartile Deviation

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{11.57 - 10.58}{11.57 + 10.58}$$

$$= \frac{0.99}{22.15} = 0.045$$

THE MEAN DEVIATION

The mean deviation or the average deviation is defined as the mean of the absolute deviations of observations from some suitable average which may be the arithmetic mean, the median or the mode. The difference ($X - \text{average}$) is called deviation and when we ignore the negative sign, this deviation is written as $|X - \text{average}|$ and is read as mod deviations. The mean of these mod or absolute deviations is called the mean deviation or the mean absolute deviation. Thus for sample data in which the suitable average is the $\overline{X}$, the mean deviation ($M.D$) is given by the relation:

$$M.D = \frac{\Sigma|X - \overline{X}|}{n}$$

For frequency distribution, the mean deviation is given by

$$M.D = \frac{\Sigma f|X - \overline{X}|}{\Sigma f}$$

When the mean deviation is calculated about the median, the formula becomes

$$M.D\ (\text{about median}) = \frac{\Sigma f|X - \text{Median}|}{\Sigma f}$$

The mean deviation about the mode is

$$M.D\,(\text{about mode}) = \frac{\Sigma f|X - \text{Mode}|}{\Sigma f}$$

For a population data the mean deviation about the population mean is

$$M.D = \frac{\Sigma f|X - \mu|}{\Sigma f}$$

The mean deviation is a better measure of absolute dispersion than the range and the quartile deviation.

A drawback in the mean deviation is that we use the absolute deviations $|X - \text{average}|$ which does not seem logical. The reason for this is that $\Sigma(X - \overline{X})$ is always equal to zero. Even if we use median or mode in place of $\overline{X}$, even then the summation $\Sigma(X - \text{median})$ or $\Sigma(X - \text{mode})$ will be zero or approximately zero with the result that the mean deviation would always be either zero or close to zero. Thus the very definition of the mean deviation is possible only on the absolute deviations.

The mean deviation is based on all the observations, a property which is not possessed by the range and the quartile deviation. The formula of the mean deviation gives a mathematical impression that is a better way of measuring the variation in the data. Any suitable average among the mean, median or mode can be used in its calculation but the value of the mean deviation is minimum if the deviations are taken from the median. A series drawback of the mean deviation is that it cannot be used in statistical inference.

Coefficient of the Mean Deviation:

A relative measure of dispersion based on the mean deviation is called the coefficient of the mean deviation or the coefficient of dispersion. It is defined as the ratio of the mean deviation to the average used in the calculation of the mean deviation. Thus

$$\text{Coefficient of } M.D\,(\text{about mean}) = \frac{\text{Mean Deviation from Mean}}{\text{Mean}}$$

$$\text{Coefficient of } M.D\,(\text{about median}) = \frac{\text{Mean Deviation from Median}}{\text{Median}}$$

$$\text{Coefficient of } M.D\,(\text{about mode}) = \frac{\text{Mean Deviation from Mode}}{\text{Mode}}$$

Example:

Calculate the mean deviation form (1) arithmetic mean (2) median (3) mode in respect of the marks obtained by nine students gives below and show that the mean deviation from median is minimum.

NOTES

Marks (out of 25): 7, 4, 10, 9, 15, 12, 7, 9, 7

Solution:

After arranging the observations in ascending order, we get

Marks: 4, 7, 7, 7, 9, 9, 10, 12, 15

$$\text{Mean} = \frac{\Sigma X}{n} = \frac{80}{9} = 8.89$$

$$\text{Median} = \text{Value of}\left(\frac{n+1}{2}\right)th \text{ item} = \text{Value of}\left(\frac{9+1}{2}\right)th \text{ item}$$

$$= \text{Value of}\,(5)th \text{ item} = 9$$

$$\text{Mode} = 7 \text{ (Since 7 is repeated maximum number of times)}$$

Marks X	$\lvert X - \bar{X}\rvert$	$\lvert X - \text{Median}\rvert$	$\lvert X - \text{Mode}\rvert$
4	4.89	5	3
7	1.89	2	0
7	1.89	2	0
7	1.89	2	0
9	0.11	0	2
9	0.11	0	2
10	1.11	1	3
12	3.11	3	5
15	6.11	6	8
Total	21.11	21	23

$$M.D \text{ from mean} = \frac{\Sigma|X - \bar{X}|}{n} = \frac{21.11}{9} = 2.35$$

$$M.D \text{ from median} = \frac{\Sigma|X - \text{Median}|}{n} = \frac{21}{9} = 2.33$$

$$M.D \text{ from mode} = \frac{\Sigma|X - \text{Mode}|}{n} = \frac{23}{9} = 2.56$$

From the above calculations, it is clear that the mean deviation from the median has the least value.

Example:

Calculate the mean deviation from mean and its coefficients from the following data.

Size of Items	3–4	4–5	5–6	6–7	7–8	8–9	9–10
Frequency	3	7	22	60	85	32	8

Solution:

The necessary calculation is given below:

Size of Items	X	f	fX	$\lvert X-\bar{X}\rvert$	$f\lvert X-\bar{X}\rvert$
3–4	3.5	3	10.5	3.59	10.77
4–5	4.5	7	31.5	2.59	18.13
5–6	5.5	22	121.0	1.59	34.98
6–7	6.5	60	390.0	0.59	35.40
7–8	7.5	85	637.5	0.41	34.85
8–9	8.5	32	272.0	1.41	45.12
9–10	9.5	8	76.0	2.41	19.28
Total		217	1538.5		198.53

$$\text{Mean} = \bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{1538.5}{217} = 7.09$$

$$M.D \text{ from mean} = \frac{\Sigma|X-\bar{X}|}{n} = \frac{198.53}{217} = 0.915$$

$$\text{Coefficient of } M.D \text{ (mean)} = \frac{M.D \text{ from Mean}}{\text{Mean}} = \frac{0.915}{7.09} = 0.129$$

STANDARD DEVIATION

The standard deviation is defined as the positive square root of the mean of the square deviations taken from arithmetic mean of the data.

For the sample data the standard deviation is denoted by and is defined as:

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

For a population data the standard deviation is denoted by (sigma) and is defined as:

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

For frequency distribution the formulas becomes

$$S = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{\Sigma f}} \text{ or } \sigma = \sqrt{\frac{\Sigma f(X-\mu)^2}{\Sigma f}}$$

NOTES

The standard deviation is in the same units as the units of the original observations. If the original observations are in grams, the value of the standard deviation will also be in grams.

The standard deviation plays a dominating role for the study of variation in the data. It is a very widely used measure of dispersion. It stands like a tower among measure of dispersion. As far as the important statistical tools are concerned, the first important tool is the mean and the second important tool is the standard deviation . It is based on all the observations and is subject to mathematical treatment.

It is of great importance for the analysis of data and for the various statistical inferences.

However some alternative methods are also available to compute standard deviation. The alternative methods simplify the computation. Moreover in discussing these methods we will confirm ourselves only to sample data because sample data rather than whole population confront mostly a statistician.

Actual Mean Method:

In applying this method first of all we compute arithmetic mean of the given data either ungroup or grouped data. Then take the deviation from the actual mean. This method is already defined above. The following formulas are applied:

For Ungrouped Data	For Grouped Data
$S=\sqrt{\frac{\Sigma\left(X-\bar{X}\right)^2}{n}}$	$S=\sqrt{\frac{\Sigma f\left(X-\bar{X}\right)^2}{\Sigma f}}$

This method is also known as direct method

Assumed Mean Method:

1. We use following formulas to calculate standard deviation:

For Ungrouped Data	For Grouped Data
$S=\sqrt{\frac{\Sigma D^2}{n}-\left(\frac{\Sigma D}{n}\right)^2}$	$S=\sqrt{\frac{\Sigma fD^2}{\Sigma f}-\left(\frac{\Sigma fD}{\Sigma f}\right)^2}$

Where $D=X-A$ and A is any assumed mean other than zero. This method is also known as short-cut method.

2. If A is considered to be zero then the above formulas are reduced to the following formulas:

For Ungrouped Data	For Grouped Data
$S=\sqrt{\frac{\Sigma X^2}{n}-\left(\frac{\Sigma X}{n}\right)^2}$	$S=\sqrt{\frac{\Sigma fX^2}{\Sigma f}-\left(\frac{\Sigma fX}{\Sigma f}\right)^2}$

3. If we are in a position to simplify the calculation by taking some common factor or divisor from the given data the formulas for computing standard deviation are:

For Ungrouped Data	For Grouped Data
$S=\sqrt{\frac{\Sigma U^2}{n}-\left(\frac{\Sigma U}{n}\right)^2}\times c$	$S=\sqrt{\frac{\Sigma fU^2}{\Sigma f}-\left(\frac{\Sigma fU}{\Sigma f}\right)^2}\times c \text{ or } h$

Where

$$U=\frac{X-A}{h \text{ or } c}=\frac{D}{h \text{ or } c}$$

h = Class Interval and

c = Common Divisor

This method is also called method of step-deviation.

Example:

Calculate the standard deviation for the following sample data using all methods: 2, 4, 8, 6, 10, and 12.

Solution:

Method-I: Actual Mean Method

X	$(X-\bar{X})^2$
2	$(2-7)^2=25$
4	$(4-7)^2=9$
8	$(8-7)^2=1$
6	$(6-7)^2=1$
10	$(10-7)^2=9$
12	$(12-7)^2=25$
$\Sigma X=42$	$\Sigma(X-\bar{X})^2=70$

$$\bar{X}=\frac{\Sigma X}{n}=\frac{42}{6}=7$$

NOTES

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

$$S=\sqrt{\frac{70}{6}}=\sqrt{\frac{35}{3}}=3.42$$

Method-II: Taking Assumed Mean as 6

X	$D=(X-6)$	D^2
2	–4	16
4	–2	4
8	2	4
6	0	0
10	4	16
12	6	36
Total	$\Sigma D=6$	$\Sigma D^2=76$

$$S=\sqrt{\frac{\Sigma D^2}{n}-\left(\frac{\Sigma D}{n}\right)^2}$$

$$S=\sqrt{\frac{76}{6}-\left(\frac{6}{6}\right)^2}=\sqrt{\frac{70}{6}}$$

$$S=\sqrt{\frac{35}{3}}=3.42$$

Method-III: Taking Assume Mean as Zero

X	X^2
2	4
4	16
8	64
6	36
10	100
12	144
$\Sigma X=42$	$\Sigma X^2=364$

$$S=\sqrt{\frac{\Sigma X^2}{n}-\left(\frac{\Sigma X}{n}\right)^2}$$

$$S=\sqrt{\frac{364}{6}-\left(\frac{42}{6}\right)^2}$$

$$S=\sqrt{\frac{70}{6}}=\sqrt{\frac{35}{3}}=3.42$$

Method-IV: Taking ^{2} as common divisor or factor

X	$U=(X-4)/2$	U^2
2	–1	1
4	0	0
8	2	4
6	1	1
10	3	9
12	4	16
Total	$\Sigma U=9$	$\Sigma U^2=31$

$$S=\sqrt{\frac{\Sigma U^2}{n}-\left(\frac{\Sigma U}{n}\right)^2}\times c$$

$$S=\sqrt{\frac{31}{6}-\left(\frac{9}{6}\right)^2}\times 2$$

$$S=\sqrt{2.92}\times 2=3.42$$

Example:

Calculate standard deviation from the following distribution of marks by using all the methods.

Marks	No. of Students
1–3	40
3–5	30
5–7	20
7–9	10

NOTES

Solution:

Method-I: Actual Mean Method

Marks	f	X	fX	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
1–3	40	2	80	4	160
3–5	30	4	120	0	0
5–7	20	6	120	4	80
7–9	10	8	80	16	160
Total	100		400		400

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{400}{100} = 4$$

$$S = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{\Sigma f}} = \sqrt{\frac{400}{100}} = \sqrt{4} = 2$$

= 2 Marks

Method-II: Taking assumed mean as 2

Marks	f	X	$D=(X-2)$	fD	fD^2
1–3	40	2	0	0	0
3–5	30	4	2	60	120
5–7	20	6	4	80	320
7–9	10	8	6	60	160
Total	100			200	800

$$S = \sqrt{\frac{\Sigma fD^2}{\Sigma f} - \left(\frac{\Sigma fD}{\Sigma f}\right)^2} = \sqrt{\frac{800}{100} - \left(\frac{200}{100}\right)^2}$$

$S = \sqrt{8-4} = \sqrt{4} = 2$ Marks

Method-III: Using assumed mean as Zero

Marks	f	X	fX	fX^2
1–3	40	2	80	160
3–5	30	4	120	480
5–7	20	6	120	720
7–9	10	8	80	640
Total	100		400	2000

$$S=\sqrt{\frac{\Sigma fX^2}{\Sigma f}-\left(\frac{\Sigma fX}{\Sigma f}\right)^2}=\sqrt{\frac{2000}{100}-\left(\frac{400}{100}\right)^2}$$

$$S=\sqrt{20-16}=\sqrt{4}=2 \text{ Marks}$$

Method-IV: By taking as the common divisor

Marks	f	X	$U=(X-2)/2$	fU	fU^2
1–3	40	2	–2	–80	160
3–5	30	4	–1	–30	30
5–7	20	6	0	0	0
7–9	10	8	1	10	10
Total	100			–100	200

$$S=\sqrt{\frac{\Sigma fU^2}{\Sigma f}-\left(\frac{\Sigma fU}{\Sigma f}\right)^2}\times h=\sqrt{\frac{200}{100}-\left(\frac{-100}{100}\right)^2}\times 2$$

$$S=\sqrt{2-1}\times 2=\sqrt{1}\times 2=1\times 2=2 \text{ Marks}$$

Coefficient of Standard Deviation:

The standard deviation is the absolute measure of dispersion. Its relative measure is called standard coefficient of dispersion or coefficient of standard deviation. It is defined as:

Coefficient of Standard Deviation

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

Coefficient of Variation:

The most important of all the relative measure of dispersion is the coefficient of variation. This word is variation not variance. There is no such thing as coefficient of variance. The coefficient of variation $(C.V)$ is defined as:

Coefficient of Variation

$$(C.V)=\frac{S}{\bar{X}}\times 100$$

Thus $C.V$ is the value of S when $\bar{X}$ is assumed equal to 100. It is a pure number and the unit of observations is not mentioned with its value. It is written in percentage form like 20% or 25%. When its value is 20%, it means that when the mean of the observations is assumed equal to 100, their standard deviation will be 20. The $C.V$ is used to compare the dispersion in different sets of data particularly the data which differ in their means or differ in the units of measurement. The wages of workers may be in rupees and the consumption of fruits in their families may be in kilograms. The standard deviation of wages in rupees cannot be compared with the standard deviation of amounts of fruits in kilograms. Both the standard deviations need to be converted into coefficient of variation for compari-

NOTES

son. Suppose the value of $C.V$ for wages is 10% and the values of $C.V$ for kilograms of fruit is 25%. This means that the wages of workers are consistent. Their wages are close to the overall average of their wages. But the families consume fruits in quite different quantities. Some families use very small quantities of fruits and some others use large quantities of fruits. We say that there is greater variation in their consumption of fruits. The observations about the quantity of fruits are more dispersed or more variant.

Example:

Calculate the coefficient of standard deviation and coefficient of variation for the following sample data: 2, 4, 8, 6, 10, and 12.

Solution:

X	$(X-\bar{X})^2$
2	$(2-7)^2=25$
4	$(4-7)^2=9$
8	$(8-7)^2=1$
6	$(6-7)^2=1$
10	$(10-7)^2=9$
12	$(12-7)^2=25$
$\Sigma X=42$	$\Sigma(X-\bar{X})^2=70$

$$\bar{X}=\frac{\Sigma X}{n}=\frac{42}{6}=7$$

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

$$S=\sqrt{\frac{70}{6}}=\sqrt{\frac{35}{3}}=3.42$$

Coefficient of Standard Deviation

$$=\frac{S}{\bar{X}}=\frac{3.42}{7}=0.49$$

Coefficient of Variation

$$(C.V)=\frac{S}{\bar{X}}\times 100=\frac{3.42}{7}\times 100=48.86\%$$

Example:

Calculate coefficient of standard deviation and coefficient of variation from the following distribution of marks:

Marks	No. of Students
1–3	40
3–5	30
5–7	20
7–9	10

Solution:

Marks	f	X	fX	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
1–3	40	2	80	4	160
3–5	30	4	120	0	0
5–7	20	6	120	4	80
7–9	10	8	80	16	160
Total	100		400		400

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{400}{100} = 4$$

$$S = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{\Sigma f}} = \sqrt{\frac{400}{100}} = \sqrt{4} = 2 \text{ Marks}$$

Coefficient of Standard Deviation

$$= \frac{S}{\bar{X}} = \frac{2}{4} = 0.5$$

Coefficient of Variation

$$(C.V) = \frac{S}{\bar{X}} \times 100 = \frac{2}{4} \times 100 = 50\%$$

Uses of Coefficient of Variation:

- Coefficient of variation is used to know the consistency of the data. By consistency we mean the uniformity in the values of the data/distribution from arithmetic mean of the data/distribution. A distribution with smaller than the other is taken as more consistent than the other.

$C.V$ is also very useful when comparing two or more sets of data that are measured in different units of measurement.

NOTES

THE VARIANCE

Variance is another absolute measure of dispersion. It is defined as the average of the squared difference between each of the observations in a set of data and the mean. For a sample data the variance is denoted is denoted by S^2 and the population variance is denoted by σ^2 (sigma square).

The sample variance S^2 has the formula:

$$S^2 = \frac{\Sigma\left(X - \bar{X}\right)^2}{n}$$

Where $\bar{X}$ is sample mean and n is the number of observations in the sample.

The population variance σ^2 is defined as:

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Where μ is the mean of population and N is the number of observations in the data. It may be remembered that the population variance σ^2 is usually not calculated. The sample variance S^2 is calculated and if need be, this S^2 is used to make inference about the population variance.

The term $\Sigma\left(X - \bar{X}\right)^2$ is positive, therefore S^2 is always positive. If the original observations are in centimeter, the value of the variance will be (centimeter) 2. Thus the unit of S^2 is the square of the units of the original measurement.

For a frequency distribution the sample variance S^2 is defined as:

$$S^2 = \frac{\Sigma f\left(X - \bar{X}\right)^2}{\Sigma f}$$

For s frequency distribution the population variance σ^2 is defined as:

$$\sigma^2 = \frac{\Sigma f(X - \mu)^2}{\Sigma f}$$

In simple words we can say that variance is the square of standard deviation.

$$\text{Variance} = (\text{Standrad Deviation})^2$$

Example:

Calculate the variance for the following sample data: 2, 4, 8, 6, 10, and 12.

Solution:

X	$(X-\bar{X})^2$
2	$(2-7)^2 = 25$
4	$(4-7)^2 = 9$
8	$(8-7)^2 = 1$
6	$(6-7)^2 = 1$
10	$(10-7)^2 = 9$
12	$(12-7)^2 = 25$
$\Sigma X = 42$	$\Sigma(X-\bar{X})^2 = 70$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{42}{6} = 7$$

$$S^2 = \frac{\Sigma(X-\bar{X})^2}{n}$$

$$S^2 = \frac{70}{6} = \frac{35}{3} = 11.67$$

$$\text{Variance} = S^2 = 11.67$$

Example:

Calculate variance from the following distribution of marks:

Marks	No. of Students
1–3	40
3–5	30
5–7	20
7–9	10

NOTES

Solution:

Marks	f	X	fX	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
1–3	40	2	80	4	160
3–5	30	4	120	0	0
5–7	20	6	120	4	80
7–9	10	8	80	16	160
Total	100		400		400

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{400}{100} = 4$$

$$S^2 = \frac{\Sigma f(X-\bar{X})^2}{\Sigma f} = \frac{400}{100} = 4$$

$$\text{Variance} = S^2 = 4$$

Sheppard Corrections:

In grouped data the different observations are put into the same class. In the calculation of variation or standard deviation for grouped data, the frequency f is multiplied with X which is the mid-point of the respective class. Thus it is assumed that all the observations in a class are centered at X. But this is not true because the observations are spread in the said class. This assumption introduces some error in the calculation of S^2 and S. The value of S^2 and S can be corrected to some extent by applying Sheppard correction. Thus

$$S^2(\text{corrected}) = S^2 - \frac{h^2}{12}$$

$$S(\text{corrected}) = \sqrt{S^2 - \frac{h^2}{12}}$$

Where h is the uniform class interval

This correction is applied in grouped data which has almost equal tails in the start and at the end of the data. If a data a longer tail on any side, this correction is not applied. If size of the class interval h is not the same in all classes, the correction is not applicable.

Corrected Coefficient of Variation:

When the corrected standard deviation is used in the calculation of the coefficient of variation, we get what is called the corrected coefficient of variation. Thus

$$= \frac{S(\text{corrected})}{\bar{X}} \times 100$$

Example:

Calculate Sheppard correction and corrected coefficient of variation from the following distribution of marks by using all the methods.

Marks	No. of Students
1 – 3	40
3 – 5	30
5 – 7	20
7 – 9	10

Solution:

Marks	f	X	fX	$U=(X-2)/2$	fU	fU^2
1 – 3	40	2	80	–2	–80	160
3 – 5	30	4	120	–1	–30	30
5 – 7	20	6	120	0	0	0
7 – 9	10	8	80	1	10	10
Total	100		400		–100	200

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{400}{100} = 4$$

$$S^2 = \left[\frac{\Sigma fU^2}{\Sigma f} - \left(\frac{\Sigma fU}{\Sigma f}\right)^2\right] \times h = \left[\frac{200}{100} - \left(\frac{-100}{100}\right)^2\right] \times 2$$

$$S^2 = [2-1] \times 2 = 1 \times 2 = 2$$

$$S^2(\text{corrected}) = S^2 - \frac{h^2}{12} = 2 - \frac{4}{12} = \frac{5}{3} = 1.67$$

$$S(\text{corrected}) = \sqrt{S^2 - \frac{h^2}{12}} = \sqrt{1.67} = 1.29$$

NOTES

$$= \frac{S(\text{corrected})}{\bar{X}} \times 100$$

$$= \frac{1.29}{4} \times 100 = 32.25$$

Combined Variance:

Like combined mean, the combined variance or standard deviation can be calculated for different sets of data. Suppose we have two sets of data containing n_1 and n_2 observations with means $\bar{X}_1$ and $\bar{X}_2$, and variances S_1^2 and S_2^2. If $\bar{X}_c$ is the combined mean and S_c^2 is the combined variance of $n_1 + n_2$ observations, then combined variance is given by

$$S_c^2 = \frac{n_1 S_1^2 + n_2 S_2^2 + n_1 (\bar{X}_1 - \bar{X}_c)^2 + n_2 (\bar{X}_2 - \bar{X}_c)^2}{n_1 + n_2}$$

It can be written as

$$S_c^2 = \frac{n_1 \left[S_1^2 + (\bar{X}_1 - \bar{X}_c)^2 \right] + n_2 \left[S_2^2 + (\bar{X}_2 - \bar{X}_c)^2 \right]}{n_1 + n_2}$$

Where

$$X_c^2 = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

The combine standard deviation S_c can be calculated by taking the square root of S_c^2.

Example:

For a group of 50 male workers the mean and standard deviation of their daily wages are Rs. 63 and Rs. 9 respectively. For a group of 40 female workers these values are Rs. 54 and Rs. 6 respectively. Find the mean and variance of the combined group of 90 workers.

Solution:

Here

$n_1 = 50$,

$\bar{X}_1 = 63$,

$S_1^2 = 81$

$n_2 = 40,$

$\bar{X}_2 = 54,$

$S_2^2 = 36$

Combined Arithmetic Mean

$$= \bar{X}_c^2 = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

$$= \bar{X}_c^2 = \frac{50(63) + 40(54)}{50 + 40} = \frac{5310}{90} = 59$$

Combined Variance $S_c^2 = \dfrac{n_1\left[S_{n_1}^2 + \left(\bar{X}_1 - \bar{X}_c\right)^2\right] + n_2\left[S_2^2 + \left(\bar{X}_2 - \bar{X}_c\right)^2\right]}{n_1 + n_2}$

$$= \frac{50\left[81 + (63 - 59)^2\right] + 40\left[36 + (54 - 59)^2\right]}{50 + 40}$$

$$= \frac{4850 + 2440}{90} = \frac{7290}{90} = 81$$

SKEWNESS

In everyday language, the terms "skewed" and "askew" are used to refer to something that is out of line or distorted on one side. When referring to the shape of frequency or probability distributions, "skewness" refers to asymmetry of the distribution. A distribution with an asymmetric tail extending out to the right is referred to as "positively skewed" or "skewed to the right," while a distribution with an asymmetric tail extending out to the left is referred to as "negatively skewed" or "skewed to the left." Skewness can range from minus infinity to positive infinity.

Karl Pearson (1895) first suggested measuring skewness by standardizing the difference between the mean and the mode, that is,

$$sk = \frac{\mu - \text{mode}}{\sigma}.$$

Population modes are not well estimated from sample modes, but one can estimate the difference between the mean and the mode as being three times the difference between the mean and the median (Stuart & Ord, 1994), leading to the following estimate of skewness:

NOTES

$$sk_{est} = \frac{3(M - \text{median})}{s}$$

Many statisticians use this measure but with the '3' eliminated, that is, $sk = \frac{(M - \text{median})}{s}$. This statistic ranges from -1 to +1. Absolute values above 0.2 indicate great skewness (Hildebrand, 1986).

Skewness has also been defined with respect to the third moment about the mean: $\gamma_1 = \frac{\Sigma(X-\mu)^3}{n\sigma^3}$, which is simply the expected value of the distribution of cubed z scores. Skewness measured in this way is sometimes referred to as "Fisher's skewness." When the deviations from the mean are greater in one direction than in the other direction, this statistic will deviate from zero in the direction of the larger deviations. From sample data, Fisher's skewness is most often estimated by:

$$g_1 = \frac{n\Sigma z^3}{(n-1)(n-2)}$$

For large sample sizes (n > 150), g1 may be distributed approximately normally, with a standard error of approximately $\sqrt{6/n}$. While one could use this sampling distribution to construct confidence intervals for or tests of hypotheses about y1, there is rarely any value in doing so.

It is important for behavioral researchers to notice skewness when it appears in their data. Great skewness may motivate the researcher to investigate outliers. When making decisions about which measure of location to report (means being drawn in the direction of the skew) and which inferential statistic to employ (one which assumes normality or one which does not), one should take into consideration the estimated skewness of the population. Normal distributions have zero skewness. Of course, a distribution can be perfectly symmetric but far from normal. Transformations commonly employed to reduce (positive) skewness include square root, log, and reciprocal transformations.

The skewness of a distribution is defined as the lack of symmetry. In a symmetrical distribution, the Mean, Median and Mode are equal to each other and the ordinate at mean divides the distribution into two equal parts such that one part is mirror image of the other. If some observations, of very high (low) magnitude, are added to such a distribution, its right (left) tail gets elongated.

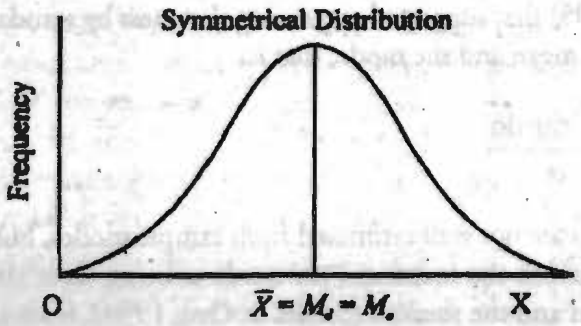
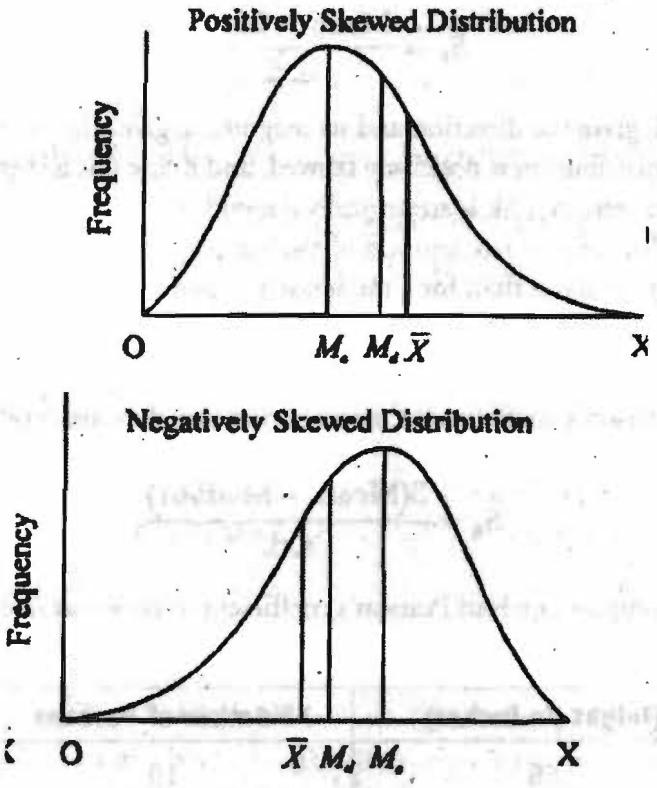


Figure 2:



These observations are also known as extreme observations. The presence of extreme observations on the right hand side of a distribution makes it positively skewed and the three averages, viz., mean, median and mode, will no longer be equal. We shall in fact have Mean > Median > Mode when a distribution is positively skewed. On the other hand, the presence of extreme observations to the left hand side of a distribution make it negatively skewed and the relationship between mean, median and mode is: Mean < Median < Mode. In Fig. 2 we depict the shapes of positively skewed and negatively skewed distributions. The direction and extent of skewness can be measured in various ways. We shall discuss four measures of skewness in this chapter.

KARL PEARSON'S MEASURES OF SKEWNESS

In Fig. 2 you noticed that the mean, median and mode are not equal in a skewed distribution. The Karl Pearson's measure of skewness is based upon the divergence of mean from mode in a skewed distribution.

Since Mean = Mode in a symmetrical distribution, (Mean - Mode) can be taken as an absolute measure of skewness. The absolute measure of skewness for a distribution depends upon the unit of measurement. For example, if the mean = 2.45 metre and mode = 2.14 metre, then absolute measure of skewness will be 2.45 metre - 2.14 metre = 0.31 metre. For the same distribution, if we change the unit of measurement to centimetres, the absolute measure of skewness is 2.45 centimetre - 2.14 centimetre = 0.31 centimetre. In order to avoid such a problem Karl Pearson takes a relative measure of skewness.

A relative measure, independent of the units of measurement, is defined as the Karl

NOTES

Pearson's Coeficient of Skewness Sk, given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{s.d.}}$$

The sign of Sk gives the direction and its magnitude gives the extent of skewness. If Sk > 0, the distribution is positively skewed, and if S, < 0 it is negatively skewed. So far we have seen that Sk is strategically dependent upon mode. If mode is not defined for a distribution we cannot find Sk. But empirical relation between mean, median and mode states that, for a moderately symmetrical distribution, we have:

Hence Karl Pearson's coefficient of skewness is defined in terms of median as

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\text{s.d.}}$$

Example 1: Compute the Karl Pearson's coefficient of skewness from the following data:

Height (in inches)	Number of Persons
58	10
59	18
60	30
61	42
62	35
63	28
64	16
65	8

Table for the computation of mean and s.d.

Height (X)	$u = X - 61$	No. of persons (f)	fu	fu^2
58	- 3	10	- 30	90
59	- 2	18	-36	72
60	- 1	30	- 30	30
61	0	42	0	0
62	1	35	35	35
63	2	28	56	112
64	3	16	48	144
65	4	8	32	128
Total		187	75	611

$$\text{Mean} = 61 + \frac{75}{187} = 61.4$$

$$\text{s.d.} = \sqrt{\frac{611}{187} - \left(\frac{75}{187}\right)^2} = 1.76$$

To find mode, we note that height is a continuous variable. It is assumed that the height has been measured under the approximation that a measurement on height that is, e.g., greater than 58 but less than 58.5 is taken as 58 inches while a measurement greater than or equal to 58.5 but less than 59 is taken as 59 inches. Thus the given data can be written as

Height (in inches)	No. of persons
57.5 - 58.5	10
58.5 - 59.5	18
59.5 - 60.5	30
60.5 - 61.5	42
61.5 - 62.5	35
62.5 - 63.5	28
63.5 - 64.5	16
64.5 - 65.5	8

By inspection, the modal class is 60.5 - 61.5. Thus, we have

$$l_m = 60.5,\ \Delta_1 = 42 - 30 = 12, \Delta_2 = 42 - 35 = 7 \text{ and } h = 1.$$

$$\therefore \quad \text{Mode} = 60.5 + \frac{12}{12+7} \times 1 = 61.13$$

Hence, the Karl Pearson's coeficient of skewness:

$$S_k = \frac{61.4 - 61.13}{1.76} = 0.153.$$

Thus the distribution is positively skewed.

BOWLEY'S MEASURE OF SKEWNESS

This measure is based on quartiles. For a symmetrical distribution, it is seen that Q1 and Q3 are equidistant ftom median. Thus (Q3 - Md) - (Md - Q1) can be taken as an absolute measure of skewness. A relative measure of skewness, known as Bowley's coefficient (SQ), is given by

$$S_Q = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}$$

$$= \frac{Q_3 - 2M_d + Q_1}{Q_3 - Q_1}$$

The Bowley's coefficient for the data on heights given in following Table is computed below:

NOTES

Height (in inches)	No. of persons (*f*)	Cumulative Frequency
57.5 - 58.5	10	10
58.5 - 59.5	18	28
59.5 - 60.5	30	58
60.5 - 61.5	42	100
61.5 - 62.5	35	135
62.5 - 63.5	28	163
63.5 - 64.5	16	179
64.5 - 65.5	8	187

Computation of Q1:

Since N/4 = 46.75, the first quartile class is 59.5 - 60.5. Thus

l_{Q1}, = 59.5, C = 28, f_{Q1} = 30 and h = 1.

$$\therefore \quad Q_1 = 59.5 + \frac{46.75 - 28}{30} \times 1 = 60.125.$$

Computation of M_d (Q_2):

Since N/2 = 93.5, the median class is 60.5 - 61.5. Thus

l_m = 60.5, C= 58, f_m = 42 and h = 1.

$$\therefore \quad M_d = 60.5 + \frac{93.5 - 58}{42} \times 1 = 61.345.$$

Computation of Q_3:

Since 3N/4 = 140.25, the third quartile class is 62.5 - 63.5. Thus

l_{Q3}, = 62.5, C = 135, f_{Q3} = 28 and h = 1.

$$\therefore \quad Q_3 = 62.5 + \frac{140.25 - 135}{28} \times 1 = 62.688.$$

Hence, Bowley's coefficient $S_Q = \frac{62.688 - 2 \times 61.345 + 60.125}{62.688 - 60.125} = 0.048$.

KELLY'S MEASURE OF SKEWNESS

Bowley's measure of skewness is based on the middle 50% of the observations because it leaves 25% of the observaticins on each extreme of the distribution. As an improvement over Bowley's measure, Kelly has suggested a measure based on P10 and, P90 so that only 10% of the observations on each extreme are ignored.

Kelly's coefficient of skewness, denoted by Sp is given by

$$S_P = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})}$$

$$= \frac{P_{90} - 2.P_{50} + P_{10}}{P_{90} - P_{10}}$$

Note that $P_{50} = M_d$ (median).

The value of Sp, for the data given in above Table, can be computed as given below.

Computation of P_{10}:

Since 10N/100 = 10 x 87 / 100 = 18.7. 10th percentile lies in the class 58.5 - 59.5. Thus

$$\therefore \quad P_{10} = 58.5 + \frac{18.7 - 10}{18} \times 1 = 58.983.$$

Computation of P_{90}:

Since 90N / 100 = 90 x 187 / 100 = 168.3, 90th percentile lies in the class 63.5 - 64.5. Thus

$l_{p90} = 63.5$, C = 163, $f_{p90} = 16$ and h = 1.

$$S_P = \frac{63.831 - 2 \times 61.345 + 58.983}{63.831 - 58.983} = 0.026.$$

Hence Kelly's Coefficient:

$$P_{90} = 63.5 + \frac{168.3 - 163}{16} \times 1 = 63.831.$$

It may be noted here that although the coefficient Sk, SQ and Sp, are not comparable, however, in the absence of skewness, each of them will be equal to zero.

KURTOSIS

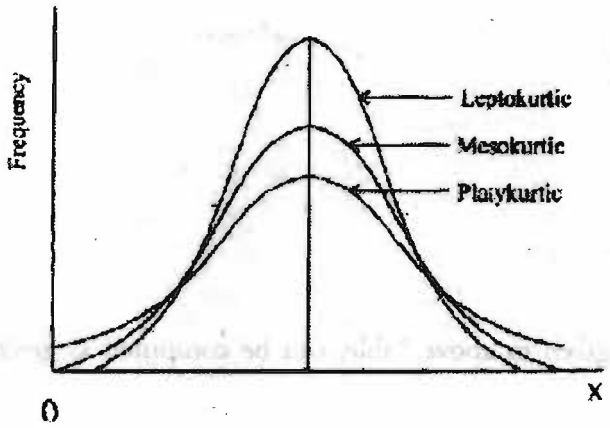
Kurtosis is another measure of the shape of a distribution. Whereas skewness measures the lack of symmetry of the frequency curve of a distribution, kurtosis is a measure of the relative peakedness of its frequency curve. Various frequency curves can be divided into three categories depending upon the shape of their peak. The three shapes are termed as Leptokurtic, Mesokurtic and Platykurtic as shown in Fig.

NOTES

Check your progress:

1) Define dispersion?
2) Why CV is useful?
3) What role standard deviation plays in the study of variation of data?

NOTES



A measure of kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

a coefficient given by Karl Pearson.

The value of $\beta 2 = 3$ for a mesokurtic curve. When $\beta 2 > 3$, the curve is more peaked than the mesokurtic curve and is termed as leptokurtic. Similarly, when $\beta 2 < 3$, the curve is less peaked than the mesokurtic curve and is called as platykurtic curve.

Example: The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Examine the skewness and kurtosis of the distribution. To examine skewness, we compute $\beta 1$.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.031$$

Since $\mu 3 > 0$ and $\beta 1$ is small, the distribution is moderately positively skewed.

Kurtosis is given by the coefficient

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3.0 .$$

Hence the curve is mesokurtic.

SUMMARY

- Dispersion is an important feature of the observations and it is measured with the help of the measures of dispersion, scatter or variation.
- The idea of dispersion is important in the study of wages of workers, prices of commodities, standard of living of different people, distribution of wealth, distribution of land among framers and various other fields of life.
- Absolute Measures of Dispersion gives us an idea about the amount of dispersion in a set of observations. They give the answers in the same units as the units of the original observations.

- Relative Measure of Dispersion is calculated for the comparison of dispersion in two or more than two sets of observations. These measures are free of the units in which the original data is measured.
- Range is defined as the difference between the maximum and the minimum observation of the given data.
- Quartile deviation is based on the lower quartile Q_1 and the upper quartile Q_3. The difference $Q_3 - Q_1$ is called the inter quartile range. The difference $Q_3 - Q_1$ divided by 2 is called semi-inter-quartile range or the quartile deviation. Thus, Quartile Deviation (Q.D)

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

- The mean deviation or the average deviation is defined as the mean of the absolute deviations of observations from some suitable average which may be the arithmetic mean, the median or the mode.
- The standard deviation is defined as the positive square root of the mean of the square deviations taken from arithmetic mean of the data.
- Variance is another absolute measure of dispersion. It is defined as the average of the squared difference between each of the observations in a set of data and the mean.
- Lack of symmetry is called Skewness. If a distribution is not symmetrical then it is called skewed distribution. The skewness may be positive or negative.
- The skewness value can be positive or negative, or even undefined. Qualitatively, a negative skew indicates that the tail on the left side of the probability density function is longer than the right side and the bulk of the values (possibly including the median) lie to the right of the mean. A positive skew indicates that the tail on the right side is longer than the left side and the bulk of the values lie to the left of the mean. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically but not necessarily implying a symmetric distribution.
- In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population.
- One common measure of kurtosis, originating with Karl Pearson, is based on a scaled version of the fourth moment of the data or population, but it has been argued that this measure really measures heavy tails, and not peakedness. For this measure, higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations. It is common practice to use an adjusted version of Pearson's kurtosis, the excess kurtosis, to provide a comparison of the shape of a given distribution to that of the nor-

NOTES

mal distribution. Distributions with negative or positive excess kurtosis are called platykurtic distributions or leptokurtic distributions respectively.

ANSWER TO CHECK YOUR PROGRESS

1. The degree to which numerical data tend to spread about an average value is called the dispersion or variation of the data.
2. Coefficient of Variance is also very useful when comparing two or more sets of data that are measured in different units of measurement.
3. The standard deviation plays a dominating role for the study of variation in the data. It is a very widely used measure of dispersion.

TEST YOURSELF

1) What do you mean by Dispersion?
2) Explain Range and Coefficient of Range.
3) Write a short note on:
   i) Quartile deviation
   ii) Mean deviation
   iii) Standard deviation
4) Following are the marks of 10 students of a college. Find the range and the coefficient of range. Marks: 400, 450, 520, 380, 485, 495, 575, 440, 410, 460
5) Calculate the mean deviation from mean and its coefficients from the following data.

Size	4 - 5	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10
Frequency	4	6	20	60	75	22

6) Calculate coefficient of standard deviation and coefficient of variation from the following distribution of marks:

Marks	No. of Students
3 - 5	50
5 - 7	40
7 - 9	30
9 - 11	20

7) Explain different measures of Skewness.
8) Write a short note on Kurtosis.

4 REGRESSION AND CORRELATION

NOTES

Chapter Includes :

- INTRODUCTIO
- REGRESSION
- CORRELATION
- RANGE OF COR LATION
- TYPES OF CORRELATION
- CORRELATION COEFFICIENT
- METHODS OF STUDYING CORRELATION
- RANK CORRELATION COEFFICIENT — (SPEARMAN'S)
- METHOD OF CONCURRENT DEVIATIONS
- STEPS
- TYPES OF DATA

Learning Objective :

After going through this chapter, you should be able to:

- Understand tegression
- Explain range of correlation
- Know types of correlation
- Understand methods of studying correlation

NOTES

Introduction:

Quantitative techniques are important tools of analysis in today's research in Economics. These tools can be broadly divided into two classes: mathematical tools and statistical tools. Economic research is often concerned with theorizing of some economic phenomenon. Different mathematical tools are employed to express such a theory in a precise mathematical form. This mathematical form of economic theory is what is generally called a in arithmetical model. A major purpose of the formulation of a mathematical model is to subject it to further mathematical treatment to gain a deeper understanding of the economic phenomenon that the researcher may be primarily interested in. However, the theory so developed needs to be tested in the real-world situation. In other words, the usefulness of a mathematical model depends on its empirical verification. Thus, in economic research, often a researcher is hard-pressed to put the mathematical model in such a form that it can render itself to empirical verification. For this purpose, various statistical techniques have been found to be extremely useful. We should note here, that often such techniques have been appropriately modified to suit the purposes of the economists. Consequently, a very rich and powerful area of economic analysis known as Econometrics has grown over the years. We may provide a working definition of Econometrics here. It may be described as the application of statistical tools in the quantitative analysis of economic phenomena. We may mention here that econometricians have not only provided important tools for economic analysis but also their contributions have significantly enriched the subject matter of Statistical Science in general. Today, no researcher can possibly ignore the need for being familiar with econometric tools for the purpose of serious empirical economic analysis.

The concepts of correlation and regression form the core of regression models. In this chapter we are going to put the two concepts in the perspective of empirical research in Economics. Here, our emphasis will be on examining how the applications of these two concepts are important in studying the possibility of relationship that may exist among economic variables.

REGRESSION

The term regression literally means a backward movement. Francis Galton first used the term in the late nineteenth century. He studied the relationship between the height of parents and that of children. Galton observed that although tall parents had tall children and similarly short parents had short children in a statistical sense, but in general the children's height tended towards an average value. In other words, the children's height moved backward or regressed to the average. However, now the term regression in statistics has nothing to do with its earlier connotation of a backward movement.

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to a very important concept known as 'Regression Analysis'.

Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is

NOTES

one of the most important statistical tools which is extensively used in almost all sciences – Natural, Social and Physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for the estimation of demand and supply graphs, cost functions, production and consumption functions and so on.

Prediction or estimation is one of the major problems in almost all the spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profits, income etc. are of very great importance to business professionals. Similarly, population estimates and population projections, GNP, Revenue and Expenditure etc. are indispensable for economists and efficient planning of an economy.

Regression analysis was explained by M. M. Blair as follows:

"Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data."

Regression analysis can be described as the study of the dependence of one variable on another or more variables. In other words, we can use it for examining the relationship that may exist among certain variables. For example, we may be interested in issues like how the aggregate demand for money depends upon the aggregate income level in an economy. We may employ regression technique to examine this. Here, Aggregate demand for money is called the dependent variable and aggregate income level is called the independent variable. Consequently, we have a simple demand for money function. In this context, we present the following table to show some of the terms that are also used in the literature in place of dependent variable and independent variable.

Table 1: Classifying Terms for Variables in Regression Analysis

Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Regressand	Regressor
Predictand	Predictor
Endogenous Variable	Exogenous Variable
Controlled Variable	Control Variable
Target Variable	Control Variable
Response Variable	Stimulus Variable

It is now important to clarify that the terms dependent and independent do not necessarily imply a causal connection between the two types of variables. Thus, regression analysis per-se is not really concerned with causality analysis. A causal connection has to be established first by some theory that is outside the parlance of the regression analysis. In our earlier example of consumption function and the present example of demand for money function we have theories like Keynesian income hypothesis and transaction demand for money. On the basis of such theo-

NOTES

ries perhaps we can employ regression technique to get some preliminary idea of some causal connection involving certain variables. In fact, causality study is now a highly specialized branch of econometrics and goes far beyond the scope of the ordinary regression analysis.

A major purpose of regression analysis is to predict the value of one variable given the value of another or more variables. Thus, we may be interested in predicting the aggregate demand of money from a given value of aggregate income.

We should be clear that by virtue of the very nature of economics and other branches of social science, the concern is a statistical relationship involving some variables rather than an exact mathematical relationship as we may obtain in natural science.

Consequently, if we are able to establish some kind of a relationship between an independent variable X and a dependent variable Y, it can be expected to give us only sort of an average value of Y for a given value of X. This kind of a relationship is known as a statistical or stochastic relationship. Regression method is essentially concerned with the analysis of such kind of a stochastic relationship.

From the above discussion, it should be clear that in our context, the dependent variable is assumed to be stochastic or random. In contrast, the independent variables are taken to be non-stochastic or non-random. However, we must mention here that at an advanced level, even the independent variables are assumed to be stochastic.

If a regression relationship has just one independent variable, it is called a two variable or simple regression. On the other hand, if we have more than one independent variable in it, then it is multiple regressions.

Correlation and Regression

Earlier we made a reference to the conceptual difference between correlation and regression. We may discuss it here. In regression analysis, we examine the **nature** of the relationship between the dependent and the independent variables. Here, as stated earlier, we **try** to estimate the average value of one variable from the given values of other variables. In correlation, on the other hand, our focus is on the measurement of the strength of such a relationship. Consequently, in regression, we classify the variables in two classes of dependent and independent variables. In correlation, the treatment of the variables is rather symmetric; we do not have such kind of a classification. Finally, in regression, at our level, we take the dependent variable as random or stochastic and the independent variables as non-random or fixed. In correlation, in contrast, all the variables **are** implicitly taken to be random in nature.

Simple Regression

Here, we **are** focusing on just one independent variable. The first thing that we have to do is to specify the relationship between X and **Y.** Let us assume that there is a linear relationship between the two variables like:

$$Y = a + bX$$

NOTES

The concept of linearity, however, requires some clarification. Moreover, there can be various types of intrinsically non-linear relationships also. The treatment of such relationships is beyond our **scope.** Our purpose is to estimate the constants a and b from empirical observations on X and Y.

The Method of Least Squares

Usually, we have a sample of observations of a given size say, n. If we plot the n pairs of observations, we obtain a scatter-plot, as it is known in the literature. **An** example of a scatter-plot is presented below.

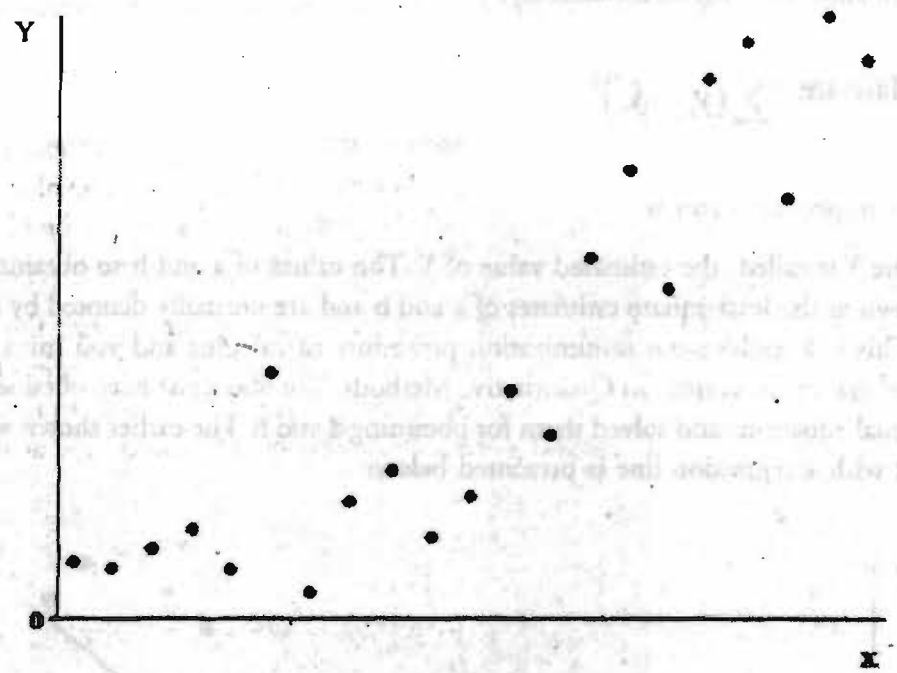


Figure: Scatter Plot

A visual inspection of the scatter-plot makes it clear that for different values of X, the corresponding values of Yare not aligned on a straight line. **As** we have mentioned earlier, in regression, we are concerned with an inexact or statistical relationship.

And this is the consequence of such a relationship. Now, the **constants** a and **b** are respectively the intercept and slope of the straight line described by the abovementioned **linear** equation and several straight lines with different pairs of the values (a, **b)** can be passed through the above scatter. Our concern is the choice of a particular pair as the estimates of a and **b** for the regression equation under consideration. Obviously, this calls for an objective criterion.

Such a criterion is provided by the method of least squares. The philosophy behind the least squares method is that we should fit in a straight line through the scatter plot, in such a manner that the vertical differences between the observed values of Y and the corresponding values obtained from the straight line for different values of X, called **errors,** are minimum. The line fitted in such a fashion is called the regression line. The values of a and b obtained from the

regression line are taken to be the estimates of the intercept and slope (regression coefficient) of the regression equation.

NOTES

The values of Y obtained from regression line are called the estimated values of Y. A stylized scatter-plot with a straight line fitted in it is presented below:

The method of least square requires that we should choose our a and b in such a manner that sum of the squares of the vertical differences between the actual values or observed values of Y and the ones obtained from the straight line is **minimum. Putting** mathematically,

$$\text{Minimize} \quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

with respect to a and b

where $\hat{Y}$ is called the estimated value of Y. The values of a and b so obtained are known as the least-square estimates of a and b and are normally denoted by $\hat{a}$ and $\hat{b}$. This is a well-known minimization procedure of calculus and you must have done that in the course on Quantitative Methods. You also must have obtained the normal equations and solved them for obtaining $\hat{a}$ and $\hat{b}$. The earlier shown scatter plot with a regression line is presented below:

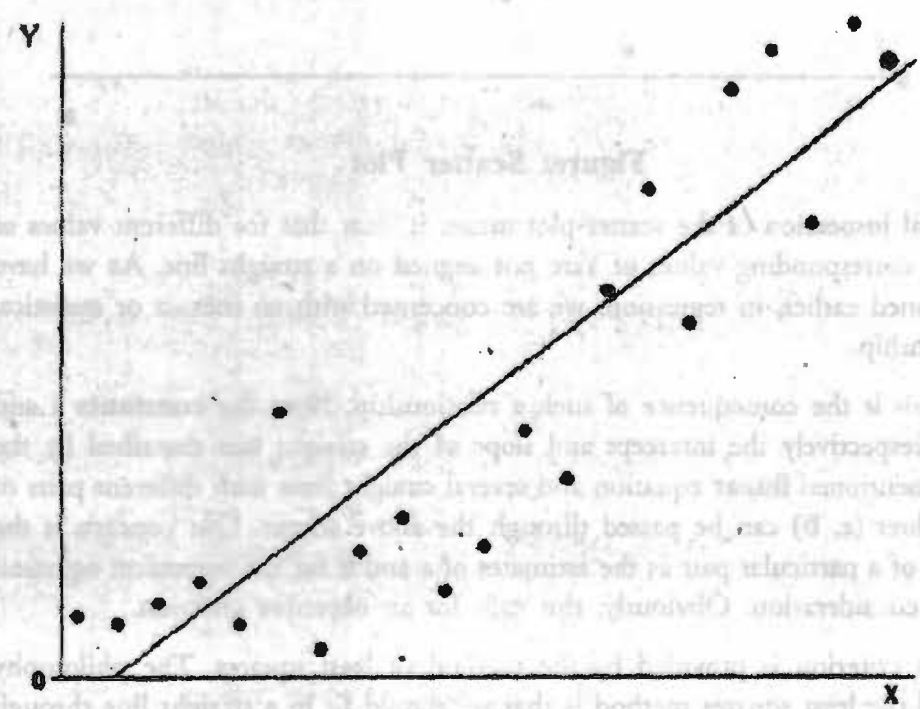


Figure: Scatter plot with the regression line

This regression line, obviously, has a negative intercept. If we recapitulate, the two normal equations that we obtained from the above-mentioned procedure are given by

$$\sum Y = na + b\sum X$$

And

$$\sum XY = a\sum X = b\sum X^2$$

After solving the two equations simultaneously we obtain the least square estimates

$$\hat{b} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

And

$$\hat{a} = \bar{Y} - \hat{b}\,\bar{X}.$$

In the regression analysis, the slope coefficient assumes special significance. It measures the rate of change of the dependent variable with respect to the independent variable. As a result, it is this constant that indicates whether there exists a relationship between X and Y or not. The regression equation

$$Y = a + bX$$

is in fact called the regression of Y on X, the slope b of this equation is termed as the regression coefficient of Y on X. It is also denoted by b_{yx} . A glance at the expression of the regression coefficient Yon X makes it quite clear that the above expression can also be written as

$$\hat{b} = r\frac{\sigma_Y}{\sigma_X}$$

Thus, putting the values of a and b, the regression equation of Y on X can be written as

$$Y - \bar{Y} = b_{YX}(X - \bar{X}) = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

Reverse Regression

Suppose, in another regression relationship X acts as the dependent variable and Y as the independent variable. Then that relationship is called the regression of X on Y. Here, we should dewtely avoid the temptation of expressing X in terms of Y from the regression equation of Y on X to obtain that of X on Y and trying to mechanically extract the least square estimates of its constants from the already

NOTES

known values of B and i . The regression of X on Y is in fact intrinsically different from that of Y on X. Geometrically speaking, in regression of X on Y, we **minimize** the sum of the squares of the horizontal distances as against the minimization of the sum of the squares of the vertical distances in Yon X, for obtaining the least square estimates. If our regression equation of X on Y is given by

$$X = a' + b'Y,$$

then its least square estimates are given by the criterion:

Minimize

$$\sum_{i=1}^{n}(X_i - \hat{X}_i)^2$$

with respect to a' and b'

By applying the usual minimization procedure, we obtain the following two normal equations:

$$\sum X = na' + b'\sum Y$$

And

$$\sum XY = a'\sum Y + b'\sum Y^2$$

We can simultaneously solve these two equations to get the least square estimates

$$\hat{b}' = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (Y - \bar{Y})^2} = r\frac{\sigma_X}{\sigma_Y}$$

And

$$\hat{a}' = \bar{X} - \hat{b}'\bar{Y}$$

The slope b' of the regression of X on Y is called the regression coefficient of X on Y. It measures the rate of change of X with respect to Y, in order to distinguish it clearly from the regression coefficient of Y on X; we also use the symbol b, for it.

Putting the values of a' and b', the regression equation of X on Y can be written as

$$X - \bar{X} = b_{XY}(Y - \bar{Y}) = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

To highlight the inherent difference between the two kinds of regression, the

regression of Y on X is sometimes termed as the direct regression and that of the X on Y is called the reverse regression. Maddala (2002) gives an example of direct regression and reverse regression in connection with the issue of gender bias in the offer of emoluments. Let us assume that the variable X represents qualifications and the variable Y represents emoluments. We may be interested in finding whether males and females with the same qualifications receive the same emoluments or not. We may examine this by running the direct regression of Y on X. Alternatively, we may be curious about if males and females with the same emoluments possess the same qualifications or not. We may investigate into this by running the reverse regression of X on Y. Thus, it is perhaps valid to **run** both the regressions in order to have a clear insight into the question of gender bias in emoluments.

Properties

Let us now briefly consider some of the properties of the regression.

1) The product of the two regression coefficients is **always** equal to the square of the correlation coefficient:

$$b_{YX} \times b_{XY} = r\frac{\sigma_Y}{\sigma_X} \times r\frac{\sigma_X}{\sigma_Y} = r^2$$

2) The two regression coefficients have the same sign. In fact, the sign of the two coefficients depends upon the sign of the correlation coefficient. Since the standard deviations of both X and Y are, by definition, positive; if correlation coefficient is positive, both the regression coefficients are positive and similarly, if correlation coefficient happens to be negative, both the regression coefficients become negative.
3) The two regression lines always intersect each other at the point $(\bar{X}, \bar{Y})$.
4) When $r = \pm 1$, there is an exact linear relationship between X and Y and in that case, the two regression lines coincide with each other.
5) When r = 0, the two regression equations reduce to $Y = \bar{Y}$ and $Y = \bar{Y}$. In such a situation, neither Y nor X can be estimated from their respective regression equations.

As mentioned earlier, coefficient of determination is an important concept in the context of regression analysis. However, the concept will be more contextual if we discuss it in the next unit.

Example:

From the following results, obtain the two regression equations and the estimate of the yield of crop, when the rainfall is 22 cm; and the rainfall, when the yield is 600 kg.

	Yield in kg	Rainfall in cm
Mean	508.4	26.7
Standard Deviation	36.8	4.6

NOTES

Co-efficient of correlation between yield and rainfall = 0.52.

Let Y be yield and X be rainfall. So, for estimating the yield, we have to run the regression of Y on X and for the purpose of estimating the rainfall, we have to use the regression of X on Y.

We have,

$$\bar{X} = 26.7$$

$$\bar{Y} = 508.4$$

$$\sigma_X = 4.6$$

$$\sigma_Y = 36.8$$

$$r = 0.52$$

Hence Regression coefficients:

$$b_{YX} = 0.52 \times \frac{36.8}{4.6} = 4.16$$

$$b_{XY} = 0.52 \times \frac{4.6}{36.8} = 0.065$$

Hence, the regression equation of Y on X is:

$$Y - 508.4 = 4.16(X - 26.7)$$

$$or\ Y = 4.16X + 397.33$$

Similarly the regression equation of X on Y is:

$$X - 26.7 = 0.065(Y - 508.4)$$

$$or\ X = 0.065Y - 6.346$$

When X = 22, $Y = 4.16 \times 22 + 397.33 = 488.8$

When Y = 600, $Y = 0.065 \times 600 - 6.346 = 32.7$

Hence, the estimated yield of crop is 488.8 kg and the estimated rainfall is 32.7 cm.

CORRELATION

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect.

NOTES

People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5" is less than the average weight of people 5'6", and their average weight is less than that of people 5'7", etc. Correlation can tell you just how much of the variation in peoples' weights is related to their heights.

Correlation is a technique which measures the strength of association between two variables. Both the variables X and Y may be random or may be that one variable is independent (non-random) and the other to be correlated are dependent. When the changes in one variable appear to be linked with the changes in the other variable, the two variables are said to be correlated. When the two variables are meaningfully related and both increase or both decrease simultaneously, then the correlation is termed as positive. If increase in any one variable is associated with decrease in the other variable, the correlation is termed as negative or inverse. Suppose marks in Mathematics are denoted by X and marks are Statistics are denoted by Y. If small values of X appear with small values of Y and large values of X come with large values of Y, then correlation is said to be positive. If X stands for marks in English and Y stands for marks in Mathematics, it is possible that small values of X appear with large values of Y. It is a case of negative correlation.

Although this correlation is fairly obvious your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data.

According to W.I. King, "Correlation means that between two series or groups of data there exists some casual connections."

According to Croxton and Cowden, "The appropriate statistical tool for discovering and measuring the relationship of quantitative nature and expressing it in brief formula is known as correlation."

RANGE OF CORRELATION

Correlation is computed into what is known as the correlation coefficient, which ranges between -1 and +1. Perfect positive correlation (a correlation co-efficient of +1) implies that as one security moves, either up or down, the other security will move in lockstep, in the same direction.

Alternatively, perfect negative correlation means that if one security moves in either direction the security that is perfectly negatively correlated will move by an equal amount in the opposite direction. If the correlation is 0, the movements of the securities is said to have no correlation, it is completely random. If one security moves up or down there is as good a chance that the other will move either up or down, the way in which they move is totally random

Check your progress:

1) What is Regression Analysis?
2) Define Correlation.

USEFULNESS OF CORRELATION

1) Correlation is very useful to economists to study the relationship between variables, like price and quantity demanded. To businessmen, it helps to

NOTES

estimate costs, sales, price and other relative variables.

2) Some variables show some kind of relationship; correlation analysis helps in measuring the degree of relationship between the variables like supply and demand, price and supply, income and expenditure, etc.

3) The relation between variables can be verified and tested for significance, with the help of correlation analysis. The effect of correlation is to reduce the range of uncertainty of our prediction.

4) The coefficient of correlation is a relative measure and we can compare the relationship between variables, which are expressed in different units.

5) Sampling error can also be calculated.

6) Correlation is the basis for the concept of regression and ration of variation.

7) The decision making is heavily felicitated by reducing the range of uncertainty and hence empowering the predictions.

TYPES OF CORRELATION

There are several different correlation techniques. The Survey System's optional Statistics Module includes the most common type, called the Pearson or product-moment correlation. The module also includes a variation on this type called partial correlation. The latter is useful when you want to look at the relationship between two variables while removing the effect of one or two other variables.

Like all statistical techniques, correlation is only appropriate for certain kinds of data. Correlation works for quantifiable data in which numbers are meaningful, usually quantities of some sort. It cannot be used for purely categorical data, such as gender, brands purchased, or favorite color.

Linear Correlation:

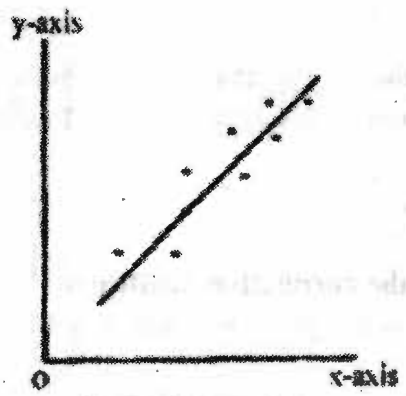
Correlation is said to be linear if the ratio of change is constant. The amount of output in a factory is doubled by doubling the number of workers is the example of linear correlation.

In other words it can be defined as if all the points on the scatter diagram tends to lie near a line which are look like a straight line, the correlation is said to be linear, as shown in the figure

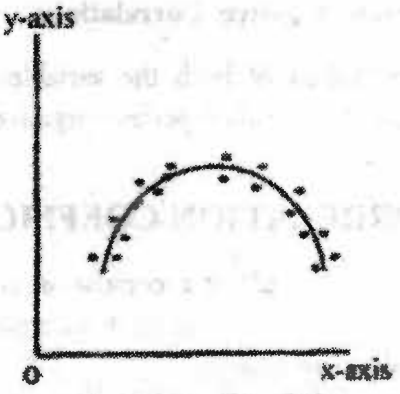
Non Linear (Curvilinear) Correlation:

Correlation is said to be non linear if the ratio of change is not constant. In other words it can be defined as if all the points on the scatter diagram tends to lie near a smooth curve, the correlation is said to be non linear (curvilinear), as shown in the figure.

NOTES



Linear Correlation



Non Linear Correlation

Positive Correlation:

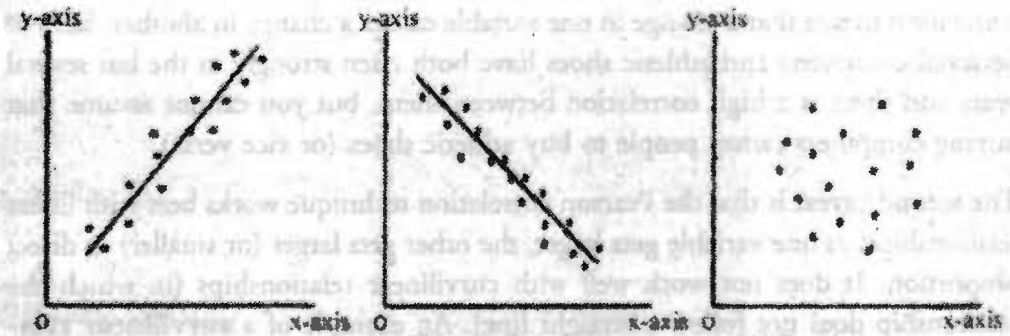
The correlation in the same direction is called positive correlation. If one variable increase other is also increase and one variable decrease other is also decrease. For example, the length of an iron bar will increase as the temperature increases.

Negative Correlation:

The correlation in opposite direction is called negative correlation, if one variable is increase other is decrease and vice versa, for example, the volume of gas will decrease as the pressure increase or the demand of a particular commodity is increase as price of such commodity is decrease.

No Correlation or Zero Correlation:

If there is no relationship between the two variables such that the value of one variable change and the other variable remain constant is called no or zero correlation.



Perfect Correlation:

If there is any change in the value of one variable, the value of the others variable is changed in a fixed proportion, the correlation between them is said to be perfect correlation. It is indicated numerically as +1 and -1.

Perfect Positive Correlation:

If the values of both the variables are move in same direction with fixed proportion is called perfect positive correlation. It is indicated numerically as +1.

NOTES

Perfect Negative Correlation:

If the values of both the variables are move in opposite direction with fixed proportion is called perfect negative correlation. It is indicated numerically as -1.

CORRELATION COEFFICIENT

The main result of a correlation is called the **correlation coefficient** (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).

While correlation coefficients are normally reported as r = (a value between -1 and +1), squaring them makes then easier to understand. The square of the coefficient (or r square) is equal to the percent of the variation in one variable that is related to the variation in the other. After squaring r, ignore the decimal point. An r of .5 means 25% of the variation is related (.5 squared =.25). An r value of .7 means 49% of the variance is related (.7 squared = .49).

A correlation report can also show a second result of each test - statistical significance. In this case, the significance level will tell you how likely it is that the correlations reported may be due to chance in the form of random sampling error. If you are working with small sample sizes, choose a report format that includes the significance level. This format also reports the sample size.

A key thing to remember when working with correlations is never to assume a correlation means that a change in one variable causes a change in another. Sales of personal computers and athletic shoes have both risen strongly in the last several years and there is a high correlation between them, but you cannot assume that buying computers causes people to buy athletic shoes (or vice versa).

The second caveat is that the Pearson correlation technique works best with linear relationships: as one variable gets larger, the other gets larger (or smaller) in direct proportion. It does not work well with curvilinear relationships (in which the relationship does not follow a straight line). An example of a **curvilinear relationship** is age and health care. They are related, but the relationship doesn't follow a straight line. Young children and older people both tend to use much more health care than teenagers or young adults. Multiple regression (also included in the Statistics Module) can be used to examine curvilinear relationships, but it is beyond the scope of this article.

Coefficient of Correlation:

The degree or level of correlation is measured with the help of correlation coefficient or coefficient of correlation. For population data, the correlation coefficient is denoted by ρ. The joint variation of X and Y is measured by the

covariance of X and Y. The covariance of X and Y denoted by $Cov(X, Y)$ is defined as:

$$Cov(X,Y) = E[X - E(X)][Y - E(Y)]$$

The $Cov(X, Y)$ may be positive, negative or zero. The covariance has the same units in which X and Y are measured. When $Cov(X, Y)$ is divided by σ_x and σ_y, we get the correlation coefficient ρ. Thus

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

ρ is free of the units of measurement. It is a pure number and lies between -1 and +1. If $\rho = \pm 1$, it is called as perfect correlation. If $\rho = -1$, it is called perfect negative correlation. If there is no correlation between X and Y, then X and Y are independent and $\rho = 0$, For sample data the correlation coefficient denoted by "r" is a measure of strength of the linear relation between X and Y variables where "r" is a pure number and lies between -1 and +1. On the other hand Karl Pearson's coefficient of correlation is:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

Properties of Coefficient of Correlation

- The correlation coefficient is symmetrical with respect to X and Y i.e. $\tau_{XY} = \tau_{YX}$.
- The correlation coefficient is the geometric mean of the two regression coefficients. $r = \sqrt{b_{YX} \times b_{XY}}$ or $r = \sqrt{b \times \hat{d}}$.
- The correlation coefficient is independent of origin and unit of measurement i.e. $\tau_{XY} = \tau_{UV}$.
- The correlation coefficient lies between -1 and +1. i.e. $-1 \leq \tau \leq +1$.

Examples:

Calculate and analyze the correlation coefficient between the number of study hours and the number of sleeping hours of different students.

Number of Study hours	2	4	6	8	10
Number of sleeping hours	10	9	8	7	6

NOTES

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
2	10	-4	+2	-8	16	4
4	9	-2	+1	-2	4	1
6	8	0	0	0	0	0
8	7	+2	-1	-2	4	1
10	6	+4	-2	-8	16	1
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma(X-\bar{X}) = 0$	$\Sigma(Y-\bar{Y}) = 0$	$\Sigma(X-\bar{X})(Y-\bar{Y}) = -20$	$\Sigma(X-\bar{X})^2 = 40$	$\Sigma(Y-\bar{Y})^2 = 10$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{30}{5} = 6 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{40}{5} = 8$$

$$r_{XY} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2 \Sigma(Y-\bar{Y})^2}} = \frac{-20}{20} = -1$$

There is perfect negative correlation between the number of study hours and the number of sleeping hours.

Example:

From the following data, compute the coefficient of correlation between *X* and *Y*:

	X Series	*Y* Series
Number if Items	15	15
Arithmetic Mean	25	18
Sum of Square Deviations	136	138

Summation of products of deviations of *X* and *Y* series from their arithmetic means = **122.**

Solution:

Here $n = 15$, $\bar{X} = 25$, $\bar{Y} = 18$, $\Sigma(X-\bar{X})^2 = \Sigma(Y-\bar{Y})^2 = 138$

$\Sigma(X-\bar{X})^2(Y-\bar{Y})^2 = 122$ and hence

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2 \Sigma(Y-\bar{Y})^2}} = \frac{122}{\sqrt{(136)(138)}} = \frac{122}{137} = 0.89$$

NOTES

METHODS OF STUDYING CORRELATION:

SCATTER DIAGRAM METHOD:

This is the simplest device for ascertaining whether two variables are related is to prepare a dot chart called scatter diagram. The given data are plotted on a graph paper in the form of dots.

Merits and demerits of the method;-

Merits- It is the simplest form of studying correlation.

It is not influenced by the size of extreme items whereas most of mathematical methods are influenced by extreme figures.

Demerits; - We can get idea of correlation but we cannot find out exact degree of correlation.

KARL PEARSON'S CORRELATION COEFFICIENT (r):-

In statistics, the Pearson product-moment correlation coefficient (r) is a common measure of the correlation between two variables X and Y. When measured in a population the Pearson Product Moment correlation is designated by the Greek letter rho (?). When computed in a sample, it is designated by the letter "r" and is sometimes called "Pearson's r." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between variables. A correlation of -1 means that there is a perfect negative linear relationship between variables. A correlation of 0 means there is no linear relationship between the two variables. Correlations are rarely if ever 0, 1, or -1. If you get a certain outcome it could indicate whether correlations were negative or positive.

Mathematical Formula:—

The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson.

The mathematical formula for computing r is:

$$r = \frac{\frac{\sum XY}{N} - (\frac{\sum X}{N})(\frac{\sum Y}{N})}{\sqrt{\frac{\sum X^2}{N} - (\frac{\sum X}{N})^2 * \frac{\sum Y^2}{N} - (\frac{\sum Y}{N})^2}}$$

Where:

N = represents the number of pairs of data

Σ = denotes the summation of the items indicated

ΣX = denotes the sum of all X scores

ΣX^2= indicates that each X score should be squared and then those squares summed

NOTES

$(\Sigma X)^2$= indicates that the X scores should be summed and the total squared. [avoid confusing ΣX^2 (the sum of the X squared scores) and $(\Sigma X)^2$ (the square of the sum of the X scores]

ΣY= denotes the sum of all y-scores

ΣY^2 = indicates that each Y score should be squared and then those squares summed

$(\Sigma Y)^2$= indicates that the Y scores should be summed and the total squared

ΣXY= indicates that each X score should be first multiplied by its corresponding Y score and the product (XY) summed

The numerator in equation 1 equals the mean of XY ($\overline{XY}$) minus the mean of X ($\overline{X}$) times the mean of Y ($\overline{Y}$); the denominators are the standard deviation for X (SD_X) and the standard deviation for Y (SD_Y). [See; *How to compute and interpret measures of variability: the range, variance and standard deviation*] Thus, Pearson's formula can be written as:

$$r = \frac{XY - \overline{XY}}{SDX * SDY}$$

Example

Compute the correlation coefficient (r) for the height-weight data shown in Figure 1. Pertinent calculations are given in Table 1.

Table 1. Height and weight of a sample of college age males.

Height, cm (X)	Weight, kg (Y)	X * Y	X2	Y2
174	61	10614	30276	3721
175	65	11375	30625	4225
176	67	11792	30976	4489
177	68	12036	31329	4624
178	72	12816	31684	5184
182	74	13468	33124	5476
183	80	14640	33489	6400
186	87	16182	34596	7569
189	92	17388	35721	8464
193	95	18335	37249	9025
ΣX=1813	ΣY=761	ΣXY=138646	ΣX^2=329069	ΣY^2=59177
$\Sigma X/N$=181.3	$\Sigma Y/N$=76.1	$\Sigma XY/N$=13864.6	$\Sigma X^2/N$=32906.9	$\Sigma Y^2/N$=5917.7

NOTES

$$r = \frac{\bar{XY} - \bar{X}\bar{Y}}{SDX * SDY}$$

$$r = \frac{138646 - (181.3)(76.1)}{\sqrt{329069 - (181.3)^2 * (5917.7) - (76.1)^2}}$$

$$r = \frac{67.67}{68.605}$$

$$r = 0.986$$

Interpreting Pearson's Correlation Coefficient

The usefulness of the correlation depends on its size and significance. If *r* reliably differs from 0.00, the *r*-value will be statistically significant (i.e., does not result from a chance occurrence.) implying that if the same variables were measured on another set of similar subjects, a similar *r*-value would result. If *r* achieves significance we conclude that the relationship between the two variables was not due to chance.

How to Evaluate a Correlation

The values of *r* always fall between -1 and +1 and the value does not change if all values of either variable are converted to a different scale. For example, if the weights of the students in Figure 1 were given in pounds instead of kilograms, the value of *r* would not change (nor would the shape of the scatter plot.)

The size of any correlation generally evaluates as follows:

Correlation Value	Interpretation
≤0.50	Very low
0.51 to 0.79	Low
0.80 to 0.89	Moderate
≥0.90	High (Good)

A high (or low) negative correlation has the same interpretation as a high (or low) positive correlation. A negative correlation indicates that high scores in one variable are associated with low scores in the other variable.

Main Characteristics of Karl Pearson's Coefficient of Correlation:

a) It is an ideal measure of correlation and is independent of the units of X and Y.

b) It is independent of change of origin and scale.

c) It is based on all the observations.

d) It varies between -1 and +1:

NOTES

r = -1, when there is a perfect negative correlation

r = 0, when there is no correlation

r - +1, when there is a perfect positive correlation.

e) It does not tell anything about cause and effect relationship.

f) It is somehow difficult to calculate.

g) It requires some interpretation.

RANK CORRELATION COEFFICIENT: — (SPEARMAN'S)

Using ranks rather than actual observation gives coefficient of rank correlation. This measure is useful when quantitative measure for certain factors cannot be fixed but individual in a group can be arranged in order thereby obtaining for each individual a number indicating his/her rank in the group.

Spearman's rank correlation coefficient is defined as: -

$$R = 1 - (6\Sigma D2) / \{N (N2 - 1)\}$$

Where, R denotes rank coefficient of correlation and D refers to difference of rank between items in two series. The values of this coefficient interpreted in the same way as Karl Pearson's correlation coefficient, ranges between +1 and –1. Where, R is +1 there is complete agreement in order of rank & ranks are in same direction. Where R is –1, there is complete agreement in order of rank & they are in opposite direction.

Features: -

1) The sum of differences of rank between 2 variables shall be zero i.e., $\Sigma D = 0$.

2) It is distribution free or non parametric because no strict assumption as made about the form of population from which sample observations are drawn.

3) Spearman's correlation coefficient is nothing but Karl Pearson's correlation coefficient between ranks. Hence, it can be interpreted in same manner as Pearsonian correlation coefficient.

There are two types of problems in rank correlation:

1) **Where ranks are given:** - The following steps are required for computing rank correlation:

   - Take difference of 2 ranks i.e., (R1 - R2) & denote these differences by D.
   - Square these differences and obtain the total$\Sigma D2$.
   - Apply the formula

$$R = 1 - (6\Sigma D2) /\{N (N2 - 1)\}$$

2) **Where ranks are not given:** - When we are given actual data & not ranks, it will be necessary to assign ranks by taking either highest value as 1 or lowest value as 1 & rest follow same method as above.

METHOD OF CONCURRENT DEVIATIONS

Sometimes it is desired to study the correlation between two series in a very casual manner and in such cases no particular attention is needed so far as precision is concerned. In such cases it is enough to calculate the coefficient of concurrent deviations. In this method correlation is calculated between the direction of deviations and not their magnitudes. As such only the direction of deviations is taken into account in the calculation of this coefficient, and their magnitude is ignored.

To calculate the coefficient of concurrent deviations, the deviations are not calculated from any average or by the method of moving averages, but only their direction from the previous period, is noted down.

The coefficient of concurrent deviation or coefficient of correlation by the concurrent deviation method is given by the formula

$$r_c = \pm\Sigma(\pm\ ((2c-n)/n)$$

where, r_c = Coefficient of concurrent deviation

c = number of concurrent deviations

n = number of pairs of deviations compared=N-1

N = number of pairs of observations

STEPS

In this method, we take into account the direction of change in the values of two variables X and Y. To find r_c we proceed as follows:

- Take the first value of X as base and look whether the second value is greater than (i.e., increasing), less than (i.e., decreasing)or equal (i.e. constant) to first value. If the second value is greater than second value, mark a (+) sign against it, if the second value is less than: value, mark a (-) sign against it, and if it is equal to first value, put zero or a (=) sign against it. Similarly second value is the base for the third value and so on. Represent this column by a symbol D_x.
- Similarly obtain the column D_y from Y-series.
- Multiply D_x and D_y [pairs of deviations having like signs will be positive (+) otherwise (-)].
- Find out the value of c = number of (+) signs in the column D_x D_y.
- Finally use the formula:

$$r_c = \pm\sqrt{(\pm\ ((2c-n)/n)}$$

i) If 2c - n / n is negative, then we take negative sign inside and outside the symbol (√).

NOTES

Check your progress:

3) What do you mean by range of correlation?

4) What is coefficient of correlation?

NOTES

ii) If 2c – n / n is positive, then we take positive sign inside and outside the symbol (√).

TYPES OF DATA

We conclude this unit by discussing the types of data that may be used for the purpose of economic analysis in general and regression analysis in particular. We can use three kinds of **data** for the empirical verification of any economic phenomenon. They are: time series, cross section, pooled or panel data.

Time Series Data

A time series is a collection of the values of a variable that are observed at different points of time. Generally, the interval between two successive points of time remains fixed. In other words, we collect data at regular time intervals. Such data may be collected daily, weekly, monthly, quarterly or annually. We have for example, daily data series for gold price, weekly money supply figures, monthly price index, quarterly GDP series and annual budget data. Sometimes, we may have the same data in more than one time interval series; for example, both quarterly and annual GDP series may be available. The time interval is generally called the frequency of the time series. It should be clear that the above-mentioned list of time intervals is by no means an exhaustive one. There can be, for example, an hourly time series like that of stock price sensitivity index. Similarly, we may have decennial population census figures. We should note that conventionally, if the frequency is one year or more, it is called a low frequency time series. On the other hand, if the frequency is less than one year, it is termed as a high frequency time series. A major problem with time series is what is known as non-stationary data. The presence of non-stationarity is the main reason for nonsense correlation that we talked about in connection with our discussion on correlation.

Cross Section Data

In cross section data, we have observations for a variable for different units at the same point of time. For example, we have the state domestic product figures for different states in India for a particular year. Similarly, we may collect various stock price figures at the same point of time in a particular day. Cross section data are also not free from problems. One main problem with this kind of data is that of the heterogeneity that we shall refer to in the next unit.

Pooled Data

Here, we may have time series observations for various cross sectional **units.** For example, we may have time series of domestic product of each state for India and we may have a panel of such series. This is why such kind of a data set is called panel data. Thus, in this kind of data, we combine the element of time series with that of cross section data. One major advantage with such kind of data is that we may have quite a large data set and the problem of degrees of freedom that mainly arises due to the non-availability of adequate **data can** largely be overcome. Recently, the treatment of panel data **has** received much attention in empirical economic analysis.

SUMMARY

NOTES

- Regression models occupy a central place in empirical economic analysis. These models **are** essentially based on the concepts of correlation and regression. Correlation is a quantitative measure of the strength of the linear relationship that may exist among some variables. The existence of a high degree of correlation, however, is not necessarily the evidence of a meaningful relationship. It only suggests that the data **are** not inconsistent with the possibility of such kind of a relationship.
- Regression on the other hand focuses on the direction of a linear relationship. Here, one is concerned with the dependence of one variable on other variables. Regression, in itself, does not suggest any **causal** relationship. Correlation and regression, both are concerned with a statistical or stochastic relationship as against a mathematical or an exact relationship. In the conventional regression analysis, the dependent variable is treated to be stochastic or random, whereas, the independent variables are taken to, be non-stochastic in nature. The constants of a regression equation are estimated from the empirical observations by using the least square technique. In a two variable regression equation, there is one dependent variable and one independent variable. The slope coefficient of a regression equation is called the regression coefficient. It measures the rate of change of the dependent variable with respect to the independent variable. The distinction between the concept of direct regression and that of the reverse regression is crucial in the regression analysis. Sometimes by running both the kinds of regression, important insight can be gained in the empirical economic analysis. In multiple regression, there are at least two independent variables.
- Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. Regression analysis can be described as the study of the dependence of one variable on another or more variables. In other words, we can use it for examining the relationship that may exist among certain variables.
- Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.
- Correlation is a technique which measures the strength of association between two variables.
- Correlation is said to be linear if the ratio of change is constant. The amount of output in a factory is doubled by doubling the number of workers is the example of linear correlation.
- The correlation in opposite direction is called negative correlation, if one variable is increase other is decrease and vice versa, for example, the volume of gas will decrease as the pressure increase or the demand of a particular commodity is increase as price of such commodity is decrease.

NOTES

- If there is any change in the value of one variable, the value of the others variable is changed in a fixed proportion, the correlation between them is said to be perfect correlation. It is indicated numerically as +1 and -1.
- A correlation report can also show a second result of each test - statistical significance. In this case, the significance level will tell you how likely it is that the correlations reported may be due to chance in the form of random sampling error. If you are working with small sample sizes, choose a report format that includes the significance level. This format also reports the sample size.
- Using ranks rather than actual observation gives coefficient of rank correlation. This measure is useful when quantitative measure for certain factors cannot be fixed but individual in a group can be arranged in order thereby obtaining for each individual a number indicating his/her rank in the group.

Answers to check your progress

1) "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data."
2) Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people.
3) Correlation is computed into what is known as the correlation coefficient, which ranges between -1 and +1. Perfect positive correlation (a correlation co-efficient of +1) implies that as one security moves, either up or down, the other security will move in lockstep, in the same direction.
4) The degree or level of correlation is measured with the help of correlation coefficient or coefficient of correlation. For population data, the correlation coefficient is denoted by ρ.

Test Yourself

1) What do you mean by the term regression analysis?
2) Define Correlation and Range of Correlation.
3) What are the types of Correlation?
4) Explain 'Correlation Coefficient' and properties of Correlation Coefficient.
5) What do you mean by Karl Pearson's Correlation Coefficient (r)?
6) Explain main characteristics of Karl Pearson's Coefficient of Correlation.

5 INDEX NUMBERS AND TIME SERIES

Chapter Includes :

- INTRODUCTION
- INDEX NUMBERS
- SELECTION OF COMMODITIES
- COLLECTION OF PRICE DATA
- SELECTION OF THE SUITABLE AVERAGE
- WEIGHTED INDEX NUMBERS
- WHOLESALE PRICE INDEX NUMBERS
- CONSUMER PRICE INDEX NUMBER
- ANALYSIS OF TIME SERIES
- COMPONENTS OF TIME SERIES

Learning Objective :

After going through this chapter, you should be able to:

- Understant concept of Index Numbers
- Explain selectioin of commodities
- Discuss weighted index numbers
- Understand analysis of time series

NOTES

INTRODUCTION

Index numbers are meant to study the change in the effects of such factors which cannot be measured directly. According to Bowley, "Index numbers are used to measure the changes in some quantity which we cannot observe directly". For example, changes in business activity in a country are not capable of direct measurement but it is possible to study relative changes in business activity by studying the variations in the values of some such factors which affect business activity, and which are capable of direct measurement.

Index numbers are commonly used statistical device for measuring the combined fluctuations in a group related variables. If we wish to compare the price level of consumer items today with that prevalent ten years ago, we are not interested in comparing the prices of only one item, but in comparing some sort of average price levels. We may wish to compare the present agricultural production or industrial production with that at the time of independence. Here again, we have to consider all items of production and each item may have undergone a different fractional increase (or even a decrease). How do we obtain a composite measure? This composite measure is provided by index numbers which may be defined as a device for combining the variations that have come in group of related variables over a period of time, with a view to obtain a figure that represents the 'net' result of the change in the constitute variables.

Index numbers may be classified in terms of the variables that they are intended to measure. In business, different groups of variables in the measurement of which index number techniques are commonly used are (i) price, (ii) quantity, (iii) value and (iv) business activity. Thus, we have index of wholesale prices, index of consumer prices, index of industrial output, index of value of exports and index of business activity, etc. Here we shall be mainly interested in index numbers of prices showing changes with respect to time, although methods described can be applied to other cases. In general, the present level of prices is compared with the level of prices in the past. The present period is called the current period and some period in the past is called the base period.

INDEX NUMBERS

Index numbers are statistical measures designed to show changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc. A collection of index numbers for different years, locations, etc., is sometimes called an index series.

Simple Index Number:

A simple index number is a number that measures a relative change in a single variable with respect to a base.

Composite Index Number:

A composite index number is a number that measures an average relative changes in a group of relative variables with respect to a base.

NOTES

Types of Index Numbers:

Following types of index numbers are usually used:

Price index Numbers:

Price index numbers measure the relative changes in prices of a commodities between two periods. Prices can be either retail or wholesale.

Quantity Index Numbers:

These index numbers are considered to measure changes in the physical quantity of goods produced, consumed or sold of an item or a group of items.

Uses of Index Numbers

The main uses of index numbers are given below:

- Index numbers are used in the fields of commerce, meteorology, labour, industrial, etc.
- The index numbers measure fluctuations during intervals of time, group differences of geographical position of degree etc.
- They are used to compare the total variations in the prices of different commodities in which the unit of measurements differs with time and price etc.
- They measure the purchasing power of money.
- They are helpful in forecasting the future economic trends.
- They are used in studying difference between the comparable categories of animals, persons or items.
- Index numbers of industrial production are used to measure the changes in the level of industrial production in the country.
- Index numbers of import prices and export prices are used to measure the changes in the trade of a country.
- The index numbers are used to measure seasonal variations and cyclical variations in a time series.

Limitations of Index Numbers

- They are simply rough indications of the relative changes.
- The choice of representative commodities may lead to fallacious conclusions as they are based on samples.
- There may be errors in the choice of base periods or weights etc.
- Comparisons of changes in variables over long periods are not reliable.
- They may be useful for one purpose but not for other.
- They are specialized types of averages and hence are subject to all those limitations with which an average suffers from.

NOTES

Construct Price Index Numbers:

The following steps are considered for the construction of price index numbers:

Object:

The first and the most important steps in the construction of index numbers is to decide the object for making the index numbers of prices. The prices may be retail or whole-sale. The index numbers of retail prices are called the consumer price index (CPI) numbers and if the whole-sale prices are taken into consideration, the index numbers are called the whole-sale price index numbers. The index numbers of prices may be calculated for a certain locality, for a certain class of people like textile workers or office clerks etc. The index numbers may be required for geographical regions like districts or provinces etc. First of all we decide the purpose of making the index numbers. Once the purpose is decided, then we decide about the scope and the area or the people who are to be considered.

SELECTION OF COMMODITIES

A list of important commodities is prepared. Those commodities are taken into account which is commonly consumed by the consumers. There is no hard and fast rule about the number of commodities. Only those commodities are considered on which a reasonable amount is spent. The commodities on which the expenditure is meager or they are used only occasionally are not included in the list. Thus the commodities which are representative of the tastes and customs of the people are taken in the list. Dr. Irving Fisher has said that 20 commodities is a small number and 50 commodities is a reasonable number. For construction of wholesale price index numbers, about 80 commodities are taken in the list and for retail-price index numbers about 300 commodities are considered. Sometimes the index numbers of very important commodities like wheat, rice, oil, ghee etc. are calculated. These index numbers are based on about one dozen commodities and are called sensitive price index numbers.

COLLECTION OF PRICE DATA

The most important and difficult step is the collection of prices. The prices are to be taken from the field. For retail price index numbers, retail prices are needed. The prices change from place to place and from time to time. On different shops the prices are different. In actual practice there are many difficulties. Usually some representative shops from where the consumers mostly purchase their items are selected and the prices are taken from those shops. The prices are taken on daily basis and then the weekly and monthly averages are calculated. Finally quarterly or yearly averages are calculated. Some commodities are sold in different varieties. Rice, sugar, mangoes, etc. have different variables which are sold on different prices. This problem is solved by assigning due weights to different varieties and then weighted average prices is calculated. Sometimes different varieties are treated as different commodities.

For whole-sale prices, the prices are taken from the whole-sale markets, fractions depots and the whole-sale agencies. The whole-sale prices are usually stable, there-

fore these prices are not taken on daily basis. The price reporting is done on weekly or monthly basis depending upon the nature of the commodity. The prices of some commodities are controlled by the government. These prices are reported whenever some change takes place.

Selection of Base Period

The prices of the commodities in the current period are to be compared with the prices of some period in the past. This period in the past is called the base period or the reference period. The base period is decided statistical division of Government. This period should not be in the remote past. The period which is economically stable and is free of disturbances and strikes is taken as the base period.

Fixed Base Method

In fixed base method, a particular year is generally chosen arbitrarily and the prices of the subsequent years are expressed as relatives of the prices of the base year. Sometimes instead of choosing a single year as the base, a period of a few years is chosen and the average price of this period is taken as the base year's price. The year which is selected as a base should be normal year or in other words, the price level in this year should neither be abnormally low nor abnormally high. If an abnormal year is chosen as the base, the price relatives of the current year calculated on its basis would give misleading conclusions. For example, a year in which war was at its peak, say the year 1965, is chosen as a base year, the comparison of the price level of the subsequent years to the prices of 1965 is bound to give misleading conclusions. The reason is that the price level in the year 1965 was abnormally high. In order to remove this difficulty associated with the selection of a normal year, the average price of a few years is sometimes taken as the base price. The fixed base method is used by the Government in the calculation of national index numbers.

In Fixed Base,

$$\text{Price relative for current year} = \frac{\text{Price of Current Year}}{\text{Price of Base Year}} \times 100$$

$$P_{on} = \frac{P_n}{P_o} \times 100$$

Example:

Find index numbers for the following data taking 1980 as base year.

Years	1980	1981	1982	1983	1984	1985	1986	1987
Price	40	50	60	70	80	100	90	110

NOTES

Solution:

Years	Price	Index No's 1980 as base $Pon = \frac{P_n}{P_o} \times 100$
1980	40	$\frac{40}{40} \times 100 = 100$
1981	50	$\frac{50}{40} \times 100 = 125$
1982	60	$\frac{60}{40} \times 100 = 150$
1983	70	$\frac{70}{40} \times 100 = 175$
1984	80	$\frac{80}{40} \times 100 = 200$
1985	100	$\frac{100}{40} \times 100 = 250$
1986	90	$\frac{90}{40} \times 100 = 225$
1988	110	$\frac{110}{40} \times 100 = 275$

Chain Base Method

In this method, there is no fixed base period. The year immediately preceding the one for which price index have to be calculated is assumed as the base year. Thus, for the year1994 the base year would be 1993, for 1993 it would be 1992 for 1992 it would be 1991 and so on. In his way there is no fixed base. It goes on changing. The chief advantage of this method is that the price relatives of a year can be compared with the price level of the immediately preceding year. Businessmen mostly interested in comparison of this type rather than in comparison relating to distant past. Yet another advantage of the chain base method is that it is possible to include new items in an index number or to delete old times which are no more important. In fixed base method it is not possible. But chain base method has drawback that comparison cannot be made over a long period.

In Chain Base,

Link relative of current years

$$= \frac{\text{Price in the Current Year}}{\text{Price in the preceding Year}} \times 100$$

$$P_{n-1,n} = \frac{P_n}{P_{n-1}} \times 100$$

NOTES

Example:

Find index numbers for the following data taking 1980 as base year.

Years	1974	1975	1976	1977	1978	1979
Price	18	21	25	23	28	30

Solution:

Years	Price	Link Relatives $Pon = \frac{P_n}{P_{n-1}} \times 100$	Chain Indices
1974	18	$\frac{18}{18} \times 100 = 100$	100
1975	21	$\frac{21}{18} \times 100 = 116.67$	$\frac{100 \times 116.67}{100} = 116.67$
1976	25	$\frac{25}{21} \times 100 = 119.05$	$\frac{116.67 \times 119.05}{100} = 138.9$
1977	23	$\frac{23}{25} \times 100 = 92$	$\frac{138.9 \times 92}{100} = 127.79$
1978	28	$\frac{28}{23} \times 100 = 121.74$	$\frac{127.79 \times 121.74}{100} = 155.57$
1979	30	$\frac{30}{28} \times 100 = 107.14$	$\frac{155.57 \times 107.17}{100} = 166.68$

SELECTION OF THE SUITABLE AVERAGE

There are different averages which can be used in averaging the price relatives or link relatives of different commodities. Experts have suggested that the geometric mean should be calculated for averaging these relatives. But as the calculation of the geometric mean is difficult, it is mostly avoided and the arithmetic mean is commonly used. In some cases the median is used to remove the effect of the wile observations.

Selection of Suitable Weights

In calculation of price index numbers all commodities are not of equal importance. In order to give them due importance, commodities are given due weights. Weights are of two kinds (a) Implicit weights, (b) Explicit weights. In the first kind of weights are not explicitly assigned to any commodity but the commodity to which greater importance is attached is repeated a number of times. A number of varieties of such commodities are included in the index number as separate items. Thus, if an index number wheat is to receive a weight of 3 and rice a weight of 2, three varieties of wheat and two varieties rice would be included in this method weights are not apparent, but items are implicitly weighted. Such weights are known as implicit weights. In the second kind weights are explicitly assigned to commodi-

NOTES

ties. Only one variety of the commodity included in the construction of index number but its price relative is multiplied by the figure of weight assigned to it. Explicit weights are decided on some logical basis. For example, if wheat and rice are to be weighted in accordance with the value of their net output and if the ratio of their net output is 5:2, wheat would receive a weight of five and rice of two. Such weights are called explicit weights. Sometimes the quantities which are consumed are used as weights. These are called quantity weights. The amount spent on different commodities can also be use as their weights. These are called the value weights.

Unweighted Index Numbers

There are two methods of constructing unweighted index numbers.

- Simple Aggregative Method
- Simple Average of Relative Method

Simple Aggregative Method

In this method, the total of the prices of commodities in a given (current) years is divided by the total of the prices of commodities in a base year and expressed as percentage.

$$P_{on} = \frac{\sum P_n}{\sum P_o} \times 100$$

Simple Average of Relatives Method

In this method, we compute price relative or link relatives of the given commodities and then use one of the averages such as arithmetic mean, geometric mean, median etc. If we use arithmetic mean as average, then

$$P_{on} = \frac{1}{n} \sum \left(\frac{P_n}{P_o} \right) \times 100$$

The simple average of relative method is very simple and easy to apply is superior to simple aggregative method. This method has only disadvantage that it gives equal weight to all items.

Example:

The following are the prices of four different commodities for 1990 and 1991. Compute a price index by (1) Simple aggregative method and (2) Average of price relative method by using both arithmetic mean and geometric mean, taking 1990 as base.

Commodity	Cotton	Wheat	Rice	Gram
Price in 1990	909	288	767	659
Price in 1991	874	305	910	573

NOTES

Solution:

The necessary calculations are given below:

Commodity	Price in 1990 P_o	Price in 1991 P_n	Price Relative $P = \frac{P_n}{P_o} \times 100$	log P
Cotton	909	874	$\frac{874}{909} \times 100 = 69.15$	1.9829
Wheat	288	305	$\frac{305}{288} \times 100 = 105.90$	2.0249
Rice	767	910	$\frac{910}{767} \times 100 = 118.64$	2.0742
Gram	659	573	$\frac{573}{659} \times 100 = 86.95$	1.9393
Total	ΣP_o = 2623	ΣP_n = 2662	$\Sigma P = 407.64$	$\Sigma \log P$ = 8.0213

(1) Simple Aggregative Method:

$$P_{on} = \frac{\Sigma P_n}{\Sigma P_o} \times 100 = \frac{2662}{2623} \times 100 = 101.49$$

(2) Average of Price Relative Method (using arithmetic mean):

$$P_{on} = \frac{1}{n}\Sigma\left(\frac{P_n}{P_o}\right) \times 100 = \frac{1}{4}(407.64) = 101.91$$

Average of Price Relative Method (using geometric mean)

$$P_{on} = antilog\left(\frac{\Sigma \log P}{n}\right) = antilog\left(\frac{8.0213}{4}\right) = 101.23$$

WEIGHTED INDEX NUMBERS

When all commodities are not of equal importance, we assign weight to each commodity relative to its importance and index number computed from these weights is called weighted index numbers.

Laspeyre's Index Number:

In this index number the base year quantities are used as weights, so it also called base year weighted index.

NOTES

$$P_{on} = \frac{\sum P_n q_o}{\sum P_o q_o} \times 100$$

Paasche's Index Number:

In this index number, the current (given) year quantities are used as weights, so it is also called current year weighted index.

$$P_{on} = \frac{\sum P_n q_n}{\sum P_o q_n} \times 100$$

Fisher's Ideal Index Number:

Geometric mean of Laspeyre's and Paasche's index numbers is known as Fisher's ideal index number. It is called ideal because it satisfies the time reversal and factor reversal test.

$$P_{on} = \sqrt{\text{Laspeyre's Index} \times \text{Paashe's Index}}$$

$$P_{on} = \sqrt{\frac{\sum P_n q_o}{\sum P_o q_o} \times \frac{\sum P_n q_n}{\sum P_o q_n}} \times 100$$

Marshal-Edgeworth Index Number:

In this index number, the average of the base year and current year quantities are used as weights. This index number is proposed by two English economists Marshal and Edgeworth.

$$P_{on} = \left(\frac{\sum P_n q_o + \sum P_n q_n}{\sum P_o q_o + \sum P_o q_n} \right) \times 100$$

$$P_{on} = \frac{\sum P_n (q_o + q_n)}{\sum P_o (q_o + q_n)} \times 100$$

Example:

Compute the weighted aggregative price index numbers for 1981 with 1980 as base year using (1) Laspeyre's Index Number (2) Paashe's Index Number (3) Fisher's Ideal Index Number (4) Marshal Edgeworth Index Number.

Commodity	Prices		Quantities	
	1980	1981	1980	1981
A	10	12	20	22
B	8	8	16	18
C	5	6	10	11
D	4	4	7	8

NOTES

Solution:

Commodity	Prices		Quantity		P_1q_0	P_0q_0	P_1q_1	P_0q_1
	1980	1981	1980	1981				
	P_0	P_1	q_0	q_1				
A	10	12	20	22	240	200	264	220
B	8	8	16	18	128	128	144	144
C	5	6	10	11	60	50	66	55
D	4	4	7	8	28	28	32	32
					ΣP_1q_0 = 456	ΣP_0q_0 = 406	ΣP_1q_1 = 506	ΣP_0q_1 = 451

Laspeyre's Index Number:

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100 = \frac{456}{406} \times 100 = 112.32$$

Paashe's Index Number:

$$P_{01} = \frac{\Sigma P_1 q_1}{\Sigma P_0 q_1} \times 100 = \frac{506}{451} \times 100 = 112.20$$

Fisher's Ideal Index Number:

$$P_{01} = \sqrt{\text{Laspeyre's Index} \times \text{Paashe's Index}}$$

$$P_{01} = \sqrt{112.32 \times 112.20} = 112.26$$

Marshal Edgeworth Index Number:

$$P_{01} = \left(\frac{\Sigma P_1 q_0 + \Sigma P_1 q_1}{\Sigma P_0 q_0 + \Sigma P_0 q_1} \right) \times 100$$

$$P_{01} = \left(\frac{456 + 506}{406 + 451} \right) \times 100 = \frac{962}{856} \times 100 = 112.38$$

Check your progress:

1. Define Index Numbers?
2. What is fixed base method?

WHOLESALE PRICE INDEX NUMBERS

The wholesale price index numbers indicate the general condition of the national economy. They measure the change in prices of products produced by different sectors of an economy. The wholesale prices of major items manufactured or produced are included in the construction of these index numbers.

NOTES

The federal bureau of statistics has been constructing and releasing wholesale price index (WPI) in USA since 1961 – 1962. The list of wholesale items consists of four major groups.

- Food
- Fuel lighting and lubricants
- Manufactures
- Raw material

CONSUMER PRICE INDEX NUMBER

Consumer price index number is measured the changes in the prices paid by the consumers for purchasing a special "basket" of goods and services during the current year as compared to the base year. The basket of goods and services will contain items like (1) Food (2) House Rent (3) Clothing (4) Fuel and Lighting (5) Education (6) Miscellaneous like washing, transport, newspaper etc. Consumer price index number is also called cost of living index numbers or retail price index number.

Construction of Consumer Price Index Numbers:

The following steps are involved in the construction of consumer price index numbers.

1) **Class of People:** The first step in the construction of consumer price index (CPI) is that the class of people should be defined clearly. It should be decided whether the cost of living index number is being prepared for the industrial workers, middle or lower class salaried people living in a particular area. It is therefore necessary to specify the class of people and locality where they reside

2) **Family Budget Inquiry:** The next step in the construction of consumer price index number is that some families should be selected randomly these families provided information about food, clothing, house rent, miscellaneous etc. The inquiry include questions on family size, income, the quality and quantity consumed and the money spent on them and the weights are assigned in proportions to the expenditure on different items.

3) **Price Data:** The next step is to collect the data on retail prices of the selected commodities for the current period and the base period these prices should be obtained from the shops situated in that locality for which the index numbers are prepared.

4) **Selection of Commodities:** The next step is the selection of the commodities to be included. We should select those commodities which are mostly used by that class of people.

Methods of Consumer Price Index Numbers

There are two methods for the compute of consumer price index numbers. (a) Aggregate expenditure method (2) Family Budget Method

Aggregate Expenditure Method:

In this method, the quantities of commodities consumed by the particular group in the base year are estimated and these figures or their proportions are used as weights. Then the total expenditure on each commodity for each year is calculated. The price of the current year is multiplied by the quantity or weight of the base year. These products are added. Similarly for the base year total expenditure on each commodity is calculated by multiplying the quantity consumed by its price in the base year. These products are also added. The total expenditure of the current year is divided by the total expenditure of the base year and the resulting figure is multiplied by 100 to get the required index numbers. In this method, the current period quantities are not used as weights because these quantities change from year to year.

$$P_{on} = \frac{\sum P_n q_o}{\sum P_o q_o} \times 100$$

Where,

P_n Represent the price of the current year,

P_0 Represents the price of the base year and

q_0 Represents the quantities consumed in the base year.

Family Budget Method:

In this method, the family budgets of a large number of people are carefully studied and the aggregate expenditure of the average family on various items is estimated. These values are used as weights. Current year's price are converted into price relatives on the basis of base year's prices and these prices relatives are multiplied by the respective values of the commodities, in the base year. The total of these products is divided by the sum of the weights and the resulting figure is the required index numbers.

$$P_{on} = \frac{\sum IW}{\sum W}$$

Where,

$$I = \frac{P_n}{P_0} \times 100 \text{ and } W = P_o q_o$$

Example:

Construct the consumer price index number for 1988 on the basis of 1987 from the following data using: (1) Aggregate expenditure method (2) Family budget method.

NOTES

Commodity	Quantity Consumed in 1987	Unit	Prices	
			1987	1988
A	6 quintal	quintal	315.75	316.00
B	6 quintal	quintal	305.00	208.00
C	1 quintal	quintal	416.00	419.00
D	6 quintal	quintal	528.00	610.00
E	4 kg	kg	12.00	11.50
F	1 quintal	quintal	1020.00	1015.00

Solution:

(1) Consumer price index number of 1988 by Aggregate expenditure method:

Commodity	Quantity Consumed 1987 q_0	Unit	Prices		P_1q_0	P_0q_0
			1987 P_0	1988 P_1		
A	6 quintal	quintal	315.75	316.00	1896	1894.5
B	6 quintal	quintal	305.00	208.00	1848	1830.0
C	1 quintal	quintal	416.00	419.00	419	416.0
D	6 quintal	quintal	528.00	610.00	3660	3168.0
E	4 kg	kg	12.00	11.50	46	48.0
F	1 quintal	quintal	1020.00	1015.00	1015	1020.0
					ΣP_1q_0 =8884	ΣP_0q_0 =8376.5

Consumer price index number of 1988 is

$$P_{01} = \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100 = \frac{8884}{8376.5} \times 100 = 106.06$$

(2) Consumer price index number of 1988 by Family Budget Method:

C	1 quintal	416.00	419.00	416.0	100.72	41899.52
D	6 quintal	528.00	610.00	3168.0	115.53	365999.04
E	4 kg	12.00	11.50	48.0	95.83	4599.84
F	1 quintal	1020.00	1015.00	1020.0	99.51	101500.20
				ΣW = 8376.5		ΣWI = 888393.56

Consumer price index number of 1988 is

$$P_{01} = \frac{\Sigma WI}{\Sigma W} \times 100 = \frac{888393.56}{8376.5} \times 100 = 106.06$$

ANALYSIS OF TIME SERIES

Introduction of Time Series:

The statistical data is recorded with its time of occurrence is called a time series. The yearly output of wheat recorded for the last twenty five years, the weekly average price of eggs recorded for the last 52 weeks, the monthly average sales of a firm recorded for the last 48 months or the quarterly average profits recorded for the last 40 quarter etc., are example of time series data. It may be observed that this data undergoes changes with the passage of time. A number of factors can be isolated with contribute to the changes occurring overtime in such series.

In the field of economics and business, for example, income, imports exports, production, consumption, prices these data are depend on time. And all of these data are pretentious by seasonal changes as well as regular cyclical changes over the time period. To evaluate the changes in business and economics, the analysis of time series plays an important role in this regard. It is necessary to associated time with time series because time is one basic variable in time series analysis.

COMPONENTS OF TIME SERIES

The factors that are responsible to bring about changes in a time series, also called the components of time series, are as follows:

1. Secular Trend (or General Trend)
2. Seasonal Movements
3. Cyclical Movements
4. Irregular Fluctuations

Secular Trend: The secular trend is the main component of atime series which results from long term effect of socio-economics and political factors. This trend may show the growth or decline in a time series over a long period. This is the type of tendency which continues to persist for a very long period. Prices, export and imports data, for example, reflect obviously increasing tendencies over time.

NOTES

Seasonal Trend: These are short term movements occurring in a data due to seasonal factors. The short term is generally considered as a period in which changes occur in a time series with variations in weather or festivities. For example, it is commonly observed that the consumption of ice-cream during summer us generally high and hence sales of an ice-cream dealer would be higher in some months of the year while relatively lower during winter months. Employment, output, export etc. are subjected to change due to variation in weather. Similarly sales of garments, umbrella, greeting cards and fire-work are subjected to large variation during festivals like Valentine's Day, Eid, Christmas, New Year etc. These types of variation in a time series are isolated only when the series is provided biannually, quarterly or monthly.

Cyclic Movements: These are long term oscillation occurring in a time series. These oscillations are mostly observed in economics data and the periods of such oscillations are generally extended from five to twelve years or more. These oscillations are associated to the well known business cycles. These cyclic movements can be studied provided a long series of measurements, free from irregular fluctuations is available.

Irregular Fluctuations: These are sudden changes occurring in a time series which are unlikely to be repeated, it is that component of a time series which cannot be explained by trend, seasonal or cyclic movements. It is because of this fact these variations some-times called residual or random component. These variations though accidental in nature, can cause a continual change in the trend, seasonal and cyclical oscillations during the forthcoming period. Floods, fires, earthquakes, revolutions, epidemics and strikes etc, are the root cause of such irregularities.

Analysis of Time Series:

The object of the time series analysis is to identify the magnitude and direction of trend, to estimate the effect of seasonal and cyclical variations and to estimate the size of the residual component. This implies the decomposition of a time series into its several components. Two lines of approach are usually adopted in analyzing a given time series, namely,

(i) The additive model .

(ii) The multiplicative model

Thus, if we denote the time series by Y, the secular trend by T, the seasonal or short term periodic movements by S, the long term cyclical movements by C and the irregular or residual component by R, then the additive model can be described as

$$Y = T + S + C + R$$

While, the multiplicative model can be describe as

$$Y = T \times S \times C \times R$$

The additive model is generally used when the time series is spread over a short time span or where the rate of growth or decline in the trend is small. The multiplicative model, which is more in use than the additive model, is generally

NOTES

used whenever the time span of the series is large or the rate of growth or decline is would be

$$Y - T = S + C + R$$

or

$$\frac{Y}{T} = S \times C \times R$$

Similarly, a de-trended, de-seasonalized series may be obtained as

$$Y - T - S = C + R$$

or

$$\frac{Y}{T \times S} = C \times R$$

It is not always necessary that the time series may include all four types of variations, rather one or more of these components might be missing altogether. For example, using annual data the seasonal component may be ignored, while in a time series of short span, having monthly or quarterly observations, the cyclical component may be ignored.

Analyzing of Secular Trend:

A number of different methods are available to estimate the trend; however, suitability of these methods largely depends on the nature of the data and the purpose of the analysis. To measure a trend which can be represented as a straight line or some type of smooth curve, the following are the commonly employed methods.

(1) Freehand smooth curves,

(2) Semi-average method,

(3) Moving average method, and

(4) Mathematical curve fitting

Generally speaking, when the time series is available for a short span of time, in which seasonal variation might be important, the freehand and semi-average methods are employed. If the available series is spread over a long time span, having annual data, where long term cyclic might be important, the moving average method and the mathematical curve fitting are generally employed.

Methods of Free Hand Curve:

It is familiar concept, briefly described for drawing frequency curves. In case of a time series a scatter diagram of the given observations is plotted against time on the horizontal axis and a free hand smooth curve is drawn through the plotted points. The curve is so drawn that most of the points concentrate around the curve, however, smoothness should not be scarified in trying to let the points

NOTES

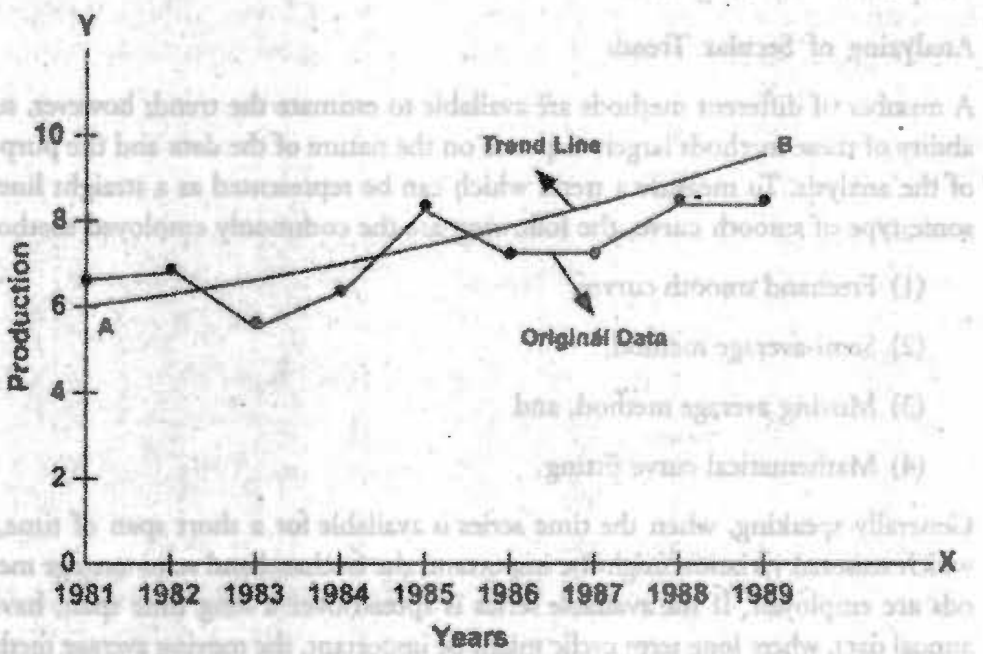
exactly fall on the curve. It would be better to draw a straight line through the plotted points instead of a curve, if possible. The curve fitted by this method eliminates the short term and long term oscillations and the irregular movements from the time series and elevate the general trend. After having drawn such a curve or line the trend values or the estimated Y values, which may be denoted by Y_t can be read from the graph corresponding to each time period.

Once of the major disadvantage of this method is that different individual would draw curves or lines which would differ in slope and intercept and hence no two conclusions would be identical. The trend values so obtained will differ from individual to individual. However, it is the most simple and quickest method of isolating the trend. This method is generally employed in such situation where the scatter diagram of the original data conforms to some well define trend.

Example: Measure the trend by method of freehand curve from the given data of production of wheat in a particular area of the world.

Years	1981	1982	1983	1984	1985	1986	1987	1988	1989
Production Million Metric Tons	6.6	6.9	5.6	6.3	8.4	7.2	7.2	8.5	8.5

Solution:



We observe that the graph of the original data does not show any closeness to any type of curve. It looks like increasing very slowly in straight (linear) manner. Thus we draw a line AB as an approximation to the original graph. The line AB represents the trend line and from this we read the trend values for the given years.

Merits and Demerits of Free hand Curve:

Merits : This method is very simple and easy to understand. It is applicable for linear and non-linear trends. It gives us a idea about the rise and fall of the time

series. For every long time series, the graph of the original data enables us to decide about the application of more mathematical models for the measurement of trend. A monthly data of 5 years has 60 values. A graph of these values may suggest that the trend is linear for the first two years (24 values) and for the next 3 years, it is non- linear. We accordingly apply the linear approach on the first 24 values and the curvilinear techniques on the next 36 values.

Demerits : It is not mathematical in nature. Different persons may draw a different trend. The method does not appeal to a common man because it seems as if it is something rough and crude.

Methods of Semi Averages:

This method is also simple and relatively objective than the free hand method. The data is divided in two equal halves and the arithmetic mean of the two sets of values of Y is plotted against the center of the relative time span. If the numbers of observations are even the division into halves will be straight forward, however, if the number of observations are odd, then the middle most i.e., $\left(\frac{n+1}{2}\right)^{th}$, item is dropped. The two points so obtained are joined through a straight line which shows the trend. The trend values of Y i.e., $\hat{Y}$ can then be read from the graph corresponding to each time period.

The arithmetic mean, since greatly affected by extreme values, is subjected to misleading values hence the trend obtained by plotting by means might be distorted. However, if extreme values are not apparent, this method may be fruitfully employed. To understand the estimation of trend, using the above noted two methods, consider the following worked example.

Example : Measure the trend by the method of semi-average by using the following table given below. Also write the equation of the trend line with origin at 1984-85.

Years	Value in Million
1984 – 85	18.6
1985 – 86	22.6
1986 – 87	38.1
1987 – 88	40.9
1988 – 89	41.4
1989 – 90	40.1
1990 – 91	46.6
1991 – 92	60.7
1992 – 93	57.2
1993 – 94	53.4

NOTES

Solution:

Years	Values	Semi-Totals	Semi-Average	Trend Values
1984 – 85	18.6			28.664 – 3.656 = 25.008
1985 – 86	22.6			32.32 – 3.656 = 28.664
1986 – 87	38.1	161.6	32.32	32.32
1987 – 88	40.9			32.32 + 3.656 = 35.976
1988 – 89	41.4			35.976 + 3.656 = 39.632
1989 – 90	40.1			39.632 + 3.656 = 43.288
1990 – 91	46.6			43.288 + 3.656 = 46.944
1991 – 92	60.7	253.0	5.60	50.60
1992 – 93	57.2			50.60 + 3.656 = 54.256
1993 – 94	53.4			54.256 + 3.656 = 57.912

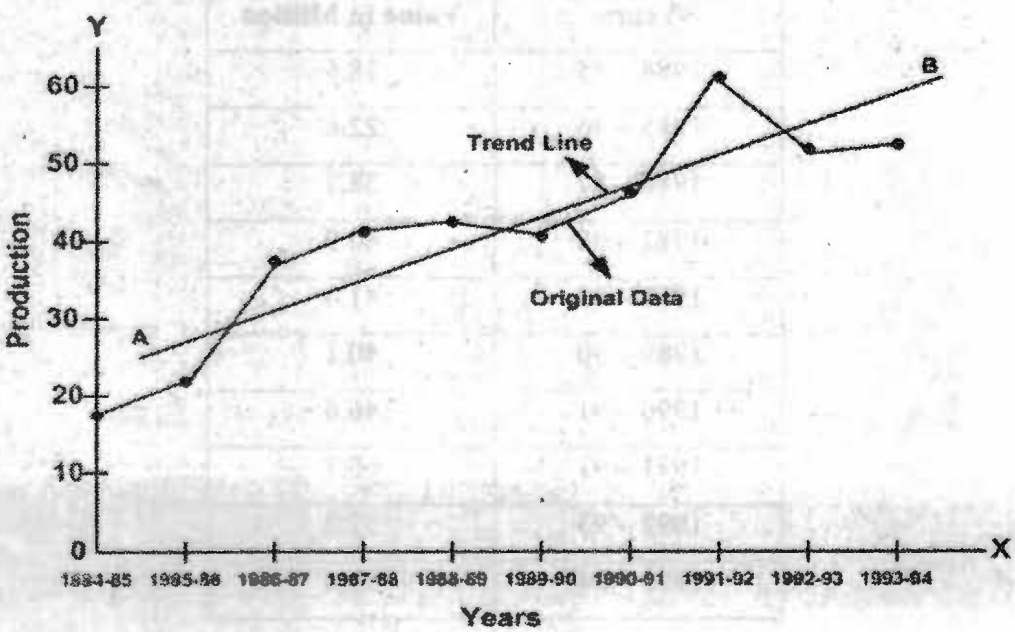
Trend for 1991 – 92 = 50.60

Trend for 1986 – 87 = 32.32

Increase in trend in 5 years = 18.28

Increase in trend in 1 year = 3.656

Trend for one year is 3.656. It is called slope of the trend line and is denoted by *b*. Thus, *b*= 3.656. The trend for 1987 – 88 is calculated by adding 3.656 to 32.32 and similar calculations are done for the subsequent years. Trend for 1985 – 86 is less than the trend for 1986 – 87. Thus trend for 1985 – 86 is 32.32 – 3.656 = 28.664. Trend for the year 1984 – 85 = 25.008. This is called the intercept because 1984 – 85 is the origin. Intercept is the value of *Y* when $X = 0$. Intercept is denoted by *a*. The equation of trend line is $\hat{Y} = a + bX = 25.008 + 3.656X$ (1984 – 85 = 0) where $\hat{Y}$ shows the trend values. This equation can be used to calculate the trend values of the time series. It can also be used for forecasting the future values of the variable.



NOTES

Merits and Demerits of Semi Average Methods:

Merits : This method is very simple and easy to understandable and also it does not require much of calculations.

Demerits : The method is used only when the trend in linear or almost linear. For non-linear trend this method is not applicable. It is used on the calculation of average and the average is affected by extreme values. Thus if there is some very large value or very small value in the time series, that extreme value should either be omitted or this method should not be applied. We can also write the equation of the trend line.

Methods of Moving Averages:

Suppose that there are n times periods denoted by $t_1, t_2, t_3, \ldots\ldots, t_n$, and the corresponding values of Y variable are $Y_1, Y_2, Y_3, \ldots\ldots Y_n$. First of all we have to decide the period of the moving averages. For short time series, we use period of 3 or 4 values. For long time series, the period may be 7, 10 or more. For quarterly time series, we always calculate averages taking 4-quarters at a time. In monthly time series, 12-monthly moving averages are calculated. Suppose the given time series is in years and we have decided to calculate 3-years moving average. The moving averages denoted by $a_1, a_2, a_3, \ldots\ldots a_{n-2}$ are calculated as below:

Years (t)	Variable (Y)	3-Years moving totals	3-Years moving averages
t_1	Y_1	——	——
t_2	Y_2	$Y_1 + Y_2 + Y_3$	$\frac{Y_1 + Y_2 + Y_3}{3} = a_1$
t_3	Y_3	$Y_2 + Y_3 + Y_4$	$\frac{Y_2 + Y_3 + Y_4}{3} = a_2$
t_4	Y_4	⋮	⋮
⋮	⋮	⋮	⋮
t_{n-2}	Y_{n-2}	⋮	⋮
t_{n-1}	Y_{n-1}	$Y_{n-2} + Y_{n-1} + Y_n$	$\frac{Y_{n-2} + Y_{n-1} + Y_n}{3} = a_{n-2}$
t_n	Y_n	——	——

The average of the first 3 values is $\frac{Y_1 + Y_2 + Y_3}{3}$ and is denoted by a_1. It is written against the middle year t_2. We leave the first value Y_1 and calculate the average for the next three values. The average is $\frac{Y_2 + Y_3 + Y_4}{3} = a_2$ and is written against the

Check your progress:

3. What is time series analysis?
4. Name different components of time series.

NOTES

middle years t_3. The process is carried out to calculate the remaining moving averages. 4-years moving averages are calculated as under:

Years (t)	Variable (Y)	3-Years moving averages	3-Years moving averages centered
t_1	Y_1	-----	----
t_2	Y_2	$\frac{Y_1+Y_2+Y_3+Y_4}{4}=a_1$	----
t_3	Y_3	$\frac{Y_2+Y_3+Y_4+Y_5}{4}=a_2$	$\frac{a_1+a_2}{2}=A_1$
t_4	Y_4	$\frac{Y_3+Y_4+Y_5+Y_6}{4}=a_3$	$\frac{a_2+a_3}{2}=A_2$
t_5	Y_5		

The first average is a_1 which is calculated as $\frac{Y_1+Y_2+Y_3+Y_4}{4}=a_1$. It is written against the middle of t_3 and t_4. The two averages a_1 and a_2 are further averaged to get an average $\frac{a_1+a_2}{2}=A_1$, which refers to the center of t_3 and is written against t_3. This is called centering of the 4-years moving averages. The process is continued till the end of the series to get 4-yearsmoving average centered. The moving averages of some proper period smooth out the short term fluctuations and the trend is measured by the moving averages.

Example: Compute 5-years, 7-years and 9-years moving averages for the following data.

Years	1990	1991	1992	1993	1994	1995	1996	19997	1998	1999	2000
Values	2	4	6	8	10	12	14	16	18	20	22

Solution:

The necessary calculations are given below:

		5-Years Moving		7-Years Moving		9-Years Moving	
Years	Values	Total	Average	Total	Average	Total	Average
1990	2	---	---	---	---	---	---
1991	4	---	---	---	---	---	---
1992	6	30	6	---	---	---	---
1993	8	40	8	56	8	---	---
1994	10	50	10	70	10	90	10

NOTES

1995	12	60	12	84	12	108	12
1996	14	70	14	98	14	126	14
1997	16	80	16	112	16	—	
1998	18	90	18	—	—	—	—
1999	20	—	—	—	—	—	—
2000	22	—	—	—	—	—	—

Example: Compute 4-years moving average centered for the following time series:

Years	1995	1996	1997	1998	1999	2000	2001	2002
Production	80	90	92	83	87	96	100	110

Solution:

The necessary calculations are given below:

Year	Production	4-Years Moving Total	4-Years Moving Average	2-values Moving Total	4-years Moving Average Centered
1995	80	—	—	—	—
1996	90	345	86.25	—	—
1997	92	352	88.00	174.25	87.125
1998	83	358	89.50	177.50	88.750
1999	87	366	91.50	181.00	90.500
2000	96	393	98.25	189.75	94.875
2001	100	—	—	—	—
2002	110	—	—	—	—

Merits and Demerits of Moving Average Methods:

Merits: Moving averages can be used for measuring the trend of any series. The method is applicable for linear as well as non-linear trends.

Demerits: The trend obtained by moving averages is, in general, neither a straight line nor some standard curve. For this reason the trend cannot be extended for forecasting the future values. Trend values are not available for some periods in the start and some values at the end of the time series. The method is not applicable for short time series.

SUMMARY

- Index numbers are meant to study the change in the effects of such factors which cannot be measured directly.
- Index numbers are commonly used statistical device for measuring the

NOTES

combined fluctuations in a group related variables. Index numbers are statistical measures designed to show changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc. A collection of index numbers for different years, locations, etc., is sometimes called an index series.

- A simple index number is a number that measures a relative change in a single variable with respect to a base.
- A composite index number is a number that measures an average relative changes in a group of relative variables with respect to a base.
- Consumer price index number is measured the changes in the prices paid by the consumers for purchasing a special "basket" of goods and services during the current year as compared to the base year.
- The statistical data is recorded with its time of occurrence is called a time series. The yearly output of wheat recorded for the last twenty five years, the weekly average price of eggs recorded for the last 52 weeks, the monthly average sales of a firm recorded for the last 48 months or the quarterly average profits recorded for the last 40 quarter etc., are example of time series data. It may be observed that this data undergoes changes with the passage of time. A number of factors can be isolated with contribute to the changes occurring overtime in such series.

ANSWER TO CHECK YOUR PROGRESS

1. Index numbers are statistical measures designed to show changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc.
2. In fixed base method, a particular year is generally chosen arbitrarily and the prices of the subsequent years are expressed as relatives of the prices of the base year.
3. The statistical data is recorded with its time of occurrence is called a time series. The yearly output of wheat recorded for the last twenty five years, the weekly average price of eggs recorded for the last 52 weeks, the monthly average sales of a firm recorded for the last 48 months or the quarterly average profits recorded for the last 40 quarter etc., are example of time series data.
4. The factors that are responsible to bring about changes in a time series, also called the components of time series, are as follows:
   1. Secular Trend (or General Trend)
   2. Seasonal Movements
   3. Cyclical Movements
   4. Irregular Fluctuations

TEST YOURSELF

1) What are the Index Numbers?
2) What are the steps involved for the construction of price index numbers?
3) Explain methods of constructing unweighted index numbers.
4) What do you mean by Weighted Index Numbers?
5) Write a short note on:

i) Wholesale Price Index Numbers

ii) Consumer price index number

6) Explain different components of Time Series.

6 PROBABILITY

NOTES

Chapter Includes :

- INTRODUCTION:
- RANDOM EXPERIMENT:
- NOT EQUALLY LIKELY OUTCOMES:
- REDUCED SAMPLE SPACE
- RANDOM VARIABLE AND PROBABILITY DISTRIBUTION
- ADDITIVE LAW OF PROBABILITY:
- ADDITIVE LAW OF PROBABILITY FOR MUTUALLY EXCLUSIVE EVENTS:
- MULTIPLICATION LAW OF PROBABILITY FOR INDEPENDENT EVENTS:
- CONDITIONAL PROBABILITY:
- THEOREMS OF MULTIPLICATION LAW OF PROBABILITY AND CONDITIONAL PROBABILITY:
- BAYE'S THEOREM:
- BINOMIAL DISTRIBUTION
- NORMAL DISTRIBUTION
- POISSON DISTRIBUTION
- SAMPLING:
- MEANING OF SAMPLING:
- SAMPLING ERRORS:
- SAMPLING DISTRIBUTION:

Learning Objective :

After going through this chapter, you should be able to:

- Understand concept of Probability.
- Learn random experiment, sample space.
- Understand addition and multiplication law of probability.
- Explain Baye's Theorem.
- Discuss Binomial, Normal and Poison Distribution.
- Understand sampling, sampling design and frame

NOTES

Introduction

We live in the world of uncertainties. A man is surrounded by situations which are not fully under his control. The nature commands these situations. A person on a road does not know whether or not he will reach his destination safely. A patient in the hospital is never sure about his survival after a delicate operation. What will be the weather conditions tomorrow, nothing is known with certainty but we always like to have an idea about the weather conditions in future. A flight will be in time, the road will be clear or there will be some traffic jam. We face this of problem in our daily life. Man is always curious to know as to what will happen in future. The things which happen are important for the man today. These things are based on what is called chance or probability. If we have some numerical measure of uncertainty, this measure is called probability. We may find a numerical measure for a bulb to be defective, some numerical measure for the rain to fall. The belief or confidence associated with a certain situation can also be measured. It is also called probability. In statistics there are various situations where uncertainty in involved.

Such situations need the application of probability. Probability is widely and rightly used in statistical decisions. The areas of statistics where probability is used are called the areas of statistical inference. Statistical inference is not possible without the use of probability. Probability is also used in different fields of life where uncertainty is involved. Knowledge of probability is used in space research, astronomy, business, weather studies, economics, genetics and various other fields of life. It is simple to explain various concepts of probability with the help of set theory. Thus we shall use here the set theory notation.

Probability theory can be understood as a mathematical model for the intuitive notion of uncertainty. Without probability theory all the stochastic models in Physics, Biology, and Economics would either not have been developed or would not be rigorous. Also, probability is used in many branches of pure mathematics; even in branches one does not expect this, like in convex geometry.

Random Experiment

The word experiment or random experiment is used for a situation of uncertainty about which we want to have some observations. The actual results of the uncertain situation is called outcome or sample point. In the random experiment, nothing can say with certain about the outcome. An experiment may consist of one or more observations. If there is only a single observation, the term random trail or simply trial is used. A bulb may be selected from a factory to examine if it is defective or not. A single bulb selected is a trial. We can select any number of bulbs. The number of observations will be equal to the number of bulbs. A random experiment has the following properties:

1. The experiment can be repeated any number of times. We may select one or more items for inspection. The number of repetitions is called the size of the experiment. In statistics, the size of the random experiment plays a major role in statistical inference.

NOTES

2. A random trial consists of at least two possible outcomes. If a basket contains all the defective bulbs, a selected bulb will be certainly defective. It has only one possible outcome. It is not a random trial. If the basket contains some good and some defective bulbs, a selected bulb will be good or defective. In this case there are two possible outcomes. Thus selecting a bulb from such a basket is a random trial.

3. Nothing can be said with certainly about the outcome of the random trial or random experiment. If a sample of four bulbs is selected, may be one bulb is defective. When another sample of four bulbs from the same lot is selected, may be all bulbs are defective. Thus the result of the experiment cannot be predicted even if the experiment is repeated a large number of times.

Sample Space:

A complete list of all possible outcomes of a random experiment is called sample space or possibility space and is denoted by *S*. Each outcome is called element of the sample space. A sample space may be containing any number of outcomes.
If it contains finite number of outcomes, it is called finite or discrete sample space. When two bulbs are selected from a lot, the possible outcomes are four which can be counted as:

1. both bulbs are defective
2. first is defective and second is good
3. first is good and second is defective
4. both are good

Here the sample space is discrete. When the possibilities of the sample space cannot be contained, it is called continuous. The number of possible readings of temperature from 45°C to 46 °C will make a continuous sample space.

Sample space is the basic term in the theory of probability. We shall discuss some sample spaces in this tutorial. It is not always possible to make the sample space. If it contains very large number of points, we cannot register all the outcomes but we must understand as to how we can make the sample space. The outcomes of the sample space are written within the {}. Some simple sample spaces are discussed below:

A coin is tossed:

When a coin is tossed, it has two possible outcomes. One is called head and the other is called tail. Anyone of the two faces may be called head. To be brief, head is denoted by *H* and tail is denoted by *T*. Thus the sample space consists of head and tail. In set theory notation, we can write *S* as:

S = {head, tail} or $S = \{H, T\}$

NOTES

Two coins tossed:

When two coins are tossed, there are four possible outcomes. Let H_1 and T_1 denote the head and tail on the first coin and H_2 and T_2 denote the head and tail on the second coin respectively. The sample space S can be written in the form as

$$S = \{(H_1, H_2), (H_1, T_2), (T_1, H_2), (T_1, T_2)\}$$

It may be noted that a sample space of throw of two coins has 4 possible points. A sample space of 3 coins will have $2^3 = 8$ possible points and for n coins, the number of possible points will be 2^n.

A die is thrown:

An ordinary die which is used in games of chances has six faces. These six faces contain 1,2,3,4,5,6 dots on them. Thus for a single throw of a die, the sample space has 6 possible outcomes which are:

$$S = \{1,2,3,4,5,6\}$$

Two dice thrown:

A die has six faces. Each face of the first die can occur with all the six faces of the second die. Thus there are $6 \times 6 = 36$ possible pairs or points when two dice are tossed together.

These 36 pairs are written below as:

$$S = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

If 3 dice are thrown, the sample space will have $6^3 = 216$ possible points, each point being a triplet of 3 digits.

Event of Probability:

Any part of the sample space is called an event of a probability. An event may contain one or more than one outcomes. When an event consists of a single outcome (sample points), it is called a simple event. An event which has two or more outcomes is called a compound event. The sample points contained in an event are written within brackets { }. If we consider a single face when a die is thrown, it is a simple event. Getting 6 on a die when thrown is called the occurrences of a simple event. If the event is any prime number on the die, the event consists of the points 2, 3, 5. It is a compound event and consists of three simple events which are {2}, {3} and {5}. When two dice are thrown, the pair (1, 1) is a single outcome in the sample space S and is therefore a simple event. The event "total is 3" consist of two outcomes that is (1, 2) and (2, 1). Thus "total is 3" is a compound event.

If a random experiment can produced $n$sample points, it has $n$simple events. Throw of a single die has 6 simple events and a throw of two dice produced 36simpleevents.Theemptyset ϕisalsoaneventbutitisnotasimpleevent.The samplespace Sisacompoundeventiscalledacertainevent.

Equally Likely Outcomes:

The outcomes of a sample space are called equally likely if all of them have the same chance of occurrence. It is very difficult to decide whether or not the outcomes are equally likely. But in this tutorial we shall assume in most of the experiments that the outcomes are equally likely. We shall apply the assumption of equally likely in the following cases:

(1) Throw of a coin or coins:

When a coin is tossed, it has two possible outcomes called head and tail. We shall always assume that head and tail are equally likely if not otherwise mentioned. For more than one coin, it will be assumed that on all the coins, head and tail are equally likely.

(2) Throw of a die or dice:

Throw of a single die can be produced six possible outcomes. All the six outcomes are assumed equally likely. For any number of dice, the six faces are assumed equally likely.

(3) Playing Cards:

There are 52 cards in a deck of ordinary playing cards. All the cards are of the same size and are therefore assumed equally likely.

4) Balls from a Bag:

There are many situation in probability in which some objects are selected from a certain container. The objects of the container are assumed to be equally likely. A famous example is the selection of a few balls from a bag containing balls of different colors. The balls of the bag are assumed equally likely.

Not Equally Likely Outcomes

When all the outcomes of a sample space do not have equal chance of occurrence, the outcomes are called not equally likely. When a matchbox is thrown, all the six faces are not equally likely. If a bag contains balls of different sizes and a ball is selected at random, all the balls are not equally likely.

Mutually Exclusive Events:

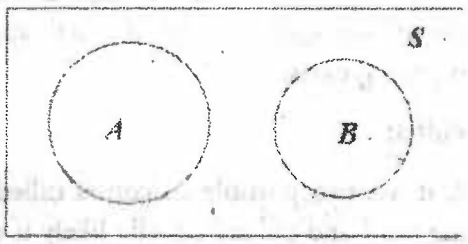
Two events are called mutually exclusive or disjoint if they do not have any outcome common between them. If the two events $A$and $B$are mutuallyexclusive, then $A \cap B = \phi$(nullset). Forthreemutuallyexclusiveevents A,$B$and C,we have $A \cap B \cap C = \phi$. Suppose there is a sample space Sas:

$S = \{1,2,3,4,5,6,7,8,9,10\}$

Let $A = \{3,6,9\}$ and $B = \{5,10\}$

NOTES

Here $A \cap B = \phi$. Thus A and A are mutually exclusive events. Both A and B belong to the same sample space but they are completely different and both cannot happen at the same time. A class of students may contain first grade, second grade and third grade. When a student is selected from the class, he will be any one of the three groups of students. Thus three groups of students are disjoint or mutually exclusive. When the two events A and B are mutually exclusive, we can show them with the help of a Venn diagram. The Venn diagram as shown in the figure that $A \cap B = \phi$.



$A \cap B = \phi$

Not Mutually Exclusive Events

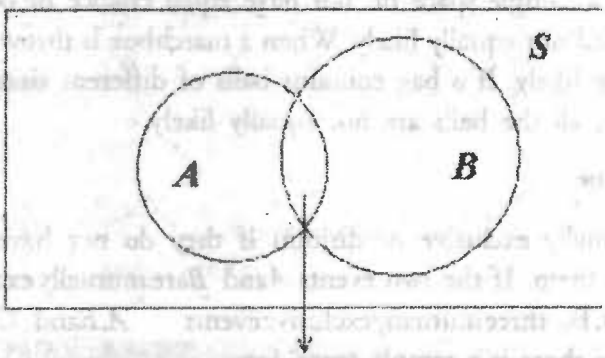
Two events are called not mutually exclusive if they have al least one outcome common between them. If the two events A and B are not mutually exclusive events, then $A \cap B = \phi$. Similarly, A, B and C are not mutually exclusive events if $A \cap B \cap C = \phi$. Thus they must have at least one common point between them. Consider a sample space:

$S = \{1,2,3,4,5,6,7,8,9,10,11\}$

Let $A = \{2,3,5,7,11\}$ and $B = \{1,3,5,7,9,11\}$

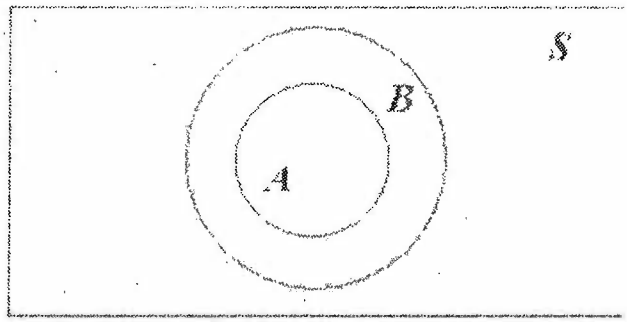
Here $A \cap B = \{3,5,7,11\}$

Thus, $A \cap B \neq \phi$ i.e. $A \cap B$ exist. Here A and B are not mutually exclusive events. $A \cap B$ consist of outcomes which are common to both A and B. As shown in the figure a Vann diagram in which A and B are not mutually exclusive events. Some area under A is common with B. If the event A is a part of the event B, then $A \cap B = A$. This is shown in the figure:



$A \cap B$

NOTES



$A \cap B = A$\

Exhaustive and Complementary Events:

Exhaustive Events:

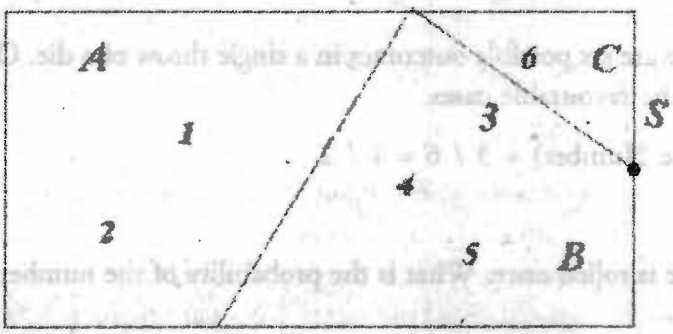
When a sample space *S* is partitioned into some mutually exclusive events such that their union is the sample space itself then the events are called exhaustive events or collectively events.

Suppose a die is tossed and the sample space is

$S = \{1,2,3,4,5,6\}$

Let $A = \{1,2\}$ $B = \{3,4,5\}$ $C = \{6\}$

Hence the events *A*,*B* and *C* are mutually exclusive because $A \cap B \cap C = \phi$ and $A \cup B \cup C = S$. As shown in the figure three events *A*,*B* and *C* which are exhaustive.



$A \cap B \cap C = \phi$ and $A \cup B \cup C = S$

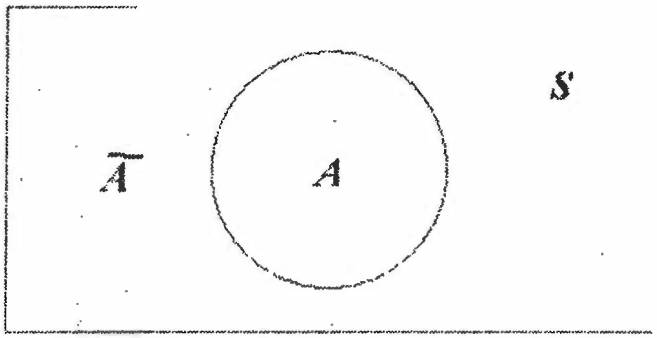
Complementary Events:

If *A* is an event defined in the sample space *S*, then $S - A$ is denoted by $\overline{A}$ and is called complement of *A*.

Thus, $\overline{A} = S - A$ or $A \cap \overline{A} = S$

In the figure shown that the event *A* and the complement of *A*.

NOTES



Example: A die is rolled once. Find the probability of getting a 5.

There are six possible ways in which a die can fall, out of these only one is favourable to the event.

$$P(5) = 1 / 6$$

Example: A coin is tossed once. What is the probability of the coin coming up with head?

Solution: The coin can come up either 'head' (H) or a tail (T). Thus, the total possible outcomes are two and one is favourable to the event.

So,

$$P(H) = 1 / 2$$

Example: A die is rolled once. What is the probability of getting a prime number?

Solution: There are six possible outcomes in a single throw of a die. Out of these; 2, 3 and 5 are the favourable cases.

$$P \text{ (Prime Number)} = 3 / 6 = 1 / 2$$

Example: A die is rolled once. What is the probability of the number '7' coming up?

What is the probability of a number 'less than 7' coming up?

Solution: There are six possible outcomes in a single throw of a die and there is no face of the die with mark 7.

$$P \text{ (number 7)} = 0 / 6 = 0$$

[**Note:** That the probability of impossible event is zero]

As every face of a die is marked with a number less than 7,

$$P (< \text{ or } = 7) = 6 / 6 = 1$$

[**Note:** That the probability of an event that is certain to happen is 1]

Example: In a simultaneous toss of two coins, find the probability of

(i) getting 2 heads (ii) exactly 1 head

NOTES

Solution: Here, the possible outcomes are

HH, HT, TH, TT.

i.e., Total number of possible outcomes = 4.

(i) Number of outcomes favourable to the event (2 heads) = 1 (i.e., HH).

P (2 heads) = 1 / 4

(ii) Now the event consisting of exactly one head has two favourable cases, namely HT and TH.

P (exactly one head) = 2 / 4 = 1 / 2

Example: In a single throw of two dice, what is the probability that the sum is 9?

Solution: The number of possible outcomes is 6 × 6 = 36. We write them as given below:

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

Now, how do we get a total of 9. We have:

3 + 6 = 9

4 + 5 = 9

5 + 4 = 9

6 + 3 = 9

In other words, the outcomes (3, 6), (4, 5), (5, 4) and (6, 3) are favourable to the said event, i.e., the number of favourable outcomes is 4.

Hence, P (a total of 9) = 4 / 36 = 1 / 9

Example: From a bag containing 10 red, 4 blue and 6 black balls, a ball is drawn at random. What is the probability of drawing

(i) a red ball ? (ii) a blue ball ? (iii) not a black ball ?

Solution: There are 20 balls in all. So, the total number of possible outcomes is 20. (Random drawing of balls ensure equally likely outcomes)

(i) Number of red balls = 10

P (a red ball) = 10 / 20 = 1 / 2

NOTES

(ii) Number of blue balls = 4

P (a blue ball) = 4 / 20 = 1 / 5

(iii) Number of balls which are not black = 10 + 4 = 14

P (not a black ball) = 14 / 20 = 7 / 10

Example: A card is drawn at random from a well shuffled deck of 52 cards. If A is the event of getting a queen and B is the event of getting a card bearing a number greater than 4 but less than 10, find P(A) and P (B).

Solution: Well shuffled pack of cards ensures equally likely outcomes.

Hence, the total number of possible outcomes is 52.

(i) There are 4 queens in a pack of cards.

P(A) = 4 / 52 = 1 / 13

(ii) The cards bearing a number greater than 4 but less than 10 are 5,6, 7,8 and 9.

Each card bearing any of the above number is of 4 suits diamond, spade, club or heart. Thus, the number of favourable outcomes = 5 × 4 = 20

= 20 / 52 = 10 / 26 = 5 / 13

Example: What is the chance that a leap year, selected at random, will contain 53 Sundays?

Solution: A leap year consists of 366 days consisting of 52 weeks and 2 extra days. These two extra days can occur in the following possible ways.

(i) Sunday and Monday
(ii) Monday and Tuesday
(iii) Tuesday and Wednesday
(iv) Wednesday and Thursday
(v) Thursday and Friday
(vi) Friday and Saturday
(vii) Saturday and Sunday

Out of the above seven possibilities, two outcomes, e.g., (i) and (vii), are favourable to the event

P (53 Sundays) = 2 / 7

REDUCED SAMPLE SPACE

Sometimes the samplespace S is reduced in size and is called reduced sample space. The symbol S, may be used for a reduced sample space. Suppose a die has been thrown and we have been informed that the experiment has produced an even face. This type of information is called the additional information. Thus the reduced sample space is determined by the additional information.

NOTES

In this example the information has disclosed that even face has occurred. If it becomes known as to which even face has occurred, then it is no more s situation ofprobability. When theinformationis that the face is even, then there is still something hidden from the experimenter. The actual outcome is not known to the observer. In this case the reduced sample space S_r is $S_r = \{2,4,6\}$.

Relative Frequency:

The term relative frequency is used for the ratio of the observed frequency of some outcome and the total frequency of the random experiment. Suppose arandomexperimentisrepeated Ntimesandsomeoutcomesisobserved $f$times, then the ratio $\frac{f}{N}$ is called therelativefrequencyof the outcome which has beenobserved f times. Some examples of relative frequencies are given here:

- We select bulbs from a certain big lot to examine whether they are good or defective. We take, say 100 such bulbs and examine them. Sixty bulbs are found defective. The symbol Nmay beused for 100and the symbol $f$may beused for theobservedfrequencywhich is60. Thus the

$$\text{Relative frequency} = \frac{f}{N} = \frac{60}{100} = 0.6$$

- We are interested to know whether a coin is unbiased (true) or not. We toss the coin say 200 times and note that the number of heads. In 200 tosses, the number of heads may be, say 110. The relative frequency of thisexperimentfor number of heads is $\frac{110}{200}$ which is not $\frac{1}{2}$. As we shall see later, the probability of head is usually written as $\frac{1}{2}$. It is just as assumption and of course a big assumption. If we repeat the sameexperimentagain, thenumberofheadsmaybelessthanormorethan 110 as observed in the first experiment. This is what happens inrandomexperiments.

- A die is thrown and we are interested in the ace (face 1). We throw the die say 600 times and ace is observed 12 times. Thus the relative frequency of aces is $\frac{12}{60} = \frac{1}{5}$. For an ideal die one should expect that the number of aces would be $\frac{60}{6} = 10$. At some later stage we would like to know more about the ratio $\frac{12}{60}$. This ratio is not something constant. A next random experimentwith the same die may produce a completely different result.

NOTES

Definition of Probability:

Probability is something strange and it has been defined in different manners. We can define probability in objective or subjective manner. Let us first use objective approach to define probability.

The Classical Definition of Probability:

This definition is for equally likely outcomes. If an experiment can produced N mutually exclusive and equally likely outcomes out of which n outcomes are favorable to the occurrence of event A, then the probability of A is denoted by P(A) and is defined as the ratio $\frac{n}{N}$. Thus the probability of A is given by y

$$P(A) = \frac{\text{Number of otcomes favorable to } A}{\text{Number of possible outcomes}} = \frac{n}{N}$$

This definition can be applied in a situation in which all possible outcomes and the outcomes in the events A can be counted. This definition is due to P.S. Laplace (1749 – 1827). The classical definition is also called the priori definition of probability. The word priori is from prior and is used because the definition of base on the previous knowledge that the outcomes are equally likely. When a coin is tossed, the probability of head is assumed to be $\frac{1}{2}$. This probability of $\frac{1}{2}$ is based on this classical definition of probability. It is assumed that the two faces of the coin are equally likely. In practical life the people do believe that a coin will do justice when it is tossed. In the playgrounds, the participating teams toss the coin to start the match. A coin in which probability of head is assumed to be equal to the probability of tail is called a true or uniform or an unbiased coin. But it is an all assumption. The probability of a certain event is a number which lies between 0 and 1. If the event does not contain any outcome, it is called impossible event and its probability is zero. If the event is as big as the sample space, the probability of the event is one because

$$P(\text{the event}) = \frac{\text{Number of otcomes in the event}}{\text{total number of outcomes}} = \frac{N}{N} = 1$$

When probability of an event is one, it is called a "Sure" or "Certain", event.

Criticism:

The classical definition of probability has always been criticized for the following reasons:

1. This definition assumes that the outcomes are equally likely. The term equally likely is almost as difficult as the word probability itself. Thus the definition uses the circular reasoning.
2. The definition is not applicable when the numbers of outcomes are not equally likely.

NOTES

3. The definition is also not applicable when the total number of outcomes is infinite or it is difficult to count the total outcomes or the outcomes favorable to the event. It is difficult to count the fish in the ocean. Thus it is difficult to find the probability of catching a fish of some weight say more than one kilogram.

Example:

One day 20 files were presented to an income tax officer for disposal. Five files contained bogus entries. All the files were thoroughly mixed and there was no indication about bogus files. What is the probability that one file with bogus entries is selected.

Solution:

Here all possible outcomes = 20

Let A be the event that the file has bogus entries.

Thus, number of favourable outcomes = 5

Here we shall apply the classical definition of probability. All the 20 files are assumed to be equally likely for the purpose of selecting a file.

Probability of selecting a file with bogus entries is written as P(A)

$$P(A) = \frac{\text{Number of otcomes favorable to } A}{\text{Number of possible outcomes in } S} = \frac{n(A)}{n(S)}$$

$$= \frac{5}{10} = \frac{1}{4}$$

Example:

A fair die is thrown. Find the probabilities that the face on the die is (1)Maximum (2) Prime (3)Multiple of 3 (4)Multiple of 7 .

Solution:

There are 6 possible outcomes when a die is tossed. We assumed that all the 6 faces are equally likely. The classical definition of probability is to be applied here.

The sample space is S = {1,2,3,4,5,6}, $n(S) = 6$

(1) Let A be the event that the face is maximum. Thus,

A = {6}, $n(A) = 1$

Therefore, $P(A) = \frac{n(A)}{n(S)} = \frac{1}{6}$

(2) Let B be the event that the face is maximum.

Thus, B = {2,3,5}, $n(B) = 3$

Therefore, $P(B) = \frac{n(B)}{n(S)} = \frac{3}{6} = \frac{1}{2}$

NOTES

(3) Let C be the event that the face is maximum. Thus,

$C = \{3,6\}, \quad n(C) = 2$

Therefore, $P(C) = \frac{n(C)}{n(S)} = \frac{2}{6} = \frac{1}{3}$

(4) Let D be the event that the face is maximum. Thus,

$D = \phi, \quad n(D) = 0$

Therefore, $P(D) = \frac{n(D)}{n(S)} = \frac{0}{6} = 0$ (not possible)

Subjective Probability:

A person may some confidence or belief regarding the occurrence of some event, Say *A*. The numerical measure of this confidence is called the subjective probability of the occurrence of *A*. This probability is based on the experience, intelligence and knowledge of the person who determining the probability in some situation. For example, we may be interested to know whether a certain political system will succeed in a country or not. The probability of success in this situation cannot be determined by objective definitions of probability.

The assessment of this probability is made by some expert. This approach can be applied in real world situations. This probability is subjective in nature. Different persons may have different probabilities for the same situation at the same time.

RANDOM VARIABLE AND PROBABILITY DISTRIBUTION

Generation of Random Numbers:

Introduction:

The word random is used quite commonly in our daily life. One may or may not know its meaning but whenever it is used, it conveys the sense for which it is used. An apple may be picked up from a shop at random. The customers enter a shop not according to some preplan; they enter the shop in a random manner. The vehicles cross a zebra crossing in a random manner. The teacher does not check the note books of all students, he checks some of the note books selected at random. Somebody is intelligent by birth, somebody is healthy by birth, somebody has a slip on the road, somebody meets an accident on the road, and somebody gets an attack of influenza. There is something random about all this. When a coin or a die is tossed so that it can fall any face freely, it is a random fall. Many situations in practical life are of random nature. Their ultimate results are based on chance. A small boy is familiar with the classical idea of lotterymethod , which is centuries old and has been used for the selection of a random sample. Nobody has so far discovered a better method of selecting a sample from the population. Most modern methods of selecting a sample are based on the theory of random selection by lottery. The random sample is the basis of the statistical inference. Thus randomness is the central idea of thestudywhich is carried out to know something about unknown situations.

Generation of Random Numbers:

In our counting system, there are ten basic digits which are used for counting purposes. These digits are 0, 1, 2 ... 9. We can make integers of any size with the help of these digits. The figure 53792 is made up of five digits 2, 3, 5, 7 and 9. We shall use these digits to make a set of numbers called table of random numbers. Suppose we select ten paper slips and on each slip we write a different digit. Thus each slip represents a digit. We select any one of these slips at random and note down its digit on a paper. We return the slip to the main lot and select aslip again. The digit on the second slip is also noted along with the first digit (row – wise) or below the first digit (column – wise). We continue this process of selecting, recording and replacing each selected slip. On each selection the probability of selection of each digit is 1/10. Thus each digit has equal probability of selection. We get a set of digits called random digits. If the first digit is 5, second is 7, third is 5 and fourth is 0, we can write them in a row as 5750 or 57 50. We can also write in a column as below:

5
7
5
0

When the first row or column is completed, we can write the selected digits in the second row or second column. In this manner a table of any size spread over a number of pages can be obtained. This is called table of random numbers. One small table of random numbers is given below:

51	22	09	12	72	12	40	92
72	45	35	50	23	39	74	44
57	18	73	31	11	75	88	75
92	69	46	75	56	82	77	66
38	32	12	93	95	68	84	87
95	71	80	36	82	16	48	38

This table is written with two digits in two columns together. This is one way of writing the digits. One can write 3 – digit or 4 – digit columns. Anybody can make a table of random numbers. A good table of random numbers contains 0, 1, 2 ... 9 almost equal numbers of times. The students shall learn in higher classes that therandom numberscan be made for each probability distribution. Therandom numbersunderdiscussion, infectaretherandomnumbersfromadiscreteuniform distribution over the interval (0, 9).

Application of Random Numbers:

There are many uses of random numbers in statistics. One of the important uses is in the selection of a simple random sample from a finite population. Suppose there are 80 students in a class and we want to select 8 students at random. The students can be numbered from 01 to 80. Then we consult the random numbertable. Let us see on table. We shall read two digit columns. If any

NOTES

number is between 01 and 80, we shall note it down for the sample. If any random number is above 80, we shall ignore it because it is not in our population. We can read the random number table from any place. We may move column – wise or row – wise. Let us read the random numbers from Table. We read the first column. The random numbers are *51,* 72, 57, 38, 22, 45, 18, 69. Two random numbers 92 and 95 have been ignored because they are above 80. The remaining eight random numbers represent those eight students who have been selected for the sample. If the number of units in the population is in hundreds, say 800 we shall read three – digit column of the random number table. If we have 100 units in the population, we shall number them from 00 to 99 so that we have to read 2 – digit columns. If we number them as 001, 002 ... 100, we shall have to read 3 – digit columns. In three digit column, most of the random numbers will be above 100 and a lot of time will be wasted in getting the required size of the sample. Further use of random number table for the selection of a simple random sample will be discussed in sampling.

Random number table can also be used to generate data without performing the actual experiment. Suppose we want to toss a coin 10 times to see the number of heads. We can do it by

(i) tossing the coin and counting the number of heads.

(ii) By using random number table. The even digits 0, 2, 4, 6, 8 will stand for the head and the odd digits 1, 3, 5, 7, 9 will be for the tail. The first ten digits of the first row of pervious table are reproduced below. **H** is for head and **T** is for tail.

Digits	5	1	2	2	0	9	1	2	7	2
Outcomes	T	T	H	H	H	T	T	H	T	H

There are 5 heads and 5 tails. It is just a chance that number of heads is equal to the number of tails. If we read the next row, the result in general would be different.

This process of getting the results of an experiment from random number table is called simulation. In simulation the actual experiment is not performed.

Examples of Random Numbers:

Example:

Two balanced coins are to be tossed **10** times to record the number of heads each times. Use random number table to record the possible observations. Write the frequency distribution of the observed number of heads.

Solution:

Two digits of a random number table will represent the result of a throw of two coins. We shall take ten pairs of random numbers for **10** throws of two coins.

From given table. We take **10** pairs of random digits and count the number of heads. Even digit will indicate head (**H**) and an odd digit will indicate tail (**T**).

NOTES

Random pairs:	51	22	09	12	72	12	40	92	72	95
Number of even digits:	0	2	1	1	1	1	2	1	1	0
No. of heads:	0	2	1	1	1	1	2	1	1	0

The observed frequency distribution of number of heads is

Number of Heads	Frequency
0	2
1	6
2	2
Total	10

If two coins are actually tossed 10 times, the observed frequency distribution may or may not agree with the above distribution.

Note: If 3 coins are to be tossed, we shall take 3 digits to represents a throw of 3 coins.

Example:

Assume that a fair die is to be rolled ten times. Without rolling the die, obtain the possible outcomes using random digits.

Solution:

Here the six digits 1, 2, 3, 4, 5, 6 will represent the six faces of the die. We shall ignore the random digit 0 and the digits above 6.

From the above table we have the random digits

5 1 2 2 1 2 2 1 2 4

Thus we assume that the first throw has given 5 on the die, second throw has given 1 on the die and the 10th throw has given 4 on the die.

Random Variable:

A set of numerical values assigned to the all possible outcomes of the random experiment are called random variable. The random variable can be briefly written as r.v. If we write A, B,..., F on the six faces of a die, these letters are not a r.v.
If we write some numerical values on the six faces of a die like 1, 2, 3,..., 6, we have a set of values called r.v. Suppose we select two bulbs from a certain lot having good and defective bulbs. Let G stand for good and D stand for defective.
There are four possible outcomes which are GG, GD, DG and DD. Each outcome can be assigned some numerical value. Let us count number of defective bulbs in each outcome. We can write

NOTES

Outcome	No. of Defective Bulbs
GG	0
GD	1
DG	1
DD	2

Thus the numerical values **0, 1, 2** are the values of the random variable where random variable is the number of defective bulbs in this discussion. A random variable is denoted by a capital letter X. Here X is the number of defective bulbs. The small letters $x_1, x_2, \ldots, x_n$ are used for the specific values of the random variable. A random variable is also called chance variable. If we have two or more than two random variables we can use the letters X, Y, Z for them. A random variable may be discrete or continuous.

Discrete Random Variable:

A random variable X is called discrete if it can assume finite number of values. If two bulbs are selected from a certain lot, the number of defective bulbs may be **0, 1** or **2**. The range of the variable is from **0** to **2** and random variable can take some selected values in this range. The number of defective bulbs cannot be **1.1** or 1 or 3 etc. This random variable can take only the specific values which are **0, 1** and **2**. When two dice are rolled the total on the two dice will be **2, 3, ..., 12**. The total on the two dice is a discrete random variable.

Discrete Probability Distribution:

Suppose a discrete random variable X can assume the values $x_1, x_2, x_3, \ldots, x_n$ with corresponding probabilities $p(x_1), p(x_2), p(x_3), \ldots, p(x_n)$. The set of ordered pairs $[x_1, p(x_1)], [x_2, p(x_2)], [x_3, p(x_3)], \ldots, [x_n, p(x_n)]$ is called the probability distribution or probability function of random variable X. A probability distribution can be presented in a tabular form showing the values of the random variable X and the corresponding probabilities donated by $p(x_i)$ or $f(x_i)$. The above information can be collected in the form of a table below called the probability distribution of the random variable X. The probability that random variable X will take the value x_i is denoted by $p(x_i)$ where $p(x_i) = P(X = x_i)$.

Values of random variable (xi)	x_1	x_2	x_3	...	x_i	....	x_n
Probability $p(x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	...	$p(x_i)$	...	$p(x_n)$

The probability of some interval can be calculated by adding the probabilities of all points in the interval. For example $P[x_1 < X < x_4] = P(X = x_2) + P(X = x_3)$.

This type of addition will not be possible in continuous random variable for finding the probability of an interval. Therein we shall use the integral calculus for finding the probability of an interval.

The probability distribution can also be described in the form of an equation for $f(x_i)$ with a list of possible values of the random variable X. Some probability distributions in the form of equations are

(i) $f(x_i) = \frac{1}{6}$ for xi = 1,2,3,....,6

For each value of X the probability is 1/6. It is the probability distribution when a fair die is rolled.

(ii) $f(x_i) = \left(\frac{3}{x}\right)\left(\frac{1}{2}\right)^3$ for x = 0,1,2,3

Example:

A digit is selected from the first 8 natural numbers. Write the probability distribution of X where X is the number of factors (divisors) of digits.

Solution:

It is assumed here that the probability of selection for each digit is $\frac{1}{8}$. We can write:

Digit	Factors	Number of factors	Probability
1	1	1	1/8
2	1, 2	2	1/8
3	1, 3	2	1/8
4	1, 2, 4	3	1/8
5	1, 5	2	1/8
6	1, 2, 3, 6	4	1/8
7	1, 7	2	1/8
8	1, 2, 4, 8	4	1/8

The information in the above table can be summarized as below in the form of table of probability distribution.

Number of factors r.v. (x_i)	1	2	3	4	Total
$p(x_i)$	1/8	4/8	1/8	2/8	1

The probability of X = 2 is 4/8 which has been obtained by adding 1/8 four times. Similarly the probability of 4 is 2/8 which has been obtained by adding 1/8 two times.

NOTES

Example:

A factory is producing bulbs out of which **25%** are defective. Two bulbs are selected from this factory for inspection. Write the probability distribution of number of defective bulbs.

Solution:

Let X denotes the number of defective bulbs. Let D denote the defective and G denote the good bulbs.

Here $P(D)=\frac{25}{100}=\frac{1}{4}$ and $P(G)=\frac{75}{100}=\frac{3}{4}$

It is assumed here that the two bulbs are selected independently because there is very large number of bulbs in the factory. According to multiplication law of independent events, we have

P(both defective) = $P(D_1D_2)=P(D_1).P(D_2)=\frac{1}{4}\times\frac{1}{4}=\frac{1}{16}$

P(first defective and second good) = $P(D_1G_2)=P(D_1).P(G_2)=\frac{1}{4}\times\frac{3}{4}=\frac{3}{16}$

P(first good and second defective) = $P(G_1D_2)=P(G_1).P(D_2)=\frac{3}{4}\times\frac{1}{4}=\frac{3}{16}$

P(both good) = $P(G_1G_2)=P(G_1).P(G_2)=\frac{3}{4}\times\frac{3}{4}=\frac{9}{16}$

The possible outcomes, the random variable X and the corresponding probabilities are put in the following tables:

Outcomes (S)	Number of defectives (r.v.)	Probability
D_1D_2	2	$\frac{1}{4}\times\frac{1}{4}=\frac{1}{16}$
D_1D_2	1	$\frac{1}{4}\times\frac{3}{4}=\frac{3}{16}$
D_1D_2	1	$\frac{3}{4}\times\frac{1}{4}=\frac{3}{16}$
D_1D_2	0	$\frac{3}{4}\times\frac{3}{4}=\frac{9}{16}$

The above information can be collected as below in the form of a probability distribution of the random variable X.

Random Variable (x_i)	0	1	2
$p(x_i)$	9/16	6/16	1/16

NOTES

Continuous Random Variable:

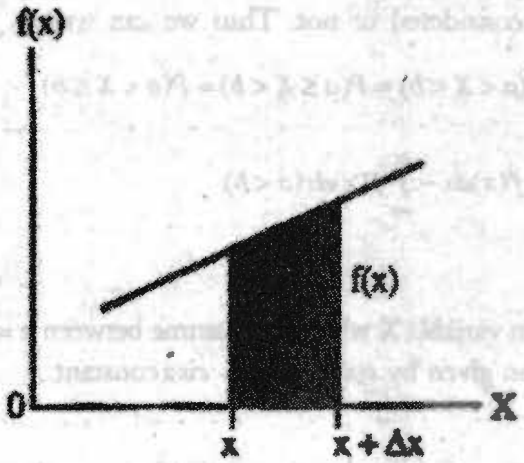
A random variable is called continuous if it can assume all possible values in the possible range of the random variable. Suppose the temperature in a certain city in the month of June in the past many years has always been between 35°C to 45°C centigrade. The temperature can take any value between the ranges 35°C to 45°C. The temperature on any day may be 40.15°C or 40.16°C or it may take any value between 40.15°C and 40.16°C. When we say that the temperature is 40°C, it means that the temperature lies between somewhere between 39°C to 40°C. Any observation which is taken falls in the interval. There is nothing like an exact observation in the continuous variable. In discrete random variable the values of the variable are exact like 0, 1, 2 good bulbs. In continuous random variable the value of the variable is never an exact point. It is always in the form of an interval, the interval may be very small.

Some examples of the continuous random variables are:

1. The computer time (in seconds) required to process a certain program.
2. The time that a poultry bird will gain the weight of 1.5 kg.
3. The amount of rain falls in the certain city.
4. The amount of water passing through a pipe connected with a high level reservoir.
5. The heat gained by a ceiling fan when it has worked for one hour.

Probability Density Function:

The probability function of the continuous random variable is called probability density function of briefly p.d.f. It is denoted by $f(x)$ where $f(x)$ is the probability that the random variable X takes the value between x and $x + \Delta x$ where Δx is a very small change in X.



If there are two points 'a' and 'b' then the probability that the random variable will take the value between a and b a given by the management.

$$P(a \le X \le b) = \int_a^b f(x)dx$$

NOTES

Where 'a' and 'b' are the points between $-\infty$ and $+=$. The quantity $f(x)dx$ is called probability differential. The number of possible outcomes of a continuous random variable is uncountable infinite. Therefore, a probability of zero is assigned to each point of the random variable. Thus $P(X = x) = 0$ for all values of X. This means that we must calculate a probability for a continuous random variable over an internal and not for any particular point. This probability can be interpreted as an area under the graph between the interval from a to b. When we say that the probability is zero that a continuous random variable assumes a specific value, we do not necessarily mean that a particular value cannot occur. We in fact, mean that the point (event) is one of an infinite number of possible outcomes. Whenever we have to find the probability of some interval of the continuous random variable, we can use any one of these two methods.

1. Integral calculus.
2. Area by geometrical diagrams (this method is easy to apply when $f(x)$ is a simple linear function).

Properties of Probability Density Function:

The probability density function $f(x)$ must have the following properties.

1. It is non-negative i.e. $f(x) \geq 0$ for all x.
2. Total Area $= \int_{-\infty}^{\infty} f(x)dx = 1$
3. $(X = c) = \int_{c}^{c} f(x)dx = 0$ Where c is any constant
4. As probability of area for X = c (constant), therefore P(X = a) = P(X = b). If we take an interval a to b. It makes no difference whether end points of the interval are considered or not. Thus we can write:

   $$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$

5. $P(a \leq X \leq b) = \int_{b}^{a} f(x)dx - \int_{-\infty}^{\infty} f(x)dx (a < b)$

Example:

A continuous random variable X which can assume between x = 2 and 8 inclusive, has a density function given by $c(x+3)$ where c is a constant.

(a) Calculate c

(b) P(3 < X < 5)

(c) P(X $\geq$ 4)

Solution:

$$f(x) = c(x+3), 2 \le x \le 8$$

(a) $f(x)$ will be density functions if

(i) $f(x) \ge 0$ for every x and

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$.

If $c \ge 0$, $f(x)$ is clearly ≥ 0 for every x in the given interval. Hence for $f(x)$ to be density function, we have

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_{2}^{8} c(x+3)dx = c\left[\frac{x^2}{2} + 3x\right]_2^8$$

$$= c\left[\frac{(8)^2}{2} + 3(8) - \frac{(2)^2}{2} - 3(2)\right] = c[32 + 24 - 2 - 6] = c[48]$$

So that $c = \frac{1}{48}$

Therefore, $f(x) = \frac{1}{48}(x+3),\ 2 \le x \le 8$

(b) $P(3 < X < 5) = \int_{3}^{5} \frac{1}{48}(x+3)dx = \frac{1}{48}\left[\frac{x^2}{2} + 3x\right]_3^5$

$$= \frac{1}{48}\left[\frac{(5)^2}{2} + 3(5) - \frac{(3)^2}{2} - 3(3)\right] = \frac{1}{48}\left[\frac{25}{2} + 15 - \frac{9}{2} - 9\right]$$

$$= \frac{1}{48}[14] = \frac{7}{24}$$

(c) $P(X \ge 4) = \int_{4}^{8} \frac{1}{48}(x+3)dx = \frac{1}{48}\left[\frac{x^2}{2} + 3x\right]_4^8$

$$= \frac{1}{48}\left[\frac{(8)^2}{2} + 3(8) - \frac{(4)^2}{2} - 3(4)\right] = \frac{1}{48}[32 + 24 - 8 - 12$$

$$= \frac{1}{48}[36] = \frac{3}{4}$$

ADDITIVE LAW OF PROBABILITY

Let A be the event of getting an odd number and B be the event of getting a prime number in a single throw of a die. What will be the probability that it is either an odd number or a prime number?

NOTES

NOTES

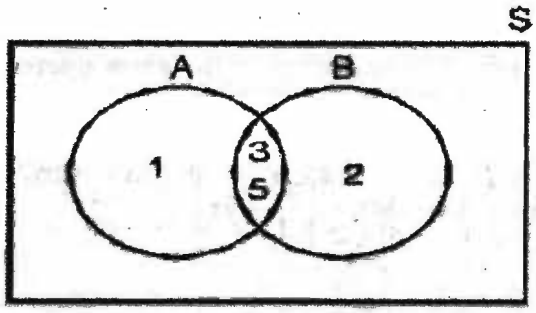
In a single throw of a die, the sample space would be

S = { 1, 2, 3, 4, 5, 6 }

The outcomes favourable to the events A and B are

A = { 1, 3, 5 }

B = { 2, 3, 5 }



The outcomes favourable to the event 'A or B' are

A U B = {1, 2, 3, 5 }.

Thus, the probability of getting either an odd number or a prime number will be

P(A or B) = 4 / 6 = 2 / 3

To discover an alternate method, we can proceed as follows:

The outcomes favourable to the event A are 1, 3 and 5.

P (A) = 3 / 6

Similarly,

P (B) = 3 / 6

The outcomes favourable to the event 'A and B' are 3 and 5. Hence,

P (A and B) = 2 / 6

Now, P(A) + P(B) - P(A and B) = 3 / 6 + 3 / 6 – 2 / 6

4 / 6 = 2 / 3 = P (A or B)

Thus, we state the following law, called additive rule, which provides a technique for finding the probability of the union of two events, when they are not disjoint.

For any two events A and B of a sample space S,

P (A or B) = P (A) + P (B) - P (A and B)

or P (AUB)=P (A)+ P (B)-P (A $\cap$ B) .....(ii)

Example: A card is drawn from a well-shuffled deck of 52 cards. What is the probability that it is either a spade or a king?

Solution: If a card is drawn at random from a well-shuffled deck of cards, the likelihood of any of the 52 cards being drawn is the same. Obviously, the sample space consists of 52 sample points.

NOTES

If A and B denote the events of drawing a 'spade card' and a 'king' respectively, then the event

A consists of 13 sample points, whereas the event B consists of 4 sample points. Therefore,

$$P(A) = 13 / 52$$

$$P(B) = 4 / 52$$

The compound event ($A \cap B$) consists of only one sample point, viz.; king of spade. So,

$$P(A \cap B) = 1 / 52$$

Hence, the probability that the card drawn is either a spade or a king is given by

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

$$= 13 / 52 + 4 / 52 - 1 / 52$$

$$= 16 / 52 = 4 / 13$$

Example: In an experiment with throwing 2 fair dice, consider the events

A: The sum of numbers on the faces is 8

B: Doubles are thrown.

What is the probability of getting A or B?

Solution: In a throw of two dice, the sample space consists of $6 \times 6 = 36$ sample points.

The favourable outcomes to the event A (the sum of the numbers on the faces is 8) are

$$A = \{ (2, 6), (3, 5), (4, 4), (5, 3), (6, 2) \}$$

The favourable outcomes to the event B (Double means both dice have the same number) are

$$B = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$$

$$A \cap B = \{ (4,4) \}.$$

Now $P(A) = 5 / 36$,

$P(B) = 6 / 36$,

$P(A \cap B) = 1 / 36$

Thus, the probability of A or B is

$$P(A \quad B) = 5 / 36 + 6 / 36 - 1 / 36$$

$$= 10 / 36 = 5 / 18$$

NOTES

ADDITIVE LAW OF PROBABILITY FOR MUTUALLY EXCLUSIVE EVENTS

We know that the events A and B are mutually exclusive, if and only if they have no outcomes in common. That is, for mutually exclusive events,

$$P(A \text{ and } B) = 0$$

Substituting this value in the additive law of probability, we get the following law:

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

Example: In a single throw of two dice, find the probability of a total of 9 or 11.

Solution: Clearly, the events - a total of 9 and a total of 11 are mutually exclusive.

Now P (a total of 9) = P [(3, 6), (4, 5), (5, 4), (6, 3)] = 4 / 36

P (a total of 11) = P [(5, 6), (6, 5)] = 2 / 36

Thus, P (a total of 9 or 11) = 4 / 36 + 2 / 36

= 1 / 36

Example: The probabilities that a student will receive an A, B, C or D grade are 0.30, 0.35, 0.20 and 0.15 respectively. What is the probability that a student will receive at least a B grade?

Solution: The event at least a 'B' grade means that the student gets either a B grade or an A grade.

P (at least B grade) = P (B grade) + P (A grade)

= 0.35 + 0.30

= 0.65

Example: Prove that the probability of the non-occurrence of an event A is 1 - P (A).

i.e., P (not A) = 1 - P (A) or, P (A) = 1 - P (A).

Solution: We know that the probability of the sample space S in any experiment is 1. Now, it is clear that if in an experiment an event A occurs, then the event (A) cannot occur simultaneously, i.e., the two events are mutually exclusive.

Also, the sample points of the two mutually exclusive events together constitute the sample space S. That is,

$$A \cup \bar{A} = S$$

Thus, $P(A \cup \bar{A}) = P(S)$

$P(A) + P(\bar{A}) = 1$ (A and $\bar{A}$ are mutually exclusive and S is sample space)

$P(\bar{A}) = 1 - P(A)$, which proves the result.

This is called the law of complementation.

Law of complementation:

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

NOTES

Example: Find the probability of the event getting at least 1 tail, if four coins are tossed once.

Solution: In tossing of 4 coins once, the sample space has 16 samples points.

P (at least one tail) = P (1or 2 or 3 or 4 tails)

= 1 - P (0 tail) (By law of complementation)

= 1 - P (H H H H)

The outcome favourable to the event four heads is 1.

Hence, P (H H H H) = 1 / 16

Substituting this value in the above equation, we get

P (at least one tail) = 1 – 1 / 16 = 15 / 16

In many instances, the probability of an event may be expressed as odds - either odds in favour of an event or odds against an event.

If A is an event:

The odds in favour of A = P (A) / $P(\bar{A})$ or P (A) to $P(\bar{A})$ where P (A) is the probability of the event A and $P(\bar{A})$ is the probability of the event 'not A'.

Similarly, the odds against A are

$P(\bar{A})$ / P (A) or P (A) to $P(\bar{A})$

Example: The probability of the event that it will rain is 0.3. Find the odds in favour of rain and odds against rain.

Solution: Let A be the event that it will rain.

P (A) = .3

By law of complementation,

$P(\bar{A})$ = 1 - .3 =.7 .

Now, the odds in favour of rain are 0.3 / 0.7 or 3 to 7 (or 3 : 7).

The odds against rain are 0.7 / 0. 3 or 7 to 3.

When either the odds in favour of A or the odds against A are given, we can obtain the probability of that event by using the following formulae

If the odds in favour of A are *a* to *b*, then

P (A) = a / a +b

NOTES

If the odds against A are a to b, the P (A) = b / a + b

This can be proved very easily.

Suppose the odds in favour of A are *a* to *b*. Then, by the definition of odds,

$$\frac{P(A)}{P(\overline{A})} = \frac{a}{b}$$

From the law of complementation,

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

Therefore,

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

or b P (A) = a - a P (A)

or (a + b) P (A) = a or

$$P(A) = \frac{a}{a+b}$$

Similarly, we can prove that

$$P(A) = \frac{b}{a+b}$$

when the odds against A are b to a.

Example: Determine the probability of A for the given odds

(a) 3 to 1 in favour of A

(b) 7 to 5 against A.

Solution: (a) P (A) = 3 / 3 + 1 = 3 / 4

(b) P (A) = 7 / 7 + 5 = 5 / 12

MULTIPLICATION LAW OF PROBABILITY FOR INDEPENDENT EVENTS

Two events A and B are said to be independent, if the occurrence or non-occurrence of one does not affect the probability of the occurrence (and hence non-occurrence) of the other.

Can you think of some examples of independent events?

The event of getting 'H' on first coin and the event of getting 'T' on the second coin in a simultaneous toss of two coins are independent events.

What about the event of getting 'H' on the first toss and event of getting 'T'

NOTES

on the second toss in two successive tosses of a coin? They are also independent events.

Let us consider the event of 'drawing an ace' and the event of 'drawing a king' in two successive draws of a card from a well-shuffled deck of cards without replacement.

Are these independent events?

No, these are not independent events, because we draw an ace in the first draw with probability 4 / 52. Now, we do not replace the card and draw a king from the remaining 51 cards and this affect the probability of getting a king in the second draw, i.e., the probability of getting a king in the second draw without replacement will be 4 / 51

Note: If the cards are drawn with replacement, then the two events become independent.

Is there any rule by which we can say that the events are independent?

How to find the probability of simultaneous occurrence of two independent events?

If A and B are independent events, then

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

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

Thus, the probability of simultaneous occurrence of two independent events is the product of their separate probabilities.

Note: The above law can be extended to more than two independent events, i.e.,

$$P(A \cap B \cap C...) = P(A) \times P(B) \times P(C)...$$

On the other hand, if the probability of the event 'A' and 'B' is equal to the product of the probabilities of the events A and B, then we say that the events A and B are independent.

Example: A die is tossed twice. Find the probability of a number greater than 4 on each throw.

Solution: Let us denote by A, the event 'a number greater than 4' on first throw. B be the event 'a number greater than 4' in the second throw. Clearly A and B are independent events.

In the first throw, there are two outcomes, namely, 5 and 6 favourable to the event A.

$$P(A) = 2 / 6 = 1 / 3$$

Similarly, $P(B) = 1 / 3$

Hence, $P(A \text{ and } B) = P(A) \cdot P(B)$

$$= 1 / 3 \cdot 1 / 3 = 1 / 9$$

NOTES

Example: Two balls are drawn at random with replacement from a box containing 15 red and 10 white balls. Calculate the probability that

(a) both balls are red.

(b) first ball is red and the second is white.

(c) one of them is white and the other is red.

Solution:

(a) Let A be the event that first drawn ball is red and B be the event that the second ball drawn is red. Then as the balls drawn are with replacement, therefore

$P(A) = 15 / 25 = 3 / 5, \quad P(B) = 3 / 5$

As A and B are independent events

Therefore, $P(\text{both red}) = P(A \text{ and } B)$

$= P(A) \times P(B)$

$= 3 / 5 \times 3 / 5 = 9 / 25$

(b) Let A : First ball drawn is red.

B : Second ball drawn is white.

$\therefore P(A \text{ and } B) = P(A) \times P(B)$

$= 3 / 5 \times 2 / 5 = 6 / 25.$

(c) If WR denotes the event of getting a white ball in the first draw and a red ball in the second draw and the event RW of getting a red ball in the first draw and a white ball in the second draw.

Then as 'RW' and WR' are mutually exclusive events, therefore

Probability

$\therefore P(\text{a white and a red ball})$

$= P(WR \text{ or } RW)$

$= P(WR) + P(RW)$

$= P(W)\,P(R) + P(R)\,P(W)$

$= 2 / 5 \,.\, 3 / 5 + 3 / 5 \,.\, 2 / 5$

$6 / 25 \,.\, 6 / 25 = 12 / 25$

CONDITIONAL PROBABILITY

Suppose that a fair die is thrown and the score noted. Let A be the event, the score is 'even'.

Then,

$A = \{2, 4, 6\}$

$P(A) = 3 / 6 = 1 / 2$

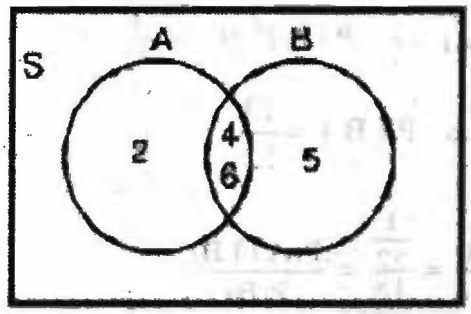
Now suppose we are told that the score is greater than 3. With this additional information what will be P (A) ?

NOTES

Let B be the event, 'the score is greater than 3'. Then B is {4, 5, 6}. When we say that B has occurred, the event 'the score is less than or equal to 3' is no longer possible. Hence the sample space has changed from 6 to 3 points only. Out of these three points 4, 5 and 6; 4 and 6 are even scores.

Thus, given that B has occurred, P (A) must be 2/ 3

Let us denote the probability of A given that B has already occurred by P (A | B) .



Again, consider the experiment of drawing a single card from a deck of 52 cards. We are interested in the event A consisting of the outcome that a black ace is drawn.

Since we may assume that there are 52 equally likely possible outcomes and there are two black aces in the deck, so we have

$$P(A) = 2 / 52$$

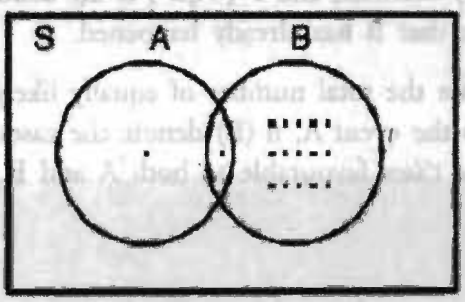
However, suppose a card is drawn and we are informed that it is a spade. How should this information be used to reappraise the likelihood of the event A?

Clearly, since the event B "A spade has been drawn" has occurred, the event "not spade" is no longer possible. Hence, the sample space has changed from 52 playing cards to 13 spade cards. The number of black aces that can be drawn has now been reduced to 1.

Therefore, we must compute the probability of event A relative to the new sample space B.

Let us analyze the situation more carefully.

The event A is " a black ace is drawn'. We have computed the probability of the event A knowing that B has occurred. This means that we are computing a probability relative to a new sample space B. That is, B is treated as the universal set. We should consider only that part of A which is included in B.



NOTES

Hence, we consider A $\cap$ B (see figure).

Thus, the probability of A. given B, is the ratio of the number of entries in A$\cap$ B to the number of entries in B. Since n (A$\cap$ B) = 1 and n (B) = 13 , then

$$P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{1}{13}$$

Notice that,

$$n(A \cap B) = 1 \Rightarrow P(A \cap B) = \frac{1}{52}$$

$$n(B) = 13 \Rightarrow P(B) = \frac{13}{52}$$

$$P(A|B) = \frac{1}{13} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{P(A \cap B)}{P(B)}$$

This leads to the definition of conditional probability as given below:

Let A an B be two events defined on a sample space S. Let P(B) > 0, then the conditional probability of A, provided B has already occurred, is denoted by P(A|B) and mathematically written as :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0$$

The symbol P(A | B) is usually read as "the probability of A given B".

THEOREMS OF MULTIPLICATION LAW OF PROBABILITY AND CONDITIONAL PROBABILITY:

Theorem 1: For two events A and B,

$$P(A \cap B) = P(A).P(B|A),$$

and $P(A \cap B) = P(B).P(A|B)$,

where P (B|A) represents the conditional probability of occurrence of B, when the event A has already occurred and P (A|B) is the conditional probability of happening of A, given that B has already happened.

Proof: Let n (S) denote the total number of equally likely cases, n (A) denote the cases favourable to the event A, n (B) denote the cases favourable to B and n (A $\cap$ B) denote the cases favourable to both A and B.

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

NOTES

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

$$P(A\cap B)=\frac{n(A\cap B)}{n(S)} \quad \text{.....(1)}$$

For the conditional event A|B , the favourable outcomes must be one of the sample points of B, i.e., for the event A|B, the sample space is B and out of the n (B) sample points, n (A ∩ B) pertain to the occurrence of the event A, Hence,

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

Rewriting (1), we get

$$P(A\cap B)=\frac{nB}{nS}=\frac{nAB}{nB}=P(B).P(A|B)$$

Similarly, we can prove

$$P(A \cap B) = P(A).P(B \mid A)$$

Note: If A and B are independent events, then

$$P(A \mid B) = P(A) \text{ and } P(B \mid A) = P(B)$$

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

Theorem 2: Two events A and B of the sample space S are independent, if and only if

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

Proof: If A and B are independent events,

then $P(A \mid B) = P(A)$

We know that $P(A \mid B) = P(A \cap B) / P(B)$

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

Hence, if A and B are independent events, then the probability of 'A and B' is equal to the product of the probability of A and probability of B.

Conversely, if $P(A \cap B) = P(A).P(B)$, then

$$P(A \mid B)=\frac{P(A\cap B)}{P(B)} \text{ gives}$$

$$P(A \mid B)=\frac{P(A)P(B)}{P(B)}=P(A)$$

That is, A and B are independent events.

NOTES

BAYE'S THEOREM

Bayes' Theorem is a theorem of probability theory originally stated by the Reverend Thomas Bayes. It can be seen as a way of understanding how the probability that a theory is true is affected by a new piece of evidence. It has been used in a wide variety of contexts, ranging from marine biology to the development of "Bayesian" spam blockers for email systems. In the philosophy of science, it has been used to try to clarify the relationship between theory and evidence. Many insights in the philosophy of science involving confirmation, falsification, the relation between science and pseudoscience, and other topics can be made more precise, and sometimes extended or corrected, by using Bayes' Theorem.

Begin by having a look at the theorem, displayed below. Then we'll look at the notation and terminology involved.

$$P(T \setminus E) = \frac{P(E|T) \times P(T)}{P(E|T) \times P(T) + P(E|\neg T) \times P(\neg T)}$$

In this formula, T stands for a theory or hypothesis that we are interested in testing, and E represents a new piece of evidence that seems to confirm or disconfirm the theory. For any proposition S, we will use P(S) to stand for our degree of belief, or "subjective probability," that S is true. In particular, P (T) represents our best estimate of the probability of the theory we are considering, prior to consideration of the new piece of evidence. It is known as the *prior probability* of T.

What we want to discover is the probability that T is true supposing that our new piece of evidence is true. This is a *conditional probability*, the probability that one proposition is true provided that another proposition is true. For instance, suppose you draw a card from a deck of 52, without showing it to me. Assuming the deck has been well shuffled, I should believe that the probability that the card is a jack, P(J), is 4/52, or 1/13, since there are four jacks in the deck. But now suppose you tell me that the card is a face card. The probability that the card is a jack, given that it is a face card, is 4/12, or 1/3, since there are 12 face cards in the deck. We represent this conditional probability as P(J|F), meaning the probability that the card is a jack *given that* it is a face card.

BINOMIAL DISTRIBUTION

The binomial distribution is applicable for counting the number of out-comes of a given type from a prespecied number n independent trials, each with two possible outcomes, and the same probability of the outcome of interest, p. The distribution is completely determined by n and p.

The Binomial distribution is given by

$$p(r; N, p) = \binom{N}{r} p^r (1-p)^{N-r}$$

where the variable r with $0 \leq r \leq N$ and the parameter N (N > 0) are integers and the parameter p $(0 \leq p \leq 1)$ is a real quantity.

The distribution describes the probability of exactly r successes in N trials if the probability of a success in a single trial is p (we sometimes also use q = 1 ? p, the probability for a failure, for convenience). It was first presented by Jacques Bernoulli in a work which was posthumously published.

NOTES

NORMAL DISTRIBUTION

A normal distribution is a continuous distribution that is "bell-shaped". Data are often assumed to be normal. Normal distributions can estimate probabilities over a continuous interval of data values.

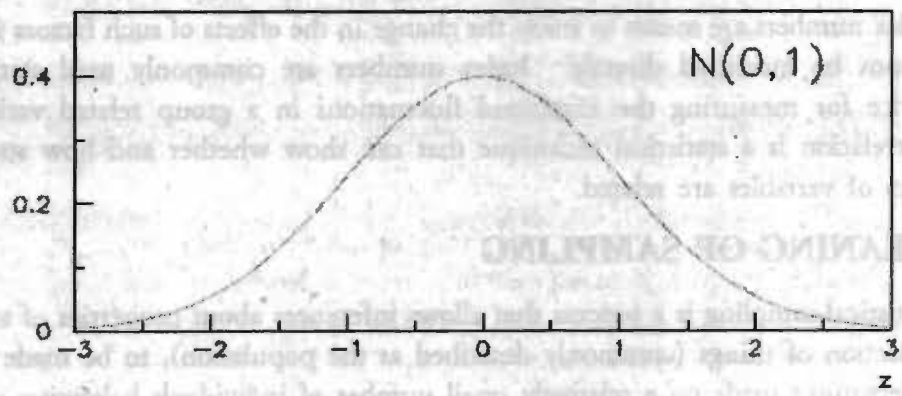
In a normal distribution, data are most likely to be at the mean. Data are less likely to farther away from the mean. Are the people around more likely to be short, tall, or average in height?

All normal distributions can be converted into a standard normal distribution. A standard normal distribution is a normal distribution with a mean=0 and standard deviation = 1.

The normal distribution or, as it is often called, the Gauss distribution is the most important distribution in statistics. The distribution is given by

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where ? is a location parameter, equal to the mean, and ? the standard deviation. For ? = 0 and ? = 1 we refer to this distribution as the standard normal distribution. In many connections it is sufficient to use this simpler form since ? and ? simply may be regarded as a shift and scale parameter, respectively. In figure we show the standard normal distribution.



POISSON DISTRIBUTION

The Poisson distribution is given by

$$p(r;\mu) = \frac{\mu^r e^{-\mu}}{r!}$$

NOTES

where the variable r is an integer $(r \geq 0)$ and the parameter ? is a real positive quantity.

It is named after the french mathematician Simeon Denis Poisson (1781–1840) who was the first to present this distribution in 1837 (implicitly the distribution was known already in the beginning of the 18th century).

As is easily seen by comparing two subsequent r-values the distribution increases up to $r + 1 < \mu$ and then declines to zero. For low values of μ it is very skewed (for $\mu < 1$ it is J-shaped).

The Poisson distribution describes the probability to find exactly r events in a given length of time if the events occur independently at a constant rate μ. An unbiased and efficient estimator of the Poisson parameter μ for a sample with n observations x_i is $\hat{\mu} = x$, the sample mean, with variance

For $\mu \to \infty$ the distribution tends to a normal distribution with mean μ and variance μ.

The Poisson distribution is one of the most important distributions in statistics with many applications.

SAMPLING

Sampling may be defined as the selection of some part of an aggregate or totality on the basis of which a judgment or inference about the aggregate or totality is made. In other words, it is the process of obtaining information about an entire population by examining only a part of it. In most of the research work and surveys, the usual approach happens to be to make generalization or to draw inferences based on samples about the parameters of population from which the samples are taken.

Index numbers are meant to study the change in the effects of such factors which cannot be measured directly. Index numbers are commonly used statistical device for measuring the combined fluctuations in a group related variables. Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

MEANING OF SAMPLING

Statistical sampling is a process that allows inferences about properties of a large collection of things (commonly described as the population), to be made from observations made on a relatively small number of individuals belonging to the population (the sample). Statistical sampling is conceptually different from the activity of merely collecting individual samples, or specimens. In the latter case, specimens can be collected and measured to describe characteristics of those specimens only, with little or no ability to generalize to the population. In conducting statistical

sampling, one is attempting to make inferences to the population. The use of valid statistical sampling techniques increases the chance that a set of specimens (the sample, in the collective sense) is collected in a manner that is representative of the population. Statistical sampling also allows a quantification of the precision with which inferences or conclusions can be drawn about the population.

Sample Design:

In sample studies, we have to make a plan regarding the size of the sample, selection of the sample, collection of the sample data and preparation of the final results based on the sample study. The whole procedure involved is called the sample design. The term sample survey is used for a detailed study of the sample. In general, the term sample survey is used for any study conducted on the sample taken from some real world data.

Sampling Frame:

A complete list of all the units of the population is called the sampling frame. A unit of population is a relative term. If all the workers in a factory make a population, a single worker is a unit of the population. If all the factories in a country are being studied for some purpose, a single factory is a unit of the population of factories. The sampling frame contains all the units of the population. It is to be defined clearly as to which units are to be included in the frame. The frame provides a base for the selection of the sample.

Advantages of Sampling:

Sampling has some advantages over the complete count. These are:

1) **Need for Sampling:** Sometimes there is a need for sampling. Suppose we want to inspect the eggs, the bullets, the missiles and the tires of some firm. The study may be such that the objects are destroyed during the process of inspection. Obviously, we cannot afford to destroy all the eggs and the bullets etc. We have to take care that the wastage should be minimum. This is possible only in sample study. Thus sampling is essential when the units under study are destroyed.

2) **Saves Time and Cost:** As the size of the sample is small as compared to the population, the time and cost involved on sample study are much less than the complete counts. For complete count huge funds are required. There is always the problem of finances. A small sample can be studied in a limited time and total cost of sample study is very small. For complete count, we need a big team of supervisors and enumeration who are to be trained and they are to be paid properly for the work they do. Thus the sample study requires less time and less cost.

3) **Reliability:** If we collect the information about all the units of population, the collected information may be true. But we are never sure about it. We do not know whether the information is true or is completely false. Thus we cannot say anything with confidence about the quality of information. We say that the reliability is not possible. This is a very important

NOTES

advantage of sampling. The inference about the population parameters is possible only when the sample data is collected from the selected sample.

4) Sometimes the experiments are done on sample basis. The fertilizers, the seeds and the medicines are initially tested on samples and if found useful, then they are applied on large scale. Most of the research work is done on the samples.

5) Sample data is also used to check the accuracy of the census data.

Limitations of Sampling

Sometimes the information about each and every unit of the population is required. This is possible only through the complete enumeration because the sample will not serve the purpose. Some examples in which the sampling is not allowed are:

i) To conduct the elections, we need a complete list of the votes. The candidates participating in the election will not accept the results prepared from a sample. With increase in literacy, the people may become statistical minded and they may become willing to accept the results prepared from the sample. In advanced countries the opinion polls are frequently conducted and unofficially the people accept the results of sample survey.

ii) Tax is collected from all the tax payers. A complete list of all the tax payers is required. The telephone, gas and electricity bills are sent to all the consumers. A complete list of the owners of land and property is always prepared to maintain the records. The position of stocks in factories requires complete entries of all the items in the stock.

Equal Probability:

The term equal probability is frequently used in the theory of sampling. This term is quite often not understood correctly. It is thought to be close to 'equal' in meaning. It is not true always. Suppose there is a population of 50 (N= 50) students in a class. We select any one student. Every student has probability 1/50 of being selected. Then a second student is selected. Now, there are 49 students in the population and every student has 1/49 probability of being selected. When the first student is selected, all the students have equal (1/50) chance of selection and when the second student is selected, again all the students have equal (1/49) chance of selection. But 1/50 is not equal to 1/49. Thus equal probability of selection means the probability when the individual is selected from the remaining available units in the population. At the time of selecting a unit, the probability of selection is equal. It is called equal probability of selection.

Known Probability:

In sampling theory the term known probability is used in random (probability) sampling. Let us explain it by taking an example. Suppose there are 300 workers in a certain factory out of which 200 are skilled and 100 are non-skilled. We have to select one sample (sub-sample) out of skilled workers and one sample out of unskilled workers. When the first worker out of skilled workers is selected, each

worker has a probability of selection equal to 1/200. Similarly when the first worker out of un-skilled workers is selected, each worker has a probability of selection equal to 1/100. Both these probabilities are known,though they are not equal.

Non-Zero Probability:

Suppose we have a population of 500 students out of which 50 are non-intelligent. We have decided to select an intelligent student from the population. The probability of selecting an intelligent student is 1/450 which is non-zero. In this example, we have decided to exclude the non-intelligent students from the population for the purpose of selecting a sample. Thus probability of selecting a non-intelligent student is zero.

Probability and Non Probability Sampling:

The term probability sampling is used when the selection of the sample is purely based on chance. The human mind has no control on the selection or non-selectionof the units forthe sample. Every unit of the population has known nonzero probability of being selected for the sample. The probability ofselectionmay b equal or unequal but it should be non-zero and should be known. The probability sampling is also called the random sampling (not simple random sampling). Some examples of random sampling are:

1. Simple random sampling.
2. Stratified random sampling.
3. Systematic random sampling.

Innon-probabilitysampling,thesampleisnotbasedonchance.Itisratherdetermined by some person. We cannot assign to an element of population the probability of its being selected in the sample. Somebody may use his personal judgment in the selection of the sample. In this case the sampling is called judgment sampling. A drawback in non-probability sampling is that such a sample cannot be used to determine the error. Any statistical method cannot be used to draw inference from this sample. But it should be remembered that judgment sampling becomes essentialin some situations. Suppose we have to take a small sample from a big heap of coal. We cannot make a list of all the pieces of coal. The upper part of the heap will have perhaps big pieces of coal. We have to use our judgment in selecting a sample to have an idea about the quality of coal. The non- probability sampling is also called non-random sampling.

Sampling with replacement:

Sampling is called with replacement when a unit selected at random from the population is returned to the population and then a second element is selected at random. Whenever a unit is selected, the population contains all the same units. A unit may be selected more than once. There is no change at all in the size of the population at any stage. We can assume that a sample of any size can be selected from the given population of any size. This is only a theoretical concept and in practical situations the sample is not selected by using this scheme of selection. Suppose the population size N = 5 and sample size n = 2, and sampling is done

NOTES

with replacement. Out of 5 elements, the first element can be selected in 5 ways. The selected unit is returned to the main lot and now the second unit can also be selected in 5 ways. Thus in total there are 5×5 = 25 samples or pairs which are possible. Suppose a container contains 3 good bulbs denoted by G_1,G_2 and G_3 and 2 defective bulbs denoted by D_1 and D_2. If any two bulbs are selected with replacement, there are 25 possible samples listed between in table.

	G_1	G_2	G_3	D_1	D_2
G_1	G_1G_1	G_1G_2	G_1G_3	G_1D_1	G_1D_2
G_2	G_2G_1	G_2G_2	G_2G_3	G_2D_1	G_2D_2
G_3	G_2G_1	G_3G_2	G_3G_3	G_3D_1	G_3D_2
D_1	D_1G_1	D_1G_2	D_1G_3	D_1D_1	D_1D_2
D_2	D_2G_1	D_2G_2	D_2G_3	D_2D_1	D_2D_2

Sampling without replacement:

Sampling is called without replacement when a unit is selected at random from the population and it is not returned to the main lot. First unit is selected out of a population of size N and the second unit is selected out of the remaining population of N – 1 units and so on. Thus the size of the population goes on decreasing as the sample size n increases. The sample size n cannot exceed the population size N. The unit once selected for a sample cannot be repeated in the same sample. Thus all the units of the sample are distinct from one another. A sample without replacement can be selected either by using the idea of permutations or combinations. Depending upon the situation, we write all possible permutations or combinations. If the different arrangements of the units are to be considered, then the permutations (arrangements) are written to get all possible samples. If the arrangement of units is of no interest, we write the combinations to get all possible samples.

Combination:

Let us again consider a lot (population) of 5 bulbs with 3 good (G_1,G_2 and G_3) and 2 defective (G_1 and G_2) bulbs. Suppose we have to select two bulbs in any order there are ${}^5C_2 = \frac{5!}{2!3!} = 10$ possible combinations or samples. These combinations (samples) are listed as:

G_1G_2, G_1G_3, G_2G_3, G_1D_1, G_1D_2, G_2D_1, G_2D_2, G_3D_1, G_3D_2, D_1D_2.

There are 10 possible samples and each of them has probability of selection equal to 1/10. The selected sample will be any one of these 10 samples. The sample selected in this manner is also called simple random sample. In general, the number of samples by combinations is equal to ${}^NC_n = \frac{N!}{n!(N-n)!}$.

Permutation:

Each combination generates a number of arrangements (permutations). Thus in general the number of permutations is greater than the number of combinations. In the previous example of bulbs, if the order of the selected bulbs is to be considered then the number of samples by permutations is given by ${}^5P_2 = \frac{5!}{(5-2)!} = 20$.

Thesesamplesare:

G_1G_2	G_2G_1	G_1G_3	G_3G_1	G_2G_3	G_3G_2	G_1D_1	D_1G_1	G_1D_2	D_2G_1
G_2D_1	D_1G_2	G_2D_2	D_2G_2	G_3D_1	D_1G_3	G_3D_2	D_2G_3	D_1D_2	D_2D_1

Eachsamplehasprobabilityofselectionequalto1/20.Theselectedsamplekeeping in view the order of the bulbs will be any one of these 20 samples. Asampleselectedin this manner is also called simple randomsamplebecause eachsamplehas equal probability of being selected.

Simple random Sampling:

Simple random sample (SRS) is a special case of a random sample. A sample is called simple random sample if each unit of the population has an equal chance of being selected for the sample. Whenever a unit is selected for the sample, the units of the population are equally likely to be selected. It must be noted that the probability of selecting the first element is not to be compared with the probability of selecting the second unit. When the first unit is selected, all the units of the population have the equal chance of selection which is $1/N$. When the second unit is selected, all the remaining $(N\ 1)$ units of the population have $1/(N\ 1)$ chance of selection.

Another way of defining a simple random sample is that if we consider all possible samples of size n, then each possible sample has equal probability of being selected.

If sampling is done with replacement, there are N^n possible samples and each sample has probability of selection equal to $\frac{1}{N^n}$. If sampling is done withoutreplacementwith thehelpofcombinationsthenthereare ${}^NC_n$ possible samples and each sample has probability of selection equal to $\frac{1}{{}^NC_n}$. If samples are made with permutations, each sample has probability of selection equal to $\frac{1}{{}^NP_n}$ Strictly speaking, the sample selected by without replacement is called simple random sample.

Difference between Random Sample and Sample Random Sample:

If each unit of the population has known (equal or un-equal) probability of

NOTES

selection in the sample, the sample is called a random sample. If each unit of the population has equal probability of being selected for the sample, the sample obtained is called simple random sample.

Selection of Sample Random Sample:

A simple random sample is usually selected by without replacement. The following methods are used for the selection of a simple random sample:

- **Lottery Method.** This is an old classical method but it is a powerful technique and modern methods of selection are very close to this method. All the units of the population are numbered from 1 to *N*. This is called sampling frame. These numbers are written on the small slips of paper or the small round metallic balls. The paper slips or the metallic balls should be of the same size otherwise the selected sample will not be truly random. The slips or the balls are thoroughly mixed and a slip or ball is picked up. Again the population of slips is mixed and the next unit is selected. In this manner, the number of slips equal to the sample size *n* is selected. The units of the population which appear on the selected slips make the simple random sample. This method of selection is commonly used when size of the population is small. For a large population there is a big heap of paper slips and it is difficult to mix the slips properly
- **Using a Random Number Table.** All the units of the population are numbered from 1 to *N* or from 0 to *N*–1. We consult the random number table to take a simple random sample. Suppose the size of the population is 80 and we have to select a random sample of 8 units. The units of the population are numbered from 01 to 80. We read two-digit numbers from the table of random numbers. We can take a start from any columns or rows of the table. Let us consult random number table given in this content. Two-digit numbers are taken from the table. Any number above 80 will be ignored and if any number is repeated, we shall not record it if sampling is done without replacement. Let us read the first two columns of the table. The random number from the table is 10, 37, 08, 12, 66, 31, 63 and 73. The two numbers 99 and 85 have not been recorded because the population does not contain these numbers. The units of the population whose numbers have been selected constitute the simple random sample. Let us suppose that the size of the population is 100. If the units are numbered from 001 to 100, we shall have to read 3-digit random numbers. From the first 3 columns of the random number table, the random numbers are 100, 375, 084, 990 and 128 and so on. We find that most of the numbers are above 100 and we are wasting our time while reading the table. We can avoid it by numbering the units of the population from 00 to 99. In this way, we shall read 2-digit numbers from the table. Thus if N is 100, 1000 or 10000, the numbering is done from 00 to 99, 000 to 999 or 0000 to 9999.
- **Using the Computer.** The facility of selecting a simple random sample is available on the computers. The computer is used for selecting a sample of prize-bond winners, a sample of Hajj applicants, and a sample of applicants for residential plots and for various other purposes.

SAMPLING ERRORS

NOTES

Suppose we are interested in the value of a population parameter, the true value of which is θ but is unknown. The knowledge about θ can be obtained either from a sample data or from the population data. In both cases, there is a possibility of not reaching the true value of the parameter. The difference between the calculated value (from sample data or from population data) and the true value of the parameter is called error. Thus error is something which cannot be determined accurately if the population is large and the units of the population are to be measured. Suppose we are interested to find the total production of wheat in Pakistan in a certain year. Sufficient funds and time are at our disposal and we want to get the 'true' figure about production of wheat. The maximum we can do is that we contact all the farmers and suppose all the farmers give maximum cooperation and supply the information as honestly as possible. But the information supplied by the farmers will have errors in most of the cases. Thus we may not be able to identify the 'true' figure. In spite of all efforts, we shall be in darkness. The calculated or the observed figure may be good for all practical purposes but we can never claim that a true value of the parameter has been obtained. If the study of the units is based on 'counting' may be we can get the true figure of the population parameter. There are two kinds of *errors* (i) sampling errors or random errors (ii) non-sampling errors.

Sampling Errors:

These are the errors which occur due to the nature of sampling. The sample selected from the population is one of all possible samples. Any value calculated from the sample is based on the sample data and is called sample statistic. The sample statistic may or may not be close to the population parameter. If the statistic is $\hat{\theta}$ and the true value of the population parameter is θ, then the difference $\hat{\theta}-\theta$ is called sampling error. It is important to note that a statistic is a random variable and it may take any value. A particular example of sampling error is the difference between the sample mean $\bar{X}$ and the population mean μ. Thus sampling error is also a random term. The population parameter is usually not known; therefore the sampling error is estimated from the sample data. The sampling error is due to the reason that a certain part of the population goes to the sample. Obviously, a part of the population cannot give the true picture of the properties of the population. But one should not get the impression that a sample always gives the result which is full of errors. We can design a sample and collect the sample data in a manner so that the sampling errors are reduced. The sampling errors can be reduced by the following methods: (1) by increasing the size of the sample (2) by stratification.

Reducing the Sampling Errors:

1. **By increasing the size of the sample.** The sampling error can be reduced by increasing the sample size. If the sample size n is equal to the population size N, then the sampling error is zero.

NOTES

2. **By Stratification.** When the population contains homogeneous units, a simple random sample is likely to be representative of the population. But if the population contains dissimilar units, a simple random sample may fail to be representative of all kinds of units, in the population. To improve the result of the sample, the sample design is modified. The population is divided into different groups containing similar units. These groups are called strata. From each group (stratum), a sub-sample is selected in a random manner. Thus all the groups are represented in the sample and sampling error is reduced. It is called stratified-random sampling. The size of the sub-sample from each stratum is frequently in proportion to the size of the stratum. Suppose a population consists of 1000 students out of which 600 are intelligent and 400 are non-intelligent. We are assuming here that we do have this much information about the population. A stratified sample of size n =100 is to be selected. The size of the stratum is denoted by N_1 and N_2 respectively and the size of the samples from each stratum may be denoted by n_1 and n_2. It is written as under:

Stratum No.	**Size of stratum**	**Size of sample from each stratum**
1	$N_1 = 600$	$n_1 = \frac{n \times N_1}{N} = \frac{100 \times 600}{1000} = 60$
2	$N_2 = 400$	$n_1 = \frac{n \times N_2}{N} = \frac{100 \times 400}{1000} = 40$
	$N_1 + N_2 = N = 1000$	$n_1 + n_2 = n = 100$

The size of the sample from each stratum has been calculated according to the size of the stratum. This is called proportional allocation. In the above sample design, the sampling fraction in the population is $\frac{n}{N} = \frac{100}{1000} = \frac{1}{10}$ and the sampling fraction in both the strata is also $\frac{1}{10}$. Thus this design is also called fixed sampling fraction. This modified sample &sign is frequently used in sample surveys. But this design requires some prior information about the units of the population. On the basis of this information, the population is divided into different strata. If the prior information is not available then the stratification is not applicable.

Non- sampling Errors:

There are certain sources of errors which occur both in sample survey as well as in the complete enumeration. These errors are of common nature. Suppose we study each and every unit of the population. The population parameter under study is the population mean and the true value of the parameter is μ which is unknown. We hope to get the value of μ by a complete count of all the units of the population. We get a value called 'calculated' or 'observed' value of the population mean. This observed value may be denoted

by μ_{cal}. The difference between μ_{cal} and μ(true) is called non-sampling error. Even if we study the population units under ideal conditions, there may still be the difference between the observed value of the population mean and the true value of the population mean. Non-sampling errors may occur due to many reasons. Some of them are:

NOTES

- The units of the population may not be defined properly. Suppose we have to carry out a study about skilled labor force in our country. Who is a skilled person? Some people do more than one job. Some do the secretariat jobs as well as the technical jobs. Some are skilled but they are doing the job of un-skilled worker. Thus it is important to clearly define the units of the population otherwise there will be non-sampling errors both in the population count and the sample study.
- There may be poor response on the part of respondents. The people do not supply correct information about their income, their children, their age and property etc. These errors are likely to be of high magnitude in population study than the sample study. To reduce these errors the respondents are to be persuaded.
- The things in human hand are likely to be mishandled. The enumerators may be careless or they may not be able to maintain uniformity from place to place. The data may not be collected properly from the population or from the sample. These errors are likely to be more serious in the population data than the sample data.
- Another serious error is due to 'bias'. Bias means an error on the part of the enumerator or the respondent when the data is being collected. Bias may be intentional or un-intentional. An enumerator may not be capable of reporting the correct data. If he has to report about the condition of crops in different areas after heavy rainfalls, his assessments may be biased due to lack of training or he may be inclined to give wrong reports. Bias is a serious error and cannot be reduced by increasing the sample size. Bias may be present in the sample study as well as the population study.

SAMPLING DISTRIBUTION

Suppose we have a finite population and we draw all possible simple random samples of size n by without replacement or with replacement. For each sample we calculate some statistic (sample mean $\overline{X}$ or proportion $\hat{p}$ etc.). All possible values of the statistic make a probability distribution which is called the sampling distribution. The number of all possible samples is usually very large and obviously the number of statistics (any function of the sample) will be equal to the number of sample if one and only one statistic is calculated from each sample. In fact, in practical situations, the sampling distribution has very large number of values. The shape of the sampling distribution depends upon the size of the sample and the nature of the population and the statistic which is calculated from all possible simple random samples. Some of the famous sampling distributions are:

NOTES

(1) Binomial distribution.

(2) Normal distribution.

(3) t-distribution.

(4) Chi-square distribution.

(5) F-distribution.

These distributions are called the derived distributions because they are derived from all possible samples.

Standard Error:

The standard deviation of some statistic is called the standard error of that statistic. If the statistic is $\bar{X}$, the standard deviation of all possible values of $\bar{X}$ is called standard error of $\bar{X}$ which may be written as S.E. ($\bar{X}$)or $\sigma_{\bar{x}}$. Similarly, if the sample statistic is proportion $\hat{p}$, the standard deviation of all possible, values of $\hat{p}$ is called standard error of $\hat{p}$ and is denoted by $\sigma_{\hat{p}}$ or S.E. ($\hat{p}$).

Sampling Distribution of $\bar{X}$:

The probability distribution of all possible values of $\bar{X}$ calculated from all possible simple random samples are called the sampling distribution of $\bar{X}$. In brief, we shall call it distribution of $\bar{X}$. The mean of this distribution is called expected value of $\bar{X}$ and is written as E($\bar{X}$)or $\mu_{\bar{x}}$. The standard deviation (standard error) of this distribution is denoted by S.E. ($\bar{X}$)or $\sigma_{\bar{x}}$ and the variance of $\bar{X}$ is denoted by *Var* ($\bar{X}$) or $\sigma^2_{\bar{x}}$. The distribution of $\bar{X}$ has some important properties as under:

- An important property of the distribution of $\bar{X}$ is that it is a normal distribution when the size of the sample is large. When the sample size n is more than 30, we call it a large sample size. The shape of the population distribution does not matter. The population may be normal or non-normal, the distribution of $\bar{X}$ is normal for $n>30$. But this is true when the number of samples is very large. As the distribution of random variable $\bar{X}$ is normal, $\bar{X}$ can be transformed into standard normal variable Z where $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$.

  The distribution of $\bar{X}$ has the t-distribution when the population is normal and $n \leq 30$. Diagram (a) shows the normal distribution and diagram (b) shows the t-distribution.

NOTES

Diagram (a)

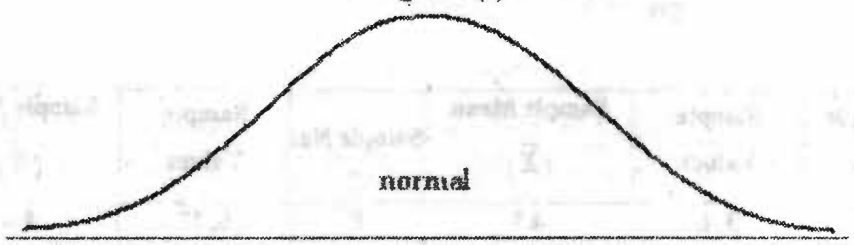
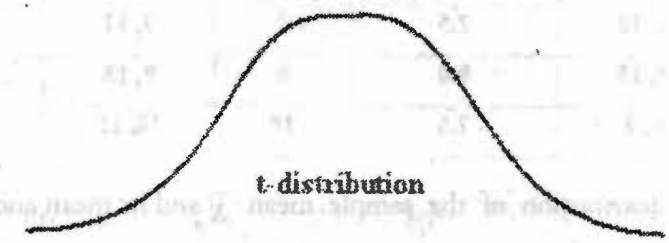


Diagram (b)



- The mean of the distribution of $\overline{X}$ is equal to the mean of the population. Thus $E(\overline{X}) = \mu_{\bar{x}} = \mu$ (Population mean). This relation is true for small as well as large sample size in sampling without replacement and with replacement.
- The standard error (standard deviation) of $\overline{X}$ is related with the standard deviation of population σ through the relations:

$$\text{S.E. } (\overline{X}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This is true when population is infinite which means N is very large or the sampling is done with replacement from finite or infinite population.

$$\text{S.E. } (\overline{X}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

This is true when sampling is without replacement from finite population. The above two equations between $\sigma_{\bar{x}}$ and σ are true both for small as well as large sample sizes.

Example:

Draw all possible samples of size **2** without replacement from a population consisting of **3, 6, 9, 12, and 15**. Form the sample distribution of sample means and verify the results.

Solution:

We have population values **3, 6, 9, 12, 15**, population size N = 5 and sample size n = 2. Thus, the number of possible samples which can be drawn without replacement is

NOTES

$$\binom{N}{n}=\binom{5}{2}=10$$

Sample No.	Sample Values	Sample Mean ($\bar{X}$)	Sample No.	Sample Values	Sample Mean ($\bar{X}$)
1	3, 6	4.5	6	6, 12	9.0
2	3, 9	6.0	7	6, 15	10.5
3	3, 12	7.5	8	9, 12	10.5
4	3, 15	9.0	9	9, 15	12.0
5	6, 9	7.5	10	12, 15	13.5

The sampling distribution of the sample mean $\bar{X}$ and its mean and standard deviation are:

$\bar{X}$	f	$f(\bar{X})$	$\bar{X}f(\bar{X})$	$\bar{X}^2f(\bar{X})$
4.5	1	1/10	4.5/10	20.25/10
6.0	1	1/10	6.0/10	36.00/10
7.5	2	2/10	15.0/10	112.50/10
9.0	2	2/10	18.0/10	162.00/10
10.5	2	2/10	21.0/10	220.50/10
12.0	1	1/10	12.0/10	144.00/10
13.5	1	1/10	13.5/10	182.25/10
Total	10	1	90/10	877.5/10

$$E(\bar{X})=\Sigma\bar{X}f(\bar{X})=\frac{90}{10}=9$$

$$\text{Var}(\bar{X})=\Sigma\bar{X}^2f(\bar{X})-\left[\Sigma\bar{X}f(\bar{X})\right]^2=\frac{887.5}{10}-\left(\frac{90}{10}\right)^2=87.75-81=6.75$$

The mean and variance of the population are:

X	3	6	9	12	15	$\Sigma X = 45$
X^2	9	36	81	144	225	$\Sigma X^2 = 495$

$$\mu=\frac{\Sigma X}{N}=\frac{45}{5}=9 \text{ and}$$

$$\sigma^2=\frac{\Sigma X^2}{N}-\left(\frac{\Sigma X}{N}\right)^2=\frac{495}{5}-\left(\frac{45}{5}\right)^2=99-81=18$$

Verification:

(i) $E(\bar{X}) = \mu = 9$

(ii) $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{18}{2}\left(\frac{5-2}{5-1}\right) = 6.75$

NOTES

Example:

If random samples of size three and drawn without replacement from the population consisting of four numbers **4, 5, 5, 7**. Find sample mean $\bar{X}$ for each sample and make sampling distribution of $\bar{X}$. Calculate the mean and standard deviation of this sampling distribution. Compare your calculations with population parameters.

Solution:

We have population values **4, 5, 5, 7,** population size $N = 4$ and sample size $n = 3$. Thus, the number of possible samples which can be drawn without replacement is

$$\binom{N}{n} = \binom{4}{3} = 4$$

		$(\bar{X})$
1	4, 5, 5	14/3
2	4, 5, 7	16/3
3	4, 5, 7	16/3
4	5, 5, 7	17/3

The sampling distribution of the sample mean $\bar{X}$ and its mean and standard deviation are:

$\bar{X}$	f	$f(\bar{X})$	$\bar{X}f(\bar{X})$	$\bar{X}^2 f(\bar{X})$
14/3	1	¼	14/12	196/36
16/3	2	2/4	32/12	512/36
17/3	1	¼	17/12	289/36
Total	4	1	63/12	997/36

$$\mu_{\bar{X}} = \Sigma \bar{X} f(\bar{X}) = \frac{63}{12} = 5.25$$

$$\sigma_{\bar{X}} = \sqrt{\Sigma \bar{X}^2 f(\bar{X}) - \left[\Sigma \bar{X} f(\bar{X})\right]^2} = \sqrt{\frac{997}{36} - \left(\frac{63}{12}\right)^2} = 0.3632$$

Check your progress:

1. What is sample space?
2. What are mutually exclusive events?
3. Define Sampling.

NOTES

The mean and standard deviation of the population are:

X	4	5	5	7	$\Sigma X = 21$
X^2	16	25	25	49	$\Sigma X^2 = 115$

$$\mu = \frac{\Sigma X}{N} = \frac{21}{4} = 5.25 \text{ and}$$

$$\sigma^2 = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{115}{4} - \left(\frac{21}{4}\right)^2} = 1.0897$$

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{1.0897}{\sqrt{3}}\sqrt{\frac{4-3}{4-1}} = 0.3632$$

Hence $\mu_X = \mu$ and $\sigma_X = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

SUMMARY

- A complete list of all possible outcomes of a random experiment is called sample space or possibility space and is denoted by *S*.
- Two events are called mutually exclusive or disjoint if they do not have any outcome common between them.
- **Addition Law of Probability:** For any two events A and B of a sample space S

  P (A or B) = P (A) + P (B) - P (A and B)
- **Additive Law of Probability for Mutually Exclusive Events:** If A and B are two mutually exclusive events, then**Probability**

  P (A or B) = P (A U B) = P (A) + P (B).
- **Odds in Favour of an Event:** If the odds for A are a to b, then

  P(A) = a / a + b

  If odds against A are a to b, then

  P(A) = b / a + b
- Two events A and B are said to be independent, if the occurrence or non-occurrence of one does not affect the probability of the occurrence (and hence non-occurrence) of the other.

  If A and B are independent events, then

  P (A and B) = P (A). P (B)

  or P(A $\cap$ B) = P (A) . P (B)

NOTES

- For two events A and B,

$$P(A \cap B) = P(A)\, P(B \mid A), \quad P(A) > 0$$

$$\text{or } P(A \cap B) = P(B)\, P(A \mid B), \quad P(B) > 0$$

where P (B|A) represents the conditional probability of occurrence of B, when the event A has already happened and P (A|B) represents the conditional probability of happening of A, given that B has already happened.

- Sampling may be defined as the selection of some part of an aggregate or totality on the basis of which a judgment or inference about the aggregate or totality is made.
- Statistical sampling is a process that allows inferences about properties of a large collection of things (commonly described as the population), to be made from observations made on a relatively small number of individuals belonging to the population (the sample).
- In sample studies, we have to make a plan regarding the size of the sample, selection of the sample, collection of the sample data and preparation of the final results based on the sample study. The whole procedure involved is called the sample design. The term sample survey is used for a detailed study of the sample.
- A complete list of all the units of the population is called the sampling frame. A unit of population is a relative term. If all the workers in a factory make a population, a single worker is a unit of the population. If all the factories in a country are being studied for some purpose, a single factory is a unit of the population of factories. The sampling frame contains all the units of the population. It is to be defined clearly as to which units are to be included in the frame. The frame provides a base for the selection of the sample.

Answer To Check Your Progress

1. A complete list of all possible outcomes of a random experiment is called sample space or possibility space and is denoted by *S*.
2. Two events are called mutually exclusive or disjoint if they do not have any outcome common between them.
3. Sampling may be defined as the selection of some part of an aggregate or totality on the basis of which a judgment or inference about the aggregate or totality is made.

Test Yourself

1. A die is rolled once. Find the probability of getting 3.
2. A coin is tossed once. What is the probability of getting the tail ?

NOTES

3. What is the probability of the die coming up with a number greater than 3 ?
4. In a simultaneous toss of two coins, find the probability of getting ' at least' one tail.
5. From a bag containing 15 red and 10 blue balls, a ball is drawn 'at random'. What is the probability of drawing (i) a red ball ? (ii) a blue ball ?
6. If two dice are thrown, what is the probability that the sum is (i) 6 ? (ii) 8? (iii) 10? (iv) 12?
7. If two dice are thrown, what is the probability that the sum of the numbers on the two faces is divisible by 3 or by 4?
8. If two dice are thrown, what is the probability that the sum of the numbers on the two faces is greater than 10?
9. What is the probability of getting a red card from a well shuffled deck of 52 cards?
10. If a card is selected from a well shuffled deck of 52 cards, what is the probability of drawing (i) a spade ? (ii) a king ? (iii) a king of spade?
11. A pair of dice is thrown. Find the probability of getting (i) a sum as a prime number (ii) a doublet, i.e., the same number on both dice (iii) a multiple of 2 on one die and a multiple of 3 on the other.
12. Three coins are tossed simultaneously. Find the probability of getting (i) no head (ii) at least one head (iii) all heads
13. Write a short note on Baye's Theorem.
14. What do you mean by the term 'Sampling'? Give its advantages and limitations.
15. Explain the concept of Sampling Distribution.